SIAM J. CONTROL
Vol. 7, No. 1, February 1969

VARIATIONAL PROBLEMS INVOLVING FUNCTIONAL DIFFERENTIAL EQUATIONS*

H. T. BANKS†

Introduction. In recent years an interest in variational problems or optimal control problems involving delayed systems has arisen. In particular, a number of papers have been written on problems involving systems with a time lag in the state variable. More general cases where the system has some type of functional dependency have also been investigated. Existence of optimal solutions for such problems has been discussed in several works [2], [6], [16]. The purpose of this paper is to obtain necessary conditions (in the form of an integrated maximum principle) for problems with quite general nonlinear functional differential systems. These systems will include as special cases many integro-differential systems and time lag (variable or constant) systems.

In this paper integrals will be understood to be Lebesgue or Lebesgue–Stieltjes integrals. Similarly, when speaking of a measurable function, we shall mean a Lebesgue measurable function unless it is specifically stated otherwise. By a solution of a (functional) differential equation will be meant an absolutely continuous (A.C.) function which satisfies the equation almost everywhere with respect to Lebesgue measure. Vector matrix notation will be employed throughout and we shall not distinguish between a vector and its transpose when it is clear what is meant. The notation $|A|$ will denote the Euclidean norm of A in whatever space A lies.

In § 1, we shall formulate a control problem for functional differential equation systems and state a maximum principle. In § 2, results needed in the proof of this principle will be given. The proof of the maximum principle will be presented in § 3. The proof involves a consideration of general extremal problems and uses a generalization of the idea of quasi-convex families due to Gamkrelidze [7]. (See also [15].) Finally, in § 4 we shall discuss several examples of systems to which our results are applicable.

1. Notation, formulation of problem, and a maximum principle.

Throughout this paper, we shall assume that t_0 and α_0 are fixed in R^1 with $-\infty < \alpha_0 < t_0$. Let $I = [\alpha_0, a)$ be a bounded interval containing $[\alpha_0, t_0]$ and put $I' = (t_0, a)$. If $\mathscr{G}$ is an open convex region in R^l (possibly all of R^l), we shall denote by $C(I, \mathscr{G})$ the space of bounded continuous l-vector functions on I into $\mathscr{G}$ with the uniform topology. That is, $C(I, \mathscr{G})$ will be considered as a (topological) subspace of $C(I, R^l)$.

* Received by the editors December 21, 1967, and in revised form May 22, 1968.

† Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. This research was supported in part by the National Aeronautics and Space Administration under Grant 40-002-015 and in part by the United States Air Force under Grant AF-AFOSR-693-67.

For any set X contained in $\mathscr{G}$, define $AC(I, X)$ to be the subset of $C(I, \mathscr{G})$ consisting of all bounded absolutely continuous l-vector functions on I into X.

In the discussion below, we shall be considering n-vector functionals $F(x(\cdot), t)$, where $F: C(I, \mathscr{G}) \times I' \to R^n$, with $\mathscr{G}$ fixed. By the notation $F(x(\cdot), t)$, we shall mean that for each fixed t in I',

$$F(\cdot, t): C([\alpha_0, t], \mathscr{G}) \to R^n,$$

so that $F(x(\cdot), t)$ may depend on any or all of the values $x(\tau), \alpha_0 \leqq \tau \leqq t$. If, for each $t \in I'$, F has a Fréchet differential dF (see [5, p. 92]) with respect to x, then $dF[x(\cdot), t; \cdot]$ is a bounded linear map from $C([\alpha_0, t], R^l)$ into R^n. We shall then write $\|dF[x(\cdot), t; \cdot]\| \leqq m(t)$ to mean $|dF[x(\cdot), t; \psi]| \leqq m(t)\|\psi\|_t$ for each $\psi \in C([\alpha_0, t], R^l)$ and $t \in I'$, where

$$\|\psi\|_t = \sup\{|\psi(s)| : s \in [\alpha_0, t]\}.$$

Note. In the discussions below, we shall not always distinguish between $C(I, \mathscr{G})$ and $C([\alpha_0, t], \mathscr{G})$. For example, instead of saying that for each t, $F(x(\cdot), t)$ is C^1 w.r.t. x in $C([\alpha_0, t], \mathscr{G})$, we shall say simply that F is C^1 in x on $C(I, \mathscr{G})$. It will be clear what is meant. In fact, given any continuous x defined on $[\alpha_0, \tau]$ (contained in I) into $\mathscr{G}$, then x may be considered as an element of $C(I, \mathscr{G})$ by the convention $x(t) \equiv x(\tau)$ for $t \geqq \tau$. Conversely, any x in $C(I, \mathscr{G})$ is also in $C([\alpha_0, t], \mathscr{G})$ for each $t \in I'$.

Having introduced the notation discussed above, we shall consider the following optimal control problem:

Minimize $J[\bar{\phi}, u, \bar{x}, t_1] = \int_{t_0}^{t_1} f^0(\bar{x}(\cdot), u(t), t)\, dt$ over $\bar{\Phi} \times \Omega \times C(I, R^{n-1}) \times I'$ subject to

$$\text{(i)}\quad \dot{\bar{x}}(t) = \bar{f}(\bar{x}(\cdot), u(t), t), \qquad t \in [t_0, t_1],$$
$$\bar{x}(t) = \bar{\phi}(t), \qquad t \in [\alpha_0, t_0],$$
$$\text{(ii)}\quad (\bar{x}(t_0), \bar{x}(t_1), t_1) \in \mathscr{T}.$$

The following assumptions and definitions are made: $\bar{\mathscr{G}}$ is a fixed open convex region in R^{n-1}, $\bar{x}$ is an $(n-1)$-vector function, $f = (f^0, \bar{f}) = (f^0, f^1, \cdots, f^{n-1})$ is an n-vector function defined on $C(I, \bar{\mathscr{G}}) \times \mathscr{U} \times I'$, where $\mathscr{U} \subset R^r$. Each f^i is assumed C^1 in $\bar{x}$ and Borel measurable in u, t. $\bar{\Phi}$ and Ω are defined by $\bar{\Phi} = AC([\alpha_0, t_0], \bar{\mathscr{G}})$ and $\Omega = \{u : u$ is measurable on I' and $u(t) \in U(t)$ for $t \in I'\}$, where U is a given mapping of I' into subsets of $\mathscr{U}$. $\mathscr{T}$ is a given C^1 manifold in R^{2n-1} of dimension less than $2n - 1$ satisfying $\mathscr{T} \subset \bar{\mathscr{G}} \times \bar{\mathscr{G}} \times I'$.

We also assume that given $\bar{X}$ compact, $\bar{X} \subset \bar{\mathscr{G}}$, and $u \in \Omega$, there exists an m in $L_1(I')$ such that

$$|f(\bar{x}(\cdot), u(t), t)| \leqq m(t),$$

$$\|df[\bar{x}(\cdot), u(t), t; \cdot]\| \leqq m(t)$$

for each $t \in I'$, and $\bar{x} \in AC(I, \bar{X})$, where df is the Fréchet differential of f w.r.t. $\bar{x}$.

Under the above assumptions, we have the following necessary conditions for $(\bar{\phi}^*, u^*, \bar{x}^*, t_1^*)$ to be optimal (i.e., a solution of the above problem).

THEOREM 1. *Let $(\bar{\phi}^*, u^*, \bar{x}^*, t_1^*)$ be optimal for the control problem with general functional system equations. Suppose that t_1^* is a regular point of $f(\bar{x}^*(\cdot), u^*(t), t)$. Then there exists a nontrivial n-vector function $\lambda(t) = (\lambda^0(t), \bar{\lambda}(t))$ of bounded variation on $[t_0, t_1^*]$, continuous at t_1^*, satisfying*:

(i) $\lambda^0(t) = \text{const.} \leqq 0, \lambda(t_1^*) \neq 0,$

$$\bar{\lambda}(t) + \int_t^{t_1} \lambda(\beta)\bar{\eta}^*(\beta, t)\, d\beta = \bar{\lambda}(t_1^*)$$

for $t \in [t_0, t_1^*)$, *where* $\bar{\eta}^*(t, s)$ *is an* $n \times (n-1)$ *matrix such that*

$$df[\bar{x}^*(\cdot), u^*(t), t; \bar{\psi}] = \int_{\alpha_0}^{t} d_s\bar{\eta}^*(t, s)\bar{\psi}(s)$$

for all $\bar{\psi} \in C([\alpha_0, t], R^{n-1})$ *and each* $t \in (t_0, t_1^*]$;

(ii) $\displaystyle\int_{t_0}^{t_1} \lambda(t)f(\bar{x}^*(\cdot), u^*(t), t)\, dt \geqq \int_{t_0}^{t_1} \lambda(t)f(\bar{x}^*(\cdot), u(t), t)\, dt$ *for all* $u \in \Omega$;

(iii) *the* $(2n-1)$*-vector*

$$\left(-\bar{\lambda}(t_0) + \int_{t_0}^{t_1^*} \lambda(\beta)\{\bar{\eta}^*(\beta, \alpha_0) - \bar{\eta}^*(\beta, t_0)\}\, d\beta, \bar{\lambda}(t_1^*), -\lambda(t_1^*)\cdot f^*(t_1^*)\right)$$

is orthogonal to $\mathcal{T}$ *at* $(\bar{x}^*(t_0), \bar{x}^*(t_1^*), t_1^*)$, *where* $f^*(t_1^*) = f(\bar{x}^*(\cdot), u^*(t_1^*), t_1^*)$.

The inequality in (ii) above is a maximum principle in integral form. In general, it is very difficult to use this condition to completely determine the optimal control u^*. However, an example where an integral maximum principle may be used to determine the optimal control for certain delayed systems is given in [1]. Applications of the results given in Theorem 1 above to a biological model will be discussed in a future paper.

Under additional hypotheses, one can obtain a pointwise maximum principle from (ii) above. For example, in the case that $f(\bar{x}(\cdot), u, t)$ is continuous in all arguments and the mapping $U(t)$ is such that $U(t) = \mathbf{U}$ for $t \in I'$, where $\mathbf{U}$ is a fixed subset of R^r, then one can show that (ii) implies a Pontryagin-type maximum principle. That is,

$$\lambda(t)\cdot f(\bar{x}^*(\cdot), u^*(t), t) = \sup\{\lambda(t)\cdot f(\bar{x}^*(\cdot), u, t) : u \in \mathbf{U}\}$$

holds almost everywhere on $[t_0, t_1^*]$.

In the next section, we give some preliminary results which will be needed to prove Theorem 1.

2. Preliminary results. The first result in this section is concerned with the existence of solutions to an integral equation. The multiplier (or adjoint) system for our control problem will be an equation of the type considered in this theorem.

THEOREM 2. *Let $\mathcal{J}'$ be a bounded open interval in R^1. Let the $n \times n$ matrix $N(\sigma, t)$ be measurable in σ on $\mathcal{J}'$ satisfying $|N(\sigma, t)| \leqq r(\sigma)$ for every t in $[0, T]$, where $(0, T] \subset \mathcal{J}'$ and $r \in L_1(\mathcal{J}')$. Let $N(\sigma, t)$ be of bounded variation on $[0, T]$ as a* *function of t, satisfying* $\bigvee_{t=0}^{T} N(\sigma, t) \leqq r(\sigma)$ *for each* σ *in* $\mathscr{I}'$. *Let* $F(z, t)$ *be defined on* $R^n \times [0, T]$ *into* R^n, *continuous in z for each t, of bounded variation in t with* $\bigvee_{t=0}^{T} F(z, t) \leqq h(z)$, *where h is a bounded measurable function on* R^n. *Furthermore, suppose there exists a bounded measurable* $\gamma(t)$ *on* $[0, T]$ *such that* $|F(z, t)| \leqq |z|\gamma(t)$. *Let* ξ *be a constant n-vector. Then the system*

$$z(0) = \xi,$$

$$z(t) + \int_0^t F(z(\sigma), t)N(\sigma, t)\, d\sigma = \xi, \qquad t \in (0, T],$$

has a solution that is of bounded variation on $[0, T]$.

Proof. Since this is a somewhat standard result, we give only a brief outline of the proof. For $k = 1, 2, \cdots$, we define the sequence of functions

$$z_k(t) = \begin{cases} \xi, & t \in [0, T/k], \\ \xi - \displaystyle\int_0^{t - T/k} F(z_k(\sigma), t)N(\sigma, t)\, d\sigma, & t \in (T/k, T]. \end{cases}$$

Then using the hypotheses on F and N, it is not difficult to show that $\{z_k\}$ is uniformly bounded on $[0, T]$ and, in fact, $\bigvee_{t=0}^{T} z_k(t)$ is uniformly bounded. It follows from a well-known theorem of Helly that $\{z_k\}$ has a convergent subsequence, which we again call $\{z_k\}$, such that $z_k(t) \to z(t)$ for every t in $[0, T]$ and the limit function z is in $BV[0, T]$. Use of the hypotheses of the theorem and dominated convergence shows that z satisfies the above system.

If, in addition to the assumptions in Theorem 2, we assume that there exists a bounded measurable function $p(t)$ on $[0, T]$ such that $|F(z_1, t) - F(z_2, t)| \leqq |z_1 - z_2|p(t)$ for all z_1, z_2 in R^n, then it is not hard to show that the system in Theorem 2 has a unique solution on $[0, T]$.

In the proof of Theorem 1, we shall be concerned with an $n \times n$ matrix function $\eta(t, s)$ which is measurable in (t, s), of bounded variation in s on $[\alpha_0, t]$ for each $t \in I'$, and satisfies $|\eta(t, s)| \leqq \tilde{m}(t)$ for each $s \in [\alpha_0, t]$, $\bigvee_{s=\alpha_0}^{s=t} \eta(t, s) \leqq \tilde{m}(t)$ for each $t \in I'$, where $\tilde{m} \in L_1(I')$. Furthermore, η will be such that $\eta(t, s) = 0$ for $s \geqq t$ and $\eta(t, s) = \eta(t, \alpha_0)$ for $s < \alpha_0$. With these hypotheses on η, one can show without difficulty that from Theorem 2 it follows that the system (for $t \in I'$, $E =$ the $n \times n$ identity matrix)

$$\Gamma(t, t) = E,$$

$$\Gamma(\sigma, t) + \int_\sigma^t \Gamma(\beta, t)\eta(\beta, \sigma)\, d\beta = E, \qquad \sigma \in [t_0, t),$$

has a unique solution $\Gamma(\sigma, t)$ which is of bounded variation in σ on $[t_0, t]$. It is easy to see that $\Gamma(\sigma, t)$ is continuous in σ at $\sigma = t$.

Our next result is a type of variation of constants formula for functional differential equations. Special cases of this representation theorem have appeared in other works, although not always in a correct form. We thus give a complete statement and proof of this more general representation result.

THEOREM 3. *Let η be as described above and let $t_1 \in I'$. Let the n-vector functions ϕ and C be given with $\phi \in AC([\alpha_0, t_0], R^n)$ and $C \in L_1(t_0, t_1)$. For each t in $(t_0, t_1]$, let $\Gamma(\sigma, t)$ be the matrix solution to*

$$\begin{aligned} &\Gamma(t,t) = E, \\ &\Gamma(\sigma,t) + \int_\sigma^t \Gamma(\beta,t)\eta(\beta,\sigma)\,d\beta = E, \qquad t_0 \leqq \sigma < t. \end{aligned} \tag{2.1}$$

Then the n-vector solution z to

$$\begin{aligned} \dot z(\sigma) &= \int_{\alpha_0}^{\sigma} d_s\eta(\sigma,s)z(s) + C(\sigma), \qquad t_0 \leqq \sigma \leqq t_1, \\ z(\sigma) &= \phi(\sigma), \qquad \alpha_0 \leqq \sigma \leqq t_0, \end{aligned} \tag{2.2}$$

is given for $t > t_0$ by

$$\begin{aligned} z(t) = \Gamma(t_0,t)\phi(t_0) &+ \int_{t_0}^{t} \Gamma(\beta,t)\,d\beta \int_{\alpha_0}^{t_0} d_s\eta(\beta,s)\phi(s) \\ &+ \int_{t_0}^{t} \Gamma(\beta,t)C(\beta)\,d\beta. \end{aligned} \tag{2.3}$$

Proof. It is not difficult to see that the existence of a solution to (2.2) is guaranteed by Theorems 1 and 2 in [1]. For $t > t_0$ we have

$$\begin{aligned} z(t) &= \Gamma(t_0,t)z(t_0) + \int_{t_0}^{t} d_\beta\{\Gamma(\beta,t)z(\beta)\} \\ &= \Gamma(t_0,t)\phi(t_0) + \int_{t_0}^{t} d_\beta\Gamma(\beta,t)z(\beta) + \int_{t_0}^{t} \Gamma(\beta,t)\dot z(\beta)\,d\beta \\ &= \Gamma(t_0,t)\phi(t_0) + \int_{t_0}^{t} d_\beta\Gamma(\beta,t)z(\beta) \\ &\quad + \int_{t_0}^{t} \Gamma(\beta,t)\left\{\int_{\alpha_0}^{\beta} d_s\eta(\beta,s)z(s) + C(\beta)\right\} d\beta \\ &= \Gamma(t_0,t)\phi(t_0) + \int_{t_0}^{t} \Gamma(\beta,t)\,d\beta \int_{\alpha_0}^{t_0} d_s\eta(\beta,s)\phi(s) \\ &\quad + \int_{t_0}^{t} \Gamma(\beta,t)C(\beta)\,d\beta + \int_{t_0}^{t} d_\beta\Gamma(\beta,t)z(\beta) \\ &\quad + \int_{t_0}^{t} \Gamma(\beta,t)\,d\beta \int_{t_0}^{\beta} d_s\eta(\beta,s)z(s). \end{aligned}$$

Consider

$$\int_{t_0}^{t} \Gamma(\beta,t)\,d\beta \int_{t_0}^{\beta} d_s\eta(\beta,s)z(s).$$

Since $\eta(\beta, s) = 0$ for $s \geqq \beta$, this integral may be written

$$\int_{t_0}^{t} \Gamma(\beta, t)\, d\beta \int_{t_0}^{t} d_s\eta(\beta, s)z(s).$$

Under the conditions which Γ, η, z satisfy, it is possible to use a slight modification of an unsymmetric Fubini-type theorem of Cameron and Martin [3] to interchange the order of integration in this integral. We obtain

$$\int_{t_0}^{t} \left\{ d_s \int_{t_0}^{t} \Gamma(\beta, t)\eta(\beta, s)\, d\beta \right\} z(s) = \int_{t_0}^{t} \left\{ d_s \int_{s}^{t} \Gamma(\beta, t)\eta(\beta, s)\, d\beta \right\} z(s)$$

since $\eta(\beta, s) = 0$ for $\beta \leqq s$. We then have

$$\begin{aligned} z(t) = \Gamma(t_0, t)\phi(t_0) &+ \int_{t_0}^{t} \Gamma(\beta, t)\, d\beta \int_{\alpha_0}^{t_0} d_s\eta(\beta, s)\phi(s) \\ &+ \int_{t_0}^{t} \Gamma(\beta, t)C(\beta)\, d\beta \\ &+ \int_{t_0}^{t} d_s\left\{\Gamma(s, t) + \int_{s}^{t} \Gamma(\beta, t)\eta(\beta, s)\, d\beta\right\} z(s). \end{aligned}$$

But the last integral vanishes since Γ satisfies (2.1). This gives the desired representation.

Before presenting the final preliminary result, we must define a special subset of $C(I, R^l)$. Recalling the definition of $AC(I, X)$ given in § 1, we define

$$AC(I, X)_K = \{x \in AC(I, X) : |\dot{x}(t)| \leqq K(t) \text{ a.e. on } I\}$$

for any nonnegative $L_1(I)$ function $K(t)$. We then have the following lemma.

Lemma 2.1. *Let X be any compact convex subset of R^l. Let K be a nonnegative $L_1(I)$ function, where I is a finite interval. Then $AC(I, X)_K$ is a compact convex subset of $C(I, R^l)$.*

Proof. The convexity follows directly from the convexity of X. Using the fact that

$$|x(\tau_1) - x(\tau_2)| \leqq \int_{\tau_1}^{\tau_2} K(t)\, dt \tag{2.4}$$

for every $x \in AC(I, X)_K$ and $\tau_1, \tau_2 \in I$, it is not difficult to show that $AC(I, X)_K$ is a closed subset of $C(I, R^l)$. That it is bounded follows from (2.4) and the fact that X is compact. Although we cannot use the Arzela-Ascoli theorem here (I is not compact), the hypotheses of a similar theorem (see [5, Theorem IV.6.5]) are satisfied. It thus follows that $AC(I, X)_K$ is conditionally compact (and hence compact since it is closed) in $C(I, R^l)$.

3. Extremal theory. In this section we shall present the proof of Theorem 1. In order to do this, we shall first formulate and give results for an abstract extremal problem. Application of this extremal theory to our control problem will yield the results given in Theorem 1. The abstract extremal problem is of interest itself since it will yield results for many control problems other than the one discussed in this paper.

Let α_0, t_0, I and I' be as defined in § 1. Let $\mathscr{G}$ be a given fixed open convex region in R^n (possibly all of R^n). Let us denote by $\mathscr{C}$ the subset $C(I, \mathscr{G})$ of $C(I, R^n)$. Denote by $\mathscr{F}$ a family of n-vector functionals $F(x(\cdot), t)$, where $F: \mathscr{C} \times I' \to R^n$. We are using here the notation and conventions explained in § 1. If k is any positive integer, we define

$$P^k = \left\{ \alpha \in R^k : \alpha^i \geqq 0 \text{ for } i = 1, \cdots, k \text{ and } \sum_1^k \alpha^i = 1 \right\}.$$

We then make the following lengthy but essential definition.

Definition 1. A family $\mathscr{F}$ is *absolutely quasi-convex* (A.Q.) if the following conditions are satisfied:

(a) Each $F(x(\cdot), t)$ in $\mathscr{F}$ is C^1 in x for fixed $t \in I'$ and measurable on I' for fixed $x \in \mathscr{C}$.

(b) Given any $F \in \mathscr{F}$ and any compact convex X contained in $\mathscr{G}$, there exists an $m \in L_1(I')$ (m depending on X, F) such that

$$|F(x(\cdot), t)| \leqq m(t),$$

$$\|dF[x(\cdot), t; \cdot]\| \leqq m(t)$$

for all $t \in I'$ and $x \in AC(I, X)$, where dF is the Fréchet differential of F w.r.t. x.

(c) For every compact convex X contained in $\mathscr{G}$, nonnegative K in $L_1(I)$, finite collection $F_1, \cdots, F_k$ in $\mathscr{F}$, and $\varepsilon > 0$, there exists for each $\alpha \in P^k$ an F_α in $\mathscr{F}$ (depending on X, K, the F_i, and ε) satisfying

$$|F_\alpha(x(\cdot), t)| \leqq \sum_1^k m_i(t),$$

$$\|dF_\alpha[x(\cdot), t; \cdot]\| \leqq \sum_1^k m_i(t)$$

for each $\alpha \in P^k$, $t \in I'$ and $x \in AC(I, X)$ (where the m_i are the $L_1(I')$ functions described in (b) above depending on X and F_i), so that

$$G(x(\cdot), t, \alpha) \equiv \sum_1^k \alpha^i F_i(x(\cdot), t) - F_\alpha(x(\cdot), t)$$

satisfies:

$$|G(x(\cdot), t, \alpha)| \leqq 2 \sum_1^k m_i(t),$$

$$\|dG[x(\cdot), t, \alpha; \cdot]\| \leqq 2 \sum_1^k m_i(t)$$

for all $x \in AC(I, X)$, $\alpha \in P^k$ and $t \in I'$. Also

$$\left| \int_{\tau_1}^{\tau_2} G(x(\cdot), t, \alpha)\, dt \right| < \varepsilon$$

for all $\alpha \in P^k$, $[\tau_1, \tau_2] \subset I'$ and $x \in AC(I, X)_K$.

(c′) If $\{\alpha_i\}_{i=1}^{\infty}$ is a sequence in P^k such that $\alpha_i \to \bar{\alpha} \in P^k$, then $\{G(x(\cdot), t, \alpha_i)\}_1^{\infty}$ converges in measure on I' to $G(x(\cdot), t, \bar{\alpha})$ for each $x \in AC(I, X)_K$.

Next let Φ be the class of A.C. n-vector functions on $[\alpha_0, t_0]$ into $\mathscr{G}$. That is, $\Phi = \{\phi : \phi \in AC([\alpha_0, t_0], \mathscr{G})\}$. For $F \in \mathscr{F}$ and $\phi \in \Phi$, we shall consider solutions to

$$(3.1)\qquad \begin{aligned} \dot{x}(t) &= F(x(\cdot), t), && t > t_0, \\ x(t) &= \phi(t), && t \in [\alpha_0, t_0]. \end{aligned}$$

If $z(t)$, $\alpha_0 \leqq t \leqq \tau_1$, is a solution to (3.1) for $(F, \phi) \in \mathscr{F} \times \Phi$, we define the $(2n+1)$-vector $q_z = (z(t_0), z(\tau_1), \tau_1)$. Let Q be the set of all such q_z for solutions to (3.1) for $(F, \phi) \in \mathscr{F} \times \Phi$.

Let $\mathscr{N}$ be a given C^1 manifold in R^{2n+1} with boundary $\mathscr{M} = \partial\mathscr{N}$. For $q \in \mathscr{M}$, let $\mathscr{N}_T(q)$ be the tangent half-plane to $\mathscr{N}$ at q and let $\mathscr{M}_T(q)$ be the tangent plane to $\mathscr{M}$ at q.

Definition 2. A solution $z(t)$, $\alpha_0 \leqq t \leqq \tau$, to (3.1) corresponding to $(F, \phi) \in \mathscr{F} \times \Phi$ is called an $\mathscr{F}, \mathscr{N}, \Phi$ *extremal* if

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

(ii) there is a neighborhood V of q_z such that $V \cap Q \cap \mathscr{N} \subset \mathscr{M}$.

Let M be an arbitrary but fixed positive function in $L_1(\alpha_0, t_0)$. Define $\delta\phi(M)$ to be the set of A.C. n-vector functions $\delta\phi$ on $[\alpha_0, t_0]$ into R^n satisfying $|\delta\phi(t)| \leqq M(t)$ a.e. on $[\alpha_0, t_0]$.

Given an $\mathscr{F}, \mathscr{N}, \Phi$ extremal $\hat{x}(t)$, $\alpha_0 \leqq t \leqq t_1$, corresponding to $(\hat{F}, \hat{\phi})$ in $\mathscr{F} \times \Phi$, we shall denote by δF the elements in $[\mathscr{F}] - \hat{F}$, where $[\mathscr{F}]$ is the convex hull of $\mathscr{F}$. That is, $\delta F = \sum_1^k \alpha^i F_i - \hat{F}$, where $\alpha \in P^k$, k arbitrary. For δF in $[\mathscr{F}] - \hat{F}$ and $\delta\phi \in \delta\Phi(M)$, let δx denote the solution to

$$(3.2)\qquad \begin{aligned} \delta\dot{x}(t) &= d\hat{F}[\hat{x}(\cdot), t; \delta x] + \delta F(\hat{x}(\cdot), t) && \text{on} \quad [t_0, t_1], \\ \delta x(t) &= \delta\phi(t) && \text{on} \quad [\alpha_0, t_0]. \end{aligned}$$

The existence of a unique solution of (3.2) is guaranteed by previous results of the author (see [1, Theorems 1 and 2]).

With the above definitions in mind, we then define the set $\hat{\mathscr{K}}$ contained in R^{2n+1} by $\hat{\mathscr{K}} = \{(\delta\phi(t_0), \delta x(t_1) + \delta t\hat{F}(\hat{x}(\cdot), t_1), \delta t) : \delta\phi \in \delta\Phi(M), \delta t \in R^1, \delta x$ is the solution to (3.2) for δF in $[\mathscr{F}] - \hat{F}$ and $\delta\phi \in \delta\Phi(M)\}$.

Theorem 4. *Suppose $\mathscr{F}$ is absolutely quasi-convex. Let $\hat{x}(t)$, $\alpha_0 \leqq t \leqq t_1$, be an $\mathscr{F}, \mathscr{N}, \Phi$ extremal corresponding to $(\hat{F}, \hat{\phi}) \in \mathscr{F} \times \Phi$. Suppose t_1 is a regular point for $\hat{F}(\hat{x}(\cdot), t)$. Then there exists a nonzero $(2n+1)$-dimensional vector ζ such that:*

(i) *ζ is orthogonal to $\mathscr{M}_T(q_{\hat{x}})$;*

(ii) *$\zeta \cdot w \geqq 0$ for all $(2n+1)$-vectors w such that $w + q_{\hat{x}} \in \mathscr{N}_T(q_{\hat{x}})$;*

(iii) *$\zeta \cdot p \leqq 0$ for all $p \in \hat{\mathscr{K}}$.*

Proof. Let M and $\delta\Phi(M)$ be as defined previously. Let $\hat{x}(t)$, $\alpha_0 \leqq t \leqq t_1$, be the $\mathscr{F}, \mathscr{N}, \Phi$ extremal corresponding to $(\hat{F}, \hat{\phi}) \in \mathscr{F} \times \Phi$. That is,

$$(3.3)\qquad \begin{aligned} \dot{\hat{x}}(t) &= \hat{F}(\hat{x}(\cdot), t), && [t_0, t_1], \\ \hat{x}(t) &= \hat{\phi}(t), && [\alpha_0, t_0]. \end{aligned}$$

Let X be a fixed compact convex subset of $\mathscr{G}$ chosen so that each $\hat{x}(t)$, $\alpha_0 \leqq t \leqq t_1$, is an interior point of X. Let $\delta F = \sum_1^k \alpha^i F_i - \hat{F}$ represent an arbitrary element of $[\mathscr{F}] - \hat{F}$. Let $\hat{m}, m_i, i = 1, \cdots, k$, be the $L_1(I')$ functions in Definition 1(b) corresponding to $\hat{F}$, $F_i, i = 1, \cdots, k$, respectively and X. Define an $L_1(I)$ function K by

$$K(t) = \begin{cases} 3(\hat{m}(t) + \sum_1^k m_i(t)), & t \in I', \\ |\dot{\hat{\phi}}(t)| + M(t), & t \in [\alpha_0, t_0]. \end{cases}$$

Since $\mathscr{F}$ is A.Q., one can use the definition to show that given any ε, $0 \leqq \varepsilon \leqq 1$, there exists a function $G_\varepsilon(x(\cdot), t)$ defined on $\mathscr{C} \times I'$ into R^n such that:

$$(\hat{F} + \varepsilon\delta F + G_\varepsilon) \in \mathscr{F}, \tag{3.4}$$

$$\begin{aligned} |G_\varepsilon(x(\cdot), t)| &\leqq 2(\hat{m}(t) + \sum_1^k m_i(t)), \\ \|dG_\varepsilon[x(\cdot), t; \cdot]\| &\leqq 2(\hat{m}(t) + \sum_1^k m_i(t)) \end{aligned} \tag{3.5}$$

for all $x \in AC(I, X)$ and $t \in I'$,

$$\left| \int_{\tau_1}^{\tau_2} G_\varepsilon(x(\cdot), t)\, dt \right| < \varepsilon^2 \tag{3.6}$$

for every τ_1, τ_2 in I' and $x \in AC(I, X)_K$.

If $z(t)$ is any solution to

$$\begin{aligned} \dot{z}(t) &= \hat{F}(z(\cdot), t) + \varepsilon\delta F(z(\cdot), t) + G_\varepsilon(z(\cdot), t), & t &> t_0, \\ z(t) &= \hat{\phi}(t) + \varepsilon\delta\phi(t) & &\text{on } [\alpha_0, t_0], \end{aligned}$$

where $\delta\phi \in \delta\Phi(M)$, then from the definition of K we get $|\dot{z}(t)| \leqq K(t)$ so that the inequality in (3.6) holds for such solutions z over intervals on which they exist. Note that if δF is not fixed, but is allowed to range over $[\delta F_1, \cdots, \delta F_v]$, then K can be chosen independent of the particular δF in this set. For then there exist $F_1, \cdots, F_k$ such that $\delta F_j = \sum_1^k a^{ji} F_i - \hat{F}$, $j = 1, \cdots, v$, so that K will depend on $F_1, F_2, \cdots, F_k, \hat{F}$ but not on a particular $\delta F \in [\delta F_1, \cdots, \delta F_v]$.

We next consider "perturbations" of the system (3.3). For arbitrary $\delta t \in R^1$, $\delta\phi \in \delta\Phi(M)$, $\delta F \in [\mathscr{F}] - \hat{F}$, and $0 \leqq \varepsilon \leqq 1$, we consider the system

$$\begin{aligned} \dot{z}(t, \varepsilon) &= \hat{F}(z(\cdot, \varepsilon), t) + \varepsilon\delta F(z(\cdot, \varepsilon), t) + G_\varepsilon(z(\cdot, \varepsilon), t), & t &> t_0, \\ z(t, \varepsilon) &= \hat{\phi}(t) + \varepsilon\delta\phi(t) & &\text{on } [\alpha_0, t_0]. \end{aligned} \tag{3.7}$$

Lemmas similar to Lemmas 4.2 and 4.3 in [1] can be proved for this system (the proofs of Lemmas 4.2 and 4.3 are changed only slightly). One then has that for $\varepsilon > 0$ sufficiently small the solution $z(t, \varepsilon)$ to (3.7) exists on $[t_0, t_1 + \varepsilon|\delta t|]$, where it has the form

$$z(t, \varepsilon) = \hat{x}(t) + \varepsilon\delta x(t) + o(\varepsilon) \tag{3.8}$$

with δx satisfying the linear variational system

$$\text{(3.9)}\qquad \begin{aligned} \delta\dot{x}(t) &= d\hat{F}[\hat{x}(\cdot),t;\delta x] + \delta F(\hat{x}(\cdot),t), && t\in[t_0, t_1 + \varepsilon|\delta t|],\\ \delta x(t) &= \delta\phi(t), && t\in[\alpha_0, t_0]. \end{aligned}$$

Let $\hat{\mathcal{N}}_T = \mathcal{N}_T(q_{\hat{x}})$, $\hat{\mathcal{M}}_T = \mathcal{M}_T(q_{\hat{x}})$, and $\hat{\mathcal{N}}_T - q_{\hat{x}} = \{w\in R^{2n+1} : w = w^* - q_{\hat{x}}$, where $w^*\in\hat{\mathcal{N}}_T\}$. Then proceeding next exactly as in the proof of Theorem 5 in [1] (the proofs of Lemmas 4.4 and 4.5 in [1] are carried out with only slight modifications) one gets that the convex sets $\hat{\mathcal{K}}$ and $\hat{\mathcal{N}}_T - q_{\hat{x}}$ can be separated by a hyperplane $\mathcal{H}$ through the origin in R^{2n+1}. Choosing ζ to be a nonzero normal to $\mathcal{H}$ such that

$$\text{(3.10)}\qquad \zeta\cdot p \leqq 0 \leqq \zeta\cdot w$$

for all $p\in\hat{\mathcal{K}}$ and $w\in\hat{\mathcal{N}}_T - q_{\hat{x}}$, we have that (ii) and (iii) of Theorem 4 hold. It is easy to show that (i) also holds since $\hat{\mathcal{M}}_T - q_{\hat{x}}$ is a plane through the origin which is contained in $\hat{\mathcal{N}}_T - q_{\hat{x}}$.

Before applying these results to our control problem, we state a lemma that will be needed. The proof of this lemma will not be given here since it is essentially the same as the proof due to Gamkrelidze of Lemma 4.1 in [7].

LEMMA 3.1. *Let X be a compact convex subset of $\mathcal{G}$, K a nonnegative $L_1(I)$ function, and $\varepsilon > 0$. Let $F_j(x(\cdot),t)$, $j = 1,\cdots,k$, be mappings from $AC(I,X)\times I'$ into R^n that are measurable in t for fixed x and C^1 in x for fixed t. Assume there exists an $m(t)$ in $L_1(I')$ such that $|F_j(x(\cdot),t)| \leqq m(t)$, $\|dF_j[x(\cdot),t;\cdot]\| \leqq m(t)$ for all $x\in AC(I,X)$ and $t\in I'$, $j = 1,\cdots,k$. Let $\mathcal{Y}$, a subset of $AC(I,X)$, be the compact convex set defined by $\mathcal{Y} = AC(I,X)_K$. Let $p_j(t)$, $j = 1,\cdots,k$, be given nonnegative real-valued measurable functions on I' satisfying $\sum_1^k p_j(t) = 1$ a.e. on I'. Then it is possible to subdivide I' into sufficiently small disjoint subintervals E_i, $i = \pm 1, \pm 2,\cdots$, and to assign to each E_i one of the functions $F_1,\cdots,F_k$, which we shall denote by F_{E_i}, so that the function $F(x(\cdot),t)$ defined by*

$$F(x(\cdot),t) = F_{E_i}(x(\cdot),t)$$

for $t\in E_i$, $i = \pm 1, \pm 2,\cdots$, and $x\in AC(I,X)$ satisfies

$$\left|\int_{\tau_1}^{\tau_2}\left\{\sum_{j=1}^k p_j(t)F_j(x(\cdot),t) - F(x(\cdot),t)\right\}dt\right| < \varepsilon$$

for every τ_1,τ_2 in I' and $x\in\mathcal{Y}$.

Now consider the optimal control problem of § 1 under the assumptions made there. Suppose that $(\bar{\phi}^*, u^*, \bar{x}^*, t_1^*)$ is a solution to this problem. Put

$$x^*(t) \equiv \begin{pmatrix} x^{0*}(t)\\ \bar{x}^*(t)\end{pmatrix},\qquad \alpha_0 \leqq t \leqq t_1^*,$$

where

$$x^{0*}(t) \equiv \begin{cases} \displaystyle\int_{t_0}^{t} f^0(\bar{x}^*(\cdot), u^*(\sigma),\sigma)\,d\sigma, & t\in[t_0, t_1^*],\\ 0, & t\in[\alpha_0, t_0]. \end{cases}$$

Now define $f(x(\cdot),u,t) = f(\bar{x}(\cdot),u,t)$, where $x = (x^0,\bar{x}) \in C(I,\mathscr{G})$ with $\mathscr{G} \equiv R^1 \times \bar{\mathscr{G}}$. (That is, hereafter we shall write f as a function of x even though it does not really depend on x^0.) Put $\mathscr{F} = \{F(x(\cdot),t) : F(x(\cdot),t) = f(x(\cdot),u(t),t) \text{ for } u \in \Omega\}$ and $\Phi = \{\phi = (\phi^0,\bar{\phi}) : \bar{\phi} \in \bar{\Phi} \text{ and } \phi^0 \in AC([\alpha_0,t_0],R^1)\}$.

Let γ^0, ξ^0, τ represent scalars and γ, ξ represent $(n-1)$-vectors. Then define $\mathscr{N} \subset R^{2n+1}$ to be all $(\gamma^0,\gamma,\xi^0,\xi,\tau)$ with (γ,ξ,τ) near $(\bar{x}^*(t_0),\bar{x}^*(t_1^*),t_1^*)$ satisfying

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

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

With the above definitions in mind one can prove (the proof is exactly the same as the proof of Lemma 5.1 in [1]) that x^* is an $\mathscr{F}$, $\mathscr{N}$, Φ extremal. Furthermore, the class $\mathscr{F}$ defined above is absolutely quasi-convex. The proof that $\mathscr{F}$ is A.Q. uses Lemma 3.1 and the arguments are similar to those for the proof of a similar result in [2].

Thus, we can apply Theorem 4 to the control problem in the above formulation. Let $\delta\Phi = \delta\Phi(M)$ be as defined above. For $(\bar{\phi}^*,u^*,\bar{x}^*,t_1^*)$ optimal, and $\delta\phi \in \delta\Phi$, α in P^k, $\{u_i\}_1^k$ in Ω, we denote by δx the solution to

$$\begin{aligned}
&\delta\dot{x}(t) = df[x^*(\cdot),u^*(t),t;\delta x] + \sum_1^k \alpha^i f(x^*(\cdot),u_i(t),t) - f(x^*(\cdot),u^*(t),t) \\
(3.11)\qquad & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{on} \quad [t_0,t_1^*], \\
&\delta x(t) = \delta\phi(t) \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{on} \quad [\alpha_0,t_0],
\end{aligned}$$

where $x^* = (x^{0*},\bar{x}^*)$. Define $\delta\mathscr{X}$ to be the set of all such solutions for $\delta\phi$ in $\delta\Phi$, $\{u_i\}_1^k$ any finite collection in Ω, and $\alpha \in P^k$, k an arbitrary positive integer. Then we have the following theorem.

THEOREM 5. *Let $(\bar{\phi}^*,u^*,\bar{x}^*,t_1^*)$ be optimal. Suppose t_1^* is a regular point for $f(\bar{x}^*(\cdot),u^*(t),t)$. Then there exists a nonzero $(2n+1)$-vector $\zeta = (b_0,b_1,a_1) = (b_0^0,b_0^1,\cdots,b_0^{n-1},b_1^0,b_1^1,\cdots,b_1^{n-1},a_1) = (b_0^0,\bar{b}_0,b_1^0,\bar{b}_1,a_1)$ such that:*

(i) *the* $(2n-1)$-vector $(\bar{b}_0,\bar{b}_1,a_1)$ *is orthogonal to* $\mathscr{T}$ *at* $(\bar{x}^*(t_0),\bar{x}^*(t_1^*),t_1^*)$;

(ii) $b_1^0 \leqq 0$;

(iii) $b_1 \cdot f(\bar{x}^*(\cdot),u^*(t_1^*),t_1^*) + a_1 = 0$;

(iv) $b_0 \cdot \delta\phi(t_0) + b_1 \cdot \delta x(t_1^*) \leqq 0$ *for arbitrary* $\delta\phi$ *in* $\delta\Phi$, δx *in* $\delta\mathscr{X}$ (δx *corresponding to* $\delta\phi$).

Proof. Theorem 5 follows almost immediately from Theorem 4. Statements (iii) and (iv) are a direct consequence of (iii) in Theorem 4. Statement (ii) follows from (ii) of Theorem 4 since the $(2n+1)$-vector $w = (0_n,-1,0_{n-1},0)$ is such that $q_{\bar{x}} + w \in \mathscr{N}_T(q_{\bar{x}})$ for the above defined $\mathscr{N}$. Finally, if (γ_T,ξ_T,τ_T) is any tangent vector to $\mathscr{T}$ at $(\bar{x}^*(t_0),\bar{x}^*(t_1^*),t_1^*)$, then the vector $(0,\gamma_T,0,\xi_T,\tau_T)$ is tangent to $\mathscr{M}$ (as defined above) at $q_{x^*} = (0,\bar{x}^*(t_0),x^{0*}(t_1^*),\bar{x}^*(t_1^*),t_1^*)$. Hence condition (i) of Theorem 5 follows from (i) of Theorem 4.

Remarks. First let us note that the results of Theorem 5 imply that $b_1 \neq 0$. For if b_1 were zero, using (iii) and (iv) would give $a_1 = 0$ and $b_0 = 0$ and hence $\zeta = (b_0,b_1,a_1) = 0$, a contradiction!

We also point out that if $\bar{\Phi}$ were some given convex class of A.C. initial functions for the control problem, one could still prove a result similar to Theorem 5. In the statement and proof of Theorem 4 then, Φ would be some specified convex class. The set $\delta\Phi(M)$ would be replaced by $\Phi - \hat{\phi}$ and for $\delta\phi$ in $[\delta\phi_1, \cdots, \delta\phi_l]$, $K(t)$ would be chosen so that $K(t) \geqq |\hat{\phi}(t)| + \sum_1^\nu |\dot{\phi}_i(t)|$, where $\delta\phi_j = \sum_{i=1}^\nu \alpha^{ji}\phi_i - \hat{\phi}$, $j = 1, \cdots, l$. In the application of Theorem 4 to the control problem, Φ would be defined as above in terms of the given class $\bar{\Phi}$ for the control problem. Then we would have $\delta\Phi = \Phi - \phi^*$, where $\phi^* = (0, \bar{\phi}^*)$, in the statement of Theorem 5. It should be noted however that the result about $b_1 \neq 0$ mentioned above does not in general hold for this problem.

To complete the proof of Theorem 1, it remains only to use Theorem 3 of § 2 to obtain a representation for solutions δx of system (3.11). The results of Theorem 1 will then follow easily from Theorem 5 above.

Consider $df[\bar{x}^*(\cdot), u^*(t), t; \cdot]$. Since for each $t \in I'$, $df[\bar{x}^*(\cdot), u^*(t), t; \cdot]$ is a bounded linear operator on $C([\alpha_0, t], R^{n-1})$ into R^n, we have by the Riesz theorem that there exists an $n \times (n-1)$ matrix function $\bar{\eta}^*(t, s)$ such that

$$df[\bar{x}^*(\cdot), u^*(t), t; \bar{\psi}] = \int_{\alpha_0}^t d_s\bar{\eta}^*(t, s)\bar{\psi}(s)$$

for all $\bar{\psi} \in C([\alpha_0, t], R^{n-1})$ and $t \in I'$.

Defining $NBV[\alpha_0, t]$ as all g satisfying: (i) g is of bounded variation on $[\alpha_0, t]$, (ii) $g(t) = 0$, (iii) g is continuous from the right; we may then assert the existence of a unique $\bar{\eta}^*(t, \cdot)$ in $NBV[\alpha_0, t]$ such that the above equation holds.

Let $\bigvee_{s=\alpha_0}^t \bar{\eta}^*(t, s)$ denote the variation of $\bar{\eta}^*(t, \cdot)$ on $[\alpha_0, t]$. Then a further consequence of the Riesz theorem is that there exists a constant $D > 0$ such that

$$\bigvee_{s=\alpha_0}^t \bar{\eta}^*(t, s) \leqq D\|df[\bar{x}^*(\cdot), u^*(t), t; \cdot]\|.$$

Thus we have

$$\bigvee_{s=\alpha_0}^t \bar{\eta}^*(t, s) \leqq Dm^*(t) = \tilde{m}(t),$$

where $m^* \in L_1(I')$.

If we adopt the notation explained above, then we may write

$$df[x^*(\cdot), u^*(t), t; \psi] = \int_{\alpha_0}^t d_s\eta^*(t, s)\psi(s) \tag{3.12}$$

for $\psi \in C(I, R^n)$, where $\eta^*(t, s)$ is the $n \times n$ matrix function

$$\begin{pmatrix} 0 & \\ \vdots & \bar{\eta}^*(t, s) \\ 0 & \end{pmatrix}.$$

The integral in (3.12) is the Lebesgue–Stieltjes integral. We shall assume that

$\eta^*(t, s)$ has been extended as follows:

$$\eta^*(t, s) = \begin{cases} \eta^*(t, t) = 0 & \text{for} \quad s > t, \\ \eta^*(t, \alpha_0) & \text{for} \quad s < \alpha_0. \end{cases}$$

This may be done without affecting (3.12).

Using (3.12) and the fact that $df[x^*(\cdot), u^*(t), t; \psi]$ is measurable on I' for each $\psi \in C(I, R^n)$, we find it is not hard to show that $\eta^*(t, s)$ is measurable in (t, s). Furthermore, for each σ in $[\alpha_0, t]$,

$$\begin{aligned} |\eta^*(t, \sigma)| &= |\eta^*(t, t) - \eta^*(t, \sigma)| \\ &\leqq |\eta^*(t, t) - \eta^*(t, \sigma)| + |\eta^*(t, \sigma) - \eta^*(t, \alpha_0)| \\ &\leqq \bigvee\nolimits_{s=\alpha_0}^{t} \eta^*(t, s) \leqq \tilde{m}(t). \end{aligned}$$

We thus find that η^* satisfies the hypotheses of Theorem 3. We then have that the elements of $\delta\mathscr{X}$ have the form, for $t > t_0$,

$$\begin{aligned} \delta x(t) &= \Gamma(t_0, t)\delta\phi(t_0) \\ &\quad + \int_{t_0}^{t} \Gamma(\beta, t)\, d\beta \int_{\alpha_0}^{t_0} d_s\eta^*(\beta, s)\, \delta\phi(s) \\ &\quad + \int_{t_0}^{t} \Gamma(\beta, t)\left\{\sum_1^k \alpha^i f(x^*(\cdot), u_i(\beta), \beta) - f(x^*(\cdot), u^*(\beta), \beta)\right\} d\beta, \end{aligned} \tag{3.13}$$

where $\delta\phi \in \delta\Phi$, $\alpha \in P^k$, $u_i \in \Omega$ and Γ satisfies (2.1) with $\eta = \eta^*$.

We next define multipliers λ by

$$\lambda(\sigma) = b_1\Gamma(\sigma, t_1^*), \qquad t_0 \leqq \sigma \leqq t_1^*, \tag{3.14}$$

where $b_1 \neq 0$ is as described in Theorem 5. Then, since Γ satisfies (2.1) with $\eta = \eta^*$, we have that λ satisfies

$$\lambda(t_1^*) = b_1,$$

$$\lambda(\sigma) + \int_{\sigma}^{t_1^*} \lambda(\beta)\eta^*(\beta, \sigma)\, d\beta = b_1, \qquad t_0 \leqq \sigma < t_1^*.$$

Furthermore, λ is continuous at t_1^* (see the remarks following Theorem 2). Hence, there is an interval $(\beta, t_1^*]$ on which λ does not vanish. We also have that

$$\lambda^0(\sigma) = \lambda^0(t_1^*) = b_1^0 \leqq 0$$

since the first column in η^* is zero.

Using (3.13), (3.14) and (iv) of Theorem 5 gives

$$\begin{aligned} &\{b_0 + \lambda(t_0)\} \cdot \delta\phi(t_0) + \int_{t_0}^{t_1^*} \lambda(\beta)\, d\beta \int_{\alpha_0}^{t_0} d_s\eta^*(\beta, s)\delta\phi(s) \\ &\quad + \int_{t_0}^{t_1^*} \lambda(\beta)\left\{\sum_1^k \alpha^i f(x^*(\cdot), u_i(\beta), \beta) - f(x^*(\cdot), u^*(\beta), \beta)\right\} d\beta \leqq 0 \end{aligned}$$

for arbitrary $\delta\phi \in \delta\Phi$, $\alpha \in P^k$, $\{u_i\}_1^k$ in Ω, k arbitrary. Since $\delta\phi$ and α, $\{u_i\}$ are independent, this may be written

$$\int_{t_0}^{t_1^*} \lambda(\beta)\{f(x^*(\cdot), u(\beta), \beta) - f(x^*(\cdot), u^*(\beta), \beta)\}\, d\beta \leqq 0 \tag{3.15}$$

for arbitrary $u \in \Omega$, and

$$\{b_0 + \lambda(t_0)\} \cdot \delta\phi(t_0) + \int_{t_0}^{t_1^*} \lambda(\beta)\, d\beta \int_{\alpha_0}^{t_0} d_s\eta^*(\beta, s)\delta\phi(s) \leqq 0 \tag{3.16}$$

for arbitrary $\delta\phi \in \delta\Phi$.

Since $-\delta\phi$ is in $\delta\Phi$ whenever $\delta\phi$ is, (3.16) may be written

$$\{b_0 + \lambda(t_0)\} \cdot \delta\phi(t_0) + \int_{t_0}^{t_1^*} \lambda(\beta)\, d\beta \int_{\alpha_0}^{t_0} d_s\eta^*(\beta, s)\delta\phi(s) = 0 \tag{3.17}$$

for arbitrary $\delta\phi \in \delta\Phi$.

Interchanging the order of integration (which again is possible by an unsymmetric Fubini theorem), we can write (3.17) as

$$\{b_0 + \lambda(t_0)\} \cdot \delta\phi(t_0) + \int_{\alpha_0}^{t_0} \left\{ d_s \int_{t_0}^{t_1^*} \lambda(\beta)\eta^*(\beta, s)\, d\beta \right\} \delta\phi(s) = 0 \tag{3.18}$$

for arbitrary $\delta\phi \in \delta\Phi$.

Defining the n-vector function H by

$$H(s) = \int_{t_0}^{t_1^*} \lambda(\beta)\eta^*(\beta, s)\, d\beta, \qquad \alpha_0 \leqq s \leqq t_0, \tag{3.19}$$

and taking $\delta\phi$ with jth component equal to 1, all other components zero in (3.18), yields

$$b_0^j + \lambda^j(t_0) + H^j(t_0) - H^j(\alpha_0) = 0.$$

Hence

$$b_0 + \lambda(t_0) + H(t_0) - H(\alpha_0) = 0 \tag{3.20}$$

or

$$b_0 = -\lambda(t_0) + \int_{t_0}^{t_1^*} \lambda(\beta)\{\eta^*(\beta, \alpha_0) - \eta^*(\beta, t_0)\}\, d\beta. \tag{3.21}$$

Combining the results of the above discussions with Theorem 5 completes the proof of Theorem 1.

Let us make a few observations about this theorem. Return now to (3.17) and (3.18). We consider two cases.

Case 1. Suppose the matrix function $\eta^*(t, s)$ is such that $d_s\eta^*(t, s) = v^*(t, s)\, ds$ on $[\alpha_0, t_0]$. That is, $\eta^*(t, s)$ is A.C. in s. Then one can use (3.17) to show (using

Lemma 2 of [1] and arguments similar to those in [1])

$$b_0 + \lambda(t_0) = 0$$

and

$$\int_{t_0}^{t_1^*} \lambda(\beta) v^*(\beta, s)\, d\beta = 0 \quad \text{for almost every } s \text{ in}[\alpha_0, t_0]. \tag{3.22}$$

Then (iii) of Theorem 1 would become: $(-\bar{\lambda}(t_0), \bar{\lambda}(t_1^*), -\lambda(t_1^*)\cdot f^*(t_1^*))$ is orthogonal to $\mathscr{T}$.

Case 2. Suppose the function H defined by (3.19) is such that $dH(s) = h(s)\, ds$ on $[\alpha_0, t_0]$. That is, H is A.C. Then (3.18) may be written

$$\{b_0 + \lambda(t_0)\} \cdot \delta\phi(t_0) + \int_{\alpha_0}^{t_0} h(s)\delta\phi(s)\, ds = 0$$

for arbitrary $\delta\phi \in \delta\Phi$. One can then show that this implies $h(s) = 0$ a.e. on $[\alpha_0, t_0]$ and, hence,

$$b_0 + \lambda(t_0) = H(\alpha_0) - H(t_0) = 0.$$

Again (iii) of Theorem 1 would take the form stated in Case 1 above.

Case 2 includes the case where the functional is a functional involving only lags (see the discussion below).

Let us point out that it is not difficult to show Case 1 implies Case 2. However, the converse is not true as will be seen in the case of lag problems.

4. Examples of functional systems. In this section we shall discuss briefly several examples of systems to which our results are applicable. For our first example, let us consider systems where the functional dependence is in terms of lags. This type of system has been considered in detail in [1]. We just remark here that the results in Theorem 1 agree with those previous results found in [1]. For simplicity, let f contain a single constant lag. That is, consider $f(\bar{x}(\cdot), u(t), t) = g(\bar{x}(t), \bar{x}(t-\theta), u(t), t)$, where $g = g(\bar{\xi}_1, \bar{\xi}_2, u, t)$ is a mapping of $R^{n-1} \times R^{n-1} \times \mathscr{U} \times I'$ into R^n. Then $\bar{\eta}^*(t, s)$ of Theorem 1 has the form

$$\bar{\eta}^*(t, s) = \begin{cases} 0, & s \geqq t, \\ -g_{\bar{\xi}_1}^*(t), & t - \theta \leqq s < t, \\ -g_{\bar{\xi}_1}^*(t) - g_{\bar{\xi}_2}^*(t), & s < t - \theta, \end{cases}$$

where $g^*(t) = g(\bar{x}^*(t), \bar{x}^*(t-\theta), u^*(t), t)$ and $g_{\bar{\xi}_i}$ is the Jacobian matrix. Note that in this case $\eta^*(t, s)$ is not A.C. in s. However (see 3.19),

$$H(s) = -\int_{s+\theta}^{t_1^*} \lambda(\beta) g_{\bar{\xi}_2}^*(\beta)\, d\beta - \int_{t_0}^{t_1^*} \lambda(\beta) g_{\bar{\xi}_1}^*(\beta)\, d\beta,$$

where

$$g_{\xi_i} = \begin{pmatrix} 0 & \\ \vdots & g_{\bar{\xi}_i} \\ 0 & \end{pmatrix},$$

so that Case 2 above holds while Case 1 does not.

For this type of problem, the multipliers are usually given as A.C. functions satisfying a set of advanced differential-difference equations. These advanced equations are just the differentiated form of the equations for λ given in (i) of Theorem 1. In the case of lags, one can show that the multipliers λ of Theorem 1 are actually A.C. and satisfy the equations in (i) in differentiated form. Thus the results of Theorem 1 agree with the known results whenever we deal with a system with lags.

A second type of functional to which our results can be applied is of the form

$$f(\bar{x}(\cdot), u(t), t) = \int_{\alpha_0}^{t} a(t, s) g(\bar{x}(s), u(t), t)\, ds,$$

where a is a scalar function on $R^1 \times R^1$ and $g = g(\bar{\xi}, u, t)$ is a mapping from $R^{n-1} \times \mathscr{U} \times I'$ into R^n. Such systems arise in the study of reactor dynamics [13]. In this example we find

$$\bar{\eta}^*(t, s) = -\int_{s}^{t} a(t, \sigma) g_{\bar{\xi}}(\bar{x}^*(\sigma), u^*(t), t)\, d\sigma$$

so that Case 1 of the previous section holds. As in the example for simple lags discussed above, one can also for this example show that the multipliers are actually A.C. and satisfy the equations in (i) of Theorem 1 in differentiated form. In addition to the results stated in Theorem 1, for this example we also obtain (see 3.22) the necessary condition

$$\int_{t_0}^{t_1^*} \lambda(\beta) a(\beta, s) g_{\bar{\xi}}(\bar{x}^*(s), u^*(\beta), \beta)\, d\beta = 0$$

for almost every s in $[\alpha_0, t_0]$.

Finally, the systems studied by Hale [10] and Halanay [9],

$$\dot{x}(t) = g(x_t, u(t), t),$$

where x_t denotes the values $x(t + \sigma)$, $-\tau \leqq \sigma \leqq 0$, are included in our general systems. In this example the function η^* would have the additional property that $\eta^*(t, s) = \eta^*(t, t - \tau)$ for $s \leqq t - \tau$.

Acknowledgment. The author wishes to express his appreciation to the referees for several remarks which aided the clarity of exposition of this paper.

REFERENCES

[1] H. T. BANKS, *Necessary conditions for control problems with variable time lags*, this Journal, 6 (1968), pp. 9–47.

[2] ———, *Optimal control problems with delays*, Doctoral thesis, Division of Mathematical Sciences, Purdue University, Lafayette, Indiana, 1967.

[3] R. H. CAMERON AND W. T. MARTIN, *An unsymmetric Fubini theorem*, Bull. Amer. Math. Soc., 47 (1941), pp. 121–125.

[4] D. H. CHYUNG AND E. B. LEE, *Linear optimal systems with time delay*, this Journal, 4 (1966), pp. 548–575.

[5] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators*, vol. 1, Interscience, New York, 1958.

[6] AVNER FRIEDMAN, *Optimal control for hereditary processes*, Arch. Rational Mech. Anal., 15 (1964), pp. 396–416.

[7] 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.

[8] L. M. GRAVES, *The Theory of Functions of Real Variables*, 2nd ed., McGraw-Hill, New York, 1956.

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

[10] J. K. HALE, *Sufficient conditions for stability and instability of autonomous functional-differential equations*, J. Differential Equations, 1 (1965), pp. 452–482.

[11] T. H. HILDEBRANDT, *On integrals related to and extensions of the Lebesgue integral*, Bull. Amer. Math. Soc., 24 (1917), pp. 177–202.

[12] S. LANG, *Introduction to Differentiable Manifolds*, Interscience, New York, 1962.

[13] J. J. LEVIN AND J. A. NOHEL, *A system of nonlinear integro-differential equations*, Michigan Math. J., 13 (1966), pp. 257–270.

[14] I. P. NATANSON, *Theory of Functions of a Real Variable*, Frederick Ungar, New York, 1955.

[15] L. W. NEUSTADT, *An abstract variational theory with applications to a broad class of optimization problems, I, II*, this Journal, 4 (1966), pp. 505–527; 5 (1967), pp. 90–137.

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

[17] V. VOLTERRA, *Theory of Functionals*, Blackie, London, 1930.

SIAM J. CONTROL
Vol. 7, No. 1, February 1969

DIFFERENCE APPROXIMATIONS IN OPTIMAL CONTROL PROBLEMS*

B. M. BUDAK, E. M. BERKOVICH AND E. N. SOLOV'EVA†

Abstract. The convergence of solutions of discrete extremum problems to the solution of the continuous optimal control problem is considered, both in the sense of the optimal value of the functional and of the optimal control. In the latter case "nonwell-posed" extremum problems are discussed.

In the majority of cases the solution of optimal control problems, in which a functional defined on the solutions of differential equations is to be minimized, requires the application of numerical methods.

In such cases, as usual, the original "differential" problem is actually replaced by a difference problem, and there arises the question of the convergence of the solution of the difference problem to the solution of the differential problem.

In the present note this question is studied for one class of the problems mentioned.

Suppose that we are required to solve the problem of minimizing the functional

$$J(\mathbf{x},\mathbf{u}) = \int_{t_0}^{T} g(\mathbf{x}(\tau),\mathbf{u}(\tau),\tau)\,d\tau + \Phi(\mathbf{x}(T)) \tag{1}$$

defined on the solutions of the Cauchy problem

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x},\mathbf{u}(t),t), \quad t_0 \leqq t \leqq T, \qquad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{2}$$

corresponding to all possible admissible controls $\mathbf{u}$ in a certain class U. Here, $\mathbf{x} = (x^1, \cdots, x^N)$, $\mathbf{u} = (u^1, \cdots, u^r)$, $\mathbf{f} = (f^1, \cdots, f^N)$ are vectors; $g(\mathbf{x},\mathbf{u},t)$ and $\Phi(\mathbf{x})$ are scalar-valued functions.

Having divided the interval $t_0 \leqq t \leqq T$ by the partition points $\Sigma_{tn}: t_0 = t_{n0} < t_{n1} < \cdots < t_{ni} < t_{ni+1} < \cdots < t_{nn-1} < t_{nn} = T$ into partial intervals of length $\tau_{ni} = t_{ni+1} - t_{ni}$, $i = 0, 1, \cdots, n-1$, we replace problem (1), (2) by the approximate difference problem of minimizing the discretized functional:

$$J_n(\mathbf{x}_n,\mathbf{u}_n) = \sum_{i=0}^{n-1} g(\mathbf{x}_{ni},\mathbf{u}_{ni},t_{ni})\tau_{ni} + \Phi(\mathbf{x}_{nn}) \tag{1'}$$

on the solutions of the Cauchy difference problem

$$\mathbf{x}_{ni+1} = \mathbf{x}_{ni} + \mathbf{f}(\mathbf{x}_{ni},\mathbf{u}_{ni},t_{ni})\tau_{ni}, \quad i = 0, 1, \cdots, n-1, \qquad \mathbf{x}_{n0} = \mathbf{x}_0 \tag{2'}$$

* Originally published in Vestnik Moskovskogo Universiteta, Seriya I: Matematika, Mekhanika, 1968, no. 2, pp. 41–55. Submitted on May 20, 1967. This translation into English has been prepared by N. H. Choksy.

Translated and printed for this Journal under a grant-in-aid by the National Science Foundation.

† Mathematical Physics Department, Moscow State University, Moscow.

corresponding to all possible admissible controls $\mathbf{u}_n$ in a certain class U_n. Here $\mathbf{u}_n$ is a vector-valued function of the integral argument i, where we take the values $\mathbf{u}_{n0}, \mathbf{u}_{n1}, \cdots, \mathbf{u}_{nn-1}$, respectively, when $i = 0, 1, \cdots, n-1$.

We say that the discrete control $\mathbf{u}_n$ is piecewise-constantly extended on $[t_0, T]$ if it is interpreted as $\mathbf{u}_n(t) = \mathbf{u}_{ni}$ when $t_{ni} \leqq t < t_{ni+1}, i = 0, 1, \cdots, n-1$. By assuming that $\tau_n = \max_{0 \leqq i \leqq n-1} \tau_{ni} \to 0$ as $n \to +\infty$, we study the question of the convergence of the solution of problem (1′), (2′) to the solution of problem (1), (2); we assume here that the functions $\mathbf{f}(\mathbf{x}, \mathbf{u}, t)$, $g(\mathbf{x}, \mathbf{u}, t)$, $\Phi(\mathbf{x})$ have been defined for all $\mathbf{x}, \mathbf{u}$ and for all $t \in [t_0, T]$, that the control $\mathbf{u}(t)$ has been defined for all $t \in [t_0, T]$ and that the following requirements are fulfilled:

(a) There exist constants $A_i, i = 1, 2, \cdots, 11$, such that[1]

$$|\mathbf{f}(\mathbf{x}, \mathbf{u}, t)| \leqq A_1|\mathbf{x}| + A_2|\mathbf{u}|^2 + A_3, \tag{3}$$

$$\begin{aligned} &|\mathbf{f}(\mathbf{x}^*, \mathbf{u}^*, t) - \mathbf{f}(\mathbf{x}^{**}, \mathbf{u}^{**}, t)| \\ &\qquad \leqq A_4|\mathbf{x}^* - \mathbf{x}^{**}| + A_5(|\mathbf{u}^*| + |\mathbf{u}^{**}|)(|\mathbf{u}^* - \mathbf{u}^{**}|) + A_6|\mathbf{u}^* - \mathbf{u}^{**}|^2, \end{aligned} \tag{4}$$

$$|g(\mathbf{x}, \mathbf{u}, t)| \leqq A_7|\mathbf{x}| + A_8|\mathbf{u}|^2 + A_9. \tag{5}$$

(b) $\mathbf{f}(\mathbf{x}, \mathbf{u}, t)$ and $g(\mathbf{x}, \mathbf{u}, t)$ are piecewise uniformly continuous in t on $[t_0, T]$ with points of discontinuity which in number are bounded uniformly relative to the piecewise-constant $\mathbf{x}$ and $\mathbf{u}$; furthermore, $\Phi(\mathbf{x})$ and $g(\mathbf{x}, \mathbf{u}, t)$ are continuous in $\mathbf{x}$, uniformly in $\mathbf{u}$ and t, $t \in [t_0, T]$, and $g(\mathbf{x}, \mathbf{u}, t)$ satisfies the condition

$$|g(\mathbf{x}, \mathbf{u}^*, t) - g(\mathbf{x}, \mathbf{u}^{**}, t)| \leqq A_{10}(|\mathbf{u}^*| + |\mathbf{u}^{**}|)(|\mathbf{u}^* - \mathbf{u}^{**}|) + A_{11}|\mathbf{u}^* - \mathbf{u}^{**}|^2. \tag{6}$$

(c) The control $\mathbf{u}(t)$ is either square summable on $[t_0, T]$ and satisfies the constraint

$$\|\mathbf{u}\|_{L_2[t_0, T]} \leqq K_1 = \text{const.} < +\infty \tag{7}$$

or is bounded and measurable and satisfies the estimate[2]

$$\|\mathbf{u}\|_{C[t_0, T]} \leqq K_2 = \text{const.} < +\infty. \tag{8}$$

In the first case $\mathbf{u}(t)$ belongs to a closed ball B_{K_1} in $L_2[t_0, T]$; in the second, to the set B_{K_2} of bounded measurable functions satisfying inequality (8); B_{K_1} and B_{K_2} are closed and convex in $L_2[t_0, T]$.

Remark. In inequalities (3) and (5), instead of $|\mathbf{x}|$ we can take $|\mathbf{x}|^m$ or $e^{|\mathbf{x}|^m}$, where m is an arbitrary positive integer; in this case a constraint arises on the length of the interval $[t_0, T]$.

However, if the control is bounded and measurable and satisfies (8), then in (3) and (5) instead of $|\mathbf{u}|^2$ we can take $|\mathbf{u}|^m$ or $e^{|\mathbf{u}|^m}$, where m is an arbitrary positive

[1] $|\mathbf{x}| = \sum_{k=1}^{N} |x^k|$; $\quad |\mathbf{u}| = \sum_{k=1}^{r} |u^k|$.

[2] $\|\mathbf{u}\|_{C[t_0, T]} = \sup_{t_0 \leqq t \leqq T} |\mathbf{u}(t)|$.

integer. All the subsequent theorems remain valid under the changes indicated. In our investigation of the convergence of difference approximations in optimal control problems, we need the following rather obvious theorem.

THEOREM 1. *If conditions* (a), (b), (c) *are fulfilled, then*: (i) *the solution of the Cauchy problem* (2) *exists and is absolutely continuous and unique for any* $\mathbf{u} \in B_{K_1}(B_{K_2})$; *moreover, for every* $\varepsilon > 0$ *we can find* $\delta = \delta(\varepsilon) > 0$ *such that when the condition*

$$\|\mathbf{u}^* - \mathbf{u}^{**}\|_{L_2[t_0,T]} < \delta(\varepsilon), \qquad \mathbf{u}^*, \mathbf{u}^{**} \in B_{K_1}(B_{K_2}), \tag{9}$$

is fulfilled, we have

$$\|\mathbf{x}^* - \mathbf{x}^{**}\|_{C[t_0,T]} < \varepsilon; \tag{10}$$

(ii) *the functional* (1) *has been defined on all* $\mathbf{u} \in B_{K_1}$ (B_{K_2}) *and is continuous in the metric of* $L_2[t_0, T]$; *i.e., for every* $\varepsilon' > 0$ *we can find* $\delta'(\varepsilon') > 0$ *such that when the condition*

$$\|\mathbf{u}^* - \mathbf{u}^{**}\|_{L_2[t_0,T]} < \delta'(\varepsilon'), \qquad \mathbf{u}^*, \mathbf{u}^{**} \in B_{K_1}(B_{K_2}) \tag{11}$$

is fulfilled we have

$$|J(\mathbf{u}^*) - J(\mathbf{u}^{**})| < \varepsilon'. \tag{12}$$

Proof. The validity of assertion (i) can be established as, for example, in [1]. The validity of assertion (ii) results easily from the validity of assertion (i) and from the properties of functions $g(\mathbf{x}, \mathbf{u}, t)$ and $\Phi(\mathbf{x})$. The theorem is proved.

THEOREM 2. *Let the conditions of Theorem* 1 *be satisfied and let* $\mathbf{u}^*(t)$ *be the optimal control for functional* (1) *in* B_{K_1} (*or in* B_{K_2}), *while* $\mathbf{x}^*(t) = \mathbf{x}(\mathbf{u}^*)$ *is the optimal trajectory corresponding to it by means of* (2), *i.e.*,

$$\begin{aligned} J^* = J(\mathbf{x}(\mathbf{u}^*), \mathbf{u}^*) &= \int_{t_0}^{T} g(\mathbf{x}^*(\tau), \mathbf{u}^*(\tau), \tau)\, d\tau + \Phi(\mathbf{x}^*(T)) \\ &= \inf_{\mathbf{u}} \left\{ \int_{t_0}^{T} g(\mathbf{x}(\tau), \mathbf{u}(\tau), \tau)\, d\tau + \Phi(\mathbf{x}(T)) \right\}. \end{aligned} \tag{13}$$

Further, let $\mathbf{u}_n^*$ *be the optimal control in* B_{K_1} (*or in* B_{K_2}) (*if it is piecewise-constantly extended on* $[t_0, T]$) *for the functional* (1′), *while* $\mathbf{x}_n^* = \mathbf{x}_n(\mathbf{u}_n^*)$ *is the optimal trajectory corresponding to it by virtue of* (2′), *i.e.*,

$$\begin{aligned} J_n^* = J_n(\mathbf{x}_n(\mathbf{u}_n^*), \mathbf{u}_n^*) &= \sum_{i=0}^{n-1} g(\mathbf{x}_{ni}^*, \mathbf{u}_{ni}^*, t_{ni})\tau_{ni} + \Phi(\mathbf{x}_{nn}^*) \\ &= \inf_{\mathbf{u}_n} \left\{ \sum_{i=0}^{n-1} g(\mathbf{x}_{ni}, \mathbf{u}_{ni}, t_{ni})\tau_{ni} + \Phi(\mathbf{x}_{nn}) \right\}, \end{aligned} \tag{14}$$

where the infimum in (13) *and* (14) *is taken over* B_{K_1} (*or* B_{K_2}). *Then, if* $\tau_n = \max_{0 \leqq i \leqq n-1} \tau_{ni} = O(n^{-1})$ *as* $n \to +\infty$, *then*

$$\lim_{n \to +\infty} J_n^* = J^*. \tag{15}$$

The proof of this theorem is based on the following three lemmas.

LEMMA 1. *If* $\tilde{J}_n^* = \inf_{\mathbf{u}_n} J(\mathbf{x}(\mathbf{u}_n), \mathbf{u}_n)$, *where* $\mathbf{x}(\mathbf{u}_n)$ *is the solution of problem* (2) *corresponding to the piecewise-constantly extended discrete control* $\mathbf{u}_n$, *and if the infimum is taken over all those admissible controls belonging to* B_{K_1} (B_{K_2}) *and corresponding to an arbitrary fixed partition* Σ_{in}, *then, under the hypothesis of Theorem* 1,

$$\lim_{n\to+\infty} \tilde{J}_n^* = J^* \quad \text{if} \quad \tau_n \to 0 \quad \text{as} \quad n \to +\infty. \tag{16}$$

Proof of Lemma 1. By using the set of continuous functions in B_{K_1} (in B_{K_2}), which is everywhere dense in the metric of $L_2[t_0, T]$, we choose a continuous function $\mathbf{u}^{**}(t) \in B_{K_1}$ (B_{K_2}) such that from the smallness of the norm $\|\mathbf{u}^* - \mathbf{u}^{**}\|_{L_2[t_0,T]}$ it follows by (11) and (12) that

$$|J(\mathbf{u}^*) - J(\mathbf{u}^{**})| < \frac{\varepsilon}{2}, \tag{17}$$

where $\varepsilon > 0$ is a preassigned arbitrary number. We shall then take τ_n so small that the oscillation of $\mathbf{u}^{**}(t)$ on all the partial intervals $[t_{ni}, t_{ni+1}]$, $i = 0, 1, \cdots, n-1$, is less than a sufficiently small $\delta' > 0$. Then, the piecewise-constant function $\mathbf{u}_n^{\delta'}(t)$, coinciding on each of the partial intervals $[t_{ni}, t_{ni+1}]$, $i = 0, 1, \cdots, n-1$, with that value of $\mathbf{u}^{**}(t)$ on $[t_{ni}, t_{ni+1}]$ which is closest to zero, will approximate $\mathbf{u}^{**}(t)$ uniformly on $[t_0, T]$ with accuracy within δ'. If δ' is sufficiently small, then by assertion (ii) of Theorem 1 we have

$$|J(\mathbf{u}^{**}) - J(\mathbf{u}_n^{\delta'})| < \frac{\varepsilon}{2}. \tag{18}$$

Adding (17) and (18) we get

$$|J(\mathbf{u}^*) - J(\mathbf{u}_n^{\delta'})| < \varepsilon. \tag{19}$$

Obviously,

$$J(\mathbf{u}^*) \leqq J(\mathbf{u}_n^{\delta'}) \leqq J(\mathbf{u}^*) + \varepsilon. \tag{20}$$

Let the functional $J(\mathbf{u}_n) = J(\mathbf{x}(\mathbf{u}_n), \mathbf{u}_n)$, where $\mathbf{x}(\mathbf{u}_n)$ is the solution of the Cauchy problem (2) corresponding to the piecewise-constantly extended $\mathbf{u}_n$, have as its greatest lower bound over such $\mathbf{u}_n$ the number

$$\tilde{J}_n^* = \inf_{\mathbf{u}_n} J(\mathbf{x}(\mathbf{u}_n), \mathbf{u}_n).$$

Then from (20) it follows that

$$J^* = J(\mathbf{u}^*) \leqq \tilde{J}^* \leqq J(\mathbf{u}_n^{\delta'}) = J(\mathbf{x}(\mathbf{u}_n^{\delta'}), \mathbf{u}_n^{\delta'}) \leqq J(\mathbf{u}^*) + \varepsilon = J^* + \varepsilon$$

for all sufficiently large n (and, consequently, for sufficiently small

$$\tau_n = \max_{0 \leqq i \leqq n-1} \tau_{ni}).$$

Lemma 1 is proved.

Let $\tilde{x}(t) = \mathbf{x}(\mathbf{u}_n)$ and $\mathbf{x}_n = \mathbf{x}_n(\mathbf{u}_n)$ denote the solutions of problems (2) and (2′) corresponding to the same control $\mathbf{u}_n$, piecewise-constantly extended to obtain $\mathbf{x}(\mathbf{u}_n) = \tilde{x}(t)$.

Lemma 2. *Under the hypotheses of Lemma* 1,

$$z_{ni} = |\tilde{x}(t_{ni}) - \mathbf{x}_{ni}| \underset{i}{\rightrightarrows} 0 \quad as \quad n \to +\infty \tag{21}$$

if

$$\tau_n = \max_{0 \leqq i \leqq n-1} \tau_{ni} = O(n^{-1}) \quad as \quad n \to +\infty. \tag{22}$$

Proof of Lemma 2. We have

$$\tilde{x}(t_{ni+1}) = \tilde{x}(t_{ni}) + \int_{t_{ni}}^{t_{ni+1}} \mathbf{f}(\tilde{x}(\tau), \mathbf{u}_{ni}, \tau)\, d\tau, \tag{23}$$

$$\mathbf{x}_{ni+1} = \mathbf{x}_{ni} + \int_{t_{ni}}^{t_{ni+1}} \mathbf{f}(\mathbf{x}_{ni}, \mathbf{u}_{ni}, t_{ni})\, d\tau. \tag{24}$$

Subtracting (24) from (23) and going over to inequalities, we get

$$\begin{aligned} z_{ni+1} \leqq z_{ni} &+ \int_{t_{ni}}^{t_{ni+1}} |\mathbf{f}(\tilde{x}(\tau), \mathbf{u}_{ni}, \tau) - \mathbf{f}(\tilde{x}(t_{ni}), \mathbf{u}_{ni}, \tau)|\, d\tau \\ &+ \int_{t_{ni}}^{t_{ni+1}} |\mathbf{f}(\tilde{x}(t_{ni}), \mathbf{u}_{ni}, \tau) - \mathbf{f}(\mathbf{x}_{ni}, \mathbf{u}_{ni}, \tau)|\, d\tau \\ &+ \int_{t_{ni}}^{t_{ni+1}} |\mathbf{f}(\mathbf{x}_{ni}, \mathbf{u}_{ni}, \tau) - \mathbf{f}(\mathbf{x}_{ni}, \mathbf{u}_{ni}, t_{ni})|\, d\tau \\ &\leqq z_{ni} + A_4 \int_{t_{ni}}^{t_{ni+1}} |\tilde{x}(\tau) - \tilde{x}(t_{ni})|\, d\tau + A_4 \int_{t_{ni}}^{t_{ni+1}} |\tilde{x}(t_{ni}) - \mathbf{x}_{ni}|\, d\tau \\ &+ \int_{t_{ni}}^{t_{ni+1}} |\mathbf{f}(\mathbf{x}_{ni}, \mathbf{u}_{ni}, \tau) - \mathbf{f}(\mathbf{x}_{ni}, \mathbf{u}_{ni}, t_{ni})|\, d\tau \\ &\leqq (1 + A_4\tau_n) z_{ni} + r_{ni}\tau_n, \end{aligned} \tag{25}$$

where $r_{ni} = o(1)$ for all $i = 0, 1, \cdots, n-1$ except, possibly, for certain values $i_1, i_2, \cdots, i_s$, where $s \leqq m_0 = \text{const.}$ for all $n = 1, 2, \cdots$; for such $i = i_1, i_2, \cdots, i_s$ we have $r_i = O(1)$ because $\mathbf{f}(\mathbf{x}, \mathbf{u}, t)$ is piecewise continuous in t with a uniformly bounded number of points of discontinuity. Taking into account that $\tau_n \leqq C/n$ and $z_{n0} = 0$, from (25) we obtain the estimate

$$0 \leqq z_{ni} \leqq o(1)[e^{A_4C} - 1] + \tau_n O(1) m_0\, e^{A_4C}, \tag{26}$$

whence it follows that $z_{ni} \underset{i}{\rightrightarrows} 0$ as $n \to +\infty$ (and $\tau_n = O(n^{-1})$). Lemma 2 is proved.

Lemma 3. *Under the hypotheses of Lemmas* 1 *and* 2 *and under conditions* (a), (b), (c),

$$|J(\mathbf{x}(\mathbf{u}_n), \mathbf{u}_n) - J_n(\mathbf{x}_n(\mathbf{u}_n), \mathbf{u}_n)| \leqq \alpha_n \to 0 \quad as \quad n \to +\infty. \tag{27}$$

Proof of Lemma 3. We have

$$
\begin{aligned}
&|J(\mathbf{x}(\mathbf{u}_n), \mathbf{u}_n) - J_n(\mathbf{x}_n(\mathbf{u}_n), \mathbf{u}_n)| \\
&\quad \leqq \left| \int_{t_0}^{T} g(\tilde{x}(\tau), \mathbf{u}_n, \tau)\, d\tau - \sum_{i=0}^{n-1} g(\mathbf{x}_{ni}, \mathbf{u}_{ni}, t_{ni})\tau_{ni} \right| + |\Phi(\tilde{x}(T)) - \Phi(\mathbf{x}_{nn})| \\
&\quad \leqq \left| \sum_{i=0}^{n-1} \int_{t_{ni}}^{t_{ni+1}} [g(\tilde{x}(\tau), \mathbf{u}_{ni}, \tau) - g(\mathbf{x}_{ni}, \mathbf{u}_{ni}, t_{ni})]\, d\tau \right| + |\Phi(\tilde{x}(T)) - \Phi(\mathbf{x}_{nn})| \qquad (28) \\
&\quad \leqq \sum_{i=0}^{n-1} \int_{t_{ni}}^{t_{ni+1}} |g(\tilde{x}(\tau), \mathbf{u}_{ni}, \tau) - g(\mathbf{x}_{ni}, \mathbf{u}_{ni}, \tau)|\, d\tau \\
&\quad\quad + \sum_{i=0}^{n-1} \int_{t_{ni}}^{t_{ni+1}} |g(\mathbf{x}_{ni}, \mathbf{u}_{ni}, \tau) - g(\mathbf{x}_{ni}, \mathbf{u}_{ni}, t_{ni})|\, d\tau + |\Phi(\tilde{x}(T)) - \Phi(\mathbf{x}_{nn})|.
\end{aligned}
$$

By virtue of the properties of $g(\mathbf{x}, \mathbf{u}, t)$ and $\Phi(\mathbf{x})$ and of relation (21) in Lemma 2, we get (27) from (28). Here we use the points of discontinuity of $g(\mathbf{x}, \mathbf{u}, t)$ in t just as we used the points of discontinuity of $\mathbf{f}(\mathbf{x}, \mathbf{u}, t)$ in t when proving the preceding lemma.

We now go on to the proof of Theorem 2. In (27) we first substitute $\tilde{\mathbf{u}}_n^*$ for $\mathbf{u}_n$, for which

$$J(\mathbf{x}(\mathbf{u}_n^*), \mathbf{u}_n^*) = \inf_{\mathbf{u}_n} J(\mathbf{x}(\mathbf{u}_n), \mathbf{u}_n) = \tilde{J}_n^*,$$

and next we substitute $\mathbf{u}_n^*$, for which

$$J_n(\mathbf{x}_n(\mathbf{u}_n^*), \mathbf{u}_n^*) = \inf_{\mathbf{u}_n} J_n(\mathbf{x}_n(\mathbf{u}_n), \mathbf{u}_n) = J_n^*.$$

Then, by (27) we get

$$\tilde{J}_n^* - J_n^* \geqq \tilde{J}_n^* - J_n(\mathbf{x}_n(\tilde{\mathbf{u}}_n^*), \tilde{\mathbf{u}}_n^*) \geqq -\alpha_n, \qquad (29)$$

$$\tilde{J}_n^* - J_n^* \leqq J(\mathbf{x}(\mathbf{u}_n^*), \mathbf{u}_n^*) - J_n^* \leqq \alpha_n. \qquad (30)$$

From (29) and (30) we find

$$|\tilde{J}_n^* - J_n^*| \leqq \alpha_n \to 0 \quad \text{as} \quad n \to +\infty. \qquad (31)$$

Adding (16) and (31) we get (15). Theorem 2 is proved.

THEOREM 3. *The sequence* $\{\mathbf{u}_n^*\}$, *where the* $\mathbf{u}_n^*$ *are determined from the condition*

$$\inf_{\mathbf{u}_n} J_n(\mathbf{x}_n(\mathbf{u}_n), \mathbf{u}_n) = J_n(\mathbf{x}(\mathbf{u}_n^*), \mathbf{u}_n^*), \qquad (32)$$

is a minimizing sequence for the functional $J(\mathbf{x}(\mathbf{u}), \mathbf{u})$ *in* B_{K_1} (B_{K_2}).

Proof. From (15) and (27) we get that

$$|J(\mathbf{x}(\mathbf{u}_n^*), \mathbf{u}_n^*) - J^*| \to 0 \quad \text{as} \quad n \to +\infty \qquad (33)$$

and $0 < \tau_n < Cn^{-1}$. The theorem is proved.

COROLLARY TO THEOREM 3. *If the functional* $J(\mathbf{x}(\mathbf{u}), \mathbf{u})$ *is sufficiently well-behaved* (*see* [2]), *then from the convergence in* (33) *there follows the convergence of the minimizing sequence of* $\mathbf{u}_n^*$ *to* u^*.

This will hold automatically, for example, if:

(i) $g(\mathbf{x}, \mathbf{u}, t)$ is convex in $\mathbf{x}$ and uniformly convex in $\mathbf{u}$, i.e., if

$$g\left(\frac{\mathbf{x}_1 + \mathbf{x}_2}{2}, \frac{\mathbf{u}_1 + \mathbf{u}_2}{2}, t\right) \leqq \frac{1}{2}g(\mathbf{x}_1, \mathbf{u}_1, t) + \frac{1}{2}g(\mathbf{x}_2, \mathbf{u}_2, t) - \delta(\|\mathbf{u}_1 - \mathbf{u}_2\|)$$

for any $\mathbf{x}_1, \mathbf{x}_2, \mathbf{u}_1, \mathbf{u}_2$ and $t \in [t_0, T]$; $\delta(\tau)$ is continuous, $\delta(0) = 0$, $\delta(\tau) > 0$ when $\tau > 0$;

(ii) $\Phi(\mathbf{x})$ is convex in $\mathbf{x}$, i.e.,

$$\Phi\left(\frac{\mathbf{x}_1 + \mathbf{x}_2}{2}\right) \leqq \frac{1}{2}\Phi(\mathbf{x}_1) + \frac{1}{2}\Phi(\mathbf{x}_2)$$

and, further,

(iii) $\mathbf{f}(\mathbf{x}, \mathbf{u}, t) = A(t)\mathbf{x} + B(t)\mathbf{u}(t)$.

Then the functional $J(\mathbf{x}(\mathbf{u}), \mathbf{u})$ will be uniformly convex, i.e.,

$$J\left(\frac{\mathbf{u}_1 + \mathbf{u}_2}{2}\right) \leqq \frac{1}{2}J(\mathbf{u}_1) + \frac{1}{2}J(\mathbf{u}_2) - \delta(\|\mathbf{u}_1 - \mathbf{u}_2\|)$$

for all $\mathbf{u}_1$ and $\mathbf{u}_2$. It is well known that if the set U of controls $\mathbf{u}$ is closed and convex, while the functional $J(\mathbf{x}(\mathbf{u}), \mathbf{u})$ is uniformly convex, then every minimizing sequence $\{\mathbf{u}_n\}$ converges to a unique minimum point $\mathbf{u}^*$ of the functional $J(\mathbf{x}(\mathbf{u}), \mathbf{u})$ on this set and, moreover,

$$\delta(\|\mathbf{u}_n - \mathbf{u}^*\|) \leqq J(\mathbf{u}_n) - J(\mathbf{u}^*) \quad \textit{for all} \quad \mathbf{u}_n \in U.$$

In particular, if $\delta(z) = \gamma z^2, \gamma = \text{const.} > 0$, then the functional is said to be strongly convex and we get

$$(33') \qquad \|\mathbf{u}_n - \mathbf{u}^*\|^2 \leqq \frac{1}{\gamma}[J(\mathbf{u}_n) - J^*].$$

Remark 2. The results obtained in Theorems 1–3 can be easily extended to the case when there are no minimizing elements in the functionals but there are greatest lower bounds and, consequently, minimizing sequences, and also to the case of other boundary conditions at $t = t_0$ and $t = T$.

Remark 3. If the optimal control is piecewise continuous, if $\mathbf{u}^*(t)$, $\mathbf{f}(\mathbf{x}, \mathbf{u}, t)$ and $g(\mathbf{x}, \mathbf{u}, t)$ satisfy all the formulated conditions, if Hölder conditions with index $\alpha, 0 < \alpha < 1$, are satisfied in t on all the intervals of continuity in t, and if, furthermore, $g(\mathbf{x}, \mathbf{u}, t)$ and $\Phi(\mathbf{x})$ satisfy a Lipschitz condition in $\mathbf{x}$, then we get

$$(34) \qquad |J_n^* - J^*| \leqq C\tau_n^\alpha.$$

Thus in the case of strong convexity of the functional, from (33') and (34) we find

$$(35) \qquad \|\mathbf{u}_n^* - \mathbf{u}^*\| \leqq \frac{C}{\gamma}\tau_n^\alpha.$$

Remark 4. It is of interest to investigate the speed of convergence of the solution of the approximating difference optimization problem to the solution of the original differential optimization problem as a function of the smoothness of the differential problem and of the approximation order of the difference scheme. In the case of analytic $\mathbf{u}^*$, $\mathbf{f}$, g, Φ, it can be shown that when the approximation of the difference scheme is of order $O(\tau_n^p)$, the speed of convergence of the solution of the difference optimization problem to the solution of the differential optimization problem also equals $O(\tau_n^p)$.

Remark 5. If the functional $J(\mathbf{u})$ is not uniformly convex and reaches an absolute minimum on B_{K_2} at several different elements of B_{K_2}, then in general there is no basis for expecting that every minimizing sequence for $J(\mathbf{u})$ will converge strongly to one of these elements. Following A. N. Tikhonov we shall say that the optimal control problem is not well-posed if there exists a minimizing sequence which does not converge strongly to any of the elements at which the functional achieves an absolute minimum. Using an idea analogous to the ideas in [3]–[7], [9] we consider one variant of the regularization of difference optimal control problems (1′), (2′) approximating an original nonwell-posed optimal control problem (1), (2), assuming that the functional (1) is convex. This version of regularization ensures the strong convergence in $L_2[t_0, T]$ of the minimizing sequence for (1), (2), obtained on the basis of difference approximations (1′), (2′), to the element of smallest norm at which the functional (1) reaches an absolute minimum in B_{K_2}, if the regularization parameter is compatible with the errors admissible in the solution of problem (1′), (2′).

We note first that the set $U^* \subseteqq B_{K_2}$ of elements at each of which $J(\mathbf{u})$ reaches an absolute minimum in B_{K_2} is convex because of the assumption of convexity of $J(\mathbf{u})$. By assuming that the hypothesis of Theorem 1 is satisfied we conclude that the set U^* will be (strongly) closed by virtue of the continuity of $J(\mathbf{u})$. Since a closed, convex set in $L_2[t_0, T]$ is strongly closed, while a convex, continuous functional is lower semicontinuous, there exists an element $\mathbf{u}^*_{\min} \in U^*$ which is smallest in norm; moreover, this element is unique.

DEFINITION 1. The function $\mathbf{u}(t)$ will be called *generalized uniformly continuous* on the interval $[t_0, T]$ if the set of its points of discontinuity on $[t_0, T]$ can be covered by a finite system of intervals whose total length is arbitrarily small.

DEFINITION 2. The set of functions $\{\mathbf{u}(t)\}$ will be called *generalized equicontinuous* on $[t_0, T]$ if for every $\eta > 0$ and for each function $\mathbf{u}(t)$ from $\{\mathbf{u}(t)\}$ we can find a finite system of intervals of total length less than η, covering the points of discontinuity of the function, such that for any $\varepsilon > 0$ we can find a $\delta = \delta(\varepsilon) > 0$ which possesses the following property: whatever t' and t'' we choose from $[t_0, T]$, lying outside the system of intervals indicated for the function $\mathbf{u}(t) \in \{\mathbf{u}(t)\}$ and satisfying the condition $|t' - t''| < \delta(\varepsilon)$, there is satisfied the inequality $|\mathbf{u}(t') - \mathbf{u}(t'')| < \varepsilon$. Everywhere in what follows we shall assume that the set U^* is generalized equicontinuous.

Let $\mathbf{u}_n(t)$ be a piecewise-constantly extended discrete control $\mathbf{u}_n$, in accordance with the definition given above. Then

$$J_n(\mathbf{u}_n) = J(\mathbf{u}_n(t)) + \Theta_n(\mathbf{u}_n(t)),$$

and it is not difficult to prove that under the conditions stated above,

$$|\Theta_n(\mathbf{u}_n)| < \gamma_n \to 0 \quad \text{as} \quad n \to +\infty, \qquad \tau_n = O(n^{-1}).$$

If the difference problem (2') is solved with roundoff, if the functional $J_n(\mathbf{u}_n)$ is computed with roundoff and if, moreover, the roundoff error is of order $\delta_n = o(n^{-1})$ as $n \to +\infty$, then, analogous to what was done in [8], we get

$$J_{n\delta_n}(\mathbf{u}_n) = J(\mathbf{u}_n(t)) + \Theta_{n\delta_n}(\mathbf{u}_n(t)). \tag{36}$$

Here $J_{n\delta_n}(\mathbf{u}_n)$ is the value of the discretized functional $J_n(\mathbf{u}_n)$ computed with roundoff, on the rounded-off solution of problem (2'), while

$$|\Theta_{n\delta_n}(\mathbf{u}_n)| \leqq \gamma_{n\delta_n} \to 0 \quad \text{as} \quad n \to +\infty, \qquad \tau_n = O(n^{-1}), \qquad \delta_n = o(n^{-1}).$$

We now consider the regularized discrete functional

$$J_{n\delta_n\alpha}(\mathbf{u}_n) = J_{n\delta_n}(\mathbf{u}_n) + \alpha\|\mathbf{u}_n\|^2,$$

where the norm is taken in $L_2[t_0, T]$. By virtue of (36) we have

$$J_{n\delta_n\alpha}(\mathbf{u}_n) = J(\mathbf{u}_n(t)) + \Theta_{n\delta_n}(\mathbf{u}_n) + \alpha\|\mathbf{u}_n\|^2. \tag{37}$$

Let $\mathbf{u}^*(t) \in U^*$; we denote the piecewise-constantly extended discrete control defined by the relations

$$\mathbf{u}^*_{ni} = \mathbf{u}^*(t_{ni}), \qquad i = 0, 1, \cdots, n-1,$$

by $(\mathbf{u}^*)_n = (\mathbf{u}^*(t))_n$. Then by virtue of the generalized equicontinuity of the family U^*,

$$\Delta_n J = \{J((\mathbf{u}^*)_n) - J(\mathbf{u}^*)\} \to 0 \quad \text{as} \quad n \to +\infty, \qquad \tau_n = O(n^{-1}),$$

and uniformly relative to $\mathbf{u}^*(t) \in U^*$.

Given some sequence $\varepsilon_n \downarrow 0$ as $n \to +\infty$, we choose some sequence of values of the regularizing parameter $\{\alpha_n\}$ satisfying the following two requirements:[3]

$$\alpha_n \downarrow 0 \qquad\qquad \text{as} \quad n \to +\infty,$$
$$(2\gamma_{n\delta_n} + \varepsilon_n + |\Delta_n J|')/\alpha_n \to 0 \quad \text{as} \quad n \to +\infty.$$

For each n we choose some difference net Σ_{tn}, satisfying the condition $\tau_n = O(n^{-1})$ as $n \to +\infty$, and $\mathbf{u}_n^{(\alpha_n,\delta_n,\varepsilon_n)} \in B_{K_2}$ for which

$$J^*_{n\delta_n\alpha_n} = \inf_{B_{K_2}} J_{n\delta_n\alpha_n} \leqq J_{n\delta_n\alpha_n}(\mathbf{u}_n^{(\alpha_n,\delta_n,\varepsilon_n)}) < J^*_{n\delta_n\alpha_n} + \varepsilon_n \tag{38}$$

on the chosen net Σ_{tn}.

THEOREM 4. *Under the conditions formulated above,* $\{\mathbf{u}_n^{(\alpha_n,\delta_n,\varepsilon_n)}\}$ *is a minimizing sequence for the functional* $J(\mathbf{u})$ *on* B_{K_2}, *i.e.,* $J(\mathbf{u}_n^{(\alpha_n,\delta_n,\varepsilon_n)}) \to J^* = \inf_{B_{K_2}} J(\mathbf{u})$ *as* $n \to +\infty$.

[3] $|\Delta_n J|' = \sup_{\mathbf{u}^* \in U^*} |\Delta_n J|$ as $n \to +\infty$.

Proof. We take some $\mathbf{u}^*(t) \in U^*$; then

$$J(\mathbf{u}^*) = J^*$$

and

$$\begin{aligned}
0 &\leqq J(\mathbf{u}_n^{(\alpha_n,\delta_n,\varepsilon_n)}(t)) - J(\mathbf{u}^*) \\
&\leqq J(\mathbf{u}_n^{(\alpha_n,\delta_n,\varepsilon_n)}(t)) + \Theta_{n\delta_n}(\mathbf{u}_n^{(\alpha_n,\delta_n,\varepsilon_n)}(t)) + \gamma_{n\delta_n} + \alpha_n \|\mathbf{u}_n^{(\alpha_n,\delta_n,\varepsilon_n)}(t)\|^2 - J(\mathbf{u}^*) \\
&\leqq J_{n\delta_n\alpha_n}(\mathbf{u}_n^{(\alpha_n,\delta_n,\varepsilon_n)}) + \gamma_{n\delta_n} - J(\mathbf{u}^*) \\
&\leqq J^*_{n\delta_n\alpha_n} + \varepsilon_n + \gamma_{n\delta_n} - J(\mathbf{u}^*) \\
&< J_{n\delta_n\alpha_n}((\mathbf{u}^*)_n) + \varepsilon_n + \gamma_{n\delta_n} - J(\mathbf{u}^*) \\
&= J((\mathbf{u}^*)_n) - J(\mathbf{u}^*) + \Theta_{n\delta_n}((\mathbf{u}^*)_n) + \alpha_n \|(\mathbf{u}^*)_n\|^2 + \varepsilon_n + \gamma_{n\delta_n} \\
&\leqq \{|\Delta_n J|' + 2\gamma_{n\delta_n} + \varepsilon_n + \alpha_n \|(\mathbf{u}^*)_n\|^2\} \\
&\to 0 \quad \text{as} \quad n \to +\infty,
\end{aligned}$$

i.e.,

$$\lim_{n \to +\infty} J(\mathbf{u}^{(\alpha_n,\delta_n,\varepsilon_n)}) = J(\mathbf{u}^*). \tag{39}$$

Theorem 4 is proved.

Remark 6. By virtue of (39) and (32) we also have

$$\lim_{n \to +\infty} J_{n\delta_n\alpha_n}(\mathbf{u}_n^{(\alpha_n,\delta_n,\varepsilon_n)}) = J(\mathbf{u}^*). \tag{40}$$

THEOREM 5. *The sequence* $\{\mathbf{u}_n^{(\alpha_n,\delta_n,\varepsilon_n)}\}$ *converges strongly to* $\mathbf{u}^*_{\min}$ *under the conditions formulated above.*

Proof. By virtue of (37) and (38), for any $\mathbf{u}^*(t) \in U^*$ we have

$$\begin{aligned}
J_{n\delta_n\alpha_n}(\mathbf{u}_n^{(\alpha_n,\delta_n,\varepsilon_n)}) &< J^*_{n\delta_n\alpha_n} + \varepsilon_n \leqq J_{n\delta_n\alpha_n}((\mathbf{u}^*)_n) + \varepsilon^n \\
&= J((\mathbf{u}^*)_n) - \Theta_{n\delta_n}((\mathbf{u}^*)_n) + \alpha_n \|(\mathbf{u}^*)_n\|^2 + \varepsilon_n \\
&\leqq J(\mathbf{u}^*) + \{J((\mathbf{u}^*)_n) - J(\mathbf{u}^*)\} + \gamma_{n\delta_n} + \alpha_n \|(\mathbf{u}^*)_n\|^2 + \varepsilon_n \\
&\leqq J(\mathbf{u}^*) + \gamma_{n\delta_n} + |\Delta_n J|' + \varepsilon_n + \alpha_n \|(\mathbf{u}^*)_n\|^2,
\end{aligned} \tag{41}$$

$$\begin{aligned}
J_{n\delta_n\alpha_n}(\mathbf{u}_n^{(\alpha_n,\delta_n,\varepsilon_n)}) &= J(\mathbf{u}_n^{(\alpha_n,\delta_n,\varepsilon_n)}(t)) + \alpha_n \|\mathbf{u}_n^{(\alpha_n,\delta_n,\varepsilon_n)}(t)\|^2 + \Theta_{n\delta_n}(\mathbf{u}_n^{(\alpha_n,\delta_n,\varepsilon_n)}) \\
&\geqq J(\mathbf{u}_n^{(\alpha_n,\delta_n,\varepsilon_n)}) + \alpha_n \|\mathbf{u}_n^{(\alpha_n,\delta_n,\varepsilon_n)}\|^2 - \gamma_{n\delta_n}.
\end{aligned} \tag{42}$$

From (41) and (42) we find

$$\begin{aligned}
\alpha_n \|\mathbf{u}_n^{(\alpha_n,\delta_n,\varepsilon_n)}\|^2 &\leqq J(\mathbf{u}^*) - J(\mathbf{u}_n^{(\alpha_n,\delta_n,\varepsilon_n)}) + 2\gamma_{n\delta_n} + \varepsilon_n + |\Delta_n J|' + \alpha_n \|(\mathbf{u}^*)_n\|^2 \\
&\leqq 2\gamma_{n\delta_n} + \varepsilon_n + |\Delta_n J|' + \alpha_n \|(\mathbf{u}^*)_n\|^2,
\end{aligned} \tag{43}$$

whence

$$\|\mathbf{u}_n^{(\alpha_n,\delta_n,\varepsilon_n)}\|^2 \leqq \frac{2\gamma_{n\delta_n} + \varepsilon_n + |\Delta_n J|'}{\alpha_n} + \|(\mathbf{u}^*)_n\|^2. \tag{44}$$

The sequence $\{\mathbf{u}_n^{(\alpha_n,\delta_n,\varepsilon_n)}\}$ is bounded in norm in $L_2[t_0, T]$ and so we can select a weakly convergent subsequence $\{\mathbf{u}_{n'}^{(\alpha_{n'},\delta_{n'},\varepsilon_{n'})}\}$; we denote its weak limit by $\check{\mathbf{u}}$. Taking into account the strong convergence of $(\mathbf{u}^*)_n$ to $\mathbf{u}^*$ and passing to the limit as $n = n' \to +\infty$ in (44), we get

$$(45) \qquad \limsup_{n' \to +\infty} \|\mathbf{u}_{n'}^{(\alpha_{n'},\delta_{n'},\varepsilon_{n'})}\| \leqq \|\mathbf{u}^*\|.$$

On the other hand, because of the weak lower semicontinuity of the norm we have

$$(46) \qquad \|\mathbf{u}\| \leqq \liminf_{n' \to +\infty} \|\mathbf{u}_{n'}^{(\alpha_{n'},\delta_{n'},\varepsilon_{n'})}\|.$$

Comparing (45) and (46) we find

$$(47) \qquad \|\check{\mathbf{u}}\| \leqq \|\mathbf{u}^*\|.$$

By virtue of the weak closedness of B_{K_2} we have $\check{\mathbf{u}} \in B_{K_2}$. But by the weak lower semicontinuity of $J(\mathbf{u})$ and because of (39) we have

$$(48) \qquad J(\check{\mathbf{u}}) \leqq \liminf_{n' \to +\infty} J(\mathbf{u}_{n'}^{(\alpha_{n'},\delta_{n'},\varepsilon_{n'})}) = \lim_{n' \to +\infty} J(\mathbf{u}_{n'}^{(\alpha_{n'},\delta_{n'},\varepsilon_{n'})}) = J(\mathbf{u}^*).$$

Consequently,

$$(49) \qquad J(\check{\mathbf{u}}) = J(\mathbf{u}^*), \qquad \check{\mathbf{u}} \in U^*.$$

The $\mathbf{u}^* \in U^*$ occurring in the right-hand side of (47) is arbitrary and, therefore, $\check{\mathbf{u}}$ is the element in U^* which is smallest in norm, i.e., $\check{\mathbf{u}} = \mathbf{u}^*_{\min}$. Setting $\mathbf{u}^* = \mathbf{u}^*_{\min}$ and $\check{\mathbf{u}} = \mathbf{u}^*_{\min}$ in (45) and (46) we get

$$(50) \qquad \|\mathbf{u}^*_{\min}\| \leqq \liminf_{n' \to +\infty} \|\mathbf{u}_{n'}^{(\alpha_{n'},\delta_{n'},\varepsilon_{n'})}\| \leqq \limsup_{n' \to +\infty} \|\mathbf{u}_{n'}^{(\alpha_{n'},\delta_{n'},\varepsilon_{n'})}\| \leqq \|\mathbf{u}^*_{\min}\|.$$

Consequently,

$$(51) \qquad \lim_{n' \to +\infty} \|\mathbf{u}_{n'}^{(\alpha_{n'},\delta_{n'},\varepsilon_{n'})}\| = \|\mathbf{u}^*_{\min}\|.$$

On the other hand we have proved that

$$(52) \qquad \mathbf{u}_{n'}^{(\alpha_{n'},\delta_{n'},\varepsilon_{n'})} \xrightarrow{\text{weakly}} \mathbf{u}^*_{\min} \quad \text{as} \quad n' \to +\infty.$$

From (51) and (52), we have

$$(53) \qquad \|\mathbf{u}_{n'}^{(\alpha_{n'},\delta_{n'},\varepsilon_{n'})} - \mathbf{u}^*_{\min}\| \to 0 \quad \text{as} \quad n' \to +\infty,$$

i.e., $\mathbf{u}_{n'}^{(\alpha_{n'},\delta_{n'},\varepsilon_{n'})}$ converges strongly to $\mathbf{u}^*_{\min}$. If it should happen that the whole sequence $\{\mathbf{u}_n^{(\alpha_n,\delta_n,\varepsilon_n)}\}$, unlike the subsequence $\{\mathbf{u}_{n'}^{(\alpha_{n'},\delta_{n'},\varepsilon_{n'})}\}$, does not converge strongly to $\mathbf{u}^*_{\min}$, then we would find $\eta_0 > 0$ and a subsequence $\{\mathbf{u}_{n''}^{(\alpha_{n''},\delta_{n''},\varepsilon_{n''})}\}$ such that the relation

$$(54) \qquad \|\mathbf{u}_{n''}^{(\alpha_{n''},\delta_{n''},\varepsilon_{n''})} - \mathbf{u}^*_{\min}\| \geqq \eta_0 \quad \text{as} \quad n'' \to +\infty,$$

would be satisfied. Because the subsequence $\{\mathbf{u}_{n''}^{(\alpha_{n''},\delta_{n''},\varepsilon_{n''})}\}$ is bounded in norm, from it we can select a weakly convergent subsequence $\{\mathbf{u}_{n'''}^{(\alpha_{n'''},\delta_{n'''},\varepsilon_{n'''})}\}$, as in the case of the subsequence $\{\mathbf{u}_{n'}^{(\alpha_{n'},\delta_{n'},\varepsilon_{n'})}\}$ we could establish the strong convergence of $\{\mathbf{u}_{n'''}^{(\alpha_{n'''},\delta_{n'''},\varepsilon_{n'''})}\}$ to $\mathbf{u}^*_{\min}$, and we would arrive at a contradiction. Theorem 5 is proved.

Remark 7. The results obtained in Theorems 1 and 2 allow us to prove the convergence of the difference approximations in optimal control problems when the difference net is introduced not only on the t-axis but also in the phase space of the coordinates $\mathbf{x}$ and in the space of the controls $\mathbf{u}$; in particular, this leads to the proof of the convergence of the discrete version of dynamic programming for solving problem (1), (2) .

On each axis x^s, $s = 1, 2, \cdots, N$, we introduce the net $\Sigma_{x^s n}$ of nodes $\{x^s_{nk}\}$, $k = 0, \pm 1, \pm 2, \cdots, \pm N^s_{xn}$, where $x^s_{nk'} < x^s_{nk''}$ when $-N^s_{xn} \leqq k' < k'' \leqq N^s_{xn}$, $s = 1, 2, \cdots, N$, $n = 0, 1, \cdots$. In the space $E^N = E^N(x^1, \cdots, x^N)$ we define the net Σ_{xn} as the metric product[4] of the nets $\Sigma_{x^s n}$, $s = 1, 2, \cdots, N$; i.e., its nodes are the endpoints of the vectors $\check{\mathbf{x}} = (\check{\mathbf{x}}^1, \cdots, \check{\mathbf{x}}^N)$ in which $\check{\mathbf{x}}^s$ coincides with some of the nodes $\{x^s_{nk}\}$, $|k| \leqq N^s_{xn}$, $s = 1, \cdots, n$. Thus, the whole net Σ_{xn} contains $2^N \Pi^N_{s=1} N^s_{xn}$ N-dimensional nodes, represented by the endpoints of the vectors $\check{\mathbf{x}} = (\check{\mathbf{x}}^1, \check{\mathbf{x}}^2, \cdots, \check{\mathbf{x}}^N)$.

The net Σ_{xn} in E^N will be called compatible with the net Σ_{tn} on the t-axis, $t_0 = t_{n0} < \cdots < t_{nn} = T$, if the following requirement is fulfilled. Let $\tau_n = \max_{0 \leqq i \leqq n-1} \tau_{ni}$, where

$$\tau_{ni} = t_{ni+1} - t_{ni}, \qquad i = 0, 1, \cdots, n-1,$$

while

$$\Delta x_n = \max_{\substack{|k| \leqq N^s_{xn} \\ s = 1, \cdots, N}} \Delta x^s_{nk},$$

where $\Delta x^s_{nk} = x^s_{nk+1} - x^s_{nk}$; then we require that

$$\Delta x_n = o(\tau_n) = \tau_n o_n(1), \tag{55}$$

where $o_n(1) \to 0$ as $n \to +\infty$.

We introduce, further, the operation of the "projection" of the vector $\mathbf{x} = (x^1, \cdots, x^N)$ onto the net Σ_{xn}, whereby we associate with the vector $\mathbf{x} = (x^1, \cdots, x^N)$ the vector $\check{\mathbf{x}} = \langle \mathbf{x} \rangle = (\langle x^1 \rangle, \cdots, \langle x^N \rangle)$, where $\langle x^s \rangle$ denotes the node of the net $\Sigma_{x^s n}$ closest to x^s. If it happens that two nodes x^s_{nk} and x^s_{nk+1} are closest to x^s (i.e., they are equidistant from x^s, $x^s - x^s_{nk} = x^s_{nk+1} - x^s$), then we shall choose the "upper" one of them, i.e., x^s_{nk+1}. In the space $E^r = E^r(u^1, \cdots, u^r)$ we introduce, further, the difference net $\Sigma_{un} = (\Sigma_{u^1 n}, \cdots, \Sigma_{u^r n})$ such that

$$\Delta u_n = \max_{\substack{k \\ s = 1, \cdots, r}} (u^s_{nk+1} - u^s_{nk}) = o_n(1) \to 0 \quad \text{as} \quad n \to +\infty \tag{56}$$

(here, for each of the nets $\Sigma_{u^s n}$ it is assumed that the inequality $u^s_{nk'} < u^s_{nk''}$ is fulfilled for its nodes if $k' < k''$). By $\check{\mathbf{u}}_n = (\check{u}^1, \check{u}^2, \cdots, \check{u}^r)$ we denote vectors whose endpoints coincide with the nodes of the net Σ_{un}.

Let us replace problem (1′), (2′) by the problem of minimizing the functional

$$J_n(\check{\mathbf{x}}(\check{\mathbf{u}}_n), \check{\mathbf{u}}_n) = \sum_{i=1}^{n-1} g(\check{\mathbf{x}}_{ni}, \check{\mathbf{u}}_{ni}, t_{ni}) \tau_{ni} + \Phi(\check{\mathbf{x}}_{nn}) \tag{57}$$

[4] We shall assume that the net Σ_{xn} has been previously extended in such a way that the discrete version of the dynamic programming process is satisfied.

on the solutions of the modified Cauchy difference problem

$$\tilde{\mathbf{x}}_{ni+1} = \check{\mathbf{x}}_{ni} + \mathbf{f}(\check{\mathbf{x}}_{ni}, \check{\mathbf{u}}_{ni}, t_{ni})\tau_{ni}, \qquad i = 0, 1, \cdots, n-1, \tag{58}$$

$$\check{\mathbf{x}}_{n+1} = \langle \tilde{\mathbf{x}}_{ni+1} \rangle, \qquad \check{\mathbf{x}}_{n0} = \langle \mathbf{x}_0 \rangle \tag{59}$$

corresponding to all possible admissible controls:

$$\mathbf{u}_n \in \textstyle\sum_{un} \cap B_{K_1} \quad (\textstyle\sum_{un} \cap B_{K_2}).$$

If we introduce the operation of the projection of the vector $\mathbf{u}_n = (u^1, \cdots, u^r)$ onto the net Σ_{un}, whereby we associate with the vector $\mathbf{u}_n$ the vector $\check{\mathbf{u}}_n = \langle \check{\mathbf{u}}_n \rangle = (\langle u_n^1 \rangle, \cdots, \langle u_n^r \rangle)$, where $\langle u^s \rangle$ is the node of net $\Sigma_{u^s n}$ closest to u^s (or the "upper" one of two equidistant closest ones), then by virtue of (56) we have

$$|\mathbf{u}_n - \langle \mathbf{u}_n \rangle| = o_n(1) \to 0 \quad \text{as} \quad n \to +\infty. \tag{60}$$

As $\mathbf{u}_n$ ranges over B_{K_1} (B_{K_2}), then $\langle \mathbf{u}_n \rangle$ automatically ranges over $\Sigma_{un} \cap B_{K_1}$ $(\Sigma_{un} \cap B_{K_2})$. Therefore, in problem (58), (59) we can set $\check{\mathbf{u}} = \langle \mathbf{u}_n \rangle$, interpreting it as the problem of minimizing the functional

$$J_n(\check{\check{\mathbf{x}}}, \langle \mathbf{u}_n \rangle) = \sum_{i=0}^{n-1} g(\check{\check{\mathbf{x}}}_{ni}, \langle \mathbf{u}_{ni} \rangle, t_{ni})\tau_{ni} + \Phi(\check{\check{\mathbf{x}}}_{nn}) \tag{61}$$

on the solutions of the modified Cauchy difference problem

$$\tilde{\tilde{\mathbf{x}}}_{ni+1} = \check{\check{\mathbf{x}}}_{ni} + \mathbf{f}(\check{\check{\mathbf{x}}}_{ni}, \langle \mathbf{u}_{ni} \rangle, t_{ni})\tau_{ni}, \qquad i = 0, 1, \cdots, n-1, \tag{62}$$

$$\check{\check{\mathbf{x}}}_{ni+1} = \langle \tilde{\tilde{\mathbf{x}}}_{ni+1} \rangle, \qquad \check{\check{\mathbf{x}}}_{n0} = \langle \mathbf{x}_0 \rangle \tag{63}$$

corresponding to all possible admissible controls

$$\mathbf{u}_n \in B_{K_1} \quad (B_{K_2}).$$

Using the difference form of Gronwall's lemma and conditions (55) and (60), we find that when $\tau_n = O(n^{-1})$, $n \to +\infty$, we have

$$|\check{\mathbf{x}}_{ni+1} - \mathbf{x}_{ni+1}|, \qquad |\check{\check{\mathbf{x}}}_{ni+1} - \mathbf{x}_{ni+1}| \leqq A_4^{-1} o_n(1)[e^{A_4(t_{ni+1} - t_0)} - 1], \tag{64}$$

$$\lim_{n \to +\infty} \check{J}_n^* = J^*, \tag{65}$$

where $\mathbf{x}_n$ is the solution of problem (1′), (2′), $\check{\mathbf{x}}_n$ is the solution of problem (58), (59), $\check{\check{\mathbf{x}}}_n$ is the solution of problem (62), (63), while

$$\check{J}_n^* = \inf_{\mathbf{u}_n \in B_{K_1} (B_{K_2})} J_n(\check{\check{\mathbf{x}}}_n, \langle \mathbf{u}_n \rangle) = \inf_{\mathbf{u}_n \in \Sigma_{un} \cap B_{K_1} (\Sigma_{un} \cap B_{K_2})} J_n(\check{\mathbf{x}}_n, \check{\mathbf{u}}_n). \tag{66}$$

If

$$\check{J}_n^* = J_n(\check{\mathbf{x}}, \check{\mathbf{u}}^*), \qquad \text{where} \quad \check{\mathbf{u}}_n^* \in \Sigma_{un} \cap B_{K_1} \quad (\Sigma_{un} \cap B_{K_2}), \tag{67}$$

then, as above, it turns out that $\{\mathbf{u}_n^*\}$ is the minimizing sequence for the functional $J(\mathbf{x}(\mathbf{u}), \mathbf{u})$.

REFERENCES

[1] C. Catathéodory, *Vorlesungen über reele Funktionen*, 2nd ed., Teubner, Leipzig, 1927.

[2] E. S. Levitin and B. T. Polyak, *Methods for minimization in the presence of constraints*, U.S.S.R. Comput. Math. and Math. Phys., 6 (1966), pp. 787–823 (of Russian original).

[3] A. N. Tikhonov, *Methods for the regularization of optimal control problems*, Soviet Math. Dokl., 6 (1965), pp. 761–763.

[4] ———, *The stability of algorithms for the solution of degenerate systems of linear algebraic equations*, U.S.S.R. Comput. Math. and Math. Phys., 5 (1965), pp. 181–188.

[5] ———, *Non-well-posed problems of optimal planning*, Ibid., 6 (1966), pp. 114–127.

[6] ———, *The stability of the problem of optimizing functionals*, Ibid., 6 (1966), pp. 631–634 (of Russian original).

[7] E. S. Levitin and B. T. Polyak, *The convergence of minimizing sequences in problems of conditional extremum*, Soviet Math. Dokl., 7 (1966), pp. 764–767.

[8] B. M. Budak and A. D. Gorbunov, *On multipoint difference methods of solving the Cauchy problem for the equation $y' = f(x, y)$*, Vestnik Moskov. Univ. Ser. I Mat. Mekh., 1961, no. 4, pp. 10–19.

[9] V. A. Morozov, *Methods for solving unstable problems*, Lecture notes (mimeograph), Moscow State University, 1967.

SIAM J. CONTROL
Vol. 7, No. 1, February 1969

DISCRETE APPROXIMATIONS TO CONTINUOUS OPTIMAL CONTROL PROBLEMS*

JANE CULLUM†

Abstract. It is demonstrated that if P is a continuous optimal control problem whose system of differential equations is linear in the control and the state variables, and whose control and state variable constraint sets are convex, a direct method of determining an optimal solution of P exists. It is demonstrated that such a "continuous" problem can be replaced by a sequence offinite-dimensional, "discrete" optimization problems in which the control and state variable constraints are treated directly. The approximation obtained relates the respective optimal solutions.

1. Introduction. An optimal control problem P is *continuous* if it is representable by a system of ordinary differential equations, an integral cost, and control and state variable constraints. The standard method of solution [1], [2], [10], [11] consists of three steps. First, P is replaced by a sequence $\{P_k\}_1^\infty$ of continuous optimal control problems partially or completely unconstrained in the state. These problems are obtained from P by introducing penalty functions [17] corresponding to the intermediate and/or the terminal state constraints of P. Second, heuristic reasoning is used to determine an integer K such that optimal solutions of P_K are "close" to optimal solutions of P. Third, the two-point boundary value problem resulting from the application of the maximum principle to P_K is solved. (The control constraints are eliminated during the application of the maximum principle.) A theoretical justification of this procedure for certain continuous problems is given by Russell [17]. The consequences of the application of this procedure to more general continuous optimal control problems are discussed in [20].

The object of this paper is to demonstrate theoretically that if the differential system of P is linear in the control and the state, P can be solved directly. It is not necessary to introduce penalty functions or the maximum principle. The control and state variable constraints in P can be treated directly; and the determination of the solution of a two-point boundary value problem that involves discontinuous functions and an unstable system of equations can be eliminated. In particular, it is demonstrated that such a continuous problem can be "approximated" by a sequence $\{P_m\}_1^\infty$ of finite-dimensional, discrete, optimization problems. The approximation obtained relates respective optimal solutions. The discrete optimization problems can be solved by mathematical programming. The numerical solution of the discrete approximations is considered in [19].

* Received by the editors July 5, 1967, and in final revised form November 1, 1968.

† IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598. This work was jointly sponsored by the Air Force Office of Scientific Research under Contract AF 49(638-1474) and by IBM.

The idea of replacing an infinite-dimensional optimization problem by a sequence of finite-dimensional optimization problems is very old [3]. Approximations of the type discussed in this paper have been successfully implemented by Rosen [14], [15]. However, literature relating solutions of the associated finite-dimensional problems to solutions of the original problem is limited. Generally, it is merely assumed that a desirable relationship exists.

Before proceeding, it is necessary to state precisely the meaning of the phrase "the sequence of problems $\{P_m\}_1^\infty$ is an approximation to the problem P." Consider the following properties:

(a) For large m, trajectories admissible for P_m exist.

(b) As $m \to \infty$, the optimal cost of P_m converges to the optimal cost of P.

(c) Any sequence of trajectories $\{x_m\}_1^\infty$, with x_m admissible for a corresponding $P_{k(m)}$, $k(m) \to \infty$ as $m \to \infty$, and with costs convergent to the optimal cost of P, "converges" to an optimal trajectory of P.

DEFINITION 1.1. If a sequence of problems $\{P_m\}_1^\infty$ satisfies (a), (b) and (c) for some problem P, then this sequence is said to be an *approximation to P*.

Two types of approximations are considered. The first is obtained by discretizing the solution formula for the system of differential equations associated with P. The second is obtained by directly replacing the differential equations by difference equations. In both cases, the cost is replaced by a simple summation.

It is proved that if the original system of linear differential equations is completely controllable, and the cost function is convex, then the family of finite-dimensional optimization problems generated by the introduction of either type of discretization contains a sequence that approximates P.

2. Problem statement and notation. Throughout this paper, P will denote a continuous optimal control problem of the following type.

System equations.

$$\dot{\hat{x}} = \hat{f}(\hat{x}, u, t) \quad \text{a.e.} \quad \text{on} \quad [0, t_F]; \tag{2.1}$$

System constraints:

$$\begin{aligned} &u(t) \in U(t) \subset E^r \quad \text{a.e.} \quad \text{on} \quad [0, t_F], \\ &\hat{x}(t) \in A(t) \subseteq E^n, \quad t \in [0, t_F], \\ &t \in [0, t_F] \subset [0, T] = \bar{I}_0, \\ &\hat{x}(0) \in G_0 \subseteq E^n, \\ &\hat{x}(t_F) \in G_1 \subseteq E^n; \end{aligned} \tag{2.2}$$

Objective: Minimize

$$C(u) = \int_0^{t_F} f^0(\hat{x}, u, t)\, dt. \tag{2.3}$$

The control and the state variable constraint sets may vary with time, and the time duration is not specified other than that it must be less than T. The corresponding problem obtained by augmenting the differential system will also be denoted by P, i.e., $\dot{x} = f(\hat{x}, u, t)$, where $f = (f^0, \hat{f})$. In most of the paper the unaugmented problem will be under consideration, but the hats will be dropped to simplify the notation. It is always assumed that the problem P being considered has an optimal solution.

The following notation is used throughout the paper. E^m denotes m-dimensional Euclidean space. I_0 denotes an open interval in E^1 and $\bar{I}_0$ its closure. For any set $B \subseteq E^m$ and $\delta > 0$, $U(B, \delta)$ denotes the set $\{x | d(x, B) \leqq \delta\}$, where $d(x, B)$ denotes the Euclidean distance of x to B. Int B denotes the interior of B. A and U denote the mappings $t \to A(t)$ and $t \to U(t)$ for $t \in [0, T]$. (A, γ) and (U, γ) denote the mappings $t \to U(A(t), \gamma)$ and $t \to U(U(t), \gamma)$ for $t \in [0, T]$. The set $\{y | y = f(\hat{x}, u, t), u \in U(t)\}$ is denoted by $f(\hat{x}, U(t), t)$. $|\cdot|$, $\|\cdot\|_{L_2}$ and $\|\cdot\|_C$ denote, respectively, the Euclidean norm, the Lebesgue norm and the norm of uniform convergence on the appropriate spaces [6]. Finally, convergence in the strong and weak L_2-topologies is denoted by $\xrightarrow{s}$ and $\xrightarrow{w}$, respectively.

3. Approximations.

DEFINITION 3.1. For $m = 1, 2, \cdots$ let $\bar{t}_m$ be a $(k(m) + 1)$-dimensional vector with $\bar{t}_m = (0, t_m^1, \cdots, t_m^{k(m)})$, and $t_m^{j-1} \leqq t_m^j$, $j = 1, \cdots, k(m)$. For each m, let $\bar{x}_m$ and $\bar{u}_m$, respectively, be single-valued mappings of the sequences $\{0, t_m^1, \cdots, t_m^{k(m)}\}$ and $\{0, t_m^1, \cdots, t_m^{k(m)-1}\}$ into E^n and E^r. Let $\bar{I}_m = [0, t_m^{k(m)}]$. For each m, define x_m to be the piecewise linear extension of $\bar{x}_m$ to $\bar{I}_m$ and u_m to be the piecewise constant extension of $\bar{u}_m$ (taken to be right-continuous) to $\bar{I}_m$.

The sequence of triplets $(\bar{x}_m, \bar{u}_m, \bar{t}_m)$, $m = 1, 2, \cdots$, is an approximation of type (a) ((b) or (c)) to a pair of functions (x_0, u_0) defined on an interval $\bar{I}_0 = [0, T]$ if as $m \to \infty$, $I_m \to I_0$ and $x_m(t) \to x_0(t)$ for almost every $t \in I_0$ (it is of type (a) and $u_m \xrightarrow{w} u_0$ or it is of type (a) and $u_m \xrightarrow{s} u_0$).

Example 3.1. Before proceeding, consider the following example [9]. The discrete problems generated by discretizing, as indicated earlier, do not yield the desired approximations. Define P as follows.

Equations:

$$\dot{x} = \begin{pmatrix} \sin 2\pi u \\ \cos 2\pi u \\ -1 \end{pmatrix}, \qquad x \in E^3, \quad u \in E^1; \tag{3.1}$$

Constraints:

$$|u| \leqq 1, \quad x(0) = (0, 0, 1), \quad x(t_f) = (0, 0, 0); \tag{3.2}$$

Objective: Minimize the final time t_f.

Define P_m, $m = 1, 2, \cdots$, as follows.

Equations:

$$x_m^{k+1} - x_m^k = \begin{pmatrix} (\sin 2\pi u_m^k)\dfrac{1}{m} \\ (\cos 2\pi u_m^k)\dfrac{1}{m} \\ -\dfrac{1}{m} \end{pmatrix}, \qquad k = 0, 1, \cdots, \quad m = 1, 2, \cdots; \tag{3.3}$$

Constraints:

$$\begin{gathered} |u_m^k| \leqq 1, \qquad k = 0, 1, \cdots, \quad m = 1, 2, \cdots, \\ x_m^0 = (0, 0, 1), \qquad x_m^{k_F^m} = (0, 0, 0); \end{gathered} \tag{3.4}$$

Objective: For each m, minimize the final time k_F^m/m.

Clearly, the optimal cost for each m is 1. Set $\bar{t}_m = (0, 1/m, \cdots, 1)$. Define $\bar{u}_m$ to be the natural mapping of $\{0, 1/m, \cdots, (m-1)/m\}$ onto $\{u_m^0, \cdots, u_m^{m-1}\}$, where

$$u_m^k = (-1)^k \tfrac{1}{4}, \qquad k = 0, 1, \cdots, m-1, \quad m = 2, 4, \cdots.$$

Then the corresponding trajectory $\bar{x}_m$ of P_m is the natural mapping of $\{0, 1/m, \cdots, 1\}$ onto $\{x_m^0, \cdots, x_m^m\}$, where

$$x_m^k = \begin{cases} (0, 0, 1 - k/m) & \text{if } k \text{ is even,} \\ (1/m, 0, 1 - k/m) & \text{if } k \text{ is odd.} \end{cases}$$

It is clear that each $(\bar{x}_m, \bar{u}_m)$ is an optimal pair for P_m, $m = 2, 4, \cdots$, and the corresponding sequence $\{(\bar{x}_m, \bar{u}_m, \bar{t}_m)\}_1^\infty$ is an approximation of type (a) to the function $x_0(t) = (0, 0, 1 - t)$ on $[0, 1]$. But, x_0 is not even admissible for P. Observe that the sequence $\{P_{2m}\}_1^\infty$ satisfies (a) and (b) in Definition 1.1.

Remarks. Intuitively, one would expect a discretization procedure to be permissible if the original problem is well-posed. (Recall that a problem is well-posed if small changes in the data yield problems that have optimal solutions that are "close" to optimal solutions of the original problem.) Essentially, it is demonstrated that any problem that was proved in [4] to be well-posed can be approximated by finite-dimensional optimization problems obtained by discretizing. The results and proofs parallel the results and proofs in [4].

Theorem 3.1 states that if a problem P is "convex," any sequence of pairs, composed of trajectories and controls admissible for discrete problems obtained by directly replacing the differential equations of P by difference equations, contains limit points; and any limit point is a pair admissible for P. In § 4, for linear problems it is proved that there exist subsequences of the discrete problems for which admissible pairs exist and for which the corresponding sequence of optimal costs converges to the optimal cost of P. These results obviously imply that approximations in the sense of Definition 1.1, of the type desired, exist. Recall that it is always assumed that P has an optimal solution.

In Theorem 3.1, P is a fixed-time, continuous optimal control problem defined by (2.1)–(2.3). The sequence $\{P_m\}_1^\infty$ is defined as follows:

Let γ_m and σ_m, $m = 1, 2, \cdots$, be sequences of real numbers such that $\gamma_m \downarrow 0$ and $\sigma_m \downarrow 0$. Define $A_m(t) = U(A(t), \gamma_m)$ and $U_m(t) = U(U(t), \sigma_m)$. Then P_m, $m = 1, 2, \cdots$, denotes the following problem.

Equations:

$$x_m^{k+1} - x_m^k = f(x_m^k, u_m^k, t_m^k)\frac{T}{m}, \qquad k = 0, 1, \cdots, m; \tag{3.5}$$

Constraints:

$$\begin{aligned} &\hat{x}_m^0 \in G_0, \qquad \hat{x}_m^m \in G_1, \\ &\hat{x}_m^k \in A_m^k \equiv A_m(t_m^k), \\ &u_m^k \in U_m^k \equiv U_m(t_m^k), \\ &t_m^k = \frac{kT}{m}; \end{aligned} \tag{3.6}$$

Objective: Minimize the first component of x_m^m.

A trajectory of P_m and a control that generates this trajectory are denoted by the pair $(\bar{x}_m, \bar{u}_m)$, where $\bar{x}_m$ $(\bar{u}_m)$ denotes the mapping of the sequence $\{0, t_m^1, \cdots, T\}$ $(\{0, t_m^1, \cdots, t_m^{m-1}\})$ onto the sequence $\{x_m^0, \cdots, x_m^m\}$ $(\{u_m^0, \cdots, u_m^{m-1}\})$.

Allowances for relaxations in the control and state space constraints are included in (3.6) because such enlargements may be necessary to insure reachability.

Observe that the hypotheses in Theorem 3.1 coincide with the hypotheses of the existence theorem given in [16]; in fact, the proof of Theorem 3.1 parallels the proof in [16], and therefore is not given.

THEOREM 3.1. *Let P satisfy the following hypotheses:*

(a) *Given any compact sets $U \subset E^r$ and $T \subset E^1$, there exists a function g bounded on bounded sets and $O(s)$ as $s \to \infty$ such that for all $(\hat{x}, u, t) \in (E^n \times U \times T), |\hat{f}(\hat{x}, u, t)| \leqq g(|\hat{x}|)$.*

(b) *$f = (f^0, \hat{f})$ is continuous.*

(c) *For each t, the sets $A(t)$ and $U(t)$ are compact, and the set $A(t)$ is convex. The mappings U and A are continuous in the respective Hausdorff metric topologies.*

(d) *G_0 and G_1 are compact.*

(e) *For each $(\hat{x}, t)$, the set $f(\hat{x}, U(t), t)$ is convex. ($f(\hat{x}, u, t)$ is linear in u and for each t, $U(t)$ is convex.)*

Then any sequence $\{(\bar{x}_m, \bar{u}_m)\}_1^\infty$ of pairs admissible for the corresponding problems $\{P_m\}_1^\infty$ contains a subsequence that is an approximation of type (a) (*type* (b)) *to a pair (x_0, u_0) admissible for P. If the cost of $(\bar{x}_m, \bar{u}_m)$ converges to the optimal cost of P as $m \to \infty$, then (x_0, u_0) is an optimal solution of P; moreover, if P has a unique optimal solution, then the original sequence is itself an approximation of the type indicated.*

Remarks. Several comments that apply to the entire paper should be made. First, the difference equation approximation chosen in Theorem 3.1 and used throughout the paper is the simple explicit Cauchy–Euler formula [8]. Obviously, more sophisticated schemes with better truncation errors and stability properties could be used. The results obtained will be valid for most of these schemes. A Cauchy–Euler scheme was chosen to simplify the notation and hence to simplify the proofs.

Second, again to simplify notation, the sets $A(t)$ were each assumed to be compact. It is clear, since the function f is required to satisfy growth and continuity conditions, that "compact" can be replaced by "closed" and the arguments restricted to an appropriate compact sphere. Third, the subintervals used in the discretizing are chosen to be of length T/m. It is clear that many other choices are equally suitable.

Fourth, since Warga [18] has proved that there is a one-to-one correspondence between free-time and fixed-time problems, and since this correspondence preserves properties (a) and (e) in Theorem 3.1, no loss of generality results in considering only fixed-time problems.

Fifth, since the time is fixed, and the length of the subintervals is T/m, the triplet notation $(\bar{x}_m, \bar{u}_m, \bar{t}_m)$ has been shortened to $(\bar{x}_m, \bar{u}_m)$.

The two lemmas needed to extend the proof in [16] to a proof of Theorem 3.1 are presented. The first lemma demonstrates that the points reachable by P and P_m, $m = 1, 2, \cdots$, in time T are uniformly bounded. J_m^k denotes either the open, closed, right-open and left-closed, or vice versa, interval with endpoints $(k - 1)T/m$ and kT/m. If the type of interval being considered is significant, the type will be stated explicitly.

LEMMA 3.1. *Let S_m denote the set of points reachable by P_m; then*

$$S_m = \left\{x \,\middle|\, x = \sum_{j=0}^{m-1} f(x^j, u^j, t^j)\frac{T}{m} + x^0, x^0 \in G_0, u^j \in U_m^j, x^j \in A_m^j\right\} . \tag{3.7}$$

If G_0 is bounded and if there exists a function g bounded on bounded sets and $O(s)$ as $s \to \infty$ such that

$$|f(\hat{x}, u, t)| \leqq g(|\hat{x}|) \quad \textit{for} \quad (\hat{x}, u, t) \in \left\{\bigcup_{m=1}^{\infty} A_m(t) \times \bigcup_{m=1}^{\infty} U_m(t) \times [0, T]\right\},$$

then the set $\bigcup_{m=1}^{\infty} S_m$ is bounded.

Proof. The argument is analogous to the argument in the continuous case and is given in [5].

LEMMA 3.2. *Let the sets $U_m(t)$ and $A_m(t)$, $t \in \bar{I}_0$, $m = 1, 2, \cdots$, be convex and satisfy the hypotheses in Theorem* 3.1. *Let there exist measurable functions $\{u_m\}_1^\infty$ and continuous functions $\{x_m\}_1^\infty$ such that for $t \in J_m^j$, $u_m(t) \in U_m(t_m^{j-1})$ and $x_m(t)$* $\in$ (convex hull $\{A_m(t_m^{j-1}) \cup A_m(t_m^j)\}$), *$j = 1, \cdots, m$ and $m = 1, 2, \cdots$. If $u_m \xrightarrow{w} u_0$ $(x_m \xrightarrow{s} x_0)$ on $\bar{I}_0$, then* (a) *$u_0(t) \in U(t)$* (b) *$(x_0(t) \in A(t))$ a.e. on $\bar{I}_0$.*

Proof.

(a) Given $\gamma > 0$ there exists $\sigma > 0$ such that $U(t') \subseteq U(U(t), \gamma)$ whenever $|t - t'| < \sigma$ and $t, t' \in \bar{I}_0$. There exists M such that $\sigma_m < \gamma$ and $1/m < \sigma$ for all

$m > M$. Hence for $m > M, j = 1, \cdots, m$, and almost every $t \in J_m^{j+1}$,

$$u_m(t) \in U_m(t_m^j) \subseteq U(U(t), 2\gamma).$$

But $S_\gamma = \{u | u$ measurable, $u(t) \in U(U(t), \gamma)$ a.e. on $\bar{I}_0\}$ is weakly closed. Hence, $u_0 \in S_\gamma$ for all $\gamma > 0$. But $\bigcap_{\gamma > 0} S_\gamma = S$.

(b) It is clear that (b) follows from (a).

4. Convergence of costs. The discussion of the convergence of the costs is restricted to problems with unaugmented systems that are linear in the state and control variables.

4.1. Time-optimal control problems. First, consider a linear, time-optimal control problem P. Since the proof of the convergence of the costs uses families of solutions of the differential equations, it is not advantageous to convert P into a nonlinear fixed-time problem. In this section the unaugmented formulation is always used (the hats are omitted).

P is defined by (2.1)–(2.3) with

$$\begin{aligned} f(x, u, t) &= C(t)x + D(t)u, \\ G_0 &= \{x_0\}, \qquad G_1 = \{x_F\}, \\ f_0(x, u, t) &= 1. \end{aligned} \tag{4.1}$$

It is assumed that (if for each $t \in [0, T]$, $U(t)$ and $A(t)$ are convex, compact sets; (ii) the mappings U and A are continuous in the respective Hausdorff metric topologies; (iii) the matrix functions C and D belong to $L_\infty(\bar{I}_0)$.

Discrete problems generated by discretizing the solution formula are considered first. Let $\{\gamma_m\}_1^\infty$ be a sequence of positive real numbers monotonically decreasing to zero. For $m = 1, 2, \cdots$ define $A_m(t) = U(A(t), \gamma_m)$ and $U_m(t) = U(t)$. (An allowance for relaxation of the state space constraint sets is included to insure reachability.)

Let X denote the matrix solution of $\dot{X} = C(t)X$ a.e. on $[0, T]$ with $X(0) = I$. Define $\tilde{P}_m$, $m = 1, 2, \cdots$, as follows.

Equations:

$$x_m^{k+1} - x_m^k = E_m^k x_m^k + F_m^k u_m^k, \qquad k = 0, 1, \cdots; \tag{4.2}$$

Constraints:

$$x_m^0 = x_0, \quad x_m^{k_F} = x_F, \quad u_m^k \in U_m^k, \quad x_m^k \in A_m^k, \qquad k = 0, 1, \cdots; \tag{4.3}$$

Objective: Minimize the final time $t_m^{k_F}$.

In (4.2) and (4.3),

$$\begin{aligned} E_m^k &= (X^{k+1}(X^k)^{-1} - I), \qquad F_m^k = X^{k+1} \int_{J_m^{k+1}} X^{-1} D, \\ U_m^k &= U(t_m^k), \qquad A_m^k = U(A(t_m^k), \gamma_m), \\ J_m^k &= [t_m^{k-1}, t_m^k], \qquad X^k = X(t_m^k), \qquad t_m^k = \frac{kT}{m}. \end{aligned} \tag{4.4}$$

Remarks. First, it is clear that (in the sense that a maximum amount of problem structure has been preserved) the sequence $\{\tilde{P}_m\}_1^\infty$ represents an optimal discretization of the problem P. However, such a discretization requires complete knowledge of the fundamental solution of the differential system associated with P, knowledge which is not readily available if the system is nonautonomous. Observe that if the sets $U(t)$ do not depend on time, then $\{\tilde{P}_m\}_1^\infty$ is essentially the restriction of P to controls that are piecewise constant on successive subintervals of equal length T/m. That is, if a control is admissible for $\tilde{P}_m$, then the piecewise constant extension of it is admissible for P, the trajectory $\mathbf{x}_m$ of P generated by this extension has the property that $\mathbf{x}_m(t_m^k) = x_m^k$, $k = 0, 1, \cdots$, and, in fact, the cost for this control is the same for both problems. Finally, observe that Theorem 3.1 is not directly applicable to the $\{\tilde{P}_m\}_1^\infty$. It is applicable directly to the equivalent fixed-time problem and its "direct" discretizations. However, the proof of the extension is not difficult (see [5]).

THEOREM 4.1. *The conclusions of Theorem* 3.1 *are valid for P and the sequence* $\{\tilde{P}_m\}_1^\infty$.

Theorem 4.2 demonstrates the "continuity" of the optimal cost under discrete perturbations. It demonstrates the existence of a sequence of finite-dimensional optimization problems, each of which has admissible pairs and for which the corresponding sequence of optimal costs converges to the optimal cost of P as $m \to \infty$. Recall that it is always assumed that an optimal solution of P exists.

THEOREM 4.2. *Let P satisfy* (2.1)–(2.3) *and* (4.1) *and let $\tilde{P}_m((A, \gamma_m), U)$, $m = 1, 2, \cdots$, denote the discrete optimal control problems defined in* (4.2), (4.3) *and* (4.4). *Let t^* denote the minimum time to reach x_F. If*

(a) *the equations* (4.1) *are completely controllable* [9],

(b) *there exist δ and $\varepsilon > 0$ such that on $[t^*, t^* + \varepsilon]$ there exists an admissible control $\tilde{u}$ with*

$$\inf_{[t^*, t^* + \varepsilon]} d(\tilde{u}(t), \partial U(t)) \geqq \delta > 0 \tag{4.5}$$

and

$$A(t)x_F + B(t)\tilde{u}(t) = 0 \quad a.e., \tag{4.6}$$

then there exists a sequence $l(m) \to \infty$ as $m \to \infty$ such that the optimal cost of $\tilde{P}_{l(m)}((A, \gamma_m), U)$ converges to the optimal cost of P as $m \to \infty$.

Remark. The fundamental ideas in the proof have been used by Neustadt [13] to derive this result for $U(t) \equiv (|u| \leqq 1)$, $A = E^n$ and $x_F = 0$. Observe that if x_F is an equilibrium point of the system $\dot{x} = A(t)x$ and $0 \in \operatorname{Int} U(t)$ for all $t \in T$, as in [13], (4.6) is satisfied. Statement (4.6) guarantees that if x_F is reachable in time t^*, then it is reachable in time $t^* + \varepsilon$ for all small ε. Inequality (4.5) and the controllability are used to prove that x_F is an interior point of the set of attainability in time $t^* + \varepsilon$ for small ε. The fact that x_F is such an interior point permits the desired discretization. Example 4.1 shows that if (4.6) is not satisfied, then the required membership may not occur.

Example 4.1. Let P be defined as follows:

$$\dot{x}_1 = x_2, \quad |u| \leqq 1, \quad |x_1| \leqq 10, \quad |x_2| \leqq 10;$$
$$\dot{x}_2 = u, \quad x_0 = (2, 2) \quad x_F = (1, \sqrt{6}) \quad t_0 = 0;$$

Minimize the final time t_F.

Clearly, x_F is reachable from x_0 since both points are on the curve $y_1 = -\frac{1}{2}(y_2)^2 + 4$ generated by $u \equiv -1$. Hence, an optimal time t^* and pair (x^*, u^*) exist [19]. Suppose there existed an admissible pair $(\tilde{x}, \tilde{u})$ mapping x_0 onto x_F in time $t^* + \varepsilon$. Clearly, $\tilde{x}_i(t^*) > x_i^*(t^*)$, $i = 1, 2$, and $\tilde{x}(t)$ cannot decrease to x_F until $\tilde{x}_2(t)$ decreases to zero. Hence, $\varepsilon \geqq 2$. Hence, x_F is not reachable in time $t^* + \varepsilon$ for any ε between 0 and 1.

Lemma 4.1 demonstrates the plausibility of replacing the family of functions satisfying the control constraints by a subfamily of step functions of a particular type. In Lemma 4.1, for $k = 1, \cdots, m - 1$ the interval J_m^k is right-open and left-closed. J_m^m is closed. Recall that if $U(t) \equiv U$, the discretization of the solution formula corresponds to restricting the admissible controls to a certain family of step functions.

LEMMA 4.1. *Let* $\bar{I}_0 = [0, 1]$. *Let U be a mapping from* $[0, 1]$ *into the compact and convex subsets of E^r that is continuous in the Hausdorff metric topology. For any function $u_0 \in L_2(\bar{I}_0)$ with $u_0(t) \in U(t)$, $t \in \bar{I}_0$, there exists a sequence of right-continuous step functions $\{u_m\}_1^\infty$ such that* (i) *u_m is constant on J_m^k, $k = 1, \cdots, m$,* (ii) $u_m(t_m^k) \in U(t_m^k)$, $k = 0, \cdots, m - 1$, *and* (iii) $u_m \xrightarrow{s} u_0$.

Proof. Since U is continuous on $[0, 1]$, for any $\varepsilon > 0$ there exists $\gamma(\varepsilon) > 0$ such that $|t - t'| < \gamma(\varepsilon)$ implies that $U(t) \subseteq U(U(t'), \varepsilon)$. Furthermore, the sets $U(t), t \in [0, 1]$, are uniformly bounded by some number K. First, prove that u_0 is approximable by a continuous function that approximately satisfies the control constraints. By Luzin's theorem, given $\varepsilon > 0$ [12, p. 106] there exists a closed set $T(\varepsilon) \subseteq [0, 1]$ such that u_0 restricted to $T(\varepsilon)$, denoted by $u_0|T(\varepsilon)$, is continuous and $\mu(T'(\varepsilon)) < \min(\gamma(\varepsilon), \varepsilon/4K^2)$, where $T'(\varepsilon) = [0, 1] - T(\varepsilon)$, and μ denotes Lebesgue measure. But $T'(\varepsilon)$ is open and hence is the countable union of pairwise disjoint intervals [12, p. 48]. Therefore, the function $u_0|T(\varepsilon)$ can be extended to a continuous function φ_ε on $[0, 1]$ by polygonal extension over each interval $(a_j, b_j), j = 1, 2, \cdots$, in $T'(\varepsilon)$. But, $\mu(a_j, b_j) < \gamma(\varepsilon)$; therefore, $\varphi_\varepsilon(t) \in U(U(t), \varepsilon)$ for all t in I_0. Clearly, $\|u_0 - \varphi_\varepsilon\|_{L_2} \leqq 4K^2\mu(T'(\varepsilon)) < \varepsilon$.

Next define $u_\varepsilon(t)$ to be the unique point [6] in $U(t)$ closest to $\varphi_\varepsilon(t)$. Then, $|u_\varepsilon(t) - \varphi_\varepsilon(t)| \leqq \varepsilon$ for all $t \in [0, 1]$.

By the continuity of φ_ε there exists an $m(\varepsilon)$ such that for $m \geqq m(\varepsilon)$ the oscillation of φ_ε on J_m^k, $k = 1, \cdots, m$, is less than ε. For $t \in J_{m(\varepsilon)}^k$, $k = 1, \cdots, m(\varepsilon)$, define $u_{m(\varepsilon)}(t) = u_\varepsilon(t_{m(\varepsilon)}^{k-1})$. Then $\|u_{m(\varepsilon)} - u_0\|_{L_2} < 3\varepsilon$. This completes the proof.

The following lemma was used by Neustadt [13] without proof. A proof is given because all of the ensuing arguments are based on this lemma.

LEMMA 4.2. *If* $\mathscr{S} = \{x | x = \sum_1^{n+1} a_i x_i, \sum_1^{n+1} a_i = 1, a_i \geqq 0\}$ *is an n-dimensional simplex and y_0 is an interior point of $\mathscr{S}$, then there exist circles $|x - x_i| < \sigma_i$ such that any set of vectors y_i, $1 \leqq i \leqq n + 1$, such that $|y_i - x_i| < \sigma_i$ generates an n-simplex that contains y_0.*

Proof. The simplex generated by the x_i is nondegenerate. Therefore, the system of $n + 1$ equations in $n + 1$ unknowns

$$\begin{aligned} &\sum_{i=1}^{n+1} \lambda_i x_i^j = 0, \qquad j = 1, \cdots, n, \\ &\sum_{i=1}^{n+1} \lambda_i = 0 \end{aligned} \tag{4.7}$$

has the unique solution $\lambda_i = 0$, $1 \leqq i \leqq n + 1$ [7, p. 14]. Therefore, the matrix

$$A = \begin{pmatrix} x_1^1 & \cdots & x_{n+1}^1 \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ x_1^n & & x_{n+1}^n \\ 1 & & 1 \end{pmatrix}$$

is invertible. Since y_0 is in the interior of the simplex, the α_i such that $y_0 = \sum_{i=1}^{n+1} \alpha_i x_i$ satisfy $0 < \alpha_i < 1$.

Let S denote the unit sphere in E^n. Consider the matrix equation

$$A(\varepsilon, \rho)\beta = \begin{pmatrix} x_1^1 + \varepsilon\rho_1^1 & \cdots & x_{n+1}^1 + \varepsilon\rho_{n+1}^1 \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ 1 & \cdots & 1 \end{pmatrix} \beta = \begin{pmatrix} y_0^1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{pmatrix} \tag{4.8}$$

for any set of vectors $\rho_i \in S$, $i = 1, \cdots, n + 1$. Since A^{-1} exists, there exists $\varepsilon_0 > 0$ such that $[A(\varepsilon, \rho)]^{-1}$ exists for all $\varepsilon < \varepsilon_0$ and all $\rho = (\rho_1, \rho_2, \cdots, \rho_{n+1}) \in S^{n+1} = (S \times S \times \cdots \times S)$. The corresponding solution of (4.8) is

$$\beta(\varepsilon, \rho) = [A(\varepsilon, \rho)]^{-1} \begin{pmatrix} y_0 \\ 1 \end{pmatrix}.$$

But, $0 < \beta^i(0, \rho) < 1$ for all i. Hence, by the continuity of $[A(\varepsilon, \rho)]^{-1}$ at $\varepsilon = 0$, which is uniform in ρ for $\rho \varepsilon S^{n+1}$, there exists $\bar{\varepsilon}_0 \leqq \varepsilon_0$ such that $0 \leqq \beta^i(\varepsilon, \rho) \leqq 1$ for all $\varepsilon < \bar{\varepsilon}_0$ and all $\rho \in S^n$, $1 \leqq i \leqq n + 1$. This completes the proof.

In the following discussion, $\Omega(\sigma, K, L)$ $(\tilde{\Omega}_m(\sigma, K, L))$ will denote the points in E^n reachable in time σ by trajectories and controls of P (P_m) with values at time t in $K(t)$ and $L(t)$, respectively (at time t_m^k in $K(t_m^k)$ and $L(t_m^k)$). Moreover, $\bar{x}_m$ will denote a trajectory of P_m, and $\bar{u}_m$ will denote a control that generates such a trajectory. x_m will denote the piecewise linear extension of $\bar{x}_m$ and u_m will denote the piecewise constant (right-continuous) extension of $\bar{u}_m$. $\mathbf{x}$ will denote the trajectory of the original problem P generated by a control u.

Proof of Theorem 4.2. By hypothesis, there exist positive real numbers δ and ε_0 and a measurable function $\tilde{u}$ satisfying (4.5) and (4.6). Let γ be given and let M

be a bound for the sets $U(t), t \in [0, T]$. Then there exists $\varepsilon(\gamma) < \varepsilon_0$ such that for $t \in [t^*, t^* + \varepsilon(\gamma)]$ and $|u| \leqq M + 1$,

$$\left| X(t) \int_{t^*}^{t} X^{-1} D u \right| \leqq \gamma. \tag{4.9}$$

It is proved that for $\varepsilon < \varepsilon(\gamma)$, there exists a sphere containing x_F such that every point in this sphere is reachable from x_0 in time $t^* + \varepsilon$ by a trajectory of P that at time t, for $t \in [0, t^* + \varepsilon]$, is in the set $U(A(t), \gamma)$.

Let $u(x_F)$ denote a control that generates x_F in time t^*. Clearly, $\{u| \, |u - \tilde{u}(t)| < \delta\} \subseteq U(t)$ for all $t \in [t^*, t^* + \varepsilon_0]$. Let S denote the unit sphere in E^n. From the controllability hypothesis [4], there exists a scalar a such that for every $\rho \in S$ there exists a measurable function u_ρ with $|u_\rho| \leqq \delta/2$ and

$$\int_{t^*}^{t^* + \varepsilon} X^{-1} D u_\rho = X^{-1}(t^* + \varepsilon) a \rho. \tag{4.10}$$

Using the fact that

$$x_F = X(t^* + \varepsilon) X^{-1}(t^*) [x_F + \int_{t^*}^{t^* + \varepsilon} X(t^*) X^{-1} D \tilde{u}], \tag{4.11}$$

one easily obtains that

$$x_F + a\rho = X(t^* + \varepsilon) \left[x_0 + \int_0^{t^* + \varepsilon} X^{-1} D u \right], \tag{4.12}$$

where

$$u = \begin{cases} u(x_F) & \text{on} \quad [0, t^*), \\ \tilde{u} + u_\rho & \text{on} \quad [t^*, t^* + \varepsilon]. \end{cases} \tag{4.13}$$

(Observe that a is a function of ε and δ.) Fix m; then there exists $\varepsilon = \varepsilon(\gamma_m/2) < \varepsilon_0$ such that $x_F \in \operatorname{Int} \Omega(t^* + \varepsilon, (A, \gamma_m/2), U)$. There exists a simplex in this set with vertices $Q_1, \cdots, Q_{n+1}$, containing x_F in its interior. By Lemma 4.2, there exist scalars $\eta_i > 0$, $i = 1, \cdots, n + 1$, such that if $|R_i - Q_i| < \eta_i$, $i = 1, \cdots, n + 1$, then x_F is in the convex hull of the R_i.

Each point Q_i is generated by a control function $u(i)$, admissible for P and defined on the interval $[0, t^* + \varepsilon]$. In $\tilde{P}_l$, $t_l^k = kT/l$. Let $t_l = R(l)T/l$ denote the largest multiple of T/l that is less than or equal to $t^* + \varepsilon$. By Lemma 4.1 (if $t^* + \varepsilon$ is not an integral multiple of T/l, then slight modifications in the argument are required), there exist sequences of step functions $u(i, l)$, $i = 1, \cdots, n + 1$, $l = 1, 2, \cdots$, such that for each fixed i and l, $u(i, l)$ is right-continuous, constant on the intervals J_l^k, $k = 1, \cdots, R(l) + 1$, and satisfies the control constraints in P at each time t_l^k, $k = 1, \cdots, R(l) + 1$, and for each i, the sequence of restrictions of the $u(i, l)$ to the interval $[0, t^* + \varepsilon]$ converges in the strong L_2-topology to $u(i)$.

It is clear that there exists an H such that

$$|X(T)|^2 \int_0^T |X^{-1} D|^2 \leqq H.$$

Furthermore, there exists $l(m)$ such that for $l > l(m)$ and $i = 1, \cdots, n + 1$,

$$|\mathbf{x}(i, t^* + \varepsilon) - \mathbf{x}(i, t_l)| < \tfrac{1}{2}\eta_i \tag{4.14}$$

and

$$H\|u(i, l) - u(i)\|_{L_2}^2 < \min(\tfrac{1}{2}\gamma_m, \tfrac{1}{2}\eta_i). \tag{4.15}$$

(Recall that $\mathbf{x}(i)$ and $\mathbf{x}(i, l)$ denote, respectively, the trajectories of P generated by $u(i)$ and $u(i, l)$.) Therefore, from (4.14) and (4.15), for $t \in [0, t^* + \varepsilon]$, $1 \leqq i \leqq n + 1$ and $l > l(m)$,

$$\begin{aligned} |\mathbf{x}(i, l, t) - \mathbf{x}(i, t)| &\leqq \tfrac{1}{2}\gamma_m, \\ |\mathbf{x}(i, l, t_l) - Q_i| &< \eta_i. \end{aligned} \tag{4.16}$$

Let $\bar{u}(i, l)$ $(\bar{u}(i))$ denote the natural projection of $u(i, l)$ $(u(i))$ onto the sequence $\{0, t_l^1, \cdots, t_l^{R(l)-1}\}$, and $\bar{x}(i, l)$ $(\bar{x}(i))$ denote the trajectory of P_l generated by $\bar{u}(i, l)$ $(\bar{u}(i))$ with the initial value x_0.

Let R_i^l, $i = 1, \cdots, n + 1$, denote the value of the trajectory $\bar{x}(i, l)$ at $R(l)T/l$. But, by construction, for $i = 1, \cdots, n + 1, \mathbf{x}(i, l, t_l) = R_i^l$. Therefore, for $l \geqq l(m)$, the points $\mathbf{x}(i, l, t_l)$, $1 \leqq i \leqq n + 1$, are in the set of attainability of $\tilde{P}_l((A, \gamma_m), U)$ in time t_l. But this set is convex; so from Lemma 4.2, x_F is also in this set. Therefore, the optimal cost of the problem $\tilde{P}_l((A, \gamma_m), U)$ is less than or equal to $t^* + 2\varepsilon$ for $l \geqq l(m)$.

Hence, there exists a sequence of integers $l(m) \to \infty$ as $m \to \infty$ such that the limit superior of the sequence of optimal costs of the corresponding problems $\tilde{P}_{l(m)}((A, \gamma_m), U)$ is less than or equal to t^*.

Observe that if $A = E^n$, $l(m) = m$.

Suppose there exists a sequence of integers $l(m) \to \infty$ as $m \to \infty$ such that the limit inferior of the sequence of optimal costs of the corresponding problems $\tilde{P}_{l(m)}((A, \gamma_m), U)$ is less than t^*. Then there exists a subsequence of triplets $(\bar{x}_{l(m)}, \bar{u}_{l(m)}, \bar{t}_{l(m)})$, admissible for $\tilde{P}_{l(m)}((A, \gamma_m), U)$, $m = 1, 2, \cdots$, such that the terminal times $t_{l(m)}^{k(l(m))} \to \bar{t} < t^*$. $\bar{x}_{l(m)}$ is a mapping of the sequence

$$\bar{t}_{l(m)} = \{0, t_{l(m)}^1, \cdots, t_{l(m)}^{k(l(m))}\}$$

onto a sequence $\{x^0, x_{l(m)}^1, \cdots, x_{l(m)}^{k(l(m))}\}$. (Without loss of generality the subsequence has been denoted by $l(m)$.) Recall that for each m, $x_{l(m)}$ and $u_{l(m)}$ denote, respectively, the piecewise linear and piecewise constant (right-continuous) extensions of $\bar{x}_{l(m)}$ and $\bar{u}_{l(m)}$ to the interval $[0, t_{l(m)}^{k(l(m))}]$. By construction, the trajectory of $P, \mathbf{x}_{l(m)}$, generated by $u_{l(m)}$, is such that for $k = 0, 1, \cdots, k(l(m))$, $\mathbf{x}_{l(m)}(t_{l(m)}^k) = x_{l(m)}^k$. Consider the restrictions of the functions $u_{l(m)}$ to the interval $[0, \bar{t}]$. (It may be necessary for some m to extend $u_{l(m)}$ to this interval.) The sequence of restrictions and extensions is sequentially compact in the weak L_2-topology. Therefore, there exists a $u_0 \in L_2$ and a subsequence (without loss of generality denoted by $l(m)$) such that $\{u_{l(m)}\}_1^\infty$ converges weakly to u_0. It is clear from the linearity that the corresponding sequence of trajectories of P, $\{\mathbf{x}_{l(m)}\}_1^\infty$, converges pointwise on $[0, \bar{t}]$ to the trajectory x_0 of P generated by u_0. By Lemma 3.2, x_0 and u_0 are admissible for P. Hence, x_F is reachable in time $\bar{t}$.

THEOREM 4.3. *Theorem* 4.2 *is valid with* $\tilde{P}_m$ *replaced by the direct discretization* P_m *defined in* (3.5) *and* (3.6), *if the following uniform condition is satisfied*:

Let $\bar{I}_0 = [0, T]$ *be given. Let* $\{u_m\}_1^\infty$ *be any sequence of functions piecewise constant on successive intervals of equal length* T/m *and satisfying* $u_m(t_m^k) \in U_m(t_m^k)$ *for* $k = 0, 1, \cdots, m - 1$. *Let* $\bar{u}_m$ *denote the natural projection of* u_m *onto the sequence* $\{0, t_m^1, \cdots, t_m^{m-1}\}$. *Then if* x_m *denotes the piecewise linear extension of the trajectory* $\bar{x}_m$ *of* P_m *generated by* $\bar{u}_m$, *and* $\mathbf{x}_m$ *denotes the trajectory of* P *generated by* u_m *and all the trajectories have the same initial value,* $\|x_m - \mathbf{x}_m\|_C \to 0$ *as* $m \to \infty$.

Proof of Theorem 4.3. The proof parallels the proof of Theorem 4.2. As in the proof of Theorem 4.2, given γ_m there exists an $\varepsilon < \varepsilon_0$ and an open sphere containing x_F such that every point in this sphere is reachable from x_0 in time $t^* + \varepsilon$ by a trajectory of P that at time t for $t \in [0, t^* + \varepsilon]$ is in the set $U(A(t), \gamma_m/4)$.

Hence, $x_F \in \operatorname{Int} \Omega(t^* + \varepsilon, (A, \gamma_m/4), U)$ and there exists a simplex in this set with vertices $Q_1, \cdots, Q_{n+1}$, containing x_F in its interior. By Lemma 4.2, there exist scalars $\eta_i > 0$, $i = 1, \cdots, n + 1$, such that $|R_i - Q_i| < \eta_i$, $1 \leqq i \leqq n + 1$, implies that x_F is in the convex hull of the R_i.

As in the proof of Theorem 4.2, there exists an $l(m)$ such that for $t \in [0, t^* + \varepsilon]$, $i = 1, \cdots, n + 1$ and $l > l(m)$,

$$\begin{aligned} |\mathbf{x}(i, l, t) - \mathbf{x}(i, t)| &< \tfrac{1}{4}\gamma_m, \\ |\mathbf{x}(i, l, t_l) - Q_i| &< \tfrac{1}{2}\eta_i, \end{aligned} \tag{4.17}$$

where t_l denotes the largest multiple of T/l smaller than or equal to $t^* + \varepsilon$, and $\mathbf{x}(i)$ and $\mathbf{x}(i, l)$, $i = 1, \cdots, n + 1$ and $l = 1, 2, \cdots$, denote the trajectories of P corresponding to the controls $u(i)$ that generate the vertices Q_i of the simplex and the controls $u(i, l)$ that approximate $u(i)$ that were obtained in Lemma 4.1.

By hypothesis, there exists an integer $l^*(m) \geqq l(m)$ such that for all $l > l^*(m)$, $i = 1, \cdots, n + 1$,

$$\|x(i, l) - \mathbf{x}(i, l)\|_C < \min(\tfrac{1}{4}\eta_i, \tfrac{1}{4}\gamma_m). \tag{4.18}$$

Hence, for $i = 1, \cdots, n + 1$ and $l > l^*(m)$, $|x(i, l, t_l) - Q_i| < \eta_i$. Therefore, by Lemma 4.2, $x_F \in \Omega_l(t_l, (A, \gamma_m), U)$ for $l \geqq l^*(m)$ and therefore, the optimal cost of $P_l((A, \gamma_m), U)$ is less than or equal to t_l which is less than or equal to $t^* + 2\varepsilon$.

Hence, there exists a sequence of integers $l(m) \to \infty$ as $m \to \infty$ such that the limit superior of the sequence of optimal costs of the corresponding problems $P_{l(m)}((A, \gamma_m), U)$ is less than or equal to t^*. (Observe that if $A = E^n$, $l(m) = m$.)

Conversely, suppose there exists a sequence of integers $l(m) \to \infty$ as $m \to \infty$ such that the limit inferior of the sequence of optimal costs of the corresponding problems $P_{l(m)}((A, \gamma_m), U)$ is less than t^*. Therefore, as in the proof of Theorem 4.2, there exists a sequence of triplets $(\bar{x}_{l(m)}, \bar{u}_{l(m)}, \bar{t}_{l(m)})$, $m = 1, 2, \cdots$, admissible for the corresponding problems $P_{l(m)}((A, \gamma_m), U)$ and with costs convergent to $\bar{t} < t^*$. As before, a control u_0, admissible for P, is obtained such that if $\mathbf{x}_0$ denotes the trajectory of P generated by u_0, then $\{\mathbf{x}_{l(m)}\}_1^\infty$ converges pointwise on the interval $[0, \bar{t}]$ to $\mathbf{x}_0$. But by hypothesis, $\|\mathbf{x}_{l(m)} - x_{l(m)}\|_C \to 0$. Hence $\{x_{l(m)}\}_1^\infty$ converges pointwise to $\mathbf{x}_0$ on $[0, \bar{t}]$, and by Lemma 3.2, $\mathbf{x}_0$ is admissible.

Remarks. The hypothesis of uniformity required in Theorem 4.3 is satisfied if P is autonomous. The proof is straightforward and given in [6].

Theorems 3.1 through 4.3 demonstrate that any "continuous" linear time-optimal control problem whose system of differential equations is completely controllable, whose terminal state trajectory value is an equilibrium point of the differential equations (in the sense described earlier) and that has an optimal solution, can be approximated, in the sense of Definition 1.1, by sequences of finite-dimensional optimization problems obtained by discretizing the solution formula of the system of differential equations or by replacing the differential equations directly by a system of difference equations.

4.2. Linear problems with convex costs. In this section linear fixed-time problems with weakly sequentially lower semicontinuous costs are considered. The results parallel the results in § 4.1.

In the following discussion $(C \times L_2)(\bar{I}_0)$ denotes the space of all pairs of functions (x, u) defined on $\bar{I}_0$ with x belonging to the n-fold product of the space of all real-valued continuous functions defined on I_0 with itself, and u belonging to the r-fold product of the space of all real-valued L_2-functions defined on $\bar{I}_0$ with itself. $(C \times \bar{U})(\bar{I}_0)$ denotes that subset of $(C \times L_2)(\bar{I}_0)$ with $u \in \bar{U}$. In the following discussion, $\bar{U}$ denotes the subset of L_2 consisting of those functions u with $u(t) \in U(t)$ for a.e. $t \in [0, T]$, and those step functions with range in $\bigcup_{t \in [0,T]} U(t)$.

Let $\tilde{P}_m(A_m, U_m)$, $m = 1, 2, \cdots$, denote the family of discrete optimal control problems defined in (4.2) and (4.3) with the cost function of time replaced by the simple summation

$$C_m(\bar{u}_m) = \sum_{k=0}^{m-1} f^0(x_m^k, u_m^k, t_m^k)T/m \tag{4.19}$$

and with $A_m = (A, \gamma_m)$ and $U_m = (U, \sigma_m)$ for some sequences $\gamma_m \downarrow 0$ and $\sigma_m \downarrow 0$ as $m \uparrow \infty$.

THEOREM 4.4. *Let P be a fixed-time continuous optimal control problem defined by* (2.1), (2.2) *and* (4.1) *with* $C(u) = \int_0^T f^0(x, u, t)\,dt$. *If the system of differential equations is completely controllable, and the cost is convex and continuous on* $(C \times \bar{U})(\bar{I}_0)$, *then there exist sequences* $l(m)$, $k(m) \to \infty$ *as* $m \to \infty$ *such that the optimal cost of the problem* $\tilde{P}_{k(m)}(A_m, U_{l(m)})$ *converges to the optimal cost of* P *as* $m \to \infty$.

Remarks. The hypotheses on the cost imply that the cost is weakly sequentially lower semicontinuous on the subset $(C \times \bar{U})(\bar{I}_0)$ (see [4]). Simple examples of admissible costs are the expenditures of fuel and energy.

Let $S(\beta, K, L)$ $(\tilde{S}_m(\beta, K, L))$ denote the set of points reachable in time T by trajectories and controls of P $(\tilde{P}_m)$ with values at time $t \in [0, T]$ in $K(t)$ and $L(t)$, respectively (at time t_m^k in $K(t_m^k)$ and $L(t_m^k)$), and costs less than or equal to β.

The notation $\bar{x}_m, \bar{u}_m, \mathbf{x}_m, x_m, u_m$ has the same meaning as in § 4.1.

Proof of Theorem 4.4. It is clear that for a given time interval $[0, T]$ and an initial point x^0, trajectories of $\dot{x} = C(t)x + D(t)u$ are Lipschitz continuous on any set of uniformly bounded measurable functions. Therefore, there exists a K such that for any two functions u_1 and u_2 in $\bar{U}$ and the corresponding trajectories $\mathbf{x}_1$ and $\mathbf{x}_2$ of P that they generate, $\|\mathbf{x}_1 - \mathbf{x}_2\|_C \leqq K\|u_1 - u_2\|_{L_2}$. Therefore, given $\varepsilon > 0$ there exists $\delta(\varepsilon) > 0$ such that $\|u_1 - u_2\|_{L_2} < \delta(\varepsilon)$ with u_1 and u_2 in $\bar{U}$ implies that

$$|C(u_1) - C(u_2)| < \varepsilon. \tag{4.20}$$

Let β^* denote the optimal cost of P and (x_0, u_0) denote an optimal pair. Since the cost is convex, and each of the sets $A(t)$ and $U(t)$ is convex, for each γ, β, σ the set $S(\beta, (A, \gamma), (U, \sigma))$ is convex.

The proof is very similar to the proof of Theorem 4.2. Fix m; then there exists $\delta_m > 0$ such that for every $t \in [0, T]$ and $\delta < \delta_m$,

$$|X(t) \int_0^t X^{-1}D\delta| < \tfrac{1}{2}\gamma_m. \tag{4.21}$$

By the controllability hypothesis, for each δ, there exists an $a(\delta) > 0$ such that for every ρ in the unit sphere S of E^n, there exists a function u_ρ with $|u_\rho(t)| \leqq \delta$ for every $t \in [0, T]$ and

$$a(\delta)\rho = X(T) \int_0^T X^{-1}Du_\rho. \tag{4.22}$$

Since $\sigma_m \downarrow 0$ there exists a $l(m)$ such that $\sigma_{l(m)} < \min(\delta_m, \delta(\varepsilon))$. Hence, there exists an open sphere that is contained in $S(\beta^* + \varepsilon, (A, \gamma_m/2), (U, \sigma_{l(m)}))$ and contains x_F. That is, if $\rho \in S$ and $y = (x_F + a(\sigma_{l(m)})\rho)$, then there exists a control u_ρ with $|u_\rho(t)| \leqq \sigma_{l(m)}$ for $t \in [0, T]$ such that if $u(y) = u_0 + u_\rho$,

$$y = X(T) \int_0^T X^{-1}Du(y)$$

and $|C(u_0) - C(u(y))| < \varepsilon$.

Therefore, there exists a simplex containing x_F in its interior and contained in $S(\beta^* + \varepsilon, (A, \gamma_m/2), (U, \sigma_{l(m)}))$ with vertices $Q_1, \cdots, Q_{n+1}$. By Lemma 4.2, there exist scalars $\eta_i > 0$, $i = 1, \cdots, n + 1$, such that $|R_i - Q_i| < \eta_i$ for all i implies that x_F is in the convex hull of the R_i. Each Q_i is generated by a control $u(i)$ such that

$$u(i, t) \in U(U(t), \sigma_{l(m)}) \quad \text{a.e. in} \quad [0, T]. \tag{4.23}$$

By Lemma 4.1, for each $i = 1, \cdots, n + 1$ there exist sequences of piecewise constant (right-continuous) functions $u(i, l)$, $l = 1, 2, \cdots$, such that for each i, the sequence $\{u(i, l)\}_1^\infty \xrightarrow{s} u(i)$, and for each i and l and a.e. $t \in [0, T]$, $u(i, l, t) \in U(U(t), \sigma_{l(m)})$. Therefore, there exists $k_1(m)$ such that for all $l > k_1(m)$ and $1 \leqq i \leqq n + 1$,

$$\|u(i, l) - u(i)\|_{L_2} < \min(\gamma_m/2K, \eta_i/2K, \sigma_{l(m)}). \tag{4.24}$$

Recall that $\mathbf{x}(i, l)$ and $\mathbf{x}(i)$, respectively, denote the trajectory of P generated by

$u(i, l)$ and $u(i)$. Then for $i = 1, \cdots, n + 1$ and $l > k_1(m)$,

$$\|\mathbf{x}(i, l) - \mathbf{x}(i)\|_C < \tfrac{1}{2}\gamma_m,$$
$$|\mathbf{x}(i, l, T) - Q_i| < \eta_i, \tag{4.25}$$
$$|C(u(i, l)) - C(u(i))| < \varepsilon.$$

Let $\bar{u}(i, l)\,(\bar{u}(i))$ denote the natural projection of $u(i, l)\,(u(i))$ onto the sequence $\{0, t_l^1, \cdots, t_l^{R(l)-1}\}$, and $\bar{x}(i, l)\,(\bar{x}(i))$ denote the trajectory of P_l generated by $\bar{u}(i, l)\,(\bar{u}(i))$ with the initial value x_0.

Let $\tilde{x}(i, l)$ denote the piecewise constant (right-continuous) extension of $\bar{x}(i, l)$ to $[0, T]$. Since trajectories of P and $\tilde{P}_l$, $l = 1, 2, \cdots$, are uniformly bounded, for $1 \leqq i \leqq n + 1$, $\|\tilde{x}(i, l) - \mathbf{x}(i, l)\|_C \to 0$ as $l \to \infty$. Therefore, since the cost is continuous, there exists $k(m) \geqq k_1(m)$ such that for $l > k(m)$,

$$|C_l(\bar{u}(i, l)) - C(u(i, l))| < \varepsilon. \tag{4.26}$$

Let R_i^l, $i = 1, \cdots, n + 1$, denote the terminal value of the trajectory $\bar{x}(i, l)$ of P_l generated by $\bar{u}(i, l)$. But, for $i = 1, \cdots, n + 1$ and $l \geqq k(m)$, $\mathbf{x}(i, l, T) = R_i^l$. Therefore, for $l \geqq k(m)$,

$$R_i^l \in S_l(\beta^* + 3\varepsilon, (A, \gamma_m), (U, \sigma_{l(m)}))$$

and

$$|R_i^l - Q_i| < \eta_i. \tag{4.27}$$

Hence, by Lemma 4.2, x_F is in the convex hull of the points $R_i^{k(m)}$, $1 \leqq i \leqq n + 1$. But, the set $S_{k(m)}(\beta^* + 3\varepsilon, (A, \gamma_m), (U, \sigma_{l(m)}))$ is convex, so x_F is also in this set.

Hence, there exist sequences of integers $l(m)$ and $k(m) \to \infty$ as $m \to \infty$ such that the limit superior of the sequence of optimal costs of the corresponding problems $\tilde{P}_{k(m)}((A, \gamma_m), (U, \sigma_{l(m)}))$ is less than or equal to β^*.

Conversely, suppose there exists a sequence of integers $l(m) \to \infty$ as $m \to \infty$ and sequences of scalars $\gamma_m \downarrow 0$ and $\sigma_m \downarrow 0$, such that the limit inferior of the sequence of optimal costs of the corresponding problems $P_{l(m)}((A, \gamma_m), (U, \sigma_m))$ is less than β^*. Denote $l(m)$ and all subsequences of $l(m)$ by m for simplicity of notation. As in the proof of Theorem 4.2, there exists a sequence of pairs $(\bar{x}_m, \bar{u}_m)$ (pairs can be used since the time is fixed), $m = 1, 2, \cdots$, admissible for the corresponding problems $P_m((A, \gamma_m), (U, \sigma_m))$ and with costs convergent to $\bar{\beta} < \beta^*$, that is an approximation of type (b) to a pair (x_0, u_0) admissible for P. All that is necessary is to prove that the cost of u_0 is $\bar{\beta}$.

Since the cost is weakly sequentially lower semicontinuous on $(C \times \bar{U})([0, T])$, $\{u_m\}_1^\infty \xrightarrow{w} u_0$, and $\{x_m\}_1^\infty$ converges uniformly to x_0 (see [6, pp. 268–269]),

$$\int_0^T f^0(x_0, u_0, t) \leqq \liminf_m \int_0^T f^0(x_m, u_m, t). \tag{4.28}$$

But, $\|x_m - \tilde{x}_m\|_C \to 0$, and f^0 is uniformly continuous on any compact subset of

$E^n \times E^r \times [0, T]$, so

$$\lim_m \left[\sum_{k=0}^{m-1} f^0(x_m^k, u_m^k, t_m^k)T/m - \int_0^T f^0(x_m, u_m, t) \right] = 0. \tag{4.29}$$

Therefore,

$$C(u_0) \leqq \liminf_m C_m(\bar{u}_m) = \bar{\beta} < \beta^*, \tag{4.30}$$

which is a contradiction.

THEOREM 4.5. *Theorem* 4.4 *is valid with $\tilde{P}_m$ replaced by the direct discretization P_m used in Theorem* 3.1, *if the hypothesis of Theorem* 4.3 *is satisfied.*

Proof of Theorem 4.5. The proof parallels the proof of Theorem 4.4, with additions analogous to those made in obtaining the proof of Theorem 4.3 from the proof of Theorem 4.2.

5. Summary. It has been proved theoretically that certain "continuous" linear optimal control problems can be replaced by sequences of finite-dimensional "discrete" optimal control problems. These "discrete" problems can be solved by mathematical programming techniques, and the solutions obtained approximate optimal solutions of the original problem. The implementation of these results is discussed in [19].

Acknowledgments. The author wishes to express her appreciation to C. D. Cullum for instigating this investigation and to the reviewers who made many helpful suggestions.

REFERENCES

[1] A. V. BALAKRISHNAN AND L. W. NEUSTADT, *Computing Methods in Optimization Problems*, Academic Press, New York, 1964.

[2] A. BRYSON AND W. DENHAM, *A steepest ascent method for solving optimum programming problems*, Trans. ASME Ser. E. J. Appl. Mech., 29 (1962), pp. 247–257.

[3] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, vol. 1, Interscience, New York, 1953.

[4] J. CULLUM, *Perturbations and approximations to continuous optimal control problems*, Mathematical Theory of Control, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, New York, 1967 pp. 156–169.

[5] ———, *Discrete approximations to continuous optimal control problems*, IBM Res. Rep. RC 1858, Yorktown Heights, New York, 1967.

[6] N. DUNFORD AND J. SCHWARTZ, *Linear Operators*, vol. 1, Interscience, New York, 1964.

[7] H. G. EGGLESTON, *Convexity*, Cambridge University Press, London, 1958.

[8] S. K. GODUNOV AND V. S. RYABENKI, *Theory of Difference Schemes*, North-Holland, Amsterdam, 1964.

[9] E. G. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

[10] G. LEITMANN, *Optimization Techniques*, Academic Press, New York, 1962.

[11] R. MCGILL, *Optimal control, inequality state constraints, and the generalized Newton-Raphson algorithm*, this Journal, 3 (1965), pp. 291–298.

[12] I. P. NATANSON, *Theory of Functions of a Real Variable*, vol. 1, Frederick Ungar, New York, 1964.

[13] L. W. NEUSTADT, *Discrete time optimal control systems*, Internat. Symposium on Nonlinear Differential Equations and Nonlinear Mechanics, Academic Press, New York, 1963, pp. 267–283.

[14] J. B. ROSEN, *Optimal control and convex programming*, Proc. IBM Scientific Computing Symposium on Control Theory and Its Applications, 1964, IBM Data Processing Division, White Plains, New York, 1966, pp. 223–238.

[15] J. B. ROSEN AND R. MEYER, *Solution of nonlinear two-point boundary value problems by linear programming*, Proc. Conference on the Mathematical Theory of Control, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, New York, 1967, pp. 71–81.

[16] E. ROXIN, *On the existence of optimal controls*, Michigan Math. J., 9 (1962), pp. 109–119.

[17] D. L. RUSSELL, *Penalty functions and bounded phase coordinates*, this Journal, 2 (1965), pp. 409–422.

[18] J. WARGA, *Necessary conditions for minimum in relaxed variational problems*, J. Math. Anal. Appl., 4 (1962), pp. 129–145.

[19] M. CANON AND J. CULLUM, *The numerical solution of linear, continuous optimal control problems by discrete approximations*, to appear.

[20] J. CULLUM, *Penalty functions and nonconvex continuous optimal control problems*, Second Internat. Conference on Computing Methods in Optimization Problems, San Remo, Italy, 1968, Academic Press, New York, 1969.

SIAM J. Control
Vol. 7, No. 1, February 1969

THE DECOUPLING OF MULTIVARIABLE SYSTEMS BY STATE FEEDBACK*

ELMER G. GILBERT†

1. Introduction. The objective of this paper is to develop a comprehensive theory for the decoupling of multivariable systems by state feedback. We begin by giving a preliminary formulation of the decoupling problem, discussing certain aspects of its solution, reviewing previous research, and indicating the contributions of this paper.

Consider the linear dynamical system with input u, output y, and state x:

$$\frac{dx}{dt} = Ax + Bu(t), \qquad y(t) = Cx. \tag{1.1}$$

Here t is time, $u(t)$ and $y(t)$ are real m-vectors, x is a real n-vector, and A, B and C are real, constant matrices of appropriate size. Often one is interested in applying feedback control in order to implement certain control objectives. For example, one might use the control law $u = \mathscr{L}_F y + \mathscr{L}_I v$, where $v(t)$, a real m-vector, is the input to the closed-loop system and $\mathscr{L}_F$ and $\mathscr{L}_I$ are linear operators. With suitable assumptions on initial conditions this leads to $y = \mathscr{L}_C v$, where $\mathscr{L}_C$ is a linear operator which represents the closed-loop system. A common control objective is to "decouple" the closed-loop system by making $\mathscr{L}_C$ be diagonal, i.e., causing $y_i = \mathscr{L}_C^i v_i, i = 1, \cdots, m$, where y_i and v_i are respectively the ith components of y and v. Early efforts in this direction relied on transfer-function descriptions for $\mathscr{L}_F$ and $\mathscr{L}_I$ and were characterized by a lack of rigor and of solid results. In this paper we consider control laws of the form originally proposed by Morgan [1]:

$$u(t) = Fx + Gv(t), \tag{1.2}$$

where F and G are real, constant matrices of appropriate size. This control law (state feedback) admits a precise problem formulation and is of real interest in applications.

The desire to decouple raises four questions: (a) Is decoupling possible? (b) What is the class of control laws which decouple? (c) What is the class of decoupled closed-loop systems? (d) What is the correspondence between elements of the classes mentioned in (b) and (c)? These four questions constitute the decoupling problem as it is treated in this paper.

Partial answers to the decoupling problem have been obtained. Morgan [1] gave a sufficient condition for decoupling (CB nonsingular) and under this con-

* Received by the editors November 8, 1967, and in revised form April 29, 1968.

† Computer, Information and Control Engineering, University of Michigan, Ann Arbor, Michigan 48104. This research was supported by the United States Air Force under Grant AF-AFOSR 814-66.

dition defined a rather restrictive class of control laws which decouple. These results were extended somewhat by Rekasius [2]. More recently Falb and Wolovich [3] gave necessary and sufficient conditions for decoupling, thus answering question (a). They also described a (restricted) class of control laws which decouple, which subsumes the classes introduced in [1] and [2]. Still more recently they obtained necessary and sufficient conditions on F and G for decoupling [4]. While this answers question (b), their conditions are in a cumbersome algebraic form which makes them difficult to apply when n is large. For several simple examples they have also characterized the class of decoupled closed-loop systems.

This paper extends the results outlined above to obtain more or less complete answers to questions (a), (b), (c) and (d). In addition the method of attack makes clearer the general nature of the decoupling problem and should lead to the solution of other interesting problems in nonlinear control and optimal control.

The paper is organized as follows. In § 2 we introduce notation, give a precise problem formulation, and state some important formulas. In § 3 it is shown that certain closed-loop-system properties are invariant with respect to F. These invariants lead naturally to necessary conditions for decoupling and an important matrix discovered by Falb and Wolovich [3]. The general approach to the decoupling problem is to treat an equivalent problem of simple structure. The required equivalence is introduced in § 4, along with the notion of an integrator decoupled system. The material in § 3 and § 4 yields an alternate proof of a theorem of Falb and Wolovich, which appears in § 5. Section 6 establishes a canonical form for integrator-decoupled systems which is the key to the main results, which are summarized in the theorems of § 7. In § 8 we discuss briefly the significance of the main results.

2. Problem formulation and basic formulas. Matrices, which we generally denote by capital letters, have real elements unless explicit dependence on the complex variable s is indicated. The notation $A = [\ ]$ will denote a partitioning of the matrix A into matrices or elements. We use I_n for the $n \times n$ identity matrix, E_i for the ith row of I_m, and 0 for the number zero or any null matrix.

An *m-input, m-output, n-th order system* S is the triple $\{A, B, C\}$, where A, B, C are respectively matrices of size $n \times n$, $n \times m$, $m \times n$. Although it is not really essential, we shall assume as is usual in the literature that $m \leqq n$. The *transfer function* of S is

$$H(s) = C(I_n s - A)^{-1}B. \tag{2.1}$$

Clearly $H(s)$ is an $m \times m$ rational matrix in the complex variable s. If s is interpreted as the Laplace transform variable, the relation of $H(s)$ to the Laplace transform solution of (1.1) is obvious.

In a similar way we introduce notation appropriate to the description of the closed-loop system arising from (1.2). A *control law* is the pair $\{F, G\}$, where the matrices F, G are respectively $m \times n$, $m \times m$. We say $S(F, G) = \{A + BF, BG, C\}$ is the system S with the control law $\{F, G\}$. The transfer function of $S(F, G)$ is

$$H(s, F, G) = C(I_n s - A - BF)^{-1}BG. \tag{2.2}$$

DEFINITION 1. The system $S(F, G)$ is *decoupled* if $H(\cdot, F, G)$ is diagonal and nonsingular.

This definition of decoupling is equivalent to the one given by Falb and Wolovich. By using it we give precise meaning to the questions raised in § 1.

The following formulas and notation are basic to our subsequent developments. By extending the well-known expansion for $(I_n s - A)^{-1}$, cf. [5, pp. 82–85], to $(I_n s - A - BF)^{-1}$, we have

$$H(s, F, G) = q(s, F)^{-1}(CBs^{n-1} + CR_1(F)Bs^{n-2} + \cdots + CR_{n-1}(F)B)G, \tag{2.3}$$

where

$$q(s, F) = s^n - q_1(F)s^{n-1} - \cdots - q_n(F) = \det(I_n s - A - BF), \tag{2.4}$$

$$R_0(F) = I_n, \qquad R_i(F) = (A + BF)R_{i-1}(F) - q_i(F)I_n, \quad i = 1, \cdots, n - 1. \tag{2.5}$$

Alternatively (2.5) may be replaced by

$$\begin{aligned} &R_0(F) = I_n, \\ &R_1(F) = (A + BF) - q_1(F)I_n, \\ &R_2(F) = (A + BF)^2 - q_1(F)(A + BF) - q_2(F)I_n, \\ &\vdots \\ &R_{n-1}(F) = (A + BF)^{n-1} - q_1(F)(A + BF)^{n-2} - \cdots - q_{n-1}(F)I_n. \end{aligned} \tag{2.6}$$

We adapt the above formulas to S by writing $H(s) = H(s, 0, I_m)$ and using the notations $q(s) = q(s, 0)$, $q_i = q_i(0)$, $R_i = R_i(0)$.

Occasionally it will be necessary to work with several systems concurrently, say S and $\bar{S}$. In these cases the notation developed above is extended in the obvious way, e.g., $\bar{q}(s, \bar{F}) = \det(I_n s - \bar{A} - \bar{B}\bar{F})$.

3. *F*-invariants.

In this section we study properties of $S(F, G)$ which are not affected by changes in F.

DEFINITION 2. An *F-invariant* of S is any property of $S(F, G)$ which for any fixed G does not depend on F.

Denote the ith row of $H(\cdot, F, G)$ by $H_i(\cdot, F, G)$ and define the integer $d_i(F, G)$ and the $1 \times m$ row matrix $D_i(F, G)$ as follows: if $H_i(\cdot, F, G) = 0$, $d_i(F, G) = n - 1$ and $D_i(F, G) = 0$; if $H(\cdot, F, G) \neq 0$, $d_i(F, G)$ is the integer j such that $\lim_{s\to\infty} s^{j+1}H_i(s, F, G)$ is nonzero and finite and $D_i(F, G) = \lim_{s\to\infty} s^{j+1}H_i(s, F, G)$. From (2.3) it is clear that $0 \leqq d_i(F, G) \leqq n - 1$.

PROPOSITION 1. *For* $i = 1, \cdots, m$, $d_i(F, G)$ *and* $D_i(F, G)$ *are F-invariants of S. In particular*: $D_i(F, G) = D_i G$ *and, for* G *nonsingular*, $d_i(F, G) = d_i$, *where* $d_i = d_i(0, I_m)$ *and* $D_i = D_i(0, I_m)$.

Proof. Let C_i be the ith row of C. From (2.3) and (2.6) with $F = 0$, $G = I_m$, it follows that

$$D_i = C_i A^{d_i} B \tag{3.1}$$

and

$$d_i = \begin{cases} 0, & C_iB \neq 0, \\ j, & C_iB = 0, \end{cases} \tag{3.2}$$

where j is the largest integer from $\{1, \cdots, n-1\}$ such that $C_iA^kB = 0$ for $k = 0, 1, \cdots, j-1$. Then by (2.6), $C_iR_k(F)B = 0, k = 0, 1, \cdots, d_i - 1$ and $C_iR_{d_i}(F)B = D_i$. From this the proposition is true by (2.3).

In engineering terms Proposition 1 says that certain "high frequency" gain properties of the closed-loop system are F-invariants. Falb and Wolovich introduce d_i and D_i (which they call B_i^*) via (3.1) and (3.2). However, they do not bring up the notion of invariance or attach physical meaning to these quantities. For future use we form the $m \times m$ matrix

$$D = \begin{bmatrix} D_1 \\ \vdots \\ D_m \end{bmatrix}. \tag{3.3}$$

The general question of F-invariants will not be developed here, although additional invariants are known. For instance, it is possible to prove the following.

PROPOSITION 2. *Let $h(s) = q(s) \det H(s)$. Then $h(s)$ is a polynomial in s of degree not greater than $n - m$ and*

$$h(s, F, G) = q(s, F) \det H(s, F, G) = h(s) \det G.$$

4. Integrator decoupled systems and control law equivalence. The key to the solution of the decoupling problem is a canonical representation of integrator decoupled systems. In this section integrator decoupled systems are defined and it is shown how they are related to the decoupling problem.

DEFINITION 3. $S = \{A, B, C\}$ is *integrator decoupled* (ID) if $D = \Gamma$, where Γ is diagonal and nonsingular, and $C_iA^{d_i+1} = 0, i = 1, \cdots, m$.

Denote the diagonal elements of Γ by $\gamma_1, \cdots, \gamma_m$. Then we have the following result.

PROPOSITION 3. *If S is* ID, *then $H(\cdot)$ is diagonal and has diagonal elements*

$$h_i(s) = \gamma_i s^{-d_i-1}, \qquad i = 1, \cdots, m.$$

Proof. Write

$$H_i(s) = q(s)^{-1}(C_iBs^{n-1} + C_iR_1Bs^{n-2} + \cdots + C_iR_{n-1}B). \tag{4.1}$$

Application of $C_iA^{d_i}B = D_i = \gamma_iE_i, C_iA^kB = 0$ for $k \neq d_i$ and (2.6) with $F = 0$ then gives

$$H_i(s) = q(s)^{-1}(s^{n-1-d_i} - q_1s^{n-2-d_i} - \cdots - q_{n-1-d_i})\gamma_iE_i. \tag{4.2}$$

Now from the Cayley–Hamilton theorem $C_iA^{n+j}B - q_1C_iA^{n+j-1}B - \cdots - q_nC_iA^jB = 0$, where j is any nonnegative integer. Taking $j = 0$ and using $C_iA^kB = 0$, we see that $k \neq d_i$ and $C_iA^{d_i}B \neq 0$ imply $q_{n-d_i} = 0$. Similarly, by taking $j = 1, \cdots, d_i$, we have $q_{n-d_i+1}, \cdots, q_n = 0$. Thus $q(s) = s^n - q_1 s^{n-1} - \cdots - q_{n-d_i-1} s^{d_i+1}$ and by (4.2) the proof is complete.

By Proposition 3 the transfer properties of an ID system are such that the ith output is the $(d_i + 1)$-fold integral of the ith input. This justifies the terminology, integrator decoupled. The example

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

shows that the converse of Proposition 3 is not true.

To establish the connection between ID systems and the decoupling problem we introduce the following definition.

DEFINITION 4. $S = \{A, B, C\}$ and $\bar{S} = \{\bar{A}, \bar{B}, \bar{C}\}$ are *control law equivalent* (CLE) if a one-to-one correspondence between $\{F, G\}$ and $\{\bar{F}, \bar{G}\}$ can be established such that, for this correspondence, $H(\cdot, F, G) = \bar{H}(\cdot, \bar{F}, \bar{G})$.

Remark 1. If the decoupling problem has been solved for S, it has been solved for $\bar{S}$.

Remark 2. Control law equivalence is transitive, i.e., if S and $\bar{S}$ are CLE and $\bar{S}$ and $\bar{\bar{S}}$ are CLE, then S and $\bar{\bar{S}}$ are CLE.

PROPOSITION 4. *Consider the system $S = \{A, B, C\}$, where D is nonsingular. Let A^* denote the $m \times n$ matrix*

$$A^* = \begin{bmatrix} C_1 A^{d_1+1} \\ \vdots \\ C_m A^{d_m+1} \end{bmatrix}. \tag{4.3}$$

Then the systems S and $\bar{S} = S(-D^{-1}A^*, D^{-1})$ are CLE. Furthermore $\bar{S}$ is ID and $\bar{d}_i = d_i, \bar{D}_i = E_i$ for $i = 1, \cdots, m$.

Proof. The one-to-one correspondence between $\{F, G\}$ and $\{\bar{F}, \bar{G}\}$, $DF + A^* = \bar{F}$ and $DG = \bar{G}$, proves the CLE property since then $A + BF = \bar{A} + \bar{B}\bar{F}$, $BG = \bar{B}\bar{G}$ and $C = \bar{C}$. The last part of the proposition follows by direct calculation of $\bar{d}_i$, $\bar{C}_i \bar{A}^{\bar{d}_i+1}$ and $\bar{D}_i$.

5. Necessary and sufficient conditions for decoupling. From the results of § 3 and § 4 we obtain by different means the theorem of Falb and Wolovich [3].

THEOREM 1. *S can be decoupled if and only if D is nonsingular. If $\{F, G\}$ decouples $S(F, G)$, $G = D^{-1}\Lambda$, where the $m \times m$ matrix Λ is diagonal and nonsingular.*

Proof. If $H(\cdot, F, G)$ is decoupled, then $H_i(\cdot, F, G) = h_i(\cdot, F, G)E_i$, where $h_i(\cdot, F, G) \neq 0$. This together with Proposition 1 implies $D_i G = \lambda_i E_i$, $i = 1, \cdots, m$. The numbers $\lambda_1, \cdots, \lambda_m$ are all nonzero. Suppose to the contrary. Then for some i, $D_i G = 0$. But if $H(\cdot, F, G)$ is to be nonsingular, G must be nonsingular. This implies $D_i = 0$ and $d_i = n - 1$, and from (3.1), (3.2), (2.6) and (2.3) we obtain $H_i(\cdot, F, G) = 0$, which contradicts the nonsingularity of $H(\cdot, F, G)$. From $D_i G = \lambda_i E_i$, $\lambda_i \neq 0$, $i = 1, \cdots, m$, we have $DG = \Lambda = \operatorname{diag}(\lambda_1, \cdots, \lambda_m)$, Λ nonsingular, which proves the necessary conditions. Sufficiency follows from Propositions 3 and 4 and the control law $\{-D^{-1}A^*, D^{-1}\}$.

Falb and Wolovich [3] prove Theorem 1 by manipulating some rather involved algebraic expressions. Besides being simpler, our proof has the advantage that it makes clear that the necessary conditions have their origin in the F-invariants of Proposition 1. The matrix A^* was used also by Falb and Wolovich in their sufficiency proof.

Since the nonsingularity of D plays such an important role in the decoupling problem, it deserves some special comment. It is easy to see that $\det H(\cdot) = 0$ implies $\det D = 0$. In this case we say S has *strong inherent coupling* and it is obvious that no control law can effect decoupling. If $\det D \neq 0$, we say S has *no inherent coupling*. If $\det D = 0$ and $\det H(\cdot) \neq 0$, we say S has *weak inherent coupling*. Systems which have weak inherent coupling cannot be decoupled by state feedback, but other control laws can achieve decoupling. We shall not pursue this issue in depth here, but the following indicates one path which can be taken.

The system $S = \{A, B, C\}$,

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

has weak inherent coupling because $\det H(\cdot) \neq 0$. We now form a new system $\tilde{S} = \{\tilde{A}, \tilde{B}, \tilde{C}\}$ which is related to S in the following way:

$$\tilde{A} = \begin{bmatrix} A & BK_2 \\ 0 & \bar{A} \end{bmatrix}, \quad \tilde{B} = \begin{bmatrix} BK_1 \\ \bar{B} \end{bmatrix}, \quad \tilde{C} = [C \quad 0],$$

where $\bar{A}, \bar{B}, K_1, K_2$ are respectively $\bar{n} \times \bar{n}, \bar{n} \times m, m \times m, m \times \bar{n}$ matrices. $\tilde{S}$ may be interpreted as the dynamical system (state $\tilde{x} = \begin{bmatrix} x \\ \bar{x} \end{bmatrix}$, input $\bar{u}(t)$) arising from the interconnection of (1.1) and

$$\frac{d\bar{x}}{dt} = \bar{A}\bar{x} + \bar{B}\bar{u}(t), \tag{5.1}$$
$$u(t) = K_1\bar{u}(t) + K_2\bar{x}.$$

Thus (5.1) acts as a precompensator for (1.1). If we choose $\bar{n} = 1$ and

$$\bar{A} = [0], \quad \bar{B} = [0 \quad 1], \quad K_1 = \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}, \quad K_2 = \begin{bmatrix} 0 \\ -1 \end{bmatrix},$$

it is easily verified that $\tilde{S}$ has no inherent coupling $\left(\tilde{D} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}\right)$. In general it is always possible to precompensate a system S which has weak inherent coupling so as to obtain a system $\tilde{S}$ with no inherent coupling. The development which shows this is tedious but straightforward.

6. Canonically decoupled systems. If in (1.1) we change the coordinate system by writing $\bar{x} = Tx$, where T is a nonsingular matrix, we obtain a new system $\bar{S} = \{\bar{A}, \bar{B}, \bar{C}\} = \{TAT^{-1}, TB, CT^{-1}\}$. This motivates what follows.

DEFINITION 5. S and $\bar{S}$ are *similar* if there exists a nonsingular $n \times n$ matrix T such that $TA = \bar{A}T$, $TB = \bar{B}$ and $C = \bar{C}T$.

Remark 3. If S and $\bar{S}$ are similar, $q(\cdot) = \bar{q}(\cdot)$ and $H(\cdot) = \bar{H}(\cdot)$. Thus $D = \bar{D}$ and $d_i = \bar{d}_i$, $i = 1, \cdots, m$. Furthermore if S is ID, $\bar{S}$ is ID.

Remark 4. Similar systems are CLE. This is a consequence of the correspondence $F = \bar{F}T$ and $G = \bar{G}$.

When S and $\bar{S}$ satisfy the conditions in Definition 5 we shall use the terminology that *T carries S into $\bar{S}$.*

For a canonically decoupled system (to be defined shortly) the decoupling problem has a particularly simple form. The main result of this section (Theorem 2) is to show that every ID system is similar to a canonically decoupled system. By Remarks 2 and 4 and Proposition 4, this means that if S can be decoupled ($\det D \neq 0$) it is possible to find a canonically decoupled system which is CLE to S. Thus by Remark 1 the treatment of the decoupling problem for S is simplified.

DEFINITION 6. $S = \{A, B, C\}$ is *canonically decoupled* (CD) if the following conditions are satisfied:

(i) The matrices A, B and C have the partitioned form:

$$A = \begin{bmatrix} A_1 & 0 & 0 & \cdots & 0 & 0 & A_1^u \\ 0 & A_2 & 0 & \cdots & 0 & 0 & A_2^u \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & A_m & 0 & A_m^u \\ A_1^c & A_2^c & A_3^c & \cdots & A_m^c & A_{m+1} & A_{m+1}^u \\ 0 & 0 & 0 & \cdots & 0 & 0 & A_{m+2} \end{bmatrix}, \quad \begin{array}{l} A_i \text{ is } p_i \times p_i, \\ A_i^c \text{ is } p_{m+1} \times p_i, \\ A_i^u \text{ is } p_i \times p_{m+2}, \end{array}$$

$$B = \begin{bmatrix} b_1 & 0 & \cdots & 0 \\ 0 & b_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & b_m \\ b_1^c & b_2^c & \cdots & b_m^c \\ 0 & 0 & \cdots & 0 \end{bmatrix}, \quad \begin{array}{l} b_i \text{ is } p_i \times 1, \\ b_i^c \text{ is } p_{m+1} \times 1, \end{array}$$

$$C = \begin{bmatrix} c_1 & 0 & \cdots & 0 & 0 & c_1^u \\ 0 & c_2 & \cdots & 0 & 0 & c_2^u \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & c_m & 0 & c_m^u \end{bmatrix}, \quad \begin{array}{l} c_i \text{ is } 1 \times p_i, \\ c_i^u \text{ is } 1 \times p_{m+2}, \end{array}$$

where $p_i \geqq d_i + 1$, $i = 1, \cdots, m$.

(ii) For $i = 1, \cdots, m$ the matrices A_i, b_i and c_i have the partitioned form:

$$A_i = \begin{bmatrix} \begin{bmatrix} 0 & I_{d_i} \\ 0 & 0 \end{bmatrix} & 0 \\ \Upsilon_i & \Phi_i \end{bmatrix}, \quad \begin{array}{l} \Upsilon_i \text{ is } r_i \times (d_i + 1), \\ \Phi_i \text{ is } r_i \times r_i, \end{array}$$

$$b_i = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \gamma_i \\ \beta_i \end{bmatrix}, \quad \beta_i = \begin{bmatrix} \beta_{i1} \\ \beta_{i2} \\ \vdots \\ \beta_{ir_i} \end{bmatrix},$$

$$c_i = [1 \quad 0 \quad \cdots \quad 0],$$

where $r_i = p_i - 1 - d_i$.

(iii) For $i = 1, \cdots, m$ the p_i column matrices $b_i, Ab_i, \cdots, A^{p_i - 1}b_i$ are linearly independent.

(iv) Let $p = \sum_{i=1}^{m} p_i$. If $p_{m+1} \neq 0$ and the n-row $\eta = [\eta_1 \quad \eta_2 \quad \cdots \quad \eta_n]$ is such that $\eta_{p+1}, \cdots, \eta_{p+p_{m+1}}$ are not all zero, then the row matrix function $\eta(I_n s - A)^{-1}B$ has at least two nonzero elements.

To proceed we need some additional notation and terminology. Let $\mathcal{Q}$ denote the n-dimensional space of n element row matrices. For $i = 1, \cdots, m$ define

(6.1) $\mathcal{Q}_i = \{\eta | \eta \in \mathcal{Q}; \eta A^j B_k = 0 \text{ for } k = 1, \cdots, m, \ k \neq i \text{ and } j = 0, \cdots, n-1\}$,

where B_i is the ith column of B. According to accepted practice we say $S = \{A, B, C\}$ is controllable if the nm columns $A^j B_k, j = 0, \cdots, n-1, k = 1, \cdots, m$, span the n-dimensional linear space of n columns.

LEMMA 1. *Assume $S = \{A, B, C\}$ is ID and controllable. Then for $i = 1, \cdots, m$ the following conditions are satisfied*:

(i) *$\mathcal{Q}_i$ is a row invariant subspace of A, i.e., $\eta \in \mathcal{Q}_i$ implies $\eta A \in \mathcal{Q}_i$*;

(ii) *$\mathcal{Q}_i \cap \mathcal{Q}_j = \{0\}$ for $j = 1, \cdots, m, j \neq i$*;

(iii) *$C_i, C_iA, \cdots, C_iA^{d_i}$ are linearly independent elements of $\mathcal{Q}_i$.*

Proof. To prove (i) we need only to show that $\eta \in \mathcal{Q}_i$ implies $\eta A^n B_k = 0$ for $k = 1, \cdots, m, \ k \neq i$. But this follows from the Cayley-Hamilton theorem ($A^n = q_1 A^{n-1} + \cdots + q_n I_n$) and the definition of $\mathcal{Q}_i$. Assume $\eta \in \mathcal{Q}_i \cap \mathcal{Q}_j$ for $i \neq j$. Then from the definition of $\mathcal{Q}_i$ and $\mathcal{Q}_j$ it is apparent that $\eta A^j B_k = 0$ for $j = 0, \cdots, n-1, \ k = 1, \cdots, m$. By the controllability of S this implies $\eta = 0$ and (ii) is true. From (3.1) and Definition 3, $C_i A^{d_i} \neq 0$ and $C_i A^{d_i + k} = 0$ for $k = 1, 2, \cdots$. Now assume $\rho_0 C_i + \rho_1 C_i A + \cdots + \rho_{d_i} C_i A^{d_i} = 0$, where $\rho_0, \cdots, \rho_{d_i}$ are scalars. Postmultiply this equation by A^{d_i} and obtain $\rho_0 C_i A^{d_i} = 0$ which implies $\rho_0 = 0$. By multiplying by successively lower powers of A we obtain $\rho_0, \rho_1, \cdots, \rho_{d_i} = 0$, which implies $C_i, \cdots, C_i A^{d_i}$ are linearly independent. From (3.1) and Definition 3 it follows that $C_i A^j B = 0$ for $j \geqq 0, \ j \neq d_i$ and

$C_iA^{d_i}B = \gamma_iE_i$. These conditions show that $C_iA^j \in \mathscr{Q}_i$, where j is any nonnegative integer. Thus (iii) is proved.

Because of Lemma 1 there exists a linear space $\mathscr{Q}_{m+1} \subset \mathscr{Q}$ such that the direct sum $\mathscr{Q}_1 \oplus \mathscr{Q}_2 \oplus \cdots \oplus \mathscr{Q}_{m+1} = \mathscr{Q}$. We adopt the notation

$$p_i = \dim \mathscr{Q}_i, \qquad i = 1, \cdots, m+1, \qquad p = \sum_{i=1}^{m} p_i. \tag{6.2}$$

Clearly $p_{m+1} = n - p$ is uniquely defined by S although $\mathscr{Q}_{m+1}$ is not (unless $p_{m+1} = 0$). Moreover from part (iii) of Lemma 1 it is clear that $p_i \geqq d_i + 1$, $i = 1, \cdots, m$. Now we can state the following lemma.

LEMMA 2. *Assume $S = \{A, B, C\}$ is ID, controllable, and $p_{m+1} \neq 0$. Let $\eta \in \mathscr{Q}$ and write $\eta \in \sum_{i=1}^{m+1} \xi_i$, where $\xi_i \in \mathscr{Q}_i$ for $i = 1, \cdots, m+1$. If $\xi_{m+1} \neq 0$, then there exist at least two integers from the set $\{1, \cdots, m\}$, say q and r, such that $\eta A^jB_q \neq 0$ for at least one $j \in \{0, \cdots, n-1\}$ and $\eta A^jB_r \neq 0$ for at least one $j \in \{0, \cdots, n-1\}$.*

Proof. If $\eta A^jB_k = 0$ for all $j = 0, \cdots, n-1$ and $k = 1, \cdots, m$, the controllability of S would imply $\eta = 0$. Thus there is at least one integer from $\{1, \cdots, m\}$, say q, such that $\eta A^jB_q \neq 0$ for all $j = 0, \cdots, n-1$. If q were the only such integer then $\eta \in \mathscr{Q}_q$. But this would imply $\xi_{m+1} = 0$ and thus the lemma is proved.

PROPOSITION 5. *Assume S is ID and controllable. Then S is similar to a CD system $\bar{S}$, where $\bar{p}_i = p_i, i = 1, \cdots, m+1$ and $\bar{p}_{m+2} = 0$.*

Proof. We use the results of Lemmas 1 and 2 and in Definition 6 replace S by $\bar{S}$. First we form the matrix

$$Q = \begin{bmatrix} Q_1 \\ \vdots \\ Q_{m+1} \end{bmatrix}, \tag{6.3}$$

where the rows of the $p_i \times n$ matrix Q_i are a basis for $\mathscr{Q}_i$. Because of the definition of $Q_1, \cdots, Q_{m+1}$, the rows of Q, which we denote by $q_1^*, \cdots, q_n^*$, are a basis for $\mathscr{Q}$. If we define $\bar{A}$ by $\bar{A}Q = QA$, the elements of the ith row of $\bar{A}$ are the components of q_i^*A with respect to the basis $q_1^*, \cdots, q_n^*$. Using this and part (i) of Lemma 1, we see that $\bar{A}$ has the structure of Definition 6, part (i). To be more specific define

$$Q_i = \begin{bmatrix} C_i \\ C_iA \\ \vdots \\ C_iA^{d_i} \\ q_{\sigma_i}^* \\ \vdots \\ q_{\rho_i}^* \end{bmatrix}, \tag{6.4}$$

where $q_{\sigma_i}^*, \cdots, q_{\rho_i}^*$ are any rows which extend $C_i, C_iA, \cdots, C_iA^{d_i}$ to form a basis for $\mathscr{Q}_i$. This, together with $C_iA^{d_i+1} = 0$, gives $\bar{A}_i$ the structure of Definition 6, part (ii). Moreover, $\bar{d}_i = d_i$ and $\bar{p}_i = p_i$.

Now define $\bar{B} = QB$. Using (6.3), (6.4), and the definition of $\mathcal{Q}_i$ gives $\bar{B}$ the structure of Definition 6, part (i). The further structure indicated in part (ii) follows from (6.4) and the fact that $C_iA^{d_i}B_i = \gamma_i$. Define $\bar{C}$ by $C = \bar{C}Q$. Then (6.3) and (6.4) give $\bar{C}$ the structure of Definition 6, parts (i) and (ii).

From the foregoing it is obvious that Q carries S into $\bar{S}$. It remains to show that parts (iii) and (iv) of Definition 6 are true for $\bar{S}$. Suppose that (iii) is not true. Then from the form of $\bar{A}$ and $\bar{B}$ indicated in (i), the columns $\bar{A}^j\bar{B}_k, j = 0, \cdots, n-1$, $k = 1, \cdots m$, do not span the n-dimensional column space. Because $\bar{A}^j\bar{B}_k = QA^jB_k$ this implies S is not controllable. This contradiction proves that (iii) holds for $\bar{S}$. To prove that $\bar{S}$ satisfies part (iv) of Definition 6 we note that

$$\eta(I_ns - A)^{-1}B = \bar{\eta}(I_ns - \bar{A})^{-1}\bar{B},$$

where $\eta = \bar{\eta}Q$. By the definition of Q, η satisfies the condition $\xi_{m+1} \neq 0$ (see Lemma 2 for notation) if and only if $\bar{\eta}_{p+1}, \cdots, \bar{\eta}_n$ are not all zero. Thus we need only to show that $\eta(I_ns - A)^{-1}B$ has at least two nonzero elements if $\xi_{m+1} \neq 0$. Using

$$\eta(I_ns - A)^{-1}B_k = q(s)(s^{n-1}\eta B_k + s^{n-2}\eta R_1B_k + \cdots + \eta R_{n-1}B_k),$$

we easily see that if $\eta A^jB_k \neq 0$ for at least one $j \in \{0, \cdots, n-1\}$, then $\eta(I_ns - A)^{-1}B_k \not\equiv 0$. Since the kth element of $\eta(I_ns - A)^{-1}B$ is $\eta(I_ns - A)^{-1}B_k$, Lemma 2 gives the desired result.

Proposition 5 requires S to be controllable. To remove this restriction we need the following lemma.

LEMMA 3. *For the n-th order system* $S = \{A, B, C\}$ *let* $n_c = \dim \mathscr{C}$, *where* $\mathscr{C}$ *is the subspace spanned by the mn columns* $A^jB_k, j = 0, \cdots, n-1, k = 1, \cdots, m$. *Then S is similar to* $\tilde{S} = \{\tilde{A}, \tilde{B}, \tilde{C}\}$, *where* :

(i)

$$\tilde{A} = \begin{bmatrix} A^c & A^u \\ 0 & A^U \end{bmatrix}, \quad \begin{array}{l} A^c \textit{ is } n_c \times n_c, \\ A^u \textit{ is } n_c \times (n - n_c), \\ A^U \textit{ is } (n - n_c) \times (n - n_c), \end{array}$$

$$\tilde{B} = \begin{bmatrix} B^c \\ 0 \end{bmatrix}, \quad B^c \textit{ is } n_c \times m,$$

$$\tilde{C} = [C^c \quad C^u], \quad C^c \textit{ is } m \times n_c, \quad C^u \textit{ is } m \times (n - n_c),$$

(ii) $S^c = \{A^c, B^c, C^c\}$ *is controllable*,

(iii) *if S is ID*, S^c *is ID and* $\Gamma^c = \Gamma$.

Proof. Parts (i) and (ii) are well known [6], [7] and may be established by taking the first n_c columns of a nonsingular matrix L to be a basis for $\mathscr{C}$. Then $T_1 = L^{-1}$ carries S into $\tilde{S}$. Part (iii) follows from Remark 3 and direct calculation of $\tilde{C}_i\tilde{A}^{\tilde{d}_i}\tilde{B}$ and $\tilde{C}_i\tilde{A}^{\tilde{d}_i+1}$ in terms of A^c, B^c and C^c.

The steps required to construct a CD representation of an ID system can now be summarized. Let S be an nth order ID system. Apply Lemma 3 obtaining $\mathscr{C}$ and thence a matrix T_1 which carries S into $\tilde{S}$. Since S^c is ID and controllable, Proposition 5 is applicable with S^c taking the role of S in Proposition 5. The matrix

Q which appears in the proof of Proposition 5 will in this case be $n_c \times n_c$. Define $\hat{p}_{m+2} = n - n_c$ and

$$T_2 = \begin{bmatrix} Q & 0 \\ 0 & I_{\hat{p}_{m+2}} \end{bmatrix}.$$

Then direct calculation shows that T_2 carries $\tilde{S}$ into the CD system $\hat{S}$, where $\hat{p}_i = \dim \mathcal{Q}_i^c$, $i = 1, \cdots, m + 1$. Thus $T_2 T_1$ carries S into $\hat{S}$ and we have a constructive proof of the promised result.

THEOREM 2. *Every ID system is similar to a CD system.*

7. Principal results. In this section we characterize the solution of the decoupling problem for CD systems and then by means of Theorem 2 extend these results to general systems.

THEOREM 3. *If S is CD, the control law $\{F, G\}$ decouples S if and only if*

$$F = \begin{bmatrix} \theta_1 & 0 & 0 & \cdots & 0 & 0 & \theta_1^u \\ 0 & \theta_2 & 0 & \cdots & 0 & 0 & \theta_2^u \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \theta_m & 0 & \theta_m^u \end{bmatrix},$$

where θ_i is $1 \times p_i$ and θ_i^u is $1 \times p_{m+2}$, and

$$G = \operatorname{diag}(\lambda_1, \cdots, \lambda_m), \qquad \lambda_i \neq 0, \quad i = 1, \cdots, m.$$

Proof. Sufficiency follows by substitution which shows $H(\cdot, F, G)$ is diagonal. In fact, the ith diagonal element of $H(s, F, G)$ is given by

$$h_i(s, F, G) = c_i(I_{p_i}s - A_i - b_i\theta_i)^{-1} b_i \gamma_i \lambda_i. \tag{7.1}$$

The necessity of $G = \operatorname{diag}(\lambda_1, \cdots, \lambda_m)$ and $\lambda_i \neq 0$ is an obvious consequence of Theorem 1. To prove the necessity of the condition on F we write

$$H(s, F, G) = H(s)(I_m - F(I_n s - A)^{-1}B)^{-1}G, \tag{7.2}$$

an identity which is derived by straightforward manipulation of the two obvious identities: $(I_n s - A)^{-1}(I_n - BF(I_n s - A)^{-1})^{-1}B = (I_n s - A - BF)^{-1}B$ and $(I_n - BF(I_n s - A)^{-1})B = B(I_m - F(I_n s - A)^{-1}B)$. Since $H(\cdot)$ is diagonal, (7.2) implies $F(I_n s - A)^{-1}B$ must be diagonal if $S(F, G)$ is to be decoupled. By partitioning F into rows and using Definition 6, this leads to the required conditions on F.

THEOREM 4. *Assume S is CD and the control law $\{F, G\}$ has the form indicated in Theorem 3. Then:*

(i) $$h_i(s, F, G) = \frac{\alpha_i(s)\gamma_i\lambda_i}{\psi_i(s, \sigma_i)}, \qquad i = 1, \cdots, m,$$

where $\alpha_i(s) = s^{r_i} - \alpha_{i1}s^{r_i - 1} - \cdots - \alpha_{ir_i}$ and

$$\psi_i(s, \sigma_i) = s^{p_i} - \sigma_{i1}s^{p_i - 1} - \cdots - \sigma_{ip_i}, \qquad \sigma_i = [\sigma_{ip_i} \cdots \sigma_{i1}];$$

(ii) $\alpha_i(s) = \det(I_{r_i}s - \Phi_i)$;

(iii) $\theta_i = (\sigma_i - \pi_i)V_i$, *where* V_i *is a* $p_i \times p_i$ *nonsingular matrix which depends only on* A_i *and* b_i, *and the* $1 \times p_i$ *matrix* $\pi_i = [0 \cdots 0\, \alpha_{ir_i} \cdots \alpha_{i1}]$;

(iv) $q(s, F) = \alpha_{m+1}(s)\alpha_{m+2}(s)\prod_{i=1}^{m} \psi_i(s, \sigma_i)$, *where* $\alpha_i(s) = \det(I_{p_i}s - A_i)$, $i = m + 1, m + 2$.

Proof. From (7.1) it is apparent that $h_i(\cdot, F, G)$ may be interpreted as the transfer function of the system $S^i(\theta_i, \lambda_i)$, where $S^i = \{A_i, b_i, c_i\}$. Since S^i is a controllable single-input, single-output system, the theory developed by Bass and others (see, e.g., Morgan [8], [9]) may be applied. We summarize this theory in the following lemma.

LEMMA 4. *Assume* $S = \{A, b, c\}$ *is single-input, single-output, order* n *and controllable. Then the transfer function of* $S(\theta, \lambda)$ *has the form*

$$H(s, \theta, \lambda) = \frac{\omega(s)\lambda}{\psi(s, \sigma)},$$

where $\omega(s)$ *is a polynomial of degree* $n - 1$ *or less and* $\psi(s, \sigma) = s^n - \sigma_1 s^{n-1} - \cdots - \sigma_n$. *Let* $\sigma = [\sigma_n \cdots \sigma_1]$, $\pi = [q_n \cdots q_1]$ *and define the matrix* $K = [k_1 \cdots k_n]$, *where the columns* $k_i = R_{n-i}b, i = 1, \cdots, n$. *Then* K *is nonsingular and* $\theta K = \sigma - \pi$.

Except for the form of $\alpha_i(s)$, application of Lemma 4 to S^i proves part (i) of the theorem. From the form of A_i it is clear that $q_i(s) = \det(I_{p_i}s - A_i) = s^{d_i+1} \cdot \det(I_{r_i}s - \Phi_i) = s^{d_i+1}\alpha_i(s) = s^{p_i} - \alpha_{i1}s^{p_i-1} - \cdots - \alpha_{ir_i}s^{p_i-r_i}$, where we have used the notation of (ii). Letting V_i correspond to K^{-1} of Lemma 4 verifies (iii). The remaining part of (i) follows by noting that

$$h_{ii}(s, 0, 1) = \omega_i(s)\psi_i^{-1}(s, 0) = \omega_i(s)(s^{d_i-1}\alpha_i(s))^{-1} = \gamma_i s^{-d_i-1}.$$

Part (iv) is obtained by observing that $I_n s - A - BF = W(s)$ can be written

$$W = \begin{bmatrix} W_{11} & W_{12} \\ 0 & W_{22} \end{bmatrix},$$

where $W_{22} = I_{p_{m+2}}s - A_{m+2}$. Thus $q(s, F, G) = \det W_{11} \det W_{22}$. Finally, W_{11} is quasi-triangular and is easily expanded to give

$$\det W_{11} = \det(I_{p_{m+1}}s - A_{m+1}) \prod_{i=1}^{m} \psi_i(s, \sigma_i).$$

THEOREM 5. *Assume* $S = \{A, B, C\}$ *can be decoupled. If* $\{F, G\}$ *decouples* S, *the diagonal elements of* $H(\cdot, F, G)$ *have the form given in part* (i) *of Theorem* 4 *where the integers* p_i *and* r_i *and the polynomials* $\alpha_i(s)$ *are uniquely determined by* S, *and* $\gamma_i = 1, i = 1, \cdots, m$. *Furthermore,* $q(s, F)$ *has the form given in part* (iv) *of Theorem* 4, *where* $\alpha_{m+1}(s)$ *and* $\alpha_{m+2}(s)$ *are polynomials of degree* p_{m+1} *and* p_{m+2} *uniquely determined by* S. *The class of control laws which decouple* S *can be characterized by* $G \in \mathcal{G}$ *and* $F \in \mathcal{F}$, *where* $\mathcal{G}$ *is an* m*-dimensional linear space and* $\mathcal{F}$ *is a* $(\sum_{i=1}^{m} p_i + mp_{m+2})$*-dimensional linear manifold. More specifically, there exist matrices* $G_i, J_1^i, \cdots, J_{p_i}^i$, $i = 1, \cdots, m$, *and an* (mp_{m+2})*-dimensional linear space*

$\mathscr{F}^u$, which are uniquely determined by S, such that

$$G = \sum_{i=1}^{m} \lambda_i G_i, \tag{7.3}$$

$$F = -D^{-1}A^* + \sum_{i=1}^{m} \sum_{k=1}^{p_i} (\sigma_{ik} - \pi_{ik})J_k^i + F^u, \tag{7.4}$$

where $F^u \in \mathscr{F}^u$, $\pi_{ik} = \alpha_{ik}$, $k = 1, \cdots, r_i$, and $\pi_{ik} = 0, k = r_i + 1, \cdots, p_i$.

Proof. The form of $H(\cdot, F, G)$ follows immediately from the CLE property between S and a CD system (Proposition 4 and Theorem 2) and Theorem 4. The one-to-one control law correspondence associated with this CLE property yields

$$G = D^{-1}\Lambda, \qquad \Lambda = \operatorname{diag}(\lambda_1, \cdots, \lambda_m), \tag{7.5}$$

and

$$F = -D^{-1}A^* + D^{-1}\begin{bmatrix} \theta_1 & 0 & 0 & \cdots & 0 & 0 & \theta_1^u \\ 0 & \theta_2 & 0 & \cdots & 0 & 0 & \theta_2^u \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \theta_m & 0 & \theta_m^u \end{bmatrix} T_2T_1, \tag{7.6}$$

where T_1 and T_2 are the nonsingular matrices which arise in the proof of Theorem 2. Results (7.3) and (7.4) are a direct consequence of (7.5) and (7.6), Theorem 3 and Theorem 4. Substitution of (7.6) into det $(I_ns - A - BF)$ leads to the expression for $q(s, F)$.

8. Discussion. Theorem 5 establishes all the data needed for the design of a decoupled multivariable system. The class of decoupled systems is given and convenient formulas for computing F and G for an arbitrarily specified decoupled system within the class exist. The general approach for obtaining the data for these formulas (p_i, r_i, α_{ij}, π_{ij}, A^*, D, G_i, J_j^i) should be clear from the foregoing developments. But for nontrivial cases of S, hand calculations are not practical. For this reason a computer program is now being written. Given A, B and C, it will generate all the necessary data. This program, along with some example applications, will be reported in a subsequent paper.

It is also possible to determine stability and decide when decoupling by output feedback [4] is possible. We say the system $S(F, G)$ is *stable* if $q(s, F)$ is Hurwitz, i.e., all roots of $q(s, F) = 0$ have negative real parts. Clearly the design of a stable decoupled system is impossible if either $\alpha_{m+1}(s)$ or $\alpha_{m+2}(s)$ is not Hurwitz. If both $\alpha_{m+1}(s)$ and $\alpha_{m+2}(s)$ are Hurwitz, $S(F, G)$ can be made stable by appropriate choice of $\sigma_1, \cdots, \sigma_m$. Using the design formulas of Falb and Wolovich [3], we see that the stability question is more critical. It can be shown that these formulas lead to

$$h_i(s, F, G) = \frac{\lambda_i}{s^{d_i+1} + m_{i1}s^{d_i} + \cdots + m_{i(d_i+1)}}, \qquad i = 1, \cdots, m.$$

Thus $S(F, G)$ can be stable if and only if the $\alpha_i(s), i = 1, \cdots, m + 2$, are Hurwitz.

We say S is decoupled by *output feedback* if there exists a pair of $m \times m$ matrices $\{K, G\}$ such that $S(KC, G)$ is decoupled. The motivation for decoupling by output feedback is clear since it corresponds to replacing (1.2) by

$$u(t) = Ky(t) + Gv(t).$$

If $\{K, G\}$ is to output decouple S, KC must have the form of F in Theorem 5. This means it may not be possible to output decouple S even if D is nonsingular. If $\{K, G\}$ output decouples S, linear constraint equations on $\sigma_1, \cdots, \sigma_m$ may be imposed. The details of the analysis which gives these results are straightforward and are therefore omitted.

Still other questions arise: what is the effect of parameter variations and disturbance inputs, what should be done if D is singular or $\alpha_{m+1}(s)$ and $\alpha_{m+2}(s)$ are not Hurwitz, can dynamic estimators of x be used to supply x when only y is available, how are constraints on control effort imposed, what happens when (1.2) is replaced by a sampled-data version, what form does the theory take if S is time varying, can the σ_i be chosen by solving an optimization problem for S_i, what is the effect of using nonlinear feedback on the system S_i. Some of these questions will be explored in later papers.

9. Acknowledgments. The author thanks P. L. Falb and W. A. Wolovich for the opportunity to inspect early drafts of their papers. Many interesting numerical examples of multivariable systems were supplied by B. S. Morgan. These helped develop insight which lead to the main results.

REFERENCES

[1] B. S. Morgan, Jr., *The synthesis of linear multivariable systems by state variable feedback*, Proc. 1964 JACC, Stanford, California, pp. 468–472.

[2] Z. V. Rekasius, *Decoupling of multivariable systems by means of state feedback*, Proc. Third Allerton Conference on Circuit and System Theory, Monticello, Illinois, 1965, pp. 439–448.

[3] P. L. Falb and W. A. Wolovich, *On the decoupling of multivariable systems*, Proc. 1967 JACC, Philadelphia, Pennsylvania, pp. 791–796.

[4] ———, *Decoupling in the design of multivariable control systems*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 651–659.

[5] F. R. Gantmacher, *The Theory of Matrices I*, Chelsea, New York, 1959.

[6] R. E. Kalman, *Canonical structure of linear dynamical systems*, Proc. Nat. Acad. Sci. U.S.A., 48 (1962), pp. 596–600.

[7] L. A. Zadeh and C. A. Desoer, *Linear System Theory*, McGraw-Hill, New York, 1963.

[8] B. S. Morgan, Jr., *The Synthesis of Single Variable Systems by State Variable Feedback*, Proc. First Allerton Conference on Circuit and System Theory, University of Illinois, 1963, pp. 509–520.

[9] ———, *Sensitivity analysis and synthesis of multivariable systems*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 506–512.

SIAM J. CONTROL
Vol. 7, No. 1, February 1969

CERTAIN RELATIONS BETWEEN THE BELLMAN AND KROTOV FUNCTIONS FOR DYNAMIC PROGRAMMING PROBLEMS*

I. V. GIRSANOV†

Abstract. Every Krotov function in some dynamic programming problem is identical with a Bellman function in another problem with the same conditions and with a return function of a less "sharp" optimum. The Bellman function is the envelope of a set of Krotov functions.

Let $\{v_t\} = \{y_t, u_t\}$, $0 \leqq t \leqq T$, be a controllable sequence in which the phase coordinates y_t and the control u_t are connected by the relations

$$y_{t+1} = f(t, y_t, u_t), \qquad t = 0, \cdots, T-1, \tag{1}$$

under the constraints

$$v_t \in V_t \qquad (\text{or} \quad u_t \in U_t(y_t), \quad y_t \in Y_t). \tag{2}$$

We call the part $v_s^t = \{v_\tau, s \leqq \tau \leqq t\}$ of the sequence a segment of the admissible trajectory from s to t, and when $s = 0$ and $t = T$, simply the admissible trajectory. We denote the class of all admissible trajectories by V_0^T. Let

$$I^\circ(v_s^t) = \sum_s^t f^\circ(\sigma, y_\sigma, u_\sigma) = \sum_s^t f^\circ(\sigma, u_\sigma).$$

We call an admissible trajectory $\bar{v}_0^T$ *I°-optimal* if

$$I^\circ(\bar{v}_0^T) \leqq I^\circ(v_0^T)$$

for any admissible trajectory. The phase coordinate y is called *attainable at the instant* t if $y_t = y$ for some $v_0^T \in V_0^T$. We denote the set of all points attainable at the instant t by the symbol Y^t.

The Bellman equation yields necessary and sufficient conditions for the optimality of $\bar{v}_0^T$. To formulate these conditions we define the extended set of admissible trajectories, retaining all of requirements (1) and (2) except $y_0 \in Y_0$. Correspondingly, we introduce the concept of extended attainable sets $\tilde{Y}^t$. For $y \in \tilde{Y}^t$ we set

$$B(t, y) = \min_{y_t = y} I^\circ(v_t^T), \qquad B(T+1, y) = 0.$$

* Originally published in Vestnik Moskovskogo Universiteta, Seriya I: Matematika, Mekhanika, 1968, no. 2, pp. 56–59. Submitted on November 22, 1966. This translation into English has been prepared by N. H. Choksy.

Translated and printed for this Journal under a grant-in-aid by the National Science Foundation.

† Department of Probability Theory and Mathematical Statistics, Moscow State University, Moscow. (Deceased.)

Then

$$\max_{u\in U_t(y)} [B(t, y) - B(t + 1, f(t, y, u)) - f^{\circ}(t, y, u)] = 0 \tag{3}$$

and every admissible trajectory v_0^T, connected with the function $B(t, y)$ by (3), is optimal.

We note right away that every segment v_t^T of an admissible trajectory, which satisfies the Bellman equation on the time interval $[t, T]$, minimizes $I^{\circ}(v_t^T)$. Thus, we have obtained the optimal trajectories for an entire class of minimum problems.

To verify the optimality of an individual trajectory $\bar{v}_0^T$, we can replace (3) by a less stringent system of relations.

THEOREM 1. *Let $\bar{v}_0^T$ be an admissible trajectory, let*

$$I^{\alpha}(v_s^t) = \sum_s^t f^{\alpha}(\sigma, v_{\sigma})$$

and let the function $f^{\alpha}(\sigma, v)$ be such that, for all admissible v_0^T,

$$I^{\alpha}(v_0^T) - I^{\alpha}(\bar{v}_0^T) \leqq I^{\circ}(v_0^T) - I^{\circ}(\bar{v}_0^T). \tag{4}$$

Let $\bar{v}_0^T$ satisfy the Bellman equation with the functions f^{α} instead of f°. Then, $\bar{v}_0^T$ is I°-optimal.

The proof of the theorem follows in an obvious manner from the fact that $\bar{v}_0^T$, as a solution of the Bellman equation, minimizes $I^{\alpha}(v_0^T)$, and that by virtue of (4) there holds the relation

$$I^{\circ}(v_0^T) - I^{\circ}(\bar{v}_0^T) \geqq I^{\alpha}(v_0^T) - I^{\alpha}(\bar{v}_0^T) \geqq 0.$$

The restriction (4) is most essential. It is satisfied, for example, for those f^{α} for which

$$f^{\alpha}(t, y, u) - f^{\alpha}(t, \bar{y}_t, \bar{u}_t) \leqq f^{\circ}(t, y, u) - f^{\circ}(t, \bar{y}_t, \bar{u}_t) \tag{5}$$

on V_t.

It turns out that in this case the function $\phi(t, y) = -B^{\alpha}(t, y)$ satisfies the sufficient optimality condition introduced by V. F. Krotov [1], [2]. He has proved that for the optimality of the trajectory $\bar{v}_t = (\bar{y}_t, \bar{u}_t)$ it is sufficient that when $0 \leqq t \leqq T - 1$, for some $\phi(t, y)$,

$$\begin{aligned} \phi(t + 1, \bar{y}_{t+1}) - \phi(t, \bar{y}_t) - f^{\circ}(t, \bar{y}_t, \bar{u}_t) &= \max_{y,u\in V_t} [\phi(t + 1, f(t, y, u)) \\ &\qquad - \phi(t, y) - f^{\circ}(t, y, u)], \\ \phi(T, \bar{y}_T) + f^{\circ}(T, \bar{y}_T, \bar{u}_T) &= \min_{y,u\in V_T} \phi(T, y) + f^{\circ}(T, y, u). \end{aligned} \tag{6}$$

By virtue of (3), on V_t

$$0 \geqq \phi(t + 1, f(t, y_t, u_t)) - \phi(t, y_t) - f^{\alpha}(t, y_t, u_t),$$

$$0 = \phi(t + 1, f(t, y_t, u_t)) - \phi(t, \bar{y}_t) - f^{\alpha}(t, \bar{y}_t, \bar{u}_t),$$

whence, taking (5) into account, we get

$$\phi(t+1, f(t, \bar{y}_t, \bar{u}_t)) - \phi(t, \bar{y}_t) - f^\circ(t, \bar{y}_t, \bar{u}_t)$$
$$\geqq \phi(t+1, f(t, y_t, u_t)) - \phi(t, y_t) - f^\circ(t, y_t, u_t),$$

which, together with $\phi(T+1, y) \equiv 0$, are equivalent to Krotov's conditions [1].

It is easy to show that the converse is also true, i.e., that the Krotov function $\phi(t, y)$ of the original problem can be obtained as the Bellman function of the problem of minimizing $I^\alpha(v_0^T)$, where

$$f^\alpha(t, y, u) = f^\circ(t, y, u) + \sup_{u \in V_t(u)} [\phi(t+1, f(t, y, u)) - \phi(t, y) - f^\circ(t, y, u)].$$

To establish a more precise relation we note that by adding a certain function $\phi_\alpha(t)$ to $f^\alpha(t, v)$ we can make the left-hand side of (6) vanish and consider that

$$\phi(T+1, y) \equiv 0. \tag{7}$$

The transformed function ϕ will be called the *normalized Krotov function.*

THEOREM 2. *The normalized Krotov function* $\phi(t, y)$ *coincides* (*up to sign*) *with the Bellman function* $-B^\circ(t, y)$ *at all points of the optimal trajectory* $\bar{v}_0^T$ *and satisfies on* Y^t *the relations*

$$\phi(t+1, \bar{y}_{t+1}) - \phi(t, \bar{y}_t) - f^\circ(t, \bar{y}_t, \bar{u}_t) = 0, \tag{8}$$

$$\max_{u \in U_t(y_t)} \phi(t+1, y_{t+1}) - \phi(t, y_t) - f^\circ(t, y_t, u_t) \leqq 0, \tag{9}$$

$$B^\circ(t, y) \geqq -\phi(t, y). \tag{10}$$

Proof. Relations (8) and (9) follow from the definition of the normalized Krotov function, and from (7) and (8) it ensues that

$$B^\circ(t, y_t) = -\phi(t, \bar{y}_t) \quad \text{and} \quad f^\alpha(t, \bar{y}_t, \bar{u}_t) = f^\circ(t, \bar{y}_t, \bar{u}_t).$$

Hence it is easy to derive (6) from (5).

Remark 1. The class of normalized and ordinary Krotov functions form a convex closed set. Furthermore, the lower bound of any set of Krotov functions is once again a Krotov function.

Remark 2. Relations (8) and (9) show that the class of normalized Krotov functions relative to the original problem is given by a system of inequalities which become equalities along the extremals. The Bellman function with a minus sign is in the lower envelope of this class, which suggests the possibility of finding Krotov functions with good differential properties even when the corresponding Bellman function is poorly behaved (has breaks, discontinuities). Using standard techniques, we find it is not difficult to replace the optimal trajectory v_0^T by a minimizing sequence $v_0^{(n)T}$.

The results presented above can be generalized in a natural way to the case of continuous time for the systems of the form

$$\frac{dy_t}{dt} = f(t, y_t, u_t)$$

and for the functional

$$\phi_0(y_0) + \int_0^T f^\circ(t, y_t, u_t)\,dt + \phi_T(y_T).$$

An analysis of Krotov's sufficient conditions [2] shows that the system of differential inequalities derived from them is a weakening of the Bellman equation for the original problem. Its solution, after certain normalizations, coincides with the Bellman function of the auxiliary problem.

REFERENCES

[1] V. F. Krotov, *Sufficient conditions for the optimality of discrete control systems*, Soviet Math. Dokl., 8 (1967), pp. 11–15.

[2] ———, *Methods for solving variational problems on the basis of sufficient conditions for an absolute minimum. I*, Automat. Remote Control, 23 (1962), pp. 1473–1484.

SIAM J. Control
Vol. 7, No. 1, Febraury 1969

APPLICATION OF A RESOLVENT IDENTITY TO A LINEAR SMOOTHING PROBLEM*

THOMAS KAILATH†

Abstract. By using a result of Siegert [3], we derive a new identity for the "resolvent" of a covariance function. This identity is used to obtain a simple relation between the "smoothed" and "filtered" linear least-squares estimates of a signal process in additive uncorrelated white noise.

1. A resolvent identity. Let $R(t, s)$, $t, s \in I \otimes I$, be a symmetric, continuous function on $I \otimes I$, where I is a finite interval on the real line, say $I = [0, T]$. Then the resolvent of $R(t, s)$ is defined (see, e.g., Riesz–Nagy [1] and Smithies [2]) as the function $H(t, s; u, T)$ that is the solution of the integral equation

$$H(t, s; u, T) + u \int_0^T R(t, r)H(r, s; u, T)\,dr = R(t, s), \qquad 0 \leqq t, s \leqq T, \tag{1}$$

where u is a complex number. The solution $H(t, s; u, T)$ will clearly be symmetric and continuous in (t, s) over $I \otimes I$, and it will be unique [1], [2] whenever $-u^{-1}$ is not an eigenvalue of $R(t, s)$.

The resolvent plays an important role in the Fredholm theory of integral equations and consequently also in several applied problems. In one such application, Siegert [3] derived and exploited some special formulas involving the resolvent. In the present one, we shall use one of Siegert's identities to obtain a new identity which we shall apply to a (so-called smoothing) problem in linear least-squares estimation.

The identity of Siegert[1] is [3, Equation (59)]

$$\begin{aligned}\frac{\partial H}{\partial T}(t, s; u, T) &= -uH(t, T; u, T)H(s, T; u, T)\\ &= -uH(T, t; u, T)H(T, s; u, T).\end{aligned} \tag{2}$$

We shall use it to show that if $0 \leqq t < s \leqq T$ or $0 \leqq s < t \leqq T$, then

$$H(t, s; u, T) = H(t, s; u, t) + H(s, t; u, s) - u \int_0^T H(r, t; u, r)H(r, s; u, r)\,dr, \tag{3}$$

* Received by the editors June 18, 1968, and in revised form September 6, 1968.

† Stanford Electronics Laboratories, Stanford University, Stanford, California 94305. This work was supported by the Applied Mathematics Division of the Air Force Office of Scientific Research under Contract AF 49(638)1517 and by the Joint Services Electronics Program at Stanford University. Some of this work was also done as a consultant to the Jet Propulsion Laboratories, Pasadena, California.

[1] This identity was also derived by Bellman [4], also by M. G. Krein [10], for symmetric *and* nonsymmetric $R(t, s)$.

where it is assumed that

$$H(t, s; u, r) = 0 \quad \text{for} \quad t \text{ or } s > r.$$

The operator form of (3) is often convenient. Let H_u denote the (integral) operator on $I \otimes I$ with kernel $H(t, s; u, T)$, let h_u denote the (integral) operator with kernel $H(t, s; u, t)$ and let h_u^* denote the (integral) operator with kernel $H(s, t; u, t)$. Then (because $H(t, s; u, t) = 0$ for $s > t$), h_u and h_u^* are Volterra operators that are also adjoints of each other. In this operator notation we can rewrite (3) as

$$\text{(3a)} \qquad H_u = h_u + h_u^* - uh_u^* h_u.$$

We now give the simple proof of (3) (and also (2)), but these can be omitted without loss of continuity.

Proof of (3). To prove (3) from (2), we shall define

$$N(t, s; u, T) \triangleq H(t, s; u, t) + H(s, t; u, s)$$
$$- u \int_0^T H(r, t; u, r) H(r, s; u, r)\, dr - H(t, s; u, T)$$

and show that $N(t, s; u, T) \equiv 0$. First note that

$$\frac{\partial N}{\partial T}(t, s; u, T) = -uH(T, t; u, T)H(T, s; u, T) - \frac{\partial}{\partial T}H(t, s; u, T)$$
$$= 0 \quad \text{by (3)}.$$

Therefore N has the same value for all T. But for $t < s = T$, we see that

$$N(t, s; u, T) = 0 + H(T, t; u, T) - 0 - H(t, T; u, T)$$
$$= 0,$$

by the symmetry in (t, s) of $H(t, s : u, T)$. Therefore $N(t, s; u, T)$ must be identically zero for all $0 \leqq t < s \leqq T$, and by a similar arrangement, for all $0 \leqq s < t \leqq T$. This establishes (3).

Proof of (2). For ease of reference, we include a proof of (2). First we differentiate the defining relation (1) with respect to T to get

$$\frac{\partial H}{\partial T}(t, s; u, T) = -uR(t, T)H(T, s; u, T) + u\int_0^T R(t, r)\frac{\partial H}{\partial T}(r, s; u, T)\, dr.$$

Suppose now that

$$\frac{\partial H}{\partial T}(t, s; u, T) = -uH(t, T, u, T)H(s, T; u, T) + g(t, s; u, T).$$

Substituting this expression into the previous equation and using (1) again we get

$$g(t, s; u, T) + u\int R(t, r)g(r, s; u, T)\, dr = 0,$$

which is similar to (1) with the right-hand side set equal to zero. Therefore, whenever (1) has a unique solution (i.e., $-u^{-1}$ is not an eigenvalue of $R(t, s)$) we must have

$$g(t, s; u, T) \equiv 0,$$

which proves that $\partial H/\partial T$ must be as in (2).

2. Application to the smoothing problem. We shall apply the identity (3) to the following problem:

We are given observations

$$y(t) = z(t) + v(t), \qquad t \in I = [0, T],$$

where $z(t)$ is a signal process with (E denotes expectation)

$$E(z(t)) = 0, \qquad E(z(t)z(s)) = R(t, s)$$

(a square-integrable function on $I \otimes I$) and $v(t)$ is a white noise process with

$$E(v(t)) = 0, \qquad E(v(t)v(s)) = \delta(t - s)$$

and such that

$$E(v(t)z(s)) \equiv 0, \qquad t, s \in I \otimes I. \tag{4}$$

The problem is to find an estimate $\hat{z}(t|T)$ of $z(t)$ that satisfies

$$E([z(t) - \hat{z}(t|T)]^2) = \text{minimum}$$

and

$$\hat{z}(t|T) = \text{ a linear functional of } \{y(\tau), 0 \leqq \tau \leqq T\}.$$

The quantity $\hat{z}(t|T)$ is often called a smoothed (or noncausal) estimate of $z(t)$ in contrast to the filtered (or causal) estimate $\hat{z}(t|t)$ which, as implied by the notation, is defined as

$$\begin{aligned}\hat{z}(t|t) &= \hat{z}(t|T), \quad T = t,\\ &= \text{the linear functional of } \{y(\tau), 0 \leqq \tau \leqq T\}\\ & \text{ that minimizes } E[z(t) - \hat{z}(t|t)]^2).\end{aligned}$$

The filtering problem has been widely studied and solutions have been obtained in various forms. Recursive methods of calculating $\hat{z}(t|t)$ by Kalman–Bucy filters have been given special attention. However, recursive solutions of the smoothing problem have generally been considered to be harder to obtain. In this paper we shall show that there is a simple relation between the smoothed and filtered estimates, viz.,

$$\hat{z}(t|T) = \hat{z}(t|t) + \int_t^T h^*(t, s)(y(s) - \hat{z}(s|s))\, ds, \tag{5}$$

where $h^*(t, s) =$ the impulse response of the *adjoint* of the optimum causal filter.

If

$$\hat{z}(t|t) = \int_a^t h(t,s)y(s)\,ds, \tag{6}$$

then

$$h^*(t,s) = h(s,t).$$

Since $h(s,t)$ is zero for $t > s$, $h^*(t,s)$ is zero for $s < t$ (which is consistent with the limits of integration in (5)).

The relation (5) shows that the solution to the smoothing problem is completely determined by the solution to the filtering problem. If a recursive solution is available for the filtering problem, then (5) immediately yields one for the smoothing problem. In this way (cf. [5], [6]) we have easily derived all previous smoothing formulas.

It is also worth noting that formula (5) for $\hat{z}(t|b)$ is in a form that can readily accommodate an increase in the interval of observation: if $c > b$,

$$\hat{z}(t|c) = \hat{z}(t|b) + \int_b^c h^*(t,s)(y(s) - \hat{x}(s|s))\,ds.$$

Proof of (5). We shall first show that

$$\hat{z}(t|T) = \int_0^T H(t,r;1,T)y(r)\,dr \tag{7}$$

and consequently that

$$\hat{z}(t|t) = \int_0^t H(t,s;1,t)y(s)\,ds, \tag{8}$$

where $H(t,s;1,T)$ is the $+1$-resolvent of $R(t,s)$, the covariance function of z. (We note that since $R(t,s)$ is a covariance function, $-u^{-1} = -1$ is not an eigenvalue of $R(t,s)$ and therefore the resolvent $H(t,s;1,T)$ is uniquely defined.) Then (5) will be a consequence of the relation

$$H(t,s;1,T) = H(t,s;1,t) + H(s,t;1,s) - \int_0^T H(r,t;1,r)H(r,s;1,r)\,dr \tag{9}$$

and the definition (6) of an adjoint filter. The relation (9) follows from the resolvent identity (3) with $u = 1$.

To prove (7), the projection theorem for linear least-squares estimates gives

$$E(z(t)y(s)) \equiv E(\hat{z}(t|T)y(s)), \qquad 0 \leqq s \leqq T. \tag{10}$$

But by plugging in (7) and using (4), we obtain

$$R(t,s) = H(t,s;1,T) + 1 \cdot \int_0^T H(t,r;1,T)R(r,s)\,dr, \qquad 0 \leqq t,s \leqq T. \tag{11}$$

But this is just (1) with $u = 1$. Therefore, (10) is satisfied and (7) is proven.

Some previous work. Formula (5) can, with certain additional arguments, be inferred from some work by Schweppe [7] on a different problem. For the special case of "lumped" processes $z(t)$, i.e., processes obtained by passing white noise through a (possibly time-variant) lumped linear dynamical system, formula (5) was obtained, in a different and more laborious way, by Kwakernaak [8] and Fraser [9]. We might remark that (recursive) smoothing solutions have generally been considered more difficult to obtain than (recursive) filtering solutions (cf. the reference and discussion in [8]–[9]). Finally, we note that (5) has been derived in a different way, by use of the concept of an "innovation" process, in Kailath and Frost [5], [6], where further discussion of the smoothing problem, including formulas for the mean-square error, various differential equations for $\hat{z}(t|T)$, some generalizations, and additional references, can be found. The innovations method [6] shows that the assumption (4) of $E(n(t)z(s)) \equiv 0$ can be relaxed to $E(n(t)z(s)) = 0$ for $s < t$, provided $h^*(t, s)$ in (5) is not taken, as in (6), as the adjoint of the causal filter, but is defined by $h^*(t, s) = E(\tilde{z}(t|t)\tilde{z}(s|s))$, the covariance function of the causal error $\tilde{z}(t|t) = z(t) - \hat{z}(t|t)$. In other words, we have

$$\hat{z}(t|T) = \hat{z}(t|t) + \int_t^T E(\tilde{z}(t|t)\tilde{z}(s|s))\, \nu(s)\, ds, \tag{12}$$

where

$$\nu(\cdot) = y(\cdot) - \hat{z}(\cdot|\cdot) \tag{13}$$

is known [5] as the "innovation process" of $y(\cdot)$.

3. Concluding remarks. The author obtained the special case (9) of the resolvent identity (2). Dr. L. A. Shepp of the Bell Telephone Laboratories, Murray Hill, New Jersey, provided the version (2) and gave the proof appearing in § 1, which was slightly different from the author's original proof.[2] It may be of interest here to briefly sketch how the author was led to first conjecture (not prove) relation (9).

Consider the equation (in the notation of § 2)

$$y(t) = z(t) + v(t) \tag{14}$$

$$= \hat{z}(t|t) + \tilde{z}(t|t) + v(t) = \hat{z}(t|t) + \nu(t), \tag{15}$$

say. Then it can be shown [5] that if $z(\cdot)$ and $v(\cdot)$ are normal, then so is $\nu(\cdot)$, and that in fact $\nu(\cdot)$ has the same statistics as $v(\cdot)$, i.e.,

$$E(\nu(t)\nu(s)) = E(v(t)v(s)).$$

Moreover since $\hat{z}(t|t)$ is a functional part of $\{y(s), s < t\}$, it can be regarded as (conditionally) known, given $\{y(s), s < t\}$. These two facts suggest that the process $y(\cdot)$ can be regarded in two ways, corresponding to (14) and (15). According to (14), $y(\cdot)$ is zero-mean Gaussian with covariance function $R(t, s) + \delta(t - s)$;

[2] However, it is easy to obtain (2) from (9): replace h by uh_u, H by uH_u and R by uR. Then the resolvent equation (11) goes into (1) and the identity (9) goes into (3).

according to (15), $y(\cdot)$ is Gaussian with mean-value function $\hat{z}(t|t)$ and covariance $\delta(t - s)$, Therefore (of course, very heuristically), in an obvious operator notation (with $\hat{z} = hy$) we can write the "density" of $y(\cdot)$ as

$$k \cdot \exp(-\tfrac{1}{2}\langle y, (I + R)^{-1}y\rangle = k \cdot \exp(-\tfrac{1}{2}(y - hy)'I(y - hy)),$$

and by matching the terms in y, we get

$$(I + R)^{-1} = (I - h)'(I - h). \tag{16}$$

But then if H is the smoothing filter, it is given (cf. (11)) by

$$H = R(I + R)^{-1} = I - (I + R)^{-1} = h + h'(I - h), \tag{17}$$

which is the relation (9). It might be of some interest to obtain a rigorous version of this heuristic argument.

The discrete-parameter case. When $t = 0, 1, 2, \cdots, N$, we define

$$\nu(k) = y(k) - \hat{z}(k|k - 1); \tag{18}$$

and it turns out [5] that $\nu(\cdot)$ is normal but (unlike the continuous-time case) now it has a variance different from that of $v(\cdot)$, viz.,

$$E(\nu(k)\nu(l)) = [E(v^2(k)) + E(\tilde{z}^2(k|k - 1))]\,\delta_{kl}. \tag{19}$$

Then the analogue of (16) will be somewhat more complicated and we do not obtain the simple relation (17) between the smoothing filter H and the causal filter h. However, we may note that, as in the continuous-time case, the innovations method [5]–[6] yields a simple formula

$$\hat{z}(k|N) = \hat{z}(k|k) + \sum_{k+1}^{N} \frac{E(\tilde{z}(k|k - 1)\tilde{z}(l|l - 1))}{E(\nu^2(l))}\nu(l). \tag{20}$$

Finally we might mention that all our results can be readily extended to the vector case. We have also recently found that the basic identity (3) can be used to obtain some sufficient conditions for the so-called "covariance-factorization" problem. These results will be described elsewhere.

REFERENCES

[1] F. Riesz and B-Sz. Nagy, *Functional Analysis*, Frederick Ungar, New York, 1955.

[2] F. Smithies, *Integral Equations*, Cambridge University Press, New York, 1962.

[3] A. J. F. Siegert, *A systematic approach to a class of problems in the theory of noise—Part II*, Trans. IRE Information Theory, IT-3 (1957), pp. 38–43.

[4] R. E. Bellman, *Functional equations in the theory of dynamic programming—VII: A partial differential equation for the Fredholm resolvent*, Proc. Amer. Math. Soc., 8 (1957), pp. 435–440.

[5] T. Kailath, *An innovations approach to least-squares estimation—Part I: Linear filtering in additive white noise*, IEEE Trans. Automatic Control, AC-13 (1968), to appear.

[6] T. Kailath and P. Frost, *An innovations approach to least-squares estimation—Part II: Linear smoothing in additive white noise*, Ibid., AC-13 (1968), to appear.

[7] F. Schweppe, *Evaluation of likelihood functions for Gaussian signals*, Trans. IEEE Information Theory, IT-11 (1965), pp. 61–70.

[8] H. Kwakernaak, *Optimal filtering in linear systems with time delays*, Trans. IEEE Automatic Control, AC-12 (1967), pp. 169–173.

[9] D. C. Fraser, *A new technique for the optimal smoothing of data*, Sc.D. thesis, Department of Aeronautical Engineering, Massachusetts Institute of Technology, Cambridge, 1967.

[10] M. G. Krein, *On integral equations governing differential equations of second order*, Dokl. Akad. Nauk SSSR, 97 (1954), pp. 21–24.

SIAM J. Control
Vol. 7, No. 1, February 1969

OPTIMAL REGULATION OF NONLINEAR DYNAMICAL SYSTEMS*

D. L. LUKES†

Introduction. This paper deals with the optimal control of autonomous systems of nonlinear differential equations. The control functions represent feedback devices which operate upon the instantaneous state of the system to generate control signals which automatically return the system to a prescribed state of equilibrium whenever an impulsive disturbance occurs in the state. These regulator devices are widely used in aircraft flight controls and many other control systems.

We define optimal feedback control in terms of a performance integral. The main result is Theorem 1.1 in which we prove the existence and uniqueness of an optimal feedback control and show that it generates control signals which are optimal in a certain local open-loop sense. The basic hypothesis used for the development is that the system be stabilizable.

Theorem 1.2 is quite well known but included to show that for linear systems the optimal feedback control coincides with the synthesis construction arising from the open-loop theory. This latter topic was studied by Kalman [5] in 1960 and has recently been re-examined by the author [8] and others. In [8] may be found a proof of the fact that the stabilizability of the system is equivalent to the solvability of the Kalman–Riccati matrix equation of Lemma 2.3.

One of the first serious attempts to treat nonlinear systems was made by Al'brekht [1], [2] who studied analytic systems around 1961–1963 and discovered the optimal control as a formal power series by considering Lyapunov functions. He presented some rather complicated dominated convergence arguments for the convergence of the series for a scalar control variable. A morê extensive treatment of the problem under the assumption of the complete controllability of the system was made by Brunovsky and may be found in [3] (without proofs).

The technique we use covers the analytic case and includes completely controllable systems. Any finite number of control variables are allowed and the analyticity assumption is relaxed to twice continuous differentiability. An outline of the proof has been included in [7]. The proof is carried out along the line of the Hamilton–Jacobi formalism [6] but carries the usual treatment further by actually proving that the required equations have solutions.

* Received by the editors April 3, 1968, and in revised form September 27, 1968.

† Mathematics Research Center, United States Army, University of Wisconsin, Madison, Wisconsin 53706. This work was sponsored by the Mathematics Research Center under Contract DA-31-124-ARO-D-462.

Notation. We study differential equations in finite n-dimensional real and complex number spaces R^n and C^n, respectively, using the inner product and norm notations $x \cdot y = \sum_{k=1}^{n} x_k y_k$ and $|x| = \sqrt{x \cdot \bar{x}}$ for $x = (x_1, x_2, \cdots, x_n)$ and $y = (y_1, y_2, \cdots, y_n)$ in R^n or C^n. The transpose of a real matrix M is denoted by M^* and we use the matrix norm $\|M\| = \sup_{|x|=1} |Mx|$. The notation $M > 0$ denotes that M is a real symmetric positive definite matrix.

1. Statement of the optimization problem and principal results. The problem is formulated in terms of a control system equation in R^n,

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

and a performance integral

$$J = \int_0^\infty G(x, u)\, dt.$$

Roughly speaking, we seek an r-dimensional vector feedback control function of the state x, $u = u(x)$, which makes the integral as small as possible for all initial states near the origin in R^n. This differs from the problem of optimal open-loop control in which the initial state is fixed and the integral is minimized over $u = u(t)$ in $L_2(0, \infty)$.

Hence, for each feedback control function $u = u(x)$ we consider the autonomous differential equation

$$\dot{x} = F(x, u(x)) \tag{1.1}$$

with the corresponding solution $x = x(t, x_0)$, where $x(0, x_0) = x_0$ for all initial states x_0 near the origin in R^n. Since we are interested in the dependence of the integral upon the initial state of (1.1) as well as upon the control function $u(x)$, we use the notation

$$J(x_0, u) = \int_0^\infty G(x(t, x_0), u(x(t, x_0)))\, dt. \tag{1.2}$$

1.1. The basic assumptions. Throughout the paper we assume that $F(x, u)$ and $G(x, u)$ are defined on some neighborhood of the origin in R^{n+r} and can be represented in the form

$$F(x, u) = Ax + Bu + f(x, u), \tag{1.3}$$

$$G(x, u) = x \cdot \mathfrak{U}x + 2x \cdot \mathfrak{C}u + u \cdot \mathfrak{B}u + g(x, u), \tag{1.4}$$

where A, B, $\mathfrak{U}$, $\mathfrak{B}$ and $\mathfrak{C}$ are real matrices and $f(x, u)$ and $g(x, u)$ are higher order terms to be discussed below. In (1.4) we assume $\begin{pmatrix} \mathfrak{U} & \mathfrak{C} \\ \mathfrak{C}^* & \mathfrak{B} \end{pmatrix} > 0$.

A real matrix is called a *stability matrix* if its eigenvalues all have negative real parts. In control theory we call the pair of matrices A, B in (1.3) and the control system defined by $F(x, u)$ *stabilizable* if there exists a real matrix D for which

$A + BD$ is a stability matrix. Our fundamental hypothesis in this paper is that $F(x, u)$ is stabilizable. For an extensive discussion of the stabilizability condition on A, B the reader is referred to [7] and [8].

We consider the class of feedback controls which are of the form

$$u = u(x) = Dx + h(x), \tag{1.5}$$

where $h(x)$ denotes higher order terms to be discussed shortly. The real matrices D are always selected so that $u(x)$ *stabilizes* (1.1); that is, we demand that in

$$F(x, u(x)) = (A + BD)x + Bh(x) + f(x, u(x)),$$

$A + BD$ should be a stability matrix.

We are interested in studying the optimization problem for two different sets of conditions on the higher order terms which we now state.

The analytic case. When $F(x, u)$ and $G(x, u)$ are real analytic about the origin in R^{n+r}, we understand the terms $f(x, u)$ and $g(x, u)$ in (1.3)–(1.4) to be real convergent power series about the origin beginning with second and third order terms in (x, u), respectively. In this situation we admit every $h(x)$ in (1.5) given by real power series converging about the origin and beginning with second order terms. The feedback controls are called C^ω *stabilizing controls.*

The differentiable case. We are also interested in studying the problem for which $F(x, u)$ and $G(x, u)$ are twice continuously differentiable about the origin in R^{n+r} and in which we admit higher order terms $h(x)$ in (1.5) which are at least once continuously differentiable about the origin. We assume:

$$\text{(a)} \qquad f(0, 0) = 0,$$

$$\text{(b)} \qquad \left\| \frac{\partial f(x, u)}{\partial(x, u)} \right\| \leqq C|(x, u)|^\alpha,$$

$$\text{(c)} \qquad g(0, 0) = 0,$$

$$\text{(d)} \qquad \left\| \frac{\partial^2 g(x, u)}{\partial(x, u)^2} \right\| \leqq C|(x, u)|^\alpha$$

near the origin in R^{n+r} for positive numbers α and C. The restrictions on the nonlinear feedbacks are

$$\text{(e)} \qquad h(0) = 0,$$

$$\text{(f)} \qquad \|h_x(x)\| \leqq L|x|^\beta$$

near the origin for some positive numbers β and L (depending upon h). The feedback controls given by (1.5) and satisfying the above specified conditions are called C^1 *stabilizing controls* and the control process is called a C^2 *process.*

1.2. Definition of optimal feedback control. Since we consider only stabilizing controls $u(x) = Dx + h(x)$ in (1.1), it follows that $|x(t, x_0)|$ and $|u(x(t, x_0))|$ decay exponentially toward zero for $|x_0|$ suitably small. In fact if the characteristic values

λ of $A + BD$ have negative real parts all less than a constant $-\mu$,

$$\operatorname{Re} \lambda(A + BD) < -\mu < 0;$$

then

$$|x(t, x_0)| \leqq C_1 e^{-\mu t} |x_0|$$

for $0 \leqq t < \infty$ and $|x_0|$ small, where C_1 is a positive number. Moreover,

$$|u(x(t, x_0))| \leqq C_2 |x(t, x_0)| \leqq C_1 C_2 e^{-\mu t} |x_0|$$

for $|x_0|$ small and $0 \leqq t < \infty$ for some positive number C_2. These basic estimates show that for each stabilizing control $u(x)$ the performance integral is uniformly convergent to a finite value $J(x_0, u)$ near the origin. In fact we shall show that in the analytic case $J(x_0, u)$ is real analytic in x_0 near the origin and in the differentiable case it is once continuously differentiable.

Note that one of our basic assumptions is that the Hessian matrix of $G(x, u)$ at the origin $\begin{pmatrix} \mathfrak{U} & \mathfrak{C} \\ \mathfrak{C}^* & \mathfrak{B} \end{pmatrix}$ is positive definite. Hence $J(x_0, u) > 0$ near the origin and the integral is a composite indicator of the rate at which the feedbacks return the disturbed process to its equilibrium state and the control energy expended during the operation. This is the motivation for making the following technical definition.

DEFINITION. A C^ω (C^1) stabilizing feedback control $u_*(x) = D_* x + h_*(x)$ is called *optimal* for the process (1.1) with respect to the performance integral (1.2) if for every C^ω (C^1) stabilizing feedback control $u(x) = Dx + h(x)$ there exists a neighborhood N_u of the origin in R^n in which

$$J(x_0, u_*) \leqq J(x_0, u).$$

In order to assert the uniqueness of the optimal control we shall agree to consider two controls to be the same (or equivalent) in case they coincide on some neighborhood of the origin in R^n. We can now state the main theorem of the paper.

THEOREM 1.1 (Main theorem). *For the C^ω (C^2) stabilizable control process in R^n*

$$\dot{x} = F(x, u) = Ax + Bu + f(x, u)$$

with performance integral

$$J(x_0, u) = \int_0^\infty G(x, u)\, dt = \int_0^\infty \left[\begin{pmatrix} x \\ u \end{pmatrix} \cdot \begin{pmatrix} \mathfrak{U} & \mathfrak{C} \\ \mathfrak{C}^* & \mathfrak{B} \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix} + g(x, u) \right] dt,$$

there exists an optimal C^ω (C^1) stabilizing feedback control u_*. The optimal control solves the functional equation

$$(\mathfrak{F}) \qquad F_u(x_0, u_*(x_0)) J_x(x_0, u_*) + G_u(x_0, u_*(x_0)) = 0$$

for all x_0 near the origin and is unique in that:

(i) *u_* is the unique C^ω (C^1) solution to $(\mathfrak{F})$;*

(ii) *u_* is the unique C^ω (C^1) stabilizing feedback control;*

(iii) *u_* synthesizes*[1] *the unique optimal open-loop control.*

Furthermore, $u_*(x) = D_*x + h_*(x)$ *and* $J(x_0, u_*) = x_0 \cdot P_*x_0 + j_*(x_0)$, *where the lowest order terms are given by matrices* D_* *and* $P_* > 0$ *depending upon only* $A, B, \mathfrak{U}, \mathfrak{B}$ *and* $\mathfrak{C}$.

THEOREM 1.2 (Truncated system). *For the special case of Theorem* 1.1 *in which*

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

and

$$J(x_0, u) = \int_0^\infty \begin{pmatrix} x \\ u \end{pmatrix} \cdot \begin{pmatrix} \mathfrak{U} & \mathfrak{C} \\ \mathfrak{C}^* & \mathfrak{B} \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix} dt,$$

the optimal control is $u_*(x) = D_*x$, *where* $D_* = -\mathfrak{B}^{-1}[\mathfrak{C}^* + B^*P_*]$. *Here* $P_* > 0$ *solves the matrix equation*

$$(\mathfrak{M}) \qquad \begin{aligned} &(\mathfrak{U} - \mathfrak{C}\mathfrak{B}^{-1}\mathfrak{C}^*) + P_*(A - B\mathfrak{B}^{-1}\mathfrak{C}^*) \\ &\qquad + (A - B\mathfrak{B}^{-1}\mathfrak{C}^*)^*P_* - P_*(B\mathfrak{B}^{-1}B^*)P_* = 0 \end{aligned}$$

and is the unique positive definite solution.

D_*x *is a global optimal control in the sense that in the definition of optimal feedback control we can take* N_u *to be the domain of* $J(x_0, u)$ *and in footnote* 1 *we may take* $\varepsilon = \infty$ *and* $N_* = R^n$. *Finally,* $J(x_0, u_*) = x_0 \cdot P_*x_0$.

2. Construction of the optimal control for the analytic case. In this section we develop a proof of Theorem 1.1 for the analytic case. The differentiable case is discussed in the next section. We extend the analysis into the complex space in order to show that the optimal control is analytic.

Since $F(x, u)$, $G(x, u)$ and $u(x) = Dx + h(x)$ are given by convergent real power series about the origin in R^{n+r}, they may be extended as complex analytic functions onto a neighborhood of the origin in $(n + r)$-dimensional complex space. Hence we shall consider (1.1) and (1.2) for complex initial conditions z_0.

LEMMA 2.1. *For each* C^ω *control* $u(x) = Dx + h(x)$ *there exists a positive invariant neighborhood* N_u^c *of the origin in complex n-space wherein the integral* $J(z_0, u) = z_0 \cdot Pz_0 + j(z_0)$ *is analytic in* z_0. *Here* $j(z_0)$ *is a power series beginning with third order terms and converging in* N_u^c. *The matrix* $P > 0$ *depends upon only the truncated problem* (*the data* A, B, $\mathfrak{U}$, $\mathfrak{B}$, $\mathfrak{C}$, D) *and is given by the formula*

$$P = \int_0^\infty e^{(A+BD)^*t}[\mathfrak{U} + \mathfrak{C}D + D^*\mathfrak{C}^* + D^*\mathfrak{B}D]e^{(A+BD)t}\,dt.$$

[1] That is, there exists an $\varepsilon > 0$ and a neighborhood of the origin N_* such that for each $x_0 \in N_*$ the response $x_*(t)$ satisfies

$$\dot{x}_* = F(x_*, u_*(x_*)), \quad x_*(0) = x_0, \quad x_*(t) \subseteq N_* \quad \text{for all} \quad 0 \leqq t < \infty,$$

and the corresponding control $u_*^{\text{op}\cdot}(t) = u_*(x_*(t))$ is the unique open-loop control achieving the minimum of $C(u) = \int_0^\infty G(x(t), u(t))\,dt$ among all measurable controls $u(t)$ on $0 \leqq t < \infty$ with $|u(t)| \leqq \varepsilon$ and generating trajectories $x(t)$ satisfying $\dot{x} = F(x, u(t))$, $x(0) = x_0$, $x(t) \subseteq N_*$ for all $0 \leqq t < \infty$.

$J(x_0, u)$ is a real power series. In N_u^c the functional equation

$$F(z, u(z)) \cdot J_z(z, u) + G(z, u(z)) \equiv 0$$

obtains.

Since $A + BD$ is a stability matrix, $\operatorname{Re} \lambda(A + BD) < -\mu < 0$ for some μ. Hence there is a neighborhood N_u^c of the origin in complex n-space wherein each solution $x(t, z_0)$ of (1.1),

$$\dot{x} = F(x, u(x)) = (A + BD)x + \cdots,$$

initiating at $z_0 \in N_u^c$ remains in N_u^c for all $t \geqq 0$ and satisfies the basic estimate

$$|x(t, z_0)| \leqq C_1 e^{-\mu t}|z_0| \quad \text{on} \quad 0 \leqq t < \infty.$$

The neighborhood N_u^c can be taken so small that $|u(z)| \leqq C_2|z|$ and so $|G(z, u(z))| \leqq C_3|z|^2$ for all $z \in N_u^c$ for positive numbers $C_1, C_2, C_3, \cdots$. The functions $x(t, z_0)$ and $u(x(t, z_0))$ are analytic in $z_0 \in N_u^c$ for each fixed $t \geqq 0$ and continuous in (t, z_0). Thus the integral

$$J(z_0, u) = \int_0^\infty G(x(t, z_0), u(x(t, z_0)))\, dt$$

is uniformly convergent in N_u^c and we conclude that $J(z_0, u)$ is analytic for $z_0 \in N_u^c$.

To compute the power series of $J(x_0, u)$ we must obtain $x(t, x_0)$ as a power series in x_0 in $N_u = N_u^c \cap R^n$. It is easy to see that

$$x(t, x_0) = e^{(A+BD)t}x_0 + (\text{higher terms})$$

and

$$u(x(t, x_0)) = De^{(A+BD)t}x_0 + (\text{higher terms}).$$

If termwise integration is valid for $G(x(t, x_0), u(x(t, x_0)))$, then it is easy to compute

$$J(x_0, u) = x_0 \cdot \left[\int_0^\infty e^{(A+BD)^*t}(\mathfrak{U} + \mathfrak{C}D + D^*\mathfrak{C}^* + D^*\mathfrak{B}D)\, e^{(A+BD)t}\, dt\right] x_0$$

$$+ (\text{cubic and higher order terms in } x_0).$$

In this case $J(x_0, u)$ has the required form. In order to justify this result we must estimate the higher order terms in $x(t, x_0)$. For this purpose we write

$$x(t, x_0) = x_0 + \int_0^t \hat{F}(x(s, x_0))\, ds$$

and

$$x_L(t, x_0) = x_0 + \int_0^t \hat{F}_x(0) x_L(s, x_0)\, ds,$$

where $\hat{F}(x) = F(x, u(x))$ and $\hat{F}_x(0) = A + BD$. Then the difference is

$$\Delta(t, x_0) = x(t, x_0) - x_L(t, x_0) = \int_0^t [\hat{F}(x(s, x_0)) - \hat{F}_x(0)x_L(s, x_0)]\, ds$$

and

$$\Delta = \int_0^t [\hat{F}_x(0)x(s, x_0) + \varepsilon(s, x_0) - \hat{F}_x(0)x_L(s, x_0)]\, ds$$

$$= \int_0^t [\hat{F}_x(0)\Delta + \varepsilon(s, x_0)]\, ds,$$

where $|\varepsilon(t, x_0)| \leqq C_4|x(t, x_0|^2 \leqq C_5\, e^{-2\mu t}|x_0|^2$. Therefore

$$\Delta(t, x_0) = \int_0^t e^{(A+BD)(t-s)}\varepsilon(s, x_0)\, ds$$

and

$$|\Delta(t, x_0)| \leqq C_6\, e^{-\mu t}|x_0|^2 \quad \text{for} \quad x_0 \in N_u, \quad t \geqq 0.$$

This yields the desired estimate

$$x(t, x_0) = e^{(A+BD)t}x_0 + \Delta(t, x_0).$$

Now we note

$$G(x, u(x)) = x \cdot \mathfrak{U}x + 2x \cdot \mathfrak{C}u + u \cdot \mathfrak{B}u + \gamma(x)$$

with $|\gamma(x)| \leqq C_7|x|^3$. Thus by taking N_u^c small, we have

$$G(x(t, x_0), u(x(t, x_0))) = x \cdot \mathfrak{U}x + 2x \cdot \mathfrak{C}u + u \cdot \mathfrak{B}u + \gamma(x(t, x_0))$$

and

$$\int_0^\infty |\gamma(x(t, x_0))|\, dt \leqq C_8 \int_0^\infty e^{-3\mu t}|x_0|^3\, dt = C_9|x_0|^3.$$

Hence the power series for $J(x_0, u)$ collects the linear and quadratic terms in x_0 from the expression

$$\int_0^\infty [(e^{(A+BD)t}x_0 + \Delta) \cdot \mathfrak{U}(e^{(A+BD)t}x_0 + \Delta)$$
$$+ 2(e^{(A+BD)t}x_0 + \Delta) \cdot \mathfrak{C}u(x) + u(x) \cdot \mathfrak{B}u(x)]\, dt.$$

But

$$u(x(t, x_0)) = D[e^{(A+BD)t}x_0 + \Delta] + \Delta_1(t, x_0)$$

with $|\Delta_1(t, x_0)| \leqq C_{10}|x(t, x_0)|^2 \leqq C_{11}e^{-2\mu t}|x_0|^2$. Thus the quadratic terms of $J(x_0, u)$ are just

$$J^{(2)}(x_0) = x_0 \cdot Px_0,$$

hence $J(x_0, u) = x_0 \cdot Px_0 +$ (higher terms in x_0) as required.

Clearly $P > 0$ since $\mathfrak{U} + \mathfrak{C}D + D^*\mathfrak{C}^* + D^*\mathfrak{B}D > 0$, which follows directly from

$$\begin{pmatrix} \mathfrak{U} & \mathfrak{C} \\ \mathfrak{C}^* & \mathfrak{B} \end{pmatrix} > 0.$$

Let $x(t, z_0)$ denote the solution of (1.1) in complex n-space with $x(0, z_0) = z_0$. By the uniqueness of the solutions

$$x(s, x(t, z_0)) = x(s + t, z_0)$$

for $|z_0|$ small and all $t \geqq 0$, $s \geqq 0$. Therefore

$$\begin{aligned} J(x(t, z_0), u) &= \int_0^\infty G(x(s + t, z_0), u(x(t + s, z_0)))\, ds \\ &= \int_t^\infty G(x(s, z_0), u(x(s, z_0)))\, ds. \end{aligned}$$

By the analyticity of $J(z, u)$ with respect to z near the origin in complex n-space we can differentiate the above equation and set $t = 0$ to get

$$F(z, u(z)) \cdot J_z(z, u) + G(z, u(z)) \equiv 0$$

for all $|z|$ small.

LEMMA 2.2. *In the analytic case there exists a unique analytic solution $u_*(x, p)$ to the equation*

$$F_u(x, u_*)p + G_u(x, u_*) = 0$$

near the origin in R^{2n} for which $u_*(0, 0) = 0$. Furthermore,

$$u_*(x, p) = -\tfrac{1}{2}\mathfrak{B}^{-1}(2\mathfrak{C}^*x + B^*p) + h_*(x, p),$$

where $h_*(x, p)$ is a convergent power series about $(0, 0)$ beginning with terms of second degree in (x, p).

From (1.3)–(1.4),

$$\begin{aligned} F(x, u) \cdot p + G(x, u) = [Ax + Bu + f(x, u)] \cdot p + x \cdot \mathfrak{A}x + 2x \cdot \mathfrak{C}u + u \cdot \mathfrak{B}u \\ + g(x, u), \end{aligned}$$

and we see that at the point $x = p = 0$, $u = 0$,

$$F_u(x, u)p + G_u(x, u) = 0$$

and

$$[F_u(x, u)p + G_u(x, u)]_u = 2\mathfrak{B} > 0.$$

Hence we can apply the implicit function theorem for analytic functions. Since $f(x, u)$ and $g(x, u)$ are power series beginning with terms of degree two and three, respectively, we can compute the linear terms in the series expansion of $u_*(x, p)$ to get

$$2\mathfrak{C}^*x + 2\mathfrak{B}u_*(x, p) + B^*p + (\text{higher degree terms in } (x, p)) = 0$$

which shows that

$$u_*(x, p) = -\tfrac{1}{2}\mathfrak{B}^{-1}(2\mathfrak{C}^*x + B^*p) + (\text{higher degree terms}).$$

LEMMA 2.3. *In the collection of all positive definite real symmetric $n \times n$ matrices there exists a unique solution P_* to the quadratic matrix equation*

$$(\mathfrak{U} - \mathfrak{C}\mathfrak{B}^{-1}\mathfrak{C}^*) + (A - B\mathfrak{B}^{-1}\mathfrak{C}^*)^*P + P(A - B\mathfrak{B}^{-1}\mathfrak{C}^*) - P(B\mathfrak{B}^{-1}B^*)P = 0.$$

For the proof see [8]. There the converse is also proved—namely, if this equation has a solution $P_* > 0$ for $\begin{pmatrix} \mathfrak{U} & \mathfrak{C} \\ \mathfrak{C}^* & \mathfrak{B} \end{pmatrix} > 0$, then the solution is unique and the matrices A, B are stabilizable.

LEMMA 2.4. *Suppose there exists a C^ω stabilizing feedback control $u_*(x) = D_*x + h_*(x)$ for the analytic system* (1.1) *which solves the nonlinear functional equation*

$$F_u(x, u_*(x))J_x(x, u_*) + G_u(x, u_*(x)) = 0 \tag{$\mathfrak{F}$}$$

for all x near the origin in R^n. Then:

(i) *u_* is the unique C^ω solution to* ($\mathfrak{F}$);

(ii) *u_* is the unique C^ω stabilizing feedback control*;

(iii) *u_* synthesizes the unique optimal open-loop control.*

Furthermore, $D_* = -\mathfrak{B}^{-1}[\mathfrak{C}^* + B^*P_*]$ and $J(x, u_*) = x \cdot P_*x + j_*(x)$, where P_* is defined in Lemma 2.3 *and $j_*(x)$ is a power series of higher order terms described in Lemma* 2.1.

Consider the real-valued function defined near the origin in R^{n+r},

$$Q(x, u) = F(x, u) \cdot J_x(x, u_*) + G(x, u).$$

By Lemma 2.1,

$$Q(x, u_*(x)) \equiv 0 \quad \text{near} \quad x = 0$$

and our hypothesis asserts that

$$Q_u(x, u_*(x)) \equiv 0 \quad \text{near} \quad x = 0.$$

Compute the Hessian

$$Q_{uu}(0, 0) = 2\mathfrak{B} > 0.$$

Hence there exists an $\varepsilon > 0$ such that

$$0 = Q(x, u_*(x)) \leqq Q(x, u_1);$$

that is,

$$0 = F(x, u_*(x)) \cdot J_x(x, u_*) + G(x, u_*(x)) \leqq F(x, u_1) \cdot J_x(x, u_*) + G(x, u_1)$$

provided $|x| < \varepsilon$ and $|u_1| < \varepsilon$; moreover, strict inequality holds for $u_1 \neq u_*(x)$. We take ε sufficiently small so that $G(x, u) \geqq 0$.

Now let $u_1(x) \not\equiv u_*(x)$ be a C^ω stabilizing feedback control for (1.1) and let N_1^0 be a neighborhood of the origin in R^n such that $|x| \leqq \varepsilon$ and $|u_1(x)| \leqq \varepsilon$ in N_1^0 and each response $x_*(t)$ or $x_1(t)$ to the corresponding feedback control which initiates in some neighborhood $N_1 \subseteq N_1^0$ of the origin remains in N_1^0 for all

$t \geqq 0$. Hence for all initial $x_0 \in N_1^0$,

$$0 \leqq \int_0^\infty [F(x_1(t), u_1(x_1(t))) \cdot J_x(x_1(t), u_*) + G(x_1(t), u_1(x_1(t)))] \, dt$$

which yields the result

$$0 \leqq -J(x_0, u_*) + J(x_0, u_1)$$

or

$$J(x_0, u_*) \leqq J(x_0, u_1),$$

and strict inequality holds provided $u_1(x_0) \neq u_*(x_0)$. Therefore u_* is the unique optimal feedback control and the unique C^ω solution of the functional equation $(\mathfrak{F})$.

Now choose a neighborhood $N_* \subseteq N_1^0$ of the origin which is positive invariant for the responses to the optimal control $u_*(x)$. Choose N_* so small that $|x_*(t)| \leqq \varepsilon$ on $0 \leqq t < \infty$ for all initial conditions in N_*. Now let $\hat{x}_0$ be an arbitrary fixed initial condition in N_* and consider any measurable open-loop control $\hat{u}(t)$ satisfying the conditions that $|\hat{u}(t)| \leqq \varepsilon$ and the response $\hat{x}(t) \subseteq N_*$ for all $t \geqq 0$. There is no loss in assuming that

$$C(\hat{u}) = \int_0^\infty G(\hat{x}(t), \hat{u}(t)) \, dt < \infty$$

since $G(\hat{x}(t), \hat{u}(t)) \geqq 0$. Then, as above,

$$0 = F(x, u_*(x)) \cdot J_x(x, u_*) + G(x, u_*(x))$$
$$\leqq F(x, \hat{u}(t)) \cdot J_x(x, u_*) + G(x, \hat{u}(t))$$

with strict inequality holding where $\hat{u}(t) \neq u_*(x)$. If $\hat{u}(t) = u_*(\hat{x}(t))$ almost everywhere on $0 \leqq t < \infty$, then the uniqueness theorem for differential equations asserts $\hat{x}(t) = x_*(t)$ which implies that $\hat{u}(t) = u_*(x_*(t))$ almost everywhere. Now assume that $\hat{u}(t) \neq u_*(\hat{x}(t))$ on some set with positive measure. Then

$$0 < \int_0^\infty [F(\hat{x}(t), \hat{u}(t)) \cdot J_x(\hat{x}(t), u_*) + G(\hat{x}(t), \hat{u}(t))] \, dt.$$

Since $C(\hat{u}) = \int_0^\infty G(\hat{x}, \hat{u}) \, dt < \infty$, it is easy to show that $\lim_{t\to\infty} \hat{x}(t) = 0$. Hence

$$0 < -J(\hat{x}_0, u_*) + C(\hat{u}) \quad \text{and} \quad C(u_*) = J(\hat{x}_0, u_*) < C(\hat{u}).$$

Thus $u_*(x_*(t))$ is the unique optimal open-loop control for $\hat{x}_0$ with the required constraints.

By Lemmas 2.1, 2.2,

$$u_*(x) = -\tfrac{1}{2}\mathfrak{B}^{-1}(2\mathfrak{C}^* x + 2B^* P_* x) + (\text{higher terms in } x).$$

By Lemma 2.1 we have

$$F(x, u_*(x)) \cdot J_x(x, u_*) + G(x, u_*(x)) = 0$$

for $|x|$ small. Expanding the left-hand side and using the expansion of $u_*(x)$ to collect the linear terms in x produces the equation

$$(\mathfrak{U} - \mathfrak{C}\mathfrak{B}^{-1}\mathfrak{C}^*) + P(A - B\mathfrak{B}^{-1}\mathfrak{C}^*) + (A - B\mathfrak{B}^{-1}\mathfrak{C}^*)^*P - P(B\mathfrak{B}^{-1}B^*)P = 0,$$

which is the equation of Lemma 2.3. But $P > 0$ by Lemma 2.1 and hence, by the uniqueness of the solution, $P = P_*$ and $J(x, u_*) = x \cdot P_*x + j_*(x)$.

Proof of Theorem 1.2. From the assumption that $\begin{pmatrix} \mathfrak{U} & \mathfrak{C} \\ \mathfrak{C}^* & \mathfrak{B} \end{pmatrix} > 0$ it follows from setting $u = -\mathfrak{B}^{-1}\mathfrak{C}^*x$ in the associated quadratic form that $\mathfrak{U} - \mathfrak{C}\mathfrak{B}^{-1}\mathfrak{C}^* > 0$. By defining $A_* = A + BD_* = A - B\mathfrak{B}^{-1}[\mathfrak{C}^* + B^*P_*]$, we may write the matrix equation of Lemma 2.3,

$$A_*^*P_* + P_*A_* = -[(\mathfrak{U} - \mathfrak{C}\mathfrak{B}^{-1}\mathfrak{C}^*) + P_*(B\mathfrak{B}^{-1}B^*)P_*] < 0,$$

which is the standard Lyapunov equation [8], from which it follows that A_* is a stability matrix. If we let $u_*(x) = D_*x$, then in terms of its associated quadratic form we can write the quadratic matrix equation as

$$A_*x \cdot 2P_*x + G(x, u_*(x)) = 0,$$

for all $x \in R^n$. By integrating this equation along the trajectory $\dot{x} = A_*x$, $x(0) = x_0$, where x_0 is any initial condition in R^n, we obtain the equation

$$x_0 \cdot P_*x_0 = \int_0^\infty G(x, u_*(x))\,dt.$$

But the integral is just $J(x_0, u_*)$. It is now a simple matter to verify that $u_*(x)$ satisfies the functional equation ($\mathfrak{F}$) in Lemma 2.4. The global nature of $u_*(x)$ can be concluded by carefully examining the proof of Lemma 2.4 or else by noting that $u_*(x)$ is given by the same formula as the synthesis of the optimal open-loop controls studied in [8].

2.1. A linear Hamiltonian system. In order to set up the nonlinear problem for a perturbation analysis we reformulate the solution to the truncated problem (Theorem 1.2) in terms of a Hamiltonian system.

We consider the quadratic form

$$2H_*(x, p) = x \cdot \mathfrak{U}_*x + 2x \cdot \mathfrak{C}_*p - p \cdot \mathfrak{B}_*p$$

in the two real n-vectors x and p where we define

$$\begin{aligned} \mathfrak{U}_* &= 2(\mathfrak{U} - \mathfrak{C}\mathfrak{B}^{-1}\mathfrak{C}^*), \\ \mathfrak{B}_* &= \tfrac{1}{2}B\mathfrak{B}^{-1}B^*, \\ \mathfrak{C}_* &= (A - B\mathfrak{B}^{-1}\mathfrak{C}^*)^* \end{aligned}$$

in terms of the data ($A, B, \mathfrak{U}, \mathfrak{B}$ and $\mathfrak{C}$) from the truncated problem. It generates a Hamiltonian system in R^{2n},

$$\dot{x} = \frac{\partial H_*(x, p)}{\partial p},$$

$$\dot{p} = -\frac{\partial H_*(x, p)}{\partial x},$$

which may be written in the matrix form

$$\begin{pmatrix} \dot{x} \\ \dot{p} \end{pmatrix} = \begin{pmatrix} \mathfrak{C}_*^* & -\mathfrak{B}_* \\ -\mathfrak{U}_* & -\mathfrak{C}_* \end{pmatrix} \begin{pmatrix} x \\ p \end{pmatrix}.$$

LEMMA 2.5. *The linear Hamiltonian system*

$$\begin{pmatrix} \dot{x} \\ \dot{p} \end{pmatrix} = \begin{pmatrix} \mathfrak{C}_*^* & -\mathfrak{B}_* \\ -\mathfrak{U}_* & -\mathfrak{C}_* \end{pmatrix} \begin{pmatrix} x \\ p \end{pmatrix}$$

in R^{2n} transforms into the system

$$\begin{pmatrix} \dot{y} \\ \dot{q} \end{pmatrix} = \begin{pmatrix} A_* & 0 \\ 0 & -A_* \end{pmatrix} \begin{pmatrix} y \\ q \end{pmatrix}$$

by a nonsingular real linear transformation $\begin{pmatrix} y \\ q \end{pmatrix} = M \begin{pmatrix} x \\ p \end{pmatrix}$. *The formula for M is*

$$M = \begin{pmatrix} I_n - 2QP_* & Q \\ 2P_* & -I_n \end{pmatrix},$$

where $A_* = A + BD_*$, $D_* = -\mathfrak{B}^{-1}[\mathfrak{C}^* + B^*P_*]$ and Q is the matrix solution of the equation $A_*Q + QA_*^* = -\mathfrak{B}_*$, $Q = \int_0^\infty e^{A_*t}\mathfrak{B}_*e^{A_*^*t}\,dt$. The inverse of M is given by the formula

$$M^{-1} = \begin{pmatrix} I_n & Q \\ 2P_* & 2P_*Q - I_n \end{pmatrix}.$$

The integral defining Q converges uniformly since, as we noted above, A_* is a stability matrix. Integration by parts shows that the integral solves the equation $A_*Q + QA_*^* = -\mathfrak{B}_*$. It is clear that $Q \geqq 0$ since $\mathfrak{B}_* \geqq 0$. We verify the fact that M is nonsingular and the formula for M^{-1} by factoring M:

$$M = \begin{pmatrix} I_n - 2QP_* & Q \\ 2P_* & -I_n \end{pmatrix} = \begin{pmatrix} I_n & Q \\ 0 & -I_n \end{pmatrix} \begin{pmatrix} I_n & 0 \\ -2P_* & I_n \end{pmatrix},$$

$$M^{-1} = \begin{pmatrix} I_n & 0 \\ -2P_* & I_n \end{pmatrix}^{-1} \begin{pmatrix} I_n & Q \\ 0 & -I_n \end{pmatrix}^{-1}$$

$$= \begin{pmatrix} I_n & 0 \\ 2P_* & I_n \end{pmatrix} \begin{pmatrix} I_n & Q \\ 0 & -I_n \end{pmatrix} = \begin{pmatrix} I_n & Q \\ 2P_* & 2P_*Q - I_n \end{pmatrix}.$$

The proof now reduces to the verification of the equation

$$\begin{pmatrix} I_n - 2QP_* & Q \\ 2P_* & -I_n \end{pmatrix}\begin{pmatrix} \mathfrak{C}_*^* & -\mathfrak{B}_* \\ -\mathfrak{U}_* & -\mathfrak{C}_* \end{pmatrix} - \begin{pmatrix} A^* & 0 \\ 0 & -A_*^* \end{pmatrix}\begin{pmatrix} I_n - 2QP_* & Q \\ 2P_* & -I_n \end{pmatrix} = 0.$$

This is equivalent to showing that each of the four terms

(i) $\mathfrak{C}_*^* - 2QP_*\mathfrak{C}_*^* - Q\mathfrak{U}_* - A_* + 2A_*QP_*$,
(ii) $-\mathfrak{B}_* + 2QP_*\mathfrak{B}_* - Q\mathfrak{C}_* - A_*Q$,
(iii) $2P_*\mathfrak{C}_*^* + \mathfrak{U}_* + 2A_*^*P_*$,
(iv) $-2P_*\mathfrak{B}_* + \mathfrak{C}_* - A_*^*$

is zero.

From the formula for A_* written in terms of $\mathfrak{C}_*$ and $\mathfrak{B}_*$, $A_* = \mathfrak{C}_*^* - 2\mathfrak{B}_*P_*$ which shows term (iv) is zero.

Substitution of

$$A_*Q = -\mathfrak{B}_* - QA_*^* = -\mathfrak{B}_* - Q(\mathfrak{C}_* - 2P_*\mathfrak{B}_*)$$

into (ii) shows that term is zero and in term (iv) produces

$$\begin{aligned}
&\mathfrak{C}_*^* - 2QP_*\mathfrak{C}_*^* - Q\mathfrak{U}_* - (\mathfrak{C}_*^* - 2\mathfrak{B}_*P_*) + 2[-\mathfrak{B}_* - Q(\mathfrak{C}_* - 2P_*\mathfrak{B}_*)]P_* \\
&\quad = -2Q[\tfrac{1}{2}\mathfrak{U}_* + \mathfrak{C}_*P_* + P_*\mathfrak{C}_*^* - 2P_*\mathfrak{B}_*P_*] \\
&\quad = -2Q[(\mathfrak{U} - \mathfrak{C}\mathfrak{B}^{-1}\mathfrak{C}^*) + (A - B\mathfrak{B}^{-1}\mathfrak{C}^*)^*P_* + P_*(A - B\mathfrak{B}^{-1}\mathfrak{C}^*) \\
&\qquad -P_*(B\mathfrak{B}^{-1}B^*)P_*] \\
&\quad = 0.
\end{aligned}$$

Substituting for A_* in term (iii) we have

$$2P_*\mathfrak{C}_*^* + \mathfrak{U}_* + 2(\mathfrak{C}_* - 2P_*\mathfrak{B}_*)P_* = 2[\tfrac{1}{2}\mathfrak{U}_* + \mathfrak{C}_*P_* + P_*\mathfrak{C}_*^* - 2P_*\mathfrak{B}_*P_*] = 0.$$

The following theorem restates the conclusions of Theorem 1.2 in a form which points to a proof of Theorem 1.1 by a perturbation analysis.

THEOREM 2.6. *For the linear Hamiltonian system*

$$\begin{pmatrix} \dot{x} \\ \dot{p} \end{pmatrix} = \begin{pmatrix} \mathfrak{C}_*^* & -\mathfrak{B}_* \\ -\mathfrak{U}_* & -\mathfrak{C}_* \end{pmatrix}\begin{pmatrix} x \\ p \end{pmatrix}$$

in R^{2n} there is a linear n-dimensional invariant manifold in which the origin is asymptotically stable. The manifold is described by the equation $p = 2P_*x$. Moreover, this manifold generates the optimal feedback control for the truncated problem of Theorem 1.2. That is, if we define $p_*(x) = 2P_*x$, then

$$u_*(x, p_*(x)) = -\tfrac{1}{2}\mathfrak{B}^{-1}[2\mathfrak{C}^*x + B^*p_*(x)] = D_*x$$

is the optimal control. (See Lemma 2.2.)

The motion in the manifold projects as the optimal closed-loop motion; that is, for any trajectory $\begin{pmatrix} x(t) \\ p(t) \end{pmatrix}$ in the manifold, $\dot{x} = Ax + B(D_*x)$.

The fact that the equation $p = 2P_* x$ describes the required manifold follows from Lemma 2.5 by noting that $q = 0$ if and only if $p = 2P_* x$ and recalling that $A_* = A + BD_*$ is a stability matrix. The projected motion is optimal because when $p = 2P_* x$, $y = x$ and $\dot{x} = A_* x$. The optimality of $u_*(x, p_*(x))$ is verified by the calculation

$$\begin{aligned} u_*(x, p_*(x)) &= -\tfrac{1}{2}\mathfrak{B}^{-1}(2\mathfrak{C}^* x + B^* p_*(x)) \\ &= -\mathfrak{B}^{-1}(\mathfrak{C}^* + B^* P_*)x = D_* x \end{aligned}$$

together with Theorem 1.2.

We notice that the equation $(I_n - 2QP_*)x + Qp = 0$ describes another linear n-dimensional invariant manifold in which the origin is asymptotically unstable.

2.2. A nonlinear Hamiltonian system. We now turn to the nonlinear (perturbed) control problem. Conclusions analogous to Theorem 2.6 are obtained by considering a perturbed Hamiltonian system. In terms of the given functions $F(x, u)$ and $G(x, u)$ and the function $u_*(x, p)$ defined in Lemma 2.2 we select the Hamiltonian

$$H_*(x, p) = F(x, u_*(x, p)) \cdot p + G(x, u_*(x, p))$$

and analyze the corresponding system of canonical differential equations

$$\begin{aligned} \dot{x} &= \frac{\partial H_*(x, p)}{\partial p}, \\ \dot{p} &= -\frac{\partial H_*(x, p)}{\partial x}. \end{aligned} \tag{2.1}$$

The linear part of this system is the linear Hamiltonian system previously studied in Lemma 2.5 and Theorem 2.6.

THEOREM 2.7. *For the nonlinear analytic Hamiltonian system* (2.1) *in* R^{2n} *there exists a real n-dimensional analytic invariant manifold S in which the origin is asymptotically stable.*

The proof is carried out by an analysis in $2n$-dimensional complex space. The conclusions of the theorem are drawn by restricting the calculated results to the real part of the space. We show that there exists an n-vector of real analytic functions $q_*(y)$ defined in a neighborhood of the origin in R^n such that the equation $q = q_*(y)$ defines an n-dimensional manifold $\tilde{S}$ in $\begin{pmatrix} y \\ q \end{pmatrix}$-space. The required manifold S in $\begin{pmatrix} x \\ p \end{pmatrix}$-space is obtained from $\tilde{S}$ by the nonsingular real linear transformation

$$\begin{pmatrix} x \\ p_*(x) \end{pmatrix} = M^{-1}\begin{pmatrix} y \\ q_*(y) \end{pmatrix},$$

where the formula for M^{-1} is given in Lemma 2.5, and hence S is defined in terms of the curvilinear coordinates x.

Since $u_*(x, p)$ satisfies the defining equation in Lemma 2.2,

$$F_u(x, u_*)p + G_u(x, u_*) = 0,$$

the canonical system (2.1) can be rewritten as

$$\dot{x} = F(x, u_*),$$
$$\dot{p} = -[F_x(x, u_*)p + G_x(x, u_*)],$$

where $u_*(x, p) = -\frac{1}{2}\mathfrak{B}^{-1}(2\mathfrak{C}^*x + B^*p) + h_*(x, p)$. By collecting the linear terms we see that the equations have the form

$$\begin{pmatrix} \dot{x} \\ \dot{p} \end{pmatrix} = \begin{pmatrix} \mathfrak{C}_*^* & -\mathfrak{B}_* \\ -\mathfrak{U}_* & -\mathfrak{C}_* \end{pmatrix} \begin{pmatrix} x \\ p \end{pmatrix} + r\begin{pmatrix} x \\ p \end{pmatrix},$$

where

$$r\begin{pmatrix} x \\ p \end{pmatrix} = \begin{bmatrix} Bh_*(x, p) + f(x, u_*) \\ -[2\mathfrak{C}h_*(x, p) + g_x(x, u_*) + f_x(x, u_*)p] \end{bmatrix}.$$

By Lemma 2.5 the change of variables $\begin{pmatrix} y \\ q \end{pmatrix} = M\begin{pmatrix} x \\ p \end{pmatrix}$ transforms the system into

$$\begin{pmatrix} \dot{y} \\ \dot{q} \end{pmatrix} = \begin{pmatrix} A_* & 0 \\ 0 & -A_* \end{pmatrix} \begin{pmatrix} y \\ q \end{pmatrix} + r_M\begin{pmatrix} y \\ q \end{pmatrix}, \tag{2.2}$$

where

$$r_M\begin{pmatrix} y \\ q \end{pmatrix} = Mr\left(M^{-1}\begin{pmatrix} y \\ q \end{pmatrix}\right).$$

By the mean value theorem, for every $\varepsilon > 0$ there exists a δ such that the Lipschitz condition

$$\left| r_M\begin{pmatrix} y \\ q \end{pmatrix} - r_M\begin{pmatrix} u \\ v \end{pmatrix} \right| \leqq \varepsilon \left\| \begin{pmatrix} y \\ q \end{pmatrix} - \begin{pmatrix} u \\ v \end{pmatrix} \right\|$$

holds for $\left\| \begin{pmatrix} y \\ q \end{pmatrix} \right\| \leqq \delta$ and $\left\| \begin{pmatrix} u \\ v \end{pmatrix} \right\| \leqq \delta$. Since A_* is a stability matrix, a conditional stability theorem [4, p. 329] can be applied to establish the existence of the required manifold. We outline the proof of the cited theorem insofar as it is applied here.

Let

$$U_1(t) = \begin{pmatrix} e^{tA_*} & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad U_2(t) = \begin{pmatrix} 0 & 0 \\ 0 & e^{-tA_*^*} \end{pmatrix}.$$

Then

$$\exp t\begin{pmatrix} A_* & 0 \\ 0 & -A_*^* \end{pmatrix} = U_1(t) + U_2(t) \quad \text{and} \quad \dot{U}_j = \begin{pmatrix} A_* & 0 \\ 0 & -A_*^* \end{pmatrix} U_j, j = 1, 2.$$

Let $\alpha > 0$ be chosen so that the real parts of the characteristic values of A_* are less than $-\alpha$. Then there exist positive constants k and σ such that

$$\|U_1(t)\| \leqq ke^{-(\alpha+\sigma)t}, \qquad t \geqq 0,$$
$$\|U_2(t)\| \leqq ke^{\sigma t}, \qquad t \leqq 0.$$

Consider the integral equation

$$\begin{aligned}\theta(t, a) = U_1(t)a + \int_0^t U_1(t-s)r_M(\theta(s, a))\,ds \\ - \int_t^\infty U_2(t-s)r_M(\theta(s, a))\,ds,\end{aligned} \tag{2.3}$$

where a is a constant vector. Let ε be chosen above so that $2\varepsilon k/\sigma < \frac{1}{2}$ and let $|a|$ satisfy $2k|a| < \delta$. Using successive approximations to solve (2.3) with initial approximation $\theta_{(0)}(t, a) = 0$, we readily obtain

$$|\theta_{(l+1)}(t, a) - \theta_{(l)}(t, a)| \leqq \frac{k|a|}{2^l} e^{-\alpha t},$$

which leads to the existence of a solution θ of (2.3) which satisfies

$$|\theta(t, a)| \leqq 2k|a|e^{-\alpha t}$$

The last n components of the vector a do not enter into the solution since they do not enter the successive approximations. That θ is a solution of (2.3) is immediate for $|a|$ small, since by the estimate of $\|U_2(t)\|$ the integral in (2.3) converges. It is also clear from the uniform convergence of the successive approximations that θ is continuous in (t, a) for $t \geqq 0$ and $|a|$ small. Furthermore, θ is analytic in a for fixed t. From (2.3) it follows that the first n components of $\theta(0, a)$ are $\theta_j(0, a) = a_j$, $j = 1, 2, \cdots, n$, and the latter components are given by

$$\theta_j(0, a) = -\left[\int_0^\infty U_2(-s)r_M(\theta(s, a))\,ds\right]_j,$$

$j = n + 1, \cdots, 2n$, where $[\cdot]_j$ denotes the jth component. We define the function q_* by

$$q_{*j}(a_1, a_2, \cdots, a_n) = -\left[\int_0^\infty U_2(-s)r_M(\theta(s, a))\,ds\right]_{n+j}$$

for $j = 1, 2, \cdots, n$, and the initial values $\begin{pmatrix} y \\ q \end{pmatrix} = \theta(0, a)$ satisfy the equation

$$q - q_*(y) = 0$$

in $\begin{pmatrix} y \\ q \end{pmatrix}$-space which defines an n-dimensional manifold $\tilde{S}$ in $\begin{pmatrix} y \\ q \end{pmatrix}$-space.

Each point θ_0 on $\tilde{S}$ can be written as $\theta(0, a)$ for some a and we notice that

$\theta(0, \theta_0) = \theta_0$. Let $\begin{pmatrix} y \\ q \end{pmatrix}(t, \theta)$ denote the unique solution of (2.2) for which $\begin{pmatrix} y \\ q \end{pmatrix}(0, \theta) = \theta$ in a neighborhood of the origin in $2n$-dimensional complex space. Then $\theta(t, \theta_0) = \begin{pmatrix} y \\ q \end{pmatrix}(t, \theta_0)$ for $t \geqq 0$, $\theta_0 \in \tilde{S}$ and $|\theta_0|$ small because both trajectories satisfy (2.2) and they intersect at $t = 0$. An important consequence of the previous equation is that

$$\begin{pmatrix} y \\ q \end{pmatrix}(t, \theta_0) \to 0 \quad \text{as} \quad t \to \infty$$

for $\theta_0 \in \tilde{S}$ and $|\theta_0|$ small because

$$\left|\begin{pmatrix} y \\ q \end{pmatrix}(t, \theta_0)\right| = |\theta(t, \theta_0)| \leqq 2k|\theta_0|e^{-\alpha t}.$$

Moreover we can show that the trajectories of (2.2) intersecting $\tilde{S}$ do not leave $\tilde{S}$. Let $\theta_0 \in \tilde{S}$ and $|\theta_0|$ be small. It satisfies the equation for the manifold which is

$$\theta_0 - \left[\begin{pmatrix} I_n & 0 \\ 0 & 0 \end{pmatrix}\theta_0 - \int_0^\infty U_2(-s)r_M(\theta(s, \theta_0))\, ds\right] = 0.$$

The trajectory through θ_0 at $t = 0$, $\theta(t, \theta_0) = \begin{pmatrix} y \\ q \end{pmatrix}(t, \theta_0)$ satisfies the differential equation (2.2). Differentiating the left-hand side of the equation for $\tilde{S}$ along the trajectory by first replacing $\theta(s, \theta_0)$ by $\begin{pmatrix} y \\ q \end{pmatrix}(s, \theta_0)$ and then replacing θ_0 by $\begin{pmatrix} y \\ q \end{pmatrix}(t, \theta_0)$ and using the formula

$$\begin{pmatrix} y \\ q \end{pmatrix}\left(s, \begin{pmatrix} y \\ q \end{pmatrix}(t, \theta_0)\right) = \begin{pmatrix} y \\ q \end{pmatrix}(s + t, \theta_0),$$

which follows from the uniqueness of the solutions to (2.2), we find the derivative to be zero at $t = 0$. This proves that the trajectory through θ_0 at $t = 0$ for $|\theta_0|$ small does not leave $\tilde{S}$.

Proof of Theorem 1.1 (Analytic case). To prove the main theorem it is sufficient to establish the existence of a C^ω stabilizing control which solves the functional equation ($\mathfrak{F}$) occurring in Theorem 1.1 and Lemma 2.4. The remaining conclusions of the theorem then follow as a corollary to the lemma.

We define $u_*(x) = u_*(x, p_*(x))$ about the origin in complex n-dimensional space. Recall $p_*(x)$ describes the manifold S discussed in Theorem 2.7 and $u_*(x, p)$ was defined in Lemma 2.2 by the implicit function theorem. In the proof of Theorem 2.7 we proved that the motion of (2.1) on S about the origin in complex $2n$-space satisfies

$$\begin{aligned} \dot{x} &= F(x, u_*(x)), \\ \dot{p}_*(x) &= -[F_x(x, u_*(x))p_*(x) + G_x(x, u_*(x))], \end{aligned} \tag{2.4}$$

where $x = x(t, x_0)$, $x(0, x_0) = x_0$ with $|x_0|$ small. The integral

$$\int_0^\infty G(x, u_*(x))\,dt = \int_0^\infty G(x, u_*(x, p_*(x)))\,dt$$

converges uniformly in a neighborhood of the origin in complex n-space by the estimate

$$\left\|\begin{pmatrix} x \\ p_*(x) \end{pmatrix}\right\| \leqq 2\|M^{-1}\|k\left|M\begin{pmatrix} x_0 \\ p_*(x_0) \end{pmatrix}\right|e^{-\alpha t}$$

which follows easily from the estimate of $\left|\begin{pmatrix} y \\ q \end{pmatrix}(t, \theta_0)\right|$ established in the proof of Theorem 2.7. The uniform convergence together with the continuity and analyticity of the functions in the integrand permit the following differentiation of the integral when the initial conditions are restricted to real space. We shall use the defining equation for $u_*(x, p)$,

$$F_u(x, u_*(x, p))p + G_u(x, u_*(x, p)) = 0, \tag{2.5}$$

and the differential equations (2.4) for the motion on S.

$$\begin{aligned}
\frac{\partial J(x_0, u_*)}{\partial x_0} &= \int_0^\infty \left[\frac{\partial x}{\partial x_0}\frac{\partial G(x, u_*)}{\partial x} + \frac{\partial u_*}{\partial x_0}\frac{\partial G}{\partial u_*}\right]dt \\
&= \int_0^\infty \left\{\frac{\partial x}{\partial x_0}\left[-\dot{p}_*(x) - \frac{\partial F}{\partial x}(x, u_*)p_*(x)\right] + \frac{\partial u_*}{\partial x_0}\frac{\partial G}{\partial u_*}\right\}dt \\
&= p_*(x_0) + \int_0^\infty \left[\frac{d}{dt}\frac{\partial x}{\partial x_0}\right]p_*(x)\,dt \\
&\quad + \int_0^\infty \left[\frac{\partial u_*}{\partial x_0}\frac{\partial G}{\partial u_*} - \left(\frac{\partial x}{\partial x_0}\frac{\partial F}{\partial x}\right)(x, u_*)p_*(x)\right]dt \\
&= p_*(x_0) + \int_0^\infty \frac{\partial F}{\partial x_0}(x, u_*)p_*(x)\,dt \\
&\quad + \int_0^\infty \left[\frac{\partial u_*}{\partial x_0}\left(-\frac{\partial F}{\partial u_*}\right)p_*(x) - \left(\frac{\partial x}{\partial x_0}\frac{\partial F}{\partial x}\right)p_*(x)\right]dt \\
&= p_*(x_0).
\end{aligned}$$

Thus, $J_x(x_0, u_*) = p_*(x_0)$ for $|x_0|$ small, and hence by (2.5),

$$F_u(x_0, u_*(x_0))J_x(x_0, u_*) + G_u(x_0, u_*(x_0)) = 0,$$

which proves $u_*(x)$ solves $(\mathfrak{F})$, and the proof is completed.

2.3. Calculation of the power series for $u_*(x)$ and $J(x, u_*)$. Since $u_*(x)$, $J_*(x) = J(x, u_*)$ and $F_*(x) = F(x, u_*(x))$ are analytic about the origin they can be expanded in power series:

$$\begin{aligned} u_*(x) &= u_*^{(1)}(x) + u_*^{(2)}(x) + \cdots, \\ J_*(x) &= J^{(2)}(x) + J^{(3)}(x) + \cdots, \\ F_*(x) &= F^{(1)}(x) + F^{(2)}(x) + \cdots. \end{aligned}$$

In preceding sections we saw that the lowest order terms have been computed as the solutions to the truncated problem:

$$\begin{aligned} u_*^{(1)}(x) &= D_* x, \\ J^{(2)}(x) &= x \cdot P_* x, \\ F^{(1)}(x) &= A_* x. \end{aligned}$$

The computation reduced to solving the quadratic matrix equation

$$(\mathfrak{U} - \mathfrak{C}\mathfrak{B}^{-1}\mathfrak{C}^*) + P(A - B\mathfrak{B}^{-1}\mathfrak{C}^*) + (A - B\mathfrak{B}^{-1}\mathfrak{C}^*)^*P - P(B\mathfrak{B}^{-1}B^*)P = 0$$

for $P = P_* > 0$ and then computing D_* and A_* by the formulas

$$\begin{aligned} D_* &= -\mathfrak{B}^{-1}(\mathfrak{C}^* + B^*P_*), \\ A_* &= A + BD_*. \end{aligned}$$

We now develop the procedure for computing the remaining terms in the power series. The computation of successively higher order terms reduces to solving successively higher order systems of linear algebraic equations.

By Lemma 2.1 and Theorem 1.1,

$$\begin{aligned} F(x, u_*(x)) \cdot J_x(x, u_*) + G(x, u_*(x)) &= 0, \\ F_u(x, u_*(x)) J_x(x, u_*) + G_u(x, u_*(x)) &= 0 \end{aligned}$$

about the origin. We can rewrite these equations in the form

$$\begin{aligned} [A_*x + B(u_* - D_*x) + f(x, u_*)] &\cdot J_x(x, u_*) \\ &+ x \cdot \mathfrak{U}x + 2x \cdot \mathfrak{C}u_* + u_* \cdot \mathfrak{B}u_* + g(x, u_*) = 0, \\ u_*(x) = -\tfrac{1}{2}\mathfrak{B}^{-1}&[(B + f_u)^* J_x(x, u_*) + 2\mathfrak{C}^*x + g_u]. \end{aligned}$$

Hence,

$$\begin{aligned} A_*x \cdot J_x(x, u_*) &= -[B(u_* - D_*x) + f(x, u_*)] \cdot J_x(x, u_*) \\ &\quad - 2x \cdot \mathfrak{C}u_* - u_* \cdot \mathfrak{B}u_* - g(x, u_*) - x \cdot \mathfrak{U}x, \\ u_*(x) &= -\tfrac{1}{2}\mathfrak{B}^{-1}[(B + f_u)^* J_x(x, u_*) + 2\mathfrak{C}^*x + g_u(x, u_*)]. \end{aligned}$$

Substituting the power series for u_* and $J(x, u_*)$ and then selecting the mth order terms from the former equation and the kth order terms from the latter, we obtain the equations

$$\begin{aligned} A_{*}x \cdot J_x^{(m)}(x) = & - \sum_{k=2}^{m-1} [B(u_{*} - D_{*}x) + f(x, u_{*})]^{(m-k+1)} J_x^{(k)}(x) \\ & - 2x \cdot \mathfrak{C} u_{*}^{(m-1)} - 2 \sum_{k=1}^{[(m-1)/2]} u_{*}^{(k)} \cdot \mathfrak{B} u_{*}^{(m-k)} \\ & - u_{*}^{(m/2)} \cdot \mathfrak{B} u^{(m/2)} - g^{(m)}(x, u_{*}) \end{aligned}$$

for $m = 3, 4, 5, \cdots$ and[2]

$$u_{*}^{(k)}(x) = -\tfrac{1}{2}\mathfrak{B}^{-1}[B^{*}J_x^{(k+1)}(x) + \sum_{j=1}^{k-1} (f_u)^{*(j)} J_x^{(k-j+1)}(x) + g_u^{(k)}(x, u_{*})],$$

$$k = 2, 3, \cdots .^{3}$$

The right-hand side of the former equation is independent of $u_{*}^{(m-1)}$ since its coefficient is $-[B^{*}J_x^{(2)}(x) + 2\mathfrak{C}^{*}x + 2\mathfrak{B}u^{(1)}] = 0$. Thus

$$\begin{aligned} A_{*}x \cdot J_x^{(m)}(x) = & - \sum_{k=3}^{m-1} B u_{*}^{(m-k+1)}(x) \cdot J_x^{(k)}(x) \\ & - \sum_{k=2}^{m-1} f^{(m-k+1)}(x, u_{*}) \cdot J_x^{(k)}(x) \\ & - 2 \sum_{k=2}^{[(m-1)/2]} u_{*}^{(k)}(x) \cdot \mathfrak{B} u_{*}^{(m-k)}(x) - u_{*}^{(m/2)}(x) \cdot \mathfrak{B} u_{*}^{(m/2)}(x) \\ & - g^{(m)}(x, u_{*}) \end{aligned}$$

for $m = 3, 4, 5, \cdots$.

We can show that these two equations can be solved to generate the power series for $u_{*}(x)$ and $J_{*}(x)$. Since $f(x, u)$ and $g(x, u)$ are power series beginning with terms in (x, u) of order two and three, respectively, it is clear that

$$f^{(j)}(x, u_{*}^{(1)} + u_{*}^{(2)} + \cdots) = f^{(j)}(x, u_{*}^{(1)} + u_{*}^{(2)} + \cdots + u_{*}^{(j-1)})$$

and

$$g^{(m)}(x, u_{*}^{(1)} + u_{*}^{(2)} + \cdots) = g^{(m)}(x, u_{*}^{(1)} + u_{*}^{(2)} + \cdots + u_{*}^{(m-2)}),$$

so the sequence of terms

$$\{u_{*}^{(1)}, u_{*}^{(2)}, \cdots, u_{*}^{(m-2)}; J^{(2)}, J^{(3)}, \cdots, J^{(m-1)}\}$$

determines the right-hand side of the equation for $J^{(m)}(x)$ and the terms

$$\{u_{*}^{(1)}, u_{*}^{(2)}, \cdots, u_{*}^{(k-1)}; J^{(2)}, J^{(3)}, \cdots, J^{(k+1)}\}$$

determine the right-hand side of the equation for $u_{*}^{(k)}(x)$.

[2] $[k]$ denotes the integer part of k and the term with $u^{(m/2)}$ in it is to be omitted for m odd.

[3] We use the convention that $\sum_k^l = 0$ for $l < k$.

Since A_* is a stability matrix, the equation for $J^{(m)}(x)$, which has the form

$$A_*x \cdot J_x^{(m)}(x) = H^{(m)}(x),$$

where $H^{(m)}(x)$ is a homogeneous form of degree m in x, has a unique solution

$$J^{(m)}(x) = -\int_0^\infty H^{(m)}(e^{A_*t}x)\,dt.$$

Since the two equations for $J^{(m)}(x)$ are equivalent, the former can be solved for the coefficients of the form $J^{(m)}(x)$ by equating coefficients of similar terms in x and solving the resulting linear algebraic equations. Hence, starting with $u_*^{(1)}(x) = D_*x$ and $J^{(2)}(x) = x \cdot P_*x$ we can compute consecutively the terms in the sequence

$$J^{(2)}(x), \quad u_*^{(1)}(x), \quad J^{(3)}(x), \quad u_*^{(2)}(x), \quad J^{(4)}(x), \quad u_*^{(3)}(x), \cdots$$

thereby generating the power series for $J_*(x)$ and $u_*(x)$.

From the power series the following conclusion is clear.

COROLLARY 2.7. *In the analytic case a sufficient condition for the optimal closed-loop control to be linear is that the following equations hold*:

(i) $f(x, D_*x) = 0$,
(ii) $f_u(x, D_*x) = 0$,
(iii) $g(x, D_*x) = 0$,
(iv) $g_u(x, D_*x) = 0$.

3. Construction of the optimal control for the differentiable case. The basic pattern for treating differentiable systems follows the analytic case. Hence we discuss only those points which will assist the reader in checking the details of the proof himself.

We note that by application of the mean value theorem to the basic assumptions (a)–(f) we obtain the additional inequalities:

$$\text{(g)} \qquad |f(x, u)| \leqq C|(x, u)|^{\alpha+1},$$

$$\text{(h)} \qquad |g(x, u)| \leqq C|(x, u)|^{\alpha+2},$$

$$\text{(i)} \qquad \left|\frac{\partial g(x, u)}{\partial(x, u)}\right| \leqq C|(x, u)|^{\alpha+1},$$

$$\text{(j)} \qquad |h(x)| \leqq L|x|^{\beta+1},$$

which hold near the origin. The remarks about the exponential decay of $|x(t, x_0)|$ and $|u(x(t, x_0))|$ remain valid. We use the basic estimate

$$|x(t, x_0)| \leqq C_1\, e^{-\mu t}|x_0|,$$

which we used in the proof of Lemma 2.1. Using the fact that $x(t, x_0)$ is of class C^1 with respect to x_0 near the origin, inequalities (b), (j) and the fundamental inequality of differential equations, we can easily obtain the estimate

$$\left\|\frac{\partial x(t, x_0)}{\partial x_0}\right\| \leqq C_2\, e^{-\mu t}$$

for all $t \geqq 0$ and $|x_0|$ small. It follows that the integral

$$\int_0^\infty \frac{\partial G}{\partial x_0}(x, u(x))\,dt$$

converges uniformly, hence $J(x_0, u)$ is of class C^1 and

$$\frac{\partial J(x_0, u)}{\partial x_0} = \int_0^\infty \frac{\partial G(x, u(x))}{\partial x_0}dt$$

near the origin. The lemma corresponding to Lemma 2.1 states $J(x_0, u) = x_0 \cdot Px_0 + j(x_0)$, where $j(0) = 0$, $|j_{x_0}(x_0)| \leqq C_3|x_0|^{\alpha+1}$ and hence $|j(x_0)| \leqq C_3|x_0|^{\alpha+2}$ near the origin.

The analogue of Lemma 2.2 states that $u_*(x, p)$ is of class C^1, and since $h_*(x, p) = -\frac{1}{2}\mathfrak{B}^{-1}[g_u(x, u_*) + f_u(x, u_*)p]$, using (b) and (d), we can easily show that $\|\partial h_*(x, p)/\partial(x, p)\| \leqq C_*|(x, p)|^{\alpha^*}$ near the origin for some positive numbers α_* and C_*. This last inequality together with (b) and (d) implies that the higher order term

$$r\begin{pmatrix} x \\ p \end{pmatrix} = \begin{bmatrix} Bh_*(x, p) + f(x, u_*) \\ -[2\mathfrak{C}h_*(x, p) + g_x(x, u_*) + f_x(x, u_*)p] \end{bmatrix}$$

in the nonlinear Hamiltonian system satisfies an inequality

$$\left\| \partial r\begin{pmatrix} x \\ p \end{pmatrix} \Big/ \partial(x, p) \right\| \leqq C_4|(x, p)|^{\alpha}.$$

Hence by the mean value theorem the Lipschitz condition

$$\left| r_M\begin{pmatrix} y \\ q \end{pmatrix} - r_M\begin{pmatrix} u \\ v \end{pmatrix} \right| \leqq \varepsilon\left| \begin{pmatrix} y \\ q \end{pmatrix} - \begin{pmatrix} u \\ v \end{pmatrix} \right|$$

required in the proof of Theorem 2.7 is available. The same proof establishes the existence of the required function $p_*(x)$ described in the proof of the theorem. The fact that $p_*(x)$ is of class C^1 is proved by applying Theorem 4.2 of [4, p. 333]. The remainder of the proof is the same as for the analytic case.

Remark. We note that $J(x_0, u_*)$ is of class C^2 because we proved $J_x(x, u_*) = p_*(x)$.

4. Elementary examples. We now calculate the optimal feedback controls for a few systems.

Example 1 (Natural barriers).

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

$$J = \int_0^\infty [|x|^2 + |u|^2]\,dt.$$

Since the states of the unit sphere in R^n, $(|x| = 1)$, are solutions to the system equation for every control, the domain of the optimal stabilizing control could never be extended outside the unit sphere. It is easy to verify that

$$u_*(x) = -x$$

solves ($\mathfrak{F}$) of Theorem 1.1 and hence is the optimal control. We can compute the optimized integral

$$J_*(x) = -\log(1 - |x|^2)$$

and note that the domain of u_* is the entire unit ball ($|x| < 1$).

Example 2 (A linear quadratic problem).

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= u, \\ J &= \int_0^\infty [x_1^2 + x_2^2 + u^2]\,dt. \end{aligned}$$

Transforming $F(x, u)$ and $G(x, u)$ to vector-matrix notation we obtain the matrices of the system,

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

and those of the performance integrand,

$$\begin{pmatrix} \mathfrak{U} & \mathfrak{C} \\ \mathfrak{C}^* & \mathfrak{B} \end{pmatrix} = \left(\begin{array}{cc|c} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \hline 0 & 0 & 1 \end{array}\right).$$

The quadratic matrix equation for P_* is

$$\mathfrak{U} + PA + A^*P - PBB^*P = 0,$$

which is equivalent to the scalar equations

$$\begin{aligned} -2p_{12} - p_{12}^2 &= 0, \\ 1 + 2[2p_{12} - p_{22}] - p_{22}^2 &= 0, \\ 2p_{11} - p_{12} - p_{22} - p_{12}p_{22} &= 0, \end{aligned}$$

which can be solved to produce the required positive definite matrix

$$P_* = \begin{pmatrix} \sqrt{3} & 1 \\ 1 & \sqrt{3} \end{pmatrix}.$$

From this we compute the matrix which provides the optimal control

$$D_* = -\mathfrak{B}^{-1}(\mathfrak{C}^* + B^*P_*) = -(0 \quad 1)\begin{pmatrix} \sqrt{3} & 1 \\ 1 & \sqrt{3} \end{pmatrix} = (-1 \quad -\sqrt{3}).$$

Therefore the optimal control and corresponding performance integral are

$$u_*(x) = -x_1 - \sqrt{3}x_2,$$
$$J_*(x) = \sqrt{3}x_1^2 + 2x_1x_2 + \sqrt{3}x_2^2.$$

Example 3 (Computation of a higher order term).

$$\dot{x}_1 = x_2,$$
$$\dot{x}_2 = x_1^2 - x_2^2 + \frac{u}{1 - (x_1 - x_2)u},$$
$$J = \int_0^\infty [x_1^2 + x_2^2 + \sin^2 u]\, dt.$$

Expanding the nonlinear terms about the origin we have

$$\dot{x}_1 = x_2,$$
$$\dot{x}_2 = u + x_1^2 - x_2^2 + (x_1 - x_2)u^2 + \cdots,$$
$$J = \int_0^\infty \left[x_1^2 + x_2^2 + u^2 - \frac{u^4}{3} + \cdots\right] dt.$$

We note that the truncated system is the same as Example 2, hence the linear term of the optimal control is

$$u_*^{(1)}(x) = -x_1 - \sqrt{3}x_2$$

and the first term in the expansion of $J_*(x)$ is

$$J^{(2)}(x) = \sqrt{3}x_1^2 + 2x_1x_2 + \sqrt{3}x_2^2.$$

The linear term of the optimized system is

$$A_*x = \begin{pmatrix} x_2 \\ -x_1 - \sqrt{3}x_2 \end{pmatrix}.$$

From the analysis of the power series for u_* developed in § 2 we have the equation

$$A_*x \cdot J_x^{(3)}(x) = -\begin{pmatrix} 0 \\ x_1^2 - x_2^2 \end{pmatrix} \cdot \begin{pmatrix} 2\sqrt{3}x_1 + 2x_2 \\ 2x_1 + 2\sqrt{3}x_2 \end{pmatrix},$$

which must be solved for $J^{(3)}(x)$. Letting $J^{(3)}(x) = c_1x_1^3 + c_2x_1^2x_2 + c_3x_1x_2^2 + c_4x_2^3$, we reduce the equation to the system

$$-c_2 = -2,$$
$$3c_1 - 2c_3 - \sqrt{3}c_2 = -2\sqrt{3},$$
$$2c_2 - 3c_4 - 2\sqrt{3}c_3 = 2,$$
$$c_3 - 3\sqrt{3}c_4 = 2\sqrt{3},$$

which has the solution

$$\begin{aligned} c_1 &= \tfrac{8}{21}\sqrt{3}, \\ c_2 &= 2, \\ c_3 &= \tfrac{4}{7}\sqrt{3}, \\ c_4 &= -\tfrac{10}{21}. \end{aligned}$$

Hence

$$J^{(3)}(x) = \tfrac{8}{21}x_1^3 + 2x_1^2x_2 + \tfrac{4}{7}\sqrt{3}x_1x_2^2 - \tfrac{10}{21}x_2^3.$$

From the general formula for $u_*^{(2)}(x)$ we can now compute

$$\begin{aligned} u_*^{(2)}(x) &= -\tfrac{1}{2}(0 \quad 1)J_x^{(3)}(x) \\ &= -\tfrac{1}{2}(2x_1^2 + \tfrac{8}{7}\sqrt{3}x_1x_2 - \tfrac{10}{7}x_2^2) \\ &= -x_1^2 - \tfrac{4}{7}\sqrt{3}x_1x_2 - \tfrac{5}{7}x_2^2. \end{aligned}$$

Successively higher order terms would be computed in a similar manner.

Example 4 (Equivalent systems). We call two optimization problems (F, G) and $(\tilde{F}, \tilde{G})$ *equivalent* if they generate the same optimal control u_* and the same corresponding integrals J_*. Equivalence is denoted by the notation $[F, G] \equiv [\tilde{F}, \tilde{G}]$. Several facts concerning equivalence can now be stated and are direct consequences of Theorem 1.1.

If $\rho(x) > 0$ near the origin, then

$$[\rho F, G] \equiv \left[F, \frac{G}{\rho}\right].$$

For instance, Example 1 is equivalent to

$$\begin{aligned} \dot{x} &= u, \\ J &= \int_0^\infty \left[\frac{|x|^2 + |u|^2}{1 - |x|^2}\right] dt. \end{aligned}$$

Another valid equivalence is given by

$$[F + \delta F, G + \delta G] \equiv [F, G],$$

which holds for all perturbations satisfying $\delta F(x) \cdot \partial J_*(x)/\partial x + \delta G(x) = 0$. Application of this relation to Example 2 using $\delta F = -\tfrac{1}{2}x_1x_2$ and $\delta G = \sqrt{3}x_1^2x_2 + x_1x_2^2$ generates the equivalent problem

$$\begin{aligned} \dot{x}_1 &= x_2(1 - \tfrac{1}{2}x_1), \\ \dot{x}_2 &= u, \\ J &= \int_0^\infty [x_1^2 + x_2^2 + \sqrt{3}x_1^2x_2 + x_1x_2^2]\, dt. \end{aligned}$$

As a final remark we note that two equivalent problems must have equivalent truncations.

Acknowledgment. These results are part of the author's doctoral thesis written under Lawrence Markus at the University of Minnesota.

REFERENCES

[1] E. G. Al'brekht, *On the optimal stabilization of nonlinear systems*, J. Appl. Math. Mech., 25 (1962), pp. 1254–1266.

[2] ———, *Optimal stabilization of nonlinear systems*, Mathematical Notes, vol. 4, no. 2, The Ural Mathematical Society, The Ural State University of A. M. Gor'kii, Sverdlovsk, 1963. In Russian.

[3] P. Brunovský, *On optimal stabilization of nonlinear systems*, Mathematical Theory of Control, A. V. Balakrishnan and Lucien W. Neustadt, eds., Academic Press, New York and London, 1967.

[4] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[5] R. E. Kalman, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana (2), 5 (1960), pp. 102–119.

[6] ———, *The theory of optimal control and the calculus of variations*, Mathematical Optimization Techniques, R. Bellman, ed., University of California Press, Berkeley, 1963.

[7] E. B. Lee and L. Markus, *Foundations of optimal control theory*, John Wiley, New York, 1967.

[8] D. L. Lukes, *Stabilizability and optimal control*, Funkcial. Ekvac., 11 (1968), pp. 39–50.

SIAM J. CONTROL
Vol. 7, No. 1, February 1969

THE QUADRATIC CRITERION FOR DISTRIBUTED SYSTEMS*

D. L. LUKES† AND D. L. RUSSELL‡

Introduction. Let us consider a linear control system in E^n:

$$\frac{dx}{dt} = Ax + Bu, \tag{*}$$

the control u lying in E^r. A control function $u_*(t)$ lying in $L^2[t_0, T]$ satisfies the quadratic criterion of optimality if it yields a minimum value for the cost functional

$$C(u) = \int_{t_0}^{T} [x^{Tr}(t)Wx(t) + u^{Tr}(t)Uu(t)]\, dt + x^{Tr}(T)Gx(T),$$

where it is understood that $x(t)$ is the response to the control $u(t)$ via (*) corresponding to some initial state $x(t_0) = x_0$. This problem was first studied by Kalman [1] and has recently been re-examined by Wonham [2] and Lukes [3]. An excellent expository treatment is given by Lee and Markus in their text [4].

Our purpose in this article is to explore the applicability of the quadratic criterion to distributed parameter systems or, alternatively, to differential equations in a Hilbert space. We shall show that such application is indeed possible and, perhaps more important from the engineering point of view, that the optimal controls thus found are limits of optimal controls corresponding to finite-dimensional models of the distributed or infinite-dimensional system. The theory will first be developed in the abstract situation and then applied to specific examples.

1. Problem statement.

Some basic assumptions underlying our work will be set forth in this section. Later, in order to obtain additional more specialized results, we shall set forth further hypotheses. A further basic assumption will be made at the beginning of § 3.

Let H_1 and H_2 be real or complex Hilbert spaces. The system state x will be an element of H_1 while the control vector u lies in H_2. We assume that A is a (possibly unbounded) closed linear operator defined on a dense domain $\Delta \subseteq H_1$ and generates a strongly continuous semigroup e^{At} for $t \geqq 0$. Then A^*, the adjoint of A, has the same properties and for some real μ_0, M,

$$\|e^{At}\| = \|e^{A^*t}\| \leqq M\, e^{\mu_0 t}, \qquad t \geqq 0. \tag{1.1}$$

* Received by the editors April 8, 1968, and in revised form August 30, 1968.

† Mathematics Research Center, United States Army, University of Wisconsin, Madison, Wisconsin 53706. The work of this author was sponsored by the Mathematics Research Center under Contract DA-31-124-ARO-D-462.

‡ Departments of Mathematics and Computer Sciences, University of Wisconsin, Madison, Wisconsin 53706. The work of this author was sponsored by the National Science Foundation under Grant GP-8927.

Our assumption is fulfilled, for example, if A is normal with spectrum $\sigma(A)$ lying in a left half-plane $\operatorname{Re}(\mu) \leqq \mu_0$ and then (1.1) holds with $M = 1$. For more general sufficient conditions on A see, e.g., [5], [6], [7]. In all these theories $\sigma(A)$, $\sigma(A^*)$ are required to lie in some left half-plane so that we may without loss of generality assume there is a real λ such that $(A - \lambda I)^{-1}$ and $(A^* - \lambda I)^{-1}$ exist as bounded linear operators on H_1.

Our control system is

$$\frac{dx}{dt} = Ax + \tilde{B}u, \tag{1.2}$$

where $\tilde{B}: H_2 \to H_1$ is a linear operator which can be written in the form

$$\tilde{B} = (A - \lambda I)^{-1}B, \tag{1.3}$$

$B: H_2 \to H_1$ being bounded. A control $u(t)$ defined on $[t_0, T]$ is admissible if it is measurable and

$$\int_{t_0}^{T} \|u(t)\|^2 \, dt < \infty. \tag{1.4}$$

The relevant concepts of measurability and integrability are discussed in [6, Chapter 3].

Let $x_0 \in \Delta$ be given and let $u(t)$ be an admissible control. A result proved by Balakrishnan in [8] shows that the response $x(t)$ to $u(t)$ via (1.2) which initiates at x_0 at time t_0 lies in Δ for all $t \geqq t_0$ and satisfies (1.2) for almost all $t \geqq t_0$. We may therefore form the cost functional

$$C(u) = \int_{t_0}^{T} [(x(t), \tilde{W}x(t)) + (u(t), Uu(t))] \, dt + (x(T), \tilde{G}x(T)), \tag{1.5}$$

provided $\tilde{W}$, U and $\tilde{G}$ satisfy the following conditions. We assume[1]

$$\tilde{W} = (A^* - \lambda I)W(A - \lambda I), \tag{1.6}$$

$$\tilde{G} = (A^* - \lambda I)G(A - \lambda I), \tag{1.7}$$

where W and G are bounded self-adjoint positive semidefinite operators on H and U, U^{-1} are bounded self-adjoint positive definite operators on H_2; i.e., for some positive numbers b_0, b_1,

$$b_0\|u\|^2 \leqq (u, Uu) \leqq b_1\|u\|^2. \tag{1.8}$$

Clearly $C(u) \geqq 0$ and hence, for fixed x_0, has a greatest lower bound $c_{x_0} \geqq 0$. The optimal control problem is this: show that there is an admissible control u_* such that

$$C(u_*) = c_{x_0} \tag{1.9}$$

and characterize u_* in terms of the operators appearing in (1.2) and (1.5).

[1] See remark at the end of § 5.

At first reading the assumptions made on $\tilde{B}$, $\tilde{W}$ and $\tilde{G}$ in this section may appear curious and somewhat arbitrary. That this is not the case is demonstrated by the examples of § 5.

2. Preliminary transformation. It is not easy to treat directly the problem posed in § 1. The main difficulty lies in the fact that the Kalman–Riccati equation (cf. [1], [3], [4]) involves the unbounded operators $\tilde{W}$ and $\tilde{G}$ in a way which makes analysis of its solutions very difficult. We shall see that a very simple transformation overcomes this problem—and creates some new problems.

Since $x(t) \in \Delta$ we may define $y(t) \in H_1$ by

$$x(t) = (A - \lambda I)^{-1} y(t). \tag{2.1}$$

Then since

$$x(t) = e^{A(t-t_0)}x_0 + \int_{t_0}^{t} e^{A(t-s)}(A - \lambda I)^{-1}Bu(s)\,ds, \tag{2.2}$$

we have

$$y(t) = e^{A(t-t_0)}y_0 + \int_{t_0}^{t} e^{A(t-s)}Bu(s)\,ds. \tag{2.3}$$

The vector function $y(t)$ is a "mild" solution of the differential equation $dy/dt = Ay + Bu$. The differential equation itself may not be satisfied by $y(t)$, however. Indeed $y(t)$ need not be differentiable in general. Equation (2.3) is all we have to work with.

The cost functional (1.5) now becomes

$$C(u) = \int_{t_0}^{T} [(y(t), Wy(t)) + (u(t), Uu(t))]\,dt + (y(T), Gy(T)) \tag{2.4}$$

and thus involves bounded operators only. If we now proceed formally, taking $y(t)$ to be a solution of $dy/dt = Ay + Bu$ and follow the existing theory for finite-dimensional systems,we are led to the Kalman–Riccati equation

$$-\frac{dQ}{dt} = A^*Q + QA + W - Q(BU^{-1}B^*)Q \tag{2.5}$$

with terminal condition

$$Q(T) = G. \tag{2.6}$$

Here $B^* : H_1 \to H_2$ is defined by

$$(B^*x, u) = (x, Bu), \qquad x \in H_1, \quad u \in H_2. \tag{2.7}$$

Again, however, we cannot expect "strict," i.e., differentiable solutions of (2.5).

The most we can expect is a solution of the integral equation

$$Q(t) = e^{A^*(T-t)}Ge^{A(T-t)} \tag{2.8}$$
$$- \int_t^T e^{A^*(s-t)}[-W + Q(s)(BU^{-1}B^*)Q(s)]e^{A(s-t)}\,ds$$

which is strongly continuous in t. Now the development of the finite-dimensional theory very definitely makes use of the differentiability of $y(t)$ and $Q(t)$. We shall follow the basic ideas already introduced in the existing finite-dimensional theory but we must not differentiate $y(t)$ nor $Q(t)$.

We remark that the proofs used in the finite-dimensional theory are valid essentially without modification if A is a bounded operator since $y(t)$ and $Q(t)$ are then differentiable. Thus our interest centers on unbounded A.

3. Optimal control on a finite interval. We begin this section with another assumption on the operator A. This assumption was not introduced earlier because the work of § 1 and § 2 does not require it. It is entirely probable that the results which we shall obtain are valid under much weaker conditions.

ASSUMPTION A. We assume the following properties of A: there is a sequence $\{E_k\}$ of projections on H_1 such that $E_k : H_1 \rightarrow H_{1,k}$, $H_{1,k+1} \supseteq H_{1,k}$, $\lim_{k\to\infty} E_k = I$ in the strong sense, there exists $K > 0$ such that $\|E_k\| \leqq K$ for all k, $A_k = AE_k$ is bounded and extends E_kA, $\lim_{k\to\infty} A_kx = Ax$ for $x \in \Delta$. The adjoint operators A^*, A_k^* and E_k^* have similar properties. This assumption is certainly fulfilled if A is a spectral operator. (A complete definition of this concept is given in [9] and [10].) We note that $e^{At}E_k = e^{A_kt}$ and $E_k^*e^{A^*t} = e^{A_k^*t}$.

Let t_1 be real and let $Q(t_1)$ be given as a bounded self-adjoint operator. Applying to the integral equation (2.8), with T replaced by t_1 and G replaced by $Q(t_1)$, the familiar technique of successive approximations one can show without difficulty that there is a real $t_2 < t_1$ and a positive number K such that in $[t_2, t_1]$ the integral equation has precisely one strongly continuous self-adjoint solution $Q(t)$ with $Q(t_1)$ as specified. Moreover, $Q(t)$ is uniformly bounded, i.e., $\|Q(t)\| \leqq K$, $t \in [t_2, t_1]$. In particular, this is true for $t_1 = T$ and $Q(t_1) = G$. This is a local result, however, in that $t_1 - t_2$ must be chosen sufficiently small. One of our main tasks is to obtain a global solution of the integral equation. First we prove a lemma.

LEMMA 1. *Assume $Q(t)$ is the unique strongly continuous solution of* (2.8) *on* $[t_\alpha, T]$ *and there satisfies* $\|Q(t)\| \leqq K$. *Let $y(t)$ be a solution of* (2.3) *on* $[t_\alpha, T]$ *with $u(t)$ an admissible control. Then for* $t \in [t_\alpha, T]$,

$$(y(t), Q(t)y(t)) - (y(T), Gy(T)) - \int_t^T [(y(s), Wy(s)) + (u(s), Uu(s))]\,ds \tag{3.1}$$
$$= -\int_t^T (u(s) + U^{-1}B^*Q(s)y(s), U[u(s) + U^{-1}B^*Q(s)y(s)])\,ds.$$

Proof. Let us put

$$Q_k(t) = E_k^* Q(t) E_k. \tag{3.2}$$

We note that $Q_k(t)$ converges strongly to $Q(t)$ as $k \to \infty$ and satisfies

$$\begin{aligned} Q_k(t) &= e^{A_k^*(T-t)} G_k e^{A_k(T-t)} \\ &\quad - \int_t^T e^{A_k^*(s-t)} [-W_k + E_k^* Q(s)(BU^{-1}B^*)Q(s)E_k] e^{A_k(s-t)}\, ds, \end{aligned} \tag{3.3}$$

where $W_k = E_k^* W E_k$, $G_k = E_k^* G E_k$. Since A_k, A_k^* are bounded operators, (3.3) can be differentiated in the usual manner. Then putting $y_k(t) = E_k y(t)$, we can compute

$$\begin{aligned} &\frac{d}{dt}\{(y_k(t), Q_k(t)y_k(t)) - \int_t^T [(y_k(s), W_k y_k(s)) + (u(s), Uu(s))]\, ds\} \\ &\quad = (A_k y_k(t) + E_k Bu(t), Q_k(t) y_k(t)) \\ &\qquad + (y_k(t), [-A_k^* Q_k(t) - Q_k(t) A_k - W_k + E_k^* Q(t)(BU^{-1}B^*)Q(t)E_k] y_k(t)) \\ &\qquad + (y_k(t), Q_k(t)[A_k y_k(t) + E_k Bu(t)]) + (y_k(t), W_k y_k(t)) + (u(t), Uu(t)) \\ &\quad = (E_k Bu(t), Q_k(t) y_k(t)) + (y_k(t), Q_k(t) E_k Bu(t)) \\ &\qquad + (y_k(t), E_k^* Q(t)(BU^{-1}B^*)Q(t)E_k y_k(t)) + (u(t), Uu(t)). \end{aligned} \tag{3.4}$$

Since $Q_k(T) = G_k$, it follows that

$$\begin{aligned} &(y_k(t), Q_k(t) y_k(t)) - (y_k(T), G_k y_k(T)) \\ &\qquad - \int_t^T [(y_k(s), W_k y_k(s)) + (u(s), Uu(s))]\, ds \\ &\quad = - \int_t^T [(E_k Bu(s), Q_k(t) y_k(s)) + (y_k(s), Q_k(s) E_k Bu(s)) \\ &\qquad\qquad + (y_k(s), E_k^* Q(s)(BU^{-1}B^*)Q(s)E_k y_k(s)) + (u(s), Uu(s))]\, ds. \end{aligned} \tag{3.5}$$

We then let $k \to \infty$ and use the strong convergence of $Q_k(t)$, $E_k^* Q(t)$ and $Q(t)E_k$ to $Q(t)$, G_k to G, W_k to W, the convergence of $y_k(t)$ to $y(t)$ and the Lebesgue dominated convergence theorem to obtain (3.1). Thus the proof of Lemma 1 is complete.

Lemma 1 is important in its own right and also plays an important part in the proof of the following lemma.

LEMMA 2. *The strongly continuous solution $Q(t)$ of the integral equation* (2.8) *has a unique strongly continuous extension to* $(-\infty, T]$.

Proof. Let $(t_\infty, T] = \bigcup_\alpha (t_\alpha, T]$, taken over all $t_\alpha < T$ for which (2.8) has a unique bounded self-adjoint solution $Q(t)$, strongly continuous on $(t_\alpha, T]$ with $Q(T) = G$. If $t_\infty = -\infty$, there is nothing to prove, so we now suppose that t_∞ is finite and shall show that this leads to a contradiction.

For $t_\infty < \hat{t}_0 \leqq T$ consider the initial value problem

$$\frac{d\hat{y}(t)}{dt} = [A - BU^{-1}B^*Q(t)]\hat{y}(t), \qquad \hat{y}(\hat{t}_0) = \hat{y}_0, \tag{3.6}$$

where $\hat{y}_0$ is an arbitrary element of H_1. The solution $\hat{y}(t)$ exists on $[\hat{t}_0, T]$ and satisfies

$$\hat{y}(t) = e^{A(t-\hat{t}_0)}\hat{y}_0 - \int_{\hat{t}_0}^{t} e^{A(t-s)}BU^{-1}B^*Q(s)\hat{y}(s)\,ds \tag{3.7}$$

and thus is a solution of (3.6) in the "mild" sense. Now $\hat{y}(t)$ is a solution of (2.3) for

$$\hat{u}(t) = -U^{-1}B^*Q(t)\hat{y}(t) \tag{3.8}$$

and hence, from (3.1), we have

$$(\hat{y}_0, Q(\hat{t}_0)\hat{y}_0) = (\hat{y}(T), G\hat{y}(T)) + \int_{\hat{t}_0}^{T} [(\hat{y}(s), W\hat{y}(s)) + (\hat{u}(s), U\hat{u}(s))]\,ds \geqq 0. \tag{3.9}$$

Thus $Q(t)$ is positive semidefinite on $(t_\infty, T]$. Now let

$$\hat{y}(t) = e^{A(t-\hat{t}_0)}\hat{y}_0, \tag{3.10}$$

corresponding to $u(t) \equiv 0$ in (2.3). Then (3.1) yields

$$\begin{aligned}(\hat{y}_0, Q(\hat{t}_0)\hat{y}_0) &= (\hat{y}_0, e^{A^*(T-\hat{t}_0)}Ge^{A(T-\hat{t}_0)}\hat{y}_0)\\ &\quad + \left(\hat{y}_0, \left[\int_{\hat{t}_0}^{T} e^{A^*(s-\hat{t}_0)}We^{A(s-\hat{t}_0)}\,ds\right]\hat{y}_0\right)\\ &\quad - \left(\hat{y}_0, \left[\int_{\hat{t}_0}^{T} e^{A^*(s-\hat{t}_0)}Q(s)BU^{-1}B^*Q(s)e^{A(s-\hat{t}_0)}\,ds\right]\hat{y}_0\right)\\ &\leqq M^2e^{2\mu_0(T-\hat{t}_0)}(\|G\| + (T-\hat{t}_0)\|W\|)\|\hat{y}_0\|^2,\end{aligned} \tag{3.11}$$

and we see that $\|Q(t)\|$ is uniformly bounded above on $(t_\infty, T]$.

Now we prove that $Q(t_\infty)$ may be defined as the strong limit of $Q(t)$ as $t \downarrow t_\infty$. For all $s \leqq T$ let us put

$$R_t(s) = \begin{cases} e^{A^*(s-t)}[-W + Q(s)BU^{-1}B^*Q(s)]e^{A(s-t)}, & s > t,\\ 0, & s \leqq t.\end{cases} \tag{3.12}$$

Then for $t \in (t_\infty, T]$ and $y \in H_1$,

$$Q(t)y = e^{A^*(T-t)}Ge^{A(T-t)}y - \int_{t_\infty}^{T} R_t(s)y\,ds. \tag{3.13}$$

Since A and A^* generate strongly continuous semigroups, $R_t(s)y$ converges to

$R_{t_\infty}(s)y$ as $t \downarrow t_\infty$. Now the boundedness of $\|Q(t)\|$, $\|e^{A(T-t)}\|$ and $\|e^{A^*(T-t)}\|$ on $(t_\infty, T]$ implies

$$(3.14)\qquad \|R_t(s)y\| \leqq r\|y\|, \qquad t \in (t_\infty, T],$$

for some fixed $r \geqq 0$. The Lebesgue bounded convergence theorem for vector-valued functions (see, e.g., [6]) then shows that

$$(3.15)\qquad \lim_{t \downarrow t_\infty} \int_{t_\infty}^{T} R_t(s)y\,ds = \int_{t_\infty}^{T} R_{t_\infty}(s)y\,ds$$

and, defining $Q(t_\infty)$ from (3.13) with $t = t_\infty$, we have

$$(3.16)\qquad \lim_{t \downarrow t_\infty} Q(t)y = Q(t_\infty)y, \qquad y \in H_1,$$

and

$$(3.17)\qquad \|Q(t_\infty)\| \leqq M^2 e^{2\mu_0(T-t_\infty)}\|G\| + (T - t_\infty)r.$$

Since $Q(t_\infty)$ satisfies the conditions originally placed on G, we can replace G by $Q(t_\infty)$ and T by t_∞ in (2.8) and the resulting equation will have a solution on a small interval $[\hat{t}_1, t_\infty]$. It is a simple matter to verify that this provides a strongly continuous solution $Q(t)$ on $[\hat{t}_1, T]$, thus contradicting the definition of t_∞. It follows that $t_\infty = -\infty$ and the proof is complete.

We are now able to characterize the optimal control $u_*(t)$.

THEOREM 1. *Let $Q(t)$ be the solution of* (2.8) *for $t_0 \leqq t \leqq T$. Let $x_0 \in \Delta$ and the cost functional $C(u)$ be given as in the problem statement of* § 1. *Then the control law*

$$(3.18)\qquad u_*(t) = -U^{-1}B^*Q(t)y_*(t) = -U^{-1}\tilde{B}^*(A^* - \lambda I)Q(t)(A - \lambda I)x_*(t),$$

which determines $x_*(t)$ as the solution of

$$(3.19)\qquad \frac{dx_*(t)}{dt} = [A - \tilde{B}U^{-1}\tilde{B}^*(A^* - \lambda I)Q(t)(A - \lambda I)]x_*(t), \qquad x_*(t_0) = x_0,$$

yields the unique optimal control $u_*(t)$ minimizing $C(u)$. Moreover, for each such x_0 and $y_0 = (A - \lambda I)x_0$, we have

$$(3.20)\qquad c_{x_0} = C(u_*) = (y_0, Q(t_0)y_0) = (x_0, (A^* - \lambda I)Q(t_0)(A - \lambda I)x_0).$$

Proof. For any admissible control $u(t)$ on $[t_0, T]$ we have, by Lemma 1,

$$(3.21)\qquad \begin{aligned} C(u) &= (y_0, Q(t_0)y_0) \\ &\quad + \int_{t_0}^{T} (u(s) + U^{-1}B^*Q(s)y(s), U[u(s) + U^{-1}B^*Q(s)y(s)])\,ds \end{aligned}$$

and, since U is positive definite, we have

$$(3.22)\qquad C(u) > (y_0, Q(t_0)y_0), \qquad u(t) \not\equiv -U^{-1}B^*Q(t)y(t),$$

while (3.20) holds if u_* is given by (3.18).

Remark. Since solutions $x(t)$ of (1.2) lie in Δ and $(A - \lambda I)\tilde{B}$ is bounded, the applicability of (3.18) and (3.20) is not affected by the unboundedness of A.

4. Optimal control on the infinite interval. Let us replace T by $+\infty$ and the cost functional (1.5) by

$$
\begin{aligned}
C(u) &= \int_{t_0}^{\infty} [(x(t), \tilde{W}x(t)) + (u(t), Uu(t))]\, dt \\
&= \int_{t_0}^{\infty} [(y(t), Wy(t)) + (u(t), Uu(t))]\, dt
\end{aligned}
\tag{4.1}
$$

and otherwise pose the same problem as in § 1.

Definition. We say that the system (1.2) is *optimizable relative to* $\tilde{W}$ if there is a constant $K_0 > 0$ and a bounded operator D such that solutions $y(t)$ of

$$y(t) = e^{A(t-t_0)}y_0 + \int_{t_0}^{t} e^{A(t-s)}BDy(s)\, ds, \tag{4.2}$$

corresponding to controls

$$u(t) = Dy(t) = D(A - \lambda I)x(t), \tag{4.3}$$

yield values of $C(u)$, as given by (4.1), satisfying

$$C(u) \leqq K_0\|y_0\|^2, \qquad y_0 \in H_1. \tag{4.4}$$

We note that for $u(t)$ given by (4.3), $C(u)$ will be independent of t_0. Hence the choice of t_0 is irrelevant in the definition and we take K_0 to be independent of t_0.

In the case of finite-dimensional systems it has been shown that complete controllability implies stabilizability which in turn implies optimizability. That no such simple relationship holds for the systems we are interested in will be made clear in § 5.

For each T, let $Q_T(t)$, $t \leqq T$, denote the solution of (2.8) with $G = 0$ so that

$$Q_T(T) = 0. \tag{4.5}$$

We can then prove the following lemma.

Lemma 3. *If the system* (1.2) *is optimizable relative to* $\tilde{W}$, *then there is a bounded self-adjoint positive semidefinite operator* Q_∞ *defined on* H_1 *such that* $Q_T(t)$ *converges strongly to* Q_∞ *for each fixed* t *as* $T \to \infty$.

Proof. Let $T \geqq t$ first be fixed as in § 3. Since $(y_0, Q_T(t)y_0)$ is the minimum value of the cost (2.4) with $t_0 = t$, $G = 0$, the facts that A, B, W and U are nontime-varying and W and U are nonnegative imply that

$$K_0\|y_0\|^2 \geqq (y_0, Q_T(t - \tau)y_0) \geqq (y_0, Q_T(t)y_0) \tag{4.6}$$

for all $\tau \geqq 0$. Since the integral equation for $Q(t)$ is autonomous,

$$Q_T(t - \tau) = Q_{T+\tau}(t), \tag{4.7}$$

and hence, for each fixed $t \leqq T$,

$$K_0\|y_0\|^2 \geqq (y_0, Q_{T+\tau}(t)y_0) \geqq (y_0, Q_T(t)y_0) \tag{4.8}$$

so that $Q_T(t)$ is self-adjoint, monotone increasing with T and bounded above. It is well known (see e.g., [7]) that this implies the existence of $Q_\infty(t)$ as $T \to \infty$. Now $Q_\infty(t) = Q_\infty$ is independent of t because

$$\lim_{T\to\infty} Q_T(t) = \lim_{T\to\infty} Q_{T+t}(0) = Q_\infty \tag{4.9}$$

in the strong sense. This completes the proof. For later reference we note that (4.8) implies

$$\|Q_T(t)\| \leqq K_0 \quad \text{for all finite } t, T, \quad t \leqq T. \tag{4.10}$$

THEOREM 2. *If* (1.2) *is optimizable relative to* $\tilde{W}$, *then the control law*

$$\begin{aligned} u_\infty(t) &= -U^{-1}B^*Q_\infty y_\infty(t) \\ &= -U^{-1}\tilde{B}^*(A^* - \lambda I)Q_\infty(A - \lambda I)x_\infty(t), \end{aligned} \tag{4.11}$$

which determines $x_\infty(t)$ *as a solution of*

$$\begin{aligned} \frac{dx_\infty(t)}{dt} &= [A - \tilde{B}U^{-1}\tilde{B}^*(A^* - \lambda I)Q_\infty(A - \lambda I)]x_\infty(t), \\ x_\infty(t_0) &= x_0 \in \Delta, \end{aligned} \tag{4.12}$$

yields the unique (*synthesized*) *optimal control* u_∞ *minimizing* $C(u)$ *as given by* (4.1). *Moreover, for* $y_0 = (A - \lambda I)x_0$,

$$C(u_\infty) = (y_0, Q_\infty y_0) = (x_0, (A^* - \lambda I)Q_\infty(A - \lambda I)x_0). \tag{4.13}$$

Proof. For fixed $y_0 \in H_1$ let us denote by $y_\infty(t)$ the unique solution of

$$y_\infty(t) = e^{A(t-t_0)}y_0 - \int_{t_0}^t e^{A(t-s)}BU^{-1}B^*Q_\infty y_\infty(s)\,ds \tag{4.14}$$

and by $y_T(t)$ the unique solution of

$$y_T(t) = e^{A(t-t_0)}y_0 - \int_{t_0}^t e^{A(t-s)}BU^{-1}B^*Q_T(s)y_T(s)\,ds. \tag{4.15}$$

Then

$$\begin{aligned} y_\infty(t) - y_T(t) &= \int_{t_0}^t e^{A(t-s)}BU^{-1}B^*[Q_T(s) - Q_\infty]y_\infty(s)\,ds \\ &\quad + \int_{t_0}^t e^{A(t-s)}BU^{-1}B^*Q_T(s)[y_T(s) - y_\infty(s)]\,ds. \end{aligned} \tag{4.16}$$

Using (1.1) and (4.10) we have, for $\alpha, \beta > 0$,

$$\|y_\infty(t) - y_T(t)\| \leqq \alpha\eta_T(t) + \beta\int_{t_0}^t \|y_\infty(s) - y_T(s)\|\,ds, \tag{4.17}$$

where

$$\eta_T(t) = \int_{t_0}^t \|(Q_T(s) - Q_\infty)y_\infty(s)\|\,ds. \tag{4.18}$$

By virtue of the strong convergence of $Q_T(s)$ to Q_∞ and (4.10) we are able to apply the Lebesgue bounded convergence theorem to conclude

$$\lim_{T\to\infty} \eta_T(t) = 0. \tag{4.19}$$

Furthermore, the convergence is uniform on bounded intervals since $\eta_T(t)$ is monotone increasing in t. A well-known result in ordinary differential equations (see, e.g., [11, exercise, p. 37]) shows that

$$\|y_\infty(t) - y_T(t)\| \leqq \alpha\eta_T(t) + \int_{t_0}^{t} \alpha\beta e^{\beta(t-s)}\eta_T(s)\,ds. \tag{4.20}$$

Then (4.19) and (4.20) imply that

$$\lim_{T\to\infty} \|y_\infty(t) - y_T(t)\| = 0 \tag{4.21}$$

uniformly on bounded intervals.

Let us now define controls

$$u_T(t) = -U^{-1}B^*Q_T(t)y_T(t), \tag{4.22}$$

$$u_\infty(t) = -U^{-1}B^*Q_\infty y_\infty(t). \tag{4.23}$$

The optimizability of the system (1.2) relative to $\tilde{W}$ implies the existence of controls yielding costs uniformly bounded with respect to $\|y_0\|^2$ (cf. (4.4)). Let $u(t)$ be any admissible control yielding finite cost. We shall show that this cost is greater than or equal to the cost of the control (4.23). Let $y(t)$ be the response to $u(t)$ via (2.3). From the work of § 3 we have

$$\begin{aligned}(y_0, Q_T(t_0)y_0) &= \int_{t_0}^{T} [(y_T(s), Wy_T(s)) + (u_T(s), Uu_T(s))]\,ds \\ &\leqq \int_{t_0}^{T} [(y(s), Wy(s)) + (u(s), Uu(s))]\,ds \\ &\leqq \int_{t_0}^{\infty} [(y(s), Wy(s)) + (u(s), Uu(s))]\,ds < \infty.\end{aligned} \tag{4.24}$$

Now Lemma 3 implies that

$$\lim_{T\to\infty} (y_0, Q_T(t_0)y_0) = (y_0, Q_\infty y_0). \tag{4.25}$$

Thus, if we can show that

$$\begin{aligned}\lim_{T\to\infty} &\int_{t_0}^{T} [(y_T(s), Wy_T(s)) + (u_T(s), Uu_T(s))]\,ds \\ &= \int_{t_0}^{\infty} [(y_\infty(s), Wy_\infty(s)) + (u_\infty(s), Uu_\infty(s))]\,ds,\end{aligned} \tag{4.26}$$

(4.24) will yield both the optimality of u_∞ and the equality (4.13).

From the optimality of $u_T(t)$ on $[t_0, T]$,

$$\begin{aligned}\int_{t_0}^{T} [(y_T(s), Wy_T(s)) + (u_T(s), Uu_T(s))]\,ds \\ \leqq \int_{t_0}^{T} [(y_\infty(s), Wy_\infty(s)) + (u_\infty(s), Uu_\infty(s))]\,ds\end{aligned} \tag{4.27}$$

and hence

$$\begin{aligned}\lim_{T\to\infty} \int_{t_0}^{T} [(y_T(s), Wy_T(s)) + (u_T(s), Uu_T(s))]\,ds \\ \leqq \int_{t_0}^{\infty} [(y_\infty(s), Wy_\infty(s)) + (u_\infty(s), Uu_\infty(s))]\,ds.\end{aligned} \tag{4.28}$$

All that remains in the proof of (4.26) is to establish the reverse inequality.

For $t_0 \leqq t_1 \leqq T$ we clearly have

$$\begin{aligned}\int_{t_0}^{T} [(y_T(s), Wy_T(s)) + (u_T(s), Uu_T(s))]\,ds \\ \geqq \int_{t_0}^{t_1} [(y_T(s), Wy_T(s)) + (u_T(s), Uu_T(s))]\,ds.\end{aligned} \tag{4.29}$$

For t_1 fixed it follows directly from (4.9), (4.10), (4.21), (4.22), (4.23) and the Lebesgue bounded convergence theorem that

$$\begin{aligned}\lim_{T\to\infty} \int_{t_0}^{T} [(y_T(s), Wy_T(s)) + (u_T(s), Uu_T(s))]\,ds \\ \geqq \int_{t_0}^{t_1} [(y_\infty(s), Wy_\infty(s)) + (u_\infty(s), Uu_\infty(s))]\,ds.\end{aligned} \tag{4.30}$$

Now letting $t_1 \to \infty$ in (4.30) and combining the resulting inequality with (4.28) we have (4.26).

Thus the optimality of u_∞ and the equality (4.13) have been established. All that remains is to show that u_∞ is the *unique* optimal control. For any $y_0 \in H_1$, let $\tilde{u}$ be an admissible control, with response $\tilde{y}$, yielding finite cost $C(\tilde{u})$, and differing from u_∞ on a set of positive measure. Then the set of real numbers

$$\{t | t \geqq t_0, \tilde{u}(t) + U^{-1}B^*Q_\infty\tilde{y}(t) \not\equiv 0\} \tag{4.31}$$

has positive measure. If this were not the case $y_\infty(t)$ and $\tilde{y}(t)$ would both be solutions of

$$y(t) = e^{A(t-t_0)}y_0 - \int_{t_0}^{t} e^{A(t-s)}BU^{-1}B^*Q_\infty y(s)\,ds \tag{4.32}$$

and hence, by uniqueness, would coincide. Then we would also have $\tilde{u}(t) = u_\infty(t)$, a.e., contrary to our assumption on $\tilde{u}$. From the fact that the set (4.31) has positive measure it follows that, for sufficiently large T,

$$\int_{t_0}^{T} (\tilde{u}(s) + U^{-1}B^*Q_\infty\tilde{y}(s), U[\tilde{u}(s) + U^{-1}B^*Q_\infty\tilde{y}(s)])\,ds > d > 0. \tag{4.33}$$

Taking $u(t) = \tilde{u}(t)$, $y(t) = \tilde{y}(t)$, $Q(t) = Q_T(t)$ in (3.21) and letting $T \to \infty$ we see that

$$\begin{aligned} C(\tilde{u}) &= (y_0, Q_\infty y_0) \\ &\quad + \int_0^\infty (\tilde{u}(s) + U^{-1}B^*Q_\infty\tilde{y}(s), U[\tilde{u}(s) + U^{-1}B^*Q_\infty\tilde{y}(s)])\,ds \\ &\geqq C(u_\infty) + d \end{aligned} \tag{4.34}$$

so that $\tilde{u}$ is not optimal. Thus u_∞ is the unique optimal control and Theorem 2 has been proved.

We conclude this section with a theorem showing that Q_∞ obeys the analogue of the Kalman matrix equation.

THEOREM 3. *If the operator $A^*Q_\infty + Q_\infty A$ is defined on the domain*

$$\{y \in H_1 | y \in \Delta = \operatorname{dom} A, Q_\infty y \in \Delta^* = \operatorname{dom} A^*\}, \tag{4.35}$$

then Q_∞ satisfies the Kalman operator equation

$$A^*Q_\infty + Q_\infty A + W - Q_\infty BU^{-1}B^*Q_\infty = 0 \tag{4.36}$$

in the sense that $-W + Q_\infty BU^{-1}B^*Q_\infty$ is an extension of $A^*Q_\infty + Q_\infty A$ as defined on (4.35) *to H_1.*

Remark. Thus $A^*Q_\infty + Q_\infty A$ is bounded on (4.35) and has a bounded extension to all of H_1. This does not imply that either A^*Q_∞ or $Q_\infty A$ is bounded, however.

Proof. As in the proof of Lemma 1, § 3, set

$$Q_{\infty,k} = E_k^*Q_\infty E_k, \qquad Q_{T,k}(t) = E_k^*Q_T(t)E_k. \tag{4.37}$$

Then $Q_{T,k}(t)$ satisfies the differential equation

$$-\frac{dQ_{T,k}}{dt} = A_k^*Q_{T,k} + Q_{T,k}A_k + W_k - E_k^*Q_TBU^{-1}B^*Q_TE_k \tag{4.38}$$

for $t \leqq T$ and the terminal condition $Q_{T,k}(T) = 0$. Since $Q_T(t)$ converges strongly to Q_∞ as $t \to -\infty$, we conclude that the right-hand side of (4.38) converges strongly to

$$A_k^*Q_{\infty,k} + Q_{\infty,k}A_k + W_k - E_k^*Q_\infty BU^{-1}B^*Q_\infty E_k. \tag{4.39}$$

If this limit were different from zero we could use (4.38) to show that $Q_{T,k}(t)$, and hence $Q_T(t)$, does not converge strongly as $t \to \infty$, a contradiction. We conclude that, for all k,

$$A_k^*Q_{\infty,k} + Q_{\infty,k}A_k = -W + E_k^*Q_\infty BU^{-1}B^*Q_\infty E_k. \tag{4.40}$$

As $k \to \infty$, the right-hand side of (4.40) converges strongly to $-W + Q_\infty BU^{-1}B^*Q_\infty$.

The proof of the theorem will be complete if we can show that, whenever y lies in the domain (4.35),

$$\lim_{k\to\infty} (A_k^*Q_{\infty,k} + Q_{\infty,k}A_k)y = (A^*Q_\infty + Q_\infty A)y. \tag{4.41}$$

First we compute

$$Q_\infty Ay - Q_{\infty,k}A_ky = (I - E_k^*)Q_\infty Ay + E_k^*Q_\infty A(I - E_k)y. \tag{4.42}$$

Since E_k^* converges strongly to I, $\lim_{k\to\infty} (I - E_k^*)Q_\infty Ay = 0$. On the other hand, since $y \in \Delta = \operatorname{dom} A$, Assumption A of § 3 implies that $A(I - E_k)y$ converges to zero as $k \to \infty$. Since $E_k^*Q_\infty$ is uniformly bounded, we conclude $\lim_{k\to\infty} E_k^*Q_\infty A(I - E_k)y = 0$ and hence

$$\lim_{k\to\infty} Q_{\infty,k}A_ky = Q_\infty Ay, \qquad y \in \Delta. \tag{4.43}$$

We next wish to show that, for y in (4.35),

$$\lim_{k\to\infty} A_k^*Q_{\infty,k}y = A^*Q_\infty y. \tag{4.44}$$

Now

$$\lim_{k\to\infty} Q_{\infty,k}y = \lim_{k\to\infty} E_k^*Q_\infty E_ky = Q_\infty y \tag{4.45}$$

since $E_k^*Q_\infty E_k$ converges strongly to Q_∞. On the other hand, since we have already proved (4.43), all terms except $A_k^*Q_{\infty,k}$ in (4.40) converge, when applied to the vector y, as $k \to \infty$. It follows that

$$\lim_{k\to\infty} A_k^*Q_{\infty,k}y = \lim_{k\to\infty} A^*Q_{\infty,k}y \tag{4.46}$$

exists. Combining (4.45) and (4.46) and using the well-known fact that A^* is a closed operator, we conclude that

$$\lim_{k\to\infty} A_k^*Q_{\infty,k}y = A^*Q_\infty y. \tag{4.47}$$

Putting (4.43) with (4.47) we have (4.41). The proof of the theorem is complete.

5. Applications and remarks. We shall now show that the theory developed in the preceding sections can be expected to have a reasonably wide range of application.

Consider a "linear oscillator" in a Hilbert space H:

$$\frac{d^2w}{dt^2} + Tw = \hat{B}u, \tag{5.1}$$

with control u lying in a second Hilbert space H_2. We assume that $\hat{B}$ is a bounded operator, $\hat{B}: H_2 \to H$, while T is a positive definite self-adjoint operator, possibly unbounded, defined on a domain Δ dense in H. The energy of such an oscillator is given by

$$E\left(w, \frac{dw}{dt}\right) = \frac{1}{2}(T^{1/2}w, T^{1/2}w) + \frac{1}{2}\left(\frac{dw}{dt}, \frac{dw}{dt}\right), \tag{5.2}$$

where $T^{1/2}$ denotes the positive square root of T. There are many familiar examples. If we take, for $\xi \in R^1$,

$$(Tw)(\xi) = (-1)^n \frac{d^n}{d\xi^n}\left(p(\xi)\frac{d^n w}{d\xi^n}\right) \tag{5.3}$$

and impose, via specification of boundary and smoothness conditions, appropriate requirements on Δ, T will be self-adjoint and strictly positive. The cases $n = 1$ and $n = 2$ correspond to the string and simple beam, respectively. In R^2 one can consider the membrane

$$(Tw)(\xi, \eta) = -\frac{\partial^2 w}{\partial \xi^2} - \frac{\partial^2 w}{\partial \eta^2} \tag{5.4}$$

or the plate

$$(Tw)(\xi, \eta) = \left\{\frac{\partial^2}{\partial \xi^2} + \frac{\partial^2}{\partial \eta^2}\right\}\left(\frac{\partial^2 w}{\partial \xi^2} + \frac{\partial^2 w}{\partial \eta^2}\right), \tag{5.5}$$

again with appropriate boundary and smoothness conditions. There are many other examples.

Setting

$$w = w_1, \qquad \frac{dw}{dt} = w_2, \tag{5.6}$$

we have a first order system in $H_1 = H \oplus H$:

$$\frac{d}{dt}\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} 0 & I \\ -T & 0 \end{pmatrix}\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} + \begin{pmatrix} 0 \\ \hat{B} \end{pmatrix} u. \tag{5.7}$$

If we now put

$$\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} I & I \\ iT^{1/2} & -iT^{1/2} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \tag{5.8}$$

we obtain, in place of (5.7),

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} iT^{1/2} & 0 \\ 0 & -iT^{1/2} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} \frac{-i}{2}T^{-1/2}\hat{B} \\ \frac{i}{2}T^{-1/2}\hat{B} \end{pmatrix} u, \tag{5.9}$$

and the energy is now

$$E(x_1, x_2) = (T^{1/2}x_1, T^{1/2}x_1) + (T^{1/2}x_2, T^{1/2}x_2). \tag{5.10}$$

If we set $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ and

$$A = \begin{pmatrix} iT^{1/2} & 0 \\ 0 & -iT^{1/2} \end{pmatrix}, \tag{5.11}$$

$$\tilde{B} = \begin{pmatrix} -\frac{i}{2}T^{-1/2}\hat{B} \\ \frac{i}{2}T^{-1/2}\hat{B} \end{pmatrix}, \tag{5.12}$$

$$\tilde{W} = \begin{pmatrix} T & 0 \\ 0 & T \end{pmatrix}, \quad \tilde{G} = \begin{pmatrix} T & 0 \\ 0 & T \end{pmatrix} \tag{5.13}$$

and take U to be any bounded, self-adjoint positive definite operator on H_2, then we have a system satisfying all of the conditions laid down in § 1. Moreover, since A is normal in this case, Assumption A of § 3 is certainly valid.

Theorem 1 then implies that on any finite interval $[0, T]$ there is a unique linear control for the linear oscillator (5.1) which minimizes

$$\int_0^T \left[E\left(w(t), \frac{dw}{dt} \right) + (u(t), Uu(t)) \right] dt + E\left(w(T), \frac{dw}{dt}(T) \right), \tag{5.14}$$

a cost functional involving the energy of the system during the interval $[0, T]$ and also the final energy at time T. This is true regardless of the nature of the control space H_2 and thus applies to control via finite-dimensional forces as well as infinite-dimensional (i.e., distributed) forces.

The situation on the infinite interval is more complicated. If we wish to use the cost functional

$$\int_0^\infty \left[E\left(w(t), \frac{dw}{dt} \right) + (u(t), Uu(t)) \right] dt, \tag{5.15}$$

there is some difficulty in obtaining any control yielding bounded cost. If $H_2 = H$ and $\hat{B}$ is nonsingular, we can set

$$u = -k\hat{B}^{-1}\frac{dw}{dt}, \qquad k > 0. \tag{5.16}$$

The resulting closed-loop system

$$\frac{d^2w}{dt^2} + k\frac{dw}{dt} + Tw = 0 \tag{5.17}$$

has solutions with a uniform exponential rate of decay. In this case the cost (5.15) can be uniformly bounded in terms of the initial energy and the results of § 4 apply to show that there is an optimal control relative to the cost functional (5.15). Now if the original system involves damping, as is the case in practically all physical problems, we could replace the left-hand side of (5.1) by the left-hand side of (5.17).

Then, regardless of the nature of $\hat{B}$ or the control space H_2, there will be an optimal control relative to the cost functional (5.15), for the control $u = 0$ yields bounded cost.

Frequently one wishes to consider an undamped oscillator with scalar control u. The pertinent equation is then

$$\frac{d^2w}{dt^2} + Tw = gu, \qquad g \in H. \tag{5.18}$$

It turns out that for the system (5.18) there is, in general, no control yielding a cost which is uniformly bounded relative to the initial energy (which is the norm of y after the transformation of § 2 is applied to (5.9)) and hence (5.15) cannot be used. However, it has been shown in [12] that, if appropriate assumptions are made concerning the expansion of the vector g in terms of the eigenvectors of T, then there is a control policy, namely $u = -\varepsilon(dw/dt, g)$, which with cost functional

$$\int_0^\infty \left[(w(t), w(t)) + \left(T^{-1/2}\frac{dw}{dt}, T^{-1/2}\frac{dw}{dt}\right) + u(t)^2\right] dt \tag{5.19}$$

yields a cost uniformly bounded with respect to the initial energy, and thus the existence of an optimal control relative to (5.19) is established.

Our theory also applies to the "heat equation"

$$\frac{dw}{dt} + Tw = \hat{B}u, \tag{5.20}$$

with the same assumptions on T as above. Appropriate cost functionals would involve (w, w), for example. In this case the preliminary transformation of § 2 is not necessary.

The theory which we have developed would be of little use in practice unless there is a feasible way to calculate the operators $Q(t)$ and Q_∞. Fortunately such a procedure does exist in many cases as we shall see.

For many of the operators A arising in physical problems the spectrum consists of a countable sequence $\{\lambda_n\}$ of eigenvalues, and the corresponding sequence $\{\phi_n\}$ of normalized eigenvectors forms a basis for the state space H_1; in fact the $\{\phi_n\}$ are usually a complete orthonormal set. It is then natural to take E_k to be the projection from H_1 onto the span of $\phi_1, \phi_2, \cdots, \phi_k$ with null space equal to the span of $\phi_{k+1}, \phi_{k+2}, \cdots$ and to define E_k^* similarly in terms of the unique biorthogonal sequence for $\{\phi_n\}$. One may then consider finite-dimensional approximations to the system (2.3),

$$\frac{dy_k}{dt} = A_k y_k + E_k B u, \tag{5.21}$$

where $y_k = E_k y$, and a correspondingly truncated cost functional

$$\begin{aligned} C_k(u) = \int_{t_0}^T & [(y_k(t), W_k y_k(t)) + (u(t), Uu(t))] \, dt \\ & + (y_k(T), G_k y_k(T)), \end{aligned} \tag{5.22}$$

where

$$W_k = E_k^* W E_k, \qquad G_k = E_k^* G E_k. \tag{5.23}$$

For the truncated problem (5.21), (5.22) one obtains the solution

$$u_k(t) = -U^{-1}B^*E_k^*\hat{Q}_k(t)y_k(t), \tag{5.24}$$

where $\hat{Q}_k(t)$ satisfies

$$\begin{aligned}\hat{Q}_k(t) &= e^{A_k^*(T-t)}G_k e^{A_k(T-t)}\\ &\quad - \int_t^T e^{A_k^*(s-t)}[-W_k + \hat{Q}_k(s)E_kBU^{-1}B^*E_k^*\hat{Q}_k(s)]e^{A_k(s-t)}\,ds.\end{aligned} \tag{5.25}$$

In fact, since all operators here are bounded, $\hat{Q}_k$ satisfies

$$-\frac{d\hat{Q}_k}{dt} = A_k^*\hat{Q}_k + \hat{Q}_kA_k + W_k - \hat{Q}_kE_kBU^{-1}B^*E_k^*\hat{Q}_k, \qquad \hat{Q}_k(T) = G_k. \tag{5.26}$$

Now $\hat{Q}_k(t)$ can be calculated using the known techniques available in the finite-dimensional case. If in some sense we have $\lim_{k\to\infty}\hat{Q}_k(t) = Q(t)$, then repeated calculations of $\hat{Q}_k(t)$, for larger and larger values of k, will yield increasingly good approximations to $Q(t)$. In general it is not immediately clear that this is true, but the theorems presented below cover many, if not most, of the situations which arise in practice.

Our first theorem treats the case where A is a normal operator and the operators W and G both commute with A (and hence with A^* and the projections $E_k = E_k^*$). The example of the linear oscillator (5.1) with cost functionals (5.14), (5.15) fits into this rather restrictive category of problems as will (5.20) with a cost functional of the type suggested there.

THEOREM 4. *If W and G both commute with the normal operator A, then we have*

$$(y_0, \hat{Q}_{k-1}(t)y_0) \leqq (y_0, \hat{Q}_k(t)y_0) \leqq (y_0, Q(t)y_0), \quad t \leqq T, \quad y_0 \in H_1. \tag{5.27}$$

For each $y_0 \in H_1$ and finite $t_1 < T$,

$$\|\hat{Q}_k(t)y_0 - Q(t)y_0\| \to 0 \quad \text{as} \quad k \to \infty \tag{5.28}$$

uniformly for $t \in [t_1, T]$. Consequently, for $y_0 \in H_1$,

$$(y_0, \hat{Q}_k(t)y_0) \to (y_0, Q(t)y_0) \quad \text{as} \quad k \to \infty \tag{5.29}$$

uniformly for $t \in [t_1, T]$. If the system (1.2) *is optimizable relative to $\tilde{W}$, we take $G = 0$ and then* (5.28), (5.29) *are valid uniformly for $-\infty < t \leqq T$ and*

$$\lim_{k\to\infty} \hat{Q}_{k,\infty} = Q_\infty. \tag{5.30}$$

Proof. We have, for each $t_0 \leqq T$,

$$(y_0, Q(t_0)y_0) = C(u_*), \tag{5.31}$$

where u_* is the optimal control on $[t_0, T]$ corresponding to the initial condition

$y(t_0) = y_0$ and the cost functional (2.4). Thus,

$$(5.32)\qquad (y_0, Q(t_0)y_0) = \int_{t_0}^{T} [(y_*(t), Wy_*(t)) + (u_*(t), Uu_*(t))]\,dt + (y_*(T), Gy_*(T)).$$

Let the same control u_* be used now for the truncated system (5.21) with initial condition $y_k(t_0) = E_k y_0$ and cost functional (5.22). The resulting response is $y_k(t) = E_k y_*(t)$ with cost

$$(5.33)\qquad \begin{aligned} C_k(u_*) &= \int_{t_0}^{T} [(y_k(t), W_k y_k(t)) + (u_*(t), Uu_*(t))]\,dt \\ &\quad + (y_k(T), G_k y_k(T)) \\ &= \int_{t_0}^{T} [(y_*(t), E_k W E_k y_*(t)) + (u_*(t), Uu_*(t))]\,dt \\ &\quad + (y_*(T), E_k G E_k y_*(T)) \\ &\leqq C(u_*) = (y_0, Q(t_0)y_0). \end{aligned}$$

The last inequality in (5.33) follows from

$$(5.34)\qquad \begin{aligned} &(y, E_k W E_k y) \leqq (y, Wy), \\ &(y, E_k G E_k y) \leqq (y, Gy), \qquad y \in H_1, \end{aligned}$$

which is an obvious consequence of the fact that W and G commute with the orthogonal projections E_k. Then since

$$(5.35)\qquad (E_k y_0, \hat{Q}_k(t_0)E_k y_0) = (y_0, \hat{Q}_k(t_0)y_0)$$

is the optimal cost for the truncated problem, the second inequality in (5.27) follows. An entirely analogous argument proves the first inequality.

Since for each t the $\hat{Q}_k(t)$ are increasing and bounded above by $Q(t)$, there is a nonnegative self-adjoint $\hat{Q}(t)$ such that $\hat{Q}_k(t)$ converges strongly to $\hat{Q}(t)$ for each t. An application of the Lebesgue bounded convergence theorem shows that $\hat{Q}(t)$ solves (2.8) and thus, by uniqueness, $\hat{Q}(t) = Q(t)$. Thus the convergences (5.28) and (5.29) follow. The fact that the convergence (5.29) is uniform on compact intervals follows from (5.27) and Dini's theorem. From Schwarz's inequality for bilinear forms and (5.27) we have

$$(5.36)\qquad \begin{aligned} \|[\hat{Q}_k(t) - Q(t)]y_0\|^4 &= (y_0, [Q(t) - \hat{Q}_k(t)]^2 y_0)^2 \\ &\leqq (y_0, [Q(t) - \hat{Q}_k(t)]y_0)([Q(t) - \hat{Q}_k(t)]y_0, [Q(t) - \hat{Q}_k(t)]^2 y_0) \\ &\leqq (y_0, [Q(t) - \hat{Q}_k(t)]y_0)(2\|Q(t)\|)^3 \|y_0\|^2. \end{aligned}$$

The boundedness of $\|Q(t)\|$ on compact subintervals of $(-\infty, T]$ follows from (3.11), hence the uniform convergence (5.29) applied to (5.36) proves the uniform convergence (5.28).

Now if (1.2) is optimizable relative to $\tilde{W}$, then Q_∞ exists; moreover, (4.6) and (5.27) imply that $\hat{Q}_{k,\infty}$ also exists and

$$(y_0, \hat{Q}_{k-1,\infty}y_0) \leqq (y_0, \hat{Q}_{k,\infty}y_0) \leqq (y_0, Q_\infty y_0). \tag{5.37}$$

Then there is a nonnegative self-adjoint Q_∞ such that, for $y \in H_1$,

$$\lim_{k\to\infty} \hat{Q}_{k,\infty}y = \hat{Q}_\infty y. \tag{5.38}$$

Clearly

$$(y, \hat{Q}_\infty y) \leqq (y, Q_\infty y), \qquad y \in H_1. \tag{5.39}$$

Now if there were some $\hat{y} \in H_1$ and $\varepsilon > 0$ such that

$$(\hat{y}, \hat{Q}_\infty \hat{y}) \leqq (\hat{y}, Q_\infty \hat{y}) - \varepsilon, \tag{5.40}$$

we would argue as follows. Choose $t_1 < T$ so that

$$(\hat{y}, \hat{Q}_\infty \hat{y}) \leqq (\hat{y}, Q(t_1)\hat{y}) < \frac{\varepsilon}{2}. \tag{5.41}$$

Then choose k so large that

$$(\hat{y}, Q(t_1)\hat{y}) - (\hat{y}, \hat{Q}_k(t_1)\hat{y}) < \frac{\varepsilon}{2}. \tag{5.42}$$

Then clearly

$$(\hat{y}, \hat{Q}_k(t_1)\hat{y}) > (\hat{y}, \hat{Q}_\infty \hat{y}) \geqq (\hat{y}, \hat{Q}_{k,\infty}\hat{y}), \tag{5.43}$$

which is a contradiction. Thus (5.39) must be an equality for all $y \in H_1$ and, since $Q_\infty - \hat{Q}_\infty$ is self-adjoint, this implies that

$$\hat{Q}_\infty = Q_\infty. \tag{5.44}$$

The fact that (5.28) and (5.29) now hold uniformly for all t follows readily if we compactify $(-\infty, T]$ in the usual manner by adjoining $-\infty$ and extend the functions in (5.28), (5.29) to $-\infty$ by continuity (e.g., $(y_0, Q(-\infty)y_0) = (y_0, Q_\infty y_0)$). The convergence is still monotone and Dini's theorem and then (5.36) along with (4.10) can be applied again.

Recall that the operators $\hat{Q}_k(t)$, Q_∞ not only represent optimal costs but, what is even more important insofar as engineering applications are concerned, they constitute the major part of the feedback synthesis construction. The real significance of (5.30) is that

$$\lim_{\substack{k\to\infty \\ t\to-\infty}} \hat{Q}_k(t) = Q_\infty \tag{5.45}$$

with no restriction regarding the manner in which (k, t) approaches $(\infty, -\infty)$. The following result treats more general cost functionals than Theorem 4, but a severe (from the mathematical, but certainly not the engineering, viewpoint) restriction is imposed upon the controls and (5.45) is replaced by the weaker result

$$\lim_{t \to -\infty} \left(\lim_{k \to \infty} \hat{Q}_k(t) \right) = Q_\infty \tag{5.46}$$

for the case of optimal control on an infinite interval.

Theorem 5. *If the control space H_2 is finite-dimensional we have*

$$\|\hat{Q}_k(t)y - Q(t)y\| \to 0 \quad \textit{as} \quad k \to \infty, \tag{5.47}$$

$t \leqq T, y \in H_1$. If A is normal, then the convergence (5.47) *is uniform on each finite interval $t_1 \leqq t \leqq T$.*

Proof. Let $\tilde{Q}_k(t)$ solve

$$\begin{aligned} \tilde{Q}_k(t) &= e^{A^*(T-t)} G e^{A(T-t)} \\ &\quad - \int_t^T e^{A^*(s-t)} [-W + \tilde{Q}_k(s) E_k B U^{-1} B^* E_k^* \tilde{Q}_k(s)] e^{A(s-t)}\, ds. \end{aligned} \tag{5.48}$$

Then

$$\hat{Q}_k(t) = E_k^* \tilde{Q}_k(t) E_k, \tag{5.49}$$

since both are solutions of (5.25). Now $E_k B : H_2 \to H_1$ and converges strongly as $k \to \infty$ to B. But, since H_2 is finite-dimensional, strong convergence implies convergence in the norm, and so we have

$$\lim_{k \to \infty} \|E_k B - B\| = 0. \tag{5.50}$$

But convergence in norm of a sequence of operators also implies convergence in norm of their adjoints; hence

$$\lim_{k \to \infty} \|B^* E_k^* - B^*\| = 0 \tag{5.51}$$

and thus

$$\lim_{k \to \infty} \|E_k B U^{-1} B^* E_k^* - B U^{-1} B^*\| = 0. \tag{5.52}$$

Comparing (5.48) with (2.8) we see that $Q(t)$ and $\tilde{Q}_k(t)$ satisfy equations differing only in the terms which appear in (5.52). Using (5.52), we can easily prove that

$$\lim_{k \to \infty} \|\tilde{Q}_k(t) - Q(t)\| = 0 \tag{5.53}$$

uniformly on any compact interval $t_1 \leqq t \leqq T$. The strong convergence (5.47) follows from (5.53), Assumption A and the estimate

$$\begin{aligned} \|[Q(t) - \hat{Q}_k(t)]y\| &\leqq \|[I - E_k^*]Q(t)y\| + \|E_k^*\| \, \|Q(t)\| \, \|[I - E_k]y\| \\ &\quad + \|E_k^*\| \, \|E_k\| \, \|Q_k(t) - \tilde{Q}(t)\| \, \|y\| \end{aligned} \tag{5.54}$$

easily attained from (5.49). If A is normal, E_k^* is an orthogonal projection, hence Dini's theorem applies to the second term of (5.54). Then the uniformity of the convergence (5.47) follows easily from (5.54) using the uniformity of the convergence (5.53) and the boundedness of $\|Q(t)\|$ on $[t_1, T]$.

It would seem that the strong convergence (5.28), (5.47) should be true under much more general conditions than those assumed in Theorems 4 and 5. The essential result needed is the strong convergence, as $k \to \infty$, of solutions of the integral equations

$$Q_k(t) = Q_k(t_0) + \int_{t_0}^{t} F_k(Q_k(s))\,ds \tag{5.55}$$

to the solution of

$$Q(t) = Q(t_0) + \int_{t_0}^{t} F(Q(s))\,ds, \tag{5.56}$$

given that $Q_k(t_0)$ converges strongly to $Q(t_0)$ and the coefficients in the expression $F_k(Q)$ converge strongly to the corresponding coefficients in the expression $F(Q)$. Problems of this type are studied separately in [13].

We conclude with a remark relative to the operators $\tilde{W}$ and $\tilde{G}$ described by formulas (1.6) and (1.7). It is quite possible that these operators, as written, might have domain consisting only of the zero element in H_1. This would be true, e.g., if the range of W, or G, were disjoint from the domain of A^*. In such an eventuality $\tilde{W}$ and $\tilde{G}$ need not be defined at all. We merely replace $(x(t), \tilde{W}x(t))$ in (1.5) by $((A - \lambda I)x(t), W(A - \lambda I)x(t))$. This changes nothing in the remainder of the work.

REFERENCES

[1] R. E. Kalman, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana, 5 (1960), pp. 102–119.

[2] W. M. Wonham, *On matrix quadratic equations and matrix Riccati equations*, CDS Tech. Rep. 67–5, Center for Dynamical Systems, Brown University, Providence, 1967.

[3] D. L. Lukes, *Stabilizability and optimal control*, Funkcia. Ekvac., 11 (1968), pp. 39–50.

[4] E. B. Lee and L. Markus, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

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

[6] E. Hille and R. S. Phillips, *Functional Analysis and Semigroups*, Colloquium Publications, vol. 31, American Mathematical Society, Providence, 1957.

[7] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.

[8] A. V. Balakrishnan, *Optimal control problems in Banach spaces*, this Journal, 3 (1965), pp. 152–180.

[9] N. Dunford, *Spectral operators*, Pacific J. Math., 4 (1954), pp. 321–354.

[10] W. G. Bade, *Unbounded spectral operators*, Ibid., 4 (1954), pp. 373–392.

[11] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[12] D. Russell, *Linear stabilization of the linear oscillator in Hilbert space*, J. Math. Anal. Appl., to appear.

[13] ———, *Continuity in the strong topology of operator-valued solutions of non-linear differential equations with an application to optimal control*, this Journal, 7 (1969), pp. 132–140.

SIAM J. CONTROL
Vol. 7, No. 1, February 1969

ON THE TOURS OF A TRAVELING SALESMAN*

KATTA G. MURTY†

Abstract. Adjacency properties of tours on their convex hull are discussed. A rule is given by which it can be tested whether any two tours are adjacent vertices on this convex hull or not. Based on this rule an algorithm is described for generating all the adjacent tours of a given tour.

1. Introduction. The traveling salesman problem is the problem of finding a minimal cost tour covering a set of n cities given the costs of traveling between every possible pair of cities. Here a tour is a path covering all the cities, each city being covered once and only once in the path. A precise mathematical definition of a tour is given later on.

Let us denote the cities by $1, 2, \cdots, n$. We put

$$x_{ij} = \begin{cases} 1 & \text{if in the tour the salesman goes from } i \text{ to } j, \\ 0 & \text{otherwise.} \end{cases}$$

Then the matrix $X = (x_{ij})$, which is a cyclic permutation matrix, represents the tour.

If in a tour the salesman goes from i_1 to i_2, then (i_1, i_2) is called an *arc* or *cell* in that tour.

We use the letters i, j to denote cities.

Let c_{ij} = cost of traveling from i to j, $i \neq j$; $c_{ii} = \alpha$, an arbitrarily chosen very large positive number. Then $C = (c_{ij})$ is the cost matrix for the problem and this is given. Starting from any city, the salesman can choose to go to any of the remaining $n-1$ cities initially. From that city he can go to any of the remaining $n-2$ cities and so on. Thus, the total number of distinct tours is $(n-1)!$. The set of all possible tours is denoted by T and their convex hull·by K_T. We shall use the letters t or s to denote tours.

Given any tour t_1 we describe an algorithm in this paper for generating the adjacent tours of t_1 on the convex polyhedron K_T.

2. Notation. The convex polyhedron K_A is the set of all feasible solutions

$$\begin{aligned} &X = (x_{ij}), && \text{an } n \times n \text{ matrix, where the } x_{ij} \text{ satisfy} \\ &\sum_i x_{ij} = 1, && j = 1, \cdots, n, \\ &\sum_j x_{ij} = 1, && i = 1, \cdots, n, \\ &x_{ij} \geqq 0. && \end{aligned} \tag{1}$$

* Received by the editors November 7, 1967, and in revised form May 6, 1968.

† Department of Industrial Engineering, University of Michigan, Ann Arbor, Michigan 48104. This research was supported by the National Science Foundation under Grant GP-4593 with the University of California at Berkeley.

An extreme point of K_A is called an *assignment.* Every assignment is a permutation matrix, i.e., it is an $n \times n$ matrix with a single nonzero entry equal to 1 in each row and column. We use the letters a, b to denote assignments.

Occasionally it is convenient to denote an assignment by its unit cells, i.e., the cells in the matrix X representing the assignment which have unit entries in them. All the other cells have zero entries, of course. Thus,

$$a = \{(1,j_1), \cdots, (n,j_n)\} \tag{2}$$

is an assignment, where $j_1, \cdots, j_n$ is a permutation of the numbers $1, 2, \cdots, n$.

We also write

$$(r,j_r) \in a$$

which means that in the matrix X representing the assignment a, the entry in the cell (r,j_r) is 1. The same fact is also expressed by saying that (r,j_r) is a cell in the assignment a, or that the assignment a has an allocation in the cell (r,j_r).

For any assignment a we shall denote specifically by $\{a\}$ the set of cells of a, i.e., if a is the assignment given by (2), then

$$\{a\} = \{(1,j_1), \cdots, (n,j_n)\}.$$

A *tour* is an assignment whose cells can be written down as a complete path covering all the cities and then returning to the starting point, without any subtours. In other words a tour t is an assignment whose cells can be written down as

$$t = \{1,j_1), (j_1,j_2), \cdots, (j_{n-1}, 1)\},$$

where $j_1, j_2, \cdots, j_{n-1}$ is a permutation of the numbers $2, 3, \cdots, n$. To be specific, we can say that t is a tour covering the cities $\{1, 2, \cdots, n\}$. Thus, $T \subset A$ and $K_T \subset K_A$.

By a *self-loop at a city* we mean a cell of the form (i, i). It corresponds to an allocation along the principal diagonal of the matrix X representing an assignment. Any cell of the form (i, i) is also called a *diagonal cell.* Any cell of the form (i, j) where $i \neq j$ is called a *nondiagonal cell.*

Pick any subset S of the cities $\{1, 2, \cdots, n\}$ such that $S \subset \{1, 2, \cdots, n\}$ and $S \neq \{1, 2, \cdots, n\}$. Then any tour covering the cities in S only is known as a *subtour.*

A *nontour* is an assignment which is not a tour *and which has no self-loops.* In other words it is an assignment without any allocation along the principal diagonal, whose unit cells constitute at least two subtours.

D.A. is an abbreviation for the diagonal assignment which is the assignment represented by the unit matrix.

Two assignments a_1 and a_2 are called *adjacent assignments* if the line segment joining them is an edge of the convex polyhedron K_A, i.e., if and only if every point of the form $\lambda a_1 + (1 - \lambda)a_2$ for all $0 \leqq \lambda \leqq 1$ has a unique representation as a convex combination of assignments.

Two tours t_1 and t_2 are called *adjacent tours* if the line segment joining them forms an edge of the convex polyhedron K_T, i.e., if and only if every point of the form $\lambda t_1 + (1 - \lambda)t_2$ for all $0 \leqq \lambda \leqq 1$ has a unique representation as a convex combination of tours. Since $K_T \subset K_A$, two tours which are not adjacent as assignments may be adjacent as tours.

Suppose the tour $t = \{(i_1, i_2), (i_2, i_3), \cdots, i_n, i_1)\}$. Then the tour $\bar{t} = \{(i_2, i_1), (i_3, i_2), \cdots, (i_1, i_n)\}$ is called the *reflection* of the tour t.

The θ-loop of a nonbasic cell. Consider a basis for (1) representing an assignment a. Such a basis consists of $2n - 1$ basic cells, the n cells of a which are at value 1 and $n - 1$ other independent cells which are at value 0 in the basis.

Let us try to obtain a new basis by bringing the nonbasic cell (i_1, j_1) into the basis. To do this, we put an entry of $+\theta$ in the nonbasic cell (i_1, j_1). Since the sum of all the entries in each row and column should equal 1, we should put a $-\theta$ entry somewhere else in column j_1 and row i_1. Make all these subsequent entries among the basic cells only. Taking up from column j_1, put alternate entries of $-\theta$ and $+\theta$ among columns and rows until the $+\theta$ entry in each row and column is canceled by a $-\theta$ entry. The set of all the basic cells along the $-\theta$ and $+\theta$ path is called the *θ-loop* of the nonbasic cell (i_1, j_1) in this basis. The maximum value which θ can take without the resulting solution violating the nonnegativity constraint of the x_{ij}'s is known as the *value* with which the nonbasic cell (i_1, j_1) enters the basis.

ZBC is an abbreviation for any zero-valued basic cell in any basis for (1). In any basis for (1), if a nonbasic cell (i_1, j_1) enters the basis with a value of zero, then it can be brought into the basis as a ZBC replacing any of the old ZBC's in its θ-loop. If it enters the basis with a unit value, then it can be brought into the basis by replacing one of the unit-valued cells in its θ-loop. But in this process some of the other unit-valued basic cells might become ZBC's.

3. Mathematical theory. We shall first of all look at a characterization of the set of all tours T as a subset of the set of all assignments A. This leads to the corollary that the traveling salesman problem is a special case of the general problem of finding the minimal cost adjacent vertex of a given vertex in a linear programming problem. This can be solved easily when the linear programming problem is nondegenerate. But if the given vertex is a degenerate vertex, the problem of finding its minimal cost adjacent vertex becomes very hard, which explains the difficulty in solving the traveling salesman problem.

THEOREM 1. *Considering K_A, the set of all feasible solutions to* (1), *we have*:

(i) *all tours are adjacent assignments to D.A.*;

(ii) *every nontour is not an adjacent assignment of D.A.*;

(iii) *the class of all adjacent assignments of D.A. consists of*

(a) *all the tours,*

(b) *all the subtours in a smaller number of cities with self-loops at the remaining cities.*

This theorem has been proved by Heller in [1].

(i) can be proved by taking a basis for (1) representing the D.A., with $(1, 2), (2, 3), \cdots, (n - 1, n)$ as ZBC's. In this basis for (1) if the nonbasic cell $(n, 1)$ is brought into the basis, the tour $\{(1, 2), (2, 3), \cdots, (n, 1)\}$ is obtained. Thus the tour $\{(1, 2), (2, 3), \cdots, (n, 1)\}$ is obtained by performing a single pivot in a basis for (1) representing the D.A., and hence it is an adjacent assignment of the D.A. A similar argument holds for every other tour.

(iii) is proved by a similar argument.

(ii) follows because any $2n - 1$ of the cells among those of the D.A. and any nontour are not linearly independent and hence cannot constitute a basis for (1). Thus any nontour cannot be obtained by a single pivot step in any basis representing the D.A., which proves (ii).

COROLLARY 1. *The traveling salesman problem is a special case of the following problem: given a feasible vertex V (i.e., an extreme point) in a linear programming problem, find the minimal cost adjacent vertex of V.*

Proof. Consider the assignment problem with C as the cost matrix, i.e., the problem of minimizing $Z = \sum_{i,j} c_{ij}x_{ij}$ subject to the constraints (1).

The cost of any self-loop is α, which is a very large positive number. Hence, (iii) of Theorem 1 implies that the minimal cost tour is the minimal cost adjacent assignment of D.A.

COROLLARY 2. *Consider any assignment a which has no self-loops:*

$$a = \{(i_1, j_1), \cdots, (i_n, j_n)\}, \qquad i_r \neq j_r, \quad r = 1, \cdots, n.$$

If the cells of a together with any $n - 1$ of the diagonal cells as ZBC's form a basis for the system of constraints (1), *then a must be a tour and conversely.*

Proof. This follows easily because if a contains at least two subtours, then any $2n - 1$ of the cells $\{(1, 1), \cdots, (n, n), (i_1, j_1), \cdots, (i_n, j_n)\}$ cannot constitute a basis for (1) as in (ii) of Theorem 1 and conversely.

4. Properties of nonadjacent tours. The following theorem provides a test for determining whether two given tours are adjacent tours or not.

THEOREM 2. *Two tours t_1 and t_2 are not adjacent tours if and only if it is possible to form another tour t_3, distinct from t_1 and t_2, by taking some cells out of t_1 and the others out of t_2, but no cells outside those of t_1 and t_2. Such a tour t_3 contains all the common cells of t_1 and t_2. In other words, t_1 and t_2 are not adjacent tours if and only if there exists a tour t_3, $t_3 \neq t_1$, $t_3 \neq t_2$ such that*

$$\{t_3\} \subset \{t_1\} \cup \{t_2\} \quad \textit{and} \quad \{t_1\} \cap \{t_2\} \subset \{t_3\}.$$

Proof. If t_1 and t_2 are not adjacent tours, then by definition there exists $0 < \alpha < 1$ such that

$$\alpha t_1 + (1 - \alpha)t_2 = \sum_{i=1}^{r} \beta_i s_i, \tag{3}$$

where $\beta_i > 0$, $\sum_{i=1}^{r} \beta_i = 1$, each of the s_i for $i = 1, \cdots, r$ is a tour and at least one of them, say s_1, is distinct from t_1 and t_2.

In (3) none of the s_i for $i = 1, \cdots, r$ can contain any cell outside those of t_1 and t_2 since $\beta_i > 0$ for all $i = 1$ to r.

It also implies that each of the s_i must contain all the common cells of t_1 and t_2, since $\beta_i > 0$ for $i = 1, \cdots, r$.

Hence, the tour s_1 which is distinct from t_1 and t_2 satisfies all the requirements in the proposition for the tour t_3.

On the other hand, if there exists a tour like t_3 above, then t_4 such that

$$\{t_4\} = [\{t_1\} \cap \{t_2\}] \cup [\{t_1\} \cup \{t_2\} \setminus \{t_3\}],$$

where $\setminus$ indicates set theoretic difference, represents another tour by Lemma 1, which follows. And, $\frac{1}{2}t_1 + \frac{1}{2}t_2 = \frac{1}{2}t_3 + \frac{1}{2}t_4$. Hence, t_1 and t_2 are not adjacent tours.

Definition. Consider any tour t, where

$$t = \{(i_1, i_2), (i_2, i_3), \cdots, (i_n, i_1)\}.$$

Then, a subset of t like

$$\{(i_1, i_2), \cdots, (i_{r-1}, i_r)\}$$

is called a *segment of t from i_1 to i_r*. It consists of all the cells of t along a path from i_1 to i_r in t. The arc (i_1, i_2) itself may be considered as a segment of t from i_1 to i_2.

Lemma 1. *Suppose t_1 and t_2 are two distinct tours and t_3 is another tour such that*

$$t_3 \neq t_1, \qquad t_3 \neq t_2,$$

$$\{t_3\} \supset \{t_1\} \cap \{t_2\},$$

$$\{t_3\} \subset \{t_1\} \cup \{t_2\}.$$

Then, the cells

$$\{t_4\} = [\{t_1\} \cap \{t_2\}] \cup [\{t_1\} \cup \{t_2\} \setminus \{t_3\}],$$

where $\setminus$ indicates set theoretic difference, represent another tour.

Proof. Since both $\{t_3\}$ and $\{t_4\}$ contain all the common cells of t_1 and t_2, it is sufficient to prove the lemma for the case when t_1 and t_2 have no common cells.

In $\{t_1\} \cup \{t_2\}$ there are two cells in each row and column. Of these t_3 contains one in each row and column, since t_3 is a tour. Thus, t_4, which consists of the remaining cells, contains one cell from each row and column. Hence, t_4 is an assignment.

It remains to show that in t_4 there is a path from any city to any other.

Since t_3 is a tour, it must consist of some segments of t_1 and some of t_2. Actually, it consists of alternating segments from t_1 to t_2 respectively, i.e., it may consist of a segment from i_{r_1} to i_{r_2} of t_1, then a segment from i_{r_2} to i_{r_3} of t_2, then again a segment from i_{r_3} to i_{r_4} of t_1, etc.

Thus, t_4, which consists of the remaining segments of t_1 and t_2 (after striking off those in common with t_3) contains a path from each city to each other. Hence, t_4 is a tour.

Lemma 2. *t and $\tilde{t}$, the reflection of t, are always adjacent tours for $n \geqq 3$.*

Proof. Consider

$$t = \{(1, 2)(2, 3)(3, 4)(4, 5)(5, 6)(6, 1)\},$$

$$\tilde{t} = \{(2, 1)(3, 2)(4, 3)(5, 4)(6, 5)(1, 6)\}.$$

If these are not adjacent tours, then by Theorem 2 it is possible to form a tour s distinct from t and $\tilde{t}$ from the cells $\{t\} \cup \{\tilde{t}\}$.

Suppose $(1, 2) \in s$. Then, since s contains only one cell from each row and column, $(1, 6) \notin s$ and $(3, 2) \notin s$. So, $(5, 6) \in s$, $(3, 4) \in s$. Hence, $(5, 4) \notin s$. Now, since s cannot contain any subtours, $(1, 2) \in s$ implies that $(2, 1) \notin s$. Similarly, $(6, 5) \notin s$, $(4, 3) \notin s$. Hence, $(2, 3) \in s$, $(6, 1) \in s$, $(4, 5) \in s$. Hence, $s = t$. Hence, it is not possible to form a tour distinct from t and $\tilde{t}$ with the cells of $\{t\} \cup \{\tilde{t}\}$. Therefore, by Theorem 2, t and $\tilde{t}$ are adjacent tours.

In general, by renumbering the cities, we can assume that

$$t = t^* = \{(1, 2), (2, 3), \cdots, (n - 1, n), (n, 1)\}.$$

By a construction similar to the above, we verify that the only tours that can be formed using only the cells $\{t^*\} \cup \{\tilde{t}^*\}$ are t^* and $\tilde{t}^*$. Hence, by Theorem 2, t^* and $\tilde{t}^*$ are adjacent tours. Only when $n = 2, t = \tilde{t} = \{(1, 2), (2, 1)\}$.

LEMMA 3. *Suppose $n \geqq 4$ and $r \leqq n - 3$. Let*

$$t = \{(1, i_1), (i_1, i_2), (i_2, i_3), \cdots, (i_{n-2}, i_{n-1}), (i_{n-1}, 1)\}$$

be any tour. Pick any r of the cells of t. Then there exists an adjacent tour t_1 of t containing exactly those r cells in common with t.

Proof. The tour t may be represented by the sequence

$$1 i_1 i_2 \cdots i_{n-2} i_{n-1}$$

indicating the order in which the cities are visited in the tour t.

The sequence which represents $\tilde{t}$, the reflection of t, is obtained by reversing the order in which the cities occur in the sequence representing t. Thus $\tilde{t}$ is represented by the sequence

$$i_{n-1} i_{n-2} \cdots i_2 i_1 1.$$

Case 1. Suppose the r cells which were picked constitute a segment of t from 1 to i_r, say. We wish to find an adjacent tour of t which contains this entire segment. For this we shall treat all these cities from 1 to i_r along the segment as a single block of cities. This is indicated by enclosing the segment from 1 to i_r within brackets, in the sequence representing t, which then becomes

$$[1 i_1 i_2 \cdots i_r] i_{r+1} \cdots i_{n-2} i_{n-1}.$$

We treat this entire block as if it were one location. Any arc entering this block enters at 1 and any arc leaving the block leaves from i_r. In t, the $n - r$ cells which are not on the segment from 1 to i_r form a tour in the cities $i_{r+1}, \cdots, i_{n-1}$ and the block, the reflection of which has all the properties desired of t_1. To generate it we write down the reverse sequence obtained by reversing the order of the cities $i_{r+1}, \cdots, i_{n-1}$ and the block in the sequence for t. In reversing the order of the cities, we treat the block as if it were another super-city, and we reverse its position in the sequence, but keep the order of the cities within it unchanged. This gives rise to the sequence

$$i_{n-1} i_{n-2} \cdots i_{r+1} [1 i_1 \cdots i_r].$$

The tour represented by this sequence

$$t_1 = \{(i_{n-1}, i_{n-2}), \cdots, (i_{r+2}, i_{r+1}), (i_{r+1}, 1), (1, i_1), \cdots, (i_{r-1}, i_r), (i_r, i_{n-1})\}$$

is an adjacent tour of t which has all the cells of the segment from 1 to i_r in common with t.

Case 2. Suppose the r cells which were picked constitute k nonoverlapping segments of t, say from 1 to i_{l_1}, from i_{l_2} to i_{l_3}, etc.

As before, write down the sequence representing the tour t and in that sequence represent each of the k segments above as a block:

$$[1 i_1 \cdots i_{l_1}] i_{l_1+1} \cdots [i_{l_2} \cdots i_{l_3}] \cdots .$$

Any city which is not in any block is known as an *out of block city*.

Now reverse the order of the out of block cities and the blocks in the above sequence, without changing the order of the cities inside each block. This gives a new sequence and let t_1 be the tour represented by it. Then t_1 is an adjacent tour of t and its common cells with t are exactly the r cells which were picked (contained within the blocks).

As an illustration, if

$$t = \{(1,3),(3,2),(6,5),(9,8);(2,7),(4,9),(5,1),(7,10),(8,6),(10,4)\},$$

the tour t_1 obtained by the above procedure, containing the first four cells in t, is

$$t_1 = \{(1,3),(3,2),(6,5),(9,8);(2,6),(5,9),(8,4),(4,10),(10,7),(7,1)\}.$$

LEMMA 4. *When $n \geqq 6$, it is always possible to find a pair of nonadjacent tours.*

Proof. If $n = 6$, let

$$t_1 = \{(1,2),(2,3),(3,4),(4,5),(5,6),(6,1)\},$$
$$t_2 = \{(1,3),(3,2),(2,4),(4,6),(6,5),(5,1)\},$$
$$t_3 = \{(1,2),(2,3),(3,4),(4,6),(6,5),(5,1)\},$$

and if $n > 6$, let

$$t_1 = \{(1,7),(7,8),\cdots,(n-1,n),(n,2),(2,3),(3,4),(4,5),(5,6),(6,1)\},$$
$$t_2 = \{(1,n),(n,n-1),(n-1,n-2),\cdots,(8,7),(7,3),(3,2),(2,4),(4,6),(6,5),(5,1)\},$$
$$t_3 = \{(1,7),(7,8),\cdots,(n-1,n),(n,2),(2,3),(3,4),(4,6),(6,5),(5,1)\}.$$

Then, $t_3 \neq t_1$, $t_3 \neq t_2$ and $\{t_3\} \subset \{t_1\} \cup \{t_2\}$. Hence, by Theorem 2, t_1 and t_2 are not adjacent tours.

LEMMA 5. *When $n \geqq 6$, the number of adjacent tours of any given tour is*

$$\geqq 2^n - \left[1 + n + \binom{n}{2}\right].$$

Proof. When $n \geqq 6$ and $r \leqq n - 3$, by Lemma 3 we know that there exists at least one adjacent tour of t containing exactly any selected r cells of t in common with it. Hence, if U_n is the number of adjacent tours of a given tour, then

$$U_n \geqq \sum_{n=0}^{n-3} \binom{n}{r} = 2^n - \left[1 + n + \binom{n}{2}\right].$$

This indicates that the number of adjacent tours of a given tour goes up at least in the order of 2^n. This completes the proof.

The important steps in the simplex algorithm for minimizing a linear function on a convex polyhedral set described by a set of linear inequalities are the following:

(i) An easy method has been developed by which adjacent vertices of any given vertex may be obtained.

In the simplex method this is done by bringing a nonbasic variable into the basis (one pivot step).

(ii) If the present vertex does not minimize the linear function on the solution set, then a simple criterion has been developed, by which one can obtain an adjacent vertex at which the linear function takes a value less than or equal to that at the present vertex.

In the simplex method this is done by bringing into the basis a nonbasic variable whose relative cost coefficient is negative.

Even though it is not easy to describe the convex polyhedral set K_T by a set of linear inequalities, it is possible to develop a simple method by which adjacent tours of a given tour may be obtained. This corresponds to Step (i) of the simplex method discussed above.

The method for obtaining adjacent tours of a given tour uses pivot steps on the assignment matrix, which is characterized by the set of linear constraints (1). This is discussed below.

4.1. An algorithm for generating an adjacent tour of a given tour. Any basis for the system of constrains (1) with the $n - 1$ ZBC's along the principal diagonal represents a tour by Corollary 2. Such a basis is known as a *diagonal basis* (DB) of that tour. Using the test developed in Theorem 2 and Lemma 2, an algorithm which starts with a DB of a given tour and leads to a DB of an adjacent tour is described below.

Consider a given tour t. Then, the cells of t are known as the *original basic cells* (OBC's).

Step 1. Start with any DB for t. Bring any nonbasic cell which is not a diagonal cell into the basis replacing an OBC (or a diagonal cell if this is not possible) in its row or column.

The new cells that are brought into the basis are called the *new basic cells* (NBC's),

At any stage an OBC in the row or column of an NBC is known as an *excess cell*. A row (or column) is known as a *deficit row* (*column*) if it has

(i) only one basic cell in it and if this is either a diagonal cell or an excess cell;

(ii) only two basic cells in it and if one of them is a diagonal cell and the other an excess cell.

Table 1

Step	Current basis	Excess cells	Deficit rows and columns	NBC or diagonal cell brought in	OBC or diagonal cell removed
	(1, 2) (2, 3) (3, 4) (4, 5) (5, 6) (6, 7) (7, 8) (8, 9) (9, 10) (10, 1); (1, 1) (2, 2) (3, 3) (4, 4) (5, 5) (6, 6) (7, 7) (8, 8) (9, 9)			(2, 7)	(2, 3)
1	(2, 7) (7, 8) (8, 9) (9, 10) (10, 1) (1, 2) (3, 3) (4, 4) (5, 5) (6, 6); (1, 1) (2, 2) (3, 4) (4, 5) (5, 6) (7, 7) (8, 8) (9, 9) (6, 7)	(6, 7)	row 3 col. 3	(6, 5)	(5, 5)
2	(2, 7) (6, 5) (5, 6) (7, 8) (8, 9) (9, 10) (10, 1) (1, 2) (3, 3) (4, 4); (1, 1) (2, 2) (3, 4) (6, 6) (7, 7) (8, 8) (9, 9) (6, 7) (4, 5)	(6, 7 (4, 5)	row 4 col. 3	(4, 9)	(4, 5)
3	(2, 7) (6, 5) (5, 6) (7, 8) (9, 10) (10, 1) (1, 2) (3, 3) (4, 4) (8, 9); (4, 9) (1, 1) (2, 2) (3, 4) (6, 6) (7, 7) (8, 8) (9, 9) (6, 7)	(8, 9) (6, 7)	row 8 col. 3	(8, 6)	(6, 7)
4	(2, 7) (6, 5) (7, 8) (9, 10) (10, 1) (1, 2) (3, 3) (4, 4) (5, 6) (8, 9); (4, 9) (8, 6) (1, 1) (2, 2) (3, 4) (6, 6) (7, 7) (8, 8) (9, 9)	(5, 5) (8, 9)	row 5 col. 3	(5, 1)	(1, 1)
5	(2, 7) (6, 5) (1, 2) (7, 8) (9, 10) (3, 3) (4, 4) (5, 6) (8, 9) (10, 1); (4, 9) (5, 1) (8, 6) (2, 2) (3, 4) (6, 6) (7, 7) (8, 8) (9, 9)	(5, 6) (8, 9) (10, 1)	row 10 col. 3	(10, 4)	(10, 1)
6	(2, 7) (10, 4) (4, 9) (8, 6) (6, 5) (5, 1) (1, 2) (7, 8) (9, 10) (3, 3); (2, 2) (4, 4) (6, 6) (7, 7) (8, 8) (9, 9) (3, 4) (5, 6) (8, 9)	(3, 4) (5, 6) (8, 9)	row 3 col. 3	(3, 2)	(3, 4)
7	(2, 7) (10, 4) (4, 9) (8, 6) (6, 5) (5, 1) (3, 3) (7, 8) (9, 10) (1, 2); (3, 2) (2, 2) (4, 4) (6, 6) (7, 7) (8, 8) (9, 9) (5, 6) (8, 9)	(1, 2) (5, 6) (8, 9)	row 1 col. 3	(1, 3)	(1, 2)
8	(8, 6) (6, 5) (5, 1) (1, 3) (3, 2) (2, 7) (10, 4) (4, 9) (7, 8) (9, 10); (2, 2) (3, 3) (4, 4) (6, 6) (7, 7) (8, 8) (9, 9) (5, 6) (8, 9)	(5, 6) (8, 9)		(5, 5)	(5, 6)
9	(8, 6) (6, 5) (5, 1) (1, 3) (3, 2) (2, 7) (10, 4) (4, 9) (7, 8) (9, 10); (2, 2) (3, 3) (4, 4) (5, 5) (6, 6) (7, 7) (8, 8) (9, 9) (8, 9)	(8, 9)		(1, 1)	(7, 8)
10	(10, 4) (4, 9) (1, 1) (2, 2) (3, 3) (5, 5) (6, 6) (7, 7) (8, 8) (9, 10); (8, 6) (6, 5) (5, 1) (1, 3) (3, 2) (2, 7) (4, 4) (9, 9) (8, 9)	(8, 9)	row 7	(7, 10)	(8, 9)
11	(10, 4) (4, 9) (1, 1) (2, 2) (3, 3) (5, 5) (6, 6) (7, 7) (8, 8) (9, 10); (8, 6) (6, 5) (5, 1) (1, 3) (3, 2) (2, 7) (7, 10) (4, 4) (9, 9)	(9, 10)	row 9	(9, 8)	(9, 10)
12	(1, 3) (3, 2) (2, 7) (7, 10) (10, 4) (4, 9) (9, 8) (8, 6) (6, 5) (5, 1); (1, 1) (2, 2) (3, 3) (4, 4) (5, 5) (6, 6) (7, 7) (8, 8) (9, 9)				

Subsequent steps. Bring into the basis a nonbasic cell which is not a diagonal cell and which is in a deficit row or column and not in a row or column of any NBC, replacing if possible an OBC in its row or column or otherwise a diagonal basic cell in the same row or column.

The process terminates when a DB is reached.

If at any stage a DB is not reached, but there is no deficit row or column, then the number of diagonal basic cells must be $< n - 1$. Bring a nonbasic diagonal cell back into the basis replacing an excess cell if possible, or otherwise an OBC in its θ-loop. When $n - 1$ diagonal basic cells are again in the basis, either a DB is obtained or some deficit rows and columns are created.

The steps are repeated until a DB is reached. The new DB represents the DB of an adjacent tour of t by Theorem 2.

Also let t be any tour and t_1 an adjacent tour of t. Start with a DB for t and bring successively the cells of $\{t_1\} \setminus \{t\}$ (where $\setminus$ indicates set theoretic difference) as NBC's in the above algorithm. By Theorem 2 there does not exist any other tour t_2 distinct from t and t_1 whose cells form a subset of $\{t\} \cup \{t_1\}$. Hence the above algorithm will terminate only when all the cells of t_1 are brought into the basis.

Thus by an appropriate choice of NBC's at the various steps, all the adjacent tours of a given tour can be obtained by the above algorithm.

4.2. A numerical example. Let $t = \{(1, 2), (2, 3), \cdots, (9, 10), (10, 1)\}$. Starting with a DB for t, we obtain an adjacent tour of t. The bases for (1) during the various steps of the algorithm are given in Table 1.

In the table, the basic cells at each stage of the algorithm are arranged in two groups; the cells listed before the symbol ";" are unit-valued basic cells and those that follow the ";" are ZBC's.

Since Step 12 gave a DB, the tour

$$t_1 = \{(1, 3), (3, 2), (2, 7), (7, 10), (10, 4), (4, 9), (9, 8), (8, 6), (6, 5), (5, 1)\}$$

is an adjacent tour of t.

REFERENCES

[1] I. HELLER, *On the traveling salesman's problem*, Proc. 2nd Symposium in Linear Programming, National Bureau of Standards, Washington D.C., 1955, pp. 643–665.

[2] ———, *Neighbour relations on the convex of cyclic permutations*, Pacific J. Math, 6 (1956), pp. 467–477.

[3] M. M. FLOOD, *The traveling salesman problem*, Operations Res., 4 (1956), pp. 61–75.

[4] H. W. KUHN, *On certain convex polyhedra*, Abstract, Bull. Amer. Math. Soc., 61 (1955), pp. 557–558.

[5] R. E. GOMORY, *The traveling salesman problem*, Proc. IBM Scientific Computing Symposium on Combinatorial Problems, White Plains, New York, 1964, pp. 93–121.

SIAM J. CONTROL
Vol. 7, No. 1, February 1969

CONTINUITY IN THE STRONG TOPOLOGY OF OPERATOR-VALUED SOLUTIONS OF NONLINEAR DIFFERENTIAL EQUATIONS WITH AN APPLICATION TO OPTIMAL CONTROL*

DAVID L. RUSSELL†

0. Introduction. A classical result in the theory of ordinary differential equations concerns continuity of solutions with respect to initial conditions and parameters. If for $x \in R^n$, $\mu \in R^m$, t real, we denote by $x(t, \xi, \mu)$ the solution of

$$\dot{x} = f(x, t, \mu) \tag{0.1}$$

which satisfies the initial condition

$$x(0, \xi, \mu) = \xi, \tag{0.2}$$

then the familiar result is the following. Suppose that the solution $x(t, \xi_0, \mu_0)$ exists for $t \in [0, T)$. Given $\delta > 0$, if we choose $\|\xi - \xi_0\|$ and $\|\mu - \mu_0\|$ sufficiently small, then the solution $x(t, \xi, \mu)$ exists for $t \in [0, T - \delta]$. Moreover, uniformly for all $t \in [0, T - \delta]$,

$$\lim_{\substack{\xi \to \xi_0 \\ \mu \to \mu_0}} x(t, \xi, \mu) = x(t, \xi_0, \mu_0). \tag{0.3}$$

Of course, certain rather general conditions must be imposed upon $f(x, t, \mu)$. These conditions as well as a proof of the result just indicated may be found, for example, in [1].

In this paper we wish to consider differential equations with solutions $X(t)$, where $X(t) \in \beta(B_1, B_2)$, the Banach space of all bounded linear transformations $X : B_1 \to B_2$, where B_1 and B_2 are themselves Banach spaces. Now the usual topology used in $\beta = \beta(B_1, B_2)$ is the one induced by the norm

$$\|X\| = \sup_{\substack{y \in B_1 \\ \|y\|_1 = 1}} \|Xy\|_2, \tag{0.4}$$

where $\|\cdot\|_j$ is the norm in B_j. As long as we speak of continuity with respect to this norm topology there is very little change from the results described above as we pass from finite-dimensional systems to systems having operator-valued solutions $X(t)$. There is, however, another topology in β which is of frequent interest. We say that a sequence $\{X_n\} \subseteqq \beta$ converges to $X_\infty \in \beta$ in the strong topology of β if, for every $y \in \beta_1$,

$$\lim_{n \to \infty} \|X_n y - X_\infty y\|_2 = 0. \tag{0.5}$$

* Received by the editors June 4, 1968.

† Department of Mathematics, University of Wisconsin, Madison, Wisconsin 53706. This research was supported in part by the National Science Foundation under Grant GP-6070.

We may then ask: Given a differential equation with solutions in β, does continuity with respect to initial conditions and parameters prevail if we work only within the framework of the strong topology?

It is immediately clear that if any such results are obtained they must differ somewhat from those already known for (0.1). For example, let β denote the space of bounded linear transformations from the Hilbert space l^2 into itself. An element $X \in \beta$ can be represented by an infinite matrix:

$$X = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots \\ x_{21} & x_{22} & x_{23} & \cdots \\ x_{31} & x_{32} & x_{33} & \cdots \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \end{pmatrix}. \tag{0.6}$$

Let us denote those X represented by diagonal matrices as

$$X = \operatorname{diag}(x_{11}, x_{22}, x_{33}, \cdots). \tag{0.7}$$

Consider then the differential equation

$$\dot{X} = X^2 \tag{0.8}$$

in $\beta = \beta(l^2, l^2)$. We prescribe initial conditions

$$\begin{aligned} X_n(0) &= \operatorname{diag}(\underbrace{-1, -1, \cdots, -1}_{n \text{ entries}}, 1, 1, 1, \cdots), \qquad n = 1, 2, 3, \cdots, \\ X_\infty(0) &= \operatorname{diag}(-1, -1, -1, \cdots). \end{aligned} \tag{0.9}$$

It is clear that $\lim_{n \to \infty} X_n(0) = X_\infty(0)$ in the strong topology of β. Now the corresponding solutions of (0.8) are

$$\begin{aligned} X_n(t) &= \operatorname{diag}\left(\underbrace{\frac{1}{-1-t}, \cdots, \frac{1}{-1-t}}_{n \text{ entries}}, \frac{1}{1-t}, \frac{1}{1-t}, \cdots\right), \\ X_\infty(t) &= \operatorname{diag}\left(\frac{1}{-1-t}, \frac{1}{-1-t}, \frac{1}{-1-t}, \cdots\right). \end{aligned} \tag{0.10}$$

The solution $X_\infty(t)$ exists for all $t \geqq 0$ while none of the solutions $X_n(t)$ is defined for $t \geqq 1$. This is a definite departure from our experience with finite-dimensional systems. Observe, however, that for all $t < 1, X_n(t)$ converges strongly to $X_\infty(t)$ as $n \to \infty$.

In § 1 we shall study, under fairly general assumptions, the question of strong continuity of solutions with respect to initial conditions and parameters. Because we wish to include the case of "mild solutions" (cf. [2]) of certain differential equations whose linear part involves an unbounded operator, we work with the integral equation (1.6) below. This equation is of a type general enough to include most differential equations in β which are likely to be of interest.

In § 2 we shall show that our results are of more than purely mathematical interest in that they can be applied to the study of finite-dimensional approximations to certain optimization problems for differential equations in Hilbert spaces, or, in engineering parlance, distributed parameter systems.

1. Strong continuity results for integral equations. Let $\beta = \beta(B_1, B_2)$ be the Banach space of bounded linear transformations $X : B_1 \to B_2$ equipped with the norm (0.4) and let Ω be a compact topological space. For $X \in \beta$, $\mu \in \Omega$, t, s real, $0 \leqq s \leqq t$, let $F(X, \mu, t, s)$ and $G(X, \mu, t)$ be functions

$$F : \beta \otimes \Omega \otimes R^2 \to \beta, \tag{1.1}$$

$$G : \beta \otimes \Omega \otimes R^1 \to \beta, \quad G(X, \mu, 0) \equiv X, \tag{1.2}$$

with the following properties.

PROPERTY 1. Let $X(\mu, t)$ and $X_0(\mu)$ be continuous relative to the strong topology of β for $\mu \in \Omega$, t real. Then $F(X(\mu, s), \mu, t, s)$ and $G(X_0(\mu), \mu, t)$ are both continuous relative to the strong topology of β for $\mu \in \Omega$, t and s real, $0 \leqq s \leqq t$.

PROPERTY 2. Corresponding to any set

$$0 \leqq t \leqq T, \qquad \|X\| \leqq K \tag{1.3}$$

there is a positive number $L = L(T, K)$ such that if X_1, X_2, t satisfy (1.3), then, uniformly for $\mu \in \Omega$, $0 \leqq s \leqq t \leqq T$,

$$\|F(X_2, \mu, t, s) - F(X_1, \mu, t, s)\| \leqq L\|X_2 - X_1\|, \tag{1.4}$$

$$\|G(X_2, \mu, t) - G(X_1, \mu, t)\| \leqq L\|X_2 - X_1\|. \tag{1.5}$$

THEOREM 1. *Let $X(\mu, t)$ be the solution of the integral equation*

$$X(\mu, t) = G(X_0(\mu), \mu, t) + \int_0^t F(X(\mu, s), \mu, t, s)\, ds \tag{1.6}$$

satisfying the initial condition

$$X(\mu, 0) = X_0(\mu). \tag{1.7}$$

If $X_0(\mu)$ is strongly continuous for $\mu \in \Omega$, there is a positive number T such that $X(\mu, t)$ is strongly continuous in the set $\{\mu \in \Omega, t \in [0, T]\}$.

Remark. The integral equation (1.6) clearly covers the case of a differential equation

$$\dot{X} = P(X, \mu, t), \tag{1.8}$$

where $P(X, \mu, t)$ is a polynomial in X with coefficient operators which are strongly continuous functions for $\mu \in \Omega$, $t \in R^1$. This is true because multiplication of bounded operators preserves strong continuity.

Proof of Theorem 1. The usual proof of the local existence and uniqueness of solutions of (1.6) employs the method of successive approximations. One sets

$$X_0(\mu, t) \equiv X_0(\mu), \qquad 0 \leqq t \leqq T, \tag{1.9}$$

and thereafter

$$X_{n+1}(\mu, t) = G(X_0(\mu), \mu, t) + \int_0^t F(X_n(\mu, s), \mu, t, s)\, ds. \tag{1.10}$$

Using Properties (1) and (2) above, one shows in a way which is by now familiar to all that, if T is sufficiently small,

$$\lim_{n\to\infty} \|X_n(\mu, t) - X(\mu, t)\| = 0 \tag{1.11}$$

uniformly for $0 \leqq t \leqq T$, where $X(\mu, t)$ is the unique solution of (1.6) and (1.7).

For each fixed $y \in B_1$, $X_0(\mu)y : \Omega \to B_2$ is continuous with respect to the $\|\cdot\|_2$ topology of B_2 for $\mu \in \Omega$. Since Ω is compact there is a positive number $M(y)$ such that

$$\|X_0(\mu)y\|_2 \leqq M(y), \qquad \mu \in \Omega. \tag{1.12}$$

The principle of uniform boundedness [3] may then be invoked to show that there is a positive number M such that

$$\|X_0(\mu)\| \leqq M, \qquad \mu \in \Omega. \tag{1.13}$$

Using (1.13) together with the fact that the inequalities (1.4) and (1.5) for F and G are required to hold uniformly for $\mu \in \Omega$, $0 \leqq s \leqq t \leqq T$, it is immediately evident upon examination of the method of successive approximations that T can be chosen independently of $\mu \in \Omega$ and that (1.11) holds uniformly for $\mu \in \Omega$, $t \in [0, T]$. Moreover, there is a positive number $K \geqq M$ such that

$$\|X_n(\mu, t)\| \leqq K, \qquad \mu \in \Omega, \qquad t \in [0, T], \qquad n = 0, 1, 2, \cdots. \tag{1.14}$$

For details of the method of successive approximations, we suggest [1].

Let $\mu_0 \in \Omega$, $0 \leqq s_0 \leqq t_0$. Since $X_0(\mu, t) \equiv X_0(\mu)$ converges strongly to $X_0(\mu_0)$ as μ converges to μ_0, we may use Property 1 to see that for each $y \in B_1$,

$$\lim_{\substack{\mu\to\mu_0\\ t\to t_0}} G(X_0(\mu), \mu, t)y = G(X_0(\mu_0), \mu_0, t_0)y \tag{1.15}$$

and

$$\lim_{\substack{\mu\to\mu_0\\ t\to t_0\\ s\to s_0}} F(X_0(\mu, s), \mu, t, s)y = F(X_0(\mu_0, s_0), \mu_0, t_0, s_0)y, \qquad 0 \leqq s \leqq t. \tag{1.16}$$

Combining (1.10) with (1.15) we see that

$$\lim_{\substack{\mu\to\mu_0\\ t\to t_0}} X_1(\mu, t)y = X_1(\mu_0, t_0)y, \qquad y \in B_1, \tag{1.17}$$

if and only if, for all $y \in B_1$,

$$\lim_{\substack{\mu\to\mu_0\\ t\to t_0}} \int_0^t F(X_0(\mu, s), \mu, t, s)y\, ds = \int_0^{t_0} F(X_0(\mu_0, s), \mu_0, t_0, s)y\, ds. \tag{1.18}$$

Let $t_1 > t_0$ be fixed and let $\hat{F} = F$, $0 \leqq s \leqq t \leqq t_1$, $\hat{F} = 0, t < s \leqq t_1$. Then the continuity expressed by (1.16) implies that for $0 \leqq s \leqq t_1, s \neq t_0$,

$$\lim_{\substack{t \to t_0 \\ \mu \to \mu_0}} \hat{F}(X_0(\mu, s), \mu, t, s)y = \hat{F}(X_0(\mu_0, s), \mu_0, t_0, s)y. \tag{1.19}$$

The continuity (1.16) together with the compactness of the set $\{\mu \in \Omega, 0 \leqq s \leqq t \leqq t_1\}$ implies the boundedness of $\hat{F}$ in that set. We may then apply the Lebesgue dominated convergence theorem to the integral

$$\int_0^{t_1} \hat{F}(X_0(\mu, s), \mu, t, s)y \, ds$$

to obtain the desired result (1.18). The Lebesgue dominated convergence theorem for vector-valued functions is proved in [3]. We have noted that (1.17) is implied by (1.18) and hence $X_1(\mu, t)$ is strongly continuous for $\mu \in \Omega$, $0 \leqq t \leqq T$.

One now proceeds by induction, showing, just as above, that the strong continuity of $X_n(\mu, t)$ implies that of $X_{n+1}(\mu, t)$ via (1.10). Thus $X_n(\mu, t)$ is strongly continuous for $\mu \in \Omega$, $0 \leqq t \leqq T$, $n = 0, 1, 2, \cdots$.

Now let $\varepsilon > 0$ be chosen and let $y \in B_1$. Let N_ε be chosen so that

$$\|X(\mu, t) - X_n(\mu, t)\| \leqq \frac{\varepsilon}{4\|y\|_1}, \qquad \mu \in \Omega, \qquad t \in [0, T], \tag{1.20}$$

for all $n \geqq N_\varepsilon$. Then let μ, t be chosen close enough to μ_0, t_0 in the product topology of $\Omega \otimes [0, T]$ so that

$$\|(X_{N_\varepsilon}(\mu_0, t_0) - X_{N_\varepsilon}(\mu, t))y\|_2 \leqq \varepsilon/2. \tag{1.21}$$

Then we readily verify that

$$\|X(\mu_0, t_0)y - X(\mu, t)y\|_2 \leqq \varepsilon, \tag{1.22}$$

and the proof of Theorem 1 is complete.

The result expressed by Theorem 1 is purely local in that strong continuity of $X(\mu, t)$ persists only over a sufficiently short interval $[0, T]$. The example (0.8) offered in the Introduction shows that in general we cannot expect more. However, we can obtain global results if we assume a certain boundedness of the solutions.

THEOREM 2. *Let it be known that, for all $\mu \in \Omega$, $X(\mu, t)$ exists and satisfies* (1.6) *for all $t \geqq 0$. A necessary and sufficient condition in order that $X(\mu, t)$ should be strongly continuous in the set $\{\mu \in \Omega, t \geqq 0\}$ is that there exists a nonnegative function $K(t)$ (without loss of generality increasing) such that*

$$\|X(\mu, t)\| \leqq K(\tau), \qquad \mu \in \Omega, \qquad 0 \leqq t \leqq \tau < \infty. \tag{1.23}$$

Proof. The condition is clearly necessary. If, for each $y \in B_1$, $X(\mu, t)y$ is continuous in the compact set $\{\mu \in \Omega, t \in [0, \tau]\}$, then $\|X(\mu, t)y\|_2$ is bounded there. The principle of uniform boundedness then shows that $\|X(\mu, t)\|$ is bounded in that set.

Now we shall show that the boundedness (1.23) is sufficient for global strong continuity of $X(\mu, t)$. Let $\tau > 0$ be chosen and let $t_0 \in [0, \tau)$. For $t_0 \leqq t \leqq \tau$ we have

$$X(\mu, t) = \hat{G}(\mu, t_0, t) + \int_{t_0}^{t} F(X(\mu, s), \mu, t, s)\, ds, \tag{1.24}$$

where

$$\hat{G}(\mu, t_0, t) = G(X_0(\mu), \mu, t) + \int_{0}^{t_0} F(X(\mu, s), \mu, t, s)\, ds. \tag{1.25}$$

Let us assume that $X(\mu, s)$ is strongly continuous in the set $\{\mu \in \Omega, 0 \leqq s \leqq t_0\}$. Now Property 1 implies the boundedness of $G(0, \mu, t)$ and $F(0, \mu, t, s)$, uniformly for $0 \leqq s \leqq t \leqq \tau$. Combining this with the a priori bound (1.23) and Property 2 we obtain a bound on $G(X_0(\mu), \mu, t)$ and $F(X(\mu, s), \mu, t, s)$ which is uniformly valid for $\mu \in \Omega$, $0 \leqq s \leqq t \leqq \tau$. The Lebesgue dominated convergence theorem implies that $\hat{G}(\mu, t_0, t)$ is strongly continuous for $\mu \in \Omega$, $0 \leqq t \leqq \tau$, for the integral in (1.25) involves $X(\mu, s)$ only for $0 \leqq s \leqq t_0$. We may now apply the techniques of Theorem 1, altered only very slightly, to extend the strong continuity from $[0, t_0]$ to $[0, t_0 + \hat{T}]$, where the size of $\hat{T} > 0$ depends only upon τ, not upon t_0. Starting with $t_0 = T$, the interval length found in Theorem 1, a finite number of extensions cover $[0, \tau]$ and the proof is complete.

We end this section by noting that the case of a sequence $X_k(t)$ of solutions (as, for example, (0.8)–(0.10)) is included in Theorems 1 and 2 by taking $\Omega = \{1, 2, \cdots, \infty\}$ with a neighborhood system described by:

(i) N is a neighborhood of the finite integer n if N is any subset of Ω which includes n.

(ii) N is a neighborhood of ∞ if N is a subset of Ω which contains all but finitely many n.

Clearly Ω, as thus topologized, is compact.

2. Application to quadratic optimal control problems. Let A be a normal operator defined on a Hilbert space H_1 with spectrum contained in some left half-plane:

$$\sigma(A) \subseteqq \{\mu | \operatorname{Re}(\mu) \leqq \mu_0\}. \tag{2.1}$$

Let H_2 be a second Hilbert space and $\tilde{B}: H_2 \to H_1$ a bounded linear operator which, for some $\lambda \notin \sigma(A)$, can be written in the form

$$\tilde{B} = (A - \lambda I)^{-1} B, \tag{2.2}$$

where $B: H_2 \to H_1$ is also bounded. We consider the linear ordinary differential equation

$$\frac{dx}{dt} = Ax + \tilde{B}u \tag{2.3}$$

in H_1 with initial condition

$$x(0) = x_0 \in \Delta = \operatorname{dom}(A). \tag{2.4}$$

Given a fixed time $T_1 > 0$ let us consider the problem of finding a control $u_*(t)$ lying in the set of measurable vector-valued functions

$$\left\{u:R^1 \to H_2 \,\middle|\, \int_0^{T_1} \|u(t)\|^2\, dt < \infty\right\} \tag{2.5}$$

which minimizes

$$\begin{aligned} C(u) = \int_0^{T_1} &\{((A - \lambda I)x(t), W(A - \lambda I)x(t)) + (u(t), Uu(t))\}\, dt \\ &+ ((A - \lambda I)x(T), G(A - \lambda I)x(T)) \end{aligned} \tag{2.6}$$

with respect to all controls in (2.5). Here W, U and G are bounded self-adjoint operators. W and G are required to be positive semi-definite while U is to be positive definite. Because (2.2) implies $x(t) \in \Delta$ for all t, $C(u)$ is always defined.

The above problem has been considered in detail by D. L. Lukes and the present author in [4], where it is shown that the minimizing control $u_*(t)$ is generated by the feedback law

$$u_*(t) = -U^{-1}\tilde{B}^*(A^* - \lambda I)Q(t)(A - \lambda I)x_*(t) \equiv D(t)x_*(t), \tag{2.7}$$

which gives rise to responses $x_*(t)$ satisfying

$$\frac{dx_*}{dt} = [A + \tilde{B}D(t)]x_*(t). \tag{2.8}$$

The bounded linear operator $Q(t):H_1 \to H_1$ is given for $t \leqq T_1$ as the unique strongly continuous solution of the integral equation

$$Q(t) = e^{A^*(T_1 - t)}Ge^{A(T_1 - t)} - \int_t^{T_1} e^{A^*(s-t)}[-W + Q(s)(BU^{-1}B^*)Q(s)]e^{A(s-t)}\, ds. \tag{2.9}$$

An optimal control theory formulated in such an abstract setting is of very little value to the engineer unless it represents a limiting case relative to some approximating sequence of finite-dimensional problems. In practice the spectrum of A is usually discrete, consisting of a sequence of complex eigenvalues, and the corresponding eigenvectors form a complete orthonormal set in H_1. Thus, a natural way to define finite-dimensional approximations is to restrict attention to finite-dimensional subspaces of H_1 spanned by finitely many of the eigenvectors of A. H_2 may already be finite-dimensional (this is usually the case) but if not it will also be necessary to consider finite-dimensional subspaces of it.

We arrange the eigenvalues of A in a sequence, $\lambda_1, \lambda_2, \lambda_3, \cdots$, and we let E_k denote the orthogonal projection from H_1 onto the space H_{1k} spanned by the eigenvectors of A corresponding to the first k eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_k$. The projections E_k commute with A and converge strongly to the identity as $k \to \infty$. Also, we let $\{\tilde{E}_k\}$ be a similar collection of orthogonal projections on H_2 which commute with the self-adjoint operator U. For $k = 1, 2, 3, \cdots$ we put

$$T_k = E_kTE_k, \qquad E_kT\tilde{E}_k, \qquad \tilde{E}_kTE_k \quad \text{or} \quad \tilde{E}_kT\tilde{E}_k \tag{2.10}$$

according as $T:H_1 \to H_1$, $T:H_2 \to H_1$, $T:H_1 \to H_2$ or $T:H_2 \to H_2$, respectively. Also, we set

$$x_k = E_k x, \qquad u_k = \tilde{E}_k u. \tag{2.11}$$

We consider now the finite-dimensional systems

$$\frac{dx_k}{dt} = A_k x_k + \tilde{B}_k u_k, \qquad k = 1, 2, 3, \cdots, \tag{2.12}$$

with initial conditions

$$x_k(0) = x_{0k}. \tag{2.13}$$

One then seeks for a control $u_{k*}(t)$ minimizing

$$\begin{aligned} C_k(u_k) = \int_0^{T_1} &\{((A_k - \lambda E_k)x_k(t), W_k(A_k - \lambda E_k)x_k(t)) + (u_k(t), U_k u_k(t))\}\, dt \\ &+ ((A_k - \lambda E_k)x_k(T_1), G_k(A_k - \lambda E_k)x_k(T_1)). \end{aligned} \tag{2.14}$$

Again one shows that the minimizing control $u_{k*}(t)$ is generated by

$$u_{k*}(t) = -U_k^{-1}\tilde{B}_k^*(A_k^* - \lambda E_k)\hat{Q}_k(t)(A_k - \lambda E_k)x_{k*}(t) = \hat{D}_k(t)x_{k*}(t) \tag{2.15}$$

yielding responses $x_{k*}(t)$ satisfying

$$\frac{dx_{k*}}{dt} = [A_k + \tilde{B}_k\hat{D}_k(t)]x_{k*}(t). \tag{2.16}$$

Here $\hat{Q}_k(t)$ satisfies

$$\begin{aligned} \hat{Q}_k(t) &= e^{A_k^*(T_1 - t)}G_k e^{A_k(T_1 - t)} \\ &\quad - \int_t^{T_1} e^{A_k^*(s-t)}[-W_k + \hat{Q}_k(s)(B_k U_k^{-1} B_k^*)\hat{Q}_k(s)]s^{A_k(s-t)}\, ds. \end{aligned} \tag{2.17}$$

It should be emphasized that in general $\hat{Q}_k \neq Q_k$, $\hat{D}_k \neq D_k$, $u_{k*} \neq u_{*k}$, and $x_{k*} \neq x_{*k}$.

The real test of this approximation procedure is provided by asking if, for $0 \leqq t \leqq T_1$,

$$\lim_{k \to \infty} u_{k*}(t) = u_*(t), \tag{2.18}$$

$$\lim_{k \to \infty} (A_k - \lambda E_k)x_{k*}(t) = (A - \lambda I)x_*(t), \tag{2.19}$$

Evidently both of these will be true if, again for $0 \leqq t \leqq T_1$,

$$\lim_{k \to \infty} \hat{Q}_k(t)y = Q(t)y, \qquad y \in H_1, \tag{2.20}$$

i.e., if $\hat{Q}_k(t)$ converges strongly to $Q(t)$ as $k \to \infty$.

If we reverse the time sense in (2.9) and (2.17) and attach an index "∞" to $Q(t)$, then (2.17), $k = 1, 2, \cdots$, and(2.9) satisfy all of the conditions set down in § 1. The parameter space Ω consists of $1, 2, 3, \cdots, \infty$ topologized as indicated at the end of § 1. Theorem 1 then implies the validity of (2.20) for t sufficiently close to T_1, $t \leqq T_1$. The result is obtained in the complete interval $0 \leqq t \leqq T_1$ by showing that $\|\hat{Q}_k(t)\|$, $\|Q(t)\|$ are uniformly bounded for $0 \leqq t \leqq T_1$, thus permitting the application of Theorem 2. This boundedness is rigorously demonstrated in [4] and is a consequence of the observation that the quadratic forms $(y, Q(t)y)$, $(y_k, \hat{Q}_k(t)y_k)$ represent the minimum costs associated with problems of the above type posed on the interval $[t, T_1]$ with initial conditions $x(t) = (A - \lambda I)^{-1}y$, $x_k(t) = (A_k - \lambda E_k)^{-1}y_k$.

REFERENCES

[1] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[2] F. E. BROWDER, *Nonlinear equations of evolution*, Ann. of Math., 80 (1964), pp. 485–523.

[3] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semi-Groups*, Colloquium Publications, vol. 31, American Mathematical Society, Providence, 1957.

[4] D. L. LUKES AND D. L. RUSSELL, *The quadratic criterion for distributed systems*, this Journal, 7 (1969), pp. 101–121.

SIAM J. Control
Vol. 7, No. 1, February 1969

EXISTENCE OF SADDLE POINTS IN DIFFERENTIAL GAMES*

P. VARAIYA and JIGUAN LIN†

1. Introduction. We consider games in which there are two players I and II whose respective states $x(t) \in R^n$, $y(t) \in R^m$ at time t obey the differential equations (1) and (2) respectively:

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

$$\dot{y}(t) = g(y(t), v(t), t). \tag{2}$$

The control functions u and v are constrained by $u(t) \in U$, $v(t) \in V$, where $U \subset R^p$, $V \subset R^q$ are fixed compact subsets. The game starts at time $t = 0$ in some specified initial states $x(0) = x_0$, $y(0) = y_0$ and ends at a specified time T, at which instant I receives from II a certain amount—the payoff. We consider two kinds of payoff. The payoff of the first kind is the value of a functional $\mu(x, y)$, where x and y are the trajectories of the two players. The payoff of the second kind is the smallest time t for which the triple $(x(t), y(t), t)$ belongs to a specified closed subset

$$F \subset R^n \times R^m \times R,$$

where it is assumed that $R^n \times R^m \times \{T\} \subset F$ and $T < \infty$. At each time t player I selects a control $u(t) \in U$ based upon his observations of the trajectory of II up to time t in such a way as to maximize the payoff; conversely, at each time t player II selects a control $v(t) \in V$ based upon his observations of $x(\tau)$, $0 \leqq \tau \leqq t$, in such a way as to minimize the payoff. Games with payoff of the first kind have been called games of prescribed duration [1], while games with payoff of the second kind have been called pursuit-evasion games (player I is the evader, II is the pursuer). Now it is difficult to make precise the notion of a strategy for the players which takes into account the information available to them at each instant of time. In this paper we shall propose a precise definition of a strategy (which agrees with our intuition) and we justify it by demonstrating the existence of a saddle point. Our definition is an extension of that given in [2] in a direction suggested by Roxin [3].

Whereas the technique that we use to prove the saddle-point theorems (Theorems 7, 8, 9) is borrowed to a large extent from Fleming [4], the spirit of this paper is closer to the approach of Ryll-Nardzewski [5]. In the next section we state standard assumptions on the systems (1) and (2) which guarantee compactness of the space of trajectories of the two players. In § 3 we define classes of strategies with differing information patterns and prove an important (although easy) result which allows us to compare these different classes of strategies. In § 4 we use this result to give a very simple proof of Fleming's theorem for a payoff

* Received by the editors March 14, 1968, and in revised form July 2, 1968.

† Electronics Research Laboratory, University of California, Berkeley, California 94720. This research was sponsored by the National Science Foundation under Grant GK-716.

of the first kind, namely, we show that the optimal payoff for the majorant and minorant games (see [4]) converge to the same limit V_F as the discrepancy in the information patterns vanishes. In § 5 we propose our definition of the game and show existence of saddle points for a payoff of the first kind (Theorem 7). The value of the game agrees with that of Fleming. As a corollary to this result in § 6 we obtain existence of saddle points for payoffs of the second kind. In § 7 we give one example which seems to show that our definition cannot be made more attractive.

2. Conditions on the differential systems. We make the following assumptions on the differential systems (1). Corresponding assumptions are made (but not stated) regarding (2).

(i) For each fixed t, f is continuous in (x, u) for all $(x, u) \in R^n \times U$.

(ii) There is a measurable function k, integrable on finite intervals, such that for every $u \in U$ and x, $\hat{x}$ in R^n,

$$|f(x, u, t) - f(\hat{x}, u, t)| \leqq k(t)|x - \hat{x}|.$$

(Here and throughout $|\cdot|$ denotes Euclidean norm in R^n or R^m.)

(iii) There are positive numbers M and N, and a measurable function l, integrable on finite intervals such that for every x in R^n, and u in U,

$$|f(x, u, t)| \leqq l(t)(M + N|x|).$$

(iv) Convexity condition: For every x in R^n, t in R, the set

$$f(x, U, t) = \{f(x, u, t)|u \in U\}$$

is convex.

A measurable function u (v) is said to be an admissible control if $u(t) \in U$ ($v(t) \in V$) for all t. A solution x of (1) (y of (2)) is said to be an admissible trajectory if it arises from an admissible control.

Definition. Let $X_T(x_0)$ denote the set of all admissible trajectories x of (1) which are defined on $[0, T]$ and which start at x_0 at time 0, i.e., $x(0) = x_0$. Similarly we define $Y_T(y_0)$.

We consider $X_T(x_0)$ as a subset of the Banach spaces C_T^n—the space of all continuous functions from $[0, T]$ into R^n under the max norm. Similarly, $Y_T(y_0)$ is a subset of C_T^m. The next result is well known (see, for example, [6] or [7]); the first part is a consequence of the assumption that the sets $f(x, U, t)$ and $g(y, V, t)$ are convex, whereas the second part follows from the assumption that f, g satisfy Lipschitz conditions.

Theorem 1. (i) *If $X_0 \subset R^n$ and $Y_0 \subset R^m$ are compact, then*

$$\bigcup_{x_0 \in X_0} X_T(x_0) \subset C_T^n \quad \textit{and} \quad \bigcup_{y_0 \in Y_0} Y_T(y_0) \subset C_T^m$$

are compact.

(ii) *$X_T(\cdot)$, $Y_T(\cdot)$ are continuous functions of their arguments. (Here continuity is with respect to the Hausdorff metric.)*

Let $X_0 \subset R^n$, $Y_0 \subset R^m$ be compact, and define

$$X_T = \bigcup_{x_0 \in X_0} X_T(x_0), \quad Y_T = \bigcup_{y_0 \in Y_0} Y_T(y_0).$$

THEOREM 2. *For each* $\delta \in (0, 1]$ *there exists a map* $\Pi_\delta^X : X_T \to X_T$ *with the following properties*:

(i) *If* $x \in X_T(x_0)$, *then* $\Pi_\delta^X(x) \in X_T(x_0)$.

(ii) *If* x, $\hat{x}$ *in* X_T *satisfy*

$$x(\tau) = \hat{x}(\tau) \quad \textit{for} \quad 0 \leqq \tau \leqq t,$$

then

$$\Pi_\delta^X(x)(\tau) = \Pi_\delta^X(\hat{x})(\tau) \quad \textit{for} \quad 0 \leqq \tau \leqq t + \delta T.$$

(iii) *Let*

$$\varepsilon(\delta) = \sup \{ \|x - \Pi_\delta^X(x)\| \, | x \in X_T \};$$

then

$$\lim_{\delta \to 0} \varepsilon(\delta) = 0.$$

(*Here and throughout* $\| \cdot \|$ *denotes norm in the Banach spaces* C_T^n, C_T^m.)

Remark. The idea of the proof is the following: We take $x \in X_T$. Suppose x arises from a control u. Then let $\Pi_\delta^X(x)$ be the trajectory arising from the same control u delayed by δT. Since different controls may yield the same trajectory, care has to be taken in choosing the control in order to obtain property (ii). Because the proof of this theorem does not directly contribute to our main interest it is presented as an Appendix.

3. Strategies. Let x_0, y_0 be specified initial states. Throughout this paper the symbol δ (with or without subscripts) represents a number which is equal to $1/2^n$ for some integer $n \geqq 0$. We now define three classes of strategies $A_\delta(x_0, y_0) = \{\alpha_\delta\}$, $A(x_0, y_0) = \{\alpha\}$, and $A^\delta(x_0, y_0) = \{\alpha^\delta\}$ for player I and three classes of strategies $B_\delta(x_0, y_0) = \{\beta_\delta\}$, $B(x_0, y_0) = \{\beta\}$, and $B^\delta(x_0, y_0) = \{\beta^\delta\}$ for player II.

DEFINITION. (i) $A_\delta(x_0, y_0)$ is the set of all functions $\alpha_\delta : Y_T(y_0) \to X_T(x_0)$ such that if y, $\hat{y}$ are in $Y_T(y_0)$ with $y(\tau) = \hat{y}(\tau)$ for $0 \leqq \tau \leqq i\delta T$, then $\alpha_\delta(y)(\tau) = \alpha_\delta(\hat{y})(\tau)$ for $0 \leqq \tau \leqq (i + 1)\delta T$, $i = 0, 1, \cdots, 1/\delta - 1$.

(ii) $A^\delta(x_0, y_0)$ is the set of all functions $\alpha^\delta : Y_T(y_0) \to X_T(x_0)$ such that if y, $\hat{y}$ are in $Y_T(y_0)$ with $y(\tau) = \hat{y}(\tau)$ for $0 \leqq \tau \leqq i\delta T$, then $\alpha^\delta(y)(\tau) = \alpha^\delta(\hat{y})(\tau)$ for $0 \leqq \tau \leqq i\delta T$, $i = 0, 1, \cdots, 1/\delta$.

(iii) $A(x_0, y_0)$ is the set of all functions $\alpha : Y_T(y_0) \to X_T(x_0)$ such that if y, $\hat{y}$ are in $Y_T(y_0)$ with $y(\tau) = \hat{y}(\tau)$ for $0 \leqq \tau \leqq t$, then $\alpha(y)(\tau) = \alpha(\hat{y})(\tau)$ for $0 \leqq \tau \leqq t$, $0 \leqq t \leqq T$.

The sets of strategies $B_\delta(x_0, y_0)$, $B(x_0, y_0)$ and $B^\delta(x_0, y_0)$ are defined in the same way.

It is convenient to regard the strategies A_δ, A, A^δ for I as subsets of $F(Y_T(y_0), X_T(x_0))$—the space of all functions from $Y_T(y_0)$ into $X_T(x_0)$ with the topology of pointwise convergence. Similarly, we regard B_δ, B, B^δ as subsets of the topological space $F(X_T(x_0), Y_T(y_0))$. By the Tikhonov theorem $F(X_T(x_0), Y_T(y_0))$, $F(Y_T(y_0), X_T(x_0))$ are compact.

The first part of the next result is a direct consequence of the definition, while the proof of the second part is a duplication of the arguments in Lemma 4.1 of [2].

THEOREM 3. *If* $\delta_1 \leqq \delta_2$, *then*:

(i) $A_{\delta_2} \subset A_{\delta_1} \subset A \subset A^{\delta_1} \subset A^{\delta_2}$ *and* $B_{\delta_2} \subset B_{\delta_1} \subset B \subset B^{\delta_1} \subset B^{\delta_2}$.

(ii) *The sets* A_δ, A, A^δ *are closed and hence compact subsets of* $F(Y_T(y_0), X_T(x_0))$. *Similarly the sets* B_δ, B, B^δ *are closed and hence compact subsets of* $F(X_T(x_0), Y_T(y_0))$.

Recall the map Π_δ^X and the function $\varepsilon(\delta)$ in Theorem 2.

THEOREM 4 (Approximation theorem). (i) *If* $\alpha^\delta \in A^\delta$, $\beta^\delta \in B^\delta$, *then* $(\Pi_\delta^X \circ \alpha^\delta)$ *belongs to* A_δ, *and* $(\beta^\delta \circ \Pi_\delta^X)$ *belongs to* B_δ.

(ii) $\|\alpha^\delta(y) - (\Pi_\delta^X \circ \alpha^\delta)(y)\| \leqq \varepsilon(\delta)$ *for* $\alpha^\delta \in A^\delta$, $y \in Y_T(y_0)$.

Proof. (i) is a consequence of the definition, while (ii) follows from Theorem 2.

4. Payoff of the first kind; Fleming's theorem. Let $X_0 \subset R^n$, $Y_0 \subset R^m$ be fixed compact sets. Let $X_T = \bigcup_{x_0 \in X_0} X_T(x_0)$, $Y_T = \bigcup_{y_0 \in Y_0} Y_T(y_0)$. The payoff is a continuous real-valued function μ defined on the compact space $X_T \times Y_T$. Let $x_0 \in X_0$, $y_0 \in Y_0$ be specified initial states. Following Fleming [4], for each δ we define a majorant game $G^\delta(x_0, y_0)$ and a minorant game $G_\delta(x_0, y_0)$ as follows: In the majorant game, player II picks a strategy $\beta_\delta \in B_\delta(x_0, y_0)$ and then player I picks a strategy $\alpha^\delta \in A^\delta(x_0, y_0)$. Then a *unique* pair of trajectories $(x, y) \in X_T(x_0) \times Y_T(y_0)$ is determined by $(\alpha^\delta, \beta_\delta)$ stepwise on successive intervals $[0, \delta T]$, $[\delta T, 2\delta T], \cdots, [T - \delta T, T]$ as follows:

1. By definition of β_δ, $\beta_\delta(x_1)(t) = \beta_\delta(x_2)(t)$ for $0 \leqq t \leqq \delta T$ and for all x_1,x_2 in $X_T(x_0)$. Let

$$y(t) = \beta_\delta(x_1)(t), \qquad 0 \leqq t \leqq \delta T,$$

where $x_1 \in X_T(x_0)$ is arbitrary.

2. If y_1, y_2 in $Y_T(y_0)$ are any trajectories such that $y_1(t) = y_2(t) = y(t)$ for $0 \leqq t \leqq \delta T$, then by definition of α^δ,

$$\alpha^\delta(y_1)(t) = \alpha^\delta(y_2)(t), \qquad 0 \leqq t \leqq \delta T.$$

Let

$$x(t) = \alpha^\delta(y_1)(t), \qquad 0 \leqq t \leqq \cdot\delta T,$$

where $y_1 \in Y_T(y_0)$ is any trajectory such that $y_1(t) = y(t)$ for $0 \leqq t \leqq \delta T$.

3. If x_1, x_2 in $X_T(x_0)$ are any trajectories such that $x_1(t) = x_2(t) = x(t)$ for $0 \leqq t \leqq \delta T$, then by definition of β_δ,

$$\beta_\delta(x_1)(t) = \beta_\delta(x_2)(t), \qquad 0 \leqq t \leqq 2\delta T.$$

Let

$$y(t) = \beta_\delta(x_1)(t), \qquad 0 \leqq t \leqq 2\delta T,$$

where $x_1 \in X_T(x_0)$ is any trajectory such that $x_1(t) = x(t)$ for $0 \leqq t \leqq \delta T$.

Analogous to step 2, the knowledge of y on the internal $[0, 2\delta T]$ determines via α^δ the trajectory x on $[0, 2\delta T]$. This in turn determines y via β_δ on the internal $[0, 3\delta T]$ and so on. We shall denote the dependence of (x, y) on the strategies $(\alpha^\delta, \beta_\delta)$ by sometimes writing $x = x(\alpha^\delta, \beta_\delta)$, $y = y(\alpha^\delta, \beta_\delta)$.

In the minorant game, player I first selects a strategy $\alpha_\delta \in A_\delta(x_0, y_0)$, and then II chooses some $\beta^\delta \in B^\delta(x_0, y_0)$. In a dual manner to the steps outlined above there results a unique pair of trajectories (x, y) in $X_T(x_0) \times Y_T(y_0)$ whose dependence on $(\alpha_\delta, \beta^\delta)$ will sometimes be denoted by $x = x(\alpha_\delta, \beta^\delta)$, $y = y(\alpha_\delta, B^\delta)$.

Since I tries to maximize the payoff and II tries to minimize it, we define

$$V^\delta(x_0, y_0) = \min_{\beta_\delta \in B_\delta(x_0, y_0)} \max_{\alpha^\delta \in A^\delta(x_0, y_0)} \mu(x(\alpha^\delta, \beta_\delta), y(\alpha^\delta, \beta_\delta)),$$

$$V_\delta(x_0, y_0) = \max_{\alpha_\delta \in A_\delta(x_0, y_0)} \min_{\beta^\delta \in B^\delta(x_0, y_0)} \mu(x(\alpha_\delta, \beta^\delta), y(\alpha_\delta, \beta^\delta)).$$

From Theorem 3(i) it follows that

$$V_{\delta_2}(x_0, y_0) \leqq V_{\delta_1}(x_0, y_0) \leqq V^{\delta_1}(x_0, y_0) \leqq V^{\delta_2}(x_0, y_0)$$

whenever $\delta_1 \leqq \delta_2$. Hence the two limits

$$\overline{V}(x_0, y_0) = \lim_{\delta \to 0} V^\delta(x_0, y_0), \qquad V(x_0, y_0) = \lim_{\delta \to 0} V_\delta(x_0, y_0)$$

exist.

From the definition of the strategies and the manner in which they determine the outcome of the trajectories, it should be clear that an alternate definition of V^δ, V_δ is the following characterization which is closer to that of Fleming [4]:

$$V^\delta(x_0, y_0) = \min_{y^1 \in Y_1(y_0)} \max_{x^1 \in X_1(x_0)} \min_{y^2 \in Y_2(y^1(\delta T))} \max_{x^2 \in X_2(x^1(\delta T))} \cdots \min_{y^{1/\delta} \in Y_{1/\delta}(y^{1/\delta - 1}((1-\delta)T))} \max_{x^{1/\delta} \in X_{1/\delta}(x^{1/\delta - 1}((1-\delta)T))} \mu(x, y), \tag{3}$$

$$V_\delta(x_0, y_0) = \max_{x^1 \in X_1(x_0)} \min_{y^1 \in Y_1(y_0)} \cdots \max_{x^{1/\delta} \in X_{1/\delta}(x^{1/\delta - 1}((1-\delta)T))} \min_{y^{1/\delta} \in Y_{1/\delta}(y^{1/\delta - 1}((1-\delta)T))} \mu(x, y), \tag{4}$$

where, $X_1(x_0)$ $(Y_1(y_0))$ is the set of all admissible trajectories x^1 (y^1) of (1) ((2)) defined on the interval $[0, \delta T]$ and starting at x_0 (y_0); and inductively if x^i (y^i) has been chosen, $X_{i+1}(x^i(i\delta T))$ $(Y_{i+1}(y^i(i\delta T)))$ is the set of all admissible trajectories x^{i+1} (y^{i+1}) defined on $[i\delta T, (i+1)\delta T]$ and starting at time $i\delta T$ in the state $x^i(i\delta T)$ $(y^i(i\delta T))$. The outcome (x, y) is defined by $x(t) = x^i(t)$ $(y(t) = y^i(t))$, $(i-1)\delta T \leqq t \leqq i\delta T$, $i = 1, 2, \cdots, 1/\delta$. Since the various sets of trajectories X_i, Y_i are compact and vary continuously with initial conditions (by Theorem 1), and since μ is a continuous function, it follows that V^δ, V_δ are well-defined and vary continuously with their arguments $(x_0, y_0) \in X_0 \times Y_0$.

The next lemma gives two other alternate expressions for V^δ, V_δ.

LEMMA 1.

(i)

$$V^\delta(x_0, y_0) = \max_{\alpha^\delta \in A^\delta(x_0, y_0)} \min_{\beta_\delta \in B_\delta(x_0, y_0)} \mu(x(\alpha^\delta, \beta_\delta), y(\alpha^\delta, \beta_\delta)), \tag{5}$$

$$V_\delta(x_0, y_0) = \min_{\beta^\delta \in B^\delta(x_0, y_0)} \max_{\alpha_\delta \in A_\delta(x_0, y_0)} \mu(x(\alpha_\delta, \beta^\delta), y(\alpha_\delta, \beta^\delta)). \tag{6}$$

(ii)

$$V^\delta(x_0, y_0) = \min_{\beta_\delta \in B_\delta(x_0, y_0)} \sup_{x \in X_T(x_0)} \mu(x, \beta_\delta(x)), \tag{7}$$

$$V_\delta(x_0, y_0) = \max_{\alpha_\delta \in A_\delta(x_0, y_0)} \inf_{y \in Y_T(y_0)} \mu(\alpha_\delta(y), y). \tag{8}$$

Sketch of proof. We shall prove (5) and (7). A proof of (5) can be obtained by noting that for any sets W, Z and any real-valued function γ on $W \times Z$, the following equality holds:

$$\inf_{z \in Z} \sup_{w \in W} \gamma(z, w) = \sup_{s \in S} \inf_{z \in Z} \gamma(z, s(z)),$$

where S is the set of all functions s from Z into W. This equality, together with the representation (3) of V^δ and the definitions of α^δ, β_δ, can then be used to give (5).

Evidently $V^\delta(x_0, y_0)$ is at least as large as the right-hand side of (7). On the other hand if $\alpha^\delta \in A^\delta(x_0, y_0)$, $\beta_\delta \in B_\delta(x_0, y_0)$ and if $x = x(\alpha^\delta, \beta_\delta)$, $y = y(\alpha^\delta, \beta_\delta)$ is the outcome, then

$$\mu(x, y) = \mu(x, \beta_\delta(x)),$$

and so the right-hand side of (7) is bigger than V^δ.

Following Fleming we propose the following definition.

DEFINITION. The game has a value $V_F(x_0, y_0)$ provided that the two limits

$$\overline{V}(x_0, y_0) = \lim_{\delta \to 0} V^\delta(x_0, y_0) \quad \text{and} \quad V(x_0, y_0) = \lim_{\delta \to 0} V_\delta(x_0, y_0)$$

are equal. In that case we define the (Fleming) value of the game:

$$V_F(x_0, y_0) = \overline{V}(x_0, y_0).$$

LEMMA 2. *Let $\eta > 0$. Then there is a δ^* such that for all $\delta < \delta^*$ and all* $(x_0, y_0) \in X_0 \times Y_0$,

$$0 \leqq V^\delta(x_0, y_0) - V_\delta(x_0, y_0) \leqq \eta.$$

Proof. Since μ is continuous on the compact space $X_T \times Y_T$, there is $\varepsilon^* > 0$ such that

$$|\mu(\hat{x}, y) - \mu(x, y)| \leqq \eta \tag{9}$$

whenever $\|x - \hat{x}\| \leqq \varepsilon^*$, $x, \hat{x} \in X_T$ and $y \in Y_T$. Let $\delta^* > 0$ be such that for all $\delta < \delta^*, \varepsilon(\delta) < \varepsilon^*$, where $\varepsilon(\delta)$ is the function defined in Theorem 2(iii). Now let

$\delta < \delta^*$, $(x_0, y_0) \in X_0 \times Y_0$ be fixed. Let $\alpha^\delta_{\text{opt}} \in A^\delta(x_0, y_0)$ be such that

$$(10) \qquad V^\delta(x_0, y_0) \leqq \mu(x(\alpha^\delta_{\text{opt}}, \beta_\delta), y(\alpha^\delta_{\text{opt}}, \beta_\delta)) \quad \text{for all} \quad \beta_\delta \in B_\delta(x_0, y_0).$$

The existence of $\alpha^\delta_{\text{opt}}$ follows from (5). Let $\underline{\alpha}_\delta = \Pi^X_\delta \circ \alpha^\delta_{\text{opt}}$. Then $\underline{\alpha}_\delta \in A_\delta(x_0, y_0)$ by Theorem 4(i). Let $\beta^\delta \in B^\delta(x_0, y_0)$ be arbitrary and suppose that $x \in X_T(x_0)$, $y \in Y_T(y_0)$ are such that

$$\underline{\alpha}_\delta(y) = x, \qquad \beta^\delta(x) = y.$$

Let $\hat{x} = \alpha^\delta_{\text{opt}}(y)$, and let $\underline{\beta}_\delta = \beta^\delta \circ \Pi^X_\delta$. Then $x = \Pi^X_\delta(\hat{x})$ and $\underline{\beta}_\delta \in B_\delta$ and furthermore,

$$\alpha^\delta_{\text{opt}}(y) = \hat{x}, \qquad \underline{\beta}_\delta(\hat{x}) = y.$$

It follows from (10) that

$$V^\delta(x_0, y_0) \leqq \mu(\hat{x}, y).$$

But $\|x - \hat{x}\| = \|\Pi^X_\delta(\hat{x}) - \hat{x}\| \leqq \varepsilon(\delta) \leqq \varepsilon^*$, so that by (9),

$$V^\delta(x_0, y_0) \leqq \mu(x, y) + \eta.$$

Since $\underline{\alpha}_\delta \in A_\delta$ and since $\beta^\delta \in B^\delta$ is arbitrary, it follows that

$$V^\delta(x_0, y_0) \leqq \eta + \max_{\alpha_\delta \in A_\delta} \min_{\beta^\delta \in B^\delta} \mu(x(\alpha_\delta, \beta^\delta), y(\alpha_\delta, \beta^\delta))$$
$$= \eta + V_\delta(x_0, y_0).$$

The lemma is proved.

THEOREM 5 (Fleming). *Under the assumptions* (*of* §2) *on the differential equations* (1) *and* (2),

$$(11) \qquad \overline{V}(x_0, y_0) = \underline{V}(x_0, y_0).$$

Furthermore, $V_F(\cdot, \cdot)$ *is continuous on* $X_0 \times Y_0$.

Proof. The equality (11) is a corollary of the preceding lemma, while the continuity of V_F follows from the fact that V^δ is continuous and the fact that V^δ converges uniformly to $\overline{V}$.

Remarks. The crucial point in the proof of Lemma 2 (which implies Theorem 5) is the existence of the maps Π^X_δ with the required properties. Suppose property (iii) of Theorem 2 is replaced by (iii'):

(iii') *Let*

$$\eta(\delta) = \sup\{\mu(\Pi^X_\delta(x), y) - \mu(x, y) | x \in X_T, y \in Y_T\}.$$

Then

$$\lim_{\delta \to 0} \eta(\delta) \geqq 0.$$

Theorem 5 can now be strengthened to Theorem 5', and with obvious modification the same proof applies.

THEOREM 5'. *Let* $\mu(x, y)$ *be upper semicontinuous in* x *for fixed* y *and lower semicontinuous in* y *for fixed* x. *Suppose that the differential equations* (1) *and* (2) *satisfy the assumptions of* § 2 *and suppose that for all* $\delta > 0$ *there exist maps* $\Pi_\delta^X : X_T \to X_T$ *which have properties* (i), (ii) *of Theorem* 2 *and property* (iii)′ *above. Then*

$$\underline{V}(x_0, y_0) = \overline{V}(x_0, y_0).$$

It can be shown that if μ is of the form

$$\mu(x, y) = \mu_1(x, y) + \int_0^T L(t, x(t), u(t))\,dt - \int_0^T M(t, y(t), u(t))\,dt$$

with μ_1 continuous, L, M concave in the control variables u, v varying over convex sets U, V, then $\mu(x, y)$ is upper semicontinuous in x for fixed y and lower semicontinuous in y for fixed x. Furthermore, in this case there exist maps Π_δ^X which satisfy the conditions of Theorem 5′, and hence the Fleming value is defined (compare [4, Theorem 2]).

5. The fair game: Existence of saddle points for payoffs of the first kind. In this section we propose a direct definition of a game. Our definition is in some sense a limit of the games G^δ, G_δ as δ goes to zero. However, our formulation is much closer to that of Ryll-Nardzewski [5].

As before let x_0, y_0 be specified initial states. Player I chooses a strategy $\alpha \in A(x_0, y_0)$, player II chooses a strategy $\beta \in B(x_0, y_0)$. It would be natural to define the outcome of such choice to be any pair $x \in X_T(x_0)$, $y \in Y_T(y_0)$ such that

$$\alpha(y) = x, \qquad \beta(x) = y.$$

Unfortunately, the above pair of equations may have either no solution or it may have more than one solution. The existence of a solution (but not uniqueness) can be guaranteed if α, β are required to be continuous functions; but then as we shall show in § 7 we cannot guarantee existence of optimal strategies. We therefore propose the following definition.

DEFINITION. Let $\alpha \in A(x_0, y_0)$ and $\beta \in B(x_0, y_0)$. A pair $x \in X_T(x_0)$, $y \in Y_T(y_0)$ is said to be an *outcome of* (α, β) if there is a sequence $x_n \in X_T(x_0)$, $y_n \in Y_T(y_0)$, $n = 1, 2, 3, \cdots$, such that

$$\lim_{n\to\infty} x_n = \lim_{n\to\infty} \alpha(y_n) = x, \quad \lim_{n\to\infty} y_n = \lim_{n\to\infty} \beta(x_n) = y.$$

(Evidently if α and β are continuous at y, x, respectively, then $\alpha(y) = x$, $\beta(x) = y$.)

Let $O(\alpha, \beta) = \{(x, y) | (x, y) \text{ is an outcome of } (\alpha, \beta)\}$.

THEOREM 6. *For each* $\alpha \in A$, $\beta \in B$, $O(\alpha, \beta)$ *is a nonempty closed subset of* $X_T(x_0) \times Y_T(y_0)$.

Proof. The closedness of $O(\alpha, \beta)$ follows from standard diagonal arguments. We now show that $O(\alpha, \beta)$ is nonempty. Let δ_k, $k = 1, 2, \cdots$, be a sequence decreasing to zero and let $\alpha_{\delta_k} = (\Pi_{\delta_k}^X \circ \alpha) \in A_{\delta_k}$. Let (x_k, y_k) be the pair such that

$$\alpha_{\delta_k}(y_k) = x_k, \qquad \beta(x_k) = y_k.$$

Since $X_T(x_0)$, $Y_T(y_0)$ are compact we can assume (taking subsequences if necessary) that there are $x \in X_T(x_0)$, $y \in Y_T(y_0)$ such that

$$\lim_{k\to\infty} x_k = \lim_{k\to\infty} \alpha_{\delta_k}(y_k) = x, \qquad \lim_{k\to\infty} y_k = \lim_{k\to\infty} \beta(x_k) = y.$$

But

$$\|\alpha_{\delta_k}(y_k) - \alpha(y_k)\| = \|(\Pi^X_{\delta_k} \circ \alpha)(y_k) - \alpha(y_k)\| \leqq \varepsilon(\delta_k)$$

by Theorem 4(ii). Since $\lim_{k\to\infty} \varepsilon(\delta_k) = 0$, the assertion follows.

DEFINITION. For each $\beta \in B(x_0, y_0)$, let

$$\mu^+(\beta) = \sup_{\alpha\in A(x_0,y_0)} \max_{(x,y)\in O(\alpha,\beta)} \mu(x, y),$$

and for each $\alpha \in A(x_0, y_0)$ let

$$\mu_-(\alpha) = \inf_{\beta\in B(x_0,y_0)} \min_{(x,y)\in O(\alpha,\beta)} \mu(x, y).$$

Now let

$$V^+(x_0, y_0) = \min_{\beta\in B(x_0,y_0)} \mu^+(\beta),$$

$$V_-(x_0, y_0) = \max_{\alpha\in A(x_0,y_0)} \mu_-(\alpha).$$

In order to show that the min and max in the definition of V^+, V_- actually exist, the following result will be helpful.

LEMMA 3.

$$\mu^+(\beta) = \sup_{x\in X_T(x_0)} \mu(x, \beta(x)) \tag{12}$$

and

$$\mu_-(\alpha) = \inf_{y\in Y_T(y_0)} \mu(\alpha(y), y).$$

Proof. We prove the first equality. Clearly $\mu^+(\beta)$ is at least as big as the right-hand side of (12). Now let $\alpha \in A$ and let x, x_n be in $X_T(x_0)$, y, y_n in $Y_T(y_0)$ for $n = 1, 2, \cdots$, such that

$$\lim_{n\to\infty} x_n = \lim_{n\to\infty} \alpha(y_n) = x, \quad \lim_{n\to\infty} y_n = \lim_{n\to\infty} \beta(x_n) = y.$$

Then,

$$\lim_{n\to\infty} \mu(x_n, \beta(x_n)) = \mu(x, y).$$

It follows that

$$\mu^+(\beta) \leqq \sup_{x\in X_T(x_0)} \mu(x, \beta(x)).$$

LEMMA 4. *$\mu^+(\beta)$ is a lower semicontinuous function of $\beta \in B(x_0, y_0)$. $\mu_-(\alpha)$ is an upper semicontinuous function of $\alpha \in A(x_0, y_0)$.*

Proof. We shall only prove the first half of the assertion since the proof for the second half is analogous. Let z be a real number and let

$$B_z = \{\beta | \beta \in B(x_0, y_0), \mu^+(\beta) \leqq z\}.$$

We must show that B_z is closed. Let $\{\beta(k)\}$ be a net in B_z converging to β in B, i.e., for each $x \in X_T(x_0)$, $\lim_k \beta(k)x = \beta(x)$. Let $x \in X_T(x_0)$. Then by definition $\mu(x, \beta(k)x) \leqq z$ for all k. It follows from the continuity of μ that $\mu(x, \beta(x)) \leqq z$. Hence $\mu^+(\beta) \leqq z$.

COROLLARY. *There is a* $\beta^* \in B(x_0, y_0)$, $\alpha^* \in A(x_0, y_0)$ *such that*:

(i) $\mu^+(\beta^*) \leqq \mu^+(\beta), \quad \beta \in B,$

$\mu_-(\alpha^*) \geqq \mu_-(\alpha), \quad \alpha \in A,$

(ii) $\mu^+(\beta^*) = V^+(x_0, y_0) = V_F(x_0, y_0) = V_-(x_0, y_0) = \mu_-(\alpha^*)$, *and*

(iii) $\min\limits_{(x,y)\in O(\alpha^*,\beta^*)} \mu(x, y) = \max\limits_{(x,y)\in O(\alpha^*,\beta^*)} \mu(x, y).$

Proof. (i) follows from the preceding lemma and the fact that $B(x_0, y_0)$ and $A(x_0, y_0)$ are compact spaces. Again from the same lemma and the definition of V^+ we see that

$$\begin{aligned}\mu^+(\beta^*) = V^+(x_0, y_0) &= \min_{\beta\in B(x_0,y_0)} \sup_{x\in X_T(x_0)} \mu(x, \beta(x)) \\ &\leqq \min_{\beta_\delta\in B_\delta(x_0,y_0)} \sup_{x\in X_T(x_0)} \mu(x, \beta_\delta(x)) \\ &= V^\delta(x_0, y_0),\end{aligned}$$

where the last equality is the same as (7). Similarly,

$$\mu_-(\alpha^*) = V_-(x_0, y_0) \geqq V_\delta(x_0, y_0)$$

so that (ii) follows from Theorem 5. To prove (iii) it is enough to note that by definition of μ_- and μ^+,

$$\mu_-(\alpha^*) \leqq \min_{(x,y)\in O(\alpha^*,\beta^*)} \mu(x, y) \leqq \max_{(x,y)\in O(\alpha^*,\beta^*)} \mu(x, y) \leqq \mu^+(\beta^*),$$

and then (iii) follows (ii).

We can now define the fair game and prove the existence of a saddle point. The game G is defined as follows: Player I selects a strategy $\alpha \in A(x_0, y_0)$ while II independently selects a $\beta \in B(x_0, y_0)$. The payoff is given by $\mu(x, y)$, where (x, y) is an arbitrarily chosen pair from $O(\alpha, \beta)$. The saddle-point theorem shows that the value is independent of the arbitrary choice of the outcome.

THEOREM 7 (Saddle-point theorem). *There exists* $\alpha^* \in A(x_0, y_0)$, $\beta^* \in B(x_0, y_0)$ *such that for all* $\alpha \in A(x_0, y_0)$ *and all* $\beta \in B(x_0, y_0)$,

$$\begin{aligned}\max_{(x,y)\in O(\alpha,\beta^*)} \mu(x, y) &\leqq \max_{(x,y)\in O(\alpha^*,\beta^*)} \mu(x, y) \\ &= \min_{(x,y)\in O(\alpha^*,\beta^*)} \mu(x, y) \\ &\leqq \min_{(x,y)\in O(\alpha^*,\beta)} \mu(x, y).\end{aligned}$$

Furthermore, $\mu(x, y) = V_F(x, y_0)$ *for all* $(x, y) \in O(\alpha^*, \beta^*)$.

Proof. By the definition of μ^+, μ_- we see that

$$\max_{(x,y)\in O(\epsilon,\beta^*)} \mu(x, y) \leqq \mu^+(\beta^*), \qquad \mu_-(\alpha^*) \leqq \min_{(x,y)\in O(\alpha^*,\beta)} \mu(x, y).$$

The result now follows from the previous corollary.

Definition. Given two players I and II with dynamics (1) and (2) respectively, and a continuous payoff μ of the first kind, the (Fleming) value of the game corresponding to initial conditions (x_0, y_0) will be denoted by

$$V_F(\mu; x_0, y_0).$$

Remark 1. In this paper the dynamics of the two players are separate and the payoff has the form $\mu(x, y)$. As seen in this section (Theorem 7), if μ is continuous, then a saddle point exists without introducing mixed strategies. Unfortunately, if only the conditions of Theorem 5′ are satisfied, then the conclusion of Theorem 7 may *not* hold; however, the following weaker version is true.

Theorem 7′. *Suppose that the conditions of Theorem 5′ hold. Then there exist* $\alpha^* \in A(x_0, y_0)$ *and* $\beta^* \in B(x_0, y_0)$ *such that for all* $\alpha \in A(x_0, y_0)$ *and all* $\beta \in B(x_0, y_0)$,

$$\begin{aligned} \inf_{y\in Y_T(y_0)} \mu(\alpha(y), y) &\leqq \inf_{y\in Y_T(y_0)} \mu(\alpha^*(y), y) = V_F(x_0, y_0) \\ &= \sup_{x\in X_T(x_0)} \mu(x, \beta^*(x)) \\ &\leqq \sup_{x\in X_T(x_0)} \mu(x, \beta(x)). \end{aligned}$$

Remark 2. In the light of the preceding remark and the remark at the end of § 4 the results of this paper appear to simplify and improve some results of [1] and [5], since we demonstrate existence of a saddle point without using mixed strategies. However, in [8] the dynamics are intertwined and the class of games considered is bigger; as a consequence mixed strategies are necessary to obtain a saddle point.

6. Payoff of the second kind: Pursuit-evasion games. In this section we consider payoffs of the second kind. Before we define the game we introduce a definition which will be helpful in relating these games to the games considered in the last section.

Let $F \subset R^n \times R^m \times [0, \infty)$ be a nonempty closed set. For each $T < \infty$ define the function $\mu_T : X_T(x_0) \times Y_T(y_0) \to R$ by

$$\mu_T(x, y) = \min \{|x(t) - \hat{x}| + |y(t) - \hat{y}| + |t - \hat{t}| \,|\, (\hat{x}, \hat{y}, \hat{t}) \in F, t \in [0, T]\}.$$

It is easy to show that μ_T is continuous. Evidently $\mu_T(x, y)$ is nonnegative and

$$(13) \qquad \mu_T(x, y) = 0 \quad \text{if and only if} \quad (x(t), y(t), t) \in F \quad \text{for some } t.$$

We now define the game: There is given a closed set $F \subset R^n \times R^m \times [0, \infty)$ and a $T_{\max} < \infty$ such that $(x, y, T_{\max}) \in F$ for all $(x, y) \in R^n \times R^m$. The game is played on the fixed time interval $[0, T_{\max}]$. Player I (the evader) selects a strategy $\alpha \in A(x_0, y_0)$ while II (the pursuer) independently selects a strategy $\beta \in B(x_0, y_0)$. The payoff given by $t(x, y)$, where $(x, y) \in O(\alpha, \beta)$, is chosen arbitrarily and $t(x, y)$ is the smallest capture time, i.e.,

$$t(x, y) = \min \{t | (t, x(t), y(t)) \in F\}.$$

Player I tries to maximize the payoff while II tries to minimize it. As before we define

$$V_-(x_0, y_0) = \sup_{\alpha \in A(x_0,y_0)} \inf_{\beta \in B(x_0,y_0)} \inf_{(x,y)\in O(\alpha,\beta)} t(x, y),$$

$$V^+(x_0, y_0) = \inf_{\beta \in B(x_0,y_0)} \sup_{\alpha \in A(x_0,y_0)} \sup_{(x,y)\in O(\alpha,\beta)} t(x, y).$$

THEOREM 8. $V_-(x_0, y_0) = V^+(x_0, y_0)$.

Proof. Evidently $V_-(x_0, y_0) \leqq V^+(x_0, y_0)$. Let $\varepsilon > 0$. Then from the definition of V_-, for every strategy α there is a strategy β and a $(x, y) \in (\alpha, \beta)$ such that

$$t(x, y) \leqq V_-(x_0, y_0) + \varepsilon,$$

i.e., there is a $t \leqq T_\varepsilon = V_-(x_0, y_0) + \varepsilon$ such that

$$(x(t), y(t), t) \in F. \tag{14}$$

Now define the continuous function μ_{T_ε} on the set $X_{T_\varepsilon}(x_0) \times Y_{T_\varepsilon}(y_0)$ as in the beginning of this section, and consider the game defined on the fixed time interval $[0, T_\varepsilon]$ with the continuous payoff function μ_{T_ε}. By Theorem 7 this game has a value $V_F(\mu_{T_\varepsilon}; x_0, y_0)$. However, because of (13) and the argument leading to (14), we conclude that

$$V_F(\mu_{T_\varepsilon}; x_0, y_0) = 0.$$

Going back to Theorem 7, the saddle-point property implies the existence of a strategy $\beta(\varepsilon)$ such that for every $\alpha \in A(x_0, y_0)$ and every $(x, y) \in O(\alpha, \beta(\varepsilon))$,

$$\mu_{T_\varepsilon}(x, y) = 0.$$

From (13) we can then conclude that for every $\alpha \in A(x_0, y_0)$ and every $(x, y) \in O(\alpha, \beta(\varepsilon))$,

$$t(x, y) \leqq T_\varepsilon = V_-(x_0, y_0) + \varepsilon.$$

It follows that

$$V^+(x_0, y_0) \leqq V_-(x_0, y_0) + \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, the theorem is proved.

DEFINITION. Let $T^* = V^+(x_0, y_0) = V_-(x_0, y_0)$.

THEOREM 9. *There exists a strategy $\beta^* \in B(x_0, y_0)$ such that*

$$\sup_{\alpha \in A(x_0,y_0)} \sup_{(x,y)\in O(\alpha,\beta^*)} t(x, y) = T^* \leqq \sup_{\alpha \in A(x_0,y_0)} \sup_{(x,y)\in O(\alpha,\beta)} t(x, y)$$

for all $\beta \in B(x_0, y_0)$, i.e., there exists an optimal pursuit strategy.

Proof. Consider the game defined on the fixed time interval $[0, T^*]$ with the continuous payoff function μ_{T^*}. Clearly $V_F(\mu_{T^*}; x_0, y_0) = 0$ and so there exists a strategy β^* such that for all $\alpha \in A(x_0, y_0)$ and all $(x, y) \in O(\alpha, \beta^*)$, $\mu_{T^*}(x, y) = 0$; this implies that $t(x, y) \leqq T^*$.

Unfortunately, trivial examples show that in general there does not exist a strategy $\alpha^* \in A(x_0, y_0)$ such that

$$(15) \qquad T^* = \inf_{\beta \in B(x_0,y_0)} \inf_{(x,y)\in O(\alpha^*,\beta)} t(x, y).$$

We can therefore only assert the following theorem.

THEOREM 10. *If there is a strategy $\alpha^* \in A(x_0, y_0)$ which is optimal for player* I *(i.e., satisfies* (15)*), then the pair (α^*, β^*) forms a saddle point, i.e., for all $\alpha \in A(x_0, y_0)$, $\beta \in B(x_0, y_0)$,*

$$\sup_{(x,y)\in O(\alpha,\beta^*)} t(x, y) \leqq \sup_{(x,y)\in O(\alpha^*,\beta^*)} t(x, y) = T^* = \inf_{(x,y)\in O(\alpha^*,\beta^*)} t(x, y)$$

$$\leqq \inf_{(x,y)\in O(\alpha^*,\beta)} t(x, y).$$

Various conditions can be placed on the set of trajectories and the endzone F which guarantee existence of an optimal evasion strategy α^*. One such condition is the following.

(C) As the initial states and time (x_0, y_0, t_0) approach F, the value $T^*(x_0, y_0, t_0)$ approaches 0.

In this case we can show that the function

$$T(\alpha) = \inf_{\beta \in B(x_0,y_0)} \inf_{(x,y)\in O(\alpha,\beta)} t(x, y)$$

is an upper semicontinuous function of $\alpha \in A(x_0, y_0)$, and hence there exists α^* such that $T(\alpha^*) \geqq T(\alpha)$ for all α. Evidently then $T(\alpha^*) = T^*$ and α^* satisfies (15). We now sketch a proof to show that condition (C) above implies the upper semicontinuity of $T(\alpha)$.

DEFINITION. Let $\alpha \in A(x_0, y_0)$. We say that a pair $(x, y) \in X_T(x_0) \times Y_T(y_0)$ is a possible outcome if there is a sequence y_n, $n = 1, 2, \cdots$, in $Y_T(y_0)$ converging to y such that $\alpha(y_n)$, $n = 1, 2, \cdots$, converges to x. Let $PO(\alpha)$ be the set of all possible outcomes.

It is easy to check that

$$T(\alpha) = \inf_{(x,y)\in PO(\alpha)} t(x, y).$$

Now let z be any real number and let

$$A_z = \{\alpha | \alpha \in A(x_0, y_0), T(\alpha) \geqq z\}.$$

We must show that A_z is a closed set. Let $\{\alpha(k)\}$ be a net in A_z converging to α and let $(x, y) \in PO(\alpha)$, i.e., let $\{y_n\} \subset Y_T(y_0)$ be a sequence such that y_n converges to y and $\alpha(y_n)$ converges to x. Suppose that $t(x, y) = z - \varepsilon$ for some $\varepsilon > 0$. This means that

$$(x(z - \varepsilon), y(z - \varepsilon), z - \varepsilon) \in F.$$

Since $\lim_{n\to\infty}\|y_n - y\| = 0$ and $\lim_{n\to\infty}\|\alpha(y_n) - x\| = 0$, given $\eta > 0$ there is

$N(\eta) < \infty$ sufficiently large such that[1]

$$\rho\{(\alpha(y_n)(z - \varepsilon), y_n(z - \varepsilon), z - \varepsilon), F\} < \eta$$

whenever $n > N(\eta)$. Now $\lim_k \alpha(k)(y_n) = \alpha(y_n)$. Hence for k sufficiently large,

$$\rho\{(\alpha(k)(y_n)(z - \varepsilon), y_n(z - \varepsilon), z - \varepsilon), F\} < 2\eta.$$

But then by condition (C), $T(\alpha(k)) \leqq z - \varepsilon + \gamma(\eta)$, where $\lim_{\eta \to 0} \gamma(\eta) = 0$. It follows that for all sufficiently large k, $T(\alpha(k)) < z$ which is a contradiction. Hence A_z is closed and so $T(\alpha)$ is upper semicontinuous. We can summarize our results as a theorem.

THEOREM 11. *Suppose that* (1) *and* (2) *satisfy the assumptions of* § 2 *and also suppose that condition* (C) *holds. Then there exist* $\alpha^* \in A(x_0, y_0)$, $\beta^* \in B(x_0, y_0)$ *such that for all* $\alpha \in A(x_0, y_0)$, $\beta \in B(x_0, y_0)$,

$$\begin{aligned} \sup_{(x,y) \in O(\alpha,\beta^*)} t(x, y) \leqq \sup_{(x,y) \in O(\alpha^*,\beta^*)} t(x, y) &= T^* \\ &= \inf_{(x,y) \in O(\alpha^*,\beta^*)} t(x, y) \\ &\leqq \inf_{(x,y) \in O(\alpha^*,\beta)} t(x, y). \end{aligned}$$

7. An example.

The system of equations is

$$\begin{aligned} \dot{x} &= u, \qquad |u| \leqq 1, \\ \dot{y} &= v, \qquad |v| \leqq 1, \end{aligned}$$

$x(0) = y(0) = 0$, final time $T = 1$. x, y, u, v, are real numbers; x is the state of player I, y is the state of player II. The payoff μ is just a function of the final states $x(1)$, $y(1)$ and is given by:

$$\mu(x, y) = \begin{cases} |x(1)| & \text{for } x(1)y(1) \geqq 0, \\ (1 - |y(1)|)|x(1)| & \text{for } x(1)y(1) \leqq 0. \end{cases}$$

Consider the strategy β^* for II given by $\beta(x) = -x$ for all $x \in X_1$. Then

$$\mu(x, \beta^*(x)) \leqq \tfrac{1}{4}.$$

Let $\alpha^* : Y_1 \to X_1$ be the strategy given by

$$\alpha^*(y)(t) = \begin{cases} y(t) & \text{for } t \leqq \frac{1}{2}, \\ y(\frac{1}{2}) + t & \text{for } t > \frac{1}{2} \text{ if } y(\frac{1}{2}) \geqq 0, \\ y(\frac{1}{2}) - t & \text{for } t > \frac{1}{2} \text{ if } y(\frac{1}{2}) < 0. \end{cases}$$

Then for all $y \in Y_1$,

$$\mu(\alpha^*(y), y) \geqq \tfrac{1}{4}.$$

[1] $\rho\{(x, y, t), F\} = \min\{|x - \hat{x}| + |y - \hat{y}| + |t - \hat{t}| \,|\, (\hat{x}, \hat{y}, \hat{t}) \in F\}$.

Evidently (α^*, β^*) are optimal. Furthermore, α^* is not continuous, although it can be approximated by continuous strategies; moreover, every continuous strategy is inferior to α^*.

8. Appendix. Proof of Theorem 2. Without loss of generality, we set $T = 1$ so that $\delta T = \delta$. Let $\delta > 0$ and let $u^* \in U$ be fixed. We construct the map Π_δ^X as follows. For each $x \in X_T$ choose an admissible control u^x such that:

(a) $\dot{x}(t) = f(x(t), u^x(t), t)$ a.e.,

(b) if x_1, x_2 are in X_T and $x_1(\tau) = x_2(\tau)$ for $0 \leqq \tau \leqq t$, then $u^{x_1}(\tau) = u^{x_2}(\tau)$ for $0 \leqq \tau \leqq t$.

Let u_δ^x be the admissible control defined by

$$u_\delta^x(\tau) = \begin{cases} u^* & \text{for } \quad 0 \leqq \tau \leqq \delta, \\ u^x(\tau - \delta) & \text{for } \quad \delta < \tau \leqq T. \end{cases}$$

Finally, for $x \in X_T$ let $\Pi_\delta^X(x) \in X_T$ be the solution of the differential equation corresponding to the control u_δ^x and the initial condition $\Pi_\delta^X(x)(0) = x(0)$.

Evidently the map Π_δ^X has properties (i) and (ii) of Theorem 2. The next proposition shows that Π_δ^X also has property (iii).

PROPOSITION. *For any admissible control u and $\delta > 0$ let u_δ be the admissible control defined by $u_\delta(\tau) = u^*$ for $0 \leqq \tau \leqq \delta$ and $u(\tau - \delta)$ for $\delta < \tau \leqq T$. Let $x \in X_T$ and $z \in X_T$ be defined by*

$$\dot{x}(t) = f(x(t), u(t), t), \quad x(0) = x_0,$$

$$\dot{z}(t) = f(z(t), u_\delta(t), t), \quad z(0) = x_0.$$

Then $\|x - z\|$ converges to zero as δ converges to 0 uniformly for all $x_0 \in X_0$ and all admissible controls u.

Proof. For $t \geqq \delta$,

$$x(t) - x(\delta) = \int_\delta^t f(x(\tau), u(\tau), \tau)\, d\tau,$$

$$z(t) - z(\delta) = \int_\delta^t f(z(\tau), u(\tau - \delta), \tau)\, d\tau,$$

so that

$$\begin{aligned} |x(t) - z(t)| &\leqq |x(\delta) - z(\delta)| + \left| \int_\delta^t [f(x(\tau), u(\tau), \tau) - f(z(\tau), u(\tau - \delta), \tau)] d\tau \right| \\ &\leqq |x(\delta) - z(\delta)| + \left| \int_\delta^t [f(x(\tau), u(\tau), \tau) - f(x(\tau - \delta), u(\tau - \delta), \tau - \delta)] d\tau \right| \\ &\quad + \int_\delta^t |f(x(\tau - \delta), u(\tau - \delta), \tau - \delta) - f(x(\tau), u(\tau - \delta), \tau)|\, d\tau \\ &\quad + \int_\delta^t |f(x(\tau), u(\tau - \delta), \tau) - f(z(\tau), u(\tau - \delta), \tau)|\, d\tau \\ &= I_1 + I_2 + I_3 + I_4, \quad \text{say.} \end{aligned}$$

We now obtain an upper bound for I_j, $1 \leqq j \leqq 3$.

LEMMA A1. *For $\varepsilon > 0$ there is a $\delta^* > 0$ such that for $\delta < \delta^*$,*

$$|x(\tau) - z(\tau)| \leqq \varepsilon, \qquad 0 \leqq \tau < \delta.$$

Proof. Apply the Bellman–Gronwall lemma using the boundedness assumption on f.

LEMMA A2. *For $\varepsilon > 0$ there is a $\delta^* > 0$ such that for $\delta < \delta^*$,*

$$I_2 \leqq \varepsilon.$$

Proof. Clearly $I_2 \leqq |x(t) - x(t - \delta)| + |x(\delta) - x(0)|$. Again use the Bellman–Gronwall lemma.

Let $B \subset R^n$ be a ball of radius b such that $x(t) \in B$ for all $x \in X_T$ and all $t \in [0, T]$.

LEMMA A3. *Given $\hat{\varepsilon} > 0$, there exists a $\delta^* > 0$ such that for $\delta < \delta^*$ there is a subset $F_\delta \subset [0, T]$ with measure $(F_\delta) > T - 2\hat{\varepsilon} - \delta$ and*

$$I_3 \leqq \hat{\varepsilon}T + 2(M + Nb)\int_{[0,T]\cap F_\delta^c} l(\tau)\,d\tau,$$

where F^c is the complement of F.

Proof. By an extension of a theorem due to Scorza-Dragoni [9], there exists a closed set $E \subset [0, T]$ with measure $(E) > T - \hat{\varepsilon}$ such that $f(x, u, t)$ is continuous on $B \times U \times E$. Therefore there exists a $\delta^* > 0$ such that for all $\delta < \delta^*$,

$$|f(x(\tau - \delta), u(\tau - \delta), \tau - \delta) - f(x(\tau), u(\tau - \delta), \tau)| < \hat{\varepsilon}$$

whenever $\tau \in E$, $(\tau - \delta) \in E$. Let $F_\delta = \{\tau | \tau \in E, (\tau - \delta) \in E\}$; then measure $(F_\delta) > T - 2\hat{\varepsilon} - \delta$. Also

$$\begin{aligned} I_3 &\leqq \int_\delta^T |f(x(\tau - \delta), u(\tau - \delta), \tau - \delta) - f(x(\tau), u(\tau - \delta), \tau)|\,d\tau \\ &\leqq \int_{[\delta,T]\cap F_\delta} |f(x(\tau - \delta), u(\tau - \delta), \tau - \delta) - f(x(\tau), u(\tau - \delta), \tau)|\,d\tau \\ &\quad + 2\int_{[\delta,T]\cap F_\delta^c} (M + Nb)l(\tau)\,d\tau \leqq \hat{\varepsilon}T + 2(M + Nb)\int_{[0,T]\cap F_\delta^c} l(\tau)\,d\tau. \end{aligned}$$

Proof of Proposition. Let $\varepsilon > 0$. Then from Lemmas A1–A3, there is a $\delta^* > 0$ such that for $\delta < \delta^*$, $I_1 + I_2 + I_3 \leqq 3\varepsilon$. Therefore,

$$\begin{aligned} |x(t) - z(t)| &\leqq 3\varepsilon + \int_\delta^t |f(x(\tau), u(\tau - \delta), \tau) - f(z(\tau), u(\tau - \delta), \tau)|\,d\tau \\ &\leqq 3\varepsilon + \int_\delta^t k(\tau)|z(\tau) - x(\tau)|d\tau. \end{aligned}$$

Hence by the Bellman–Gronwall lemma,

$$|x(t) - z(t)| \leqq 3\varepsilon \exp\left(\int_\delta^T k(\tau)\,d\tau\right).$$

The proposition follows.

REFERENCES

[1] W. H. Fleming, *A note on differential games of prescribed duration*, Contributions to the Theory of Games, vol. 3, Ann. of Math. Studies no. 39, Princeton University Press, Princeton, 1957, pp. 407–416.

[2] P. P. Varaiya, *The existence of solutions to a different game*, this Journal, 5 (1967), pp. 153–162.

[3] E. Roxin, *On Varaiya's definition of a differential game*, presented at the U.S.–Japan Seminar on Differential and Functional Equations, Minneapolis, 1967.

[4] W. H. Fleming, *The convergence problem for differential games*, J. Math. Anal. Appl., 3 (1961), pp. 102–116.

[5] C. Ryll-Nardzewski, *A theory of pursuit and evasion*, Advances in Game Theory, Princeton University Press, Princeton, 1964, pp. 113–126.

[6] D. Eggert and P. Varaiya, *Representation of a differential system*, Memo M 177, Electronic Research Laboratory, University of California, Berkeley; J. Differential Equations, to appear.

[7] E. Roxin, *The existence of optimal controls*, Michigan Math. J., 9 (1962), pp. 109–119.

[8] W. H. Fleming, *The convergence problem for differential games II*, Advances in Game Theory, Princeton University Press, 1964, pp. 195–210.

[9] G. S. Goodman, *On a theorem of Scorza–Dragoni and its application to optimal control*, Mathematical Theory of Control, Academic Press, New York, 1967, pp. 222–233.

SIAM J. Control
Vol. 7, No. 1, February 1969

ON THE APPROXIMATION OF INTEGRALS OF MULTIVALUED FUNCTIONS*

MARC Q. JACOBS†

1. Introduction. Let X denote a finite-dimensional real inner product space. The inner product of two elements $x, y \in X$ is denoted by $\langle x, y\rangle$. A norm $\|\cdot\|$ on X is defined by $\|x\|^2 = \langle x, x\rangle$ for $x \in X$. The metric $(x, y) \to \|x - y\|$, $x, y \in X$, will be denoted by ρ. Following Michael [23], we denote the collection of nonempty closed subsets of X by 2^X. T denotes a compact metric space on which a positive Radon measure μ is defined [5]. The collection of all integrable functions $f: T \to X$ is denoted by $\mathscr{L}(T, X, \mu)$, or simply by $\mathscr{L}$ when no confusion can arise. A measurable mapping [7], $\Omega: T \to 2^X$ is *integrable* if there is an $f \in \mathscr{L}$ such that $f(t) \in \Omega(t)$ for each $t \in T$. In this case we define $\int_T \Omega\, d\mu$ to be the set

$$\left\{\int_T f\, d\mu \,\middle|\, f \in \mathscr{L} \text{ and } f(t) \in \Omega(t) \text{ for all } t \in T\right\}.$$

This is a slightly specialized version of the integral of a multivalued function which is discussed by Aumann [1]. Aside from the applications in economics [11], such integrals are related in a natural way (see Remark 6) to the so-called attainable sets for linear control systems of the type studied in [24], [25], [26] and [19].

For integrable mappings $\Omega: T \to 2^X$, we cannot in general infer that $\int_T \Omega\, d\mu$ is closed (see [19, p. 50] and Example 1). In this paper we examine the problem of "approximating" the closure of $\int_T \Omega\, d\mu$. If Ω is majorized by a function $\phi \in \mathscr{L}(T, R, \mu)$ (R denotes the set of real numbers), and if T is nonatomic, then $\int_T \Omega\, d\mu$ is compact and convex [1], [7]. In this situation $\Omega(t)$ is compact for each $t \in T$, and the results which we obtain differ only slightly from results already obtained by Aumann [1], Castaing [7] and Debreu [11]. However, in the unbounded case, there does appear to be some novelty in our study of the integral $\int_T \Omega\, d\mu$. If x_0 is a given point of X, the optimal final value control problem for linear systems [2], [14] can be viewed as a special case of the problem of determining $x \in X$ and $f \in \mathscr{L}(T, X, \mu)$ satisfying

$$x = \int_T f d\mu \in \int_T \Omega\, d\mu,$$

* Received by the editors February 22, 1968, and in final revised form December 2, 1968.

† Center for Dynamical Systems, Applied Mathematics Division, Brown University, Providence, Rhode Island 02912. This research was supported by the National Science Foundation under Grant GU-1153 while the author was in the Mathematics Department at Rice University, Houston, Texas.

where $f(t) \in \Omega(t)$ for each $t \in T$, and such that $\rho(x_0, x) = \rho\left(x_0, \int_T \Omega\, d\mu\right)$ (see [4, Part 2, p. 149] for notation). Such a pair x, f will evidently exist if $\int_T \Omega\, d\mu$ is closed. But whether $\int_T \Omega\, d\mu$ is closed or not, there is the "relaxed problem" of estimating $\rho\left(x_0, \int_T \Omega\, d\mu\right)$, i.e., determining minimizing sequences $\{x_n\}$ and $\{f_n\}$ such that $x_n = \int_T f_n\, d\mu$ and such that $\rho(x_0, x_n) \to \rho\left(x_0, \int_T \Omega\, d\mu\right)$ as $n \to \infty$. We suggest one possible solution to this problem. Insofar as our results are related to the optimal final value control problem our results could be considered to be in the spirit of [6], [9], [10] and [17] in the sense that the method consists in determining a sequence of related but simpler optimization problems whose solutions give the required minimizing sequence. However, other than this, our results have little overlap with these papers.

Finally we mention the interesting problem of determining sets of sufficient conditions for the set $\int_T \Omega\, d\mu$ to be closed. An answer to this has been known for some time [24], [26] for certain classes of mappings Ω taking their values in the space of compact subsets of X. Somewhat more general situations were discussed in [19] and [27]. However, suitable criteria guaranteeing the closedness (and not at the same time implying compactness) of $\int_T \Omega\, d\mu$ for a reasonably generous class of mappings $\Omega : T \to 2^X$ seem still to be unavailable.[1] In [19] we made a conjecture concerning this problem which turned out to be precipitous in view of Examples 1, 2 and 3 of this paper. It appears that the approximation theorems of § 3 will be useful in attacking the above problem, and we shall pursue this matter in a later paper.

2. Additional notation and definitions. Throughout this paper μ denotes a positive Radon measure defined on a compact metric space T(see [5]). In discussing multivalued mappings we shall generally use the terminology of Michael [23]. We shall use $\mathscr{A}(X)$ to denote the collection of nonempty subsets of X, and, as stated in the introduction, 2^X to denote the collection of nonempty closed subsets of X. $\mathscr{C}(X)$ will denote the collection of nonempty compact subsets of X. If $H : T \to \mathscr{A}(X)$ is a mapping, and if S is a subset of X, then we define H^-S to be the set $\{t \in T | H(t) \cap S \neq \varnothing\}$. We say the multivalued mapping H is *measurable* if H^-F is measurable for every closed $F \subset X$. If d is any metric on X, then $\mathscr{I}^d$ is a basis for the uniform structure [4, Part 1] determined by the metric d, where

[1] Recent results of C. Olech [28], [29] do have applications to this question, and Professor Olech has recently informed the author that he is preparing a further note on this subject.

$\mathscr{J}^d = \{J^d_\varepsilon \mid \varepsilon > 0\}$ and

$$J^d_\varepsilon = \{(x, y) \in X \times X \mid d(x, y) < \varepsilon\}.$$

If V is a subset of $X \times X$, and A is a subset of X, then $V[A]$ denotes the set

$$V[A] = \{y \in X \mid \exists x \in A : (x, y) \in V\}.$$

The uniform structure generated by $\mathscr{J}^d$ determines a uniform structure on $\mathscr{A}(X)$ [23, pp. 153 and 167] which we denote by 2^d. Let $W(J^d_\varepsilon)$ denote the set

$$\{(A, B) \in \mathscr{A}(X) \times \mathscr{A}(X) \mid J^d_\varepsilon[A] \supset B \text{ and } J^d_\varepsilon[B] \supset A\}$$

for $\varepsilon > 0$. A basis for the uniform structure 2^d is simply $\{W(J^d_\varepsilon) \mid \varepsilon > 0\}$. The topology on $\mathscr{A}(X)$ determined by 2^d is called the *uniform topology on* $\mathscr{A}(X)$ *determined by d*. Also $2^d \cap (2^X \times 2^X)$ (respectively $2^d \cap (\mathscr{C}(X) \times \mathscr{C}(X))$) is the relative uniform structure on 2^X (respectively $\mathscr{C}(X)$) determined by d, and by abuse of language these two uniform spaces will be denoted by $(2^X, 2^d)$ (respectively $(\mathscr{C}(X), 2^d)$). The topology on 2^X (respectively $C(X)$) determined by 2^d is called the *uniform topology on* 2^X (respectively $\mathscr{C}(X)$) *determined by d*. The uniform space $(\mathscr{A}(X), 2^d)$ is pseudo-metrizable, but in general it is not Hausdorff and therefore not metrizable. We note that limits in $(\mathscr{A}(X), 2^d)$ are not unique, i.e., if a sequence $\{A_n\}$ in $\mathscr{A}(X)$ converges to A in $\mathscr{A}(X)$, then the sequence also converges to $\mathrm{cl}(A)$ (the closure of A). The uniform space $(2^X, 2^d)$ (and therefore also $(\mathscr{C}(X), 2^d)$) is metrizable [23, Proposition 4.1] with the Hausdorff metric [3] determined by $d^* = d/(1 + d)$. However, it will be more convenient and more effective computationally simply to view these spaces as uniform spaces. We require two additional uniform structures on $\mathscr{A}(X)$. The *upper* (respectively *lower*) *semiuniform structure on* $\mathscr{A}(X)$ *determined by d* has as basis the collection

$$\{(A, B) \in \mathscr{A}(X) \times \mathscr{A}(X) \mid J^d_\varepsilon[A] \supset B\}, \qquad \varepsilon > 0$$

(respectively $\{(A, B) \in \mathscr{A}(X) \times \mathscr{A}(X) \mid J^d_\varepsilon[B] \supset A\}, \varepsilon > 0$) (cf. [23, p. 181]). The corresponding topologies are called the *upper* (respectively *lower*) *semiuniform topologies on* $\mathscr{A}(X)$ *determined by d*.

The symbol $d(x, A)$, for $x \in X$ and $A \subset X$, which has already been alluded to in the Introduction, denotes $\inf\{d(x, a) \mid a \in A\}$.

A mapping $H: T \to \mathscr{A}(X)$ is *simple* if there is a finite partition of T, say $\{T_1, \cdots, T_n\}$, such that H has a constant value on each of the T_i. If the sets T_i are each measurable, then H is a measurable simple function, i.e., $H^- F$ is measurable for every closed $F \subset X$. If $H(t) = F_i$ for $t \in T_i$, $i = 1, 2, \cdots, n$, then we shall denote the simple function H by $\{T_1, \cdots, T_n; F_1, \cdots, F_n\}$.

We shall require two metrics on X. One is Euclidean distance, $\rho(x, y) = \|x - y\|$, $x, y \in X$, and the other is a metric which determines the topology of the one-point compactification of X [4, Part 1, p. 92]. Specifically we choose the metric ρ_∞ on X determined by ρ and stereographic projection on the Riemann sphere, viz.,

$$\rho_\infty(x, y) = \frac{\|x - y\|}{[1 + \|x\|^2]^{1/2}[1 + \|y\|^2]^{1/2}}, \qquad x, y \in X.$$

These two metrics define the same topology on X; however, they do not define the same uniform structure on X. What is more important, the uniform topologies on 2^X (or $\mathscr{A}(X)$) determined by ρ and ρ_∞ are not equivalent [21]. The uniform topologies on $C(X)$ determined by ρ and ρ_∞ are equivalent [21]. The uniform topology on 2^X (or $\mathscr{A}(X)$) determined by ρ_∞ is weaker than the uniform topology on 2^X(or $\mathscr{A}(X)$) determined by ρ (see [21, Remark (iii)]). The information we shall need concerning ρ_∞ and 2^{ρ_∞} can be found in [21]. In many practical aspects (cf. [21]) 2^{ρ_∞} is well-suited for discussing the convergence of sequences of unbounded sets in X.

We shall frequently refer to the entourages J_ε^ρ and $J_\varepsilon^{\rho_\infty}$, $\varepsilon > 0$. For simplicity we shall agree that when J_ε^ρ is meant, ρ will be suppressed and we shall only write J_ε; when $J_\varepsilon^{\rho_\infty}$ is meant we shall write J_ε^∞.

We shall reserve the symbol ω to denote the collection of positive integers.

3. Approximation theorems. The first theorem gives a criterion for the integrability of a measurable mapping $\Omega : T \to 2^X$ (cf. [1, Theorem 2]). A simple lemma is needed to prove the theorem.

LEMMA 1. *Let F be a nonempty closed subset of X. Then for each $x \in X$ there exists $f_x \in F$ such that $\rho(x, F) = \rho(x, f_x)$.*

Remark 1. That the metric ρ has the useful property mentioned in Lemma 1 is well known. This situation does not obtain if the metric ρ_∞ is substituted in the lemma. Indeed, let F denote the set $\{(0, y) \in R^2 | y \geqq 0\} \subset R^2$, and let x denote the point $(0, -n) \in R^2$, $n > 1$. Then for each $f \in F$ we have the inequality $\rho_\infty(x, f) > 1/(1 + n^2)^{1/2}$. However, the infimum of the collection $\{\rho_\infty(x, f) | f \in F\}$ is evidently $1/(1 + n^2)^{1/2}$, and consequently there is no $f \in F$ such that $\rho_\infty(x, f) = \rho_\infty(x, F)$.

THEOREM 1. *Let Ω be a measurable mapping $\Omega : T \to 2^X$. In order that Ω be integrable it is necessary and sufficient for the mapping $\alpha : T \to R$ to be integrable, where α is defined by the relation $\alpha(t) = \rho(0, \Omega(t))$, $t \in T$ (the zero element of X is denoted by* 0).

Proof. This result in essence appears in [20]. We note that the set $\{t \in T | \alpha(t) < r\}$, $r > 0$, is equal to the set $\Omega^- J_r[0]$, which is measurable. Therefore α is measurable. If Ω is integrable, then there exists $f \in \mathscr{L}(T, X, \mu)$ such that $f(t) \in \Omega(t)$ for each $t \in T$. From the definition of α it follows that $0 \leqq \alpha(t) \leqq \|f(t)\|$ for each $t \in T$, and therefore α is integrable. Conversely suppose α is integrable, and define a mapping $\Gamma : T \to \mathscr{C}(X)$ by the relation

$$\Gamma(t) = \{x \in \Omega(t) | \|x\| = \alpha(t)\}, \qquad t \in T.$$

That $\Gamma(t)$ is nonempty for each $t \in T$ follows from Lemma 1. The compactness of each $\Gamma(t)$ follows immediately from the fact that each $\Omega(t)$ is closed. The mapping Γ is measurable. The proof of this is simple but tedious (all that is needed is Theorem 2.3 in [21]) and will not be given. Now let $\xi = (x_i)$ be any basis for the vector space X. For each $t \in T$ define $f(t)$ to be the lexicographical minimum of $\Gamma(t)$ with respect to the basis ξ (see [27]). The mapping $t \to f(t)$, $t \in T$, is integrable [21, Lemma 2.5], and $f(t) \in \Omega(t)$ for each $t \in T$.

THEOREM 2. *Let $\Omega : T \to 2^X$ and $f : T \to X$ be measurable mappings. Then the mapping $t \to \rho(f(t), \Omega(t))$, $t \in T$, is measurable, and there exists a measurable mapping $g : T \to X$ such that:* (i) *$g(t) \in \Omega(t)$ for each $t \in T$, and* (ii) *$\rho(f(t), g(t)) = \rho(f(t), \Omega(t))$.*

Proof. The mapping $(t, x) \to \rho(x, \Omega(t))$, $(t, x) \in T \times X$, fulfills the conditions of Theorem 2.1 in [18, p. 623] (see [21] for the details). Thus given $\varepsilon > 0$ there exists a compact set $T_\varepsilon^1 \subset T$ such that $\mu(T \backslash T_\varepsilon^1) < \varepsilon/2$ and such that the mapping $(t, x) \to \rho(x, \Omega(t))$ restricted to $T_\varepsilon^1 \times X$ is continuous. Since $f : T \to X$ is measurable there is also a compact $T_\varepsilon^2 \subset T$ such that $\mu(T \backslash T_\varepsilon^2) < \varepsilon/2$ and such that $f | T_\varepsilon^2$ is continuous. Thus if T_ε denotes $T_\varepsilon^1 \cap T_\varepsilon^2$, then $\mu(T \backslash T_\varepsilon) < \varepsilon$ and both $f | T_\varepsilon$ and the restriction of $(t, x) \to \rho(x, \Omega(t))$ to $T_\varepsilon \times X$ are continuous. Consequently the restriction of the mapping $t \to \rho(f(t), \Omega(t))$ to T_ε is continuous. The $\varepsilon > 0$ is arbitrary, and the measurability of $t \to \rho(f(t), \Omega(t))$, $t \in T$, follows [5, p. 169]. That a mapping $g : T \to X$ exists satisfying (i) and (ii) is clear. That such a g can be chosen which is also measurable follows from extensions of Filippov's implicit functions lemma (see [7], [21] or [22]).

Remark 2. If in Theorem 2 the mapping Ω takes its values in $\mathscr{C}(X)$, then Theorem 2 remains valid when ρ is replaced by ρ_∞(cf. Remark 1).

LEMMA 2. *Let A be a nonempty closed subset of X. Let K_n, $n \in \omega$, be an increasing sequence of compact subsets of X such that $\bigcup_{n \in \omega} K_n = X$. Then the sequence*

$$\Gamma_n = \begin{cases} A \cap K_n & \text{if } \ A \cap K_n \neq \varnothing, \quad n \in \omega, \\ \{x_0\} & \text{otherwise}, \end{cases}$$

where x_0 is an arbitrary point of X, has the property that Γ_n converges to A in $(2^X, 2^{\rho_\infty})$.

Proof. The metric space (A, ρ_∞) is precompact. Thus given $\varepsilon > 0$ there exist $a_1, a_2, \cdots, a_k \in A$ such that $\bigcup_{i=1}^k J_{\varepsilon/2}^\infty[a_i] \supset A$. There exists $N_0 \in \omega$ such that $a_1, a_2, \cdots, a_k \in K_n$ whenever $n \geqq N_0$. Thus for $n \geqq N_0$ we find that

$$(1) \qquad J_{\varepsilon/2}^\infty[\Gamma_n] \supset A.$$

The dual inequality,

$$(2) \qquad J_{\varepsilon/2}^\infty[A] \supset \Gamma_n, \qquad n \geqq N_0,$$

follows immediately from the definition of Γ_n. The two inequalities (1) and (2) combine to complete the proof of the theorem.

Remark 3. In R the sequence of compact intervals $K_n = [-n, n]$, $n \in \omega$, certainly does not converge to R in $(2^R, 2^\rho)$. But K_n does converge to R in $(2^R, 2^{\rho_\infty})$.

THEOREM 3 (Egorov). *Let $\Omega_n, n \in \omega$, be a sequence of measurable mappings from T into 2^X such that $\Omega_n(t) \to \Omega(t)$ a.e. on T as $n \to \infty$ in $(2^X, 2^{\rho_\infty})$, where Ω is a mapping $\Omega : T \to 2^X$. Then*

(i) *Ω is measurable;*

(ii) *for every $\varepsilon > 0$ there exists a compact $T_\varepsilon \subset T$ such that $\mu(T \backslash T_\varepsilon) < \varepsilon$ and $\Omega_n(t)$ converges to $\Omega(t)$ in $(2^X, 2^{\rho_\infty})$ uniformly on T_ε.*

Proof. By Theorem 2.4 in [21] the functions Ω_n, $n \in \omega$, are measurable in the Bourbaki sense [5] when treated as mappings of T into the metrizable space $(2^X, 2^{\rho_\infty})$. Thus Egorov's theorem [5, p. 175] applies to give the desired conclusions.

The uniform topologies on $\mathscr{C}(X)$ determined by ρ and ρ_∞ are identical [23, Theorem 3.3]. Thus in the next theorem limits can be taken with respect to either topology.

THEOREM 4 (Dominated convergence theorem). *Let T be nonatomic. Let Ω_n, $n \in \omega$, be integrable mappings from T into $\mathscr{C}(X)$ such that* $\sup\{\|x\| \,|\, x \in \Omega_n(t)\} \leqq \phi(t)$ *for each $n \in \omega$ and each $t \in T$, where $\phi \in \mathscr{L}(T, R, \mu)$. $\mathscr{C}(X)$ can have either of the equivalent topologies determined by 2^ρ or 2^{ρ_∞}. If $\Omega_n(t) \to \Omega(t)$ as $n \to \infty$ a.e. on T, where Ω is a mapping $\Omega: T \to \mathscr{C}(X)$, then*

(i) *Ω is integrable;*

(ii) $\int_T \Omega \, d\mu$ *and* $\int_T \Omega_n \, d\mu$, *$n \in \omega$, belong to $\mathscr{C}(X)$;*

(iii) $\int_T \Omega_n \, d\mu \to \int_T \Omega \, d\mu$ *as $n \to \infty$.*

Proof. The measurability of Ω results from Theorem 3. The convergence $\Omega_n(t) \to \Omega(t)$ a.e. on T also gives

$$\sup\{\|x\| \,|\, x \in \Omega(t)\} \leqq \phi(t) \qquad \text{a.e. on } T. \tag{3}$$

Since Ω is measurable there is a measurable $f: T \to X$ such that $f(t) \in \Omega(t)$ for each $t \in T$ [21, Lemma 2.5]. The function f is also integrable by (3). Thus Ω is integrable. For conclusion (ii) we refer to Castaing [7, Theorem 7.1].

For the final conclusion, we observe that the set function $E \subset T \to \int_E 2\phi \, d\mu$, E measurable, is absolutely continuous with respect to μ (see [12]). Thus given $\varepsilon > 0$ there is a $\delta(\varepsilon) > 0$ such that $E \subset T$, E measurable and $\mu(E) < \delta(\varepsilon)$ imply $0 \leqq \int_E 2\phi \, d\mu < \varepsilon/2$.

By Theorems 3 (Egorov's theorem), there is a compact $T_\varepsilon \subset T$ such that $\mu(T \backslash T_\varepsilon) < \lambda$, where $\lambda = \min(\delta(\varepsilon), \varepsilon/(2(\mu(T) + 1)))$, and such that $\Omega_n(t) \to \Omega(t)$ as $n \to \infty$ uniformly on T_ε. Thus there exists $n_0 \in \omega$ such that $n \geqq n_0$ implies

$$J_\lambda[\Omega_n(t)] \supset \Omega(t) \quad \text{and} \quad J_\lambda[\Omega(t)] \supset \Omega_n(t), \qquad t \in T_\varepsilon. \tag{4}$$

If f is an element of $\mathscr{L}(T, X, \mu)$ such that $f(t) \in \Omega(t)$ for each $t \in T$, then for each $n \in \omega$ there exists $g_n \in \mathscr{L}(T, X, \mu)$ such $g_n(t) \in \Omega_n(t)$, $t \in T$, and $\rho(f(t), g_n(t)) = \rho(f(t), \Omega_n(t))$ for each $t \in T$ (by Theorem 2). Hence by (4) we have that $\|f(t) - g_n(t)\| < \lambda$ whenever $n \geqq n_0$ and $t \in T_\varepsilon$. Therefore whenever $n \geqq n_0$ it follows that

$$\begin{aligned} \left\| \int_T (f - g_n) \, d\mu \right\| &\leqq \int_{T_\varepsilon} \|f - g_n\| \, d\mu + \int_{T \backslash T_\varepsilon} \|f - g_n\| \, d\mu \\ &< \lambda\mu(T) + \varepsilon/2 \\ &\leqq \varepsilon/2 + \varepsilon/2 = \varepsilon. \end{aligned}$$

We thereby infer

$$J_\varepsilon\left[\int_T \Omega_n\, d\mu\right] \supset \int_T \Omega d\mu \qquad \text{whenever } n \geqq n_0. \tag{5}$$

The dual inequality

$$J_\varepsilon\left[\int_T \Omega\, d\mu\right] \supset \int_T \Omega_n\, d\mu \qquad \text{whenever } n \geqq n_0 \tag{5'}$$

is proved similarly with the aid of (4) and Theorem 2. We omit the details.

It is clear that if we only want (iii) in Theorem 4 to be true in $(\mathscr{A}(X), 2^\rho)$, then the assumption that T is nonatomic is unnecessary.

The main aspects of Theorem 4 were proved by Debreu [11] by using Rådstrom's imbedding theorem [30]. Indeed, the mappings $t \to \Omega_n(t)$, $n \in \omega$, are measurable by assumption, and thus $t \to \mathrm{co}\,(\Omega_n(t))$ ($\mathrm{co}(\Omega_n(t))$ denotes the convex hull of $\Omega_n(t)$), $n \in \omega$, are measurable by Theorem 6.2 of [7]. Since the Ω_n, $n \in \omega$, are majorized by an integrable function, the same is true of the mappings $t \to \mathrm{co}\,(\Omega_n(t))$, $n \in \omega$. Thus the conclusions of Theorem 4 could also be achieved by combining Corollary 16 of [13, p. 151] with Propositions 6.5, 7.2 and inequality (5.4) of [11]. The reader should bear in mind that we are in general using "integrable" in a different sense than Debreu, but for the situation in Theorem 4 the two notions of "integrability" agree.

THEOREM 5. *Let K_n, $n \in \omega$, be an increasing sequence of compact subsets of X such that $\bigcup_{n\in\omega} K_n = X$. Let Ω denote a measurable mapping $\Omega : T \to 2^X$ such that for all sufficiently large n, $\Omega(t) \cap K_n$ is nonempty for each $t \in T$. Then*

(i) *Ω is integrable;*

(ii) *if $\Omega \cap K_n$ denotes the mapping $t \to \Omega(t) \cap K_n$, $t \in T$, then for all sufficiently large n, $\Omega \cap K_n$ is integrable, and*

$$\lim \int_T \Omega \cap K_n\, d\mu = \int_T \Omega\, d\mu,$$

where the limit is taken in $(\mathscr{A}(X), 2^{\rho_\infty})$.

Proof. We cannot infer that $\int_T \Omega\, d\mu$ is closed, and thus the limit must be taken in $(\mathscr{A}(X), 2^{\rho_\infty})$. However, once the theorem is established one also obtains (if T is nonatomic)

$$\lim \int_T \Omega \cap K_n\, d\mu = \mathrm{cl}\left(\int_T \Omega\, d\mu\right)$$

in $(2^X, 2^{\rho_\infty})$ (see [23]). That Ω is integrable follows from Theorem 1. There exists $m_0 \in \omega$ such that $n \geqq m_0$ implies $\Omega(t) \cap K_n \neq \varnothing$ for each $t \in T$. We assume in the rest of the proof that $n \geqq m_0$. Then we have $\Omega \cap K_n$ is integrable. Also we have, for every $\varepsilon > 0$,

$$J_\varepsilon^\infty\left[\int_T \Omega\, d\mu\right] \supset \int_T \Omega \cap K_n\, d\mu. \tag{6}$$

Thus it suffices for us to show that for each $\varepsilon > 0$ there exists $n_0 \in \omega$ such that $n \geqq n_0$ implies

$$(6') \qquad J_\varepsilon^\infty\left[\int_T \Omega \cap K_n \, d\mu\right] \supset \int_T \Omega \, d\mu .$$

In order to establish the inequality (6′) we first make the observation:

(7) For each $\varepsilon > 0$ and each $x \in \int_T \Omega \, d\mu$ there corresponds a $n(\varepsilon, x) \in \omega$ and a $y_{n(\varepsilon,x)} \in \int_T \Omega \cap K_n \, d\mu$ such that $\|x - y_{n(\varepsilon,x)}\| < \varepsilon$.

Let $x \in \int_T \Omega \, d\mu$ be given, with $x = \int_T g \, d\mu$ and $g(t) \in \Omega(t)$ for each $t \in T$. Let $\bar{f}$ be a measurable function $\bar{f}: T \to X$ such that $\bar{f}(t) \in \Omega(t) \cap K_{m_0}$ for each $t \in T$. Such an $\bar{f}$ exists [21, Lemma 2.5]. Then $\bar{f}$ is also integrable. We define a sequence $g_n \in \mathscr{L}(T, X, \mu)$, $n \in \omega$, by the relation

$$g_n(t) = \begin{cases} g(t) & \text{if } \ g(t) \in K_n, \\ \bar{f}(t) & \text{if } \ g(t) \notin K_n. \end{cases}$$

Then $g_n(t) \in \Omega(t) \cap K_n$ for each $t \in T$. Also $\|g_n - g\|$ is majorized on T by the integrable function $2\sup(\|\bar{f}\|, \|g\|)$, and $\lim \|g_n(t) - g(t)\| = 0$ for $t \in T$. Hence $\lim \int_T g_n \, d\mu = \int_T g \, d\mu$ and (7) results.

Now returning to the proof of (6′), we define $A_n = \int_T \Omega \cap K_n \, d\mu$ and $A = \int_T \Omega \, d\mu$. The set A is precompact in (X, ρ_∞). Thus given $\varepsilon > 0$ there exist $a_1, \cdots, a_k \in A$ such that

$$(8) \qquad \bigcup_{i=1}^{k} J_{\varepsilon/2}^\infty [a_i] \supset A.$$

We invoke (7) to determine $y_{m_i} \in A_{m_i}$, $i = 1, 2, \cdots, k$, such that

$$(9) \qquad \rho_\infty(y_{m_i}, a_i) \leqq \|y_{m_i} - a_i\| < \varepsilon/2, \qquad i = 1, 2, \cdots, k.$$

Let n_0 denote the maximum of $\{m_0, m_1, m_2, \cdots, m_k\}$. Then for $n \geqq n_0$, (6′) is obtained.

As a corollary to Theorems 4 and 5 we obtain the following result.

COROLLARY 1. *Let K_n, $n \in \omega$, be an increasing sequence of compact sets in X such that $\bigcup_{n \in \omega} K_n = X$. Let $\Omega: T \to 2^X$ be a mapping such that $\Omega(t) \cap K_n \neq \emptyset$, $t \in T$, $n \in \omega$, and such that the mappings $t \to \Omega(t) \cap K_n$, $t \in T$, $n \in \omega$, are continuous mappings into $\mathscr{C}(X)$ with the topology induced by 2^{ρ_∞}. Then there is a sequence of measurable simple functions $S_n: T \to \mathscr{C}(X)$, $n \in \omega$, such that*

(i) $\lim S_n(t) = \Omega(t)$ *in* $(2^X, 2^{\rho_\infty})$ *for each* $t \in T$, *and*

(ii) $\lim \int_T S_n \, d\mu = \int_T \Omega \, d\mu$ *in* $(\mathscr{A}(X), 2^{\rho_\infty})$.

Proof. The mappings $t \to \Omega(t) \cap K_n$ are continuous on T, and therefore uniformly continuous on T. Thus there is a sequence of measurable simple functions $H_{nm} : T \to \mathscr{C}(X)$, n, $m \in \omega$, such that $H_{nm}(t)$ converges to $\Omega(t) \cap K_n$ as $m \to \infty$ *uniformly* on T for each $n \in \omega$. By Theorems 4 and 5 there results

$$\lim_n \lim_m \int_T H_{nm}\, d\mu = \int_T \Omega\, d\mu$$

in $(\mathscr{A}(X), 2^{\rho_\infty})$. Hence S_n can be taken to be $H_{\tau(n)n}$ for an appropriately chosen sequence $\tau(n)$ in ω.

Let us denote by τ_l (respectively τ_u) the topology on $\mathscr{A}(X)$ induced by the lower (respectively upper) semiuniform structures on $\mathscr{A}(X)$ determined by ρ_∞. Evidently a sequence A_n, $n \in \omega$, in $\mathscr{A}(X)$ converging to A in both topologies τ_u and τ_l also converges to A in the uniform topology determined by ρ_∞.

COROLLARY 2. *Let K_n, $n \in \omega$, and Ω be given as in the hypotheses of Theorem 5. Let $\Omega_n : T \to 2^X$, $n \in \omega$, be a sequence of integrable functions such that for all sufficiently large $m \in \omega$, $\Omega_n(t) \cap K_m \to \Omega(t) \cap K_m$ a.e. on T as $n \to \infty$ in $(2^X, 2^{\rho_\infty})$. Then there results*

$$\lim \int_T \Omega_n\, d\mu = \int_T \Omega\, d\mu$$

in $\mathscr{A}(X)$'with topology τ_l.

Proof. Let $\varepsilon > 0$ be given. By Theorem 5 there exists $m_0 \in \omega$ such that

$$J^{\infty}_{\varepsilon/2}\left[\int_T \Omega \cap K_m\, d\mu\right] \supset \int_T \Omega\, d\mu \quad \text{whenever} \quad m \geqq m_0.$$

From Theorem 4 we can infer that there is an $n_0 \in \omega$ such that

$$J^{\infty}_{\varepsilon/2}\left[\int_T \Omega_n \cap K_{m_0}\, d\mu\right] \supset \int_T \Omega \cap K_{m_0} \quad \text{whenever} \quad n \geqq n_0.$$

Combining these two inclusion relationships it is determined that

$$J^{\infty}_{\varepsilon}\left[\int_T \Omega_n\, d\mu\right] \supset \int_T \Omega\, d\mu \quad \text{whenever} \quad n \geqq n_0.$$

Therefore $\lim \int_T \Omega_n\, d\mu = \int_T \Omega\, d\mu$ in $\mathscr{A}(X)$ with topology τ_l.

COROLLARY 3. *Let K_n, $n \in \omega$, and Ω be given as in the hypotheses of Theorem 5. Let $\Omega_n : T \to 2^X$, $n \in \omega$, be a sequence of integrable functions such that for all sufficiently large $m \in \omega$, $\Omega_n(t) \cap K_m \to \Omega(t) \cap K_m$ a.e. on T as $n \to \infty$ in $(2^X, 2^{\rho_\infty})$. Let $\int_T \Omega_n \cap K_m\, d\mu \to \int_T \Omega_n\, d\mu$ uniformly with respect to $n \in \omega$ as $m \to \infty$ in $(\mathscr{A}(X), 2^{\rho_\infty})$. Then there results*

$$\lim \int_T \Omega_n\, d\mu = \int_T \Omega\, d\mu$$

in $(\mathscr{A}(X), 2^{\rho_\infty})$.

Proof. In view of Corollary 2 it suffices to prove that

$$\lim \int_T \Omega_n \, d\mu = \int_T \Omega \, d\mu$$

in $\mathscr{A}(X)$ with topology τ_u. Given $\varepsilon > 0$ there exists $m_0 \in \omega$ such that

$$J^{\infty}_{\varepsilon/2}\left[\int_T \Omega_n \cap K_m\right] \supset \int_T \Omega_n \, d\mu, \qquad n \in \omega, \quad m \geqq m_0.$$

By Theorem 5 there exists $n_0 \in \omega$ such that

$$J^{\infty}_{\varepsilon/2}\left[\int_T \Omega \cap K_{m_0} \, d\mu\right] \supset \int_T \Omega_n \cap K_{m_0} \, d\mu, \qquad n \geqq n_0.$$

Upon combining these two inclusion relationships we obtain

$$J^{\infty}_{\varepsilon}\left[\int_T \Omega \, d\mu\right] \supset \int_T \Omega_n \, d\mu, \qquad n \geqq n_0,$$

thereby proving $\lim \int_T \Omega_n \, d\mu = \int_T \Omega \, d\mu$ in $\mathscr{A}(X)$ with topology τ_u.

THEOREM 6. *If $\Omega : T \to 2^X$ is an integrable mapping, and if $\Omega_n : T \to 2^X$, $n \in \omega$, is a sequence of integrable functions such that* $\lim \Omega_n(t) = \Omega(t)$ *uniformly on $T \backslash N$ in $(2^X, 2^\rho)$, where $\mu(N) = 0$, then*

$$\lim \int_T \Omega_n \, d\mu = \int_T \Omega \, d\mu$$

in $(\mathscr{A}(X), 2^\rho)$.

Proof. We may assume that $\mu(T) \neq 0$. Given $\varepsilon > 0$ there exists $n_0 \in \omega$ such that

$$(10) \qquad J_\delta[\Omega(t)] \supset \Omega_n(t) \quad \text{and} \quad J_\delta[\Omega_n(t)] \supset \Omega(t)$$

for $t \in T \backslash N$, whenever $n \geqq n_0$, where $\delta = \varepsilon/(2\mu(T))$. We shall show that

$$(11) \qquad J_\varepsilon\left[\int_T \Omega \, d\mu\right] \supset \int_T \Omega_n \, d\mu \quad \text{and} \quad J_\varepsilon\left[\int_T \Omega_n \, d\mu\right] \supset \int_T \Omega \, d\mu,$$

whenever $n \geqq n_0$. Suppose $x \in \int_T \Omega \, d\mu$, with $x = \int_T f \, d\mu$, $f \in \mathscr{L}(T, X, \mu)$, and $f(t) \in \Omega(t)$ for each $t \in T$. Then by Theorem 2 we have that corresponding to each $n \in \omega$ there is a measurable function $g_n : T \to X$ such that $\rho(f(t), g_n(t)) = \rho(f(t), \Omega_n(t))$ for each $t \in T$. From (10) and the definition of g_n we must have $\|f(t) - g_n(t)\| < \delta$, when $t \in T \backslash N$ and $n \geqq n_0$. Thus we also have that $\|g_n(t)\| < \|f(t)\| + \delta$ when $t \in T \backslash N$ and $n \geqq n_0$. We thereby obtain that the g_n, $n \geqq n_0$, are integrable on T. There results

$$\left\| x - \int_T g_n \, d\mu \right\| \leqq \int_T \|f - g_n\| \, d\mu \leqq \delta\mu(T) = \frac{\varepsilon}{2} < \varepsilon,$$

whenever $n \geqq n_0$. Hence the second inclusion in (11) is true for $n \geqq n_0$. The proof of the first inclusion in (11) is analogous with the aid of (10) and Theorem 2. We omit the details.

Remark 4. In Theorem 6 we can also infer that $\lim \mathrm{cl}\left(\int_T \Omega_n \, d\mu\right) = \mathrm{cl}\left(\int_T \Omega \, d\mu\right)$ in $(2^X, 2^\rho)$ (see [23]). We are interested primarily in approximations to $\int_T \Omega \, d\mu$, and thus we would like to apply Theorem 6 to the situation in which the Ω_n are measurable simple functions. The determination of such measurable simple Ω_n can be a difficult problem. In fact there are very simple examples of integrable Ω (as in Theorem 6) which cannot be approximated, even pointwise a.e. on T, in $(2^X, 2^\rho)$ by any sequence of measurable simple functions. Consider the following example given by Cesari [8, pp. 374–375], viz., $T = [0, 1]$, μ is Lebesgue measure, $X = R^2$, with $t \to \Omega(t) \in 2^{R^2}$ given by the relation

$$\Omega(t) = \{(x, y) \in R^2 | x \geqq 0 \text{ and } 0 \leqq y \leqq tx\}$$

for $t \in [0, 1]$. As a mapping from T into 2^{R^2} with the uniform topology determined by ρ, Ω is discontinuous at each $t \in [0, 1]$ (see [8]). Moreover, the discontinuities occur in such a way that if t_n, $n \in \omega$, is any sequence in $[0, 1]$ such that $t_n \neq t_0$, $n \in \omega$, and such that $t_n \to t_0$, then $\Omega(t_n)$ will not converge to $\Omega(t_0)$ in $(2^{R^2}, 2^\rho)$ (cf. [8]). Thus if T_ε is any compact subset of $[0, 1]$ which has positive measure, the mapping $\Omega | T_\varepsilon$ will be discontinuous at some points of T_ε. Hence Ω as a mapping of T into the metrizable space $(2^{R^2}, 2^\rho)$ is *not* measurable in the Bourbaki sense [5]. Consequently there is no sequence of measurable simple functions from T into 2^X which converges pointwise almost everywhere on T to Ω in $(2^{R^2}, 2^\rho)$. The mapping Ω is measurable as a multivalued function from T into 2^{R^2} (see § 2 for definition). Indeed, Ω as a mapping from T into the metrizable space $(2^{R^2}, 2^{\rho_\infty})$ is continuous (cf. [21]).

The question of approximating measurable mappings $\Omega : T \to 2^X$ with measurable simple functions is more readily resolved in $(2^X, 2^{\rho_\infty})$, as the next theorem illustrates.

THEOREM 7. *Let 2^X have the uniform topology determined by ρ_∞. A necessary and sufficient condition that a mapping $\Omega : T \to 2^X$ be measurable is that there exists a sequence S_n, $n \in \omega$, of measurable simple functions $S_n : T \to 2^X$, $n \in \omega$, such that $S_n(t) \to \Omega(t)$ a.e. on T as $n \to \infty$.*

Proof. See [21, Corollary 2.5].

We next state some lemmas which are useful in calculating integrals of multivalued functions. These lemmas seem to be fairly well known.

LEMMA 3. *Let T be nonatomic, let E be a measurable subset of T, let real $p_i \geqq 0$, $i = 1, 2, \cdots, n$, be given such that $\sum_{i=1}^n p_i = 1$. Then there is a measurable partition $\{T_1, \cdots, T_n\}$ of E such that $\mu(T_i) = p_i \mu(E)$, $i = 1, 2, \cdots, n$.*

If A and B are subsets of X and if α is a real number, then $A + B = \{a + b | a \in A, b \in B\}$ and $\alpha A = \{\alpha a | \alpha \in A\}$. Also co (A) denotes the convex hull of A.

LEMMA 4. *Let T be nonatomic, let $\Omega : T \to 2^X$ be a constant mapping, i.e., $\Omega(t) = F \in 2^X$ for each $t \in T$, and let E be a measurable subset of T. Then*

$$\int_E \Omega \, d\mu = \mu(E) \operatorname{co}(F).$$

Proof. The inclusion

$$\int_E \Omega \, d\mu \subset \mu(E) \operatorname{co}(F)$$

follows from the convexity theorem [6, Theorem 1, p. 203]. Suppose X has dimension r, and suppose $x \in \mu(E) \operatorname{co}(F)$. Then there exist real $p_i \geqq 0$ and $f_i \in F$, $i = 1, 2, \cdots, r + 1$, such that

$$x = \mu(E) \sum_{i=1}^{r+1} p_i f_i \quad \text{and} \quad \sum_{i=1}^{r+1} p_i = 1$$

(see [3]). By Lemma 3 there is a measurable partition $\{T_1, \cdots, T_{r+1}\}$ of E such that $\mu(T_i) = p_i \mu(E)$, $i = 1, 2, \cdots, r + 1$. We define an integrable function $f : T \to F \subset X$ by the relations $f(t) = f_i$ if $t \in T_i$, $i = 1, 2, \cdots, r + 1$. Then we have

$$\int_E f \, d\mu = \sum_{i=1}^{r+1} p_i \mu(E) f_i = x \in \int_E \Omega \, d\mu .$$

The desired equality results.

As an immediate consequence of Lemma 4 and induction we get the following statement.

LEMMA 5. *Let T be nonatomic, and let $\Omega : T \to 2^X$ be a measurable simple function $\{T_1, \cdots, T_n; F_1, \cdots, F_n\}$. Then there results*

$$\int_T \Omega \, d\mu = \sum_{i=1}^{n} \mu(T_i) \operatorname{co}(F_i).$$

There is also the following fundamental result.

LEMMA 6. *If T is nonatomic, and if $\Omega : T \to 2^X$ is integrable, then $\int_T \Omega \, d\mu$ is convex.*

This is in essence a special version of Theorem 1 in [1]. The lemma can also be proved by a slight extension of Proposition 8.10 in [16] (cf. the remarks in [19] and [20]).

The following two theorems together with Lemma 6 are basic to estimating $\rho\left(x_0, \int_T \Omega \, d\mu\right)$.

THEOREM 8. *Let A and A_n, $n \in \omega$, be nonempty, closed and convex subsets of X such that* $\lim A_n = A$ *in $(2^X, 2^{\rho_\infty})$. A sequence $x_n' \in X$, $n \in \omega$, is defined by the relation*

$$\|x_n'\| = \min \{\|x\| \mid x \in A_n\}, \qquad n \in \omega, \tag{12}$$

and we define $x' \in X$ by the relation

$$(13) \qquad \|x'\| = \min\{\|x\| \,|\, x \in A\}.$$

Then x'_n is convergent and $\lim x'_n = x'$.

Proof. That the relations (12), (13) define x'_n, $n \in \omega$, and x' uniquely follows from: *a closed convex (nonempty) set in X contains a unique element of minimal norm* (see [12]). Let $\delta > 0$ be defined by the relation $\delta = (1 + \|x'\|^2)^{-1/2}$. Then there is an $N_\delta \in \omega$ such that $n \geqq N_\delta$ implies

$$(14) \qquad J^{\infty}_{\delta/2}[A_n] \supset A \quad \text{and} \quad J^{\infty}_{\delta/2}[A] \supset A_n.$$

As a result of (14) we have:

(15) If for each $n \geqq N_\delta$ we define γ_n by the relation $\gamma_n = \rho_\infty(x', A_n)$, then corresponding to each $n \geqq N_\delta$ there exists a point $x^*_n \in A_n$ such that $\rho_\infty(x', x^*_n) = \gamma_n$.

In order to prove this, we choose a "minimizing sequence" $x_{\alpha n} \in A_n$, $\alpha \in \omega$, such that $\lim_\alpha \rho(x', x_{\alpha n}) = \gamma_n$, when $n \geqq N_\delta$. By (14) there exists $y_n \in A_n$ such that $\gamma_n \leqq \rho_\infty(x', y_n) < \delta/2$ for $n \geqq N_\delta$. Therefore we may assume that $\rho_\infty(x', x_{\alpha n}) < \delta/2$, $\alpha \in \omega$, $n \geqq N_\delta$. We let $X_\infty = X \cup \{\infty\}$ denote the one-point compactification of X (see [4]). The set $S = \{x_{\alpha n} | \alpha \in \omega, n \geqq N_\delta\}$ is bounded in norm, i.e., $\{\|x_{\alpha n}\| \,|\, \alpha \in \omega, n \geqq N_\delta\}$ is a bounded set of real numbers. For if this were not the case, there would be a sequence $x_\alpha \in S$, $\alpha \in \omega$, such that $\lim x_\alpha = \infty$ in (X_∞, ρ_∞). Consequently we would have $\lim \rho_\infty(x', x_\alpha) = \delta < \delta/2$, a contradiction. The existence of the required sequence x^*_n, $n \geqq N_\delta$, in (15) can now be evinced from the facts: the set S is bounded in norm; A_n, $n \geqq N_\delta$, is closed in X; ρ_∞ is continuous. Now since S was bounded in norm, the set $\{\|x^*_n\| \,|\, n \geqq N_\delta\}$ is also bounded. Let us suppose that $\|x^*_n\| \leqq M$, $n \geqq N_\delta$, and let us define $B > 0$ to be the number $\delta/(1 + M^2)^{1/2}$. Given $\varepsilon > 0$ there is an $N_0 \in \omega$ such that $N_0 \geqq N_\delta$ and such that $n \geqq N_0$ implies

$$(16) \qquad J^{\infty}_{B\varepsilon}[A_n] \supset A \quad \text{and} \quad J^{\infty}_{B\varepsilon}[A] \supset A_n.$$

Hence for $n \geqq N_0$ we have that

$$\rho_\infty(x', x^*_n) = \frac{\|x' - x^*_n\|}{(1 + \|x'\|^2)^{1/2}(1 + \|x^*_n\|^2)^{1/2}} < B\varepsilon,$$

from whence we obtain that $n \geqq N_0$ implies

$$\|x' - x^*_n\| < B\varepsilon(1 + \|x'\|^2)^{1/2}(1 + \|x^*_n\|^2)^{1/2} \leqq \frac{B\varepsilon}{B} = \varepsilon,$$

Consequently it is determined that if $n \geqq N_0$, then $\|x^*_n\| < \varepsilon + \|x'\|$. However, since $\|x^*_n\| \geqq \|x'_n\|$, $n \in \omega$, there results

$$(17) \qquad \|x'_n\| < \varepsilon + \|x'\|, \qquad n \geqq N_0.$$

The set $H = \{\|x'_n\| \,|\, n \geqq N_0\}$ is bounded. Thus let D be chosen such that $\|x'_n\| \leqq D$, $n \in \omega$. Define $\delta^* > 0$ to be the number $(1 + D^2)^{-1/2}$. Choose $N_{\delta^*} \geqq N_0$, $N_{\delta^*} \in \omega$, such that

$$J^{\infty}_{\delta^*/2}[A_n] \supset A \quad \text{and} \quad J^{\infty}_{\delta^*/2}[A] \supset A_n, \tag{18}$$

whenever $n \geqq N_{\delta^*}$. From (18) we deduce the following result:

(19) If for each $n \geqq N_{\delta^*}$ we define γ'_n by the relation $\gamma'_n = \rho_\infty(x'_n, A)$, then corresponding to each $n \geqq N_{\delta^*}$ there exists a point $x''_n \in A$ such that $\rho_\infty(x'_n, x''_n) = \gamma'_n$.

The proof of (19) is similar to the proof of (15). For each $n \geqq N_{\delta^*}$ we pick "minimizing sequences" $y_{\alpha n} \in A$, $\alpha \in \omega$, such that $\lim_\alpha \rho_\infty(x'_n, y_{\alpha n}) = \gamma'_n$. Because of (18) we can assume that $\rho_\infty(x'_n, y_{\alpha n}) < \delta^*/2$, $\alpha \in \omega$, $n \geqq N_{\delta^*}$. The set $S' = \{y_{\alpha n} | \alpha \in \omega, n \geqq N_{\delta^*}\}$ is bounded in norm. For if this did not obtain, then a sequence $y_\nu \in S'$, $\nu \in \omega$, would exist such that $\lim y_\nu = \infty$ in (X_∞, ρ_∞). In this event there would result $\lim_\nu \rho_\infty(x'_n, y_\nu) = (1 + \|x'_n\|^2)^{-1/2} \leqq \delta^*/2$, and since $\|x'_n\| \leqq D$, $n \in \omega$, we would have $\delta^* \leqq \delta^*/2$, which is a contradiction. Thus S' is bounded in norm, A is closed, and ρ_∞ is continuous. The existence of the required $x''_n \in A$, $n \geqq N_{\delta^*}$, in (19) is now clear. Moreover, the set $\{\|x''_n\| \,|\, n \geqq N_{\delta^*}\}$ is evidently bounded. Select C such that $\|x''_n\| \leqq C$, whenever $n \geqq N_{\delta^*}$, and define $b > 0$ by the relation $b = \delta^*/(1 + C^2)^{1/2}$. Then there exists $N_1 \in \omega$ such that for $N_1 \geqq N_{\delta^*} \geqq N_0 \geqq N_\delta$,

$$J^{\infty}_{\varepsilon b}[A_n] \supset A \quad \text{and} \quad J^{\infty}_{\varepsilon b}[A] \supset A_n \tag{20}$$

whenever $n \geqq N_1$. From (19) we must have

$$\rho_\infty(x'_n, x''_n) = \frac{\|x'_n - x''_n\|}{(1 + \|x'_n\|^2)^{1/2}(1 + \|x''_n\|^2)^{1/2}} < \varepsilon b$$

whenever $n \geqq N_1$, and consequently

$$\|x'_n - x''_n\| < \varepsilon b(1 + \|x'_n\|^2)^{1/2}(1 + \|x''_n\|^2)^{1/2} \leqq \frac{\varepsilon b}{b} = \varepsilon,$$

whenever $n \geqq N_1$. Hence we obtain the estimate

$$\|x''_n\| < \varepsilon + \|x'_n\| \qquad n \geqq N_1. \tag{21}$$

Since, however, $x''_n \in A$, it follows that $\|x'\| \leqq \|x''_n\|$, $n \geqq N_1$, and this together with (21) yields

$$\|x'\| < \varepsilon + \|x'_n\| \qquad \text{whenever} \quad n \geqq N_1. \tag{22}$$

Now both (17) and (22) are true for $n \geqq N_1$, and therefore

$$\lim \|x'_n\| = \|x'\|. \tag{23}$$

It is also clear that

$$\lim \|x'_n - x''_n\| = 0. \tag{24}$$

Thus we also have that

$$\lim_{n,m} \left\| \frac{x'_n + x'_m}{2} - \frac{x''_n + x''_m}{2} \right\| = 0. \tag{24'}$$

Denoting $(x''_n + x''_m)/2$ by x''_{nm}, we observe that $x''_{nm} \in A$, because A is convex and $x''_n, x''_m \in A$.

Case 1. $\|x'\| = 0$. This is the trivial case. The conclusion of our theorem follows immediately from (23).

Case 2. $\|x'\| > 0$. Given $\varepsilon > 0$ and small enough so that $2\|x'\| > \varepsilon/(2\|x'\| + 1)$, there exists $n_0 \in \omega$ (by (24')) such that

$$(25) \qquad \|x'_m + x'_n\| + \eta > 2\|x''_{nm}\| \qquad \text{whenever} \quad n, m \geqq n_0,$$

with η defined to be $\varepsilon/(2\|x'\| + 1)$. However, x''_{nm} belongs to A and (13) reveals that $\|x''_{nm}\| \geqq \|x'\|$, and this together with (25) results in the inequality

$$(26) \qquad \|x'_m + x'_n\| > 2\|x'\| - \eta > 0, \qquad n, m \geqq n_0.$$

The norm $\|\cdot\|$ satisfies the parallelogram law, and thus from (26) we deduce that, for n, $m \geqq n_0$,

$$\begin{aligned}
\|x'_m - x'_n\|^2 &= 2\|x'_m\|^2 + 2\|x'_n\|^2 - \|x_m + x_n\|^2 \\
&\leqq 2\|x'_m\|^2 + 2\|x'_n\|^2 - (2\|x'\| - \eta)^2 \\
(27) \qquad &= 2[\|x'_m\|^2 - \|x'\|^2] + 2[\|x'_n\|^2 - \|x'\|^2] \\
&\quad + \frac{2\|x'\|\varepsilon}{2\|x'\| + 1} - \frac{\varepsilon^2}{(2\|x'\| + 1)^2} \\
&< 2[\|x'_m\|^2 - \|x'\|^2] + 2[\|x'_n\|^2 - \|x'\|^2] + \varepsilon.
\end{aligned}$$

From (27) we infer that x'_n, $n \in \omega$, is a Cauchy sequence and therefore converges to some $x \in X$. Since the x''_n are in A, (24) implies that x is a limit point of A. But A is closed, so x belongs to A. Relation (23) now yields

$$\lim \|x'_n\| = \|x\| = \|x'\|.$$

But x' satisfying (13) is unique, so $x = x'$, and $\lim x'_n = x'$.

Remark 5. If in Theorem 8, the A_n converge to A in $(2^X, 2^\rho)$, then, of course, the theorem still applies. However, in this case there is a much simpler proof. The difficulties we encounter in proving Theorem 8 are partially explained by Remark 1 and by the following difficulty. The metric ρ enjoys the useful property that if C is a nonempty convex set contained in X, then $J_\varepsilon[C]$ is convex for each $\varepsilon > 0$. This does not in general obtain for the metric ρ_∞. Indeed, if $C = \{(0, y) \in R^2 | y \in R\}$, then $J^\infty_\varepsilon[C]$ is not convex for any $\varepsilon > 0$.

The metric ρ_∞ is not invariant under translations, but nonetheless, the following theorem is true.

THEOREM 9. *If A and A_n, $n \in \omega$, are nonempty closed subsets of X such that* $\lim A_n = A$ *in* $(2^X, 2^{\rho_\infty})$, *and if* $\{x_n\}$, $n \in \omega$, *is a sequence in X such that $x_n \to x_0$ as* $n \to \infty$, *then* $\lim (x_n + A_n) = x_0 + A$ *in* $(2^X, 2^{\rho_\infty})$.

Proof. Given $m \in \omega$ there exists $N_m \in \omega$ such that

$$(28) \qquad J^\infty_{1/m}[A_n] \supset A \qquad \text{whenever} \quad n \geqq N_m.$$

Also $x_0 + A$ is precompact in (X, ρ_∞), and consequently given $\varepsilon > 0$ there corresponds $a_1, \cdots, a_k \in A$ such that

$$\bigcup_{i=1}^{k} J^{\infty}_{\varepsilon/2}[x_0 + a_i] \supset x_0 + A. \tag{29}$$

By (28) we have that given $i = 1, 2, \cdots, k$ there exists $a_{imn} \in A_n$, $n \geqq N_m$, such that $\rho_\infty(a_i, a_{imn}) < 1/m$. Hence $\lim_{m,n} a_{imn} = a_i$, and consequently $\lim_{m,n}(x_n + a_{imn}) = x_0 + a_i$, $i = 1, 2, \cdots, k$. Therefore we conclude that there exists $N_\varepsilon \in \omega$ such that for $n, m \geqq N_\varepsilon$ we have that

$$\rho_\infty(x_0 + a_i, x_n + a_{imn}) < \varepsilon/2. \tag{30}$$

From (29) and (30) we obtain that

$$J^{\infty}_{\varepsilon}[x_n + A_n] \supset x_0 + A \quad \text{whenever} \quad n \geqq N_\varepsilon. \tag{31}$$

To complete the proof we shall show that

$$J^{\infty}_{\varepsilon}[x_0 + A] \supset x_n + A_n \tag{32}$$

for all sufficiently large n. In order to establish (32) observe that for any subsequence A_{n_k} of A_n we have:

$$\text{If } a_{n_k}, k \in \omega, \text{ and if } x_{n_k} + a_{n_k} \to P \text{ as } k \to \infty, \text{ then } P \in x_0 + A. \tag{33}$$

Suppose (32) is false. Then there is an $\varepsilon > 0$, and there is a strictly increasing sequence $\lambda(n)$ in ω, such that

$$J^{\infty}_{\varepsilon}[x_0 + A] \not\supset x_{\lambda(n)} + A_{\lambda(n)}, \qquad n \in \omega. \tag{34}$$

The result in (34) implies there is a $P_n \in (x_{\lambda(n)} + A_{\lambda(n)}) \backslash J^{\infty}_{\varepsilon}[x_0 + A]$ for $n \in \omega$. Now P_n has the form $x_{\lambda(n)} + a_{\lambda(n)}$, $a_{\lambda(n)} \in A_{\lambda(n)}$, and this implies that P_n is norm-bounded if and only if $a_{\lambda(n)}$ is norm-bounded. Thus (33) reveals that P_n is not norm-bounded, and there is a subsequence of P_n (still denoted by P_n) such that P_n converges to ∞ in $X_\infty = X \cup \{\infty\}$, the one-point compactification of X. Since $A_n \to A$ in $(2^X, 2^{\rho_\infty})$, and since $a_{\lambda(n)}$ also converges to ∞, there is a sequence $a^*_{\lambda(n)} \in A$ such that $x_0 + a^*_{\lambda(n)}$ converges to ∞. Consequently there is a $K_0 \in \omega$ such that

$$\rho_\infty(x_{\lambda(n)} + a_{\lambda(n)}, x_0 + a^*_{\lambda(n)}) < \varepsilon, \qquad k \geqq K_0,$$

which contradicts the fact that $P_n \notin J^{\infty}_{\varepsilon}[x_0 + A]$ for every $n \in \omega$. This completes the proof.

Remark 6. Let μ denote Lebesgue measure on the compact interval $T = [t_0, t_1]$. Consider the standard linear control system

$$\dot{x} = A(t)x + \phi(u, t), \qquad t \in T, \tag{35}$$

with $x \in R^p$, $u \in R^q$, and A a continuous $p \times p$ matrix-valued function defined on T. The mapping $\phi: R^q \times T \to R^p$ is such that $\phi(\cdot, t)$ is continuous for each $t \in T$, and such that $\phi(u, \cdot)$ is integrable for each $u \in R^q$. Let U denote a nonempty closed subset of R^q such that $\phi(U, t)$ is closed for each $t \in T$. Let X denote the fundamental matrix satisfying the matrix differential equation $\dot{X} = AX$ with the initial condition $X(0) = I$, where I is the $p \times p$ identity matrix. Let $\mathscr{T}$ denote the set of all measurable mappings $u : T \rightarrow U$ such that the mapping $t \rightarrow \phi(u(t), t)$, $t \in T$, is integrable. Then corresponding to any $u \in \mathscr{T}$ there exists a unique absolutely continuous function $x(\cdot, u) : T \rightarrow R^p$ satisfying (35) a.e. on T and the initial condition

$$x(t_0, u) = x_0 \in R^p, \tag{36}$$

where $x_0 \in R^p$ is fixed. By variation of parameters this function is given by

$$x(t, u) = X(t)\left[x_0 + \int_{t_0}^{t} X^{-1}(\xi)\phi(u(\xi), \xi)\, d\xi\right] \tag{37}$$

for $t \in T$. A point $x \in R^p$ is *attainable* if there exists $u \in \mathscr{T}$ such that $x(t_1, u) = x$. The attainable set is the set

$$\mathscr{R} = \{x \in R^p | x \text{ is attainable}\}.$$

Now we observe that the mapping $\Omega : t \rightarrow X(t_1)X^{-1}(t)\phi(U, t) \in 2^{R^p}$, $t \in T$, is measurable. In order to prove this, consider the mapping $\Phi : t \rightarrow \phi(U, t) \in 2^{R^p}$, $t \in T$, and observe that if F is a closed set in R^p, then $\Omega^{-}F = \{t \in T | \Omega(t) \cap F \neq \varnothing\}$ $= \{t \in T | \Phi(t) \cap X(t)X^{-1}(t_1)F \neq \varnothing\}$. Consequently it suffices to verify that Φ is measurable. Let $\varepsilon > 0$ be given. Then by [18, Corollary 2.1 or Corollary 2.3] there is a compact $T_\varepsilon \subset T$ such that $\mu(T \backslash T_\varepsilon) < \varepsilon$ and such that $\phi | R^p \times T_\varepsilon$ is continuous. Let K_n, $n \in \omega$, denote a sequence of compact sets in U such that $\bigcup_{n \in \omega} K_n = U$. Let Φ_ε denote $\Phi | T_\varepsilon$, and let F be a closed set in R^p. Then we have that

$$\Phi_\varepsilon^{-} F = \{t \in T_\varepsilon | \Phi(t) \cap F \neq \varnothing\} = \bigcup_{n \in \omega} \{t \in T_\varepsilon | \phi(K_n, t) \cap F \neq \varnothing\}.$$

The sets $\{t \in T_\varepsilon | \phi(K_n, t) \cap F \neq \varnothing\}$, $n \in \omega$, are each closed. Therefore $\Phi_\varepsilon^{-} F$ is measurable. Since $\varepsilon > 0$ was arbitrary, Φ is measurable. Hence Ω is also measurable. By our assumptions we have that if u_0 is a fixed point in U, then $t \rightarrow \phi(u_0, t)$ is integrable. Consequently Ω is integrable. By implicit function theorems given in [7], [22] or [21] we see that

$$\mathscr{R} = X(t_1)x_0 + \int_T \Omega\, d\mu.$$

This makes clear the relationship between the attainable set $\mathscr{R}$ and integrals of the type we have studied.

Example 1. When $A(t)$ is skew for each $t \in T$, then $X(t)$ is a rotation [15, p. 238], and the problem of estimating $\mathscr{R}$ lends itself to geometric interpretations through the results we have obtained. We illustrate with a very simple example. The control system is given by

$$\begin{bmatrix} \dot{x}^1 \\ \dot{x}^2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} x^1 \\ x^2 \end{bmatrix} + \begin{bmatrix} 0 \\ u \end{bmatrix}, \qquad t \in T = \left[0, \frac{\pi}{2}\right],$$

$$x^1(0, u) = 0 = x^2(0, u).$$

This is the first order system corresponding to $\ddot{y} + y = u$. The set U is the set of nonnegative integers $\{0, 1, 2, \cdots\}$. The fundamental matrix $X(t)$ is determined to be

$$X(t) = \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix},$$

whereas

$$X^{-1}(t) = \begin{bmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{bmatrix}.$$

Let $\Omega(t) = \{(x^1, x^2)|x^1 = u \cos t,\ x^2 = u \sin t,\ u \in U\}$ for $t \in T = [0, \pi/2]$. The mapping $t \rightarrow \Omega(t) \in 2^{R^2}, t \in T$, is continuous in $(2^{R^2}, 2^{\rho_\infty})$. However, Ω is not continuous in $(2^{R^2}, 2^{\rho})$; the difficulties are essentially those mentioned in Cesari's example cited earlier in Remark 4. Consider the sequence of intervals:

$$T_{nk} = \left[\frac{(k-1)\pi}{2n}, \frac{k\pi}{2n}\right], \qquad n \in \omega, \quad k = 1, 2, 3, \cdots, n-1,$$

and

$$T_{nn} = \left[\frac{(n-1)\pi}{2n}, \frac{\pi}{2}\right].$$

For brevity, denote the left and right endpoints of T_{nk} by a_{nk} and b_{nk} respectively, $n \in \omega, k = 1, 2, \cdots, n$. Denote by F_{nk} the element in 2^{R^2} defined by

$$F_{nk} = \{(u \cos t, u \sin t)\,|u \in U, a_{nk} \leqq t \leqq b_{nk}\},$$

$n \in \omega$, $k = 1, 2, \cdots, n$. Denote by S_n, $n \in \omega$, the sequence of simple functions $\{T_{n1}, T_{n2}, \cdots, T_{nn}; F_{n1}, F_{n2}, \cdots, F_{nn}\}$, $n \in \omega$. Then by Lemma 5 we have that for each $n \in \omega$, $\int_T S_n \, d\mu = \{(x^1, x^2)|x^1 \geqq 0, x^2 \geqq 0\}$. If we let $K_n = \{x = (x^1, x^2) \in R^2|\ \|x\| \leqq n\}$, then the mappings $S_n, n \in \omega, \Omega$, and the sets K_n, $n \in \omega$, satisfy the hypotheses of Corollary 3, and we have that

$$\lim \int_T S_n \, d\mu = \int_T \Omega \, d\mu,$$

and therefore $\mathrm{cl}\left(\int_T \Omega \, d\mu\right) = \{(x^1, x^2)|x^1 \geqq 0, x^2 \geqq 0\}$. Clearly $(0, x^2)$ and $(x^1, 0)$ do not belong to $\int_T \Omega \, d\mu$ for $x^2 \neq 0, x^1 \neq 0$, and hence the attainable set $\mathscr{R} = \int_T \Omega \, d\mu$ is not closed. With the aid of Lemma 6 it is actually determined that $\int_T \Omega \, d\mu = \{(x^1, x^2)|x^1 > 0, x^2 > 0\} \cup \{(0, 0)\}$. Thus the conjecture in [19, p. 47] is false. Note that the outcome of this example would be the same if $\Omega(t)$ were replaced by $\mathrm{co}\,(\Omega(t))$.

Example 2. Let

$$F_1 = \{(x, y) \in R^2|y \geqq 1/x, x > 0\},$$

and let

$$F_2 = \{(x, y) \in R^2|y \geqq -1/x, x < 0\}.$$

Let μ be Lebesgue measure on $[0,1] = T$. Let $T_1 = [0,\frac{1}{2}]$, and let $T_2 = (\frac{1}{2},1]$. If Ω is the simple function $\{T_1, T_2; F_1, F_2\}$, then $\int_T \Omega\, d\mu = \frac{1}{2}F_1 + \frac{1}{2}F_2$, which is not closed.

Another interesting example which Professor Halkin showed to the author is the following.

Example 3. μ is Lebesgue measure on $[0,1] = T$. The mapping Ω is given by $t \rightarrow \Omega(t) = \{(x,y) \in R^2 | x = u, y = -u^2t^2, -\infty < u < \infty\}$. The set $\int_T \Omega\, d\mu$ is not closed, since $(1,0) \notin \int_T \Omega\, d\mu$, but for each $n \in \omega$, $(1, -1/(3n)) = \int_T f_n\, d\mu \in \int_T \Omega\, d\mu$, where $f_n(t) = (n, -n^2t^2)$ on $[0, 1/n]$ and $f_n(t) = (0,0)$ on $(1/n, 1]$.

REFERENCES

[1] R. J. AUMANN, *Integrals of set-valued functions*, J. Math. Anal. Appl., 12 (1965), pp. 1–12.

[2] A. V. BALAKRISHNAN, *An operator theoretic formulation of a class of control problems and a steepest descent method of solution*, this Journal, 1 (1963), pp. 109–127.

[3] C. BERGE, *Topological Spaces*, Macmillan, New York, 1963.

[4] N. BOURBAKI, *General Topology, Parts 1 and 2*, Addison-Wesley, Reading, Massachusetts, 1966.

[5] ———, *Intégration*, Hermann, Paris, 1965, Chaps. 1–4.

[6] T. F. BRIDGLAND, JR., *On the problem of approximate synthesis of optimal controls*, this Journal, 5 (1967), pp. 326–344.

[7] C. CASTAING, *Sur les multi-applications mesurables*, Revue Française d'Informatique et de Recherche Operationnelle, 1 (1967), pp. 91–126.

[8] 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, 413–429.

[9] J. CULLUM, *Perturbations of optimal control problems*, this Journal, 4 (1966), pp. 473–487.

[10] ———, *Perturbations and approximations of continuous optimal control problems*, Mathematical Theory of Control, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, New York, 1967, pp. 156–169.

[11] G. DEBREU, *Integration of correspondences*, Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, 1966, pp. 351–372.

[12] N. DINCULEANU, *Vector Measures*, Pergamon, New York, 1967.

[13] N. DUNFORD AND J. SCHWARTZ, *Linear Operators*, Interscience, New York, 1958.

[14] E. G. GILBERT, *An iterative procedure for computing the minimum of quadratic form on a convex set*, this Journal, 4 (1966), pp. 61–80.

[15] W. GREUB, *Linear Algebra*, 3rd ed., Springer-Verlag, New York, 1967.

[16] H. HALKIN, *On a necessary condition for optimal control of nonlinear systems*, J. Analyse Math., 12 (1964), pp. 1–82.

[17] H. HERMES, *The equivalence and approximation of optimal control problems*, J. Differential Equations, 1 (1965), pp. 409–426.

[18] M. Q. JACOBS, *Remarks on some recent extensions of Filippov's implicit functions lemma*, this Journal, 5 (1967), pp. 622–627.

[19] ———, *Attainable sets for linear systems with unbounded controls*, Mathematical Theory of Control, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, New York, 1967, pp. 46–53.

[20] ———, *Attainable sets in systems with unbounded controls*, J. Differential Equations, 4 (1968), pp. 444–459.

[21] ———, *Measurable multivalued mappings and Lusin's theorem*, Trans. Amer. Math. Soc., 134 (1968), pp. 471–481.

[22] K. KURATOWSKI AND C. RYLL-NARDZEWSKI, *A general theorem on selectors*, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys., 13 (1965), pp. 397–403.

[23] E. MICHAEL, *Topologies on spaces of subsets*, Trans. Amer. Math. Soc., 71 (1951), pp. 152–182.

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

[25] C. OLECH, *A note concerning set-valued measurable functions*, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys., 13 (1965), pp. 317–321.

[26] ———, *Extremal solutions of a control system*, J. Differential Equations, 2 (1966), pp. 74–101.

[27] ———, *Lexicographical order, range of integrals and 'bang-bang' principle*, Mathematical Theory of Control, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, New York, 1967, pp. 35–45.

[28] ———, *On the range of unbounded vector valued measure*, Math. Systems Theory, 2 (1968), pp. 251–256.

[29] ———, *Existence theorems for optimal problems with vector valued cost function*, Trans. Amer. Math. Soc., 136 (1969), pp. 1–22.

[30] H. RÅDSTROM, *An embedding theorem for spaces of convex sets*, Proc. Amer. Math. Soc., 3 (1952), pp. 165–169.

SIAM J. Control
Vol. 7, No. 2, May 1969

TOWARD A THEORY OF MANY PLAYER DIFFERENTIAL GAMES*

JAMES H. CASE†

Abstract. We consider a differential game $\mathscr{G}$ between the players $1, 2, \cdots, N$ whose state is governed by the equation $\dot{x} = f(t, x, u_1, \cdots, u_N)$, where u_i is a control vector belonging to player i, and suppose that each player i wishes to manipulate his u_i in such a way as to maximize the functional

$$J_i = K_i(t_f, x(t_f)) + \int_{t_0}^{t_f} L_i(t, x(t), u_i(t), \cdots, u_N(t))\, dt.$$

Here t_0 and t_f are, respectively, the times at which the game begins and ends. A strategy N-tuple $u_1 = \sigma_1^*(t, x), \cdots, u_N = \sigma_N^*(t, x)$ is called an equilibrium point for $\mathscr{G}$ if the inequalities

$$J_i(\sigma_1^*, \cdots, \sigma_N^*) \geqq J_i(\cdots, \sigma_{i-1}^*, \sigma_i, \sigma_{i+1}^*, \cdots)$$

hold for each $i = 1, \cdots, N$ and for each admissible strategy $u_i = \sigma_i(t, x)$ for player i. We seek methods of finding equilibrium points for the game $\mathscr{G}$.

The principal results of the paper are three: a theorem to the effect that the "value functions" are in a certain sense everywhere differentiable, a system of partial differential equations which they must satisfy and a sufficiency result which guarantees that, under certain circumstances, the "Dynamic Programming" method suggested by the preceding differential equations does indeed yield the equilibrium points of the game. In the final two sections, a significant class of games is solved explicitly, and a method of characteristics is given for the solution of the "Hamilton–Jacobi" partial differential equations.

1. Introduction. A considerable body of literature has grown up in recent years on the theory of zero-sum two player differential games, and significant applications of this theory have been made to problems of a military nature. However, for applications to other fields, such as biology or economics, it is necessary to study games which are not zero-sum and which may involve more than two players. In this paper we shall develop a technique for solving such games. Our method will reduce, in the case of zero-sum two player games, to that of Berkovitz [3], [4], [5] and Isaacs [8] and to dynamic programming if the games involve only one player. Our approach will be essentially that of Berkovitz [5] except that we shall allow a slightly more inclusive class of strategies. The rest of this introductory section will be devoted to the explanation of certain terms and notations to be used throughout the paper.

Vector matrix notation will generally be used. Vectors and matrices will be denoted by single letters. Supercripts will be used to denote the components of a vector; a subscript on a vector will indicate the player with which that vector is to be associated. Thus, for example, u_i^j is the jth component of the control vector u_i belonging to player i. Vectors will be written as matrices consisting of either one

* Received by the editors May 21, 1968, and in revised form February 12, 1969.

† Mathematics Research Center, University of Wisconsin, Madison, Wisconsin 53706. This research was supported by the Mathematics Research Center, United States Army, Madison, Wisconsin, under Contract DA-31-124-ARO-D-462.

row or one column. We shall not use a transpose symbol to distinguish between the two usages, as it will be clear from the context which is meant. The transpose of a matrix, however, will be indicated by a prime ($'$). The scalar product of two n-dimensional vectors p and x will be written $\langle p, x\rangle$, and all scalars that occur will be real. All vectors will have real components.

The functions that occur will depend on one or more of the variables $t, x, u_1, \cdots, u_N, p_1, \cdots, p_N$. If $\Phi = \Phi(t, x, u_1, \cdots, u_N, p_1, \cdots, p_N)$ is differentiable in some region R, we shall denote by $\nabla_{p_i}\Phi$ the gradient of Φ with respect to the variable p_i. If Φ takes values in E_m, and p_i is an element of E_n, $\nabla_{p_i}\Phi$ is an $m \times n$ matrix. The element in the jth row and the kth column is the real-valued function $\partial\Phi^j/\partial p_i^k$ which is defined on R.

We consider a game between the players $1, \cdots, N$ and take the state of the game to be represented by an element x of E_n. Ordinarily, x will be constrained to lie in a closed subset $\mathscr{E}$ of E_n, and the motion of x will be governed by the "kinematic equation"

$$\dot{x} = f(t, x, u_1, \cdots, u_N), \tag{1}$$

where each u_i is a control variable governed by player i. We assume that each u_i is constrained to lie in a closed subset $U_i = U_i(t, x)$ of E_{m_i} and that f is defined on an open subset A of E_M (here $M = n + m_1 + \cdots + m_N + 1$) containing $E_1 \times \mathscr{E} \times U_1 \times \cdots \times U_N$. The function f must be everywhere at least once continuously differentiable in each of its M (scalar) arguments. Next let $\mathscr{D}$ be an open subset of E_n containing $\mathscr{E}$. We suppose that for each player i there is a class Σ_i of "strategies" σ_i (which the reader should think of as piecewise smooth functions of t and x, at least until the precise definition is given in the next section), defined on $E_1 \times \mathscr{D}$, and taking values in U_i, such that for every element $\sigma = (\sigma_1, \cdots, \sigma_N)$ of $\Sigma = \Sigma_1 \times \cdots \times \Sigma_N$, the differential equation

$$\dot{x} = f(t, x, \sigma_1(t, x), \cdots, \sigma_N(t, x)) \tag{2}$$

has a unique solution $x_\sigma(t; \tau, \eta)$ thru every initial point (τ, η) in $E_1 \times \mathscr{E}$. We shall say that the game "terminates" if the curve $x_\sigma(t; \tau, \eta)$ ever strikes a certain smooth manifold $\mathscr{C}$ in $E_1 \times \mathscr{E}$, and we shall denote by $t_f = t_f(\sigma; \tau, \eta)$ the instant at which this occurs.[1] The manifold $\mathscr{C}$ is called the "terminal surface" of the game.

Next we assume that $L_1, \cdots, L_N$ are real-valued continuously differentiable functions defined on A and that $K_1, \cdots, K_N$ are real-valued and twice continuously differentiable functions defined on $E_1 \times \mathscr{D}$. Then if $\sigma \in \Sigma$ is such that the solution $x(t) = x_\sigma(t; \tau, \eta)$ of (2) strikes $\mathscr{C}$, we say that σ is "playable" and define the payoff functionals associated with the "play" $x(t) = x_\sigma(t; \tau, \eta)$ by the relations

$$\begin{aligned} J_i = J_i(\tau, \eta; \sigma_1, \cdots, \sigma_N) &= K_i(t_f, x(t_f)) \\ &\quad + \int_\tau^{t_f} L_i(t, x(t), \sigma_1(t, x(t)), \cdots, \sigma_N(t, x(t)))\, dt. \end{aligned} \tag{3}$$

[1] Ordinarily, we shall assume that $\mathscr{C}$ is of dimension n.

The object of the game is for each player i to choose his strategy $\sigma_i \in \Sigma_i$ in such a way as to maximize his own payoff J_i.

There are many definitions of "solution" for games involving more than two players, but in what follows, we shall consider only equilibrium points. A playable strategy N-tuple $\sigma^* = (\sigma_1^*, \cdots, \sigma_N^*)$ is called an equilibrium point for the game previously described, if the inequality

$$(4) \qquad J_i(\tau, \eta; \cdots, \sigma_{i-1}^*, \sigma_i, \sigma_{i+1}^*, \cdots) \leqq J_i(\tau, \eta; \cdots, \sigma_{i-1}^*, \sigma_i^*, \sigma_{i+1}^*, \cdots)$$

holds for every $\sigma_i \in \Sigma_i$ such that the left side of (4) is defined for each $i = 1, \cdots, N$. We choose this solution concept not because there is any widespread agreement as to its virtues, but merely for its apparent tractability. In the next section we address ourselves to the problem of selecting an appropriate class of strategies, which is an important one for differential games. In particular, we shall concern ourselves only with pure strategies in what follows and with games which may be expected to have pure strategy solutions. This will dictate, in particular, that all players have "perfect information" throughout the course of play.

2. The strategies. We shall not pause either to motivate or to justify our choice of the strategy sets $\Sigma_1, \cdots, \Sigma_N$. Suffice it to say that most of the games in Isaacs' book [8] do have solutions in the classes which we shall consider, and that our choice will permit a rigorous development of the theory. A relatively extensive discussion of the various properties which a class of strategies should possess is available in [6, Chap. 1], and, in particular, it is pointed out there that in most situations of interest, it is altogether nonsensical to consider strategies which depend upon t alone. A more readily available discussion of these matters will be found in [8, pp. 36–40].

We shall begin by defining a "tactic" for player i to be a continuously differentiable function $\sigma_i(t, x)$ defined on $E_1 \times \mathscr{D}$ and taking values in U_i. A "switching strategy" will then be, in intuitive terms, a finite collection $\sigma_i^1(t, x), \cdots, \sigma_i^{k_i}(t, x)$ of tactics together with a set of instructions for switching from one to another.[2] Let $M_i^1, \cdots, M_i^{l_i}$ be a collection of manifolds of dimension less than $n + 1$ in $E_1 \times \mathscr{E}$, and let $I_i(t, x)$ be a function defined on $E_1 \times \mathscr{E}$ and taking values among the numbers $1, 2, \cdots, k_i$. The manifolds $M_i^1, \cdots, M_i^{l_i}$ will be called the "switching manifolds" belonging to the strategy σ_i, and I_i will be called its "indicator function." We require that I_i be constant along any curve Γ which never leaves or enters any of the manifolds M_i^j. That is, if P and Q are two points of $E_1 \times \mathscr{E}$ which lie on each of the manifolds $M_i^{j_1}, \cdots, M_i^{j_\alpha}$ (and on no others), and if they can be joined by a curve Γ which lies on just those manifolds, then we require that $I_i(P) = I_i(Q)$. We now define the "switching strategy" σ_i to be the set $\{\sigma_i^1(t, x), \cdots, \sigma_i^{k_i}(t, x), M_i^1, \cdots, M_i^{l_i}, I_i(t, x)\}$ consisting of the k_i tactics $\sigma_i^j(t, x)$, the l_i switching manifolds M_i^j, and the indicator function $I_i(t, x)$. A player i "plays" such a strategy σ_i simply by setting his control u_i equal to $\sigma_i^{I_i(t,x)}(t, x)$. It is evident from the requirements on

[2] Henceforth, when we speak of a strategy, we shall always mean a switching strategy. No other class of strategies will be either defined or discussed.

the functions I_i, that the right-hand side of the differential equation

$$\dot{x} = f(t, x, \sigma_1^{I_1(t,x)}(t, x), \cdots, \sigma_N^{I_N(t,x)}(t, x)) \tag{5}$$

is continuously differentiable along any curve Γ which neither leaves nor enters any of the switching manifolds M_i^j belonging to the strategies σ_i of the players $i = 1, \cdots, N$. For convenience, we shall often write (5) in the abbreviated form

$$\dot{x} = f(t, x, \sigma_1, \cdots, \sigma_N). \tag{6}$$

By the strategy set Σ_i we shall mean the set of all the switching strategies σ_i previously defined.

In the next section, it will be convenient to speak not of individual strategies, but of strategy N-tuples $\sigma = (\sigma_1, \cdots, \sigma_N)$ lying in the Cartesian product $\Sigma = \Sigma_1 \times \cdots \times \Sigma_N$ of the individual strategy sets. It is clear that we may represent such a σ in the form $\sigma = \{\sigma^1, \cdots, \sigma^k, M^1, \cdots, M^l, I\}$, where each σ^j is a continuously differentiable function from $E_1 \times \mathscr{E}$ to $U_1 \times \cdots \times U_N$, each M^j is a switching manifold belonging to one of the strategies σ_i and I is a function on $E_1 \times \mathscr{E}$ taking values in the set $\{1, \cdots, k_1\} \times \cdots \times \{1, \cdots, k_N\}$, which is constant along any curve Γ which neither enters nor leaves a switching manifold M^j. In the next section, we shall assume that an N-tuple σ has been chosen from Σ in such a way that the solutions of (5) do actually reach the terminal surface $\mathscr{C}$, at least for those initial points (τ, η) which lie in a certain subset R of $E_1 \times \mathscr{E}$, and show that the resulting payoffs $J_i(\sigma, \tau, \eta)$ to the various players depend in a highly regular fashion upon the quantities τ and η.

3. The value functions. Let us now suppose that each player i has chosen a strategy σ_i from his strategy set $\sum_i$, and let us denote by σ the resulting strategy N-tuple. Write $\sigma = \{\sigma^1, \cdots, \sigma^k, M^1, \cdots, M^l, I\}$. Let $P = P(\tau, \eta)$ be the point in $E_1 \times \mathscr{E}$ with coordinates (τ, η), and let $x_\sigma(t; P) = x_\sigma(t; \tau, \eta)$ be a solution of (5) satisfying the initial condition $x_\sigma(\tau; P) = \eta$, which is defined in a time interval $\tau \leqq t \leqq \tau'$. Then the function $I(t, x_\sigma(t; P))$ will be constant in every subinterval during which none of the players $i = 1, \cdots, N$ change tactics. A point of discontinuity for $I(t, x_\sigma(t; P))$ will be called a "switching time for $x_\sigma(t; P)$," or more simply, a "switching." We denote by R the set of all initial points P in $E_1 \times \mathscr{E}$ such that (5) has a solution $x_\sigma(t; P)$ starting at $P(\tau, \eta)$ and terminating on $\mathscr{C}$ after at most a finite number of switchings.

For the construction which is to follow, we shall need certain assumptions regarding the manner in which the field of trajectories $x_\sigma(t)$ generated by σ (that is, the field of solutions of (5)) covers R. Namely, we shall assume that

(i) if $x_\sigma(t)$ and $x_\sigma^*(t)$ are two solutions of (5) such that $x_\sigma(\tau) = \eta = x_\sigma^*(t)$ at some point (τ, η) in R, then $x_\sigma(t) = x_\sigma^*(t)$ for all $t > \tau$, and

(ii) if $\tau_0 > \tau$ is the first instant at which $x_\sigma(t; P)$ lies on some M^α, then the curve segment $x_\sigma(t; P)$, $\tau \leqq t \leqq \tau_0$, is not tangent to that M^α. In particular, $x_\sigma(t; P)$, $\tau \leqq t \leqq t_f$, is not tangent to $\mathscr{C}$.

The requirement (i) of uniqueness is somewhat restrictive, as bifurcating optimal trajectories do occur in certain known examples. However, we have chosen not to treat such phenomena here on the grounds that pursuers and evaders do not bifurcate. It is to be emphasized that the requirements (i) and (ii) each are imposed in one time direction only. For we wish, in particular, to treat manifolds M where trajectories may meet and then continue on together along M. Isaacs [8, p. 156] has discovered several examples of such manifolds.

Since each point $P(\tau, \eta)$ in R is joined to $\mathscr{C}$ by a unique solution $x_\sigma(t; P)$ of (5), the payoffs $J_i(\sigma; \tau, \eta) = J_i(\sigma; P)$ are well defined by (3). We shall show that the functions $J_i(\sigma; P)$ so defined are (for fixed σ) continuous in τ along the curves $x_\sigma(t; P)$ and continuously differentiable in a certain sense which we shall make precise later.

The proof will be by induction on the number of switchings. Let us partition R into the disjoint subsets $R_0 \cup R_1 \cup R_2 \cup \cdots$, where P is in R_α if and only if the unique trajectory $x_\sigma(t; P)$ starting at P strikes $\mathscr{C}$ after exactly α switchings. Clearly, since the function I is constant along any curve Γ which never leaves nor enters any of the manifolds $M^1, \cdots, M^l$, these manifolds contain the boundaries of the sets R_α. We shall, by integrating backwards from $\mathscr{C}$ along the trajectories $x_\sigma(t; P)$, show that (3) defines functions $J_1(\sigma; P), \cdots, J_N(\sigma; P)$ throughout R_0, which have the desired regularity there. And starting at the interface $\partial R_0 \cap \partial R_1$, the procedure may be repeated in order to extend the functions in a sufficiently smooth fashion to all of R_1, and so forth.

It will be convenient to partition R in yet another way. Let $j = \{j_1, \cdots, j_\alpha\}$ be some subset of the integers $\{1, \cdots, l\}$ which index the manifolds $M^1, \cdots, M^l$ belonging to σ. We will say that P lies in S_j if it lies on each of the manifolds $M^{j_1}, \cdots, M^{j_\alpha}$, and on no others. In particular, we denote by S_ϕ the set of all those points P which do not lie on any of the manifolds $M^1, \cdots, M^l$. Then every point P in $E_1 \times \mathscr{E}$ lies on some S_j. Now, for a given j, the set $\mathscr{C} \cap S_j$ may be decomposed into its various arcwise connected components. And if $\mathscr{C} \cap S_j$ is given its relative topology, then each such component is an open set.

Let S be one of these components, say of dimension β. It is clear that the solutions of the ordinary differential equation

$$\dot{x} = f(t, x, \sigma_1^{\mu_1}(t, x), \cdots, \sigma_N^{\mu_N}(t, x)) \tag{7}$$

which pass through S form, at least in some neighborhood of S, a manifold of dimension $\beta + 1$ for each $\mu = (\mu_1, \cdots, \mu_N)$ in the range of the indicator function I belonging to σ. If $S(\mu)$ is any such manifold, we shall form a subset $S'(\mu)$ by deleting from $S(\mu)$ every point P which is not in R_0, or for which there exists no neighborhood $N_\varepsilon(P)$ in $E_1 \times \mathscr{E}$ such that $I(t, x)$ is identically equal to μ on $N_\varepsilon(P) \cap S(\mu)$. It may now be the case that $S(\mu)$ is empty for every μ. If so, none of the trajectories $x_\sigma(t; P)$ beginning in R terminate on S.

If $S'(\mu)$ is not empty, it is a submanifold of $S(\mu)$, whose boundary is composed of certain portions of the manifolds $M^1, \cdots, M^l$ and S and which is disjoint from its boundary. And since $I(P) = \mu$ for every point P in $S'(\mu)$, the solutions of (7) which fill $S'(\mu)$ and terminate on S are also solutions of (5). That is, they are among the trajectories generated by σ.

Next let P be an arbitrary point of R_0. By assumption, a unique trajectory $x_\sigma(t; P)$ joins P to $\mathscr{C}$. Let Q be the point at which $x_\sigma(t; P)$ strikes $\mathscr{C}$ and denote by $(t_f(P), s_f(P))$ the coordinates of Q on $\mathscr{C}$. Q must lie in one of the sets S_j, since every point in $E_1 \times \mathscr{E}$ does, and hence in one of the connected components S of that S_j. Moreover, since P is in R_0, $I(t, x_\sigma(t; P))$ must be constant on the interval $\tau \leqq t \leqq t_f(P)$. Thus, if we denote by μ the value which I assumes during this time, $x_\sigma(t; P)$ must in fact be a solution of (7), and P must be a point of $S'(\mu)$. If $\mathscr{S}$ were another connected component of one of the sets S_j, and P lay on $\mathscr{S}'(\nu)$ as well, then two different trajectories would start at P, contrary to our assumption (i) of uniqueness. Hence each point P in R_0 must lie on exactly one of the manifolds $S'(\mu)$. Next we observe that the functions $J_i(\sigma; P)$ defined by (3) are of class C^1 on each such manifold $S'(\mu)$.

To see that this is indeed the case, fix $S'(\mu)$ and observe that the solutions $x_\mu(t; P)$ of (7) starting there may not strike $\mathscr{C}$ tangentially. For such curves are solutions of (5) as well, and so are subject to condition (ii). Hence the quantities t_f and s_f previously defined must be continuously differentiable functions of P, in the sense of differentiation on the manifold $S'(\mu)$. And the function

$$(8) \qquad J_i^A(P) = K_i(t_f(P), s_f(P))$$

must also be a continuously differentiable function of P, which is constant along trajectories. Similarly the function

$$(9) \qquad J_i^B(P) = \int_\tau^{t_f(P)} L_i(t, x_\mu(t; P), \sigma_1^{\mu_1}(t, x_\mu(t; P)), \cdots, \sigma_N^{\mu_N}(t, x_\mu(t; P)))\, dt$$

is a C' function of the point P, since the functions L_i, σ_j^k and $x_\mu(t; P)$ all depend in a C' fashion upon P and upon each other. Hence $J_i(\sigma; \tau, \eta) = J_i^A(\tau, \eta) + J_i^B(\tau, \eta)$ is also a C' function of the point P in the manifold $S'(\mu)$. Once again, it is to be emphasized that the differentiation referred to here is only differentiation on the manifold $S'(\mu)$. The functions J_i need not even be continuous across a given manifold. And since, as we observed, the manifolds $S'(\mu)$ exhaust R_0, the desired result has been obtained for this region. We shall call each $S'(\mu)$ a "manifold of regularity" for the functions J_i. Each point P in R_0 lies on exactly one such manifold.

Next, let us consider again a particular $S'(\mu)$ and follow the solutions $x_\mu(t)$ of (7) backward through time. There are four possibilities. A curve may remain in $S'(\mu)$ for all $t < t_f$. Or there may be a first instant τ beyond which the solution $x_\mu(t)$ cannot be continued. Or there may be a first instant τ beyond which $x_\mu(\tau)$ may be defined, but beyond which the solution $x_\sigma(t; \tau, x_\mu(\tau))$ with which it agrees cannot be continued. In any of these three cases, $x_\mu(t)$ is a complete solution of the system (5), defined on the intervals $-\infty < t < t_f$ and $\tau < t < t_f$ respectively. The last possibility is that there be a first instant τ beyond which the curve $x_\sigma(t; \tau, x_\mu(t))$ may be continued (in one or more ways, since we are moving backward through time), but not without a switching. In this case we shall call $(\tau, x_\mu(\tau))$ an "end point of $S'(\mu)$" and the set of all such points the "end set" $B(\mu)$ of $S'(\mu)$. Clearly every end point of $S'(\mu)$ must lie on one or more of the switching manifolds $M^1, \cdots, M^l$ belonging to σ. That is, every end point $(\tau, x_\mu(\tau))$ of $S'(\mu)$ must lie in some S_j. Moreover, every point of $\partial R_0 \cap \partial R_1$ must either lie in one of the sets $S'(\mu)$ or else be an end point of such a set.

Now let T be either $S'(\mu)$ itself or else some subset of $B(\mu)$ which is maximal with respect to the property that I remains constant on T. In either case the functions $J_i(\sigma; \tau, \eta)$ are defined and of class C' on T. We may continue the functions $J_i(\sigma; \tau, \eta)$ into R_1 from R_0 in exactly the same way that we continued them from $\mathscr{C}$ into R_0. We merely take T in place of S and form all the "manifolds of regularity" $T'(\mu)$ which are filled with solutions of (5) which begin in R_1, pass through T and strike $\mathscr{C}$ at some point of S. Letting T and μ take on all possible values, we generate all of the trajectories which begin in R_1 and end on S, and, by varying S as before, we obtain all the trajectories which strike $\mathscr{C}$ after only a single switching. Once again, each point P in R_1 lies on exactly one "manifold of regularity" $T'(\mu)$, and the functions $J_i(\sigma; \tau, \eta)$ are of class $\mathscr{C}'$ at P in the sense of differentiation within that manifold. Continuing in this manner, we will eventually generate all the solutions of (5) which begin in R and reach $\mathscr{C}$. Each point P in R will lie in exactly one R_α, and hence on exactly one manifold of regularity for the functions $J_i(\sigma; \tau, \eta)$. These will again be of class $\mathscr{C}'$ at P, in the sense of differentiation on the appropriate manifold.

If, in particular, $\sigma = \sigma^*$ should be an equilibrium point for our game, we define the value functions $V_1, \cdots, V_N$ by

$$V_i(\tau, \eta) = J_i(\sigma^*; \tau, \eta), \qquad i = 1, \cdots, N. \tag{10}$$

Thus the value functions are defined only at those points (τ, η) from which the terminal surface is reached after at most a finite number of switchings. But they are, in the preceding sense, continuously differentiable at every such point. This fact will enable us, in the next section, to write down a system of first order partial differential equations, which the value functions must obey and whose solutions will provide us with complete information about the equilibrium point σ^*, the equilibrium strategies $\sigma_i^*(t, x)$, the equilibrium trajectories $x^*(t; \tau, \eta)$ and, of course, the value functions $V_i(\tau, \eta)$ themselves. These equations will reduce to Isaacs' "main equation" in the case of a zero-sum two player game and to the so-called "Bellman equation" if only one player is involved.

4. The Hamilton–Jacobi equations. Let us now consider a point $P(\tau, \eta)$ in R, and let S be the manifold of regularity for σ^* (and hence for the functions $V_1(t, x), \cdots, V_N(t, x)$) upon which P lies. For ease of exposition, we shall assume that S may be globally coordinatized by the quantities t and $y = y_1, \cdots, y_\beta$. Since the indicator function $I^*(t, x)$ is constant on every manifold of regularity, the equilibrium strategies $\sigma_1^*, \cdots, \sigma_N^*$ for the several players, when restricted to S, are just the tactics $\sigma_1^{I_1(P)}(t, x), \cdots, \sigma^{I_N(P)}(t, x)$, which may for our purposes be written $\sigma_1^*(t, y), \cdots, \sigma_N^*(t, y)$. For, if all the players $j \neq i$ agree to utilize the strategies $u_j = \sigma_j^*$ throughout the game, then they will utilize the tactics $u_j = \sigma_j^*(t, y)$ so long as the state of the game remains in S. Suppose that they have agreed to do so, and let $N(P)$

be a neighborhood of P which intersects no switching manifold of σ^* save, perhaps, S itself. Define the function $K(t, y)$ on $\partial N(P) \cap S$ by

$$K(t, y) = V_i(t, y) \tag{11}$$

and consider the ordinary optimization problem: Find

$$\begin{aligned} \max_{u_i \in U_i} J_i(\sigma; \tau, \xi) &= K(t_f, y(t_f)) \\ &\quad + \int_\tau^{t_f} L_i(t, y(t), \cdots, \sigma^*_{i-1}(t, y(t)), u_i, \sigma^*_{i+1}(t, y(t)), \cdots)\, dt, \end{aligned} \tag{12}$$

subject to

$$\dot{y} = g(t, y, \cdots, \sigma^*_{i-1}(t, y), u_i, \sigma^*_{i+1}(t, y), \cdots) \tag{13}$$

and

$$y(\tau) = \xi, \qquad (t_f, y(t_f)) \in \partial N(P), \tag{14}$$

where the function $g(t, y, u_1, \cdots, u_N)$ is defined by

$$f(t, x, u_1, \cdots, u_N) = g(t, y, u_1, \cdots, u_N) \tag{15}$$

for every point $(t, x) = (t, y)$ on S. By the definition of an equilibrium point, σ_i^* is an optimal strategy for the problem (12)–(14). Therefore, in particular, $\sigma_i^*(t, y)$ is optimal in the class of all those functions $u_i = \sigma_i(t, y)$ for which the solution of (13) remains on S, at least until it strikes $\partial N(P)$. Let $T_{t,y}S$ denote the tangent space to S at (t, y). It is clear that if the points (t, y) are interior to S and $\sigma_i(t, y)$ is an element of the set $UA_i(t, y)$ defined by

$$UA_i(t, y) = \{u_i \in U_i : g(t, y, \cdots, \sigma^*_{i-1}(t, y), u_i, \sigma^*_{i+1}(t, y), \cdots) \in T_{t,y}S\} \tag{16}$$

for each point (t, y) of $N(P) \cap S$, then the solution of (13) corresponding to $\sigma_i(t, y)$ does not escape from S before reaching $\partial N(P)$. Finally, we note that if $u_i \in UA_i(t, y)$, there is a vector $\tilde{g}(t, y, \cdots, \sigma^*_{i-1}(t, y), u_i, \sigma^*_{i+1}(t, y), \cdots)$ in $T_{t,y}S$ which coincides with $g(t, y, \cdots)$. Since $\tilde{g}$ is in $T_{t,y}S$, it has only β components instead of n, and we can form its inner product with other vectors of dimension β.

We now observe that the optimal payoffs $V_i(\tau, \xi) = J_i(\sigma^*; \tau, \xi)$ in the problem (12)–(14) must satisfy Bellman's equation

$$\begin{aligned} \max_{u_i \in UA_i} (L_i + \langle \nabla_\xi V_i, \tilde{g} \rangle) &= -\nabla_\tau V_i(\tau, \xi) \\ &= \max_{u_i \in UA_i} \{L_i(\tau, \xi, \cdots, \sigma^*_{i-1}(\tau, \xi), u_i, \sigma^*_{i+1}(\tau, \xi), \cdots) \\ &\quad + \langle \nabla_\xi V_i(\tau, \xi), \tilde{g}(\tau, \xi, \cdots, \sigma^*_{i-1}(\tau, \xi), u_i, \sigma^*_{i+1}(\tau, \xi), \cdots) \rangle, \end{aligned} \tag{17}$$

as is well known[3] in the event that S is of dimension $n + 1$ and P is an interior point of S. A careful derivation of this equation is to be found in the article [5] by Berko-

[3] In order that (17) make sense, it is necessary that $\nabla_\xi V_i$ and $\tilde{g}$ have β components. The notation $g(t, y, \cdots) \in T_{t,y}S$ in (16) is an abuse of language.

witz, where it is obtained for an arbitrary zero-sum two player game. But clearly it holds as well for the one player game (12)–(14). The derivation in the more general case of a lower dimensional manifold S and a point P, which may lie on the boundary of S, requires only a trivial modification of Berkovitz' proof, and we shall not bother to reproduce it here. We remark in passing, that if S is of full dimension and P is interior to S, then UA_i, ξ and $\tilde{g}$ in (17) may be replaced by U_i, η and f respectively.

Next, let us apply another often useful result of Berkovitz to the problem (12)–(14). Let us suppose that M_i is a switching manifold for player i, separating two manifolds S and S' of regularity for the functions $V_1(t, y), \cdots, V_N(t, y)$, each of dimension $\beta + 1$, and that if $i \neq j$ the function $I_j^*(t, y)$ is continuous across M_i. Also suppose that the optimal trajectories $y^*(t)$ neither enter nor leave M_i tangentially. Then M_i is called a transition manifold for player i, and the functions $\nabla_x V_i$ and $\nabla_t V_i$ are continuous across M_i. This result, too, is proved for two player games, but holds equally well for the one player game (12)–(14).

We summarize the results of the preceding two sections in the following theorem.

THEOREM 1. *Suppose that the game described above has an equilibrium point* $\sigma^* = \sigma_1^*, \cdots, \sigma_N^*$ *in the class of N-tuples of switching strategies. Then the subset* R *of points of* $E_1 \times \mathscr{E}$ *from which termination is achieved after only a finite number of switchings is partitioned into a collection of disjoint* $\mathscr{C}'$ *manifolds on each of which the value functions* $V_1, \cdots, V_N$ *are of class* $\mathscr{C}'$. *On each such manifold, the equations* (17) *hold for* $i = 1, \cdots, N$. *If* M_i *is a transition manifold for player i, then* $\nabla_x V_i$ *and* $\nabla_t V_i$ *are continuous across* M_i.

In analogy with control theory and the theory of two player games, this theorem suggests the following procedure. Let S be of dimension n, for the moment, let $p_1, \cdots, p_N$ be N arbitrary vectors in E_n, and define the N Hamiltonian functions $H_1, \cdots, H_N$ by

$$(18) \qquad H_i(t, x, u_1, \cdots, u_N, p_i) = L_i(t, x, u_1, \cdots, u_N) + \langle p_i, f(t, x, u_1, \cdots, u_N)\rangle.$$

The equations (17), which we shall henceforth call the "main equations," after Isaacs, may now be written:

$$(19) \qquad \nabla_t V_i + \max_{u_i} H_i(t, x, \cdots, \sigma_{i-1}^*(t, x), u_i, \sigma_{i+1}^*(t, x), \cdots, \nabla_x V_i) = 0.$$

Next fix t, x and $p_1, \cdots, p_N$, and consider the game over Euclidean space defined by the payoff functions $H_1, \cdots, H_N$ and the strategy sets $U_1, \cdots, U_N$. If this game has an equilibrium point $u_1 = u_1^*(t, x, p), \cdots, u_N = u_N^*(t, x, p)$, where $p = (p_1, \cdots, p_N)$ is an element of E_{nN}, then

$$(20) \qquad \begin{aligned} &H_i(t, x, u_1^*(t, x, p), \cdots, u_N^*(t, x, p), p_i) \\ &\qquad = \max_{u_i} H_i(t, x, \cdots, u_{i-1}^*(t, x, p), u_i, u_{i+1}^*(t, x, p), \cdots, p_i). \end{aligned}$$

Therefore the substitutions $u_i = u_i^*(t, x, \nabla_x V_1(t, x), \cdots, \nabla_x V_N(t, x))$ and $\sigma_j(t, x) = u_j^*(t, x, \nabla_x V_1(t, x), \cdots, \nabla_x V_N(t, x))$, $j \neq i$ into (19) yields

$$\begin{aligned}\nabla_t V_i(t,x) + H_i(t, x, u_1^*(t, x, \nabla_x V_1(t,x), \cdots, \nabla_x V_N(t,x)), \cdots,\\ u_N^*(t, x, \nabla_x V_1(t,x), \cdots, \nabla_x V_N(t,x)), \nabla_x V_i(t,x)) = 0\end{aligned} \tag{21}$$

for each $i = 1, \cdots, N$. The system (21) consists of N first order partial differential equations for the N unknown functions $V_1, \cdots, V_N$, and it is to be hoped that solving (21), subject to the boundary condition that $V_i(t,x) = K_i(t,x)$ on $\mathscr{C}$, and substituting the gradients $\nabla_x V_i$ into the functions $u_i^*(t, x, p)$ will yield the equilibrium strategies $\sigma_i^*(t, x)$. In the next section, we shall prove a theorem to the effect that the analogy is apt, for under certain circumstances the preceding program may indeed be carried out. And in the section after that, we shall study a class of games for which the equations (21), hereinafter called the Hamilton–Jacobi equations, may be solved explicitly. We remark that equations (21) hold as well on any manifold S of regularity, except that they must be written in terms of the manifold coordinates t and y instead of the E_{n+1} coordinates t, x.

5. A sufficiency theorem.[4] We shall begin this section by proving a lemma which is due, in its original form, to Carathéodory. Our sufficiency test, which represents a slight strengthening of the necessary condition (19), will then emerge as a direct consequence of the lemma. Let S be the closure of an open subset of $E_1 \times \mathscr{E}$, and let σ^* be an N-tuple of tactics (i.e., strategies whose indicator functions are constant on S) which transfer every point (τ, η) of S to $\mathscr{C}$ in such a way that the trajectories $x^*(t; \tau, \eta)$ never leave S. Next we assume that the payoffs are purely integral (that is, $K_i = 0$ for $i = 1, \cdots, N$), that the integrands are of the special form $L_i(t, x, u_i)$ and that the functions L_i have, at each point (t, x) in S, zero as a unique absolute maximum in u_i. Moreover, they assume their respective maxima at the points $u_i = \sigma_i^*(t, x)$ in U_i. That is, we assume that

$$L_i(t, x, \sigma_i^*(t,x)) = \max_{u_i \in U_i} L_i(t, x, u_i) = 0, \tag{22}$$

the maximum being unique and absolute.

LEMMA. *σ^* is an equilibrium point in the class of all those strategy N-tuples which transfer points (τ, η) of S to $\mathscr{C}$ in the required manner, and the value of each point in S to the player i is zero. That is,*

$$\begin{aligned} J_i(\sigma^*; \tau, \eta) &= 0,\\ J_i(\sigma; \tau, \eta) &\leqq 0\end{aligned} \tag{23}$$

for any other strategy N-tuple $\sigma = (\cdots, \sigma_{i-1}^*, \sigma_i, \sigma_{i+1}^*, \cdots)$ which transfers (τ, η) to $\mathscr{C}$ without leaving S.

[4] Throughout this section, we shall restrict our attention to games without switchings. It is hoped that the resulting loss in generality will be justified by the considerable simplification in exposition which it permits.

The proof of the lemma is immediate, for

$$J_i(\sigma^*;\tau,\eta) = \int_\tau^{t_f} L_i(t, x^*(t), \sigma_i^*(t, x^*(t)))\,dt = 0$$

and

$$J_i(\sigma;\tau,\eta) = \int_\tau^{t_f} L_i(t, x_\sigma(t), \sigma_i(t, x_\sigma(t)))\,dt \leqq 0,$$

since the integrand L_i is assumed nonpositive.

Now, let us single out a class of games in which the procedure described in the previous section may be justified. Most games studied by Isaacs [8] and all games studied by Pontryagin [10] are of the following class. Suppose that $H_1, \cdots, H_N$ are the Hamiltonian functions for a particular differential game $\mathscr{G}$, and suppose that for each p in E_{nN} and for each (t, x) in S there is a unique control $u^*(t, x, p) = u_1^*(t, x, p), \cdots, u_N^*(t, x, p)$ in $U_1 \times \cdots U_N$ for which the inequalities

$$\begin{aligned}(24)\qquad & H_i(t, x, \cdots, u_{i-1}^*(t, x, p), u_i, u_{i+1}^*(t, x, p), \cdots, p_i) \\ & \quad < H_i(t, x, u_1^*(t, x, p), \cdots, u_N^*(t, x, p))\end{aligned}$$

hold for every $u_i \neq u_i^*(t, x, p)$ in U_i and for each $i = 1, \cdots, N$. Then we shall say that $\mathscr{G}$ is "normal relative to S" and that $u_1^*(t, x, p), \cdots, u_N^*(t, x, p)$ are the "normalizing controls." In particular, if S is all of $E_1 \times \mathscr{E}$, we shall call $\mathscr{G}$ a normal game. We may now state the following theorem.

THEOREM 2. *Suppose that $\mathscr{G}$ is normal relative to S, and that $u_1^*(t, x, p), \cdots, u_N^*(t, x, p)$ are the normalizing controls. Let σ^* be an N-tuple of tactics which transfers every point (τ, η) in S to $\mathscr{C}$ in such a way that $x^*(t;\tau,\eta)$ never leaves S before striking $\mathscr{C}$. Also, let $V_1(t, x), \cdots, V_N(t, x)$ be solutions of the Hamilton–Jacobi equations* (21) *such that*

$$(25)\qquad V_i(t, x) = 0$$

for all points (t, x) on $\mathscr{C}$ and for each $i = 1, \cdots, N$ and such that

$$(26)\qquad \sigma_i^*(t, x) = u_i^*(t, x, \nabla_x V_1(t, x), \cdots, \nabla_x V_N(t, x))$$

for each (t, x) in S and for $i = 1, \cdots, N$. Then σ^* is an equilibrium point in the class of all strategy N-tuples which transfer points (τ, η) to $\mathscr{C}$ in the required manner, and

$$(27)\qquad J_i(\sigma^*;\tau,\eta) = V_i(\tau,\eta), \qquad i = 1, \cdots, N.$$

Proof. Let us consider the functions $p_1(t, x) = \nabla_x V_1(t, x), \cdots, p_N(t, x) = \nabla_x V_N(t, x)$, the nN vector $p(t, x) = p_1(t, x), \cdots, p_N(t, x)$ and the new payoff function $\hat{L}_i(t, x, u_i)$ defined by

$$\begin{aligned}(28)\qquad \hat{L}_i(t, x, u_i) &= \nabla_t V_i(t, x) + H_i(t, x, \cdots, u_{i-1}^*(t, x, p(t, x)), \\ & \qquad u_i, u_{i+1}^*(t, x, p(t, x)), \cdots, p_i(t, x)).\end{aligned}$$

We claim that the $\hat{L}_i$'s satisfy the conditions of the Carathéodory lemma. To see that this is so, note that

$$\begin{aligned}(29)\quad &\hat{L}_i(t, x, \sigma_i^*(t, x)) \\ &\quad = \nabla_t V_i + H_i(t, x, u_1^*(t, x, p(t, x)), \cdots, u_N^*(t, x, p(t, x)), p_i(t, x)) = 0\end{aligned}$$

because the functions $V_1(t, x), \cdots, V_N(t, x)$ satisfy the Hamilton–Jacobi equations (21) and because of relation (26). Furthermore, we observe that

$$(30)\qquad \hat{L}_i(t, x, \sigma_i^*(t, x)) = \max_{u_i \in U_i} \hat{L}_i(t, x, u_i) = 0,$$

since $\sigma_i^*(t, x) = u_i^*(t, x, p(t, x))$, so that, by the lemma[5]

$$(31)\qquad \begin{aligned}&\int_\tau^{t_f} \hat{L}_i(t, x^*(t), \sigma_i^*(t, x^*(t)))\, dt = 0, \\ &\int_\tau^{t'_f} \hat{L}_i(t, x_\sigma(t), \sigma_i(t, x_\sigma(t)))\, dt \leqq 0,\end{aligned}$$

where σ is any other strategy N-tuple of the form $\sigma = (\cdots, \sigma_{i-1}^*, \sigma_i, \sigma_{i+1}^*, \cdots)$ which transfers points (τ, η) of S to $\mathscr{C}$ in the manner required in the theorem. But

$$\begin{aligned}\hat{L}_i(t, x, u_i) &= \nabla_t V_i(t, x) + H_i(t, x, \cdots, u_{i-1}^*(t, x, p(t, x)), u_i, \\ &\qquad\qquad u_{i+1}^*(t, x, p(t, x)), \cdots, p_i(t, x)) \\ &= \nabla_t V_i(t, x) + L_i(t, x, \cdots, u_{i-1}^*(t, x, p(t, x)), \\ &\qquad\qquad u_{i+1}^*(t, x, p(t, x)), \cdots) \\ &\quad + \langle p_i(t, x), f(t, x, \cdots, u_{i-1}^*(t, x, p(t, x)), u_i, \\ &\qquad\qquad u_{i+1}^*(t, x, p(t, x)), \cdots)\rangle.\end{aligned}$$

$$\begin{aligned}\hat{L}_i(t, x^*(t), \sigma_i^*(t, x^*(t))) &= L_i(t, x^*(t), \sigma_i^*(t, x^*(t))), \cdots, \sigma_N^*(t, x^*(t))) \\ &\quad + \frac{d}{dt} V_i(t, x^*(t))\end{aligned}$$

and that

$$\begin{aligned}&\hat{L}_i(t, x_\sigma(t), \sigma_i(t, x_\sigma(t))) \\ &\quad = L_i(t, x_\sigma(t), \cdots, \sigma_{i-1}^*(t, x_\sigma(t)), \sigma_i(t, x_\sigma(t)), \sigma_{i+1}^*(t, x_\sigma(t)), \cdots) + \frac{d}{dt} V_i(t, x_\sigma(t))\end{aligned}$$

[5] Here, t'_f is the instant at which $x_\sigma(t)$ strikes $\mathscr{C}$. Notice that t'_f need not equal t_f in this argument.

whenever σ is of the class under discussion. Hence

$$
\begin{aligned}
&\int_{\tau}^{t_f} \hat{L}_i(t, x^*(t), \sigma_i^*(t, x^*(t)))\,dt = J_i(\sigma^*; \tau, \eta) - V_i(\tau, \eta) + V_i(t_f, x^*(t_f)),\\
(32)\qquad &\\
&\int_{\tau}^{t_f'} \hat{L}_i(t, x_\sigma(t), \sigma_i(t, x_\sigma(t)))\,dt = J_i(\sigma; \tau, \eta) - V_i(\tau, \eta) + V_i(t_f', x_\sigma(t_f')).
\end{aligned}
$$

But because of the hypothesis (25), $V_i(t_f, x^*(t_f))$ and $V_i(t_f', x_\sigma(t_f'))$ are equal to zero, so that we may conclude finally that

$$
\begin{aligned}
&J_i(\sigma^*; \tau, \eta) - V_i(\tau, \eta) = 0,\\
(33)\qquad &\\
&J_i(\sigma; \tau, \eta) - V_i(\tau, \eta) \leqq 0
\end{aligned}
$$

for any strategy N-tuple σ of the type considered. Since i was arbitrary, this completes the proof.

Before proceeding further, we should remark that Theorem 2 is proved under the assumptions that S is of full dimension and that the payoff functionals J_i are in integral form. This too is done for ease in exposition, and is in no way essential. In particular, a game in which terminal payoffs are present may be reduced to one with a purely integral payoff by the usual transformations of the variational calculus. It should be remarked, also, that Theorem 2 bears a striking resemblance to Isaacs' "verification theorem." Like the latter, it is only a local sufficient condition, because all the statements are relative to the subregion S of $E_1 \times \mathscr{E}$. Indeed, the principal difficulty in applying the theorem is often the finding of a suitable region S in which to apply it. In the following section, we discuss a class of games for which this difficulty does not arise.

6. A class of differential games. In this section we consider a system governed by

$$
\begin{aligned}
&\dot{x}(t) = A(t)x(t) + B_1(t)u_1(t) + \cdots + B_N(t)u_N(t) + f(t),\\
(34)\qquad &\\
&y_1(t) = C_1(t)x(t), \quad \cdots, \quad y_N(t) = C_N(t)x(t).
\end{aligned}
$$

Here $x(t)$ is called the state of the system and $y_1(t), \cdots, y_N(t)$ are called its outputs. We shall suppose that certain "output schedules" $z_1(t), \cdots, z_N(t)$ are given and that each player i wishes to keep his output $y_i(t)$ "close" to his assigned output schedule $z_i(t)$, over the (fixed) time interval $[0, T]$ during which the game is played. Rather than impose "hard" constraints on the variables $u_1, \cdots, u_N$ and x, we shall assume that $U_1, \cdots U_N$ and $\mathscr{E}$ are the entire spaces $E_{m_1}, \cdots, E_{m_N}$ and E_n, respectively, and introduce a penalty term for excessive "control effort" into the cost integrand, in order to keep the players from employing excessively large values of the control variables $u_1, \cdots, u_N$. That is, we shall assume that each player seeks to minimize an integral of the form

$$J_i = \int_0^T \{\tfrac{1}{2}\langle(z_i(t) - y_i(t)), Q_i(t)(z_i(t) - y_i(t))\rangle + \tfrac{1}{2}\langle u_i(t), R_i(t)u_i(t)\rangle\}\,dt, \tag{35}$$

where the matrix $Q_i(t)$ is a positive semidefinite symmetric matrix and is not the zero-matrix, while $R_i(t)$ is positive definite. If $Q_i(t)$ were zero, the optimal policy for i would be $u_i \equiv 0$.

We shall solve the game by the use of Theorem 2. To this end we write down the ith Hamiltonian function

$$\begin{aligned} &H_i(t, x, u_1, \cdots, u_N, p_i) \\ &\quad = \tfrac{1}{2}\langle(z(t) - C_i(t)x), Q_i(t)(z(t) - C_i(t)x)\rangle + \tfrac{1}{2}\langle u_i, R_i(t)u_i\rangle \\ &\quad + \langle f(t), p_i\rangle + \langle A(t)x, p_i\rangle + \langle B_1(t)u_1, p_i\rangle + \cdots + \langle B_N(t)u_N, p_i\rangle \end{aligned} \tag{36}$$

and observe that the normalizing control $u_i^*(t, x, p)$ will be that u_i for which the sum

$$\tfrac{1}{2}\langle u_i, R_i(t)u_i\rangle + \langle B_i(t)u_i, p_i\rangle \tag{37}$$

takes on the least value (since we are trying to minimize J_i, we minimize H_i instead of maximizing it). Since the gradient of (37) is just

$$R_i(t)u_i + B_i'(t)p_i, \tag{38}$$

and the gradient of (38) is the positive definite matrix $R_i(t)$, it is clear that $u_i^*(t, x, p)$ renders (37) a minimum if and only if it renders (38) equal to zero. Hence

$$u_i^*(t, x, p) = -R_i^{-1}(t)B_i'(t)p_i \tag{39}$$

and the Hamilton–Jacobi equations (21) may be written

$$\begin{aligned} &\nabla_t V_i + \tfrac{1}{2}\langle(z_i(t) - C_i(t)x), Q_i(t)(z_i(t) - C_i(t)x)\rangle\tfrac{1}{2}\langle S_i(t)\nabla_x V_i, \nabla_x V_i\rangle \\ &\quad + \langle f(t), \nabla_x V_i\rangle + \langle A(t)x, \nabla_x V_i\rangle \\ &\quad - \langle S_1(t)\nabla_x V_i, \nabla_x V_i\rangle - \cdots - \langle S_N(t)\nabla_x V_N, \nabla_x V_i\rangle = 0, \end{aligned} \tag{40}$$

where $S_i(t) = B_i(t)R_i^{-1}(t)B_i'(t)$, $i = 1, \cdots, N$. Motivated by the results of Kalman [9], we shall try to solve the system (40) by separating variables. That is, we shall substitute the trial solutions

$$V_i(t, x) = \tfrac{1}{2}\langle x, K_i(t)x\rangle - \langle g_i(t), x\rangle + \Phi_i(t) \tag{41}$$

into (40) in the hope that the problem will reduce to one of ordinary differential equations. If we assume that the matrices $K_i(t)$ are symmetric and differentiate (41) with respect to x and t, we obtain

$$\begin{aligned} \nabla_t V_i(t, x) &= \tfrac{1}{2}\langle x, K_i(t)x\rangle - \langle \dot{g}_i(t), x\rangle + \Phi_i(t) \\ \nabla_x V_i(t, x) &= K_i(t)x - g_i(t), \end{aligned} \tag{42}$$

$i = 1, \cdots, N$; and substitution of these expressions into (40) yields

$$\begin{aligned}&\tfrac{1}{2}\langle x, \dot{K}_i(t)x\rangle - \langle \dot{g}_i(t), x\rangle + \Phi_i(t) + \tfrac{1}{2}\langle (z_i(t) - C_i(t)x), Q_i(t)(z_i(t) - C_i(t)x)\rangle\\ &\quad + \tfrac{1}{2}\langle S_i(t)(K_i(t)x - g_i(t)), (K_i(t)x - g_i(t))\rangle + \langle A(t)x, (K_i(t)x - g_i(t))\rangle\\ &\quad - \langle S_i(t)(K_i(t)x - g_i(t)), (K_i(t)x - g_i(t))\rangle - \cdots\\ &\quad - \langle S_N(t)(K_N(t)x - g_N(t)), K_i(t)x - g_i(t))\rangle + \langle f(t), (K_i(t)x - g_i(t))\rangle = 0.\end{aligned} \tag{43}$$

After some relatively tedious manipulations, (43) may be put in the form

$$\begin{aligned}&\langle x, (\dot{K}_i(t) + C_i'(t)Q_i(t)C_i(t) + K_i(t)S_i(t)K_i(t) + 2A'(t)K_i(t)\\ &\qquad - 2K_1(t)S_1(t)K_i(t) - \cdots - 2K_N(t)S_N(t)K_i(t))x\rangle\\ &\quad - \langle x, (2g_i(t) + 2C_i'(t)Q_i(t)z_i(t) - K_i(t)S_i(t)g_i(t)2A'(t)g_i(t)\\ &\qquad - 2K_i(t)S_1(t)g_1(t) - 2K_1(t)S_1(t)g_i(t)\\ &\qquad\qquad - 2K_i(t)S_N(t)g_N(t) - 2K_N(t)S_N(t)g_i(t) + K_i(t)f(t))\rangle\\ &\quad + 2\Phi_i(t) + \langle z_i(t), - Q_i(t)z_i(t)\rangle - \langle g_i(t), S_i(t)g_i(t)\rangle\\ &\quad - 2\langle S_1(t)g_1(t), g_i(t)\rangle - \cdots - 2\langle S_N(t)g_N(t), g_i(t)\rangle - \langle f(t), g_i(t)\rangle = 0.\end{aligned} \tag{44}$$

Now the first term in (44) is a quadratic form $\langle x, \mathscr{A}(t)x\rangle$ and is unchanged if we replace $\mathscr{A}(t)$ by its symmetrization $\frac{1}{2}(\mathscr{A}(t) + \mathscr{A}'(t))$. Hence if we demand that the matrices $K_1(t), \cdots, K_N(t)$ satisfy the equations

$$\begin{aligned}\dot{K}_i(t) &= K_1(t)S_1(t)K_i(t) + K_i(t)S_1(t)K_1(t) + \cdots + K_N(t)S_N(t)K_i(t)\\ &\qquad + K_i(t)S_N(t)K_N(t)\\ &\qquad - A'(t)K_i(t) - K_i(t)A(t) - K_i(t)S_i(t)K_i(t) - C_i'(t)Q_i(t)C_i(t),\end{aligned} \tag{45}$$

$i = 1, \cdots, N$, the first term in (44) will vanish identically. Similarly, if we require that

$$\begin{aligned}\dot{g}_i(t) &= K_i(t)S_1(t)g_1(t) + K_1(t)S_1(t)g_i(t)\\ &\qquad + \cdots + K_i(t)S_N(t)g_N(t) + K_N(t)S_N(t)g_i(t)\\ &\qquad - \tfrac{1}{2}K_i(t)S_i(t)g_i(t) - A'(t)g_i(t) - C_i'(t)Q_i(t)z_i(t) + K_i(t)f(t)\end{aligned} \tag{46}$$

and

$$\begin{aligned}\dot{\Phi}_i(t) &= \langle S_1(t)g_1(t), g_i(t)\rangle + \cdots + \langle S_N(t)g_N(t), g_i(t)\rangle\\ &\qquad - \tfrac{1}{2}\langle g_i(t), S_i(t)g_i(t)\rangle - \tfrac{1}{2}\langle z_i(t), - Q_i(t)z_i(t)\rangle - \langle f(t), g_i(t)\rangle,\end{aligned} \tag{47}$$

then the last two terms in (44) will also vanish identically, so that the functions (41) are indeed solutions of (40). Moreover, if we demand that $K_1(T), g_1(T), \Phi_1(T), \cdots, K_N(T), g_N(T), \Phi_N(T)$ all vanish, $V_1(t, x), \cdots, V_N(t, x)$ satisfy the boundary conditions (25) of Theorem 2 on the terminal manifold $\mathscr{C}$ given by $t = T$. In order to compute the functions (41), we begin by solving the system of $\frac{1}{2}Nn(n + 1)$ ordinary differential equations (45), subject to the initial conditions $K_1(T) = 0 = \cdots = K_N(T)$, on some interval $(a, T]$. When this has been done, the coefficients of the nN linear equations (46) are known so that these too may be solved in $(a, T]$ subject to $g_1(T) = 0 = \cdots = g_N(T)$. Then the functions $\Phi_1(t), \cdots, \Phi_N(t)$ in (47) may be obtained by simple quadrature.

Once this has been accomplished, the value functions (41) are known in the region $(a, T) \times E_n$ of E_{n+1}. If we define strategies $\sigma_1^*(t, x), \cdots, \sigma_N^*(t, x)$ by

$$\sigma_i^*(t, x) = u_i^*(t, x, \nabla_x V_i(t, x)) = - R_i^{-1} B_i'(t)(K_i(t)x - g_i(t)), \tag{48}$$

condition (26) of Theorem 2 is satisfied, and $\sigma^* = \sigma_1^*, \cdots, \sigma_N^*$ must indeed be an equilibrium point for our game. The equilibrium trajectories $x^*(t)$ are the solutions of

$$\begin{aligned} \dot{x}^*(t) = A(t)x^*(t) - B_1(t)R_1^{-1}(t)B_1'(t)(K_1(t)x^*(t) - g_i(t)) \\ - B_N(t)R_N^{-1}(t)B_N'(t)(K_N(t)x^*(t) - g_N(t)) + f(t). \end{aligned} \tag{49}$$

In short, the solution of the game has been reduced to the solution of the three systems (45), (46) and (47) of ordinary differential equations. These systems always have solutions in an interval $(a, T]$, but since (45) is nonlinear, one cannot be certain that the number a may be chosen to be arbitrarily small. For $N = 1$, this difficulty does not arise (see [9]) but we were unable to prove a similar result for arbitrary N.

It is not clear whether or not the class of games solved previously will ever find practical application, though certainly the results for $N = 1$ are widely used. Also, Y. C. Ho and his co-workers have obtained significant results from a zero-sum version of these games. But the principal value of the preceding is to demonstrate that N-player differential games can have solutions and that the procedure described at the end of § 4 can be used to calculate them, at least in certain special cases. In the next section we shall develop a "method of characteristics," which will sometimes be useful for the solution of the Hamilton–Jacobi equation.

7. A method of characteristics. In order to discover an appropriate technique for obtaining the solutions of the Hamilton–Jacobi equations, let us recall for a moment the process by which the players governed the system (34) of the previous example. To begin with, they each calculated the solutions $V_1(t, x), \cdots, V_N(t, x)$ of (40), together with their respective gradient directions $p_1(t, x) = \nabla_x V_1(t, x), \cdots, p_N(t, x) = \nabla_x V_N(t, x)$. Then at each point (t, x) of E_{n+1}, they substituted the appropriate gradients $p_1, \cdots, p_N$ into their respective "normalizing strategies" $u_1^*(t, x, p), \cdots, u_N^*(t, x, p)$ in order to obtain the optimal controls. Thus if we wish to calculate a particular equilibrium trajectory $x^*(t) = x^*(t; \tau, \eta)$, we must be able simultaneously to calculate the values of $p_1(t, x^*(t)), \cdots, p_N(t, x^*(t))$ needed to generate the optimal "control histories" $u_i^{**}(t) = u_i^*(t, x^*(t), p(t, x^*(t)))$ associated therewith. As in control theory, this may be done by solving a system of ordinary differential equations for the functions $p(t) = p(t, x^*(t)) = p_1(t, x^*(t)), \cdots, p_N(t, x^*(t))$. To derive the appropriate equations, let S be a particular manifold of regularity in the subset R of $E_1 \times \mathscr{E}$ associated with the given strategy N-tuple $\sigma^* = (\sigma_1^*, \cdots, \sigma_N^*)$. Assume, as before, that S may be globally coordinatized by the quantities t and y. Then write the Hamilton–Jacobi equations in the form

$$\nabla_t V_i(t, y) + H_i(t, y, \sigma_1^*(t, y), \cdots, \sigma_N^*(t, u), \nabla_y V_i(t, y)) = 0, \tag{50}$$

$i = 1, \cdots, N$, obtained by substituting $\sigma_i^*(t, y) = u_i^*(t, y, p(t, y))$ into (21). We now apply the operator ∇_y to either side of (50). In order to do this, we shall need to assume the functions $V_1(t, y), \cdots, V_N(t, y)$ to be not once, but twice, continuously differentiable on S. Obvious modifications of the construction of § 3 yield such functions V_i (cf. Berkovitz [4]).

Since the calculations are somewhat lengthy, we introduce the notations $\sigma^*(t, y) = \sigma_1^*(t, y), \cdots, \sigma_N^*(t, y)$ and $\sigma^{**}(t) = \sigma^*(t, y^*(t))$. Then

$$\begin{aligned}
&\nabla_y \nabla_t V_i(t, y) + \nabla_y H_i(t, y, \sigma^*(t, y), \nabla_y V_i(t, y)) \\
&\quad = \nabla_y \nabla_t V_i(t, y) + \nabla_y \langle \nabla_y V_i(t, y), g(t, y, \sigma^*(t, y)) \rangle \\
&\qquad + \nabla_y L_i(t, y, \sigma^*(t, y)) + \nabla_u L_i(t, y, \sigma^*(t, y)) \nabla_y \sigma^*(t, y) \\
&\quad = \nabla_t \nabla_y V_i(t, y) + \nabla_y \nabla_y V_i(t, y) g(t, y, \sigma^*(t, y)) \\
&\qquad + \nabla_y L_i(t, y, \sigma^*(t, y)) + \nabla_u L_i(t, y, \sigma^*(t, y)) \nabla_y \sigma^*(t, y) \\
&\qquad + \nabla_y V_i(t, y) \nabla_y g(t, y, \sigma^*(t, y)) + \nabla_y V_i(t, y) \nabla_u g(t, y, \sigma^*(t, y)) \nabla_y \sigma^*(t, y) = 0.
\end{aligned} \tag{51}$$

But, also, we know that

$$\begin{aligned}
\frac{d}{dt} \nabla_y V_i(t, y^*(t)) &= \nabla_y \nabla_y V_i(t, y^*(t)) \dot{y}^*(t) + \nabla_t \nabla_y V_i(t, y^*(t)) \\
&= \nabla_y \nabla_y V_i(t, y^*(t)) g(t, y^*(t), \sigma^{**}(t)) + \nabla_t \nabla_y V_i(t, y^*(t)),
\end{aligned} \tag{52}$$

which is just the fourth line of (51) evaluated along $y^*(t)$. Thus, if $p_i(t) = \nabla_y V_i(t, y^*(t))$, we have for each $i = 1, \cdots, N$:

$$\begin{aligned}
\dot{p}_i(t) &= \frac{d}{dt} \nabla_y V_i(t, y^*(t)) \\
&= -\nabla_y L_i(t, y^*(t), \sigma^{**}(t)) - \nabla_y V_i(t, y^*(t)) \nabla_y g(t, y^*(t), \sigma^{**}(t)) \\
&\quad - \nabla_u L_i(t, y^*(t), \sigma^{**}(t)) \nabla_y \sigma^*(t, x^*(t)) \\
&\quad - \nabla_y V_i(t, y^*(t)) \nabla_u g(t, y^*(t), \sigma^{**}(t)) \nabla_y \sigma^*(t, y^*(t)) \\
&= -\nabla_y H_i(t, y^*(t), \sigma^{**}(t), p_i(t)) - \nabla_u H_i(t, y^*(t), \sigma^{**}(t), p_i(t)) \nabla_y \sigma^*(t, y^*(t)).
\end{aligned} \tag{53}$$

These equations reduce, in the case of a two person zero-sum game, to the second of equations (5.17) obtained by Berkovitz [5], and, together with

$$\dot{y} = g(t, y, \sigma_1^*(t, y), \cdots, \sigma_N^*(t, y)), \tag{54}$$

they make up a system of characteristics for the equations (50). In particular, if S is of full dimension, we may replace y by x and g by f in (53) and (54). In order to obtain appropriate initial data for solving equations (53), we use once again the fact that $V_i(t, y)$ is the solution of the single player game (12)–(14). Hence the "costate vector" $p_i(t) = \nabla_y V_i(t, y^*(t))$ must satisfy the well-known "transversality condition" (cf. Athans and Falb [1, p. 303] or Hestenes [7, p. 344])

$$p_i(t_f) - \nabla_y K_i(t_f, y^*(t_f)) = k_i \nabla_y h(y^*(t_f)), \tag{55}$$

where $\mathscr{C}$ is given by $\{y : h(y) = 0\}$ and t_f is the time at which $y^*(t)$ meets $\mathscr{C}$ (or some other manifold playing a like role). These results may be summed up in the form of the following maximum principle.

THEOREM 3. *Let $\sigma^* = \sigma_1^*, \cdots, \sigma_N^*$ be an equilibrium point in switching strategies for $\mathscr{C}$, and let $x^*(t)$ be the corresponding trajectory having the properties $x^*(\tau) = \eta$ and $x^*(t_f) \in \mathscr{C}$, where (τ, η) is some point of R. Then $x^*(t)$, $\tau \leqq t \leqq t_f$, is contained in some finite number of manifolds S of regularity, on each of which the functions $V_1, \cdots, V_N$ are of class C^2, and on each such manifold we have*

$$(\mathrm{M}_1) \quad \begin{aligned} &H_i(t, y^*(t), \sigma_1^{**}(t), \cdots, \sigma_N^{**}(t), p_i(t)) \\ &\quad = \max_{u_i \in UA_i} H_i(t, y^*(t), \cdots, \sigma_{i-1}^{**}(t), u_i, \sigma_{i+1}^{**}(t), \cdots, p(t)), \end{aligned}$$

$$(\mathrm{M}_2) \quad \begin{aligned} \dot{y}^*(t) &= \nabla p_i H_i(t, y^*(t), \sigma_1^{**}(t), \cdots, \sigma_N^{**}(t), p_i(t)), \\ \dot{p}_i(t) &= -\frac{\partial H_i}{\partial y}(t, y^*(t), \sigma_1^{**}(t), \cdots, \sigma_N^{**}(t), p_i(t)), \end{aligned}$$

where

$$(\mathrm{M}_3) \quad \begin{aligned} \frac{\partial H_i}{\partial y} &= \nabla_y H_i + \nabla_{u_i} H_i \nabla_y \sigma_1^*(t, y) + \cdots + \nabla_{u_N} H_i \nabla_y \sigma_N^*(t, y), \\ p_i(t_f) &= k_i \nabla_y h(y^*(t_f)) + \nabla_y K_i(t_f, y^*(t_f)), \end{aligned}$$

and if the kinematic equations (1) *are autonomous,*

$$(\mathrm{M}_4) \quad H_i(t, y^*(t), \sigma_1^{**}(t), \cdots \cdot \sigma_N^{**}(t), p_i(t)) \equiv 0.$$

Here, as always, t and y refer to coordinates on the manifold S. If S is of full dimension, y may be replaced by x in statements (M_1) to (M_4), and the set UA_i may be taken to be all of U_i. Also, it should be noted that the transversality condition (M_3) may be applied at a point $y^*(t_\alpha)$, where $y^*(t)$ passes from R_α into $R_{\alpha+1}$ (once V_i is known in R_α) simply by taking $K_i(t, y) = V_i(t, y)$ on $\partial R_\alpha \cap \partial R_{\alpha+1}$.

At this point, the principal difference between two player zero-sum games and our many player problems becomes apparent. For Isaacs' constructive procedure for solving differential games is simply to integrate the characteristics (M_2) backward in time from the terminal surface $\mathscr{C}$. And the principal reason he is so often able to do so is that, in most cases of interest, he is able to show $\partial H_i/\partial y = \nabla_y H_i$. However, for many player games, this is the case only in isolated instances,[6] and one is generally unable to eliminate the appearances of $\sigma_1^*(t, y), \cdots, \sigma_N^*(t, y)$

[6] For instance, it is not true of the linear-quadratic games discussed in the previous section.

in the right side of the equations for $p_1, \cdots, p_N$. And while this is not always an insurmountable difficulty, it does add materially to the complexity of the problem. It is our own belief that, while the study of two person zero-sum differential games may concern itself with ordinary differential equations, the many player theory must, in general, deal directly with the Cauchy problem for the Hamilton–Jacobi partial differential equations.

REFERENCES

[1] M. Athans and P. L. Falb, *Optimal Control: An Introduction to the Theory and its Applications*, McGraw-Hill, New York, 1966.

[2] R. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, 1957.

[3] L. D. Berkovitz and W. H. Fleming, *On differential games with integral payoffs*, Contributions to the Theory of Games, vol. 3, Annals of Mathematics Studies, no. 39, Princeton University Press, Princeton, 1957, pp. 413–435.

[4] L. D. Berkovitz, *A variational approach to differential games*, Advances in Game Theory, Annals of Mathematics Studies, no. 52, Princeton University Press, Princeton, 1964, pp. 127–174.

[5] L. D. Berkovitz, *Necessary conditions for optimal strategies in a class of differential games and control problems*, this Journal, 5 (1967), pp. 1–24.

[6] J. H. Case, *Equilibrium points of N-person differential games*, Doctoral thesis, University of Michigan, Ann Arbor, 1967.

[7] M. R. Hestenes, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.

[8] R. Isaacs, *Differential Games*, John Wiley, New York, 1965.

[9] R. E. Kalman, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana, 5 (1960), pp. 102–119.

[10] L. S. Pontryagin, *On the theory of differential games*, Uspekhi Mat. Nauk., 21 (1966), pp. 219–274.

SIAM J. CONTROL
Vol. 7, No. 2, May 1969

BOUNDARY CONTROLLABILITY OF NONLINEAR HYPERBOLIC SYSTEMS*

MARCO CIRINÀ†

1. Introduction. This paper is concerned with the existence of boundary controls for quasilinear systems of hyperbolic partial differential equations in two independent variables. Specificially, we consider the boundary control problem

$$z_t + A(x,t,z)z_x = f(x,t,z), \qquad (x,t) \in \Re(1,T), \tag{1.1}$$

$$z(x,0) = \phi(x), \qquad x \in [0,1], \tag{1.2}$$

$$\bar{z}(0,t) = \bar{u}(t), \qquad \underline{z}(1,t) = \underline{u}(t), \qquad t \in [0,T], \tag{1.3}$$

$$z(x,T) = 0, \qquad x \in [0,1], \tag{1.4}$$

where $\Re(1,T) = [0,1] \times [0,T]$, A, f, ϕ are given functions, A is matrix-valued, f, ϕ vector-valued and $\bar{z}$, $\underline{z}$ is an appropriate partition of the components of the vector z. The data and the coefficients of (1.1) will always be assumed to possess continuous first derivatives and so the solution will be a C^1 function satisfying the equation everywhere on its domain of definition.

Our main object is to find conditions on A, f, ϕ which insure that there is a real number $T > 0$ and functions $\bar{u}$, $\underline{u}$ on $[0,T]$ so that the solution $z = z(x,t)$ of the mixed boundary problem (1.1) to (1.3) exists and satisfies (1.4). The analogous boundary control problem for the 2 by 2 linear system reformulating the wave equation $w_{tt} = c(x)w_{xx}$ has been solved by Russell in [12]; results on the boundary controllability of linear equations are also given in [1] for the simplest wave equation and in [5] for general equations including the case of many space variables. See also [9], [3]. In [12] the construction of the controls makes explicit use of the characteristics of the equation, and the controls are eventually obtained by solving appropriate characteristic initial value (or Goursat) problems. Such a method, however, does not appear to be adequate if the number of characteristic fields is not small and the equation is not linear. On the other hand, as it will be seen, these difficulties disappear if the construction of the controls is carried out without ever considering Goursat problems. Indeed if A does not depend on z, the existence of the boundary control is an almost immediate consequence of the standard existence and uniqueness theorem for mixed boundary problems, provided that (1.1) is hyperbolic and $A(x,t)$ invertible.

An important difference between linear and nonlinear differential equations is that conditions insuring the existence of a local solution are not sufficient, in

* Received by the editors November 14, 1968, and in revised form February 10, 1969.

† Department of Mathematics, New York University, New York, New York 10012. This research was supported in part by the National Science Foundation under Grant GP6070 and in part by the United States Atomic Energy Commission under Contract AT(30-1)-1480.

general, to guarantee the existence of a solution of the latter equations on a set of preassigned size. On the other hand, in a hyperbolic initial value problem the data determine the solution completely on a certain domain. Hence, in the nonlinear case, for the existence of boundary controls it is necessary that the solution of the relevant hyperbolic problem can be extended to sets of given size. Some such extensions have been studied in [2] and the conditions found there, strengthened so as to make the time-like and the space-like variables interchangeable, will be seen to be sufficient for the existence of boundary controls in the nonlinear case. It will be also proved that, if $T > 0$ is not too small, in an appropriate Banach space there is an open set of initial data which can be brought to zero in time T.

2. Definitions, zero controllability. Let $m > 0, n > 0$ be integers; $R, R_m, R_{m \times n}$ are, respectively, the real numbers, the m-dimensional Euclidean space and the space of real matrices with m rows and n columns.

2.1. Norms. Throughout this paper, $| \cdot |$ denotes sup norms; so $|h|$ = absolute value of h if $h \in R$, $|h| = \max\{|h_i| : i = 1, \cdots, m\}$ if $h = (h_i) \in R_m$, $|h| = \max\{\sum_{j=1}^{n} |h_{ij}| : i = 1, \cdots, m\}$ if $h = (h_{ij}) \in R_{m \times n}$ and $|h| = \sup\{|h(x)| : x \in X\}$ if h is a function defined on a set X and taking values in either R, R_m or $R_{m \times n}$.

From now on $\bar{m}, \underline{m}$ denote positive integers. Fix $m = \bar{m} + \underline{m}$; if $h = (h_i) \in R_m$, $\bar{h}$ and $\underline{h}$ are the points of $R_{\bar{m}}, R_{\underline{m}}$ whose components are defined by

$$\bar{h}_i = h_i, \quad i = 1, \cdots, \bar{m}; \qquad \underline{h}_i = h_{i+m}, \quad i = 1, \cdots, \underline{m};$$

analogously if $h = (h_{ij}) \in R_{m \times m}$, $\bar{h}$ is the submatrix of h formed by the first $\bar{m}$ rows of $\underline{h}$, and $\underline{h}$ that formed by the last $\underline{m}$ rows; if h is a function taking values in R_m or $R_{m \times m}$, $\bar{h}$ and $\underline{h}$ are defined similarly. If α is a positive real number and $\mathfrak{R} \subset R_2$, $\mathfrak{R}_\alpha$ is the set defined by

$$\mathfrak{R}_\alpha = \{(x, t, w) : (x, t) \in \mathfrak{R}, w \in R_m, |w| \leqq \alpha\};$$

if $a, T \in (0, \infty]$, $\mathfrak{R} = \mathfrak{R}(a, T)$ is the set

$$\mathfrak{R} = \{(x, t) : 0 \leqq x \leqq a, x \neq \infty, 0 \leqq t \leqq T, t \neq \infty\};$$

so if a, T are real, $\mathfrak{R}$ and $\mathfrak{R}_\alpha$ are compact. If $a, T \in (0, \infty)$, $\tau = \tau(a, T)$ is the triangle

$$\tau = \left\{(x, t) : 0 \leqq t \leqq T, 0 \leqq x \leqq a - \frac{a}{T} t\right\}.$$

2.2. The class $\Sigma(\mathfrak{R}, m, \alpha)$. Fix $m = \bar{m} + \underline{m}$, $\alpha > 0$ real, $\mathfrak{R} \subset R_2$. Write $A \in \Sigma(\mathfrak{R}, m, \alpha)$ if and only if the following holds:

$A = A(x, t, w)$ is a C^1 function from $\mathfrak{R}_\alpha$ to $R_{m \times m}$, in short $A \in C^1(\mathfrak{R}_\alpha, R_{m \times m})$, and there is $S \in C^1(\mathfrak{R}_\alpha, R_{m \times m})$ such that for some $\delta > 0$

(i) $|\det S(x, t, w)| > \delta$, all (x, t, w);

(i′) $D(x, t, w) = S(x, t, w)A(x, t, w)S^{-1}(x, t, w)$ is a diagonal matrix for every (x, t, w);

(ii) the diagonal elements d_i of D satisfy $d_i(x, t, w) > \delta$, $i = 1, \cdots, \bar{m}$, $d_i(x, t, w) < -\delta$, $i = \bar{m} + 1, \cdots, m$, for all (x, t, w);

(ii′) the submatrices of $S(x, t, w)$ defined by

$$\bar{S}(x,t,w) = [\nabla_1(x,t,w), \nabla_2(x,t,w)], \quad \nabla_1(x,t,w) \in R_{\bar{m}\times\bar{m}},$$
$$\underline{S}(x,t,w) = [\nabla_1(x,t,w), \Delta_2(x,t,w)], \quad \Delta_2(x,t,w) \in R_{\underline{m}\times\underline{m}}$$

satisfy $|\det \nabla_1(x,t,w)| > \delta$, $|\det \Delta_2(x,t,w)| > \delta$ for all (x,t,w);

(iii) S, D (hence, by (i), S^{-1} and A) are bounded together with all their first partial derivatives.

In (iii) and throughout this paper S^{-1} is the map $(x, t, w) \to (S(x, t, w))^{-1}$, where the last object is the inverse of $S(x, t, w)$ in $R_{m\times m}$; A^{-1} and D^{-1} are defined analogously. Note that (i), (i′) amount to the definition of "the system (1.1) is hyperbolic on $\mathfrak{R}_\alpha$" and (ii), (ii′) are usual conditions in dealing with hyperbolic mixed boundary problems; also if $\mathfrak{R}$ is compact, (iii) is redundant and it suffices that (i), (ii) and (ii′) hold for $\delta = 0$.

2.3. The class C¹(*X*, *Y*). We write $z \in C^1(X, Y)$ as an abbreviation of "z is a continuously differentiable Y-valued function on X"; in absence of ambiguity the range space Y will be omitted. If X is an interval, $C^1(X, R_m)$ is given the following more special meaning. Let $I \subset R$ be a compact interval; $C^1(I, R_m)$ is the set of R_m-valued continuous functions ϕ on I possessing a continuous derivative on I, normed by

$$\|\phi\| = |\phi| + |\phi'|,$$

where ϕ' is the derivative of ϕ. $C_0^1 = C_0^1([0, 1], R_m)$ is the subspace of $C^1 = C^1([0, 1], R_m)$ defined by

$$C_0^1 = \{\phi : \phi \in C^1, \phi(0) = \phi'(0) = 0\}.$$

So C^1, C_0^1 are both Banach spaces.

2.4. Zero controllability. Put $\mathfrak{R} = \mathfrak{R}\ (1, \infty)$; suppose $A = A(x, t, w)$ and $f = f(x, t, w)$ are functions defined on $\mathfrak{R}_\alpha$, A is $R_{m\times m}$-valued, f is R_m-valued and $m = \bar{m} + \underline{m}$. We say that the system (1) is *zero controllable with one boundary control* if there is an open set $\Omega \subset C_0^1([0, 1], R_m)$ such that for each $\phi \in \Omega$ there exist a real number $T > 0$ and a function $\underline{u} = \underline{u}(t)$ from $[0, T]$ to R_m so that the solution $z = z(x, t)$ of (1.1), (1.2), (1.3), with $\underline{u} = 0$, exists and satisfies (1.4).

Analogously, we say that (1.1) is *zero controllable with two boundary controls* if there is an open set $\Omega \subset C^1([0, 1], R_m)$ such that for each $\phi \in \Omega$ there exist a real number $T > 0$ and functions $\bar{u} = \bar{u}(t)$, $\underline{u} = \underline{u}(t)$ from $[0, T]$ to $R_{\bar{m}}$ and $R_{\underline{m}}$, respectively, so that the solution of (1.1) to (1.3) exists and satisfies (1.4).

3. The main result. If A is of class Σ then (3.1) is zero controllable with one boundary control. This assertion is a particular case of our main result, Theorem 3.1, which will be seen to follow mainly from the fact that if $A \in \Sigma$ then the solution of (3.1) to (3.3) below exists on a preassigned rectangle whenever the data $\phi, \bar{u}, \underline{u}$ are conveniently restricted.

For $m = \bar{m} + \underline{m}$ consider the following mixed boundary problem:

$$z_t + A(x, t, z)z_x = 0, \qquad (x, t) \in \mathfrak{R}(a, T), \tag{3.1}$$

$$z(x, 0) = \phi(x), \qquad x \in [0, a], \tag{3.2}$$

$$\bar{z}(0, t) = \bar{u}(t), \quad \underline{z}(a, t) = \underline{u}(t), \qquad t \in [0, T], \tag{3.3}$$

where

$$\phi \in C^1([0, a], R_m), \quad \bar{u} \in C^1([0, T], R_{\bar{m}}), \quad \underline{u} \in C^1([0, T], R_m) \tag{3.4}$$

and $\phi, \bar{u}, \underline{u}$ satisfy the compatibility conditions

(i) $\bar{u}(0) = \bar{\phi}(0), \quad \bar{u}'(0) + \bar{A}(0, 0, \phi(0))\phi'(0) = 0,$

(ii) $\underline{u}(0) = \underline{\phi}(a), \quad \underline{u}'(0) = \underline{A}(a, 0, \phi(a))\phi'(a) = 0.$

Remark 3.1. Suppose z satisfies (3.1), $0 \leqq t < T$, and for $i = 1, \cdots, \bar{m}$ let $\xi_i = \xi_i(s)$ be defined by

$$\frac{d}{ds}\xi_i(s) = d_i(\xi_i(s), s, z(\xi_i(s), s)), \qquad s \geqq t, \quad \xi_i(t) = 0,$$

where d_i is the ith diagonal element of $D = SAS^{-1}$. If $d_i > 0$, then at $s = t$ the curve $(\xi_i(s), s), s \geqq t$, called the ith characteristic of (3.1) through $(0, t)$, enters the rectangle $\mathfrak{R} = \mathfrak{R}(a, T)$. So if $A \in \bar{\Sigma}(\mathfrak{R}, m, \alpha)$, the first of (3.3) amounts to fixing on the boundary $x = 0$ of $\mathfrak{R}$ exactly those components of z which correspond to characteristics entering $\mathfrak{R}$ there.

The following continuation result is known (see [2, Theorems 5.III and 5.I]).

LEMMA 3.1. *Fix* $0 < c_0 < \alpha, 0 < T < \infty$ *all real,* $m = \bar{m} + \underline{m}, 0 < \varepsilon < b \leqq \infty$, $\mathfrak{R} = \mathfrak{R}(b, \infty)$ *and* $A = S^{-1}DS \in \bar{\Sigma}(\mathfrak{R}, m, \alpha)$. *Conclusion: there are real numbers* $c > 0, N > 0$ *such that if* $a \in R \cap [\varepsilon, b]$, $\phi, \bar{u}, \underline{u}$ *satisfy* (3.4), $|\phi| \leqq c_0$ *and* $\max(|\phi'|, |\bar{u}'|, |\underline{u}'|) \leqq c$, *then on* $\mathfrak{R}(a, T)$ *there is a (unique) function* z *of class* C^1 *which satisfies* (3.1) *to* (3.3), *and* $|z_x| \leqq N(|\phi'| + |\bar{u}'| + |\underline{u}'|)$; *moreover for* $0 < T_1 \leqq \min(T, a/|\underline{D}|)$ *the restriction of* z *to the triangle* $\tau(a, T_1)$ *is independent of the choice of* u.

Lemma 3.1 is the main tool for proving the following theorem.

THEOREM 3.1. *Put* $\mathfrak{R} = \mathfrak{R}(1, \infty)$; *fix* $m = \bar{m} + \underline{m}$, $A = S^{-1}DS \in \bar{\Sigma}(\mathfrak{R}, m, \alpha)$, $0 < c_0 < \alpha$ *and* $\bar{u} \in C^1([0, \infty], R_{\bar{m}})$ *with bounded support. Conclusion: there is a real* $c > 0$ *such that if* $\phi \in C^1([0, 1], R_m)$, ϕ, $\bar{u}$ *satisfy* (3.4)(i), $|\phi| \leqq c_0$ *and* $\max(|\phi'|, |\bar{u}'|) \leqq c$, *then there exist* $0 < T < \infty$, $u \in C^1([0, T], R_{\underline{m}})$ *so that the solution* $z = z(x, t)$ *of* (3.1) *to* (3.3) *with* $a = 1$ *exists in* $C^1(\mathfrak{R}(1, T), R_m)$, *is unique there and, moreover, satisfies*

$$z(x, T) = 0, \quad \textit{all } x \in [0, 1]; \quad |z| \leqq \min(\alpha, 2c_0). \tag{3.5}$$

Proof. Fix c_1, c_2, t_0 real so that

$$c_0 < c_1 < c_2 < \min(\alpha, 2c_0); \quad \bar{u}(t) = 0 \quad \text{all} \quad t \geqq t_0 \geqq 0; \tag{3.6}$$

and define

$$T_0 = |D|^{-1}, \qquad T_1 = |\underline{(D^{-1})}|; \tag{3.7}$$

note that T_0 and T_1 are real, positive and $T_0 \leqq T_1$.

For each real $\delta > 0$, fix a real number $\Delta = \Delta(\delta)$ such that if h satisfies

$$h \in C^1([0, T_1], R_{\bar{m}}), \quad |h(T_1)| \leqq c_1, \quad |h'(T_1)| \leqq \delta, \tag{3.8}$$

then h has a C^1 extension H to $[0, \infty)$ satisfying

$$|H| \leqq c_1 + \delta, \quad |H'| \leqq \delta, \quad H(t) = 0, \quad \text{all} \quad t \geqq T_1 + \Delta.$$

(α3.1) Consider the mixed boundary problem

$$z_x + A^{-1}(x, t, z) z_t = 0, \qquad (x, t) \in \mathfrak{R}(1, T), \tag{3.9}$$

$$z(0, t) = \psi(t), \qquad t \in [0, T], \tag{3.10}$$

$$\bar{z}(x, 0) = \bar{\phi}(x), \qquad \underline{z}(x, T) = 0, \qquad x \in [0, 1]. \tag{3.11}$$

As it is easily checked $A^{-1} \in \bar{\Sigma}(\mathfrak{R}, m, \alpha)$; hence by Lemma 3.1 there is δ_2,

$$0 < \delta_2 \leqq \min\left(\frac{c_1 - c_0}{T_1}, c_2 - c_1\right), \tag{3.12}$$

such that if

$$\begin{gathered} T_0 \leqq T < \infty, \quad \psi \in C^1([0, T], R_m), \quad \psi(T) = \psi'(T) = 0, \\ \bar{\psi}(0) = \bar{\phi}(0), \quad \bar{\phi}'(0) + \overline{A^{-1}}(0, 0, \psi(0))\psi'(0) = 0, \\ |\psi| \leqq c_2 \quad \text{and} \quad \max(|\psi'|, |\bar{\phi}'|) \leqq \delta_2, \end{gathered} \tag{3.13}$$

there is a (unique) $z \in C^1(\mathfrak{R}(1, T))$ satisfying (3.9) to (3.11) and

$$|z| \leqq \min(\alpha, 2c_0). \tag{3.14}$$

(α3.2) For $\underline{v} = \underline{v}(t)$ satisfying

$$\underline{v} \in C^1([0, T_1], R_{\underline{m}}), \quad \underline{v}(0) = \underline{\phi}(1), \quad v'(0) + \underline{A}(1, 0, \phi(1))\phi'(1) = 0 \tag{3.15}$$

consider the mixed boundary problem

$$w_t + A(x, t, w) w_x = 0, \qquad (x, t) \in \mathfrak{R}(1, T_1), \tag{3.16}$$

$$w(x, 0) = \phi(x), \qquad x \in [0, 1], \tag{3.17}$$

$$\bar{w}(0, t) = \bar{u}(t), \quad \underline{w}(1, t) = \underline{v}(t), \qquad t \in [0, T_1]. \tag{3.18}$$

Since $A \in \bar{\Sigma}(\mathfrak{R}, m, \alpha)$, Lemma 3.1 implies that there is $\delta_1 > 0$ such that, whenever $\phi, \bar{u}, \underline{v}$ satisfy (3.4) for $a = 1$, $T = T_1$, (3.15), $|\phi| \leqq c_0$ and $\max(|\phi'|, |\bar{u}'|, |\underline{v}'|) \leqq \delta_1$, there is a function $w \in C^1(\mathfrak{R}(1, T_1), R_m)$ satisfying (3.16), (3.17), (3.18) and

$$|w_t| \leqq \delta_2, \qquad |w| \leqq c_1. \tag{3.19}$$

Hence there is c, $0 < c \leqq \delta_2$, so that if $\phi, \bar{u}$ are as in the hypotheses of the theorem, there is $\underline{v}$ satisfying (3.15) and a (unique) $w \in C^1(\mathfrak{R}(1, T_1), R_m)$ satisfying (3.16) to

(3.19). Indeed c can be taken to be any number satisfying

$$0 < c \leqq \min(\delta_2, \delta_1), \quad |A|c \leqq \delta_1,$$

and $\underline{v}$ any function satisfying (3.15) and $|\underline{v}'| \leqq c$, for instance,

$$\underline{v}(t) = \underline{\phi}(1) + t\underline{A}(1, 0, \phi(1))\phi'(1).$$

It will now be proved that for this c the conclusion of the theorem holds. To this end, let ϕ, $\bar{u}$ be as in the hypotheses; fix $\underline{v}$ as said in (α3.2) and let w be the function satisfying (3.16) to (3.19). Define

$$T_2 = T_1 + \Delta(\delta_2), \quad h(t) = w(0, t), \qquad t \in [0, T_1];$$

in view of (3.12) and the definition of $\Delta(\delta_2)$, h has a C^1 extension H to $[0, \infty]$ with $|H| \leqq c_2$, $|H'| \leqq \delta_2$ and $H(t) = 0$ for all $t \geqq T_2$. Let

$$T'_2 = \max(T_2, t_0), \qquad T = T'_2 + |\overline{(D^{-1})}|,$$

and define the function $\psi = \psi(t)$ by

$$\bar{\psi}(t) = \bar{u}(t), \qquad \underline{\psi}(t) = H(t), \qquad t \in [0, T];$$

then $\psi(t) = 0$ for $T'_2 \leqq t \leqq T$ and ψ satisfies (3.13). So let $z^* = z^*(x, t)$ be the only function in $C^1(\mathfrak{R}(1, T), R_m)$ satisfying (3.9), (3.10), (3.11) and (3.14). It will now be shown that z^* satisfies also

$$z^*(x, T) = 0, \qquad \text{all } x \in [0, 1], \tag{3.20}$$

$$z^*(x, 0) = \phi(x), \qquad \text{all } x \in [0, 1]. \tag{3.21}$$

To this end let $\tau_0 \subset \mathfrak{R}(1, T)$ be the triangle

$$\tau_0 = \{(x, t) : 0 \leqq x \leqq 1,\ T'_2 + (T'_2 - T)x \leqq t \leqq T\}$$

and consider the mixed boundary problem

$$z_x + A^{-1}(x, t, z)z_t = 0, \tag{3.22}$$

$$z(0, t) = 0, \qquad t \in [T'_2, T], \tag{3.23}$$

$$\underline{z}(x, T) = 0, \qquad x \in [0, 1]; \tag{3.24}$$

the zero function on τ_0 and the restriction of z^* to τ_0 both satisfy (3.22) on τ_0, (3.23) and (3.24); also since $T - T'_2 = |\overline{(D^{-1})}|$, the last assertion in Lemma 3.1 implies that on τ_0 there is at most one function in $C^1(\tau_0, R_m)$ which satisfies (3.22) to (3.23); hence (3.20) holds since

$$z^*(x, t) = 0, \qquad \text{all } (x, t) \in \tau_0.$$

Analogously, let

$$\tau = \{(x, t) : 0 \leqq x \leqq 1, 0 \leqq t \leqq T_1 - T_1 x\}$$

and consider the mixed boundary problem

$$z_t + A^{-1}(x,t,z)z_x = 0 \tag{3.25}$$

$$z(0,t) = \psi(t), \qquad t \in [0, T_1], \tag{3.26}$$

$$\bar{z}(x,0) = \bar{\phi}(x), \qquad x \in [0,1]. \tag{3.27}$$

In view of the definitions of ψ, w, z^*, it is easily seen that the restrictions $w|_\tau$ and $z^*|_\tau$ both belong to $C^1(\tau, R_m)$ and satisfy (3.25) on τ, (3.26) and (3.27); since $T_1 = |\underline{(D^{-1})}|$, on τ uniqueness prevails, and hence

$$w|_\tau = z^*|_\tau;$$

this proves (3.21).

Define

$$\underline{u}(t) = \underline{z}^*(1,t), \qquad t \in [0,T];$$

then $\underline{u} \in C^1([0,T], R_{\underline{m}})$. Since $z = z^*$ satisfies (3.9) to (3.11), (3.14), (3.20) and (3.21) from the definition of ψ and $\underline{u}$, it follows that z^* is the solution of (3.1) to (3.3) with $a = 1$ and satisfies (3.5). This completes the proof.

Let us note that if A does not depend on z, Theorem 3.1 holds in a stronger form since no assumption on the smallness of the data need be made (see the ending of § 5).

4. Controllability of $z_t + A(x,t,z)z_x = 0$. Let us first note that the existence proof of the boundary control $\underline{u}$, as given in Theorem 3.1, is constructive and, as it will be indicated later in this section, it is well suited as a basis for the numerical computation of such control. Some consequences of the main result will now be made explicit. The special case of Theorem 3.1 for $\bar{u} = 0$ follows.

THEOREM 4.1. *Suppose $\mathfrak{R} = \mathfrak{R}(1,\infty)$, $m = \bar{m} + \underline{m}$ and $A \in \bar{\Sigma}(\mathfrak{R}, m, \alpha)$. Then $z_t + A(x,t,z)z_x = 0$ is zero controllable with one boundary control.*

Fix m and c_0 as in Theorem 3.1 and let $c > 0$ be as given there for $\bar{u} = 0$; put $C^1 = C^1([0,1], R_m)$ and define $\Omega_0 = \Omega_0(c_0, c)$ by

$$\Omega_0 = \{\phi : \phi \in C^1, \bar{\phi}(0) = 0, |\phi| \leqq c_0, |\phi'| \leqq c\}.$$

Then Ω_0 contains nontrivial open sets of $C_0^1([0,1], R_m)$, and from the proof of Theorem 3.1 it is clear that the real number $T > 0$ produced there does not depend on the choice of ϕ in Ω_0, i.e., any ϕ in Ω can be brought to zero in time T; furthermore T cannot be too small. This is formalized in the following corollary.

COROLLARY 4.1. *Suppose the hypotheses of Theorem 4.1 hold and fix $0 < c_0 < \alpha$. Then*

(i) *there are real numbers $c > 0$ and $T > 0$ such that if $\phi \in \Omega_0(c_0, c)$ there exist $\underline{u} \in C^1([0,T], R_{\underline{m}})$ and $z \in C^1(\mathfrak{R}(1,T), R_m)$ satisfying*

$$z_t + A(x,t,z)z_x = 0, \qquad (x,t) \in \mathfrak{R}(1,T), \tag{a}$$

$$z(x,0) = \phi(x), \qquad x \in [0,1], \tag{b}$$

$$\bar{z}(0,t) = 0, \quad \underline{z}(1,t) = \underline{u}(t), \qquad t \in [0,T], \tag{c}$$

$$z(x,T) = 0, \qquad x \in [0,1]; \tag{d}$$

(ii) *if* c, T *is any such pair, then* $T \geqq |D|^{-1}$;

(iii) *if* c_0 *is sufficiently small, there are* c *and* T *having the properties in* (i) *and, in addition,*

$$T \leqq |\underline{(D^{-1})}| + 1 + |\overline{(D^{-1})}|.$$

Proof. (i) has already been seen; (iii) follows immediately from the proof of Theorem 3.1 for $\bar{u} = 0$; and to establish (ii) it suffices to notice that in $\Omega_0(c_0, c)$ there are initial data (for instance $\bar{\phi} = 0$, $\underline{\phi} = c_0/2$) for which on the triangle $\tau(1, |\underline{D}|^{-1})$ the solution of (a), (b) and the first part of (c) is a nonzero constant.

Remark 4.1. For a given $\phi \in \Omega(c_0, c)$ the control function $\underline{u}$ is by no means unique. This is due to the fact that in the construction of $\underline{u}$, see proof of Theorem 3.1, one can choose $\underline{v}$ among infinitely many functions and extend $h(t) = \underline{w}(0, t)$ in infinitely many ways. For instance, it is easily seen that h can be usefully extended by using any function in some closed convex set contained in $C^1([T_1, T_2], R_{\underline{m}})$ and containing more than one element, hence infinitely many; also in the proof of Theorem 3.1 it is shown that to each such extension H of h there correspond $\underline{u} \in C^1([0, T], R_{\underline{m}})$ such that (a) to (d) in Corollary 4.1 hold; on the other hand from the uniqueness of solution of the mixed boundary problem (a) to (c) in Corollary 4.1 it follows that the map $H \to \underline{u}$ is one to one; whence there are infinitely many $\underline{u}$ which bring the given ϕ to zero in finite time.

It will now be shown that the hyperbolic system studied so far is also zero controllable with two boundary controls. This is a consequence of Lemma 3.1 and the proof of Theorem 3.1. For $c_0 > 0$, $c > 0$ real and $C^1 = C^1([0, 1], R_m)$ define $\Omega = \Omega(c_0, c)$, a subset of C^1 with nontrivial interior, by

$$\Omega = \{\phi : \phi \in C^1, |\phi| \leqq c_0, |\phi'| \leqq c\}.$$

THEOREM 4.2. *Put* $\mathfrak{R} = \mathfrak{R}(1, \infty)$, *fix* $m = \bar{m} + \underline{m}$, $0 < c_0 < \alpha < \infty$ *and suppose* $A = S^{-1}DS \in \bar{\bar{\Sigma}}(\mathfrak{R}, m, \alpha)$. *Conclusion: there is a real* $c > 0$ *such that if* $\phi \in \Omega(c_0, c)$ *there are* $0 < T < \infty$, $\bar{u} \in C^1([0, T], R_{\bar{m}})$, $\underline{u} \in C^1([0, T], R_{\underline{m}})$ *so that the solution* $z = z(x, t)$ *of* (3.1) *to* (3.3) *with* $a = 1$ *exists in* $C^1(\mathfrak{R}(1, T), R_m)$ *and satisfies*

$$z(x, T) = 0, \quad \textit{all} \quad x \in [0, 1]; \qquad |z| \leqq \min(\alpha, 2c_0). \tag{4.1}$$

Thus $z_t + A(x, t, z)z_x = 0$ *is zero controllable with two boundary controls.*

Proof. Fix $c_0 < c_1 < c_2 < \min(\alpha, 2c_0)$ and define

$$T_0 = \frac{|D|^{-1}}{2}, \qquad T_1 = \frac{|D^{-1}|}{2}, \tag{4.2}$$

where, it is recalled, D^{-1} is the map $(x, t, w) \to (D(x, t, v))^{-1}$; so T_0, T_1 are real and $T_0 \leqq T_1$. For each real $\delta > 0$, fix a real number $\Delta = \Delta(\delta)$ such that, if h satisfies

$$h \in C^1([0, T_1], R_m), \quad |h| \leqq c_1, \quad |h'| \leqq \delta, \tag{4.3}$$

then h has a C^1 extension H to $[0, \infty]$ satisfying

$$|H| \leqq c_1 + \delta, \quad |H| \leqq \delta, \quad H(t) = 0, \qquad \text{all } t \geqq T_1 + \Delta.$$

(α4.1) Consider the pair of mixed boundary problems

$$z_x + A^{-1}(x,t,z)z_t = 0, \quad (x,t) \in \mathfrak{R}^{-1} = [0,\tfrac{1}{2}] \times [0,T], \tag{4.4}$$

$$z(\tfrac{1}{2},t) = \psi(t), \qquad t \in [0,T], \tag{4.5}$$

$$\underline{z}(x,0) = \phi(x), \quad \bar{z}(x,T) = 0, \qquad x \in [0,\tfrac{1}{2}]; \tag{4.6}$$

$$z_x + A^{-1}(x,t,z)z_t = 0, \quad (x,t) \in \mathfrak{R}^{+} = [\tfrac{1}{2},1] \times [0,T], \tag{4.4'}$$

$$z(\tfrac{1}{2},t) = \psi(t), \qquad t \in [0,T], \tag{4.5'}$$

$$\bar{z}(x,0) = \bar{\phi}(x), \qquad \underline{z}(x,T) = 0, \qquad x \in [\tfrac{1}{2},1]. \tag{4.6'}$$

Since $A^{-1} \in \bar{\Sigma}(\mathfrak{R}, m, \alpha)$, Lemma 3.1 implies that there is a δ_2,

$$0 < \delta_2 \leqq \min\left(\frac{c_1 - c_0}{T_1}, c_2 - c_1\right), \tag{4.7}$$

such that if

$$\begin{aligned} &T \in [T_0, \infty), \quad \psi \in C^1([0,T], R_m), \quad \phi \in C^1([0,1], R_m), \\ &\psi \text{ and } \phi \text{ satisfy the compatibility conditions } \psi(T) = \psi'(T) = 0, \\ &\psi(0) = \phi(\tfrac{1}{2}), \quad \phi'(\tfrac{1}{2}) + A^{-1}(\tfrac{1}{2}, 0, \phi(\tfrac{1}{2}))\psi'(0) = 0, \quad \text{and} \\ &|\psi| \leqq c_2, \quad \max(|\psi'|, |\phi'|) \leqq \delta_2, \end{aligned} \tag{4.8}$$

then there is a unique pair of functions $z_- \in C^1(\mathfrak{R}^-, R_m)$ satisfying (4.4) to (4.6), $z_+ \in C^1(\mathfrak{R}^+, R_m)$ satisfying (4.4′) to (4.6′) and, moreover,

$$|z_-|, |z_+| \leqq \min(\alpha, 2c_0). \tag{4.9}$$

(α4.2) For $\bar{v}$, $\underline{v}$ satisfying

$$\begin{aligned} &\bar{v} \in C^1([0,T_1], R_{\bar{m}}), \quad \bar{v}(0) = \bar{\phi}(0), \quad \bar{v}'(0) + \bar{A}(0,0,\phi(0))\phi'(0) = 0, \\ &\underline{v} \in C^1([0,T_1], R_{\underline{m}}), \quad \underline{v}(0) = \underline{\phi}(1), \quad \underline{v}'(0) + \underline{A}(1,0,\phi(1))\phi'(1) = 0, \end{aligned} \tag{4.10}$$

consider the mixed boundary problem

$$w_t + A(x,t,z)w_x = 0, \qquad (x,t) \in \mathfrak{R}(1, Y_1), \tag{4.11}$$

$$w(x,0) = \phi(x), \qquad x \in [0,1], \tag{4.12}$$

$$\bar{w}(0,t) = \bar{v}(t), \quad \underline{w}(1,t) = \underline{v}(t), \qquad t \in [0,T_1]. \tag{4.13}$$

since $A \in \bar{\Sigma}(\mathfrak{R}, m, \alpha)$, Lemma 3.1, in view of the reasoning made in (α3.2), implies that there is a c, $0 < c < \delta_2$, such that if $\phi \in \Omega(c_0, c)$ there exist $\bar{v}$, $\underline{v}$ satisfying (4.10) and $w \in C^1(\mathfrak{R}(1, T_1), R_m)$ satisfying (4.11), (4.12), (4.13) and

$$|w_t| \leqq \delta_2, \quad |w| \leqq c_1. \tag{4.14}$$

To see that for this c the conclusion of the theorem holds, let $\phi \in \Omega(c_0, c)$, fix $\bar{v}$, $\underline{v}$ so that what has been said in (α4.2) holds, and let w be the function satisfying (4.11) to (4.14). Define

$$T_2 = T_1 + \Delta(\delta_2), \quad T = T_2 + \frac{|D^{-1}|}{2},$$

$$h(t) = w(\tfrac{1}{2}, t), \qquad t \in [0, T_1].$$

Since h satisfies (4.3), from the definition of $\Delta(\delta_2)$ and (4.7) it follows that we can fix a function $\psi \in C^1([0, T], R_m]$ which extends h and satisfies $|\psi| \leqq c_2$, $|\psi'| \leqq \delta_2$ and $\psi(t) = 0$ if $t \in [T_2, T]$. So T, ψ, ϕ satisfy (4.8); let z_-, z_+ be the solutions of (4.4) to 4.6) and (4.4′) to (4.6′) respectively. Define

$$\bar{u}(t) = \bar{z}_-(0, t), \quad \underline{u} = \underline{z}_+(0, t), \qquad t \in [0, T];$$

then $\bar{u} \in C^1([0, T], R_{\bar{m}})$ and $\underline{u} \in C^1([0, T], R_{\underline{m}})$. By using the same uniqueness arguments already used in the proof of Theorem 3.1, it follows that z_- is the solution of

$$z_t + A(x, t, z)z_x = 0, \qquad (x, t) \in \mathfrak{R}^-,$$
$$z(x, 0) = \phi(x), \qquad x \in [0, \tfrac{1}{2}],$$
$$\bar{z}(0, t) = \bar{u}(t), \quad \underline{z}(\tfrac{1}{2}, t) = \underline{\psi}(t), \qquad t \in [0, T],$$

and satisfies

$$z_-(x, T) = 0, \quad x \in [0, \tfrac{1}{2}], \quad |z_-| \leqq \min(\alpha, 2c_0);$$

analogously, z_+ is the solution of

$$z_t + A(x, t, z)z_x = 0, \qquad (x, t) \in \mathfrak{R}^+,$$
$$z(x, 0) = \phi(x), \qquad x \in [\tfrac{1}{2}, 1],$$
$$\bar{z}(\tfrac{1}{2}, t) = \bar{\psi}(t), \quad \underline{z}(1, t) = \underline{u}(t), \qquad t \in [0, T],$$

and satisfies

$$z_+(x, T) = 0, \quad x \in [\tfrac{1}{2}, 1], \qquad |z_+| \leqq \min(\alpha, 2c_0).$$

Define z to be z_- on $\mathfrak{R}^-$, z_+ on $\mathfrak{R}^+$; then $z \in C^1(\mathfrak{R}(1, T), R_m)$, and a moment of reflection shows that z is the solution of (3.1) to (3.3) and satisfies (4.1). The theorem is thus established.

The two boundary controls $\bar{u}$ and $\underline{u}$ are not unique; this depends, as before, on the fact that there are many useful choices of $\bar{v}$ and $\underline{v}$ and many useful extensions of $w(\tfrac{1}{2}, \cdot)$. Incidentally this lack of uniqueness is most interesting since it leaves open the possibility of choosing the boundary controls $\bar{u}$ and $\underline{u}$ so as to minimize T or, for fixed T, to minimize some functional of $\bar{u}$, $\underline{u}$ and z.

The next corollary is the analogue of Corollary 4.1 and follows immediately from Theorem 4.2. It asserts in particular that if the initial data ϕ have sufficiently small derivatives and the real number $T > 0$ is not too small then ϕ can be brought to zero in time T.

COROLLARY 4.2. *Let c_0 and A be as in the hypotheses of Theorem* 4.2. *Then*

(i) *there are real numbers $c > 0$, $T > 0$ such that for each $\phi \in \Omega(c_0, c)$ there exist $\bar{u} \in C^1([0, T], R_{\bar{m}})$, $\underline{u} \in C^1([0, T], R_{\underline{m}})$ and $z \in C^1(\mathfrak{R}(1, T), R_m)$ satisfying*

$$z_t + A(x, t, z)z_x = 0, \qquad (x, t) \in \mathfrak{R}(1, T),$$

$$z(x, 0) = \phi(x), \qquad x \in [0, 1],$$

$$\bar{z}(0, t) = \bar{u}(t), \quad \underline{z}(1, t) = \underline{u}(t), \qquad t \in [0, T],$$

$$z(x, T) = 0, \qquad x \in [0, 1];$$

(ii) *if c, T is any such pair, then $T \geqq |D|^{-1}/2$;*

(iii) *if c_0 is sufficiently small, there are c, T having the properties* (i) *and, in addition,*

$$T \leqq |D^{-1}| + 1.$$

As for the numerical determination of boundary controls it is useful to observe that the proofs of the existence Theorems 3.1 and 4.2 give a general method of computation. Indeed, in the case of one boundary control $\underline{u} = \underline{u}(t)$, the computation of $\underline{u}$ is reduced by Theorem 3.1 to the numerical solution of two mixed boundary problems, namely (3.16) to (3.18) and (3.9) to (3.11). Analogously, in the case of two boundary controls, computation of $\bar{u}$ and $\underline{u}$ is reduced by Theorem 4.2 to the numerical solution of three mixed boundary problems. Therefore any numerical scheme for solving hyperbolic mixed boundary problems, such as, for instance, those in [7], [8] and [13], gives a scheme for computing boundary controls.

5. Controllability of $z_t + A(x, t, z)z_x = f(t, z)$. It will be seen that sufficient conditions for the hyperbolic system

$$z_t + A(x, t, z)z_x = f(x, t, z)$$

to be zero controllable are the usual conditions on A,f for solving the mixed boundary problem, augmented by

$$(A^{-1}f)_t = 0, \quad f_x = 0, \quad \frac{|f(x, t, z)|}{|z|} \to 0 \quad \text{as} \quad z \to 0.$$

These additional requirements are used to guarantee that for some class of data the two relevant mixed problems analogous to (3.9) to (3.11) and (3.16) to (3.18) have solutions on preassigned rectangles. Since the system studied in § 4 satisfies the preceding additional conditions, the results in this section generalize those already obtained; however, in a sense, they are also more special because the set of initial data ϕ brought to zero in finite time will be smaller, for not only $|\phi'|$ but also $|\phi|$ will be required to be small.

For α and T, positive real, define

$$B_\alpha = \{w : w \in R_m, |w| \leqq \alpha\}.$$

Suppose

(a) $f = f(t, w)$ is a C^1 function from $[0, T] \times B_\alpha$ to R_m and for each $t \leqq [0, T]$, $|f(t, w)|/|w| \to 0$ as $w \to 0$; consider the mixed boundary problem

$$z_t + A(x, t, z)z_x = f(t, z), \qquad (x, t) \in \mathfrak{R}(a, T), \tag{5.1}$$

$$z(x, 0) = \phi(x), \qquad x \in [0, a], \tag{5.2}$$

$$\bar{z}(0, t) = \bar{u}(t), \quad \underline{z}(0, t) = \underline{u}(t), \qquad t \in [0, T], \tag{5.3}$$

where

$$\phi \in C^1([0, a], R_m), \quad \bar{u} \in C^1([0, T], R_{\bar{m}}), \quad \underline{u} \in C^1([0, T], R_{\underline{m}}),$$

$$\begin{aligned} &\text{(i) } \bar{u}(0) = \bar{\phi}(0), \quad \bar{u}'(0) + \bar{A}(0, 0, \phi(0))\phi'(0) = f(0, \phi(0)), \\ &\text{(ii) } \underline{u}(0) = \underline{\phi}(a), \quad \underline{u}'(0) + \underline{A}(a, 0, \phi(a))\phi'(a) = f(0, \phi(a)). \end{aligned} \tag{5.4}$$

The following analogue of Lemma 3.1 is known (see [2, Theorems 5.II and 5.I]).

LEMMA 5.1. *Fix* $m = \bar{m} + \underline{m}$, $0 < T < \infty$, $0 < b \leqq \infty$, $A = S^{-1}DS \in \bar{\Sigma}(\mathfrak{R}, m, \alpha)$, *where* $\mathfrak{R} = \mathfrak{R}(b, T)$ *with* f *satisfying* (a), $0 < \varepsilon < b$ *and* $N > 0$ *real. Conclusion: there are real numbers* $c_0 > 0, c > 0$ *such that if* $a \in R \cap [\varepsilon, b]$, $\phi, \bar{u}$ and $\underline{u}$ *satisfy* (5.4), $|\phi| \leqq c_0$ *and* $\max(|\phi'|, |\bar{u}'|, |\underline{u}'|) \leqq c$, *then there is a unique* $z \in C^1(\mathfrak{R}(a, T), R_m)$ *which satisfies* (5.1) *to* (5.3), *and, moreover,*

$$|z| \leqq 2c_0, \qquad |z_x| \leqq N;$$

also if $0 < T_1 \leqq \min(T, a/|\underline{D}|)$, *the restriction of* z *to the triangle* $\tau(a, T_1)$ *does not depend on the choice of* $\underline{u}$.

DEFINITION. Suppose $0 < T \leqq \infty$, $0 < b \leqq \infty$ and $\mathfrak{R} = \mathfrak{R}(b, T)$; write $(A, f) \in \tilde{\bar{\Sigma}}(\mathfrak{R}, m, \alpha)$ if and only if $A = A(x, t, w) \in \bar{\Sigma}(\mathfrak{R}, m, \alpha)$, $f = f(t, w)$ satisfies (a) with $[0, T]$ replaced by $[0, T] \cap R$, and $(A^{-1}f)_t = 0$.

Remark 5.1. If the partial derivative of $(A^{-1}f)$ with respect to t vanishes everywhere on $\mathfrak{R}_\alpha$, which is trivially true if A and f are independent of t, then

$$A^{-1}(x, t, w)f(t, w) = A^{-1}(x, \tilde{t}, w)f(\tilde{t}, w), \qquad \text{all } (x, t, w), (x, \tilde{t}, w) \in \mathfrak{R}_\alpha.$$

Hence if $(A, f) \in \tilde{\bar{\Sigma}}(\mathfrak{R}, m, \alpha)$, $A^{-1}f$ can be identified with the map $\tilde{f} = \tilde{f}(x, w)$ defined by

$$\tilde{f}(x, w) = A^{-1}(x, 0, w)f(0, w), \quad (x, w) \in ([0, b] \cap R) \times B_\alpha,$$

and $\tilde{f}$ satisfies the analogue of (a), i.e., $\tilde{f}$ is of class C^1 and for each $x \in [0, b] \cap R$, $|\tilde{f}(x, w)|/|w| \to 0$ as $w \to 0$. So, if $(A, f) \in \tilde{\bar{\Sigma}}(\mathfrak{R}, m, \alpha)$, A and f satisfy the hypotheses of Lemma 5.1 and A^{-1} and f satisfy the hypotheses of Lemma 5.1 with x playing the role of t; hence Lemma 5.1, rewritten with the obvious change in notation, holds for the mixed boundary problem

$$z_x + A^{-1}(x, t, z)z_t = f(x, z), \qquad (x, t) \in \mathfrak{R}(a, T),$$

$$z(0, t) = \psi(t), \qquad t \in [0, T],$$

$$\bar{z}(x, 0) = \bar{v}(x), \quad \underline{z}(x, T) = \underline{v}(x), \qquad x \in [0, a].$$

The next theorem is analogous to Theorem 3.1; it follows from Lemma 5.1

and the preceding remark in essentially the same way in which Theorem 3.1 follows from Lemma 3.1; its proof is omitted since it is very similar to that of Theorem 3.1.

THEOREM 5.1. *Put* $\mathfrak{R} = \mathfrak{R}(1, \infty)$; *fix* $m = \bar{m} + \underline{m}$, $(A, f) \in \tilde{\bar{\Sigma}}(\mathfrak{R}, m, \alpha)$ *and* $\bar{u} \in C^1([0, \infty], R_{\bar{m}})$ *with bounded support. Conclusion*: *there are* $c_0 > 0$, $c > 0$ *real such that if* $\phi \in C^1([0, 1], R_m)$, ϕ and $\bar{u}$ *satisfy* (5.4i), $\max(|\phi|, |u|) \leqq c_0$ and $\max(|\phi'|, |\bar{u}'|) \leqq c$, *then there exist* $0 < T < \infty$ *and* $\underline{u} \in C^1([0, T], R_{\bar{m}})$ *so that the solution* $z = z(x, t)$ *of* (5.1) *to* (5.3) *with* $a = 1$ *exists in* $C^1(\mathfrak{R}(1, T), R_m)$, *is unique there and, moreover, satisfies*

$$z(x, T) = 0, \quad all \quad x \in [0, 1]; \qquad |z| \leqq 2c_0.$$

By taking $\bar{u} = 0$ in Theorem 5.1 one obtains the following corollary.

COROLLARY 5.1. *Suppose* $\mathfrak{R} = \mathfrak{R}(1, \infty)$, $m = \bar{m} + \underline{m}$ *and* $(A, f) \in \tilde{\bar{\Sigma}}(\mathfrak{R}, m, \alpha)$. *Then* $z_t + A(x, t, z)z_x = f(t, z)$ *is zero controllable with one boundary control.*

The next theorem follows from Lemma 5.1 and Theorem 5.1; its proof is omitted because it can be obtained by making minor modifications in that of Theorem 4.2.

THEOREM 5.2. *Suppose* $\mathfrak{R} = \mathfrak{R}(1, \infty)$, $m = \bar{m} + \underline{m}$ *and* $(A, f) \in \tilde{\bar{\Sigma}}(\mathfrak{R}, m, \alpha)$. *Conclusion*: *there are* $c_0 > 0, c > 0$ *real such that if* $\phi \in \Omega(c_0, c)$ *there are* $0 < T < \infty$, $\bar{u} \in C^1([0, T], R_{\bar{m}})$, $\underline{u} \in C^1([0, T], R_{\underline{m}})$ *so that the solution* $z = z(x, t)$ *of* (5.1) *to* (5.3) *with* $a = 1$ *exists in* $C^1(\mathfrak{R}(1, T), R_m)$ *and satisfies* $z(x, t) = 0$ *all* $x \in [0, 1]$. *Thus* $z_t + A(x, t, z)z_x = f(t, z)$ *is zero controllable with two boundary controls.*

We shall now consider briefly the semilinear problem

$$z_t + A(x, t)z_x = f(x, t, z), \qquad (x, t) \in \mathfrak{R}(1, T), \tag{5.5}$$

$$z(x, 0) = \phi(x), \qquad x \in [0, 1], \tag{5.6}$$

$$\bar{z}(0, t) = \bar{u}(t), \qquad \underline{z}(1, t) = \underline{u}(t), \qquad t \in [0, T]. \tag{5.7}$$

Let $m = \bar{m} + \underline{m}$, $\mathfrak{R} = \mathfrak{R}(1, \infty)$ and suppose, in obvious notation, that $A = A(x, t)$ belongs to $\bar{\Sigma}(\mathfrak{R}, m)$, f is a C^1 function from $[0, 1] \times [0, \infty) \times R_m$ to R_m bounded as are its first derivatives, and ϕ, $\bar{u}$, $\underline{u}$ are given compatible C^1 functions. Under these hypotheses it is well known that (5.5) to (5.7) has a (unique) C^1 solution z, for any $T > 0$. By using this fact in the proof of Theorem 3.1 one obtains the following corollary

COROLLARY 5.2. *Suppose* A, f, $\bar{u}$ *are given as above and* $\bar{u}$ *has compact support. Then there are a real number* $T > 0$ *and, for each given* ϕ, *a function* $\underline{u} \in C^1([0, T], R_{\bar{m}})$ *so that the solution* z *of* (5.5) *to* (5.7) *exists and satisfies* $z(x, T) = 0$ *for all* $x \in [0, 1]$.

Therefore if $T > 0$ is not too small, any $\phi \in C_0^1([0, 1], R_m)$ can be brought to zero at time T with one boundary control and, a fortiori, any $\phi \in C^1([0, 1], R_m)$ in the case of two boundary controls.

6. Example: the wave equation. Consider the following boundary control problem: find $T > 0$ and real-valued functions

$$\underline{u} = \underline{u}(t) \text{ on } [0, T], \quad w = w(x, t) \text{ on } \mathfrak{R} = [0, 1] \times [0, T]$$

such that

$$(6.1)\qquad w_{tt} = c^2(u_x)w_{xx}, \qquad (x,t)\in \Re,$$

$$(6.2)\qquad w(x,0) = f(x), \quad w_t(x,0) = h(x), \qquad x\in[0,1],$$

$$(6.3)\qquad w(0,t) = 0, \quad w_x(1,t) = \underline{u}(t), \qquad t\in[0,T],$$

$$(6.4)\qquad w(x,T) = 0, \qquad x\in[0,1],$$

where c, f and h are given real-valued functions of a real variable and f and h satisfy appropriate compatibility conditions at $x = 0$.

If c is specialized to

$$(6.5)\qquad c(q) = \left[1 + E\left(1 - \frac{1}{\sqrt{1+q^2}}\right)\right]^{1/2}, \qquad q\in R,$$

where $E > 0$ is a certain constant (Young's modulus), it is shown in [6, Chap. 3] that a function w satisfying (6.1) describes the transverse planar vibration of an elastic string. If, moreover, $\underline{u}$ is given, then the initial value problem (6.1) to (6.3) can be thought of as approximating the transverse planar vibration of a string with given initial state f, h, one end clamped at $x = 0$, and the other end free to move at $x = 1$, along the straight line orthogonal to the x-axis contained in the plane of motion and subject to the external action $\bar{u} = \bar{u}(t)$.

If $z\in R_2$, let $\bar{z}, \underline{z}$ be, respectively, the first and the second component of z; it is easily seen that the transformation

$$\bar{z} = w_t, \qquad \underline{z} = w_x$$

reduces (6.1), $\cdots$, (6.4) to

$$z_t + A(z)z_x = 0, \qquad (x,t)\in\Re,$$

$$z(x,0) = \phi(x), \qquad x\in[0,1],$$

$$\bar{z}(0,t) = 0, \quad \underline{z}(1,t) = \underline{u}(t), \qquad t\in[0,T],$$

$$z(x,T) = 0, \qquad x\in[0,1],$$

where

$$A(z) = \begin{pmatrix} 0 & -c(\bar{z}) \\ -1 & 0 \end{pmatrix}, \quad \bar{\phi} = f', \quad \underline{\phi} = h.$$

Also the eigenvalues of $A(z)$ are $\pm c(\bar{z})$; so if

$$(6.6)\qquad 0 < a < \infty, \quad c\in C^1([-a,a],R), \quad c(0)\neq 0,$$

then A satisfies all the hypotheses of Corollary 4.1. Therefore as a particular case one obtains the following proposition.

PROPOSITION. *Suppose c satisfies* (6.6). *Then the wave equation* (6.1) *is zero controllable with one boundary control.*

Thus whenever f and h are conveniently restricted there are T, u and w satisfying (6.1) to (6.4). It is clear that if (6.6) holds then (6.1) is also controllable with two boundary controls.

REFERENCES

[1] A. G. BUTKOVSKII AND L. N. POLTAVSKII, *Optimal control of wave processes*, Avtomat. i Telemekh., 27 (1966), pp. 48–53.

[2] M. CIRINÀ, *Nonlinear hyperbolic problems: a priori bounds and solutions on preassigned sets*, Rep. 960, University of Wisconsin, Math. Res. Center, Madison, Wisconsin.

[3] R. CONTI, *On some aspects of linear control theory*, Proc. Conference on Mathematical Theory of Control at the University of Southern California (1967), Academic Press, New York, 1967, pp. 285–300.

[4] H. O. FATTORINI, *Time optimal control of solutions of operational differential equations*, this Journal, 2 (1964), pp. 54–59.

[5] ———, *Boundary control systems*, this Journal, 6 (1968), pp. 349–385.

[6] G. D. JOHNSON, *On a nonlinear vibrating string*, Doctoral thesis, University of California, Los Angeles, 1967.

[7] H. B. KELLER and V. THOMEE, *Unconditionally stable difference methods for mixed problems for quasilinear hyperbolic systems in two dimensions*, Comm. Pure Appl. Math., 15 (1962), pp. 63–73.

[8] H. O. KREISS, *Difference approximations for the initial-boundary value problem for hyperbolic differential equations*, Numerical Solution of Nonlinear Differential Equations, D. Greenspan, ed., John Wiley, New York, 1966.

[9] J. L. LIONS, *Sur le contrôle optimal des systèmes decrits par des équations aux dérivées partielles (I)*, *(II)*, *(III)*, C. R. Acad. Sci. Paris, 262 (1966), pp. 661–663, 713–715, 776–779.

[10] L. MARKUS, *Controllability of nonlinear processes*, this Journal, 3 (1965), pp. 78–90.

[11] D. L. RUSSELL, *Optimal regulation of linear symmetric hyperbolic systems with finite dimensional controls*, this Journal, 4 (1966), pp. 276–294.

[12] ———, *On boundary-value controllability of linear symmetric hyperbolic systems*, Proc. Conference on Mathematical Theory of Control at the University of Southern California (1967), Academic Press, New York, 1967, pp. 312–321.

[13] V. THOMÉE, *A stable difference scheme for the mixed boundary problem for a hyperbolic first order system in two dimensions*, J. Soc. Indust. Appl. Math., 10 (1962) , pp. 229–245.

SIAM J. CONTROL
Vol. 7, No. 2, May 1969

INTEGER PROGRAMMING OVER A FINITE ADDITIVE GROUP*

FRED GLOVER†

Abstract. An algorithm is given for solving an integer program over an additive group. Computation times appear to grow more favorably with increases in the number of variables and group elements than with the dynamic programming approach proposed by Gomory. A new property satisfied by optimal solutions to the group problem is established by reference to the structure of the algorithm. Extension of the algorithm to the general integer programming problem is developed in a sequel.

1. Introduction. In this paper we give an algorithm for solving the problem:

$$\text{(I)}\qquad \text{Minimize} \quad \sum_{j=1}^{n} c_j x_j$$

$$\text{subject to} \quad \sum_{j=1}^{n} \alpha_j x_j = \alpha_0, \quad x_j \geqq 0 \quad \text{and integer,} \quad j = 1, \cdots, n,$$

where the c_j are nonnegative scalar constants, and α_0 and the $\alpha_j, j = 1, \cdots, n$, are elements of a finite additive group. We sometimes also refer to $\sum_{j=1}^{n} c_j x_j$ in matrix notation as cx, where $c = (c_1, c_2, \cdots, c_n)$ and $x = (x_1, x_2, \cdots, x_n)$. An example of (I) is the problem:

$$\begin{aligned} \text{(I}'\text{)}\qquad &\text{Minimize} \quad 3x_1 + 7x_2 + 4x_3 \\ &\text{subject to} \quad 8x_1 + 3x_2 + 5x_3 \equiv 6 \pmod{11}, \\ &\qquad\qquad x_1, x_2, x_3 \geqq 0 \quad \text{and integer.} \end{aligned}$$

Alternatively, consider the linear integer programming problem:

$$\text{(II)}\qquad \text{Minimize} \quad \sum_{j=1}^{n} c'_j x'_j$$

$$\text{subject to} \quad \sum_{j=1}^{n} a'_{ij} x'_j + y'_i = b'_i, \quad i = 1, \cdots, m,$$

$$x'_j, y'_i \geqq 0 \quad \text{and integer for all } i, j,$$

where c'_j, a'_{ij} and b'_i are integer constants.

Applying the simplex method to (II) without the integer restriction on the x'_j and y'_i yields an equivalent representation:[1]

$$\text{(II}'\text{)}\qquad \text{Minimize} \quad \sum_{j=1}^{n} c_j x_j$$

$$\text{subject to} \quad \sum_{j=1}^{n} a_{ij} x_j + y_i = b_i, \quad i = 1, \cdots, m,$$

* Received by the editors November 16, 1966, and in final revised form December 4, 1968.

† Department of Statistics and Operations Research, The University of Texas at Austin, Austin, Texas 78712. This research was supported in part by the U.S. Office of Naval Research under Contract NONR 760(24) NR 047-048.

[1] It is assumed an optimal continuous solution exists.

where the c_j, a_{ij} and b_i are rational, the x_j and y_i are obtained by renaming the x'_j and y'_i (e.g., $x_1 = y'_2$, $x_2 = x'_4$, etc.) and $c_j \geqq 0$, $b_i \geqq 0$ for all i and j. Problem (II′) then becomes an instance of (I) by dropping the restriction $y_i \geqq 0$, giving:[2]

$$\text{(III)} \quad \text{Minimize} \quad \sum_{j=1}^{n} c_j x_j$$

$$\text{subject to} \quad \sum_{j=1}^{n} a_{ij} x_j \equiv b_i \,(\text{mod } 1), \quad i = 1, \cdots, m.$$

The significance of (I) lies in the fact that under certain conditions, its solution gives an optimal solution to (II) (see § 7).

It frequently happens that all of the constraints

$$\sum_{j=1}^{n} a_{ij} x_j \equiv b_i \,(\text{mod } 1), \quad i = 1, \cdots, m,$$

can be replaced by a single constraint

$$\sum_{j=1}^{n} \alpha_j x_j \equiv \alpha_0 \,(\text{mod } D),$$

where D, α_0 and the α_j are integer constants, so that (III) becomes equivalent to the class of problems whose form is illustrated by (I′).

A variety of such problems containing from 50 to 1500 variables and from 100 to 4500 group elements have been solved with the algorithm of this paper. Computational results are reported in § 8.

2. Methods for solving (I).

Two methods have been proposed for solving (I) other than the method of this paper. The first, due to Ralph Gomory [4], is based upon a dynamic programming recursion for the knapsack problem developed by Gilmore and Gomory [1]. Refinements in this approach have also been suggested by W. W. White [8]. Computation time is estimated to be proportional to nD, where n is the number of variables and D the order of the additive group.

The second method, due to Jeremy Shapiro [6], is based on a dynamic programming recursion for the knapsack problem developed by Shapiro and Wagner [7]. No estimates of computation time are available for this method, although the method appears intuitively to be quite promising.

The method of this paper takes a different approach that departs from the dynamic programming framework. An appeal to the structure of the algorithm establishes a new property satisfied by optimal solutions to (I) (see § 7). Computation times for the method, as reported in § 8, appear to depend somewhat more favorably on n and D than a direct proportionality to nD.

This method can be considered a dual method, in that optimal solutions are generated for a sequence of right-hand sides, until a feasible solution is found.

[2] That (III) is in fact an instance of (I) derives from the elegant theory developed by Gomory in [3], [4].

3. Simplified version of the algorithm. We describe three versions of the algorithm in this and the next two sections, beginning with the simple and working toward the more complex (and more efficient). Formal justification of the principal ideas and claims is deferred to § 6.

To begin with, we eliminate degeneracy by assuming $c_j > 0$. If $c_j \geqq 0$ is rational, this can be ensured as follows: multiply c by a positive integer large enough to make all components integer in the resulting new c. Then replace all $c_j = 0$ by $c_j = 1/P$, where P is a number such that $\sum_{c_j=0} x_j \leqq P - 1$. Thus any feasible adjustments of $\{x_j \mid c_j \text{ was zero}\}$ cannot change the objective as much as a unit change in any x_j. In particular, it suffices to let $P = D$ (see § 7).[3]

The algorithm generates a sequence of solutions (vectors of nonnegative integers) $x(1), x(2), \cdots, x(i), \cdots$, where $x(i)$ is the vector $(x_1^i, x_2^i, \cdots x_n^i)$. Associated with $x(i)$ is the "cost" $c(i) = \sum_{j=1}^n c_j x_j^i$ and the group element $\alpha(i) = \sum_{j=1}^n \alpha_j x_j^i$. If $\alpha(i) = \alpha_0$, then $x(i)$ is a *feasible* solution to (I). Each $x(i)$ is generated from an earlier solution $x(p)$, called the *predecessor* of $x(i)$, by incrementing one of the components of $x(p)$ by one. Thus if x_r^p is the component of $x(p)$ that is incremented to give $x(i)$, then we may write $x(i) = x(p) + e_r$, where e_r denotes the vector with 1 in the rth component and 0's elsewhere. We observe that $c(i) = c(p) + c_r$ and $\alpha(i) = \alpha(p) + \alpha_r$.

We construct the sequence of solutions to satisfy the following conditions:

(i) If $p \neq q$, then $x(p) \neq x(q)$.

(ii) If $p < q$, then $c(p) \leqq c(q)$.

(iii) $x(i)$ is an optimal solution to (I) when α_0 is replaced by $\alpha(i)$.

(iv) The solution sequence is finite, and $\alpha(i) = \alpha_0$ for some $x(i)$ if and only if problem (I) has a feasible solution.

If we alternately interpret the α_j as ordinary column vectors, our strategy in generating the $x(i)$ may be seen to correspond quite closely to the strategy of the dual simplex method in solving the ordinary linear programming problem. In fact, the successive basic solutions determined by the pivot rules of the dual simplex method satisfy exactly the same four conditions.

We shall introduce several of the fundamental ideas of the algorithm (in a simplified form) by means of an example. Consider the problem:

$$\begin{aligned} &\text{Minimize} && 3x_1 + 7x_2 + 4x_3 \\ &\text{subject to} && 8x_1 + 3x_2 + 5x_3 \equiv 6 \pmod{11} \end{aligned}$$

given in § 1 as an instance of (I).

Table 1 shows a sequence of solutions $x(i)$ generated by the algorithm.[4] Included in the table are the costs $c(i)$, group elements $\alpha(i)$, and the indices p_i and r_i from which one may verify the relations

$$x(i) = x(p) + e_r,$$

$$c(i) = c(p) + c_r,$$

$$\alpha(i) = \alpha(p) + \alpha_r$$

[3] Here (as earlier), and throughout the paper, we let D denote the number of elements in the additive group.

[4] The specific rules of the algorithm follow the example.

for $p = p_i$ and $r = r_i$, where p_i names the predecessor of solution i, and r_i names the variable which was incremented to get $x(i)$ from $x(p_i)$.

Note that the starting solution, $x(1)$, is the 0-vector. Because $x(1)$ has no predecessor, r_1 and p_1 have not been assigned values.

TABLE 1

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
$c(i)$	0	3	4	6	7	7	8	9	10	10	11	11	12	12	13	13	14	14	14
$\alpha(i)$	0	8	5	5	2	3	10	2	10	0	7	8	10	4	7	8	4	5	6
r_i		1	3	1	1	2	3	1	1	1	1	2	1	3	1	1	1	1	2
p_i		1	1	2	3	1	3	4	5	6	7	3	8	7	9	10	11	12	6
x_1^i	0	1	0	2	1	0	0	3	2	1	1	0	4	0	3	2	2	1	0
x_2^i	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	1	0	1	2
x_3^i	0	0	1	0	1	0	2	0	1	0	2	1	0	3	1	0	2	1	0
				x				*	x	x		x	*		*	*	x	*	

To identify the contribution of each variable x_j to the generation of Table 1, we define a transition index t_j which names the next solution from which x_j will be incremented. That is, if $x(1), x(2), \cdots, x(k-1)$ denote the solutions currently generated, then $x(k)$ will be one of the solutions $x(t_1) + e_1, x(t_2) + e_2, \cdots, x(t_n) + e_n$. That is, t_j will be the predecessor the next time x_j gets incremented. All of the t_j are initially set equal to 1, so that $x(2)$ will be one of the solutions $e_1, e_2, \cdots, e_n$.

As soon as x_j is incremented, i.e., when $x(t_j) + e_j = x(k)$, then the predecessor name is changed. The next time x_j gets incremented, its predecessor will be $\bar{t}_j$ instead of t_j, where[5]

$$\bar{t}_j = \min \{i : i > t_j \text{ and } r_i \geqq j\}.$$

All that remains for the determination of $x(k)$ is the selection of the particular index r for which $x(k) = x(t_r) + e_r$. To do this we define a *next cost* $N_j = c(t_j) + c_j$ for each j. N_j is the cost $c(k)$ when $x(k)$ is generated from the predecessor $x(t_j)$, i.e., when

$$x(k) = x(t_j) + e_j.$$

Then we select the index r by

$$N_r = \min \{N_1, N_2, \cdots, N_n\}$$

and set $x(k) = x(t_r) + e_r$.

We summarize our foregoing remarks in the following description of the procedure as developed to this point.

[5] The reason for the stipulation $r_i \geqq j$ is to avoid duplications. For example, without this restriction the solution $x(5) = e_1 + e_3$ in Table 1 could be generated both as $x(2) + e_3$ and $x(3) + e_1$, since $x(3) = e_3$.

Simplified Algorithm.

1. Begin with $x(1) = 0$ and $t_j = 1$ for $j = 1, \cdots, n$. Designate the solutions currently generated by $x(i), i = 1, \cdots, k - 1$.

2. To generate the next solution $x(k)$, select r to be one of the j's, $j = 1, \cdots, n$, by the rule $N_r = \min (N_j)$. If more than one index j is a candidate for r by this criterion, let r be the smallest of these indices. Then let $x(k) = x(t_r) + e_r$.

3. Update t_r by setting it equal to its next value $_r$ (note $\bar{t}_r \leqslant k$) and repeat the foregoing process.

The reader may verify that this algorithm generates the sequence of columns of Table 1. To facilitate this verification, Table 2 supplies the successive values assumed by the t_j and the N_j.

The entries for portions of the table left blank are the same as the nearest preceding entries in the same column. The value of r ($N_r = \min \{N_j\}$) at each stage

Table 2

	j	1	2	3	
	c_j	3	7	4	
1	t_j	1	1	1	$r = 1$
	N_j	3	7	4	
2	t_j	2			$r = 3$
	N_j	6			
3	t_j			3	$r = 1$
	N_j			8	
4	t_j	3			$r = 1$
	N_j	7			
5	t_j	4			$r = 2$
	N_j	9			
6	t_j		3		$r = 3$
	N_j		11		
7	t_j			7	$r = 1$
	N_j			12	
8	t_j	5			$r = 1$
	N_j	10			
9	t_j	6			$r = 1$
	N_j	10			
10	t_j	7			$r = 1$
	N_j	11			
11	t_j	8			$r = 2$
	N_j	12			
12	t_j		6		$r = 1$
	N_j		14		
13	t_j	9			$r = 3$
	N_j	13			
14	t_j			14	$r = 1$
	N_j			16	
15	t_j	10			$r = 1$
	N_j	13			
16	t_j	11			$r = 1$
	N_j	14			
17	t_j	12			$r = 1$
	N_j	14			
18	t_j	13			$r = 2$
	N_j	15			

is indicated to the right of the appropriate portion of Table 2. It may be noted that the amount of computation required in going from one iteration to the next is very small.

This simplified procedure generates solutions that are unnecessary for solving (I), and a glance at Table 1 discloses a variety of them: $x(4)$, $x(8)$, $x(9)$, $x(10)$, $x(12)$, $x(13)$, $x(15)$, $x(16)$, $x(17)$ and $x(18)$. These solutions are *dominated* in the sense that other solutions generated earlier in the table give the same $\alpha(i)$ with as good or better values for $c(i)$. Since these solutions are superfluous, they can be dropped.

There is of course no gain in dropping the dominated solutions at this stage, since the work devoted to generating them has already been done. However, if each solution is checked to see if it is dominated before it is added to the table, then the outcome is somewhat different. The x's and *'s beneath Table 1 show the columns that would never have entered the table. The x's are attached to columns that would have been checked for inclusion in the table, but rejected, and the *'s are attached to columns that never would have been checked or generated at all since they are descendants of other dropped columns.

It is not evident that solutions can be dropped legitimately at the point at which they are discovered to be dominated, unless they are dominated by a solution with a strictly lower cost. In fact, it can be shown that dropping dominated solutions can cause the method never to generate a feasible solution to (I), let alone an optimal one, if an improper tie-breaking rule is used in the choice of r at instruction 2.

A disguised complexity in the process of dropping solutions arises from the fact that some of the t_j's can thereby become "undefined." On the other hand, from an ability to drop solutions also comes an ability to impose bounds on variables, thereby further limiting the number of solutions examined.

The procedural details for accommodating these facts are given in the next section.

4. Procedures for handling dominated solutions and upper bounds. To supplement our previous remarks we define a list $G(k)$, $k = 1, 2, \cdots, D$, where $G(k) = 0$ if none of the $x(i)$ currently generated gives $\alpha(i) = g_k$ (g_k denotes the kth group element). Otherwise, if $\alpha(p) = g_k$ for some p, then $G(k) = p$. $G(k)$ names the solution index (or "iteration") p for which the right-hand-side element, $\alpha(p)$, is g_k.

The use of the G-list in dropping dominated solutions is as follows. When preparing to generate the solution $x(k) = x(t_r) + e_r$, identify the group element g_h such that $g_h = \alpha(t_r) + \alpha_r$. Then $x(t_r) + e_r$ is permitted to be generated as $x(k)$ only if $G(h) = 0$, whereupon $G(h)$ is set equal to k. Otherwise, if $G(h) \geqq 1$, then $x(t_r) + e_r$ is dominated by the previously generated solution $x(i)$ for $i = G(h)$, and thus is not recorded in the table. Note that x_r might eventually be incremented even if $x(t_r) + e_r$ is dominated at iteration k. Therefore, whether or not it is dominated, the next step is to find the next value $\bar{t}_r$ for t_r. We now define

$$\bar{t}_r = \min \{i : i > t_r, r \leqq r_i \text{ and } G(q) = 0,$$
$$\text{where } g_q \text{ denotes the group element } \alpha(i) + \alpha_r\}.$$

As already intimated, there may not be a next value for t_r that satisfies this definition. Thus, we introduce the set $T = \{j : t_j \text{ is defined}\}$. Initially, T contains all the j, $j = 1, \cdots, n$ (since $t_j = 1$ for all j). Thereafter, the composition of T can vary. But from the results of § 6, T cannot become empty unless (I) has no solution.

We now summarize these remarks by describing an algorithm for (I) that accommodates dominated solutions.

To begin, let $T = \{j : j = 1, \cdots, n\}$, $t_j = 1$ for all $j \in T$, $G(h) = 0$ for $h = 1, \cdots, D - 1$ and $G(D) = 1$, where g_D is the "0" group element, generated by the starting solution $x(1) = 0$. If $a_0 = 0$, the problem is trivially solved by $x(1)$.

Otherwise, we denote the solutions generated at the current stage of the method by $x(1), \cdots, x(k - 1)$ and the next step is to generate $x(k)$.

ALGORITHM FOR (I).

1. If T is empty, problem (I) has no solution. Otherwise, identify the index r such that

$$N_r = \min_{j \in T} \{N_j\}.$$

If more than one j qualifies to be r, let r be the smallest of the qualifying indices.

2. Let g_h denote the group element given by $g_h = \alpha(t_r) + \alpha_r$.

   (i) If $G(h) \geqq 1$, do nothing at instruction 2. Go to instruction 3.

   (ii) If $G(h) = 0$, indicating that g_h has not previously been generated, generate the solution $x(k) = x(t_r) + e_r$ and let $G(h) = k$. If $\alpha(k) = \alpha_0$, $x(k)$ is optimal for (I) and the method stops.

3. Update t_r to its next value $\bar{t}_r$ (using the expanded definition of this section). If the updated value of t_r does not exist, remove r from T.

4. If a new solution $x(k)$ was *not* generated at instruction 2, then return to step 1 to pick up the next smallest N_j. But if a new $x(k)$ was generated, check whether any of the $j \notin T$ can be returned to T; i.e., whether $j \leqq r_k$ and $G(h) = 0$ for g_h given by $g_h = \alpha(k) + \alpha_j$. Let $t_j = k$ for all such j added back to T, and then return to step 1 to generate $x(k)$ for the next larger value of k.

We illustrate the algorithm above by applying it to the problem:

$$\begin{aligned} \text{Minimize} \quad & 3x_1 + 4x_2 + 5x_3 + 7x_4 \\ \text{subject to} \quad & 5x_1 + 9x_2 + 3x_3 + 4x_4 \equiv 1 \pmod{10}. \end{aligned}$$

Table 3 gives the sequence of solutions generated by the algorithm.

TABLE 3

i	1	2	3	4	5	6	7	8	9	10
$c(i)$	0	3	4	5	7	8	9	10	12	13
$\alpha(i)$	0	5	9	3	4	8	2	6	7	1
r_i		1	2	3	1	1	2	3	1	1
p_i		1	1	1	3	4	4	4	7	8
x_r^i	0	1	1	1	1	1	1	2	1	1
Σx_j^i	0	1	1	1	2	2	2	2	3	3

Notice that in place of recording the vector $x(i)$ for each column, as in the earlier example, we have instead recorded $\sum x_j^i$ and the value of the single variable x_r^i (for $r = r_i$). The formula for determining these values is as follows. Let $x(p)$ denote the predecessor of $x(i)$, i.e., $x(i) = x(p) + e_r$. Then $\sum x_j^i = \sum x_j^p + 1$, and $x_r^i = x_r^p + 1$ if $r_p = r$ and $x_r^i = 1$ otherwise.

The successive iterations of the algorithm that produced the columns of this table are summarized in Table 4. As before, entries for portions left blank are the same as the nearest preceding entries in the same column.

TABLE 4

	j	1	2	3	4	
	c_j	3	4	5	7	
	α_j	5	9	3	4	
	t_j	1	1	1	1	
1	N_j	3	4	5	7	$r = 1$
	$\alpha(t_j) + \alpha_j$	5	9	3	4	
	t_j	*				
2	N_j	*				$r = 2$
	$\alpha(t_j) + \alpha_j$	*				
	t_j	3	3			
3	N_j	7	8			$r = 3$
	$\alpha(t_j) + \alpha_j$	4	8			
	t_j			4		
4	N_j			10		$r = 1$
	$\alpha(t_j) + \alpha_j$			6		
	t_j	4			*	$r = 4$
5	N_j	8			*	$r = 1$
	$\alpha(t_j) + \alpha_j$	8			*	
	t_j	*	4			$r = 2$
6	N_j	*	9			$r = 2$
	$\alpha(t_j) + \alpha_j$	*	2			
	t_j	7	7			
7	N_j	12	13			$r = 3$
	$\alpha(t_j) + \alpha_j$	7	1			
	t_j			*		
8	N_j			*		$r = 1$
	$\alpha(t_j) + \alpha_j$			*		
	t_j	8				
9	N_j	13				$r = 1$
	$\alpha(t_j) + \alpha_j$	1				

The steps of the algorithm can be traced from the tables as follows. From step 1 of Table 4, $r = 1$, producing column 2 of Table 3. Thereupon, the next value for t_1 would ordinarily be 2, except that $\alpha(2) + \alpha_1 = 10 \equiv 0$, and 0 has already been generated ($\alpha(1) = 0$). Consequently, since there are no other possible values for t_1, it becomes undefined, as indicated by the asterisks in step 2.

At step 2, $r = 2$, producing column 3 of Table 3. t_1 becomes defined again ($t_1 = 3$) and the next value of t_2 is determined ($t_2 = 3$), as shown in step 3 of Table 4.

Step 4 of Table 4 is generated routinely. At step 5, however, $r = 4$ is indicated, except that $\alpha(1) + \alpha_4 = 4$ has by now been generated ($\alpha(5) = 4$), and hence the next permissible value is sought for t_4. There is none, and so t_4 becomes undefined (as indicated by the asterisks). The actual value of r at step 5 is therefore $r = 1$.

At step 6, $r = 2$ is indicated, but $\alpha(3) + \alpha_2 = 8$ has been generated ($\alpha(6) = 8$). Thus, the next value of t_2 is determined, giving $t_2 = 4$. In this case, N_2 is still minimum, and so $r = 2$ gives the correct value of r. Steps 7, 8 and 9 of Table 4 are determined similarly.

The optimal x-vector can be recovered as follows. Begin with $x = 0$ and i the index of the optimal (last) column of Table 3.

Let $x_r = x + x_r^i e_r$ (for $r = r_i$) and identify the group element $g_h = \alpha(i) - x_r^i \alpha_r$. From the G-list (or simply by scanning back through the table) locate the column i for which $g_h = \alpha(i)$ (i.e., set $i = G(h)$), and then repeat this procedure until $i = 1$.

The foregoing also works by replacing x_r^i with 1 at each step, so that the x_r^i values need not have been recorded for this purpose. Using either procedure, we see that an optimal solution is given by $x = (1 \quad 0 \quad 2 \quad 0)$.

The x_r^i and $\sum x_r^i$ values are not needed to recover the optimal solution, but can be used to serve another more fruitful purpose. Specifically, whenever a solution $x(t_r) + e_r$ is rejected as the next $x(k)$ at instruction 2, postpone this rejection, temporarily designating $x(k) = x(t_r) + e_r$. Then if $x_r^k = \sum x_j^k$, it follows that[6] $x(k) = x_r^k e_r$. Moreover, since $x(k)$ is dominated, one may reasonably guess that x will satisfy $x_r \leqq x_r^k - 1$ in all undominated solutions; this is shown to be true in Lemma 5, § 6.

We denote the upper bound so determined for x_r by U_r. After checking for such a bound, we discard, without being recorded, the dominated solution temporarily designated $x(k)$, and the process continues.

Similarly, one may check to establish an upper bound for x_r at instruction 3 when seeking the updated value $\bar{t}_r$ for t_r, since the chance to identify dominated solutions also arises there.

To make use of the upper bounds U_j thus determined, one expands the definition of $\bar{t}_j$ to

$$\bar{t}_j = \min \{i : i > t_j, j \leqq r_i, x_j^i < U_j \text{ and } G(h) = 0, \text{ where } g_h = \alpha(i) + \alpha_j\}.$$

[6] A slightly quicker way to check whether $x(k) = x_r^k e_r$ is to record a flag for each $x(k)$ which takes the value 0 if $x(k) = x_r^k e_r$ and 1 otherwise. The flag for a successor $x(q)$ of $x(k)$ is the same as for $x(k)$ if $r_q = r_k$, and is 1 if $r_q \neq r_k$.

The stipulation[7] $x_j^i < U_j$ is easily checked after checking for $j \leqq r_i$, since $x_j^i = 0$ if $j < r_i$, and x_j^i is precisely the value recorded in the table as "x_r^i" if $j = r_i$.

Had such upper bounds been computed in generating Tables 3 and 4, $U_1 = 1$ would have been determined at step 2 of Table 4, $U_4 = 0$ at step 5 and $U_3 = 2$ at step 8. Also, accounting for $U_1 = 1$ would have avoided two attempts to determine a next value for t_1 in going from step 5 to 6, and accounting for $U_4 = 0$ would have avoided repeated checks to see whether t_4 should become defined once again after step 5.

5. An accelerated version of the algorithm. We now show how to solve problem (I) by generating only a subset of the $x(i)$ produced by the algorithm in § 4.

First note that the algorithm in § 4 generates optimal solutions $x(i)$ in order of increasing cost, stopping when a solution with the desired right-hand side is reached. Let x^* be the optimal x-vector which is to be generated by the algorithm. Consider two nonnegative integer vectors x^a and x^b such that $x^* = x^a + x^b$ and $|cx^b - cx^a| = \min |cx'' - cx'|$, where x' and x'' range over all pairs of nonnegative integer x that sum to x^*. We prove in § 6 (Lemmas 6 and 7) that vectors qualifying to be x^a and x^b will be generated by the algorithm. Consequently, we hereafter denote these solutions by $x(a)$ and $x(b)$, where, say, $a < b$.

Since $c(a)$ and $c(b)$ are either equal or nearly so, it may be expected that $x(a)$ and $x(b)$ will be generated somewhat before x^*. But since $x^* = x(a) + x(b)$, it would be possible to stop immediately after generating $x(b)$, eliminating the generation of all subsequent solutions.

Let $\gamma = c(k) - c(p)$, where k and p are candidates for b and a, so that $x(k) + x(p) = x^* \cdot \gamma$ is $\geqq 0$ because $k > p$. The accelerated algorithm will generate candidates for $x(a)$ and $x(b)$ in order of increasing (or nondecreasing) γ. The trick is to know when the optimal $x(k)$ and $x(p)$, namely $x(a)$ and $x(b)$, have been generated.

Clearly, the first step is to check each time a new $x(k)$ is generated at instruction 2 to determine whether $G(q) \geqq 1$, where g_q is the group element $\alpha_0 - \alpha(k)$. If so, $x(k) + x(p)$ is a feasible solution to (I), where $p = G(q)$ $(\leqq k)$. However, $x(p) + x(k)$ may not be optimal, and, in general, several feasible solutions to (I) may be found by repeating this check for successive $x(k)$.

Let x' denote the best of these solutions. Now, if x' is not optimal, then $c(b) + c(a) < cx'$. Furthermore, until $x(b)$ is finally generated, it must be true that $c(b) \geqq N_r$, because candidates for $x(b)$ are generated in order of increasing cost. Consequently, $c(a) < cx' - N_r$, and an upper bound for $c(s)$ is found by identifying the largest $c(i)$ (call it $c(a')$) such that $c(i) < cx' - N_r$. Since $c(b) \geqq N_r$ and $c(a) \leqq c(a')$, it must be that $c(b) - c(a) \geqq N_r - c(a')$. Moreover, since $c(b)$ and $c(a)$ are as nearly the same as possible, it is also evident that $c(b) - c(a) \leqq c_m$, where $c_m = \max \{c_j\}$. Thus once $N_r - c(a') > c_m$ becomes satisfied, $x(b)$ has already been generated and x' is the optimal solution.

[7] The handling of the U_j can alternately be accommodated by requiring "$j \leqq d_i$" instead of "$j \leqq r_i$ and $x_j^i < U_j$" in the definition of the $\bar{t}_j$, where $d_i = r_i - 1$ if $x_r^i = U_r$ $(r = r_i)$, and $d_i = r_i$ otherwise.

Frequently, a smaller value can be given for c_m than $\max\{c_j\}$, thereby permitting earlier termination of the algorithm. To see this, let $\delta = c(b) - c(a)$. Then if x' is not optimal, $2c(a) + \delta < cx'$, and hence $\delta < cx' - 2c(a)$. Suppose each α_j is scanned before starting the algorithm and if $\alpha_j = \alpha_0$, the solution $x = e_j$ is admitted as a candidate for x'. This implies that if x' is not optimal, it is also true that $x(a) \neq 0$, and hence $c(a) \geqq c(1)$. Thus $\delta < cx' - 2\,c(1)$, and c_m may alternately be given by $c_m = \max\{c_j : c_j < cx' - 2\,c(1)\}$.

The cutoff level thus determined will generally succeed in stopping the algorithm considerably in advance of generating x^*. However, the whole process becomes more effective by dropping solutions that would ordinarily be retained. Specifically, when $x(p) + x(k)$ is found to be feasible for (I), both $x(p)$ and $x(k)$ and all their successors can be eliminated from further consideration (Lemma 7, § 6). This is easily accomplished for $x(k)$ simply by not recording it in the table (although $G(h)$ is assigned some positive value to permit solutions dominated by $x(k)$ to be dropped). To prevent additional successors from being generated it suffices to set $r_p = 0$ (or $r_p = -r_p$ in case it is desired to recover the value of r_p later). Similarly, one may locate successors $x(i)$ of $x(p)$ that are already generated, and set $r_i = 0$ (or $r_i = -r_i$) to assure that no more of *their* successors will be generated. Clearly this process can be carried out for as many generations of descendants of $x(p)$ as desired. However, the chances of finding a descendant of $x(p)$ beyond an immediate successor are probably remote.

The main content of the foregoing discussion can be summarized by prescribing the following changes in the constructions of the algorithm as stated in § 4.

Change in instruction 1. If no feasible solution for (I) has previously been found, the instruction remains unchanged. Otherwise, let x' denote the best solution found. Identify N_r (as before), let $c(a') = \max\{c(i) : c(i) < cx' - N_r\}$, and let $c_m = \max\{c_j\}$ (or $c_m = \max\{c_j : c_j < cx' - 2\,c(1)\}$ if the solutions $x = e_j$, $j = 1, \cdots, n$, have been included as candidates for x'). If one of the following conditions holds, then x' is optimal for (I) and the method stops:

(i) T is empty;

(ii) $c(a')$ or c_m does not exist;

(iii) $N_r - c(a') > c_m$.

Change in instruction 2. If 2(i) is applicable, the instruction is unchanged. If instruction 2(ii) is applicable, set $G(h) = k$ (as before) but postpone all other work involved in recording $x(k)$. Identify the group element $g_q = \alpha_0 - \alpha(k)$ and the index $p = G(q)$. If $p = 0$, the generation of $x(k)$ is recorded, as before, and nothing further is done. But if $p \geqq 1$, then $x(p) + x(k)$ is feasible for (I) and is designated the new x' if $c(p) + c(k) < cx'$ (letting $cx' = \infty$ if x' does not exist). Furthermore, the generation of $x(k)$ is not recorded, and if $p \neq k$, r_p is set equal to 0 (or to $-r_p$) to prevent generating new successors of $x(p)$. (One may also replace r_i by $-r_i$ for successors $x(i)$ of $x(p)$ already generated, and similarly for *their* successors, etc.)

Except for these changes, the algorithm remains the same as before. Note that the dropping of $x(k)$ and $x(p)$ specified by the changed instruction 2 may provide upper bounds for some of the x_j in the manner described in the latter part of § 4.

We trace the course taken by the accelerated version of the algorithm by examining Table 3. Since the accelerated version is the same as the original except for checking for new solutions $x(k) + x(p)$ and dropping their successors, we confine ourselves to determining the effect of these operations on the columns of the table.

The first candidates found for $x(a)$ and $x(b)$ are $x(4)$ and $x(6)$—$\alpha(4) + \alpha(6) = 11 \equiv 1$—yielding $cx' = c(4) + c(6) = 13$. We shall now verify that this is optimal. First, no successors of $x(4)$ or $x(6)$ need be generated. Thus, changing r_4 from 3 to 0 and masking over the column for $x(6)$ (which is not actually generated by the accelerated method) insures that $x(7)$ will be bypassed. Also, $x(8)$ need not be generated, since it is also a successor of $x(4)$. However, to avoid its generation would ordinarily require checking the current t_j and updating t_3 which is found to equal 4. But the method stops without generating $x(8)$ by checking the relation $N_r - c(a') > c_m$. Specifically, upon preparing to generate $x(8)$, $N_r = 10$; hence $c(a') < 13 - 10$, giving $c(a') = 0$. Also, $c_m = 7$, and the relation becomes $10 - 0 > 7$, which is true, thus signaling optimality and directing the method to stop.

The optimality of $x(4) + x(6)$ can also be verified more quickly if the preliminary scanning is used to admit each e_j as a candidate for x' (none of the e_j qualify).

Then $c_m < cx' - 2c(1) = 13 - 6$, giving $c_m = c_3 = 5$. Before updating N_r from 9 to 10, the relation $N_r - c(a') > c_m$ is $9 - 3 > 5$, the validity of which again signals optimality.

6. Theorems and proofs. We refer to the simplified (incomplete) form of the algorithm given in § 3 by stipulating that no solutions are dropped, and the complete form of the algorithm (including the use of upper bound restrictions) by stipulating that solutions *are* dropped.

LEMMA 1. *If no solutions are dropped and the algorithm is not permitted to stop upon generating α_0, then the method will generate every x-vector having finite components.*

Proof. Note that $c > 0$ implies every $j, j = 1, \cdots, n$, will be selected as r at finite intervals. Suppose $x = x'$ is not generated. Then neither would the method generate $x(i) = x' - e_u$, where u is the first nonzero component of x'. For clearly $u \leqq r_i$, which means t_u must eventually be set equal to r_i and hence x' generated. Repeating this argument implies that 0 is not generated, contrary to $x(1) = 0$.

LEMMA 2. *No solution is generated twice, whether or not some solutions are dropped.*

Proof. Let $x(q)$ be the first solution that duplicates a previous one, say $x(p)$. Then for some $h < q$ and $k < p$, $x(q)$ was generated as $x(h) + e_r$ and $x(p)$ was generated as $x(k) + e_r$, where r is the first nonzero component of $x(q)$ and $x(p)$. Thus $x(k) = x(h)$, and since $x(q)$ is the first duplicating solution, $h = k$. When $x(p)$ was generated $t_r = h$, and then t_r was increased, never to be decreased. Consequently, $x(q)$ could not have been generated from $x(h)$, contrary to assumption.

Lemma 3. *If no solutions are dropped,* $p < q$ *implies*[8]

(i) $c(p) < c(q)$ *or*

(ii) $c(p) = c(q)$ *and* $x(p) \overset{l}{>} x(q)$.

Proof. Note that either (i) or (ii) is satisfied for $q = 2$ and $p < q$ (hence $p = 1$). Suppose the lemma is true for all $q < k$ and $p < q$. We prove it true for $q = k$ and $p = h < k$. Write $x(k) = x(k') + e_u$, $x(h) = x(h') + e_v$. When $x(h)$ was generated, $c(h) = N_v$ and either $N_v < N_u$ or $N_v = N_u$ and $v < u$. If t_u (and N_u) are the same when $x(k)$ is generated as when $x(h)$ was generated, then the proof is immediate. If t_u and N_u change, then let N_u' be the new N_u when generating $x(k)$. We have $N_u' = c(t_u') + c_u$ and $N_u = c(t_u) + c_u$. But $t_u < t_u' < k$ and, by hypothesis, (i) or (ii) holds relative to $p = t_u$ and $q = t_u'$. It follows immediately that (i) or (ii) must also hold relative to $p = h$ and $q = k$.

Lemma 4. *Let S denote the sequence of solutions generated by the simplified algorithm and suppose this algorithm is modified so that occasionally solutions are not generated but bypassed (according to any rule whatsoever). The resulting sequence of solutions S' is a subsequence of S (i.e., contains a subset of the solutions of S in the same relative order).*

Proof. It is evident from Lemma 1 that S and S' are well-defined. Denote those $x(i)$ in S by $x^1(i)$ and those $x(i)$ in S' by $x^2(i)$. Let $\hat{S}$ be the subsequence of S obtained by deleting from S each $x^1(i)$ such that $x^1(i) \neq x^2(k) \in S'$. We note by Lemmas 2 and 3 that the components of $\hat{S}$ must be a permutation of those of S'. Thus, designate the smallest i such that $x^1(i) \in \hat{S}$ by $\bar{0}$, the next smallest i by $\bar{1}$, and so on. Then we wish to prove that $x^1(\bar{i}) = x^2(i)$ for all $x^2(i) \in S'$. Suppose otherwise, and let $p = \min\{i : x^1(\bar{i}) \neq x^2(i)\}$. Also identify the indices q and r such that $x^1(\bar{p}) = x^2(q)$ and $x^2(p) = x^1(\bar{r})$. It is assured by Lemma 2 that $q, r > p$. Since $\bar{i} = \bar{s}$ if and only if $i = s$ for all i and s, we have $x^1(\bar{p}) = x^1(\bar{h}) + e_u$ for some u and some $h < p$, and $x^2(p) = x^2(k) + e_v$ for some v and some $k < p$. Now, $x^1(\bar{p})$ was generated before $x^1(\bar{r})$, but $x^1(\bar{r}) = x^2(p)$ implies $x^1(\bar{r}) = x^1(\bar{k}) + e_v$. Since $k < p$, this means that when $x^1(\bar{p})$ was generated, $t_v\,(= \bar{k})$ was well-defined ($v \in T$). Thus there was a choice to make between generating $x^1(\bar{p})$ and $x^1(\bar{r})$. Similarly, $x^2(p)$ was generated before $x^2(q)$, but $x^2(q) = x^2(h) + e_u$, so that, by analogous reasoning, there was a choice to make between generating $x^2(p)$ and $x^2(q)$ when $x^2(p)$ was generated in S'. But $\bar{p} < \bar{r}$ thus implies $q < p$, providing a contradiction.

Remark. Lemma 4 establishes the validity of Lemma 3 for the case when solutions are dropped.

Lemma 5. *If a solution x' is dropped at instructions* 2 *or* 3 *of the complete algorithm, then there is no vector $x^* \geqq x'$ that is a lexicographically largest optimal solution.*

Proof. Suppose this lemma is false, and let x' be the first solution dropped that has a lexicographically largest optimal descendant. x' is dropped because there is a solution $x(i)$ already generated such that $c(i) \leqq cx'$ and $\alpha(i) = \sum \alpha_j x_j'$. Let

[8] A vector y is defined to be *lexicographically larger* than a vector z, written $y \overset{l}{>} z$, if the first nonzero component of $y - z$ is positive.

$x^* = x' + x''$ be the lexicographically largest optimal solution. Since $\alpha(i) = \sum \alpha_j x'_j$ and $c(i) \leqq cx'$ it must be true that $x(i) + x''$ is also optimal. Moreover, $c(i) = cx'$. But then $x(i)$ is lexicographically larger than x' (since x' would have been generated later than $x(i)$) and in turn $x(i) + x''$ is lexicographically larger than $x' + x''$, contrary to assumption.

Lemmas 1 to 5 immediately imply the next theorem.

THEOREM 1. *The algorithm of* § 4 *yields an optimal solution to* (I) *or verifies that no feasible solution exists, after generating at most D solutions $x(i)$, each of which is optimal for* (I) *with α_0 replaced by $\alpha(i)$.*

The succeeding results refer to the accelerated algorithm of § 5.

LEMMA 6. *Let $x(a)$ and $x(b)$, $a < b$, be two solutions such that $x(a) + x(b)$ is optimal for* (I), *and, moreover, let b be the least index $(b \geqq a)$ for which two such solutions can be found. Then $c(b) - c(a) \leqq cx' - cx''$ for all solutions $x', x'' \geqq 0$ such that $x' + x''$ is optimal for* (I) *and $cx' \geqq cx''$.*

Proof. Let x^q and x^p be two solutions qualifying as x' and x'' and minimizing $cx' - cx''$. Thus the lemma asserts $c(b) - c(a) = cx^q - cx^p$. The lemma is trivially true for $cx^p = 0$; hence suppose $cx^p > 0$.

Let $\alpha^p = \sum \alpha_j x_{pj}$ and $\alpha^q = \sum \alpha_j x_{qj}$, where x_{pj} and x_{qj} are the jth components of x^p and x^q. Define problem (I^p) to be the same as (I) with α^p replacing α_0 and (I^q) to be the same as (I) with α^q replacing α_0. Since $cx^p, cx^q < c(x^p + x^q)$, it follows from Theorem 1 and Lemma 3 that there exist $x(p)$ and $x(q)$ generated in the process of solving (I) such that $x(p)$ is optimal for (I^p) and $x(q)$ is optimal for (I^q). Suppose $p \leqq q$. By assumption $b \leqq q$, and hence $c(b) \leqq c(q)$. But since $c(q) + c(p) = c(a) + c(b)$, it follows that $c(a) \geqq c(p)$ and hence $c(q) - c(p) = c(b) - c(a)$, proving the lemma.

LEMMA 7. *Let $x(a)$ and $x(b)$ be as in Lemma* 6, *and suppose there are solutions $x(h)$ and $x(k)$, $h \leqq k < b$, such that $x(h) + x(k)$ is feasible for* (I). *Then there is no vector $z \geqq 0$ that satisfies one or more of the following four conditions*:

(i) $x(h) + z = x(a)$,

(ii) $x(h) + z = x(b)$,

(iii) $x(k) + z = x(a)$,

(iv) $x(k) + z = x(b)$.

Proof. Since $k < b$, $x(h) + x(k)$ is not optimal, and hence $c(h) + c(k) > c(a) + c(b)$. Condition (i) implies $x(h) + z + x(b)$ is optimal, hence there exists $x(v)$ such that $x(v) + x(h)$ is optimal and $c(v) < c(k)$. Consequently $v < k$, and $x(v)$ and $x(k)$ qualify to be $x(a)$ and $x(b)$, contrary to $k < b$. Conditions (ii), (iii) and (iv) lead to similar contradictions.

Lemmas 6 and 7 establish the next theorem.

THEOREM 2. *The accelerated algorithm will find an optimal solution if one exists and, in particular, will generate $x(a)$ and $x(b)$ of Lemma* 6.

7. Properties of optimal solutions.

The characteristics of the solution sequence $x(1), x(2), \cdots$, generated by the algorithm, make several properties of optimal solutions to (I) immediately evident. For example, let n_j denote the order of the subgroup generated by all multiples of α_j. Then, since $n_j\alpha_j$ is the 0-element, the solution n_je_j is dominated by $x(1)$, and $x_j^i \leqq n_j - 1$ holds for every $x(i)$ generated. The existence of optimal solutions with this property is proved by Gomory in [4].

Moreover, there are at most D of the $x(i)$ (including $x(1) = 0$), and the sum of the variables in each is only one more than in its predecessor. Thus it is evident that $\sum x_j^i \leqq D - 1$ for all $x(i)$. The existence of optimal solutions with this property is also proved by Gomory in [4].

We see that solutions satisfying *both* of the two foregoing properties exist and, in fact, are the only solutions generated by the algorithm.

More recently Gomory has proved that optimal solutions may be found that satisfy[9] $\prod_{j=1}^n (x_j + 1) \leqq D$. It may be observed that this property is somewhat stronger than $\sum x_j^i \leqq D - 1$.

We shall prove a different property that is also considerably stronger than $\sum x_j^i \leqq D - 1$ by a direct appeal to the structure of the algorithm.

THEOREM 3. *Let* (I^i) *denote problem* (I) *with* α_0 *replaced by* g_i *for* $i = 1, \cdots, D$. *Then there exists a set of optimal solutions* x^1 *for* (I^1), x^2 *for* (I^2), $\cdots$, x^D *for* (I^D), *such that* $\max\{x_1^1, x_1^2, \cdots, x_1^D\} + \max\{x_2^1, x_2^2, \cdots, x_2^D\} + \cdots + \max\{x_n^1, x_n^2, \cdots, x_n^D\} \leqq D - 1$.

Proof. Label the group elements to correspond to the $\alpha(i)$ generated by the algorithm; i.e., $g_1 = \alpha(1)$, $g_2 = \alpha(2), \cdots, g_D = \alpha(D)$, where $\alpha_0 = g_D$. Then the solutions x_i specified by the theorem are precisely the $x(i)$ generated by the algorithm of § 4. To see this, let $U_j = \max\{x_j^1, x_j^2, \cdots, x_j^D\}$ for $j = 1, \cdots, n$. The theorem asserts $\sum U_j \leqq D - 1$. Beginning with $j = 1$, delete each $x(i)$ which is derived from its predecessor by incrementing only x_1. No solution $x(k)$ is deleted in which any of the components $x_2^k, x_3^k, \cdots, x_n^k$ differs from the corresponding component of the predecessor of $x(k)$. Consequently, $x_2, x_3, \cdots, x_n$ attain their maximum values in the undeleted solutions. There are at least U_1 of the $x(i)$ to be deleted, leaving at most $D - U_1$ solutions behind (one of which is $x(1) = 0$). Moreover, in each remaining $x(k)$, $\sum_{j=2}^n x_j^k$ is only one larger than in one of the preceding $x(k)$. Thus since there are at most $D - 1 - U_1$ solutions other than $x(1) = 0$, $\sum_{j=2}^n x_j^k \leqq D - 1 - U_1$. We repeat this process for $j = 2, \cdots, n$. At each step r, $\sum_{j=r+1}^n x_j^k \leqq D - 1 - \sum_{j=1}^r U_j$. Finally, we obtain $0 \leqq D - 1 - \sum_{j=1}^n U_j$, or $\sum_{j=1}^n U_j \leqq D - 1$, as claimed.

The U_j in the proof of the preceding theorem constitute upper bounds for the x_j that apply regardless of which group element α_0 happens to be. One way to determine such a set of U_j is to apply the algorithm until every g_i is generated, and then compute $\max\{x_j^1, \cdots, x_j^D\}$ for each j. There is also a second much faster way. Suppose the definition of $\bar{t}_j$ is simplified so that, when $x(t_j) + e_j$ is generated as the solution $x(k)$, j is set equal to k. Further suppose the method is stopped only when T becomes empty. This "modified" version of the algorithm has the following features.

(i) Only solutions of the form $1e_j, 2e_j, 3e_j, \cdots$ are generated for each j.

(ii) As soon as a solution he_j is dropped (checked but not generated), j is removed from T and never returns. At this point U_j can be recorded as $h - 1$.

[9] Reported by W. W. White in [8].

It may be observed that the foregoing method will usually generate fewer (and never more) than the D solutions required to determine the U_j with the unmodified algorithm. Moreover, all comparison operations in determining the next value $\bar{t}_j$ of t_j are eliminated. No upper bounds are checked for the variables, since a variable drops from T as soon as its upper bound is attained and verified. Because of this, T also tends to shrink more rapidly than with the unmodified algorithm, reducing the number of effective problem variables. Finally, there is no need to check those $j \notin T$ to see if they should be put back into T.

By Lemma 4, the sequence of solutions generated is a subsequence of that generated if no solutions are dropped. (Here some of the solutions are "dropped" by the restrictive definition of $\bar{t}_j$.) Solutions bypassed due to dominance considerations are therefore truly dominated and would not be generated in any case. Consequently, the U_j are valid (although possibly not as restrictive as those obtained from the sequence of $x(i)$ generated by the unmodified algorithm) and satisfy $\sum U_j \leqq D - 1$ by the proof of Theorem 3.

8. Computational experience. Roughly five hundred problems have been solved with the algorithm, containing from 50 to 1500 variables and from 100 to 4500 group elements. The problems all have the form

$$\text{Minimize} \quad \sum_{j=1}^{n} c_j x_j$$

$$\text{subject to} \quad \sum_{j=1}^{n} \alpha_j x_j \equiv \alpha_0 \,(\text{mod}\, D), \quad x_j \geqq 0 \text{ and integer,}$$

where the c_j and α_j are positive integers.

The c_j were randomly generated to lie within a specified interval, and several different intervals were tested to determine the effect on computation times.

The α_j were generated by selecting α_1 randomly from the set $S = \{1, 2, \cdots, D - 1\}$, α_2 randomly from the set $S - \{\alpha_1\}$, α_3 randomly from the set $S - \{\alpha_1, \alpha_2\}$, and so on. Thus, $\alpha_p = \alpha_q$ for $p \neq q$ was avoided, although one might expect this situation to arise in practice, thereby making it possible to reduce the number of problem variables.

Representative tables of computation times follow. All times reported are in seconds of central processing time on the CDC 6600.[10] The c_j were arranged in ascending order before starting the algorithm but the time for this preliminary ordering is not included.

The tables are headed with the symbols n, D, Total, Av., Fast and Ratio. "Total" gives the time for the algorithm of § 4 to solve problem (I) for every value of α_0. That is, the algorithm is permitted to continue until $\alpha(1), \alpha(2), \cdots, \alpha(D)$ are all generated. Since, in practice, one will often be interested in solving (I) for a

[10] The code was written in FORTRAN IV.

particular value of α_0, the times to solve (I) for[11] $\alpha_0 = \alpha(D/5), \alpha(2D/5), \cdots, \alpha(D)$ were averaged to give an idea of expected computation time, and this average appears in the column headed "Av."

The "Fast" column gives the computation time for solving problem (I) with the accelerated version of the algorithm. The accelerated version was applied to (I) with $\alpha_0 = \alpha(D)$ (thus requiring more computation than with α_0 at any other value).

The "Ratio" column gives the ratio of the "Fast" column to the "Total" column, indicating the relative efficiency of the accelerated version to the version of the algorithm of § 4.

From Tables 5, 6 and 7 it may be seen that computation times tend to become longer as the relative difference between the largest and smallest c_j decreases.[12] In Tables 5 and 6 the effect of holding n constant and increasing D is an almost exactly proportional increase in the "Total" times. The increase in "Total" times in Table 7 for n and $D \geqq 1000$ is somewhat less than proportional to increases in D.

TABLE 5

		$1 \leqq c_j \leqq 400$			
n	D	Total	Av.	Fast	Ratio
500	501	.148	.096	.098	.662
	1002	.270	.175	.197	.730
	1503	.421	.274	.334	.793
1000	1000	.245	.176	.154	.629
	2002	.452	.307	.382	.845
	3003	.680	.479	.577	.849
1500	1501	.439	.288	.333	.759
	3002	.888	.559	.555	.625
	4503	1.274	.838	.855	.671

TABLE 6

		$301 \leqq c_j \leqq 700$			
n	D	Total	Av.	Fast	Ratio
500	501	.739	.507	.376	.509
	1002	1.326	.843	.409	.308
	1503	2.297	1.237	.645	.281
1000	1001	2.537	1.858	1.092	.430
	2002	4.388	2.996	1.294	.295
	3003	6.663	3.972	2.081	.312
1500	1501	4.481	3.485	1.724	.385
	3002	8.790	5.565	2.252	.256
	4503	13.916	7.996	3.717	.267

[11] The numbers $D/5$, $2D/5$, $\cdots$ were of course replaced with their nearest integers.

[12] This may be due in part to a less than optimal computer subroutine for determining N_r at each iteration.

TABLE 7

		$601 \leqq c_j \leqq 1000$			
n	D	Total	Av.	Fast	Ratio
500	501	1.226	1.017	.445	.363
	1002	2.041	1.508	.509	.249
	1503	3.146	2.060	1.207	.384
1000	1001	4.749	3.999	5.393	1.136*
	2002	6.373	5.194	1.532	.240
	3003	9.592	6.738	2.567	.268
1500	1501	8.240	7.657	1.061	.129
	3002	12.006	10.286	2.389	.199
	4503	15.471	12.201	4.544	.294

TABLE 8

n	D	Total	Fast	Ratio	
50	100	.050	.020	.400	$1 \leqq c_j \leqq 40$
	150	.061	.024	.393	
100	200	.081	.035	.436	
50	100	.056	.029	.516	$31 \leqq c_j \leqq 70$
	150	.072	.033	.458	
100	200	.153	.035	.228	
50	100	.065	.013	.200	$61 \leqq c_j \leqq 100$
	150	.088	.019	.215	
100	200	.164	.044	.268	

Also, for the ranges of c_j in which the computation times are longer (Tables 6 and 7), the "Av." and "Fast" times become increasingly favorable relative to the "Total" times. The superiority of the accelerated version of the algorithm is quite evident from the fact that the "Fast" times in Tables 6 and 7 are not only better than the "Total" times, but are also considerably better than the "Av." times. An exception occurs for $n = 1000$ and $D = 1001$ in Table 7, as indicated by the asterisk beside the "Ratio" entry. The reason for this exceptional divergence from the pattern evident in the other entries is not known.

While computation times appear to increase roughly in proportion to increases in D, they do not increase in proportion to increases in n. For example, in Table 5, a proportional increase in computation time would lead one to expect the "Total" and "Fast" times for $n = 1500$ and $D = 1501$ to be roughly 1.2 and .99 seconds (multiplying the times for $n = 500$ and $D = 1503$ by three). In contrast, they are actually .439 and .333 seconds.

More dramatic examples arise by comparing the times[13] of Table 8 to those of Tables 5, 6 and 7. At first glance the Table 8 times are very small, since most of the "Total" times are under .09 seconds and most of the "Fast" times are under .04 seconds. However, if the computational time were to increase in proportion to nD, as in the group algorithm of [4], the times for "corresponding" ranges of c_j would be greater by a factor of from 4 to 40 than the times appearing in Tables 5, 6 and 7. For example, extrapolating from $n = 50$ and $D = 150$ for $1 \leqq c_j \leqq 40$ would give a "Total" time of 54.900 seconds for $n = 1500$, $D = 4503$ in Table 5, as compared to 1.274 seconds.

Such comparisons do not yield a precise formula linking computation times and increases in n and D, both because of the effect of different ranges of c_j and because of probable shortcomings of the computer code in determining $N_r = \min \{c_j\}$ and in determining the current composition of the set T.[14] Nevertheless, without attempting to be definitive, the tables do establish definite patterns in the performance of the algorithm: in particular, that the accelerated version of the algorithm is distinctly superior to the version of § 4 and that computation times for both versions increase at a considerably more favorable rate than nD.

Acknowledgment. I am indebted to Sam Ginsburg of the University of California, Berkeley, and to Professor Stanley Zionts of the State University of New York at Buffalo for suggestions that have improved the exposition of this paper.

REFERENCES

[1] P. C. GILMORE AND R. E. GOMORY, *Multistage cutting stock problems of two and more dimensions*, Operations Res., 13 (1965).

[2] FRED GLOVER, *An algorithm for solving the linear integer programming problem over a finite additive group, with extensions to solving general linear and certain non-linear integer problems*, Rep. ORC 66–29, Operations Research Center, University of California, Berkeley, 1966.

[3] R. E. GOMORY, *An algorithm for integer solutions to linear programs*, Recent Advances in Mathematical Programming, R. L. Graves and P. Wolfe, eds., McGraw-Hill, New York, 1963.

[4] ———, *On the relation between integer and noninteger solutions to linear programs*, Proc. Nat. Acad. Sci. U.S.A., 53 (1965), pp. 260–265.

[5] ———, *Faces of an integer polyhedron*, Ibid., 57 (1967), pp. 16–18.

[6] JEREMY F. SHAPIRO, *Dynamic programming algorithms for the integer programming problem. I: The integer programming problem viewed as a knapsack type problem*, Operations Res., 16 (1968), pp. 103–121.

[7] J. F. SHAPIRO AND H. M. WAGNER, *A finite renewal algorithm for the knapsack and turnpike models*, Ibid., 15 (1967), pp. 319–341.

[8] W. W. WHITE, *On a group theoretic approach to linear integer programming*, Rep. ORC 66–27, Operations Research Center, University of California, Berkeley, 1966.

[13] "Av." times were not computed for Table 8.

[14] However, the latter should probably penalize computation times more for larger problems than for smaller ones. A more sophisticated computer code is currently being designed for a sequel to this paper which extends the group algorithm to the general integer programming problem.

SIAM J. CONTROL
Vol. 7, No. 2, May 1969

GENERALIZED KUHN–TUCKER CONDITIONS FOR MATHEMATICAL PROGRAMMING PROBLEMS IN A BANACH SPACE*

MONIQUE GUIGNARD†

Abstract. Generalized Kuhn–Tucker conditions stated in this paper correspond to the optimality conditions for mathematical programming problems in a Banach space. Constraint qualifications given before can be regarded as special cases of the present constraint qualification introduced to prove the necessity. Pseudoconvexity of the constraint set rather than convexity is required for sufficiency. In case this hypothesis fails to be satisfied, second order optimality conditions are sufficient for an isolated local optimum.

1. Introduction. Optimality conditions are given for a generalized mathematical programming problem. The constraint set, defined in a Banach space similar to that in [1], was considered in a Euclidean space in [2] and is different from that in [3]. First order necessary optimality conditions stated in the first part of Theorem 2 generalize the Kuhn–Tucker conditions [4], while the constraint qualification is a substitute for all the constraint qualifications of Kuhn–Tucker [4], of Arrow, Hurwicz and Uzawa [5] and of Abadie [6]. Sufficiency is proved for objective functions either pseudoconcave [7] or quasi-concave [8], the constraint set being now taken as pseudoconvex. In case even these weakened convexity conditions fail to be satisfied, second order optimality conditions may be sufficient for an isolated local optimum. Results similar to those in [9], [10] and [11] are stated for a more general program and in a form related to first order conditions.

2. Preliminaries. For any two topological spaces S and T, $L(S, T)$ denotes the set of all continuous linear mappings from S into T. For all $s \in S$ and for all $l \in L(S, T)$, $\langle l, s\rangle$ is $l(s)$, i.e., the value of the continuous linear mapping l at s.

We consider X, a locally convex real linear topological space, and X^*, its topological dual, i.e., for E, a one-dimensional Euclidean space, $X^* = L(X, E)$. X^* is given the uniform topology. Let M be a subset in X (respectively X^*); $\overline{M}$ denotes the closure of M and $\{M\}$ is the smallest convex subset in X (respectively in X^*) containing M. $-M$ is $\{-x : x \in M\}$. Let N be another subset of X; then $M\backslash N = \{x \in X : x \in M, x \notin N\}$.

1.1. Cones. The following definitions are given.

DEFINITION 1. $C \subset X$ (respectively $C^* \subset X^*$) is a *cone* if for $x \in C$ (respectively $x^* \in C^*$), $\alpha x \in C$ (respectively $\alpha x^* \in C^*$) for all $\alpha \geqq 0$.

DEFINITION 2a. For a cone $C \subset X$, we define

$$C^- = \{u \in X^* : \langle u, x\rangle \leqq 0 \text{ for all } x \in C\},$$

$$C^+ = \{u \in X^* : \langle u, x\rangle \geqq 0 \text{ for all } x \in C\}.$$

* Received by the editors March 22, 1968, and in revised form December 30, 1968.

† Laboratoire de Calcul, Université de Lille, Lille, France.

Definition 2b. For a cone $C^* \subset X^*$, we define

$$C^{*-} = \{x \in X : \langle u, x\rangle \leqq 0 \text{ for all } u \in C^*\},$$

$$C^{*+} = \{x \in X : \langle u, x\rangle \geqq 0 \text{ for all } u \in C^*\}.$$

Remark. C^- and C^+ (respectively C^{*-} and C^{*+}) are closed convex cones in X^* (respectively in X).

The following properties are given for the defined cones. Proofs can be found in [12] for a Euclidean space or in [13] for a locally convex linear topological space.

Property C1. *Let C be a cone; then $C^{--} = \overline{\{C\}}$. In particular, if C is closed and convex, $C^{--} = C$.*

Property C2. *Let C be a cone; then $C^- = \overline{\{C\}}^-$.*

Property C3. *Let C_1 and C_2 be closed convex cones; then $(C_1 \cap C_2)^- = \overline{C_1^- + C_2^-}$.*

1.2. Cones tangent and pseudotangent to a set. We consider X, a real Banach space, M, a nonempty set in $X, \bar{x} \in M, y \in X$. The following definitions are given.

Definition 3 (cf. [6]). The vector y is *tangent* to M at $\bar{x}$ if there exist a sequence $\{x_k\}$ contained in M and converging to $\bar{x}$ and a sequence $\{\lambda_k\}$ of nonnegative numbers such that the sequence $\{\lambda_k(x_k - \bar{x})\}$ converges to y.

Definition 4 (cf. [6]). The set $T(M, \bar{x})$ of all the vectors tangent to M at $\bar{x}$ is called the *cone tangent* to M at $\bar{x}$.

Definition 5 (cf. [3]). The set $P(M, \bar{x})$, the closure of the convex hull of $T(M, \bar{x})$, is called the *cone pseudotangent* to M at $\bar{x}$.

Let I be a set of indices, not necessarily finite, $A_i \subset X$, $i \in I$. Let $A \equiv \bigcap_{i \in I} A_i \neq \varnothing$, $\bar{x} \in A$ and $\tilde{A} \equiv \bigcup_{i \in I} A_i$. The defined cones have the following properties.

Property T1 (cf. [3]). *$T(A, \bar{x}) \subset \bigcap_{i \in I} T(A_i, \bar{x})$ and $P(A, \bar{x}) \subset \bigcap_{i \in I} P(A_i, \bar{x})$.*

Proof. For all $i \in I$, $A \subset A_i$ implies $T(A, \bar{x}) \subset T(A_i, \bar{x})$. Therefore $T(A, \bar{x}) \subset \bigcap_{i \in I} T(A_i, \bar{x}) \subset \bigcap_{i \in I} P(A_i, \bar{x})$. Since the intersection of perhaps infinitely many closed convex cones is a closed convex cone, $\bigcap_{i \in I} P(A_i, \bar{x})$ contains $P(A, \bar{x})$ which is the smallest closed convex cone containing $T(A, \bar{x})$.

Property T2. *$\bigcup_{i \in I} T(A_i, \bar{x}) \subset T(\tilde{A}, \bar{x})$ and $\bigcup_{i \in I} P(A_i, \bar{x}) \subset P(\tilde{A}, \bar{x})$.*

Proof. For all $i \in I$, $A_i \subset \tilde{A}$ implies $T(A_i, \bar{x}) \subset T(\tilde{A}, \bar{x})$ which implies that $\bigcup_{i \in I} T(A_i, \bar{x}) \subset T(\tilde{A}, \bar{x})$. Moreover, $T(A_i, \bar{x}) \subset T(\tilde{A}, \bar{x}) \subset P(\tilde{A}, \bar{x})$. Since $P(\tilde{A}, \bar{x})$ is a closed convex cone, it contains $P(A_i, \bar{x})$ which is the smallest closed convex cone containing $T(A_i, \bar{x})$ for all $i \in I$. Then $\bigcup_{i \in I} P(A_i, \bar{x}) \subset P(\tilde{A}, \bar{x})$.

1.3. Pseudoconvexity and pseudoconcavity.

Definition 6 (cf. [3]). M is *pseudoconvex at* $\bar{x}$ if for all $x \in M$, $x - \bar{x} \in P(M, \bar{x})$.

Definition 7. M is *convex* if for all x and $y \in M$ and for all λ, $0 < \lambda < 1$, $\lambda x + (1 - \lambda)y \in M$.

Convex and pseudoconvex sets have the following properties.

Property PC1. *If all A_i, $i \in I$, are pseudoconvex at $\bar{x}$, then $\tilde{A}$ is pseudoconvex at $\bar{x}$.*

Proof. For all $i \in I$, A_i is pseudoconvex at $\bar{x}$, so that $x - \bar{x} \in P(A_i, \bar{x}) \subset \bigcup_{i \in I} P(A_i, \bar{x})$ for all $x \in A_i$. Then by Property T2, $x - \bar{x} \in P(\tilde{A}, \bar{x})$.

Remark 1. The intersection of several pseudoconvex sets is not necessarily pseudoconvex, as shown by the following example. E will denote a one-dimensional Euclidean space. Let $A_1 = \{x \in E : x = 1/n, n \in N\}$ and $A_2 = \{x \in E\colon x = 1$ or $x = \pi/n, n \in N\}$. Then $A = A_1 \cap A_2 = \{x \in E : x = 1$ or $x = 0\}$ is not pseudoconvex at $\bar{x} = 0$, although A_1 and A_2 are. Here, and also later when needed, N stands for the set of all nonnegative integers.

PROPERTY PC2 (cf. [3]). *If M is convex, M is pseudoconvex at $\bar{x}$ for all $\bar{x} \in M$.*

Proof. Let $x \in M$, and let $\{\lambda_k\}$ be a sequence of positive numbers, $0 < \lambda_k \leqq 1$, converging to 0. Then there exist a sequence $\{\mu_k\}$ of nonnegative numbers $\mu_k = 1/\lambda_k$ and a sequence $\{x_k\}$ contained in M and converging to $\bar{x}$: $x_k = \bar{x} + \lambda_k(x - \bar{x})$, such that the sequence $\mu_k(x_k - \bar{x}) = \mu_k\ \lambda_k(x - \bar{x}) = x - \bar{x}$ converges to $x - \bar{x}$. Hence, $x - \bar{x}$ is a vector tangent to M at $\bar{x}$, $x - \bar{x} \in T(M, \bar{x}) \subset P(M, \bar{x})$ for all $x \in M$, and M is pseudoconvex at $\bar{x}$.

Remark 2. From Property PC1 and Property PC2, the union of several convex sets the intersection of which is not empty is pseudoconvex at any point of this intersection. This is an example of a pseudoconvex set which is not necessarily convex.

Let $\psi(x)$ be a real function of $x \in X$.

DEFINITION 8 (cf. [8]). ψ is *quasi-concave* if for all $\lambda \in E$, $\{x \in X : \psi(x) \geqq \lambda\}$ is convex.

Quasi-concave differentiable functions have the following property. ∇ will denote the differential operator.

PROPERTY PC3. *If ψ is quasi-concave and Fréchet-differentiable at $\bar{x}$, then $\langle \nabla\psi(\bar{x}), x - \bar{x}\rangle < 0$ implies $\psi(x) - \psi(\bar{x}) < 0$.*

The proof of this statement can be found in [8].

DEFINITION 9 (cf. [3] and [7]).ψ is *pseudoconcave* over M at $\bar{x}$ if ψ is Fréchet-differentiable at $\bar{x}$ and if $x \in M$, $\langle \nabla\psi(\bar{x}), x - \bar{x}\rangle \leqq 0$ implies $\psi(x) - \psi(\bar{x}) \leqq 0$.

3. First order optimality conditions. Let X be a real Banach space, A a nonempty subset in X, $\bar{x} \in A$ and $\psi(x)$ a real-valued function of $x \in X$, Fréchet-differentiable at $\bar{x}$. Consider the problem: maximize $\{\psi(x) : x \in A\}$.

THEOREM 1. *A necessary condition for $\bar{x}$ to maximize ψ over A is that $\nabla\psi(\bar{x}) \in P^-(A, \bar{x})$. It is also sufficient if ψ is pseudoconcave over A at $\bar{x}$ and A is pseudoconvex at $\bar{x}$.*

Proof. We first give the necessity proof. Let $y \in T(A, \bar{x})$. Then there exist a sequence $\{x_k\}$, $x_k \in A$ for all k, $\lim_{k\to\infty} x_k = \bar{x}$, and a sequence $\{\lambda_k\}$, $\lambda_k > 0$ for all k, such that $\lim_{k\to\infty} \lambda_k(x_k - x) = y$. If $\bar{x}$ maximizes ψ over A, $\psi(x_k) - \psi(\bar{x}) \leqq 0$ for all k. Moreover, $\psi(x_k) - \psi(\bar{x}) = \langle \nabla\psi(\bar{x}), x_k - \bar{x}\rangle + o(\|x_k - \bar{x}\|)$. Then

$$\langle \nabla\psi(\bar{x}), \lambda_k(x_k - \bar{x})\rangle \leqq \frac{-o(\|x_k - \bar{x}\|)}{\|x_k - \bar{x}\|} \cdot \lambda_k\|x_k - \bar{x}\|.$$

Let k go to infinity; then $\langle \nabla\psi(\bar{x}), y\rangle \leqq 0 \cdot \|y\| = 0$. Therefore, $\nabla\psi(\bar{x}) \in T^-(A, \bar{x}) = P^-(A, \bar{x})$ by Property C2.

The sufficiency proof is as follows. Let $x \in A$. Since A is pseudoconvex at $\bar{x}$, we have $x - \bar{x} \in P(A, \bar{x})$ and $\langle \nabla\psi(\bar{x}), x - \bar{x}\rangle \leqq 0$. But since ψ is pseudoconcave over A at $\bar{x}$, this yields $\psi(x) - \psi(\bar{x}) \leqq 0$, and $\bar{x}$ maximizes ψ over A. This completes the proof.

Let Y be another real Banach space and $a : X \to Y$ a map. Let B and C be nonempty subsets in Y and X, respectively, and assume that $A = \{x \in C : a(x) \in B\}$ is not void. We may rewrite the problem: maximize $\{\psi(x) : x \in C, a(x) \in B\}$. Suppose that $a(x)$ is Fréchet-differentiable at $\bar{x} \in A$. Let $K = \{y \in X : \langle \nabla a(\bar{x}), y\rangle \in P[B, a(\bar{x})]\}$, $H = \{h \in X^* : h = u \cdot \nabla a(\bar{x}), u \in P^-[B, a(\bar{x})]\}$.

THEOREM 2 (The generalized Kuhn–Tucker conditions). *If H is closed and G is a closed convex cone in X such that $K \cap G = P(A, \bar{x})$ and $K^- + G^-$ is closed, then a necessary condition for $\bar{x}$ to maximize ψ over A is that there exists $u \in P^+[B, a(\bar{x})]$ such that $\nabla\psi(\bar{x}) + u \cdot \nabla a(\bar{x}) \in G^-$. This condition is also sufficient if ψ is continuous, and if G is a closed convex cone in X such that $x - \bar{x} \in G$ for all $x \in A$, if A or $\Delta = \{x \in X : a(x) \in B\}$ is pseudoconvex at $\bar{x}$, and if either ψ is pseudoconcave over A at $\bar{x}$, or quasi-concave with $\nabla\psi(\bar{x}) \neq 0$.*

Remark 3. A sufficient condition for H to be closed is that the map $\nabla a(\bar{x}) : X \to Y$ have closed range (cf. [13]).

Proof. Necessity. If $\bar{x}$ maximizes ψ over A, by Theorem 1 we have $\nabla\psi(\bar{x}) \in P^-(A, \bar{x})$. Since $K^- + G^-$ is closed, by Property C3 we have $P^-(A, \bar{x}) = K^- + G^-$. Then there exists $k \in K^+$ such that $\nabla\psi(\bar{x}) + k \in G^-$. Let $v \in H^-$; then $\langle u \cdot \nabla a(\bar{x}), v\rangle \leqq 0$ for all $u \in P^-[B, a(\bar{x})]$. Suppose that $\nabla a(\bar{x})(v) \notin P[B, a(\bar{x})]$. By the strong separation theorem, and since $P[B, a(\bar{x})]$ is a cone, there exists $y \in Y^*$ such that $\langle y, \nabla a(\bar{x})(v)\rangle > 0 \geqq \langle y, w\rangle$ for all $w \in P[B, a(\bar{x})]$. Therefore, $y \in P^-[B, a(\bar{x})]$ and $y \cdot \nabla a(\bar{x}) \in H$, which contradicts $\langle y \cdot \nabla a(\bar{x}), v\rangle = \langle y, \nabla a(\bar{x})(v)\rangle > 0$. Then $\nabla a(\bar{x})(v) \in P[B, a(\bar{x})]$, i.e., $v \in K$ for all $v \in H^-$. Since H and K are closed convex cones, $H^- \subset K$ yields $H \supset K^-$. Therefore, there exists $u \in P^+[B, a(\bar{x})]$ such that $\nabla\psi(\bar{x}) + u \cdot \nabla a(\bar{x}) \in G^-$.

For sufficiency, we first prove the following lemma.

LEMMA. *If G is a closed convex cone such that $x - \bar{x} \in G$ for all $x \in A$, if either A or Δ is pseudoconvex at $\bar{x}$, if there exists $u \in P^+[B, a(\bar{x})]$ such that $\nabla\psi(\bar{x}) + u \cdot \nabla a(\bar{x}) \in G^-$, then for all $x \in A$, $\langle \nabla\psi(\bar{x}), x - \bar{x}\rangle \leqq 0$.*

Proof. For all $x \in A$, $x - \bar{x} \in G$. Therefore,

$$(1) \qquad \langle \nabla\psi(\bar{x}), x - \bar{x}\rangle \leqq \langle -u \cdot \nabla a(\bar{x}), x - \bar{x}\rangle.$$

If Δ is pseudoconvex at $\bar{x}$, $x - \bar{x} \in P(\Delta, \bar{x})$ for all $x \in A \subset \Delta$. If A is pseudoconvex at $\bar{x}$, $x - \bar{x} \in P(A, \bar{x}) \subset P(\Delta, \bar{x})$ for all $x \in A$. In both cases, $\langle \nabla a(\bar{x}), y\rangle \in T[B, a(\bar{x})]$ for all $y \in T(\Delta, \bar{x})$. By continuity and convexity, since $\nabla a(\bar{x})$ is a continuous linear map, $\langle \nabla a(\bar{x}), y\rangle \in P[B, a(\bar{x})]$ for all $y \in P(\Delta, \bar{x})$, and, in particular, for all $x \in A$. Moreover, $u \in P^+[B, a(\bar{x})]$, so that $\langle -u \cdot \nabla a(\bar{x}), x - \bar{x}\rangle \leqq 0$ for all $x \in A$, and by (1), $\langle \nabla\psi(\bar{x}), x - \bar{x}\rangle \leqq 0$ for all $x \in A$. This completes the proof.

Sufficiency. If ψ is pseudoconcave over A at $\bar{x}$, by the lemma, for all $x \in A$, $\langle \nabla\psi(\bar{x}), x - \bar{x}\rangle \leqq 0$; that is, $\psi(x) - \psi(\bar{x}) \leqq 0$.

If $\nabla\psi(\bar{x}) \neq 0$, there exists $x_1 \in X$ such that $\langle \nabla\psi(\bar{x}), x_1 - \bar{x}\rangle < 0$. Suppose indeed that $\langle \nabla\psi(\bar{x}), x - \bar{x}\rangle \geqq 0$ for all $x \in X$. Then let $x' \in X$ and $x'' = 2\bar{x} - x'$.

Then $0 \leqq \langle \nabla\psi(\bar{x}), x'' - x' \rangle = -\langle \nabla\psi(\bar{x}), x' - \bar{x} \rangle \leqq 0$, which implies $\langle \nabla\psi(\bar{x}), x' - \bar{x} \rangle = 0$ for all $x' \in X$, i.e., $\nabla\psi(\bar{x}) = 0$. Let $x \in A$. We define

$$x(\theta) = x + \theta(x_1 - x),$$

$$\bar{x}(\theta) = \bar{x} + \theta(x_1 - \bar{x})$$

for all $\theta > 0$, $\langle \nabla\psi(\bar{x}), \bar{x}(\theta) - \bar{x} \rangle = \theta\langle \nabla\psi(\bar{x}), x_1 - \bar{x} \rangle < 0$. For all $\theta \in [0, 1]$, $\langle \nabla\psi(\bar{x}), x(\theta) - \bar{x}(\theta) \rangle = (1 - \theta)\langle \nabla\psi(\bar{x}), x - \bar{x} \rangle \leqq 0$ by the lemma. Then for all θ, $0 < \theta \leqq 1$, $\langle \nabla\psi(\bar{x}), x(\theta) - \bar{x} \rangle < 0$. By the quasi-concavity of the function $\psi(x)$, we have (cf. Property PC3) $\psi(x(\theta)) < \psi(\bar{x})$, and $\theta \to 0$ implies $\psi(x) < \psi(\bar{x})$ for all $x \in A$.

4. Second order optimality conditions. Let us now consider the second order optimality conditions. First, additional notations required for the discussion will be introduced. If $l: X \to Y$ is Fréchet-differentiable at $\bar{x} \in X$, $\langle \nabla l(\bar{x}), x \rangle$ denotes the value of the mapping $\nabla l(\bar{x})$ at x. If l is twice continuously differentiable at $\bar{x}$, $\nabla^2 l(\bar{x})$ is an element of $L(X, L(X, Y))$ which can be identified with $L(X^2, Y)$ [14, p. 174]. We shall denote by $\langle \nabla^2 l(\bar{x}), (x, y) \rangle$ or $\langle\langle \nabla^2 l(\bar{x}), x \rangle, y \rangle$ the value of the mapping $\nabla^2 l(\bar{x})$ at $(x, y) \in X \times X$.

We suppose now that X is finite-dimensional Finite-dimensionality is a necessary and sufficient condition that the unit sphere $\{x \in X : \|x\| \leqq 1\}$ be compact in a Banach space (cf. [15, p. 85]). The second order conditions given next are sufficient for an isolated local optimum.

THEOREM 3. *If* (i) *ψ and a are twice continuously differentiable at $\bar{x}$,* (ii) *G is a closed convex cone in X,* (iii) *in a neighborhood of $\bar{x}$, $x \in A$ implies $x - \bar{x} \in G$,* (iv) *in a neighborhood of $a(\bar{x})$, $y \in B$ implies $y - a(\bar{x}) \in P[B, a(\bar{x})]$, then a sufficient condition that $\bar{x}$ be an isolated local optimum for ψ over A is that there exists $u \in P^+[B, a(\bar{x})]$ such that $\nabla\psi(\bar{x}) + u \cdot \nabla a(\bar{x}) \in G^-$ and for all nontrivial $h \in X$ such that $\langle \nabla\psi(\bar{x}), h \rangle = 0$ and $\langle \nabla a(\bar{x}), h \rangle \in -P[B, a(\bar{x})] \cap P[B, a(\bar{x})]$ it follows that $\langle \nabla^2\psi(\bar{x}) + u \cdot \nabla^2 a(\bar{x}), (h, h) \rangle < 0$.*

Proof. Suppose that $\bar{x}$ is not an isolated local maximum for ψ over A. Then there exists a sequence $\{x_k\}$, $x_k \in A$, $x_k \neq \bar{x}$ for all k, such that $\lim_{k\to\infty} x_k = \bar{x}$ and $\psi(x_k) \geqq \psi(\bar{x})$ for all k. Moreover, since the unit sphere is compact, we may assume that

$$\lim_{k\to\infty} \frac{x_k - \bar{x}}{\|x_k - \bar{x}\|} = h \neq 0.$$

Note, then, that $h \in T(A, \bar{x})$. In a neighborhood of $\bar{x}$, $x_k \in A$ implies $x_k - \bar{x} \in G$. Then $\nabla\psi(\bar{x}) + u \cdot \nabla a(\bar{x}) \in G^-$ yields $\langle \nabla\psi(\bar{x}) + u \cdot \nabla a(\bar{x}), x_k - \bar{x} \rangle \leqq 0$. Therefore, $\langle \nabla\psi(\bar{x}) + u \cdot \nabla a(\bar{x}), h \rangle \leqq 0$. Since $h \in T(A, \bar{x}) \subset T(\Delta, \bar{x})$, $\langle \nabla a(\bar{x}) \cdot h \rangle \in P[B, a(\bar{x})]$. We have two cases to consider:

(i) Suppose that $\langle \nabla\psi(\bar{x}), h \rangle < 0$. Then there exists a positive number N such that for all $K \geqq N$, $\psi(x_k) - \psi(\bar{x}) < 0$. Since this is a contradiction, $\langle \nabla\psi(\bar{x}), h \rangle \geqq 0$.

(ii) Suppose that

$$\langle \nabla a(\bar{x}), h\rangle \in P[B, a(\bar{x})] \backslash -P[B, a(\bar{x})].$$

Then $\langle \nabla\psi(\bar{x}), h\rangle \leqq -\langle u \cdot \nabla a(\bar{x}), h\rangle < 0$ since $u \in P^+[B, a(\bar{x})]$. By (i) this is an impossibility. Therefore, $\langle \nabla a(\bar{x}), h\rangle \in -P[B, a(\bar{x})] \cap P[B, a(\bar{x})]$, i.e., $\langle u \cdot \nabla a(\bar{x}), h\rangle = 0$ and $\langle \nabla\psi(\bar{x}), h\rangle = 0$.

Let us define $\xi(x) = \psi(x) + u \cdot a(x)$. Then $\langle \nabla^2\xi(\bar{x}), (h, h)\rangle < 0$. Since $\langle \nabla\xi(\bar{x}), x_k - \bar{x}\rangle \leqq 0$,

$$\xi(x_k) - \xi(\bar{x}) = \langle \nabla\xi(\bar{x}), x_k - \bar{x}\rangle + \tfrac{1}{2}\langle \nabla^2\xi(\bar{x}), (x_k - \bar{x}, x_k - \bar{x})\rangle + (o\|x_k - \bar{x}\|^2).$$

We have that

$$\lim_{k\to\infty} \frac{\xi(x_k) - \xi(\bar{x})}{\|x_k - \bar{x}\|^2} \leqq \tfrac{1}{2}\langle \nabla^2\xi(\bar{x}), (h, h)\rangle < 0$$

and there exists a positive integer N' such that for all $k \geqq N'$, $\xi(x_k) - \xi(\bar{x}) < 0$. But in a neighborhood of $a(\bar{x}), a(x_k) \in B$ implies $a(x_k) - a(\bar{x}) \in P[B, a(\bar{x})]$. Then $-\langle u, a(x_k) - a(\bar{x})\rangle \leqq 0$ and $\psi(x_k) + u \cdot a(x_k) < \psi(\bar{x}) + u \cdot a(\bar{x})$, i.e., $\psi(x_k) < \psi(\bar{x}) - \langle u, a(x_k) - a(\bar{x})\rangle \leqq \psi(\bar{x})$, which is impossible. Hence, such a sequence $\{x_k\}$ does not exist and $\bar{x}$ is an isolated local maximum for ψ over A.

It is hoped that one can apply these results to the theory of optimal control. Constraint qualification introduced here ensures that the "multiplier" associated with the objective function, which is encountered in most of the earlier papers dealing with a maximum principle, is positive. It is more general than the ones in [1] or [13] which are the same as in [5] for a more general problem in a more general space. In the following section, it is shown how these optimality conditions apply to mathematical programming problems.

5. Application to mathematical programming. Let X be an n-dimensional Euclidean space E^n and Y an m-dimensional Euclidean space $E^m \cdot E^r_+$ will denote $\{x \in E^r : x \geqq 0\}$, $r = m, n$. Two examples will be discussed.

Example 1. If $B = E^m_+$, the problem becomes: maximize $\{\psi(x) : a_i(x) \geqq 0, i = 1, \ldots, m, x \in C\}$. If $\bar{x} \in A = \{x \in C : a(x) \geqq 0\}$, let I and $\bar{I}$ be such that $a_j(\bar{x}) = 0$ for all $j \in I$ and $a_j(\bar{x}) > 0$ for all $j \in \bar{I}$. Then

$$P[B, a(\bar{x})] = \{u \in (E^m)^* : u \geqq 0, u \cdot a(\bar{x}) = 0\} = \{u \in (E^m)^* : u^j \geqq 0, j \in I, u^j = 0, j \in \bar{I}\},$$

$$K = \{y \in E^n : \langle \nabla a_j(\bar{x}), y\rangle \geqq 0, j \in I\},$$

$$H = \{h \in (E^n)^* : h = -\sum_{j \in I} u^j \cdot \nabla a_j(\bar{x}), u^j \geqq 0, j \in I\}.$$

Notice that both K and H are closed convex cones. G must be a closed convex cone such that $K \cap G = P(A, \bar{x})$ and $K^- + G^-$ is closed. We shall call this hypothesis, imposed upon G, the hypothesis H(G).

If $C = E^n$ (respectively E^n_+), and if this hypothesis H(G) is satisfied with $G = P(C, \bar{x})$, we obtain the usual Kuhn–Tucker conditions, since

$P^-(E^n, \bar{x}) = \{0\} \subset (E^n)^*$ and $P^-(E^n, \bar{x}) = \{y \in (E^n)^* : y \leqq 0, \langle y, \bar{x} \rangle = 0\}$. In both cases, since K and G are polyhedral cones, $K^- + G^-$ is closed [16, p. 388], so that H(G) reduces to $K \cap P(C, \bar{x}) = P(A, \bar{x})$, and this constraint qualification is actually weaker than those in [4], [5] and [6].

Kuhn and Tucker [4] defined

$$C(A, \bar{x}) = \{y \in E^n : \exists \xi : E_+ \to E^n, \xi(0) = \bar{x}, \xi'(+0) = y, \xi(\theta) \in A \ \forall \theta \in [0, 1]\},$$

and their constraint qualification was $K \cap P(C, \bar{x}) = C(A, \bar{x})$.

Arrow, Hurwicz and Uzawa [5] weakened this assumption, noting that the left-hand side was always closed and convex, and their constraint qualification was $K \cap P(C, \bar{x}) = \overline{\{C(A, \bar{x})\}}$.

Abadie [6] weakened the Kuhn–Tucker constraint qualification in a different way, assuming that $K \cap P(C, \bar{x}) = T(A, \bar{x})$.

These authors did not refer to $P(C, \bar{x})$ explicitly, however; when necessary they introduced the constraints $x \geqq 0$ as components of $a(x) \geqq 0$. (See Table 1.)

TABLE 1

Case $C = E^n$		Case $C = E^n_+$	
Problem	Optimality conditions	Problem	Optimality conditions
max $\psi(x)$ $a(x) \geqq 0$ $x \in E^n$	$\exists u \geqq 0, u \in (E^m)^*$ $\nabla\psi(\bar{x}) + u \cdot \nabla a(\bar{x}) = 0$ $u \cdot a(\bar{x}) = 0$	max $\psi(x)$ $a(x) \geqq 0$ $x \geqq 0$	$\exists u \geqq 0, u \in (E^m)^*$ $W = \nabla\psi(\bar{x}) + u \cdot \nabla a(\bar{x}) \leqq 0$ $\langle W, \bar{x} \rangle = 0$ $u \cdot a(\bar{x}) = 0$

Suppose now that H(G) is not satisfied for $G = P(C, \bar{x})$, but is satisfied for a certain closed convex cone G. We then obtain optimality conditions that could not have been written otherwise, as is shown by the following well-known example. Consider the problem: maximize $\{\psi(x_1, x_2) : x_1^3 - x_2 \geqq 0, x_2 \geqq 0\}$. Suppose $\bar{x} = (0, 0)$. Then $G = E_+ \times E$ is such that $K \cap G = P(A, \bar{x}) = \{(y_1, y_2) : y_1 \geqq 0, y_2 = 0\}$. If $\bar{x} = (0, 0)$ maximizes ψ over A, there exists $u \geqq 0$ such that $\nabla\psi(\bar{x}) + u \cdot \nabla a(\bar{x}) \in G^-$; that is,

$$\frac{\partial\psi(\bar{x})}{\partial x_1} + u \cdot \frac{\partial a(\bar{x})}{\partial x_1} \leqq 0,$$

$$\frac{\partial\psi(\bar{x})}{\partial x_2} + u \cdot \frac{\partial a(\bar{x})}{\partial x_2} = 0$$

and $u \cdot a(\bar{x}) = 0$.

In the following, Z will denote the set of all integers and N the set of all nonnegative integers.

If $C = Z^n$ or N^n, the hypothesis H(G) is satisfied with $G = \widetilde{K}^-$, i.e., the subset in E^n which is isomorphic to $K^- \subset (E^n)^*$. Then $K \cap G = K \cap \widetilde{K}^- = \{0\}$

$= P(A, \bar{x})$. Suppose, indeed, that $x \in K \cap \tilde{K}^-$; then $x \in \tilde{K}^-$ implies that $\langle x, y \rangle \leqq 0$ for all $y \in K$, and, in particular, for x; therefore $\langle x, x \rangle = 0$, i.e., $x = 0$. Then, since $\tilde{K}$ is the subset in $(E^n)^*$ which is isomorphic to $K \subset E^n$, we obtain the optimality conditions given in Table 2.

TABLE 2

Problem	Optimality conditions
$\max \psi(x)$	
$a(x) \geqq 0$	$\exists u \geqq 0$
$x \in Z^n$ or N^n	$\nabla\psi(\bar{x}) + u \cdot \nabla a(\bar{x}) \in \tilde{K}$
	$u \cdot a(\bar{x}) = 0$

Example 2. If $B = \{0\} \times E^r$, where $\{0\} \subset E^{m-r}$, let $J = \{1, \cdots, m - r\}$ and $\bar{J} = \{m - r + 1, \cdots, m\}$. Then the problem becomes: maximize $\{\psi(x) : a_j(x) = 0, j \in J, a_j(x) \geqq 0, j \in \bar{J}, x \in C\}$. Let $\bar{x} \in A = \{x \in C : a_j(\bar{x}) = 0, j \in J, a_j(\bar{x}) \geqq 0, j \in \bar{J}\}$; then $P[B, a(\bar{x})] = (E^{m-r})^* \times (E^r_+)^*$, and we obtain the optimality conditions given in Table 3.

TABLE 3

Problem	Optimality conditions
$\max \psi(x)$	
$a_j(x) \geqq 0, \quad j \in \bar{J}$	$\exists u^j \geqq 0, j \in \bar{J}$
$a_j(x) = 0, \quad j \in J$	$\exists u^j, \quad j \in J$
$x \in C$	$\nabla\psi(\bar{x}) + \sum_{j \in J \cup \bar{J}} u^j \cdot \nabla a_j(\bar{x}) \in G^-$

Note that even if the $a_j, j \in J$, are nonlinear, these conditions may be sufficient, since A or Δ need only be pseudoconvex at $\bar{x}$. But sufficiency may be derived in another way.

If sgn (u) is an element of $(E^m)^*$ such that

$$[\operatorname{sgn}(u)]^i = \begin{cases} 1 & \text{if } u^i > 0, \quad i \in J \cup \bar{J}, \\ -1 & \text{if } u^i < 0, \quad i \in J, \\ 0, 1 \text{ or } -1 & \text{if } u^i = 0, \quad i \in J \cup \bar{J}, \end{cases}$$

then let $D = \{x \in E^n : [\operatorname{sgn}(u)]^i \cdot a_i(x) \geqq 0, \ i \in J \cup \bar{J}\} \supset \Delta$. If $C \cap D \subset \bar{x} + G$, if either $C \cap D$ or D is pseudoconvex at $\bar{x}$, if either ψ is pseudoconcave over $C \cap D$ at $\bar{x}$ or quasi-concave with $\nabla\psi(\bar{x}) \neq 0$, then $\bar{x}$ maximizes ψ over $C \cap D$, and a fortiori over A.

Remark 4. A sufficient, but not necessary, condition that D be pseudoconvex at $\bar{x}$ is that $[\operatorname{sgn}(u)]^i \, a_i(x)$ is quasi-convex for all $i \in J \cup \bar{J}$.

Remark 5. We may point out another consequence of this statement. If the previous hypothesis is satisfied with $[\operatorname{sgn}(u)]^i = 0$, $i \in L \subset J \cup \bar{J}$, $L = \{i : u^i = 0\}$

$\neq \emptyset$, no assumption is required upon a_i, $i \in L$. So it is possible that ineffective constraints play no role for sufficiency of optimality conditions. However, they may be used in order to make D or $D \cap C$ pseudoconvex at $\bar{x}$.

The second order optimality conditions are given in Table 4.

TABLE 4

Problem	Second order optimality conditions
$\max \psi(x)$	
$a_j(x) \geqq 0, \quad j \in \bar{J}$	$\exists u^j \geqq 0, \quad j \in \bar{J}$
$a_j(x) = 0, \quad j \in J$	$\exists u^j, \quad j \in J$
$x \in C$	$\nabla\psi(\bar{x}) + \sum_{j \in J \cup \bar{J}} u^j \cdot \nabla a_j(\bar{x}) \in G^-$
	$\left.\begin{array}{l} h \neq 0, \\ \langle \nabla\psi(\bar{x}), h \rangle = 0, \\ \langle \nabla a_i(\bar{x}), h \rangle = 0 \text{ for all } i \in J \cup \{j \in \bar{J} : u^j > 0\} \end{array}\right\}$
	$\Rightarrow \langle \nabla^2\psi(\bar{x}) + u \cdot \nabla^2 a(\bar{x}), (h, h) \rangle < 0.$

Acknowledgments. I wish to acknowledge the very useful discussions I had with Professor P. Huard on the subject of this paper when preparing my thesis at the University of Lille. I also wish to thank Dr. O. Gurel and Dr. K. Spielberg for their helpful comments during the writing of this paper.

REFERENCES

[1] P. P. VARAIYA, *Nonlinear programming in Banach space*, SIAM J. Appl. Math., 15 (1967), pp. 284–293.

[2] M. GUIGNARD, *On the Kuhn–Tucker theory*, Sixth Symposium on Mathematical Programming, Princeton University, Princeton, August 14–18, 1967.

[3] ———, *Conditions d'optimalité et dualité en programmation mathématique*, Thèse de Doctorat de Spécialité, Université de Lille, Laboratoire de Calcul, 1967.

[4] H. W. KUHN AND A. W. TUCKER, *Nonlinear programming*, Proc. Second Berkeley Symposium on Mathematical Statistics and Probability, J. Neyman, ed., University of California Press, Berkeley, 1951, pp. 481–492.

[5] K. J. ARROW, L. HURWICZ AND H. UZAWA, *Constraint qualification in maximization problems*, Naval Res. Logist. Quart., 8 (1961), pp. 175–191.

[6] J. M. ABADIE, *Problèmes d'optimisation*, Institut Blaise Pascal, Paris, 1965.

[7] O. L. MANGASARIAN, *Pseudo-convex functions*, SIAM J. Control, 3 (1965), pp. 281–290.

[8] K. J. ARROW AND A. C. ENTHOVEN, *Quasi-concave programming*, Econometrica, 29 (1961), pp. 779–800.

[9] R. PALLU DE LA BARRIÈRE, *Compléments à la théorie des multiplicateurs en programmation non linéaire*, Rev. Française Recherche Opér., 7 (1963), pp. 163–180.

[10] G. P. MCCORMICK, *Second order conditions for constrained minima*, SIAM J. Appl. Math., 15 (1967), pp. 641–652.

[11] D. R. RICE AND M. E. THOMAS, *Sufficiency conditions in nonlinear programming*, Working paper, College of Engineering, University of Florida, 1967.

[12] W. FENCHEL, *Convex cones, sets, and functions*, Hectographed, Princeton University, Princeton, 1953.

[13] P. P. VARAIYA, *Nonlinear programming and optimal control*, ERL Tech. Memo. M-129, University of California, Berkeley, 1965.

[14] J. DIEUDONNÉ, *Foundation of Modern Analysis*, Academic Press, New York, 1960.

[15] K. YOSIDA, *Functional Analysis*, Springer-Verlag, New York, 1965.

[16] M. SIMONNARD, *Programmation linéaire*, Dunod, Paris, 1962.

[17] D. L. RUSSEL, *The Kuhn–Tucker conditions in Banach space with an application to control theory*, J. Math. Anal. Appl., 15 (1966), pp. 200–212.

[18] A. M. RUBINOV, *Necessary conditions for an extreme value and their use in the study of certain equations*, Soviet Math. Dokl., 7 (1966), pp. 978–980.

[19] M. ALTMAN, *Stationary point in non-linear programming*, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys., 12 (1964), pp. 29–35.

[20] K. RITTER, *Duality for nonlinear programming in a Banach space*, SIAM J. Appl. Math., 15 (1967), pp. 294–302.

[21] H. HALKIN, *An abstract framework for the theory of process optimization*, Bull. Amer. Math. Soc., 72 (1966), pp. 677–678.

[22] H. HALKIN AND L. W. NEUSTADT, *General necessary conditions for optimization problems*, Proc. Nat. Acad. Sci. U.S.A., 56 (1966), pp. 1066–1071.

[23] H. HALKIN, *Nonlinear nonconvex programming in an infinite dimensional space*, Mathematical Theory of Control, A. V. Balakrishnan, ed., Academic Press, New York, 1967, pp. 10–25.

[24] A. YA. DUBOVITSKII AND A. A. MILYUTIN, *Extremum problems in the presence of constraints*, Zh. Vychisl. Mat. i Mat. Fiz., 5 (1965), pp. 395–453.

[25] L. W. NEUSTADT, *An abstract variational theory with applications to a broad class of optimization problems. I. General theory*; *II. Applications*, this Journal, 4 (1966), pp. 505–527; 5 (1967), pp. 90–137.

[26] B. N. PSCHENICHNIY, *Convex programming in a normed space*, Kibernetika, 5 (1965), pp. 46–54.

[27] A. V. FIACCO, *Second order sufficient conditions for weak and strict constrained minima*, SIAM J. Appl. Math., 16 (1968), pp. 105–108.

[28] P. HUARD, *Mathématiques des programmes économiques*, Monographies de Recherche Opérationnelle, Dunod, Paris, 1964.

[29] C. RAFFIN, *Programmes linéaires d'appui d'un programme convexe, application aux conditions d'optimalité et à la dualité*, Rev. Française Recherche Opér., 13 (1968), pp. 27–60.

[30] M. CANON, C. CULLUM AND E. POLAK, *Discrete Optimal Control*, McGraw-Hill, New York, 1968.

SIAM J. CONTROL
Vol. 7, No. 2, May 1969

ON THE τ-DECOMPOSITION METHOD OF STABILITY ANALYSIS FOR RETARDED DYNAMICAL SYSTEMS*

MUO S. LEE† AND C. S. HSU‡

Abstract. This paper deals with a method of studying the effects of the amount of time delay on the stability of dynamical systems. The method of analysis involves first decomposing the positive time delay axis into many intervals within each of which the stability character remains unchanged. For the change of stability character from interval to interval the result is expressed geometrically in terms of leaving or entering of a unit circle by a curve which is determined by a rational function. The paper extends the work of Krall and extends and modifies the earlier work of Sokolov and Miasnikov.

1. Introduction. When one studies the stability of dynamical systems involving time lags [1], [2], [3], one faces, in practice, a variety of different types of problems. For instance, sometimes one may be merely interested in determining the stability of a specific system with given time lags and with given values for all the system parameters. Other times one may wish to study a family of systems with the time lags specifically given but with some of the system parameters undetermined, and one wants to investigate the dependence of the stability character of the systems upon these parameters. There is yet another class of problems which has to do with investigating the dependence of the stability character of a system upon the magnitudes of the time lags while all other system parameters are taken to be fixed. In this paper we shall deal with some problems in this class. The work is motivated by that of Sokolov and Miasnikov [4] and that of Krall [5].

Of various methods of stability analysis available for dynamical systems with time lags, the one due to Pontryagin [6] is probably one of the most general. However, when one must obtain specific criteria of stability or instability for practical purposes, the theorems of Pontryagin are very difficult to apply. Except for very simple systems [7], [8] the method has not found wide applications. For the class of problems we are interested in here, the direct application of a generalized Nyquist criterion is also not practical because numerous diagrams must be constructed. A more interesting method is one due to Sokolov and Miasnikov discussed in [4, pp. 421–426] and referred to in [5]. This method may be appropriately called a τ-decomposition method. It involves first decomposing the time lag τ-axis into intervals such that within each interval the same stability character prevails, and then investigating the change of stability character of the system as the boundary points of the intervals are crossed. The method is a highly practical one. The treatment given in [4] is, however, very incomplete and

* Received by the editors April 16, 1968, and in revised form January 6, 1969.

† East Fishkill Facility, IBM Corporation, Hopewell Junction, New York 12533.

‡ Division of Applied Mechanics, University of California, Berkeley, California 94720. This work was supported in part by the National Science Foundation under Grant GK-781.

contains some explicit assertions which can be shown to be not valid in general and never valid for very large time lag τ. It is the purpose of this paper to carry out a more complete analysis of the τ-decomposition method by which a simple geometrical criterion is established and the deficiencies in the discussions given in [4] are removed. The development still follows the approach of Sokolov and Miasnikov, but the complete analysis will also cover cases which can be encountered in applying the τ-decomposition method but have not been treated previously.

2. The exponential polynomial. Consider an nth order dynamical system with a constant time retardation τ for which the governing equation may be written in the form

$$\sum_{i=0}^{n} a_i \frac{d^i x(t)}{dt^i} + \sum_{i=0}^{m} b_i \frac{d^i x(t-\tau)}{dt^i} = 0, \tag{1}$$

where a_i and b_i are constants and $a_n \neq 0$. The dynamical system is assumed to be of retarded type [2]; hence, $n > m$ and τ nonnegative. The characteristic equation of (1) is

$$\varphi(z) \triangleq g(z)e^{\tau z} + h(z) = 0, \tag{2}$$

where

$$g(z) = \sum_{i=0}^{n} a_i z^i, \qquad h(z) = \sum_{i=0}^{m} b_i z^i. \tag{3}$$

The function $\varphi(z)$ of (2) will be called an exponential polynomial in this paper, although it is also called a characteristic quasi-polynomial in the literature.

For systems (1) the following theorems of stability are known [1], [2].

THEOREM 1. *If all zeros of $\varphi(z)$ have negative real parts, then the trivial solution of (1) is asymptotically stable.*

THEOREM 2. *If at least one zero of $\varphi(z)$ has a positive real part, then the trivial solution of (1) is unstable.*

THEOREM 3. *If $\varphi(z)$ has simple purely imaginary zeros and the remaining zeros have negative real parts, then the trivial solution of (1) is stable.*

THEOREM 4. *If there is a multiple zero among the purely imaginary zeros, then the trivial solution of (1) is unstable.*

From these theorems one sees that the test of negativeness of the real parts of all the zeros of $\varphi(z)$ is all important in studying the stability of systems (1) via characteristic equations.

Before proceeding further, some nonessential restrictions will be placed upon $\varphi(z)$. $\varphi(z)$ may be classified into three categories, namely:

Category I. $g(z)$ and $h(z)$ have common purely imaginary zeros.

Category II. $z = 0$ is a zero of $\varphi(z)$.

Category III. $g(z)$ and $h(z)$ have no common purely imaginary zeros and $z = 0$ is not a zero of $\varphi(z)$.

For Category I, the following procedure of analysis may be followed. (i) If the common purely imaginary zero is a simple one, let us denote it by iy^*. The exponential polynomial $\varphi(z)$ may then be written as

$$\varphi(z) = (z - iy^*)\varphi_1(z).$$

If iy^* is the only common purely imaginary zero, then one can take $\varphi_1(z)$ as a new characteristic exponential polynomial and proceed as in Category III. (ii) if iy^* is a multiple common zero, then the system is unstable by Theorem 4 and there is no need to proceed any further. (iii) If, besides iy^*, there are other common purely imaginary zeros, the above procedure may be repeated.

For Category II, there are three possibilities to consider. (i) If $z = 0$ is a common zero of $g(z)$ and $h(z)$, the case may be treated as in Category I. (ii) If $z = 0$ is a simple zero of $g(z) + h(z)$ but not for $g(z)$ and $h(z)$ separately, then from (2) one can easily see that $\varphi(z)$ always has a simple purely imaginary zero at $z = 0$ for any value of τ. In other words, this simple zero $z = 0$ does not leave the imaginary axis as τ varies. With the behavior of the zero $z = 0$ clear, the rest of the problem can be processed as in Category III. (iii) If $z = 0$ is a multiple zero, then the system is unstable by Theorem 4.

Since the effects on stability of the presence of common purely imaginary zeros of $g(z)$ and $h(z)$ or a zero of $z = 0$ can be ascertained separately and easily, we shall assume in the subsequent analysis that $\varphi(z)$ is in Category III.

3. Decomposition of the τ-axis. The development of the present method makes use of an idea upon which the D-decomposition method [2] and the method of Pinney [3] are also based. In the usual D-decomposition method, the parameter space is partitioned into a number of regions under the condition that the time retardation τ is kept constant. After the regions have been found, one then studies the change of stability character of the system as the boundaries of the regions are crossed. In the present investigation, since we are interested in the effects of τ on the stability, τ is allowed to vary while other parameters are kept fixed. The positive half of the τ-axis is first divided into intervals by boundary points at which purely imaginary roots of the characteristic equation exist. One then shows that within each interval the stability character of the system does not change. Next, one proceeds to determine how the number of zeros of $\varphi(z)$ with positive real parts changes as the boundary points of the intervals are crossed. Here it is found possible to identify this change with the leaving or entering of a unit circle by a simple curve defined by an algebraic expression which does not involve the time retardation at all. This, in the form of Theorem 15, is the main result of the investigation.

By defining

$$w_1(z) = e^{\tau z}, \tag{4}$$

$$w_2(z) = -\frac{h(z)}{g(z)}, \tag{5}$$

the characteristic equation (2) becomes

$$w(z) \overset{\Delta}{=} w_2(z) - w_1(z) = 0 \quad \text{or} \quad w_1(z) = w_2(z). \tag{6}$$

To find the purely imaginary zeros of $\varphi(z)$ or $w(z)$ we set $z = iy$ in (4) and (5). As z varies along the imaginary axis of the z-plane from $-i\infty$ to $+i\infty$, (4) and (5) yield two curves in the complex w-plane. Evidently, (4) maps the imaginary axis of the z-plane to a unit circle with the path going in the counterclockwise direction, and the existence of purely imaginary zeros of $w(z)$ is signified, according to (6), by the intersection of the $w_2(z)$ curve with the unit circle. This w_2-curve in the w-plane shall be called a *testing path in w* or simply a *testing path* in the subsequent analysis. For a given τ, the number of zeros of $\varphi(z)$ with positive real parts will be denoted by $N(\tau)$. In terms of this notation we cite without proof the following theorems.

THEOREM 5 (see [4],.[5]). *If the testing path $w_2(z)$ does not intersect the unit circle, then $N(\tau) = N(0)$.*

THEOREM 6. *The stability character of the system* (1) *is unaffected by the time retardation $\tau \geqq 0$, if*

$$|g(iy)| > |h(iy)| \quad \textit{for all } y. \tag{7}$$

Theorem 6 follows immediately from Theorem 5. Here we note that if the testing path does not intersect the unit circle, it is necessarily inside of the unit circle because $n > m$ in (1). We also note that as $g(z)$ and $h(z)$ are assumed not to have common purely imaginary zeros, the purely imaginary zeros of $g(z)$ cannot be zeros of $\varphi(z)$.

If the testing path does intersect the unit circle, let $\exp(i\alpha_1), \exp(i\alpha_2), \cdots,$ $\exp(i\alpha_j), \cdots$ be the points of intersection and let y_j be the value of y on $w_2(z)$ at $\exp(i\alpha_j)$. Concerning the angle measurements of α_j, we adopt the following rule: For $y_j > 0$ the angle α_j will be measured in the positive direction and, therefore, $0 \leqq \alpha_j < 2\pi$. For $y_j < 0$ the angle α_j will be measured in the negative direction and, therefore, $-2\pi < \alpha_j \leqq 0$. With this convention, the set of critical values of τ at which purely imaginary zeros of $\varphi(z)$ exist is readily seen from (4) to be

$$\tau_{jk} = \frac{1}{y_j}(\alpha_j + 2k\pi), \tag{8}$$

where $k = 0, 1, 2, \cdots$ for $y_j > 0$ and $k = 0, -1, -2, \cdots$ for $y_j < 0$. Since $z = 0$ is assumed not to be a zero of $\varphi(z)$, $y_j \neq 0$. Rearranging the nonnegative τ_{jk} into an ascending sequence $\tau_1, \tau_2, \tau_3, \cdots$, one obtains the desired boundary points of the intervals of τ. For each interval we have the following theorem which is given here with its simple proof omitted.

THEOREM 7 (see [5]). *If t_1 and t_2 are in the same open interval (τ_i, τ_{i+1}), then $N(t_1) = N(t_2)$.*

Often two or more of the τ_{jk} calculated according to (8) may coincide, each τ_{jk} corresponding to a different set of y_j, α_j and k. A typical example occurs when the purely imaginary zeros come in a pair $\pm iy_j$ with the corresponding angles $\pm\alpha_j$. This occurs, for instance, when $g(z)$ and $h(z)$ have their coefficients all real. Another distinct possibility is, of course, that the testing path may be tangent to the unit circle at the point of intersection, thus raising the possibility that the purely imaginary zero may be of multiple order. This general case has not been considered by previous writers. In rare instances, the testing path could intersect the unit circle at a specific point $w = \exp(i\alpha_j)$ more than once. Let us say that it passes through that point Q times; then the corresponding $y_{j(1)}, y_{j(2)}, \cdots, y_{j(Q)}$ are necessarily different. In that case the point of intersection $\exp(i\alpha_j)$ will be considered to be Q different intersection points, each being characterized by $(\alpha_j, y_{j(q)})$, $q = 1, 2, \cdots, Q$, and each to have its corresponding τ_{jk} calculated by (8) with y_j replaced by $y_{j(q)}$. We also remark here that the determination of y_j does not require solving any transcendental equations; it involves only calculations on polynomial expressions.

We conclude this section by referring to some assertions made in [4, p. 424]. Consider a case (Case f in [4]) where the system without retardation is stable and where there are two points of intersection K'_1 and K'_2 between the positive half of the testing path and the unit circle. Let α_1 and α_2 be the angles and let iy_1 and iy_2 be the purely imaginary zeros associated with K'_1 and K'_2, respectively. Moreover, let

$$\tau_{10} = \frac{\alpha_1}{y_1} < \frac{\alpha_2}{y_2} = \tau_{20}.$$

It is then asserted that (in the notation of this paper) the system is stable in

$$0 \leqq \tau < \tau_{10}, \quad \tau_{20} < \tau < \tau_{11}, \quad \tau_{21} < \tau < \tau_{12}, \cdots \tag{9a}$$

and unstable in

$$\tau_{10} < \tau < \tau_{20}, \quad \tau_{11} < \tau < \tau_{21}, \quad \tau_{12} < \tau < \tau_{22}, \cdots \tag{9b}$$

with continual alternation of the regions of stability and instability with increase of τ. It is also explained that the above "corresponds to the fact that since the Mikhailov curve in the presence of delay is 'wavy' with increase of τ the waves deform in such a way that the points K'_1 and K'_2 are alternately incident on the origin of coordinates. Incidence of the point K'_1 on the origin of coordinates always implies instability of the system while incidence of K'_2 returns the system each time to the stable state." Similar assertions are also made for a case where the system is unstable without delay and the testing path intersects the unit circle at two points.

A careful examination of (8) will show that in general τ_{1k} and τ_{2k}, $k = 0, 1, 2, \cdots$, do not form an alternate sequence. Sooner or later either the τ_{1k} or the τ_{2k} sequence will lag behind and the picture of alternate regions of stability and instability cannot be true any more. As a matter of fact, it will be shown later in the paper that as long as the testing path intersects the unit circle the system is always unstable for sufficiently large τ. Thus, the assertions quoted above cannot be true in general and are never true for large enough τ. The basic idea behind the approach of Sokolov and Miasnikov is, however, a most fruitful one; it is only necessary to recast the criteria of stability and instability in a somewhat different form. This will be done in a general and more complete manner in the following sections.

4. Main theorems on the variation of $N(\tau)$. The next task for us to do is to find how $N(\tau)$ varies as each boundary point τ_j is crossed by the increasing τ. As a preparatory work we first define the types of purely imaginary zeros which might be encountered in the above construction.

DEFINITION 1. Let z_0 be a zero of an analytic function $f(z)$ and

$$f(z_0) = \frac{df(z_0)}{dz} = \frac{d^2f(z_0)}{dz^2} = \cdots = \frac{d^{m-1}f(z_0)}{dz^{m-1}} = 0, \qquad \frac{d^mf(z_0)}{dz^m} \neq 0; \tag{10}$$

then z_0 is called a zero of order m.

Next we consider mapping of $w_1(z)$ and $w_2(z)$ curves in the w-plane to curves in a ζ-plane through

$$w = e^{\zeta}. \tag{11}$$

Evidently, the characteristic equation can now be written as

$$\zeta(z) \triangleq \zeta_2(z) - \zeta_1(z) = 0, \tag{12}$$

where

$$\zeta_1(z) = \tau z, \tag{13}$$

$$\zeta_2(z) = \ln\,[w_2(z)] \pm 2n\pi i = \ln\left(-\frac{h(z)}{g(z)}\right) \pm 2n\pi i, \tag{14}$$

with $\ln\,[w_2(z)]$ denoting the principal value of the logarithmic function.

Let $\zeta = \xi + i\eta$. Then the unit circle $w_1(z)$ in the w-plane is mapped into the imaginary axis of the ζ-plane, the $w_2(z)$ curve is mapped into a sequence of identical curves displaced in the vertical direction with a period of $2\pi i$, a point of intersection $w = \exp\,(i\alpha_j)$ is mapped into $\eta = \alpha_j \pm 2n\pi$, the interior of the unit circle in the w-plane is mapped into the left half of the ζ-plane, and a $w_2(z)$ curve leaving (or entering) the unit circle is mapped into a ζ_2-curve entering into the right (or left) half of the ζ-plane. Moreover, since the mapping is conformal, the angle between the unit circle and the w_2-curve at the point of intersection is preserved in the transformation. Also obvious is that if iy is a purely imaginary zero of $w(z)$ of order m, it is also a zero of $\zeta(z)$ of the same order.

Consider now a typical boundary point $\tau_H(=\tau_{jk})$ separating the intervals (τ_{H-1}, τ_H) and (τ_H, τ_{H+1}). At this value of τ there exists a purely imaginary characteristic root $z = iy_j$, which corresponds to the intersection point P of the ζ_2-curve (in the strip $2k\pi \leqq \operatorname{Im}\zeta < 2(k+1)\pi$ for $y_j > 0$ and in the strip $2(k-1)\pi < \operatorname{Im}\zeta \leqq 2k\pi$ for $y_j < 0$) with the imaginary axis at $\eta = \alpha_j + 2k\pi$. Here, k is the same integer k used in defining $\tau_H = \tau_{jk}$ through (8). If τ_H is given a small increment $\Delta\tau$, then the root iy_j will also change and, in general, will not remain purely imaginary. Our aim is to find the real part of this change Δz. The new root $z = iy_j + \Delta z$ corresponding to $\tau = \tau_H + \Delta\tau$, of course, has to satisfy (12) which we rewrite now as

$$\tau = \frac{\zeta_2(z)}{z} \overset{\Delta}{=} f(z). \tag{15}$$

What we need to do is to solve z in terms of τ in the neighborhood of ($z = iy_j$, $\tau = \tau_H$) in order to find the change Δz for a given change $\Delta\tau$. To proceed further we consider separately the different cases according to the order of the zero $z = iy_j$ involved.

Case I. *Simple zero.* Let us look at the characteristic equation (15). It is satisfied by $z = iy_j$ when $\tau = \tau_H$ and the corresponding value of ζ_2 is $iy_j\tau_H$. Now assume that $z = iy_j$ is a simple zero of (12). Then, by Definition 1,

$$\frac{d}{dz}\zeta_2(iy_j) \neq \tau_H. \tag{16}$$

This implies by (15) that

$$\left(\frac{d\tau}{dz}\right)_{z=iy_j} = \left[\frac{1}{z}\left(\frac{d\zeta_2}{dz} - \frac{\zeta_2}{z}\right)\right]_{z=iy_j} \neq 0. \tag{17}$$

This allows us to invoke the following lemma to express z in terms of τ.

LEMMA 1 (see [9]). *If $f(z)$ is regular in a neighborhood of z_0 and if $f(z_0) = \tau_0$, $df(z_0)/dz \neq 0$, then the equation $\tau = f(z)$ has a unique solution, regular in a neighborhood of τ_0, of the form*

$$z = z_0 + \sum_{n=1}^{\infty} \frac{(\tau - \tau_0)^n}{n!}\left[\frac{d^{n-1}}{dz^{n-1}}\{\psi(z)\}^n\right]_{z=z_0}, \tag{18}$$

where

$$\psi(z) = \frac{z - z_0}{f(z) - \tau_0}. \tag{19}$$

Applying (18), which is known as Lagrange's inversion formula, to our problem at hand, we obtain

$$\psi(z) = \frac{z}{-\tau_H + \sum_{m=1}^{\infty} \frac{1}{m!}\left(\frac{d^m\zeta_2}{dz^m}\right)_{z=iy_j} (z - z_0)^{m-1}} \tag{20}$$

and $\Delta\tau$ being real,

$$\operatorname{Re}(\Delta z) = \sum_{n=1}^{\infty} \frac{(\Delta\tau)^n}{n!} \operatorname{Re}\left[\frac{d^{n-1}}{dz^{n-1}}\{\psi(z)\}^n\right]_{z=iy_j}. \tag{21}$$

For our purpose of determining the change of $N(\tau)$ as τ increases across τ_H, it is not necessary to obtain the series (21) in its entirety but rather only the leading term. For the following analysis we need the real and imaginary parts of $(d^n\zeta_2/dz^n)$ at $z = iy_j$ separately. Let us write

$$\left(\frac{d^n\zeta_2}{dz^n}\right)_{z=iy_j} = (-i)^n(r_n + is_n) \quad \text{or} \quad \left(\frac{d^n\zeta_2}{dy^n}\right)_{z=iy_j} = r_n + is_n. \tag{22}$$

Case IA. Consider first the case $r_1 \neq 0$. In this case one easily finds by (20) and (21) that

$$\operatorname{Re}(\Delta z) = -\frac{y_j r_1 \Delta\tau}{r_1^2 + (s_1 - \tau_H)^2} + O((\Delta\tau)^2). \tag{23}$$

This says that for $\Delta\tau > 0$ the characteristic root $z = iy_j$ at $\tau = \tau_H$ changes into one with positive real part if $y_j r_1 < 0$ and changes into one with negative real part if $y_j r_1 > 0$. For $\Delta\tau < 0$ the reverse is true. In other words, we have the following theorem.

THEOREM 8. *If the zero $z = iy_j$, which exists for $\tau = \tau_H$, is a simple zero and if $r_1 \neq 0$ at $z = iy_j$, then as τ increases across τ_H, $N(\tau)$ increases by* 1 *if $y_j r_1 < 0$ and decreases by* 1 *if $y_j r_1 > 0$.*

Case IB. Next we consider the case $r_1 = 0$. In this case we must have $s_1 \neq \tau_H$ because otherwise $d\zeta_2/dz = \tau_H$ at $z = iy_j$ and the zero is no longer simple. We further note that when $r_1 = 0$ the first order term on the right-hand side of (23) vanishes. Therefore, one must proceed to investigate terms involving higher orders of $\Delta\tau$ in order to ascertain the change $\operatorname{Re}(\Delta z)$ with respect to $\Delta\tau$. Let us assume for the general case that $r_M \neq 0$ but all $r_m = 0$ for $m < M$. Then by (20) and (21) it is found that

$$\operatorname{Re}(\Delta z) = -\frac{r_M y_j^M}{M!(s_1 - \tau_H)^{M+1}}(\Delta\tau)^M + O((\Delta\tau)^{M+1}). \tag{24}$$

From this we see that if M is even then the changes of $\operatorname{Re}(\Delta z)$ are of the same sign for both $\Delta\tau > 0$ and $\Delta\tau < 0$; therefore, there is no change of $N(\tau)$ as τ increases across τ_H. If M is odd, then the change of $N(\tau)$ depends upon the sign of $y_j r_M$. Thus, we have established the following theorems.

THEOREM 9. *If the zero $z = iy_j$, which exists for $\tau = \tau_H$, is a simple zero, if $r_M \neq 0$ for an even M and if all r_m for $m < M$ vanish, then $N(\tau)$ suffers no change as τ increases across τ_H.*

THEOREM 10. *If the zero $z = iy_j$, which exists for $\tau = \tau_H$, is a simple zero, if $r_M \neq 0$ for an odd M and if all r_m for $m < M$ vanish, then as τ increases across τ_H, $N(\tau)$ increases by* 1 *if $y_j r_M < 0$ and decreases by* 1 *if $y_j r_M > 0$.*

Case II. *Zero of order K. $K > 1$.* Let $z = iy_j$ now be a zero of (12) of order K. By Definition 1,

$$\frac{d\zeta_2(iy_j)}{dz} = \tau_H, \quad \frac{d^2\zeta_2(iy^j)}{dz^2} = 0 \quad , \cdots, \quad \frac{d^{K-1}\zeta_2(iy_j)}{dz^{K-1}} = 0, \tag{25}$$
$$\frac{d^K\zeta_2(iy_j)}{dz^K} \neq 0.$$

When referred to (15), one finds that (25) implies

$$\frac{df(iy_j)}{dz}=0,\quad \frac{d^2f(iy_j)}{dz^2}=0\quad ,\cdots,\quad \frac{d^{K-1}f(iy_j)}{dz^{K-1}}=0, \tag{26}$$
$$\frac{d^Kf(iy_j)}{dz^K}\neq 0.$$

Therefore, in the neighborhood of $z = iy_j$,

$$\tau=\tau_H+\sum_{n=K}^{\infty}\frac{1}{n!}\frac{d^nf(iy_j)}{dz^n}(z-iy_j)^n. \tag{27}$$

Here we note that the series in $(z - iy_j)$ on the right-hand side starts with the Kth power, and the Lagrange inversion formula (18) is no longer appropriate or adequate. We need a new lemma.

LEMMA 2. *If $f(z)$ is regular in the neighborhood of z_0 with $f(z_0) = \tau_0$, and if the first $K - 1$ derivatives of $f(z)$ vanish at z_0, but $d^Kf(z_0)/dz^K \neq 0$, then the equation $f(z) = \tau$ has a K-valued solution of the form*

$$z=z_0+\sum_{n=1}^{\infty}\frac{(\tau-\tau_0)^{n/K}}{n!}\left[\frac{d^{n-1}}{dz^{n-1}}\{\psi(z)\}^n\right]_{z=z_0}, \tag{28}$$

where

$$\psi(z)=\frac{z-z_0}{[f(z)-\tau_0]^{1/K}}. \tag{29}$$

Proof. Let us introduce a complex variable t which is related to τ by

$$\tau-\tau_0=(t-t_0)^K. \tag{30}$$

Then t is a K-valued function of τ. As a function of z,

$$t=t_0+(f(z)-\tau_0)^{1/K}\overset{\Delta}{=}u(z). \tag{31}$$

Consider the K values of $(f(z) - \tau_0)^{1/K}$ one by one. Then it is easy to show that

$$\left(\frac{dt}{dz}\right)_{z=z_0}\neq 0 \tag{32}$$

and that $t(z)$ is regular in a neighborhood of z_0. Thus, Lemma 1 is applicable to $t = u(z)$ for each of the K values of $(f(z) - \tau_0)^{1/K}$,

$$z=z_0+\sum_{n=1}^{\infty}\frac{(t-t_0)^n}{n!}\left[\frac{d^{n-1}}{dz^{n-1}}\{\psi(z)\}^n\right]_{z=z_0}, \tag{33}$$

where

$$\psi(z)=\frac{z-z_0}{u(z)-t_0}. \tag{34}$$

Making use of (30) and (31) to express the series in terms of $f(z)$ and τ, we obtain the desired formulas (28) and (29). When the K values of $(f(z) - \tau_0)^{1/K}$ are taken altogether, z is seen to be a K-valued function of τ. This fact is also clearly reflected by (29). This completes the proof.

Applying Lemma 2 to the problem at hand, we find

$$\psi(z) = \frac{z^{1/K}}{\left\{\sum_{m=K}^{\infty} \frac{1}{m!} \frac{d^m \zeta_2(iy_j)}{dz^m}(z - iy_j)^{m-K}\right\}^{1/K}} \tag{35}$$

and

$$\operatorname{Re}(\Delta z) = \sum_{n=1}^{\infty} \frac{1}{n!} \operatorname{Re}\left\{(\Delta\tau)^{n/K}\left[\frac{d^{n-1}}{dz^{n-1}}\{\psi(z)\}^n\right]_{z=iy_j}\right\}. \tag{36}$$

Case IIA. Consider first the case of $r_K \neq 0$. Then by (36) and (35),

$$\operatorname{Re}(\Delta z) = \operatorname{Re}\left\{\frac{K!i^K(s_K + ir_K)y_j\Delta\tau}{r_K^2 + s_K^2}\right\}^{1/K} + O(|\Delta\tau|^{2/K}). \tag{37}$$

Let us denote

$$\frac{K!i^K(s_K + ir_K)y_j\Delta\tau}{r_K^2 + s_K^2} = \rho e^{i\theta}, \qquad 0 \leqq \theta < 2\pi. \tag{38}$$

Then

$$\operatorname{Re}(\Delta z) = \rho^{1/K}\cos\theta_n + O(|\Delta\tau|^{2/K}), \qquad n = 1, 2, \cdots, K, \tag{39}$$

where

$$\theta_n = \frac{\theta}{K} + \frac{2(n-1)\pi}{K}. \tag{40}$$

There are K values of $\operatorname{Re}(\Delta z)$ for each change $\Delta\tau$, corresponding to the order K of the zero $z = iy_j$. It can be easily shown that as long as $r_K \neq 0$, then none of the multiple values of $\operatorname{Re}(\Delta z)$ from the first term of (39) is zero and, therefore, the purely imaginary zero of order K changes into K zeros with nonvanishing real parts of order $|\Delta\tau|^{1/K}$ as τ increases or decreases from τ_H. We also note that the K values of the arguments θ_n are spaced evenly from 0 to 2π.

Consider now separately the case where K is even and the case where K is odd. For K even, there are necessarily $K/2$ number of θ_n in the range $(-\pi/2, \pi/2)$ and $K/2$ number of θ_n in the range $(\pi/2, 3\pi/2)$ no matter whether $\Delta\tau$ is positive or negative. This means we have the following theorem.

THEOREM 11. *If the zero $z = iy_j$, which exists for $\tau = \tau_H$, is a zero of even order K and if $r_K \neq 0$, then $N(\tau)$ suffers no change as τ increases across τ_H.*

For the case K is odd, we need the following results which are cited here with their straightforward proofs omitted.

LEMMA 3. *If $(K+1)/2$ is even and if $-\pi/2 < \theta < \pi/2$, then among*

$$\theta_n = \frac{\theta}{K} + \frac{2(n-1)\pi}{K}, \qquad n = 1, 2, \cdots, K, \tag{41}$$

$(K-1)/2$ number of θ_n are in the range $-\pi/2 < \theta_n < \pi/2$ and $(K+1)/2$ number of θ_n are in the range $\pi/2 < \theta_n < 3\pi/2$. If $\pi/2 < \theta < 3\pi/2$, then the reverse is true.

LEMMA 4. *If $(K+1)/2$ is odd and if $-\pi/2 < \theta < \pi/2$, then among*

$$\theta_n = \frac{\theta}{K} + \frac{2(n-1)\pi}{K}, \qquad n = 1, 2, \cdots, K, \tag{42}$$

$(K+1)/2$ number of θ_n are in the range $-\pi/2 < \theta_n < \pi/2$ and $(K-1)/2$ number of θ_n are in the range $\pi/2 < \theta_n < 3\pi/2$. If $\pi/2 < \theta < 3\pi/2$, then the reverse is true.

Now if we refer to (38) and take K to be odd, we may rewrite it as

$$\rho e^{i\theta} = \frac{K!(-1)^{(K+1)/2}(r_K - is_K)y_j\Delta\tau}{r_K^2 + s_K^2}. \tag{43}$$

If $(K+1)/2$ is even and $y_j r_K > 0$, then by Lemma 3 the number of zeros with negative real parts will exceed by one the number of zeros with positive real parts for $\Delta\tau > 0$, and the number of zeros with positive real parts will exceed that with negative real parts by one for $\Delta\tau < 0$. In other words, $N(\tau)$ will decrease by one as τ increases across τ_H. If $(K+1)/2$ is even but $y_j r_K < 0$, then the reverse is true and $N(\tau)$ will increase by one. When $(K+1)/2$ is odd, then we can apply Lemma 4 to obtain similar results.

THEOREM 12. *If the zero $z = iy_j$, which exists for $\tau = \tau_H$, is a zero of odd order K and if $r_K \neq 0$, then as τ increases across τ_H, $N(\tau)$ increases by one if $y_j r_K < 0$, and decreases by one if $y_j r_K > 0$.*

Case IIB. Next, we consider the case $r_K = 0$. Here it is desirable to study four separate cases, namely: $K = 0, 1, 2$ and $3 \pmod 4$, respectively.

Take first $K = 4N$, $N = 1, 2, 3, \cdots$. For this case, by (38) we have

$$\rho = \frac{K!|s_K y_j \Delta\tau|}{s_K^2}, \quad \theta = \begin{cases} 0 & \text{if } s_K y_j \Delta\tau > 0, \\ \pi & \text{if } s_K y_j \Delta\tau < 0. \end{cases} \tag{44}$$

Now, if $s_K y_j \Delta\tau < 0$, then all the K values of $\mathrm{Re}(\Delta z)$ calculated from the leading term of (39) can be shown to be nonzero, half positive and half negative. Thus, we have the information we need on $\mathrm{Re}(\Delta z)$. The multiple zero iy_j of order K, regarded as K zeros momentarily coalesced into one at $\tau = \tau_H$, changes into $K/2$ zeros with positive real parts and $K/2$ zeros with negative real parts. If $s_K y_j \Delta\tau > 0$ the situation is different. Based upon the leading term in (39), there will be $(K-2)/2$ number of zeros with positive real parts and $(K-2)/2$ number of zeros with negative real parts, but there will be also two (distinct) purely imaginary zeros corresponding to

$$\Delta z = \pm i\rho^{1/K} + O(|\Delta\tau|^{2/K}). \tag{45}$$

For these two values of Δz, and for these two only, the first order term of (39) or (36) does not provide us with any information about the real part of Δz; it is, therefore, necessary to go to the higher order terms of (36) in order to complete the analysis. It is useful to note that the imaginary parts of these two values of Δz are still of the order $|\Delta\tau|^{1/K}$. Let $r_{K+1} = r_{K+2} = \cdots = r_{M-1} = 0$ but $r_M \neq 0$.

Then it can be shown that the higher order calculation of Re (Δz) leads to

$$\text{(46)} \qquad \operatorname{Re}(\Delta z) = -\frac{(K-1)!}{M!}\frac{r_M y_j}{s_K y_j}(\pm\rho^{1/K})^{M-K+1} + O(|\Delta\tau|^{(M-K+2)/K}).$$

Now, if M is even so that $M - K + 1$ is odd, then there are one positive and one negative Re (Δz) from (46). Combining this result with the previous discussion, one finds that there are still $K/2$ zeros with positive real parts and $K/2$ zeros with negative real parts even for the case $s_K y_j \Delta\tau > 0$. Therefore, there will be no change in $N(\tau)$ as τ increases across τ_H.

If M is odd, then by (46) there will be two positive Re (Δz) or two negative ones depending upon the signs of $r_M y_j$ and $s_K y_j$. The results obtained by a straightforward examination of (46) may be combined with the earlier discussion as shown in Table 1.

TABLE 1

	Number of Re(Δz) > 0	Number of Re(Δz) < 0	Change in $N(\tau)$ as τ increases
$r_M y_j > 0, \quad s_K y_j > 0, \Delta\tau < 0$ $\Delta\tau > 0$	$K/2$ $(K/2) - 1$	$K/2$ $(K/2) + 1$	-1
$r_M y_j > 0, \quad s_K y_j < 0, \Delta\tau < 0$ $\Delta\tau > 0$	$(K/2) + 1$ $K/2$	$(K/2) - 1$ $K/2$	-1
$r_M y_j < 0, \quad s_K y_j > 0, \Delta\tau < 0$ $\Delta\tau > 0$	$K/2$ $(K/2) + 1$	$K/2$ $(K/2) - 1$	$+1$
$r_M y_j < 0, \quad s_K y_j < 0, \Delta\tau < 0$ $\Delta\tau > 0$	$(K/2) - 1$ $K/2$	$(K/2) + 1$ $K/2$	$+1$

Consider next the case $K = 4N + 1$, $N = 1, 2, \cdots$. For this case by (38) we have

$$\text{(47)} \qquad \rho = \frac{K!|s_K y_j \Delta\tau|}{s_K^2}, \qquad \theta = \begin{cases} \pi/2 & \text{if } s_K y_j \Delta\tau > 0, \\ 3\pi/2 & \text{if } s_K y_j \Delta\tau < 0. \end{cases}$$

Then by (39) one finds that among the K values of Re (Δz) calculated to the first order of $|\Delta\tau|^{1/K}$ there are $(k - 1)/2$ positive values, $(K - 1)/2$ negative values and one zero value. This last one corresponds to

$$\text{(48)} \qquad \Delta z = [\operatorname{sgn}(s_K y_j \Delta\tau)] i\rho^{1/K} + O(|\Delta\tau|^{2/K}).$$

For this one and this one alone we need go to the higher order terms of (36) to find the real part of Δz. Let r_M again be the first nonvanishing r_M for $m > K$. Then the higher order analysis yields

$$\text{(49)} \quad \operatorname{Re}(\Delta z) = -\frac{(K-1)!}{M!}\frac{r_M y_j}{s_K y_j}[\operatorname{sgn}(s_K y_j \Delta\tau)\rho^{1/K}]^{M-K+1} + O(|\Delta\tau|^{(M-K+2)/K}).$$

Now, if M is even so that $M - K + 1$ is even, then the sign of $\operatorname{Re}(\Delta z)$ given by (49) will be the same for $\Delta\tau > 0$ and for $\Delta\tau < 0$. This means that there will be no change in $N(\tau)$. If M is odd, then a tabulation similar to Table 1 can be carried out, leading to the conclusion that as τ increases across τ_H, $N(\tau)$ increases by 1 if $r_M y_j < 0$ and decreases by 1 if $r_M y_j > 0$.

Other cases where $K = 2$ or 3 (mod 4) can be carried out in a similar manner. The criterion expressed in terms of r_M and $r_M y_j$ remains unchanged. Thus, we have the following theorems.

THEOREM 13. *If the zero $z = iy_j$, which exists for $\tau = \tau_H$, is a multiple zero of order K, if $r_M \neq 0$ for an even M and if all r_m for $m < M$ vanish, then $N(\tau)$ suffers no change as τ increases along τ_H.*

THEOREM 14. *If the zero $z = iy_j$, which exists for $\tau = \tau_H$, is a multiple zero of order K, if $r_M \neq 0$ for an odd M and if all r_m for $m < M$ vanish, then as τ increases across τ_H, $N(\tau)$ increases by 1 if $r_M y_j < 0$ and decreases by 1 if $r_M y_j > 0$.*

Next, we study the curve $\zeta_2(z)$ in the strip $2k\pi \leqq \operatorname{Im}\zeta < 2(k+1)\pi$ or $2(k-1)\pi < \operatorname{Im}\zeta \leqq 2k\pi$ in which the intersection point P is located. The purpose of the study is to establish the local geometrical property of the ζ_2-curve near P with respect to the imaginary axis. For $z = iy_j$, ζ_2 has a purely imaginary value $2k\pi + \alpha_j$. As y_j is given a small increment dy, the corresponding change in ζ_2 is given by

$$d\zeta_2 = \sum_{n=1}^{\infty} \frac{1}{n!} \frac{d^n\zeta_2}{dz^n}(i\,dy)^n = \sum_{n=1}^{\infty} \frac{1}{n!} \frac{d^n\zeta_2}{dy^n}(dy)^n, \tag{50}$$

where the derivatives are evaluated at $z = iy_j$. Making use of (22), one finds that the real part of $d\zeta_2$ is simply

$$d\xi_2 = \operatorname{Re}(d\zeta_2) = \sum_{n=1}^{\infty} \frac{1}{n!} r_n (dy)^n. \tag{51}$$

Suppose that among $r_1, r_2, \cdots$ the first nonvanishing one is r_M; and if the direction in which y increases is taken to be the forward direction for the ζ_2-curve, then we obviously have:

(i) If M is odd and r_M is positive, the ζ_2-curve crosses the η-axis at P and goes from the left to the right.

(ii) If M is odd and r_M is negative, the ζ_2-curve crosses the η-axis at P but goes from the right to the left.

(iii) If M is even then the ζ_2-curve, although touching the imaginary axis at P, does not cross it. It remains on one side of the imaginary axis in the neighborhood of P.

Making use of these geometrical properties of the ζ_2-curve near P, and observing the mapping relationship between the ζ_2-curve and the w_2-curve and that between the unit circle in the w-plane and the imaginary axis in the ζ-plane, we may reinterpret the results established in Theorems 8 to 14 in the following manner.

THEOREM 15. *Let τ_H be a boundary point on the τ-axis, let the associated purely imaginary zero be iy_j, and let Q be the corresponding intersection point of the* *testing path $w_2(z)$ with the unit circle. If the testing path enters the unit circle at Q, then as τ increases across τ_H, $N(\tau)$ increases by 1 for $y_j > 0$ and decreases by 1 for $y_j < 0$. If the testing path leaves the unit circle at Q, then as τ increases across τ_H, $N(\tau)$ decreases by 1 for $y_j > 0$ and increases by 1 for $y_j < 0$. If in the neighborhood of Q the testing path remains on one side of the unit circle, then $N(\tau)$ suffers no change as τ increases across τ_H.*

Starting with $N(0)$ and counting the change of $N(\tau)$ at each of the boundary points τ_j, one can find $N(\tau)$ for any value of τ. The system is asymptotically stable only when $N(\tau) = 0$.

Remark 1. Theorem 15 is established on the assumption that each boundary point is associated with only one iy_j. When this is not the case, the counting procedure has to be modified. We have mentioned earlier that often two or more of the τ_{jk} calculated by (8) may coincide, each τ_{jk} corresponding to a different set of y_j, α_j and k. In this case each of the identical τ_{jk} so obtained must be treated as a separate one and its corresponding change of $N(\tau)$ calculated according to Theorem 15. The algebraic sum of all the changes of $N(\tau)$ gives then the net change of $N(\tau)$ as τ increases across this particular $\tau_H = \tau_{jk}$. Thus, for instance, when $g(z)$ and $h(z)$ have their coefficients all real so that the purely imaginary zeros come in pairs $\pm iy_j$ with their corresponding angles $\pm\alpha_j$, the change of $N(\tau)$ will always be -2, 0, $+2$, or other even integers, as each boundary point is crossed.

Remark 2. In this section the conclusions of the theorems are stated in terms of the change of $N(\tau)$ as τ increases across τ_H. When τ_H happens to be equal to zero, one cannot use directly these theorems because as τ changes from positive to negative the system changes from one of the retarded type to one of the advanced type possessing an infinite number of zeros with positive real parts. What one really wants under that circumstance is not to have $N(\tau)$ for $\tau < 0$ involved in the analysis but rather simply to find whether the purely imaginary zero which exists at $\tau = 0$ changes into a zero or zeros with positive or negative real parts for very small positive values of τ. This information is, however, not difficult to obtain because it has been explicitly given in the discussions leading to the establishment of the theorems. With this knowledge in hand and the number $N(0)$ obtained from a stability analysis of the system without delay, one can start the counting procedure of $N(\tau)$ along the whole positive τ-axis.

5. Large time retardation. If the testing path of a system does not intersect the unit circle, then by Theorem 5 the stability character of the system with any amount of delay remains the same as of that without delay. Let the system be such that the testing path does intersect the unit circle. There are two types of points of contact: points where the testing path crosses the unit circle and points where the testing path remains on one side of the unit circle in the neighborhood of the contact point. Associated with points of the second type there will be no change of $N(\tau)$; therefore, we shall ignore them. For points of intersection of the first kind, let the associated y_j be arranged in the following manner:

$$(52) \qquad y_1^- < y_2^- < \cdots < y_A^- < y_B^+ < y_{B-1}^+ < \cdots < y_2^+ < y_1^+,$$

where the superscripts $(-)$ and $(+)$ indicate the signs of y_j. Because of $n > m$ for retarded systems (1) the testing path starts and ends at the origin of the w-plane as y varies from $-\infty$ to $+\infty$. Therefore, y_j^- with j odd (even) and y_j^+ with j even (odd) are necessarily associated with points where the testing path leaves (enters) the unit circle. Moreover, $A + B$ must be even. Let the corresponding angles measured according to the convention established earlier (refer to (8)) be denoted by

$$\alpha_1^-, \alpha_2^-, \cdots, \alpha_A^-, \alpha_B^+, \alpha_{B-1}^+, \cdots, \alpha_2^+, \alpha_1^+. \tag{53}$$

DEFINITION 2. Int (a) stands for the smallest integer, including zero, which is greater or equal to the number a.

With the aid of this symbol Int (a) it is easy to show that for a given τ which is not a boundary point the number of τ_{jk} associated with y_j which exist in the interval $(0, \tau)$ is given by

$$\operatorname{Int}\left(\frac{\tau|y_j| - |\alpha_j|}{2\pi}\right). \tag{54}$$

The results contained in Theorem 15 may then be recast in the form of the following theorem.

THEOREM 16. *At a given value of τ, not one of the boundary values,*

$$N(\tau) = N(0) + U - V, \tag{55}$$

where

$$U = \sum_{j=1,3,5,\cdots}^{B \text{ or } B-1} \operatorname{Int}\left(\frac{\tau y_j^+ - \alpha_j^+}{2\pi}\right) + \sum_{j=1,3,5,\cdots}^{A \text{ or } A-1} \operatorname{Int}\left(\frac{\tau|y_j^-| - |\alpha_j^-|}{2\pi}\right), \tag{56}$$

$$V = \sum_{j=2,4,6,\cdots}^{B \text{ or } B-1} \operatorname{Int}\left(\frac{\tau y_j^+ - \alpha_j^+}{2\pi}\right) + \sum_{j=2,4,6,\cdots}^{A \text{ or } A-1} \operatorname{Int}\left(\frac{\tau|y_j^-| - |\alpha_j^-|}{2\pi}\right). \tag{57}$$

We may also write $U - V$ as

$$\begin{aligned} U - V = & \sum_{j=1,3,5,\cdots}^{A-1} \left[\operatorname{Int}\left(\frac{\tau|y_j^-| - |\alpha_j^-|}{2\pi}\right) - \operatorname{Int}\left(\frac{\tau|y_{j+1}^-| - |\alpha_{j+1}^-|}{2\pi}\right)\right] \\ & + \sum_{j=1,3,5,\cdots}^{B-1} \left[\operatorname{Int}\left(\frac{\tau y_j^+ - \alpha_j^+}{2\pi}\right) - \operatorname{Int}\left(\frac{\tau y_{j+1}^+ - \alpha_{j+1}^+}{2\pi}\right)\right] \end{aligned} \tag{58}$$

if A is even, or

$$\begin{aligned} U - V = & \sum_{j=1,3,5,\cdots}^{A-2} \left[\operatorname{Int}\left(\frac{\tau|y_j^-| - |\alpha_j^-|}{2\pi}\right) - \operatorname{Int}\left(\frac{\tau|y_{j+1}^-| - |\alpha_{j+1}^-|}{2\pi}\right)\right] \\ & + \operatorname{Int}\left(\frac{\tau|y_A^-| - |\alpha_A^-|}{2\pi}\right) + \sum_{j=1,3,5,\cdots}^{B-2} \left[\operatorname{Int}\left(\frac{\tau y_j^+ - \alpha_j^+}{2\pi}\right)\right. \\ & \left. - \operatorname{Int}\left(\frac{\tau y_{j+1}^+ - \alpha_{j+1}^+}{2\pi}\right)\right] + \operatorname{Int}\left(\frac{\tau y_B^+ - \alpha_B^+}{2\pi}\right) \end{aligned} \tag{59}$$

If A is odd. We note here that all Int (a) appearing in (58) and (59) are nonnegative. Also, because of the ordering of the magnitudes of y_j by (52), all the expressions in the brackets of (58) and (59) are nonnegative if τ is greater than a certain value T. For instance, we may choose T to be the smallest τ satisfying

$$\tau > \frac{|\alpha_j^{\pm}| - |\alpha_{j+1}^{\pm}|}{|y_j^{\pm}| - |y_{j+1}^{\pm}|} \qquad \text{for all odd } j. \tag{60}$$

Furthermore, if T or a larger one, if necessary, *also* satisfies

$$T > \frac{2\pi + |\alpha_j^{\pm}| - |\alpha_{j+1}^{\pm}|}{|y_j^{\pm}| - |y_{j+1}^{\pm}|} \qquad \text{for one odd } j, \tag{61}$$

then

$$U - V \geqq 1. \tag{62}$$

Thus, we have proved the following theorem.

THEOREM 17. *If the testing path for a system does intersect the unit circle so that the set $\{y_j\}$ is nonempty, then the system is unstable for all sufficiently large values of τ.*

This theorem obviously invalidates one of the assertions given in [4] and quoted in § 3.

6. Examples. We conclude the paper by giving two examples of application.

Example 1. Consider a dynamical system described by

$$\frac{d^2x(t)}{dt^2} + 4\pi^2x(t) + \pi^2x(t - \tau) = 0. \tag{63}$$

The characteristic equation is

$$\varphi(z) = (z^2 + 4\pi^2)e^{\tau z} + \pi^2 = 0. \tag{64}$$

For $\tau = 0, z = \pm\sqrt{5}\pi i$. Hence $N(0) = 0$. For $\tau \neq 0$ we follow the procedure given in this paper and find the purely imaginary zeros to be

$$\begin{aligned} y_1^- &= -\sqrt{5}\pi, & \alpha_1^- &= 0 & &\text{leaving,} \\ y_2^- &= -\sqrt{3}\pi, & \alpha_2^- &= -\pi & &\text{entering,} \\ y_2^+ &= \sqrt{3}\pi, & \alpha_2^+ &= \pi & &\text{leaving,} \\ y_1^+ &= \sqrt{5}\pi, & \alpha_1^+ &= 0 & &\text{entering} \end{aligned} \tag{65}$$

and the boundary points of τ to be at

$$\begin{aligned} \tau_{1k}^{\pm} &= \frac{2k}{\sqrt{5}}, & k &= 0, 1, 2, \cdots, \\ \tau_{2k}^{\pm} &= \frac{1 + 2k}{\sqrt{3}}, & k &= 0, 1, 2, \cdots. \end{aligned} \tag{66}$$

Then, by Theorem 15 we find the regions of stability and instability as follows:

$$
\begin{array}{lll}
& 0 < \tau < \frac{1}{\sqrt{3}}, \quad \frac{2}{\sqrt{5}} < \tau < \frac{3}{\sqrt{3}}, \quad \frac{4}{\sqrt{5}} < \tau, & \text{unstable,} \\
(67) & \frac{1}{\sqrt{3}} < \tau < \frac{2}{\sqrt{5}}, \quad \frac{3}{\sqrt{3}} < \tau < \frac{4}{\sqrt{5}}, & \text{asymptotically stable,} \\
& \tau = 0, \frac{1}{\sqrt{3}}, \frac{2}{\sqrt{5}}, \frac{3}{\sqrt{3}}, \frac{4}{\sqrt{5}}, & \text{stable.}
\end{array}
$$

Example 2. Next consider a dynamical system of the fifth order described by

$$
(68) \quad 10\frac{d^5x(t)}{dt^5} + 23\frac{d^4x(t)}{dt^4} + 24\frac{d^3x(t)}{dt^3} + 16\frac{d^2x(t)}{dt^2} + 6\frac{dx(t)}{dt} + x(t - \tau) = 0.
$$

The characteristic equation is

$$
(69) \qquad \varphi(z) = (10z^5 + 23z^4 + 24z^3 + 16z^2 + 6z)e^{\tau z} + 1 = 0.
$$

When $\tau = 0$, we find $N(0) = 0$ by the Routh–Hurwitz criterion. For $\tau \neq 0$, we find

$$
(70) \qquad
\begin{array}{lll}
y_1^- = -0.180, & \alpha_1^- = -1.28 & \text{leaving,} \\
y_1^+ = 0.180, & \alpha_1^+ = 1.28 & \text{entering}
\end{array}
$$

and

$$
(71) \qquad \tau_{1k}^{\pm} = \frac{1.28 + 2k\pi}{0.180}, \qquad k = 0, 1, 2, \cdots.
$$

The regions of stability and instability are found to be

$$
(72) \qquad
\begin{array}{ll}
0 < \tau < 7.1, & \text{asymptotically stable,} \\
7.1 < \tau, & \text{unstable,} \\
\tau = 0,\ 7.1 & \text{stable.}
\end{array}
$$

REFERENCES

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

[2] L. E. EL'SGOL'TS, *Introduction to the Theory of Differential Equations with Deviating Arguments*, Holden-Day, San Francisco, 1966.

[3] E. PINNEY, *Ordinary Difference-Differential Equations*, University of California Press, Berkeley, 1959.

[4] E. P. POPOV, *The Dynamics of Automatic Control Systems*, Pergamon Press, New York, 1962.

[5] A. M. KRALL, *Stability criteria for feedback systems with a time lag*, this Journal, 4 (1965), pp. 160–170.

[6] L. S. PONTRYAGIN, *On the zeros of some elementary transcendental functions*, Amer. Math. Soc. Transl., Ser. 2, 1 (1955), pp. 95–110.

[7] S. J. BHATT AND C. S. HSU, *Stability criteria for second-order dynamical systems with time lag*, J. Appl. Mech., 33 (1966), pp. 113–118.

[8] C. S. HSU AND S. J. BHATT, *Stability charts for second-order dynamical systems with time lag*, Ibid., 33 (1966), pp. 119–124.

[9] E. T. COPSON, *Theory of Functions of a Complex Variable*, Oxford University Press, London, 1950.

SIAM J. CONTROL
Vol. 7, No. 2, May 1969

ON OPTIMAL CONTROL OF THE VIBRATING STRING*

KAZIMIERZ MALANOWSKI†

Summary. The problem of minimization of total energy of the vibrating string in a given time is investigated. The boundary value control with constrained magnitude is considered.

It is shown that this optimal control problem is equivalent to the determining of the minimum of some integral functional subject to constraints on the magnitude of the argument.

The form of the optimal control function is characterized. It turns out that, in general, an optimal control of "bang-bang" type does not exist and the magnitude of the optimal control function is equal to the constraints only on some subintervals of the interval of control.

1. Problem statement. In the paper [5] the optimal control of a system described by a symmetric hyperbolic equation is investigated. The control function is a finite-dimensional one and its Euclidean norm is constrained. The time of control is given and the performance index has the form of some quadratic functional of integral form of the terminal state. It is shown in [5] that the optimal control exists and the necessary conditions of optimality in the form of some maximum principle are formulated. Moreover, in that paper the conditions for optimal control to be of "bang-bang" type [2] are analyzed.

By analogy with the systems described by ordinary differential equations these conditions are called by the author the normality conditions. In contradistinction to the ordinary differential equation the normality of the analyzed systems depends on initial and terminal states.

As an example of the application of these results the problem of minimization of total energy of the vibrating string is investigated in [5]. The normality conditions for this case are obtained, but they are characterized by means of the terminal state of the system which is not given a priori.

In the present paper the problem of determination of this optimal control of the vibrating string is investigated. The direct method analogous to that used in [4] is applied. It is shown that this problem is equivalent to the problem of minimization of some integral functional subject to the constraints on the magnitude of the argument.

The method of Lagrange functionals [1], [3] is applied to solve this last problem and the form of the optimal control function is characterized. It appears that in general an optimal control of "bang-bang" type does not exist. Generally the magnitude of the optimal control function is equal to the constraints only in some subintervals of the control interval.

* Received by the editors August 1, 1968, and in revised form October 1, 1968.

† Department of Engineering, University of California, Los Angeles, California 90024. This research was supported in part by the Air Force Office of Scientific Research, Applied Mathematics Division, United States Air Force, under Grant AF-68-1408.

Our problem can be formulated as follows: There is given a system described by the equation

$$\rho\frac{\partial^2 w}{\partial t^2} - \tau\frac{\partial^2 w}{\partial x^2} = 0, \tag{1}$$

and the following boundary conditions are satisfied:

$$w(0,t) = 0, \qquad w(l,t) = f(t). \tag{2}$$

We shall require that $df(t)/dt$ be an absolutely continuous function whose derivative $u(t) = d^2f(t)/dt^2$, which exists almost everywhere, satisfies the condition

$$|u(t)| \leqq 1. \tag{3}$$

The function $u(t)$ will be called a control function.

We introduce the change of variable

$$w(x,t) = y(x,t) - \frac{x}{l}f(t). \tag{4}$$

Hence (1) takes on the form

$$\rho\frac{\partial^2 y}{\partial t^2} - \tau\frac{\partial^2 y}{\partial x^2} = \rho\frac{x}{l}u(t) \tag{5}$$

along with the boundary conditions

$$y(0,t) \equiv 0, \qquad y(l,t) \equiv 0. \tag{6}$$

Putting

$$y_1(x,t) = \frac{\partial y(x,t)}{\partial t}, \qquad y_2(x,t) = \frac{\partial y(x,t)}{\partial x}, \tag{7}$$

we rewrite (5) in the form

$$\begin{gathered} \rho\frac{\partial y_1}{\partial t} = \tau\frac{\partial y_2}{\partial x} + \rho\frac{x}{l}u(t), \\ \frac{\partial y_2}{\partial t} = \frac{\partial y_1}{\partial x}. \end{gathered} \tag{8}$$

The boundary conditions (6) imply

$$y_1(0,t) \equiv 0, \qquad y_1(l,t) \equiv 0. \tag{9}$$

Moreover the following initial conditions are given:

$$y_1(x,0) = y_1^0(x), \qquad y_2(x,0) = y_2^0(x), \tag{10}$$

where $y_1^0(x)$ and $y_2^0(x)$ are absolutely continuous functions and satisfy

$$y_1^0(0) = 0, \qquad y_1^0(l) = 0. \tag{10a}$$

Our aim is to find a control function $u(t)$ satisfying (3) such that at the given time T the vibrating energy of the system (8)–(10) given by

$$E_v(y_1, y_2, T) = \frac{1}{2}\int_0^l [\rho y_1^2(x, T) + \tau y_2^2(x, T)]\, dx \tag{11}$$

is minimal.

It was shown in [5] that the solution of this problem exists and that the terminal state of the system corresponding to the optimal control is determined uniquely.

In the sequel a method of determining this optimal control will be given.

2. Determination of the terminal state. Without any loss of generality we can put

$$\rho = \tau = 1, \qquad l = \pi. \tag{12}$$

Using the Fourier technique we obtain [6, Chaps. 22, 23], [7, Chap. 2] the following formulas for the generalized solution of (8) under the conditions (9) and (10) at the time T:

$$\begin{aligned} y_1(x, T) = y_1^f(x) &= \frac{2}{\pi}\sum_{n=1}^{\infty}\left[(-1)^{n+1}\frac{1}{n}\sin nx \int_0^T \cos n(T - t)u(t)\, dt\right] \\ &+ \sum_{n=1}^{\infty}(-\psi_n^0 \sin nT + \phi_n^0 \cos nT)\sin nx, \end{aligned} \tag{13a}$$

$$\begin{aligned} y_2(x, T) = y_2^f(x) &= \frac{2}{\pi}\sum_{n=1}^{\infty}\left[(-1)^{n+1}\frac{1}{n}\cos nx \int_0^T \sin n(T - t)u(t)\, dt\right] \\ &+ \sum_{n=1}^{\infty}(\psi_n^0 \cos nT + \phi_n^0 \sin nT)\cos nx. \end{aligned} \tag{13b}$$

Here ϕ_n^0 and ψ_n^0 are the Fourier coefficients of the functions $y_1^0(x)$ and $y_2^0(x)$, respectively. These functions are defined on the interval $[0, \pi]$ and extended to $(-\infty, +\infty)$ as odd and even periodic functions, respectively, i.e.,

$$y_1^0(x) = \sum_{n=1}^{\infty}\phi_n^0 \sin nx, \tag{14a}$$

$$y_2^0(x) = \sum_{n=1}^{\infty}\psi_n^0 \cos nx, \qquad x \in (-\infty, +\infty). \tag{14b}$$

In the same way we extend the definition of $y_1^f(x)$ and $y_2^f(x)$ to $(-\infty, +\infty)$ requiring

$$y_1^f(-x) = -y_1^f(x), \qquad y_1^f(x + 2\pi i) = y_1^f(x), \tag{14c}$$

$$y_2^f(-x) = y_2^f(x), \qquad y_2^f(x + 2\pi i) = y_2^f(x). \tag{14d}$$

Taking into consideration the identities

$$(-1)^{n+1}\frac{1}{n}\sin nx = -\int_0^x \cos n(\xi+\pi)\,d\xi$$

and

$$(-1)^{n+1}\frac{1}{n}\cos nx = \int_0^x \sin n(\xi+\pi)\,d\xi + (-1)^{n+1}\frac{1}{n}$$

and performing elementary transformations, we rewrite (13) in the following form:

$$\begin{aligned} y_1^f(x) = {} & -2\int_0^x \sum_{n=1}^{\infty}\left(\cos n(\xi+\pi)\frac{1}{\pi}\int_0^T \cos nt\,u(T-t)\,dt\right)d\xi \\ & + \tfrac{1}{2}[y_2^0(T+x) - y_2^0(T-x)] + \tfrac{1}{2}[y_1^0(T+x) - y_1^0(T-x)], \end{aligned} \tag{15a}$$

$$\begin{aligned} y_2^f(x) = {} & 2\int_0^x \sum_{n=1}^{\infty}\left(\sin n(\xi+\pi)\frac{1}{\pi}\int_0^T \sin nt\,u(T-t)\,dt\right)d\xi \\ & + 2\sum_{n=1}^{\infty}(-1)^{n+1}\frac{1}{n\pi}\int_0^T \sin n(T-t)\,u(t)\,dt \\ & + \tfrac{1}{2}[y_2^0(T+x) + y_2^0(T-x)] + \tfrac{1}{2}[y_1^0(T+x) + y_1^0(T-x)]. \end{aligned} \tag{15b}$$

Let

$$T = 2\pi m + \varepsilon, \tag{16}$$

where $0 < \varepsilon < 2\pi$, and let us put

$$\begin{aligned} & \tfrac{1}{2}[y_2^0(T+x) - y_2^0(T-x)] + \tfrac{1}{2}[y_1^0(T+x) - y_1^0(T-x)] \\ & \quad = \tfrac{1}{2}[y_2^0(\varepsilon+x) - y_2^0(\varepsilon-x)] + \tfrac{1}{2}[y_1^0(\varepsilon+x) - y_1^0(\varepsilon-x)] = f_1(x), \end{aligned} \tag{17a}$$

$$\begin{aligned} & \tfrac{1}{2}[y_2^0(T+x) + y_2^0(T-x)] + \tfrac{1}{2}[y_1^0(T+x) + y_1^0(T-x)] \\ & \quad = \tfrac{1}{2}[y_2^0(\varepsilon+x) + y_2^0(\varepsilon-x)] + \tfrac{1}{2}[y_1^0(\varepsilon+x) + y_1^0(\varepsilon-x)] = f_2(x), \end{aligned} \tag{17b}$$

$$2\sum_{n=1}^{\infty}(-1)^{n+1}\frac{1}{n\pi}\int_0^T \sin n(T-t)\,y(t)\,dt = k. \tag{17c}$$

We introduce the periodic functions $u_i(x)$ defined on the interval $(0, 2\pi)$ as follows:

$$\begin{aligned} & u_i(x) \triangleq u(t)|_{t=2\pi(i-1)+x}, && i = 1, 2, \cdots, m, \\ & u_{m+1}(x) \triangleq \begin{cases} u(t)|_{t=2\pi m+x} & \text{for } 0 < x < \varepsilon, \\ 0 & \text{for } \varepsilon < x < 2\pi. \end{cases} \end{aligned} \tag{18}$$

Hence equations (15) can be rewritten in the form

$$\text{(19a)}\quad y_1^f(x) = -2\int_0^x \left[\sum_{n=1}^{\infty} \cos n(\xi + \pi) \sum_{i=1}^{m+1} \frac{1}{\pi}\int_0^{2\pi} u_i(\varepsilon - x)\cos nx\,dx\right] d\xi + f_1(x),$$

$$\text{(19b)}\quad y_2^f(x) = 2\int_0^x \left[\sum_{n=1}^{\infty} \sin n(\xi + \pi) \sum_{i=1}^{m+1} \frac{1}{\pi}\int_0^{2\pi} u_i(\varepsilon - x)\sin nx\,dx\right] d\xi + k + f_2(x).$$

Note that the formulas

$$\frac{1}{\pi}\int_0^{2\pi} u_i(\varepsilon - x)\sin nx\,dx \quad\text{and}\quad \frac{1}{\pi}\int_0^{2\pi} u_i(\varepsilon - x)\cos nx\,dx$$

are the respective Fourier coefficients of the function $u_i(\varepsilon - x)$. Hence taking into consideration that each of the functions $u_i(\varepsilon - x)$ is equal to the sum of its Fourier series almost everywhere (as a bounded function), we can rewrite (19) in the form

$$\text{(20a)}\quad \begin{aligned} y_1^f(x) &= -2\int_0^x \left[\sum_{i=1}^{m+1} u_i^e(\varepsilon - \pi - \xi) - c\right] d\xi + f_1(x) \\ &= -2\int_0^x \left[\sum_{i=1}^{m+1} u_i^e(\xi - \varepsilon + \pi) - c\right] d\xi + f_1(x), \end{aligned}$$

$$\text{(20b)}\quad \begin{aligned} y_2^f(x) &= 2\int_0^x \left[\sum_{i=1}^{m+1} u_i^o(\varepsilon - \pi - \xi)\right] d\xi + k + f_2(x) \\ &= -2\int_0^x \left[\sum_{i=1}^{m+1} u_i^o(\xi - \varepsilon + \pi)\right] d\xi + k + f_2(x). \end{aligned}$$

The superscripts e and o denote here respectively the even and odd component of the corresponding function, and the constant

$$c = \sum_{i=1}^{m+1} \frac{1}{2\pi}\int_0^{2\pi} u_i(x)\,dx$$

is an arbitrary parameter. Denoting

$$\text{(21)}\quad \sum_{i=1}^{m+1} u_i(x - \varepsilon + \pi) = v(x)$$

we can rewrite (20) as follows:

$$\text{(22a)}\quad y_1^f(x) = -2\int_0^x [v^e(\xi) - c]\,d\xi + f_1(x),$$

$$\text{(22b)}\quad y_2^f(x) = 2\int_0^x v^o(\xi)\,d\xi + k + f_2(x).$$

Putting $x = 0$ we obtain from (22b)

$$y_2^f(0) = k + f_2(0);$$

then (22b) can be rewritten in the form

$$y_2^f(x) - y_2^f(0) = -2\int_0^x v^o(\xi)\,d\xi + f_2(x) - f_2(0). \tag{23}$$

Integrating this equation over the interval $[0, \pi]$ and taking into consideration (7) and (9) we obtain

$$y_2^f(0) = \frac{1}{\pi}\int_0^\pi \left[2\int_0^x v^o(\xi)\,d\xi - f_2(x) + f_2(0)\right] dx. \tag{24}$$

Hence eventually (23) takes on the form

$$\begin{aligned} y_2^f(x) &= -2\int_0^x v^o(\xi)\,d\xi + f_2(x) - f_2(0) \\ &\quad + \frac{1}{\pi}\int_0^\pi \left[2\int_0^\pi v^o(\xi)\,d\xi - f_2(x) + f_2(0)\right] dx. \end{aligned} \tag{25}$$

3. Nature of optimal control. It follows from (11), (12) and (14) that the vibrational energy of the system (8) at the time T is given by

$$E(y, T) = J(v^e, v^o) = \frac{1}{4}\int_0^{2\pi} \{[y_1^f(x)]^2 + [y_1^f(x)]^2\}\,dx. \tag{26}$$

Substituting (22a) and (25) for (26) we obtain the vibrational energy expressed in terms of $v^e(x)$ and $v^o(x)$. On the other hand it follows from (3), (18) and (21) and from the fact that $v(x)$ is a periodic function that $v(x)$ is subject to the following constraints:

$$|v(x)| = |v^e(x) + v^o(x)| \leqq \phi(x) \quad \text{almost everywhere for } x \in [0, 2\pi], \tag{27}$$

where the periodic function $\phi(x)$ is defined as follows:

$$\phi(x + \varepsilon - \pi) = \begin{cases} m + 1 & \text{for } x \in [0, \varepsilon), \\ m & \text{for } x \in [\varepsilon, 2\pi). \end{cases} \tag{27a}$$

Hence the problem of minimization of vibrational energy is equivalent to the determination of a function $v(x)$ satisfying (27) which minimizes functional (26). To solve this problem we shall use the method of Lagrange functionals given in [1] and [3].

The constraints (27) can be rewritten in the form

$$\begin{aligned} [g_1(v^e, v^o)](x) &= v^e(x) + v^o(x) - \phi(x) \leqq 0, \\ [g_2(v^e, v^o)](x) &= -v^e(x) - v^o(x) - \phi(x) \leqq 0, \end{aligned} \tag{28}$$

almost everywhere for $x \in [0, 2\pi]$.

The pair of functions (v^e, v^o) can be treated as an element of the space $L^2[0, 2\pi] \times L^2[0, 2\pi]$, and the formula

$$G(v^e, v^o) = [g_1(v^e, v^o), g_2(v^e, v^o)]$$

can be considered as describing an operator from $L^2[0, 2\pi] \times L^2[0, 2\pi]$ into the same space.

Let us introduce a Lagrange function of the form

$$\begin{aligned}\Phi(v^e, v^o, \lambda) &= J(v^e, v^o) + \lambda[G(v^e, v^o)] \\ &= J(v^e, v^o) + \lambda_1[g_1(v^e, v^o)] + \lambda_2[g_2(v^e, v^o)],\end{aligned} \tag{29}$$

where $\lambda = (\lambda_1, \lambda_2)$ is a linear functional defined on the space $L^2[0, 2\pi] \times L^2[0, 2\pi]$.

The functional $J(v^e, v^o)$ and the operator $G(v^e, v^o)$ are convex (in the sense of natural order in the spaces $L^2[0, 2\pi]$). Moreover $G(v^e, v^o)$ satisfies the regularity conditions (see [1] and [3]). Hence it follows from [1] and [3] that the functions $v_0^e(x)$ and $v_0^o(x)$ minimize the functional (26) subject to (28) if and only if there exist functionals λ_1^0 and λ_2^0, such that the conditions (28) are satisfied and

$$\begin{aligned}d_{v^e}\Phi[(v_0^e, v_0^o); v^e] &= d_{v^e}J[(v_0^e, v_0^o); v^e] + \lambda_1^0\{d_{v^e}g[(v_0^e, v_0^o); v^e]\} \\ &\quad + \lambda_2^0\{d_{v^e}g[(v_0^e, v_0^o); v^e]\} = 0,\end{aligned} \tag{30a}$$

$$\begin{aligned}d_{v^o}\Phi[(v_0^e, v_0^o); v^o] &= d_{v^o}J[(v_0^e, v_0^o); v^o] + \lambda_1^0\{d_{v^o}g[(v_0^e, v_0^o); v^o]\} \\ &\quad + \lambda_2^0\{d_{v^o}g[(v_0^e, v_0^o); v^o]\} = 0,\end{aligned} \tag{30b}$$

$$\lambda_i^0 \geqq 0, \qquad i = 1, 2, \tag{30c}$$

$$\lambda_i^0[g_i(v_0^e, v_0^o)] = 0, \qquad i = 1, 2. \tag{30d}$$

The formulas $d_{v^e}J[(v_0^e, v_0^o); v^e]$ and $d_{v^e}g[(v_0^e, v_0^o); v^e]$ denote here the Fréchet differentials of the respective functional and operator at the point (v_0^e, v_0^o) in the direction $(v^e, 0)$.

The inequality in (30c) means that the respective functional assumes nonnegative values on nonnegative (in the sense of natural order) elements of the space $L^2[0, 2\pi]$.

Note that since the functional $J(v^e, v^o)$ is strictly convex, the functions $v_0^e(x)$ and $v_0^o(x)$ which minimize it are defined uniquely. From (22a) and (26) we obtain

$$\begin{aligned}d_{v^e}J[(v_0^e, v_0^o); v^e] &= -\int_0^{2\pi}\left\{-2\int_0^x \left[v_0^e(\xi) - c\right] d\xi + f_1(x)\right\}\int_0^x v^e(\xi)\, d\xi\, dx \\ &= -\int_0^{2\pi} y_{1,0}^f(x)\int_0^x v^e(\xi)\, d\xi\, dx \\ &= -\int_0^{2\pi}\int_x^{2\pi} y_{1,0}^f(\xi)\, d\xi\, v^e(x)\, dx.\end{aligned} \tag{31}$$

In the same way, from (25) and (26) we obtain

$$\begin{aligned}d_{v^o}J[(v_0^e, v_0^o); v^o] &= -\int_0^{2\pi} y_{2,0}^f(x)\left[\int_0^x v^o(\xi)\, d\xi - \int_0^{2\pi}\frac{2}{\pi}\int_0^x v^o(\xi)\, d\xi\right] dx \\ &= -\int_0^{2\pi} y_{2,0}^f(x)\int_0^x v^o(\xi)\, d\xi\, dx \\ &= -\int_0^{2\pi}\int_x^{2\pi} y_{2,0}^f(\xi)\, d\xi\, v^o(x)\, dx.\end{aligned} \tag{32}$$

From (28) we have

$$d_{v^e} g_1[(v_0^e, v_0^o); v^e] = v^e(x), \tag{33a}$$

$$d_{v^e} g_2[(v_0^e, v_0^o); v^e] = -v^e(x). \tag{33b}$$

We obtain exactly the same formulas for the differentials of the operators g_i in the direction $(0, v^o)$.

Substituting (31)–(33) into (30) and using the general form of nonnegative functional in the space $L^2[0, 2\pi]$, we obtain

$$-\int_0^{2\pi}\int_x^{2\pi} y_{1,0}^f(\xi)\,d\xi\, v^e(x)\,dx + \int_0^{2\pi} \lambda_1^0(x)v^e(x)\,dx - \int_0^{2\pi} \lambda_2^0(x)v^e(x)\,dx = 0, \tag{34a}$$

$$-\int_0^{2\pi}\int_x^{2\pi} y_{2,0}^f(\xi)\,d\xi\, v^o(x)\,dx + \int_0^{2\pi} \lambda_1^0(x)v^o(x)\,dx - \int_0^{2\pi} \lambda_2^0(x)v^o(x)\,dx = 0, \tag{34b}$$

where

$$\lambda_i^0(x) \geqq 0 \quad \text{almost everywhere for } x \in [0, 2\pi], \qquad i = 1, 2, \tag{34c}$$

$$\begin{aligned} &\int_0^{2\pi} [v_0^e(x) + v_0^o(x) - \phi(x)]\lambda_1^0(x)\,dx = 0, \\ &\int_0^{2\pi} [-v_0^e(x) - v_0^o(x) - \phi(x)]\lambda_2^0(x)\,dx = 0. \end{aligned} \tag{34d}$$

Since (34a) and (34b) must be satisfied for arbitrary functions $v^o, v^e \in L^2[0, 2\pi]$, we obtain

$$-\int_x^{2\pi} y_{1,0}^f(\xi)\,d\xi + \lambda_1^0(x) - \lambda_2^0(x) = 0, \tag{35a}$$

$$-\int_x^{2\pi} y_{2,0}^f(\xi)\,d\xi + \lambda_1^0(x) - \lambda_2^0(x) = 0 \tag{35b}$$

almost everywhere on $[0, 2\pi]$. Subtracting (35b) from (35a) we obtain

$$\int_x^{2\pi} y_{1,0}^f(\xi)\,d\xi = \int_x^{2\pi} y_{2,0}^f(\xi)\,d\xi.$$

Hence

$$y_{1,0}^f(x) = y_{2,0}^f(x). \tag{36}$$

Therefore taking (9) into consideration we have

$$y_{2,0}^f(0) = 0. \tag{37}$$

Summing (35a) and (35b) we obtain

$$-\int_x^{2\pi} [y_{1,0}^f(\xi) + y_{2,0}^f(\xi)]\,d\xi + 2\lambda_1^0(x) - 2\lambda_2^0(x) = 0. \tag{38}$$

From (17), (22a), (23) and (37) we obtain

$$
\begin{aligned}
y^f_{1,0}(x) + y^f_{2,0}(x) &= \int_0^x - 2[v_0(\xi) - c]\,d\xi + f_1(x) + f_2(x) - f_2(0) \\
&= \int_0^x - 2[v_0(\xi) - c]\,d\xi + y_2^0(\varepsilon + x) + y_1^0(\varepsilon + x) - y_2^0(\varepsilon) - y_1^0(\varepsilon) \\
&= \int_0^x \left[\frac{dy_1^0(\varepsilon + \xi)}{d\xi} + \frac{dy_2^0(\varepsilon + \xi)}{d\xi} + 2c - 2v_0(\xi)\right] d\xi \\
&= \int_0^x [F(\xi) + 2c - 2v_0(\xi)]\,d\xi,
\end{aligned} \tag{39}
$$

where

$$
F(x) = \frac{dy_1^0(\varepsilon + x)}{dx} + \frac{dy_2^0(\varepsilon + x)}{dx}. \tag{39a}
$$

Taking into consideration (36) and (39) we find that the minimal value of functional (26) can be expressed in the form

$$
J(v_0^e, v_0^o) = \frac{1}{8}\int_0^{2\pi} \left\{ \int_0^x [F(\xi) + 2c - 2v_0(\xi)]\,d\xi \right\}^2 dx. \tag{40}
$$

The value of the functional (40) depends on the constant c, which is an arbitrary parameter. This constant ought to be chosen in such a way that the value of functional (40) is minimal. Let us denote by c_0 the optimal value of c. Differentiating (40) with respect to c we obtain the following necessary condition for the optimality of $v_0(x)$:

$$
\int_0^{2\pi} \int_0^x [F(\xi) + 2c_0 - 2v_0(\xi)]\,d\xi\,dx = 0. \tag{41}
$$

Note that it follows from (27) that the minimal value of (40) is equal to zero if and only if

$$
\max_{x \in [0,2\pi]} [F(x) - 2\phi(x)] \leqq -2c_0 \leqq \min_{x \in [0,2\pi]} [F(x) + 2\phi(x)]. \tag{42}
$$

Let us assume that the condition (42) is not satisfied.

For more detailed characterization of the optimal function $v_0(x)$ we put $c = c_0$ and we substitute (39) into (38):

$$
-\int_x^{2\pi} \int_0^{\xi} [F(\eta) + 2c_0 - 2v_0(\eta)]\,d\eta\,d\xi + 2\lambda_1^0(x) - 2\lambda_2^0(x) = 0. \tag{43}
$$

Let us introduce the sets

$$
M = \{x \in [0, 2\pi] = v_0(x) - \phi(x) = 0\}, \tag{44a}
$$

$$
N = \{x \in [0, 2\pi] = -v_0(x) - \phi(x) = 0\}. \tag{44b}
$$

It follows from (28), (30d) and (44) that

$$\lambda_1^0(x) = 0 \quad \text{for } x \in [0, 2\pi] \backslash M, \tag{45a}$$

$$\lambda_2^0(x) = 0 \quad \text{for } x \in [0, 2\pi] \backslash N. \tag{45b}$$

Hence taking into consideration (34c) and (45) we obtain, from (43),

$$\int_x^{2\pi} \int_0^{\xi} [F(\eta) + 2c_0 - 2v_0(\eta)]\, d\eta\, d\xi \geqq 0 \quad \text{for } x \in M, \tag{46a}$$

$$\int_x^{2\pi} \int_0^{\xi} [F(\eta) + 2c_0 - 2v_0(\eta)]\, d\eta\, d\xi \leqq 0 \quad \text{for } x \in N, \tag{46b}$$

$$\int_x^{2\pi} \int_0^{\xi} [F(\eta) + 2c_0 - 2v_0(\eta)]\, d\eta\, d\xi = 0 \quad \text{for } x \in [0, 2\pi] \backslash (M \cup N). \tag{46c}$$

Note that the left-hand side of (46) is a continuous function of x. Therefore taking into account (41) we conclude that for each component M_i of M the following equation must be satisfied:

$$\int_{M_i} \int_0^{\xi} [F(\eta) + 2c_0 - 2v_0(\eta)]\, d\eta\, d\xi = 0. \tag{47}$$

The same equation must be satisfied for each component N_i of the set N. Then the points at which $F(x) + 2c_0 \geqq \phi(x)$ as well as such that $F(x) + 2c_0 \leqq \phi(x)$ must belong to each subset M_i. Therefore if $F(x)$ is piecewise monotonic, the number of the components M_i must be finite. Similarly in this case the number of components N_i is finite.

The effective determining of the sets M and N is difficult and in each case it has to be done numerically making use of the properties (46) and (47). The additional difficulty is that we do not know the optimal value of the parameter c, which has to be determined taking advantage of the condition (41).

Let us now assume that we have found the sets M and N and let us determine the form of optimal control.

It follows from (41), (46c) and (47) that

$$v_0(x) = \tfrac{1}{2}F(x) + c_0 \quad \text{for } x \in [0, 2\pi] \backslash (M \cup N). \tag{48}$$

The function $v_0(x)$ is uniquely determined by (44) and (48). From (18), (21), (27) and (44) we obtain

$$u_{0,i}(x) = \begin{cases} +1 & \text{for } x \in M' = \{x \in [0, 2\pi] : v_0(x + \varepsilon - \pi) = \phi(x + \varepsilon - \pi)\}, \\ -1 & \text{for } x \in N' = \{x \in [0, 2\pi] : v_0(x + \varepsilon - \pi) = -\phi(x + \varepsilon - \pi)\}, \\ & \qquad i = 1, 2, \cdots, m. \end{cases} \tag{49}$$

For $u_{0,m+1}(x)$ the same formulas are satisfied, but instead of the sets M' and N' we have $M' \cap (0, \varepsilon)$ and $N' \cap (0, \varepsilon)$, respectively. Hence on these sets the functions $u_{0,i}(x)$ are defined uniquely.

On the other hand it follows from (18), (21) and (48) that for $x \in [0, 2\pi] \backslash (M' \cup N')$ the functions $u_{0,i}(x)$ are not defined uniquely. They may be arbitrary. However, the condition (3) and the equations

$$\text{(50a)} \qquad \sum_{i=1}^{m+1} u_{0,i}(x) = \tfrac{1}{2}F(x + \varepsilon - \pi) + c_0 \quad \text{for } x \in \{[0, 2\pi]\backslash(M' \cup N')\} \cap (0, \varepsilon),$$

$$\text{(50b)} \qquad \sum_{i=1}^{m} u_{0,i}(x) = \tfrac{1}{2}F(x + \varepsilon - \pi) + c_0 \quad \text{for } x \in \{[0, 2\pi]\backslash(M' \cup N')\} \cap (\varepsilon, 2\pi)$$

must be satisfied.

In particular, we can put

$$\text{(51)} \qquad u_{0,i}(x) = u_{0,j}(x), \qquad \begin{cases} i, j = 1, 2, \cdots, m + 1 & \text{for } x \in (0, \varepsilon), \\ i, j = 1, 2, \cdots, m & \text{for } x \in (\varepsilon, 2\pi). \end{cases}$$

Hence it follows from (18) that the optimal control $u_0(t)$ is a periodic function. From (18), (49), (50) and (51) we obtain

$$u[t + 2(i - 1)\pi]$$

$$\text{(52)} \qquad = \begin{cases} +1 & \text{for } t = x \in M', \\ -1 & \text{for } t = x \in N', \\ \dfrac{1}{m + 1}[F(x)|_{x=t} + c_0] & \text{for } t = x \in \{[0, 2\pi]\backslash(M' \cup N')\} \cap (0, \varepsilon), \\ \dfrac{1}{m}[F(x)|_{x=t} + c_0] & \text{for } t = x \in \{[0, 2\pi]\backslash(M' \cup N')\} \cap (\varepsilon, 2\pi). \end{cases}$$

It follows from the previous considerations that, in the case where $M \cup N = [0, 2\pi]$, the optimal control is unique and that it is of the "bang-bang" type. Naturally this is the case when the interval $[0, 2\pi]$ does not contain any subset of positive measure on which (46c) is satisfied. Taking into consideration (7), (36), (39) and (47) we see that this condition is equivalent to the requirement that there not exist any subset of positive measure on which

$$\text{(53)} \qquad y_{1\,0}^{f}(x) = \left.\frac{\partial y_0(x, t)}{\partial t}\right|_{t=T} = 0.$$

This condition was obtained in [5] and it was called the condition of normality of the system. It follows from the previous considerations that in general this condition is not satisfied.

On the other hand, except for the case where the final vibrational energy is equal to zero, the set $M \cup N$ must be of positive measure. Hence the magnitude of $|u_0(t)|$ is equal to 1 on a subinterval of the interval of control $[0, T]$ having positive measure. Therefore it is possible to formulate the maximum principle, as it was done in [5].

REFERENCES

[1] K. J. Arrow, L. Hurwicz and H. Uzawa, *Studies in Linear and Non-linear Programming*, Stanford University Press, Stanford, California, 1958, Chap. 4.

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

[3] J. Majerczyk-Gomulka and K. Makowski, *Determination of optimal control of dynamic processes by means of Lagrange functionals*, Institute Basic Technical Problems Reports, no. 9, Warsaw, 1967.

[4] K. Malanowski, *On time optimal control of vibrating string*, Avtomat. i. Telemekh., to appear.

[5] D. L. Russel, *Optimal regulation of linear symmetric hyperbolic systems with finite dimensional controls*, this Journal, 4 (1966), pp. 276–294.

[6] S. L. Sobolev, *Partial Differential Equations of Mathematical Physics*, Pergamon Press, New York, 1964.

[7] A. N. Tikhonov and A. A. Samarsky, *Equations of Mathematical Physics*, Macmillan, New York, 1963.

SIAM J. CONTROL
Vol. 7, No. 2, May 1969

ON SECOND ORDER NECESSARY CONDITIONS OF OPTIMALITY*

E. J. MESSERLI AND E. POLAK†

1. Introduction.

In the last few years it has been shown [1], [2] that most of the problems of nonlinear programming, the calculus of variations and optimal control can be treated in a unified manner by transcribing these problems into a simple canonical form. Necessary conditions of optimality for this canonical form may then be obtained and related to the original problems through the structure of each particular problem.

For finite-dimensional problems, this canonical form is given as follows.

BASIC PROBLEM. Let $f:E^n \to E^1$ and $r:E^n \to E^m$ be continuously differentiable functions, and let Ω be a subset of E^n. Find a vector $\hat{x}$ in E^n such that (i) $\hat{x} \in \Omega$, $r(\hat{x}) = 0$ and (ii) for every x in Ω with $r(x) = 0$, $f(\hat{x}) \leqq f(x)$.

Following the convention of nonlinear programming, an $\hat{x}$ satisfying (i) will be called *feasible*, while an $\hat{x}$ satisfying both (i) and (ii) will be called an *optimal solution* to the Basic Problem.

A similar problem, common in mathematical programming, is perhaps better known.

NONLINEAR PROGRAMMING PROBLEM. Let $f:E^n \to E^1$, $r:E^n \to E^m$ and $g:E^n \to E^k$ be given functions. Find $\hat{x}$ such that $r(\hat{x}) = 0, g(\hat{x}) \leqq 0$ and $f(\hat{x}) = \min \{f(x)|r(x) = 0, g(x) \leqq 0\}$.

This problem may be put in Basic Problem form by identifying Ω as the set $\{x|g(x) \leqq 0\}$.

As a more interesting example, consider the following discrete optimal control problem: Let $f_i^0:E^l \to E^1, f_i:E^n \times E^l \to E^n, i = 0, 1, \cdots, k-1$, $g:E^n \to E^m$ be given functions, and U_i a given set in E^l for $i = 0, 1, \cdots, k-1$. Find a control sequence $(u_0, \cdots, u_{k-1})$ which minimizes $\sum_{i=0}^{k-1} f_i^0(u_i)$ subject to

(i) $y_{i+1} - y_i = f_i(y_i, u_i), i = 0, 1, \cdots, k-1$,

and

(ii) $y_0 = \hat{y}_0, g(y_k) = 0, u_i \in U_i, i = 0, 1, \cdots, k-1$.

To see that this problem may be cast in Basic Problem form, let x in E^{kl} be given by $x = (u_0, \cdots, u_{k-1})$, $f(x) = \sum_{i=0}^{k-1} f_i^0(u_i)$, $r(x) = g(y_k(x))$, where y_k is obtained by solving (i) with $y_0 = \hat{y}_0$, and, finally, $\Omega = U_0 \times U_1 \times \cdots \times U_{k-1}$.

The demonstrated generality of the Basic Problem makes it a convenient vehicle for the introduction of second order conditions of optimality. By a second order condition of optimality, we mean a condition which augments or replaces the usual first order conditions and, generally, involves a second derivative of one or more of the cost or constraint functions.

* Received by the editors April 9, 1968, and in revised form January 6, 1969.

† Department of Electrical Engineering and Computer Sciences, Electronics Research Laboratory, University of California, Berkeley, California 94720. This research was supported by the National Aeronautics and Space Administration under Grant NsG-354, Suppl. 4.

First and second order conditions are not independent; a second order condition, usually, is only meaningful when a first order condition is already satisfied. To clarify these ideas, we shall state a fundamental first order necessary condition for the Basic Problem. This requires a tractable local representation of the set Ω.

DEFINITION 1 ([1]). A convex cone $C(\hat{x}, \Omega)$ will be called a *conical approximation* to the constraint set Ω at the point $\hat{x}$ if for any collection $\{y_1, y_2, \cdots, y_k\}$ of linearly independent vectors in $C(\hat{x}, \Omega)$, there exists an $\alpha_0 > 0$ (possibly depending on $\hat{x}, y_1, \cdots, y_k$) and a continuous map $\zeta(\cdot)$ from the convex hull of $\{\alpha y_1, \cdots, \alpha y_k\}$ into $\Omega - \hat{x}$, for each $0 \leqq \alpha \leqq \alpha_0$, of the form

$$\zeta(y) = y + o(y), \tag{1}$$

where $\|o(y)\|/\|y\| \to 0$ as $\|y\| \to 0$.

If the map $\zeta(\cdot)$ is given by $\zeta(y) = y$, then $C(\hat{x}, \Omega)$ will be called a *simple* conical approximation to Ω at $\hat{x}$.

THEOREM 1 ([1]). *If $\hat{x}$ is an optimal solution to the Basic Problem and $C(\hat{x}, \Omega)$ is a conical approximation to Ω at $\hat{x}$, then there is a nonzero vector $\psi = (\psi^0, \cdots, \psi^m)$ in E^{m+1}, with $\psi^0 \leqq 0$, such that*[1]

$$\psi^0 \frac{\partial f}{\partial x}(\hat{x})(y) + \sum_{i=1}^{m} \psi^i \frac{\partial r^i}{\partial x}(\hat{x})(y) \leqq 0 \tag{2}$$

for all y in $\overline{C(\hat{x}, \Omega)}$, the closure in E^n of $C(\hat{x}, \Omega)$.

The inequality (2) may be satisfied under several different circumstances. The first is when the multiplier ψ^0 must be chosen to be zero, and hence no information about the cost function $f(\cdot)$ enters into the necessary condition (2). This occurs most often when there is only one x in Ω satisfying $r(x) = 0$ and may be avoided by introducing a regularity condition, usually called a constraint qualification [3] on $r(\cdot)$ and Ω. We shall not be concerned with this case.

A second situation for which (2) is always satisfied is that in which the vectors $\partial f(\hat{x})/\partial x$, $\partial r^1(\hat{x})/\partial x, \cdots, \partial r^m(\hat{x})/\partial x$ are linearly dependent, since one can then always choose a $\psi \neq 0$ which satisfies $\psi^0[\partial f(\hat{x})/\partial x] + \sum_{i=1}^m \psi^i[\partial r^i(\hat{x})/\partial x] = 0$, and hence (2), without reference to the optimality of $\hat{x}$. This situation does not usually lead to a second order condition unless it arises from the fact that one or more of the gradients $\partial f(\hat{x})/\partial x$, $\partial r^1(\hat{x})/\partial x, \cdots, \partial r^m(\hat{x})/\partial x$ are the zero vector. When this particular situation occurs, we shall refer to it as the *zero-gradient case.*

There is another situation in which a second order condition is meaningful, which is quite distinct from the zero-gradient case. This occurs when for every y contained in $C(\hat{x}, \Omega)$, we have $\partial f(\hat{x})(y)/\partial x = 0$ and $\partial r^i(\hat{x})(y)/\partial x = 0$ for $i = 1, \cdots, m$,

[1] It is to be understood throughout that any derivatives used are assumed to exist. The expansion of a function $f(\cdot)$ at the point $\hat{x}$ will be denoted by

$$f(\hat{x} + y) = f(\hat{x}) + \frac{\partial f}{\partial x}(\hat{x})(y) + \frac{1}{2}\frac{\partial^2 f}{\partial x^2}(\hat{x})(y, y) + \cdots;$$

$\partial f(\hat{x})/\partial x$ represents the gradient of $f(\cdot)$ at $\hat{x}$, assumed to be a column vector. Vectors will be written $x = (x^1, x^2, \cdots, x^n)$, etc., with the *exception that in E^{m+1} the component numbering will be from* $0, 1, 2, \cdots, m$.

Thus, it is possible to satisfy (2) irrespective of the choice of the vector ψ. Such vectors y are, in a sense, *critical* (see Definition 6), and second order conditions for this case correspond to examining second order effects along curves tangent to y at $\hat{x}$. Of course, any combination of the preceding three situations may occur simultaneously.

In § 2 of this paper, we survey briefly some of the known second order conditions for finite-dimensional spaces and show that they are second order conditions either for the zero-gradient case or for the critical directions case.

The major contributions of this paper are given in § 3—Theorem 6 for the critical directions case and Theorem 7 for the zero-gradient case. Both of these theorems are expressed in terms of local approximations to the set Ω, since, in well-formulated optimization problems, Ω has an interior, which ensures the existence of such approximations. Several ways by which such approximations may be constructed are given. It is also shown that most second order necessary conditions are special cases of Theorem 6 or Theorem 7.

In § 4, proofs for Theorems 6 and 7 are given. These proofs display the important fact that some of the proven techniques of first order theory, in particular the use of fixed-point theorems, can be applied to second order theory.

2. A brief survey of some second order conditions. Since our interest is in the Basic Problem, or the closely related Nonlinear Programming Problem, we shall not cover any special results from the calculus of variations [4], [5] or optimal control [6], [7, pp. 63–101]. Nor shall we be concerned with sufficiency conditions either, because in many cases the required strengthening of the necessary conditions may be obvious, or because our local approximation to the set Ω may not be sufficiently rich to describe completely the nature of Ω in the vicinity of $\hat{x}$.

Perhaps the simplest second order condition arises when the gradient of the cost function $f(\cdot)$ is zero at the optimal solution and we do not choose to isolate the equality constraints for special attention.

DEFINITION 2. Let Ω be an arbitrary set. The *tangent cone*, $TC(\hat{x}, \Omega)$, to Ω at $\hat{x}$ is defined to be the set of all y such that there exist a differentiable function $x:E^1 \to E^n$ and an $\bar{\alpha} > 0$ such that (i) $x(0) = \hat{x}$, (ii) $dx(0)/d\alpha = y$ and (iii) $x(\alpha) \in \Omega$ for $0 \leqq \alpha \leqq \bar{\alpha}$.

THEOREM 2. *If $\hat{x}$ is an optimal solution to the Basic Problem and $\partial f(\hat{x})/\partial x = 0$, then $\partial^2 f(\hat{x})(y, y)/\partial x^2 \geqq 0$ for all y in $\overline{TC(\hat{x}, \Omega')}$, where $\Omega' = \{x | r(x) = 0, x \in \Omega\}$.*

This is seen to be a simple but general result for the zero-gradient case; however its application depends on our having a characterization for $TC(\hat{x}, \Omega')$. In some cases, we may represent $TC(\hat{x}, \Omega')$ as the intersection of $TC(\hat{x}, \Omega')$ and $TC(\hat{x}, \{x | r(x) = 0\})$; this facilitates matters. However, this is not true in general.

Remark 1. Theorem 2 remains true if we replace the tangent cone, $TC(\hat{x}, \Omega)$, by the sequential tangent cone $STC(\hat{x}, \Omega)$, defined as follows.

DEFINITION 3 ([6]). Let Ω be an arbitrary set. The *sequential tangent cone*, $STC(\hat{x}, \Omega)$, to Ω at $\hat{x}$ is defined to be the set of all y such that there is a sequence $\{x_i\}_{i=1}^{\infty}$ in Ω and a sequence $\{d_i\}_{i=1}^{\infty}$ of strictly positive scalars such that (i) $x_i \to \hat{x}$ and (ii) $d_i(x_i - \hat{x}) \to y$.

For the critical vector case, we are able to obtain second order conditions without requiring the gradient of the cost function to be zero.

THEOREM 3. *Let $\hat{x}$ be an optimal solution to the Basic Problem, and let $x:E^1 \to E^n$ be any twice continuously differentiable function such that $x(0) = \hat{x}$ and $x(\alpha)$ is feasible for all $\alpha \in [0, \bar{\alpha}]$, with $\bar{\alpha} > 0$. If $df(x(0))/d\alpha = 0$, then $d^2f(x(0))/d\alpha^2 \geqq 0$.*

In general, without making additional assumptions about $r(\cdot)$ and Ω, the conditions of Theorems 2 and 3 cannot be decomposed into more structured forms.

One approach to a more structured condition is that followed by Dubovickii and Milyutin [8], [9]. Essentially, for a fixed $\tilde{y}$ satisfying $\partial f(\hat{x})(\tilde{y})/\partial x = 0$ and $\partial r^i(\hat{x})(\tilde{y})/\partial x = 0$ for $i = 1, \cdots, m$ (i.e., $\tilde{y}$ is critical) they consider the following sets:

(3) $C_0(\tilde{y}) = \{y|$ there exists an $\alpha_0 > 0$ and a function $o:[0, \alpha_0] \to E^n$, with $\lim_{\alpha \to 0} (\|o(\alpha)\|/\alpha) = 0$, such that $r(\hat{x} + \alpha\tilde{y} + \alpha^2 y + o(\alpha^2)) = 0$ for all $\alpha \in [0, \alpha_0]\}$.

(4) $C_1(\tilde{y}) = \{y|$ there exists an $\alpha_0' > 0$ such that $\hat{x} + \alpha\hat{y} + \alpha^2 u \in \Omega$ for all $\alpha \in [0, \alpha_0']\}$.

(5) $C_2(\tilde{y}) = \{y|$ there exists an $\alpha_0'' > 0$ such that $f(\hat{x} + \alpha\tilde{y} + a^2 y) < f(\hat{x})$ for all $\alpha \in (0, \alpha_0'']\}$.

THEOREM 4 ([9]). *If $\hat{x}$ is an optimal solution to the Basic Problem, then*[2] $C_0(\tilde{y}) \cap \operatorname{int}(C_1(\tilde{y})) \cap \operatorname{int}(C_2(\tilde{y})) = \emptyset$.

Whenever $C_0(\tilde{y})$ is a linear manifold and $C_1(\tilde{y})$ and $C_2(\tilde{y})$ are convex cones (possibly translated) with nonempty interiors, the condition $C_0(\tilde{y}) \cap \operatorname{int}(C_1(\tilde{y}) \cap \operatorname{int}(C_2(\tilde{y})) = \emptyset$ guarantees the existence of affine functionals, $c_0(\cdot)$, $c_1(\cdot)$, $c_2(\cdot)$, not all zero, with $c_0(\cdot)$ vanishing on $C_0(\tilde{y})$ and $c_i(\cdot)$ nonnegative on $C_i(\tilde{y})$ for $i = 1, 2$, such that $c_0(x) + c_1(x) + c_2(x) = 0$ for all x in E^n [8]. When specialized to the Nonlinear Programming Problem with rather restrictive assumptions, this gives a result similar to Theorem 5.

Finally, McCormick [10] has observed that in some cases the first order necessary conditions for the Nonlinear Programming Problem display a multiplier vector which can also be used in a second order condition.

DEFINITION 4 ([10]). Consider the Nonlinear Programming Problem. The *second order constraint qualification* is said to be satisfied at $\hat{x}$ if for each y such that $\partial r^i(\hat{x})(y)/\partial x = 0$ for $i = 1, \cdots, m$, and $\partial g^i(\hat{x})(y)/\partial x = 0$ for $i \in I(\hat{x}) \overset{\Delta}{=} \{i|g^i(\hat{x}) = 0\}$, there is a twice continuously differentiable function $x:E^1 \to E^n$ and $\bar{\alpha} > 0$ such that

(i) $x(0) = \hat{x}$, $dx(0)/d\alpha = y$,

(ii) for all $\alpha \in [0, \bar{\alpha}]$, $x(\alpha)$ is feasible, and moreover, $g^i(x(\alpha)) = 0$ for $i \in I(\hat{x})$.

THEOREM 5 ([10]). *If $\hat{x}$ is an optimal solution to the Nonlinear Programming Problem and the first order*[3] [10] *and second order constraint qualifications are*

[2] The interior of a set C is denoted by int(C).

[3] The first order constraint qualification is a statement of the Kuhn–Tucker constraint qualification [11] for a constraint set defined by both equalities and inequalities.

satisfied at $\hat{x}$, then there exist multipliers $\psi^1, \cdots, \psi^m$ and $\mu^1 \cdots, \mu^k$ with $\mu^i \leqq 0$ for $i = 1, \cdots, k$, such that

(i) $-\dfrac{\partial f}{\partial x}(\hat{x}) + \sum_{i=1}^{m} \psi^i \dfrac{\partial r^i}{\partial x}(\hat{x}) + \sum_{i=1}^{k} \mu^i \dfrac{\partial g^i}{\partial x}(\hat{x}) = 0,$

(ii) $\mu^i g^i(\hat{x}) = 0$ *for* $i = 1, 2, \cdots, k$

and

(iii) *for every $\tilde{y}$ such that $\partial r^i(\hat{x})(\tilde{y})/\partial x = 0$ for $i = 1, 2, \cdots, m$, and $\partial g^i(\hat{x})(\tilde{y})/\partial x = 0$ for $i \in I(\hat{x})$,*

$$-\frac{\partial^2 f}{\partial x^2}(\hat{x})(\tilde{y}, \tilde{y}) + \sum_{i=1}^{m} \psi^i \frac{\partial^2 r^i}{\partial x^2}(\hat{x})(\tilde{y}, \tilde{y}) + \sum_{i=1}^{k} \mu^i \frac{\partial^2 g^i}{\partial x^2}(\hat{x})(\tilde{y}, \tilde{y}) \leqq 0.$$

Conditions (i) and (ii) represent the first order necessary conditions for the Nonlinear Programming Problem, with the first order constraint qualification ensuring a nonzero cost multiplier, -1, in (i). Choosing $\tilde{y}$ such that $\partial r^i(\hat{x})(\tilde{y})/\partial x = 0$ for $i = 1, \cdots, m$ and $\partial g^i(\hat{x})(\tilde{y})/\partial x = 0$ for $i \in I(\hat{x})$, we see that (i) and (ii) imply $\partial f(\hat{x})(\tilde{y})/\partial x = 0$, i.e., $\tilde{y}$ is critical. The second order constraint qualification then leads to the third condition.

It is also clear, however, that $\partial f(\hat{x})(\tilde{y})/\partial x = 0$ also follows only from the optimality of $\hat{x}$ and the second order constraint qualification, since if $\partial f(\hat{x})(\tilde{y})/\partial x \neq 0$, then either $\partial f(\hat{x})(\tilde{y})/\partial x < 0$ or $\partial f(\hat{x})(-\tilde{y})/\partial x < 0$, and the second order qualification leads to a contradiction of optimality. Thus, it is apparent that the first order constraint qualification may be removed to obtain a slightly weaker theorem. In addition, one would expect to have a condition in terms of curves $x(\cdot)$ that are feasible for $\alpha \in [0, \bar{\alpha}]$, but do not necessarily satisfy the rather demanding condition, $g^i(x(\alpha)) = 0$ for $i \in I(\hat{x})$.

Our task in the next section will be to obtain optimality conditions without explicit assumptions relating $r(\cdot)$ and Ω, i.e., without constraint qualifications. This will be achieved by formulating the second order necessary conditions as fairly natural geometric relations between appropriate sets.

3. Second order necessary conditions. We have seen that Theorem 1, which gives first order necessary conditions of optimality, relies on local approximation to the set Ω. While this approximation can also be used for some second order conditions, it is convenient to introduce a new local approximation.

DEFINITION 5. A pair $\{C(\hat{x}, \tilde{y}, \Omega), y^*\}$ will be called a *$\tilde{y}$-directed conical approximation* to Ω at $\hat{x}$, if $C(\hat{x}, \tilde{y}, \Omega)$ is a convex cone and if for any collection $\{y_1, y_2, \cdots, y_k\}$ of vectors in $C(\hat{x}, \tilde{y}, \Omega)$, any $k - 1$ of which are linearly independent, there is an $\alpha_0 > 0$ and a continuous map $\zeta_{\tilde{y}}(\cdot, \cdot)$, (possibly depending on $\hat{x}, \tilde{y}, y^*, y_1, \cdots, y_k$) from $[0, \alpha_0] \times co\{y_1, \cdots, y_k\}$ into $\Omega - \hat{x}$, of the form

$$\zeta_{\tilde{y}}(\alpha, y) = \alpha\tilde{y} + \frac{\alpha^2}{2}(y^* + y) + o(\alpha^2, y), \tag{6}$$

where $\|o(\alpha^2, y)\|/\alpha^2 \to 0$ as $\alpha \to 0$, for all y.

We shall refer to $\{C(\hat{x}, \tilde{y}, \Omega), y^*\}$ simply as a directed conical approximation when $\tilde{y}$ is clear from the context. The special cases which arise when $C(\hat{x}, \tilde{y}, \Omega) = \{0\}$, or $y^* = 0$, or even $\tilde{y} = 0$ (or any combination of these) are not excluded from consideration.

There may, of course, be many directed conical approximations for a single $\tilde{y}$, as well as useful relations between the conical approximation defined in Definition 1 and the directed conical approximations defined in Definition 5. Thus, if $\{C(\hat{x}, \tilde{y}, \Omega), y^*\}$ is a directed conical approximation, the ray $\{y | y = \lambda \tilde{y}, \lambda \geqq 0\}$ may be regarded as a trivial conical approximation with map $\zeta(y) = \zeta(\lambda \tilde{y}) = \lambda \tilde{y} + \lambda^2(y^* + \bar{y})/2 + o(\lambda^2, \bar{y})$, where $\bar{y}$ is any vector in $C(\hat{x}, \tilde{y}, \Omega)$ and $o(\cdot, \cdot)$ is given by (5).

Conversely, we may often obtain directed conical approximations from conical approximations, the most important case being the following one.

LEMMA 1. *If $C(\hat{x}, \Omega)$ is a simple conical approximation to Ω at $\hat{x}$ and $\tilde{y}$ is any vector in $C(\hat{x}, \Omega)$, then $\{RC(\tilde{y}, C(\hat{x}, \Omega)), 0\}$ is a $\tilde{y}$-directed conical approximation to Ω at $\hat{x}$ (where for any set S and $x \in S$, we define $RC(x, S) = \{y|$ there exists a $\bar{\lambda} > 0$ such that $x + \lambda y \in S$ for $0 \leqq \lambda \leqq \bar{\lambda}\}$).*

Note that if $\tilde{y} \in C(\hat{x}, \Omega)$, then $C(\hat{x}, \Omega) \subset RC(\tilde{y}, C(\hat{x}, \Omega))$, with strict inclusion whenever $-\tilde{y} \notin C(\hat{x}, \Omega)$.

We digress to indicate several important cases for which simple conical approximations may be constructed. (For $\Omega = \{x | g^i(x) \leqq 0,\ i = 1, \cdots, k\}$ and $x \in \Omega$, the index set $I(x) \triangleq \{i | i \in \{1, \cdots, k\}$ and $g^i(x) = 0\}$.)

LEMMA 2. *Suppose that $\Omega = \{x | g^i(x) \leqq 0$ for $i = 1, \cdots, k\}$ and that $\hat{x} \in \Omega$. If the internal cone to Ω at $\hat{x}$, $IC(\hat{x}, \Omega)$, defined by*

$$IC(\hat{x}, \Omega) = \left\{ y \,\middle|\, \frac{\partial g^i}{\partial x}(\hat{x})(y) < 0, i \in I(\hat{x}) \right\}, \tag{7}$$

is not empty, then it is a simple conical approximation to Ω at $\hat{x}$.

LEMMA 3. *If $\hat{x}$ is contained in Ω and Ω^* is any set containing $\hat{x}$ such that $\Omega \cap \Omega^*$ is convex, then $RC(\hat{x}, \Omega \cap \Omega^*)$ is a simple conical approximation to Ω at $\hat{x}$.*

LEMMA 4. *If $C(\hat{x}, \Omega)$ is a conical approximation with nonempty interior* int $(C(\hat{x}, \Omega))$, *then* int $(C(\hat{x}, \Omega))$ *is a simple conical approximation to Ω at $\hat{x}$.*

Whenever Ω has the description given in Lemma 2, i.e., we are dealing with the Nonlinear Programming Problem, it is consistent with our idea of a well-formulated problem that Ω will have an interior, and hence $IC(\hat{x}, \Omega)$ will be nonempty. This situation enables us to construct an important class of directed conical approximations.

LEMMA 5. *Consider the set $\Omega = \{x | g^i(x) \leqq 0$ for $i = 1, 2, \cdots, k\}$ and let $\hat{x} \in \Omega$. Suppose that $\tilde{y} \in \overline{IC(\hat{x}, \Omega)}$, and let the index set $I(\hat{x}, \tilde{y})$ be defined by*

$$I(\hat{x}, \tilde{y}) = \left\{ i \,\middle|\, i \in I(\hat{x}), \frac{\partial g^i}{\partial x}(\hat{x})(\tilde{y}) = 0 \right\}, \tag{8}$$

and let the $\tilde{y}$-directed internal cone to Ω at $\hat{x}$, $IC(\hat{x}, \tilde{y}, \Omega)$, be defined by

$$IC(\hat{x}, \tilde{y}, \Omega) = \left\{ y \,\middle|\, \frac{\partial g^i}{\partial x}(\hat{x})(y) < 0, i \in I(\hat{x}, \tilde{y}) \right\} \tag{9}$$

(*and* $IC(\hat{x}, \tilde{y}, \Omega) = E^n$ *if* $I(\hat{x}, \tilde{y}) = \emptyset$). *Then there is a vector* $y^* \in E^n$ *such that*

$$\text{(i)} \qquad \frac{\partial^2 g^i}{\partial x^2}(\hat{x})(\tilde{y}, \tilde{y}) + \frac{\partial g^i}{\partial x}(\hat{x})(y^*) \leqq 0 \quad \textit{for} \quad i \in I(\hat{x}, \tilde{y})$$

and (ii) *the pair* $\{IC(\hat{x}, \tilde{y}, \Omega), y^*\}$ *is a* $\tilde{y}$*-directed conical approximation to* Ω *at* $\hat{x}$. *Moreover,* $IC(\hat{x}, \Omega) \subset IC(\hat{x}, \tilde{y}, \Omega)$, *with strict inclusion if* $I(\hat{x}, \tilde{y}) \neq I(\hat{x})$.

To illustrate the usefulness of Lemma 5 and to see that there are situations when y^* must be nonzero if one wishes to obtain a directed conical approximation, let $x = (x^1, x^2), \Omega = \{x | g(x) \triangleq ((x^1 - 1)^2 + (x^2)^2 - 1)/2 \leqq 0\}$ and $\hat{x} = (0, 0)$. With $\tilde{y} = (0, 1)$ it is easy to verify that there is no cone C such that $\{C, (0, 0)\}$ is a $\tilde{y}$-directed conical approximation, however $\{IC(\hat{x}, \Omega), (1, 0)\}$ is a $(0, 1)$-directed conical approximation.

We now isolate those vectors $\tilde{y}$ for which, in the context of the Basic Problem, a $\tilde{y}$-directed conical approximation will lead to a very general second order necessary condition of optimality.

DEFINITION 6. A vector $\tilde{y}$ is said to be a *critical direction* (at $\hat{x}$) for the Basic Problem if $\partial f(\hat{x})(\tilde{y})/\partial x \leqq 0$ and $\partial r^i(\hat{x})(\tilde{y})/\partial x = 0$ for $i = 1, \cdots, m$.

THEOREM 6. *Suppose that* $\hat{x}$ *is an optimal solution to the Basic Problem and that* $\tilde{y}$ *is a critical direction at* $\hat{x}$. *If* $\{C(\hat{x}, \tilde{y}, \Omega), y^*\}$ *is a* $\tilde{y}$*-directed conical approximation to* Ω *at* $\hat{x}$, *then there exists a nonzero vector* $\psi = (\psi^0, \psi^1, \cdots, \psi^m)$ *in* E^{m+1}, *satisfying* $\psi^0 \leqq 0$ *and* $\psi^0 = 0$ *if* $\partial f(\hat{x})(\tilde{y})/\partial x < 0$, *such that*

$$\text{(i)} \ \psi^0 \frac{\partial f}{\partial x}(\hat{x})(y) + \sum_{i=1}^{m} \psi^i \frac{\partial r^i}{\partial x}(\hat{x})(y) \leqq 0 \ \textit{for all} \ y \in \overline{C(\hat{x}, \tilde{y}, \Omega)},$$

$$\text{(ii)} \ \psi^0 \left(\frac{\partial^2 f}{\partial x^2}(\hat{x})(\tilde{y}, \tilde{y}) + \frac{\partial f}{\partial x}(\hat{x})(y^*) \right) + \sum_{i=1}^{m} \psi^i \left(\frac{\partial^2 r^i}{\partial x^2}(\hat{x})(\tilde{y}, \tilde{y}) + \frac{\partial r^i}{\partial x}(\hat{x})(y^*) \right) \leqq 0.$$

Remark 2. Note that inequality (2) of Theorem 1 may also be obtained from Theorem 4 by Dubovickii and Milyutin [8], [9] and Theorem 6 of this section. then, setting $\tilde{y} = 0$, $C(\hat{x}, 0, \Omega) = C(\hat{x}, \Omega)$, $y^* = 0$, and $\zeta_0(\alpha, y) = \alpha^2 y/2 + o(\alpha^2 y)$, we find that part (i) of Theorem 6 yields the same result as Theorem 1, but part (ii) carries no information. However, the inequality (i) of Theorem 6 will often hold for cones $C(\hat{x}, \tilde{y}, \Omega)$ which are much larger than any conical approximation to Ω at $\hat{x}$.

It is appropriate at this stage to comment on the crucial differences between Theorem 4 by Dubovickii and Milyutin [8], [9] and Theorem 6 of this section. Note that Theorem 4 is essentially a disjointness condition in the domain space of the map $F(\cdot) = (f(\cdot), r^1(\cdot), \cdots, r^m(\cdot))$ (i.e., in E^n), while Theorem 6 represents separation conditions in the range space of $F(\cdot)$ (i.e., in E^{m+1}), which requires simpler assumptions. Thus, to obtain from Theorem 4 inequalities of the form (i) and (ii) of Theorem 6 it is necessary to make fairly strong assumptions on each of the sets $C_0(\tilde{y})$, $C_1(\tilde{y})$ and $C_2(\tilde{y})$ (see (3), (4) and (5)). On the other hand, any time $C_1(\tilde{y})$ is of the form $C_1(\tilde{y}) = y^* + C(\tilde{y})$, where $C(\tilde{y})$ is a convex cone, we find that $\{C(\tilde{y}), y^*\}$ is a $\tilde{y}$-directed conical approximation to Ω at x^*, and we obtain (i) and (ii) of Theorem 6 immediately.

Before further discussion of Theorem 6, we consider the zero-gradient case. It is assumed that at most one gradient corresponding to an equality constraint is zero. We define the ray P in E^{m+1} by

$$P = \{w | w^0 < 0 \text{ and } w^i = 0 \text{ for } i = 1, \cdots, m\}.$$

THEOREM 7. *Suppose that $\hat{x}$ is an optimal solution to the Basic Problem and that $C(\hat{x}, \Omega)$ is a conical approximation to Ω at $\hat{x}$ with nonempty interior* $\operatorname{int}(C(\hat{x}, \Omega))$.

(i) *If $\partial f(\hat{x})/\partial x = 0$, then the ray P has no points in the interior of the set*

$$(10)\quad L_0 = \left\{ w \,\middle|\, w^0 = \frac{\partial^2 f}{\partial x^2}(\hat{x})(y, y), w^i = \frac{\partial r^i}{\partial x}(\hat{x})(y), i = 1, 2, \cdots, m, y \in \operatorname{int}(C(\hat{x}, \Omega)) \right\}.$$

(ii) *If $\partial r^1(\hat{x})/\partial x = 0$, then the ray P has no points in the interior of the set*

$$(11)\quad \begin{aligned} L_1 = \Big\{ w \,\Big|\, & w^0 = \frac{\partial f}{\partial x}(\hat{x})(y), w^1 = \frac{\partial^2 r^1}{\partial x^2}(\hat{x})(y, y), \\ & w^i = \frac{\partial r^i}{\partial x}(\hat{x})(y), i = 2, 3, \cdots, m, y \in \operatorname{int}(C(\hat{x}, \Omega)) \Big\}. \end{aligned}$$

(iii) *If $\partial f(\hat{x})/\partial x = \partial r^1(\hat{x})/\partial x = 0$, then the ray P has no points in the interior of the set*

$$(12)\quad \begin{aligned} L_2 = \Big\{ w \,\Big|\, & w^0 = \frac{\partial^2 f}{\partial x^2}(\hat{x})(y, y), w^1 = \frac{\partial^2 r^1}{\partial x^2}(\hat{x})(y, y), \\ & w^i = \frac{\partial r^i}{\partial x}(\hat{x})(y), i = 2, 3, \cdots, m, y \in \operatorname{int}(C(\hat{x}, \Omega)) \Big\}. \end{aligned}$$

Remark 3. Theorem 7 remains true even when $\operatorname{int}(C(\hat{x}, \Omega))$ is replaced by the relative interior of $C(\hat{x}, \Omega)$. Also, if only the case $\partial f(\hat{x})/\partial x = 0$ is considered, it can be shown that the following is true.

THEOREM 8 ([12]). *Suppose that $\hat{x}$ is an optimal solution to the Basic Problem and that $C(\hat{x}, \Omega)$ is a conical approximation to Ω at $\hat{x}$. If $\partial f(\hat{x})/\partial x = 0$, then the ray P has no points in the interior of the set*

$$(13)\quad L_0' = \left\{ w \,\middle|\, w^0 = \frac{\partial^2 f}{\partial x^2}(\hat{x})(y, y), w^i = \frac{\partial r^i}{\partial x}(\hat{x})(y), \quad i = 1, 2, \cdots, m, y \in C(\hat{x}, \Omega) \right\}.$$

It can be shown that theorems similar to Theorem 8 cannot be obtained for all situations covered by Theorem 7, i.e., $\operatorname{int}(C(\hat{x}, \Omega))$ *cannot* be replaced by $C(\hat{x}, \Omega)$. In fact, consideration of Example 1, with $\Omega = \{(x^1, x^2) | (x^1 - 1)^2 + (x^2 - 1)^2 - 2 \leqq 0\}$, $\hat{x} = (0, 0)$ and $C(\hat{x}, \Omega) = \{(x^1, x^2) | x^1 + x^2 \geqq 0\}$, will confirm this.

Theorems 6 and 7 represent two different approaches to second order conditions. Theorem 6 is well structured and neatly supplements the first order conditions in Theorem 1. Theorem 7, on the other hand, is in rather awkward form, since the sets L_0, L_1 and L_2 are in general neither convex nor even conical. However, in spite of this, Theorem 7 answers some questions which Theorem 6 does not, and in some cases leads to alternate expressions. We now demonstrate this.

Examination of Theorem 6 indicates that the multiplier vector ψ depends on the critical direction $\tilde{y}$. However, it is clear from Lemma 1 that we may be able to find a pair $\{C, y^*\}$ which is a directed conical approximation for more than one critical direction $\tilde{y}$. The natural question to ask, then, is whether there is a multiplier vector ψ which will satisfy the conditions of Theorem 6 for all these critical vectors $\tilde{y}$ and the given pair $\{C, y^*\}$. Unfortunately, as will be seen from the following example, this is not always true.

Example 1. Let $x = (x^1, x^2)$ and consider the Basic Problem with $f(x) = -(x^1 - x^2)^2$, $r(x) = (x^1)^2 - (x^2)^2$ and $\Omega = \{x | x^1 \geqq 0, x^2 \geqq 0\}$. Clearly, the point $\hat{x} = (0, 0)$ is an optimal solution since the only feasible points are defined by the intersection of the line $x^1 - x^2 = 0$ and the positive quadrant. Since $\partial f(\hat{x})/\partial x = \partial r(\hat{x})/\partial x = 0$, each $\tilde{y}$ in Ω is a critical direction and we may take $C(\hat{x}, \tilde{y}, \Omega) = \Omega$, $y^* = (0, 0)$. Now, since each gradient is zero, if there is a single multiplier vector ψ which satisfies Theorem 6 for all $\tilde{y}$ in Ω, it must satisfy

$$\psi^0 \frac{\partial^2 f}{\partial x^2}(\hat{x})(\tilde{y}, \tilde{y}) + \psi^1 \frac{\partial^2 r}{\partial x^2}(\hat{x})(\tilde{y}, \tilde{y}) \leqq 0$$

for all $\tilde{y} \in \Omega$, with $\psi^0 \leqq 0$. This is equivalent to requiring that the cone

$$V = \left\{ v = (v^0, v^1) \middle| v^0 = \frac{\partial^2 f}{\partial x^2}(\hat{x})(y, y), v^1 = \frac{\partial^2 r}{\partial x^2}(\hat{x})(y, y), y \in \Omega \right\}$$

be separated from the ray P. It is trivial to verify that V is the set $\{(0, 0)\} \cup \{v | v^0 < 0, v^0 + v^1 \geqq 0\} \cup \{v | v^0 < 0, v^0 - v^1 \geqq 0\}$, which cannot be separated from the ray P.

We see that in the preceding example the set Ω also serves as a simple conical approximation to Ω at $(0, 0)$. Since Ω has an interior and both $\partial f(\hat{x})/\partial x$ and $\partial r(\hat{x})/\partial x = 0$, Theorem 7 can be applied to answer the question as to when there is a single multiplier vector satisfying

$$\psi^0 \frac{\partial^2 f}{\partial x^2}(\hat{x})(y, y) + \psi^1 \frac{\partial^2 r}{\partial x^2}(\hat{x})(y, y) \leqq 0$$

for all $y \in \Omega$. In particular, it yields the following modification of Theorem 6.

THEOREM 9. *Suppose that the Basic Problem has only one equality constraint, i.e., $r: E^n \to E^1$. If $\hat{x}$ is an optimal solution to the Basic Problem, with $\partial f(\hat{x})/\partial x = \partial r(\hat{x})/\partial x = 0$, and if $C(\hat{x}, \Omega)$ is a conical approximation to Ω at $\hat{x}$ such that the set L_2 (equation (12)) is convex, then there exist scalars ψ^0 and ψ^1 not both zero with $\psi^0 \leqq 0$ such that*

$$\psi^0 \frac{\partial^2 f}{\partial x^2}(\hat{x})(y, y) + \psi^1 \frac{\partial^2 r}{\partial x^2}(\hat{x})(y, y) \leqq 0 \quad \textit{for all } y \textit{ in } \overline{C(\hat{x}, \Omega)}.$$

Proof. Since there is only a single equality constraint and both $\partial f(\hat{x})/\partial x$ and $\partial r(\hat{x})/\partial x = 0$, the set L_2 is conical and by assumption also convex. Suppose

L_2 is not separated from P. It follows that P has points in the interior of the set L_2, which contradicts Theorem 7. Thus, P and L_2 must be separated, which proves the theorem.

When $r(\cdot) \equiv 0$, we have a somewhat simpler situation and Theorem 8 leads to the following result.

THEOREM 10. *Suppose that in the Basic Problem,* $r(\cdot) \equiv 0$ *and* $\hat{x}$ *is an optimal solution with* $\partial f(\hat{x})/\partial x = 0$. *If* $\{C_\alpha(\hat{x}, \Omega) | \alpha \in A\}$, *where* A *is an index set, is any collection of conical approximations to* Ω *at* $\hat{x}$, *then*

$$\frac{\partial^2 f}{\partial x^2}(\hat{x})(y, y) \geqq 0 \quad \textit{for all } y \textit{ in } \quad \overline{\bigcup_{\alpha \in A} C_\alpha(\hat{x}, \Omega)}.$$

Proof. Let α be arbitrary in A, and let $C_\alpha(\hat{x}, \Omega)$ be the corresponding conical approximation. Applying Theorem 8, we see that the set L_0' is one-dimensional and hence the statement of the theorem is that

$$\frac{\partial^2 f}{\partial x^2}(\hat{x})(y, y) \geqq 0 \quad \text{for all } y \text{ in } C_\alpha(\hat{x}, \Omega).$$

Thus,

$$\frac{\partial^2 f}{\partial x^2}(\hat{x})(y, y) \geqq 0 \quad \textit{for all } y \textit{ in } \quad \bigcup_{\alpha \in A} C_\alpha(\hat{x}, \Omega)$$

and by continuity this is also true for the closure, which completes the proof.

Theorem 2 of § 2 may be obtained from Theorem 10. This follows by identifying an index set A by $A = TC(\hat{x}, \Omega')$, where $\Omega' = \{x | x \in \Omega,\ r(x) = 0\}$, and for each $y \in A$, i.e., $y \in TC(\hat{x}, \Omega')$, identifying a corresponding conical approximation by $C_y(\hat{x}, \Omega') = \{y' | y' = \lambda y, \lambda > 0\}$—which is a valid conical approximation since $y \in TC(\hat{x}, \Omega')$. Then $\bigcup_{y \in A} C_y(\hat{x}, \Omega') = TC(\hat{x}, \Omega')$ and if $\partial f(\hat{x})/\partial x = 0$, Theorem 2 follows from the statement of Theorem 10.

When $r(\cdot) \equiv 0$, we can also obtain a corollary of Theorem 6, similar to Theorem 10.

THEOREM 11. *Suppose that in the Basic Problem,* $r(\cdot) \equiv 0$. *If* $\hat{x}$ *is optimal and* A *is a set such that for each* $\alpha \in A$ *there is a critical direction* y_α *and corresponding directed conical approximation* $\{C(\hat{x}, \tilde{y}_\alpha, \Omega), y_\alpha^*\}$, *then*

$$\text{(i)} \quad \frac{\partial f}{\partial x}(\hat{x})(y) \geqq 0 \quad \textit{for all } y \textit{ in } \quad \overline{\bigcup_{\alpha \in A} C(\hat{x}, \tilde{y}_\alpha, \Omega)}$$

and

$$\text{(ii)} \quad \frac{\partial^2 f}{\partial x^2}(\hat{x})(\tilde{y}_\alpha, \tilde{y}_\alpha) + \frac{\partial f}{\partial x}(\hat{x})(y_\alpha^*) \geqq \quad \textit{for all } \alpha \in A.$$

Proof. Let $\alpha \in A$ be arbitrary and let $\{C(\hat{x}, \tilde{y}_\alpha, \Omega), y_\alpha^*\}$ be the corresponding $\tilde{y}_\alpha$-directed conical approximation. From Theorem 6, since $\psi = (\psi^0)$ is nonzero and satisfies $\psi^0 \leqq 0$, we may take $\psi^0 = -1$. Thus, $\partial f(\hat{x})(y)/\partial x \geqq 0$ for all $y \in C(\hat{x}, \tilde{y}_\alpha, \Omega)$ and $\partial^2 f(\hat{x})(\tilde{y}_\alpha, \tilde{y}_\alpha)/\partial x^2 + \partial f(\hat{x})(y_\alpha^*)/\partial x \geqq 0$. Since α was arbitrary, the theorem is true.

Theorem 3 of § 2 is obtained as a special case of Theorem 11. Thus, suppose $x:E^1 \to E^n$ is a twice differentiable function such that $x(\alpha)$ is contained in Ω' for α in $[0,\bar{\alpha}]$, $\bar{\alpha} > 0$, that $\hat{x} = x(0)$ is optimal and that $df(x(0))/d\alpha = 0$. Taking the critical direction to be $\bar{y} = dx(0)/d\alpha$ and the corresponding directed conical approximation to be $\{C(\hat{x}, \bar{y}, \Omega'), y^*\} = \{\{0\}, d^2x(0)/d\alpha^2\}$, we obtain from the second part of Theorem 11 that

$$\frac{\partial^2 f}{\partial x^2}(\hat{x})\left(\frac{dx(0)}{d\alpha}, \frac{dx(0)}{d\alpha}\right) + \frac{\partial f}{\partial x}(\hat{x})\left(\frac{d^2x(0)}{d\alpha^2}\right) \geqq 0,$$

which corresponds to the condition $d^2f(x(0))/d\alpha^2 \geqq 0$ of Theorem 3.

Remark 4. In view of Theorems 6 and 7, one may be inclined to think that more information about a candidate optimal solution $\hat{x}$ could be obtained and verification of the necessary conditions simplified, by transcribing the original problem into an equivalent form with simple structure or many critical directions. Thus, for any problem of the form of the Basic Problem, an equivalent problem with a single equality constraint, $\tilde{r}:E^n \to E^1$, can always be defined by letting $\tilde{r}(x) = \sum_{i=1}^m (r^i(x))^2$. Since $\partial\tilde{r}(\hat{x})/\partial x = 2\sum_{i=1}^m r^i(\hat{x})\partial r^i(\hat{x})/\partial x = 0$, we can now always apply either Theorem 7 or Theorem 6 with the set of critical directions being $\{y|\partial f(\hat{x})(y)/\partial x \leqq 0\}$. Unfortunately, Theorems 6 and 7 can be satisfied trivially for this new problem and so it is seen that the transcription does not increase the amount of information available about the optimal solutions of the original problem. (Since $\partial^2\tilde{r}(\hat{x})(y, y)/\partial x^2 = 2\sum_{i=1}^m (\partial r^i(\hat{x})(y)/\partial x)^2 \geqq 0$ for all y, in Theorem 6 we may take $\psi^0 = 0$ and $\psi^1 = -1$, while in Theorem 7, again by the preceding inequality, the ray P will have no points in the interior of L_1 or L_2.)

3.1. Applications to nonlinear programming. Theorem 6 will now be applied to the Nonlinear Programming Problem to obtain a generalization of Theorem 5. We note that as long as the total number of equality and inequality constraints is less than n (where $x \in E^n$) and Ω has an interior, critical directions with nontrivial directed conical approximations will exist. The following key theorem will therefore normally be applicable.

THEOREM 12. *If $\hat{x}$ is an optimal solution to the Nonlinear Programming Problem and if $IC(\hat{x}, \Omega)$ is not empty, then for each critical direction $\bar{y} \in \overline{IC(\hat{x}, \Omega)}$, there exists a vector y^* in E^n, multipliers $\psi^0, \psi^1, \cdots, \psi^m$ not all zero, satisfying $\psi^0 \leqq 0$ and $\psi^0 = 0$ if $\partial f(\hat{x})(\bar{y})/\partial x < 0$, and multipliers $\mu^i \leqq 0$, satisfying $\mu^i = 0$ if $i \notin I(\hat{x}, \bar{y})$, such that*

$$\text{(i)}\quad \frac{\partial^2 g^i}{\partial x^2}(\hat{x})(\bar{y}, \bar{y}) + \frac{\partial g^i}{\partial x}(\hat{x})(y^*) \leqq 0 \quad \textit{for} \quad i \in I(\hat{x}, \bar{y}),$$

$$\text{(ii)}\quad \psi^0\frac{\partial f}{\partial x}(\hat{x}) + \sum_{i=1}^m \psi^i\frac{\partial r^i}{\partial x}(\hat{x}) + \sum_{i=1}^k \mu^i\frac{\partial g^i}{\partial x}(\hat{x}) = 0,$$

$$\text{(iii)}\quad \psi^0\left(\frac{\partial^2 f}{\partial x^2}(\hat{x})(\bar{y}, \bar{y}) + \frac{\partial f}{\partial x}(\hat{x})(y^*)\right) + \sum_{i=1}^m \psi^i\left(\frac{\partial^2 r^i}{\partial x^2}(\hat{x})(\bar{y}, \bar{y}) + \frac{\partial r^i}{\partial x}(\hat{x})(y^*)\right) \leqq 0.$$

Proof. Let $\tilde{y} \in \overline{IC(\hat{x}, \Omega)}$. By Lemma 5, there is a y^* which satisfies (i), and moreover the pair $\{IC(\hat{x}, \tilde{y}, \Omega), y^*\}$ (see (9)) is a $\tilde{y}$-directed conical approximation. Thus, by Theorem 6, there is a nonzero multiplier ψ, satisfying $\psi^0 \leqq 0$ and $\psi^0 = 0$ if $\partial f(\hat{x})(\tilde{y})/\partial x < 0$, such that

$$\psi^0 \frac{\partial f}{\partial x}(\hat{x})(y) + \sum_{i=1}^{m} \psi^i \frac{\partial r^i}{\partial x}(\hat{x})(y) \leqq 0 \quad \text{for all } y \in \overline{IC(\hat{x}, \tilde{y}, \Omega)},$$

and, in addition, this multiplier satisfies condition (iii). Applying Farka's lemma [13] to the preceding inequality, there are scalars $-\mu^i \geqq 0$ for $i \in I(\hat{x}, \tilde{y})$ such that

$$\psi^0 \frac{\partial f}{\partial x}(\hat{x}) + \sum_{i=1}^{m} \psi^i \frac{\partial r^i}{\partial x}(\hat{x}) = \sum_{i \in I(\hat{x}, \tilde{y})} -\mu^i \frac{\partial g^i}{\partial x}(\hat{x}).$$

Defining $\mu^i = 0$ for $i \notin I(\hat{x}, \tilde{y})$ completes the proof.

COROLLARY 1. *If in the statement of Theorem* 12, y^* *satisfies*

$$\frac{\partial^2 g^i}{\partial x^2}(\hat{x})(\tilde{y}, \tilde{y}) + \frac{\partial g^i}{\partial x}(\hat{x})(y^*) = 0 \quad \textit{for } i \in I(\hat{x}, \tilde{y}),$$

then condition (iii) *may be replaced by*

$$\text{(iii')} \qquad \psi^0 \frac{\partial^2 f}{\partial x^2}(\hat{x})(\tilde{y}, \tilde{y}) + \sum_{i=1}^{m} \psi^i \frac{\partial^2 r^i}{\partial x^2}(x)(\tilde{y}, \tilde{y}) + \sum_{i=1}^{k} \mu^i \frac{\partial^2 g^i}{\partial x^2}(\hat{x})(\tilde{y}, \tilde{y}) \leqq 0.$$

Proof. Let μ^i be the scalars given in the statement of the theorem. Then

$$\sum_{i=1}^{k} \mu^i \left(\frac{\partial^2 g^i}{\partial x^2}(\hat{x})(\tilde{y}, \tilde{y}) + \frac{\partial g^i}{\partial x}(\hat{x})(y^*) \right) = 0,$$

and therefore this term may be added to (iii) without changing the sign of the inequality. However, from condition (ii) we have

$$\psi^0 \frac{\partial f}{\partial x}(\hat{x})(y^*) + \sum_{i=1}^{m} \psi^i \frac{\partial r^i}{\partial x}(\hat{x})(y^*) + \sum_{i=1}^{k} \mu^i \frac{\partial g^i}{\partial x}(\hat{x})(y^*) = 0,$$

which gives condition (iii'). This completes the proof.

A sufficient condition which ensures that $IC(\hat{x}, \Omega)$ is not empty and that a y^* satisfying the hypothesis of Corollary 1 exists is that the vectors $\partial g^i(\hat{x})/\partial x$, $i \in I(\hat{x})$, are linearly independent. However, while this assumption simplifies Theorem 12, we are again faced with the question of determining when the multiplier vector ψ does not depend on the critical direction. Necessary conditions holding for a class of critical directions, but involving only a single multiplier vector, may be obtained with the following assumption.

ASSUMPTION 1.[4] *Consider the Nonlinear Programming Problem. Let* $\bar{y}$ *be a critical direction at* $\hat{x}$, *satisfying* $\partial g^i(\hat{x})(\bar{y})/\partial x \leqq 0$ for $i \in I(\hat{x})$, *and define the set*

$$(14) \quad Y(\bar{y}) = \left\{ \tilde{y} | \tilde{y} \textit{ is a critical direction at } \hat{x}, \frac{\partial g^i}{\partial x}(\hat{x})(\tilde{y}) \leqq 0 \textit{ for } i \in I(\hat{x}), \right.$$
$$\left. \textit{and } I(\hat{x}, \tilde{y}) = I(\hat{x}, \bar{y}) \right\}.$$

[4] Note that Assumption 1 does not require that $IC(\hat{x}, \Omega) \neq$

It is assumed that for every $\tilde{y} \in Y(\bar{y})$, *there is a twice differentiable function* $x : E^1 \to E^n$ *and an* $\alpha_0 > 0$, *such that* (i) $x(0) = \hat{x}$, $dx(0)/d\alpha = \tilde{y}$ *and* (ii) $x(\alpha)$ *is feasible for* $\alpha \in [0, \alpha_0]$, *with* $g^i(x(\alpha)) = 0$ *for* $i \in I(\hat{x}, \tilde{y})$.

THEOREM 13. *Suppose that* $\hat{x}$ *is an optimal solution to the Nonlinear Programming Problem and that* $\bar{y}$ *is a critical direction at* $\hat{x}$ *satisfying* $\partial g^i(\hat{x})(\bar{y})/\partial x \leqq 0$ *for* $i \in I(\hat{x})$. *If Assumption* 1 *holds for the set* $Y(\bar{y})$ (*equation* (14)), *then there exist multipliers* $\psi^0, \psi^1, \cdots, \psi^m$ *satisfying* $\psi^0 \leqq 0$, *and multipliers* $\mu^1, \mu^2, \cdots, \mu^k$ *satisfying* $\mu^i \leqq 0$ *and* $\mu^i = 0$ *if* $i \notin I(\hat{x}, \bar{y})$, *such that*

$$\psi^0 \frac{\partial f}{\partial x}(\hat{x}) + \sum_{i=1}^{m} \psi^i \frac{\partial r^i}{\partial x}(\hat{x}) + \sum_{i=1}^{k} \mu^i \frac{\partial g^i}{\partial x}(\hat{x}) = 0, \tag{15}$$

$$\psi^0 \frac{\partial^2 f}{\partial x^2}(\hat{x})(\tilde{y}, \tilde{y}) + \sum_{i=1}^{m} \psi^i \frac{\partial^2 r^i}{\partial x^2}(\hat{x})(\tilde{y}, \tilde{y}) + \sum_{i=1}^{k} \mu^i \frac{\partial^2 g^i}{\partial x^2}(\hat{x})(\tilde{y}, \tilde{y}) \leqq 0 \quad \textit{for all } \tilde{y} \in \overline{Y(\bar{y})}. \tag{16}$$

Furthermore, if $IC(\hat{x}, \Omega) \neq \varnothing$, *then* $\psi \neq 0$, *and if* $IC(\hat{x}, \Omega) = \varnothing$, *then* $\mu = 0$ *and equality holds for* (16).

Remark 5. The set $\overline{Y(\bar{y})}$ consists of those critical directions $\tilde{y}$ satisfying $\partial g^i(\hat{x})(\tilde{y})/\partial x \leqq 0$ for $i \in I(\hat{x})$ and $I(\hat{x}, \tilde{y}) \supset I(\hat{x}, \bar{y})$. Note also that the case $\partial f(\hat{x})(\bar{y})/\partial x < 0$ is impossible if $\hat{x}$ is optimal and Assumption 1 holds.

Proof. For any $\tilde{y} \in Y(\bar{y})$, let $x(\cdot)$ be the twice differentiable function guaranteed by Assumption 1, and let $y^* = dx(0)/d\alpha^2$. The following conditions are readily established from Assumption 1, the optimality of $\hat{x}$ and the fact that $\tilde{y}$ is a critical direction:

$$\begin{aligned} &\frac{\partial^2 f}{\partial x^2}(\hat{x})(\tilde{y}, \tilde{y}) + \frac{\partial f}{\partial x}(\hat{x})(y^*) \geqq 0, \\ &\frac{\partial^2 r^i}{\partial x^2}(\hat{x})(\tilde{y}, \tilde{y}) + \frac{\partial r^i}{\partial x}(\hat{x})(y^*) = 0, \qquad i = 1, 2, \cdots, m, \\ &\frac{\partial^2 g^i}{\partial x^2}(\hat{x})(\tilde{y}, \tilde{y}) + \frac{\partial g^i}{\partial x}(\hat{x})(y^*) = 0, \qquad i \in I(\hat{x}, \bar{y}). \end{aligned} \tag{17}$$

Considering first the case $IC(\hat{x}, \Omega) \neq \varnothing$ and applying Theorem 12, with the critical direction given by $\bar{y}$, we obtain a nonzero vector ψ, satisfying $\psi^0 \leqq 0$, and a corresponding vector μ, satisfying $\mu^i \leqq 0$ and $\mu^i = 0$ if $i \notin I(\hat{x}, \bar{y})$, which satisfy (15). But now for any $\tilde{y} \in Y(\bar{y})$, conditions (17) imply that

$$\psi^0\left(\frac{\partial^2 f}{\partial x^2}(\hat{x})(\tilde{y}, \tilde{y}) + \frac{\partial f}{\partial x}(\hat{x})(y^*)\right) + \sum_{i=1}^{m} \psi^i\left(\frac{\partial^2 r^i}{\partial x^2}(\hat{x})(\tilde{y}, \tilde{y}) + \frac{\partial r^i}{\partial x}(\hat{x})(y^*)\right)$$

$$+ \sum_{i=1}^{k} \mu^i\left(\frac{\partial^2 g^i}{\partial x^2}(\hat{x})(\tilde{y}, \tilde{y}) + \frac{\partial g^i}{\partial x}(\hat{x})(y^*)\right) \leqq 0.$$

By (15), the terms involving y^* sum to zero, and hence the inequality (16) is obtained.

Now consider the case $IC(\hat{x}, \Omega) = \varnothing$. It can be shown [1, Lemma 2] that there exists a nonzero vector $\mu = (\mu^1, \mu^2, \cdots, \mu^k)$, with $\mu^i \leqq 0$ and $\mu^i = 0$ if $i \notin I(\hat{x})$, $i = 1, 2, \cdots, k$, such that

$$\sum_{i=1}^{k} \mu^i \frac{\partial g^i}{\partial x}(\hat{x}) = 0. \tag{18}$$

Moreover, by taking the scalar product with $\bar{y}$, it is clear that $\mu^i = 0$ if $i \notin I(\hat{x}, \bar{y})$. Therefore, for $\tilde{y} \in Y(\bar{y})$, (17) implies that

$$\sum_{i=1}^{k} \mu^i \left(\frac{\partial^2 g^i}{\partial x^2}(\hat{x})(\tilde{y}, \bar{y}) + \frac{\partial g^i}{\partial x}(\hat{x})(y^*) \right) = 0,$$

and using (18), we obtain

$$\sum_{i=1}^{k} \mu^i \frac{\partial^2 g^i}{\partial x^2}(\hat{x})(\tilde{y}, \bar{y}) = 0. \tag{19}$$

Defining $\psi = (\psi^0, \psi^1, \cdots, \psi^m) = 0$ for this case completes the proof, since (18) and (19) are the desired conditions.

A sufficient condition for the cone $IC(\hat{x}, \Omega)$ to be nonempty and for Assumption 1 to hold (for a particular $\bar{y}$) is that the vectors $\partial r^i(\hat{x})/\partial x$ for $i = 1, 2, \cdots, m$ and $\partial g^i(\hat{x})/\partial x$ for $i \in I(\hat{x}, \bar{y})$ are linearly independent. From (14) and the fact that ψ is nonzero for this case, this is also sufficient to guarantee that ψ^0 in (15) must be nonzero.

Theorem 12 is a second order necessary condition of optimality which does not depend on any explicit assumptions (i.e., constraint qualifications) relating the functions $r(\cdot)$ and $g(\cdot)$, and it can often be applied in situations for which Theorem 5 previously considered does not apply. The price paid for the freedom from explicit assumptions is the dependence of the multipliers ψ, μ on the particular critical direction under consideration, a fault which Theorem 5 does not suffer from. However, Theorem 13 also permits the use of the *same* multipliers (ψ, μ) for a class of critical directions and is in fact a generalization of Theorem 5. It is clear that although Assumption 1 is similar to the second order constraint qualification (Definition 4), the first order constraint qualification is not required. Theorem 13 also applies to a larger class of critical directions than does Theorem 5, as we shall now show.

Example 2. Consider a particular problem of the form of the Nonlinear Programming Problem, with $x = (x^1, x^2)$, $f(x) = -(x^1 - 1)^2 - (x^2)^2$, $r(x) \equiv 0$, $g^1(x) = (x^1 - 1)^2 + (x^2)^2 - 5$ and $g^2(x) = (x^1 + 1)^2 + (x^2)^2 - 5$. Since $g^1(\hat{x}) \leqq 0$, the identity $f(x) = -g^1(x) - 5$ implies $f(\hat{x}) \geqq -5$, and hence the point $\hat{x} = (0, 2)$, for which $g^1(\hat{x}) = g^2(\hat{x}) = 0$, is an optimal solution, with $I(\hat{x}) = \{1, 2\}$. Now, since $\partial g^1(\hat{x})/\partial x$ and $\partial g^2(\hat{x})/\partial x$ are linearly independent, Theorem 5 applies only to the trivial case $\bar{y} = 0$. However, we find that $\bar{y} = (-2, -1)$ is a critical direction contained in $\overline{IC(\hat{x}, \Omega)}$, with $I(\hat{x}, \bar{y}) = \{1\}$. Thus, Theorem 12 can be applied (without further assumptions) and, in addition, by trivially verifying that Assumption 1 is satisfied, Theorem 13 can be applied (where $\overline{Y(\bar{y})} = \{y | y = \lambda\bar{y}, \lambda \geqq 0\}$). In either case we obtain an inequality for $\bar{y}$ which is not obtained from Theorem 5.

To illustrate how Theorem 12 augments the first order theory for the Nonlinear Programming Problem, suppose that $\bar{x}$ is a candidate optimal solution, $\bar{y}$ is a critical direction at $\bar{x}$ and either $I(\bar{x}) = \emptyset$ or $\bar{y} \in IC(\bar{x}, \Omega)$. Then, since $I(\bar{x}, \bar{y}) = \emptyset$, (ii) of Theorem 12 requires that $\psi^0 \partial f(\bar{x})/\partial x + \sum_{i=1}^m \psi^i \partial r^i(\bar{x})/\partial x = 0$. If $\partial f(\bar{x})/\partial x, \partial r^1(\bar{x})/\partial x, \cdots, \partial r^m(\bar{x})/\partial x$ are linearly independent, we conclude, since $\psi \neq 0$, that $\bar{x}$ cannot be optimal. Similarly, if $\partial f(\bar{x})(\bar{y})/\partial x < 0$, and $\partial r^1(\bar{x})/\partial x, \cdots, \partial r^m(\bar{x})/\partial x$ are linearly independent, we conclude that $\bar{x}$ cannot be optimal. However, if (ii) of Theorem 12 is satisfied, then (iii) of Theorem 12 would have to be examined; this is an easy task whenever the multipliers are unique.

As a more explicit illustration of Theorem 12, consider the quadratic programming problem, minimize $\frac{1}{2}\langle x, Qx\rangle + \langle d, x\rangle$ subject to $Ax = b$, where Q is an $n \times n$ symmetric matrix and A is an $m \times n$ matrix with $m < n$. Assume that $\hat{x}$ is optimal and that the rows a_i, $i = 1, \cdots, m$, of A are linearly independent. Then, choosing a vector $\bar{y}$ such that $A\bar{y} = 0$, we may set $I(\hat{x}) = \emptyset$, $IC(\hat{x}, \Omega) = E^n$, $y^* = 0$. Thus, from (i) of Theorem 12, $\psi^0(Q\hat{x} + d) + \sum_{i=1}^m \psi^i a_i = 0$, and clearly ψ^0 must be strictly less than 0. From (ii) of Theorem 12 we obtain $\langle \bar{y}, Q\bar{y}\rangle \geqq 0$, and since $\langle \bar{y}, Q\hat{x} + d\rangle = 0$, then $\langle d, \bar{y}\rangle = 0$ if $Q\bar{y} = 0$. In other words, we have the necessary conditions: (i) $\langle \bar{y}, Q\bar{y}\rangle \geqq 0$ for every $\bar{y}$ such that $A\bar{y} = 0$ and (ii) $\langle d, \bar{y}\rangle = 0$ for every $\bar{y}$ such that $A\bar{y} = 0$ and $Q\bar{y} = 0$. Note that these conditions do not involve $\hat{x}$. In fact, it can be shown that the existence of one feasible solution, together with conditions (i) and (ii), are sufficient conditions for the existence of an optimal solution to the preceding quadratic programming problem.

We have thus shown that most second order necessary conditions of optimality are special cases of Theorem 6 or Theorem 7 which we shall now proceed to prove.

4. Derivation of the major theorems. In this section proofs for Theorem 6 and Theorem 7 are given. Both proofs are developed within a geometric framework and are unified by the crucial use made of the Brouwer fixed-point theorem [14, pp. 146–150], [16]. Throughout this section it is to be understood that $\hat{x}$ is an optimal solution and that the map $F: E^n \to E^{m+1}$ and the *prohibited* ray P are defined by

$$F(x) = (f(x) - f(\hat{x}), r(x)) \tag{20}$$

and

$$P = \{w \in E^{m+1} | w^0 < 0 \text{ and } w^i = 0 \text{ for } i = 1, 2, \cdots, m\}. \tag{21}$$

The optimality of $\hat{x}$ leads to the simple but fundamental necessary condition, $F(\Omega) \cap P = \emptyset$.

Proof of Theorem 6. Let $\bar{y}$ be a critical direction at $\hat{x}$, and let $\{C(\hat{x}, \bar{y}, \Omega), y^*\}$ be a $\bar{y}$-directed conical approximation to Ω at $\hat{x}$. We define a convex set K in

E^{m+1} by

$$K = \frac{\partial^2 F}{\partial x^2}(\hat{x})(\tilde{y}, \tilde{y}) + \frac{\partial F}{\partial x}(\hat{x})(y^* + C(\hat{x}, \tilde{y}, \Omega)) \tag{22}$$

and a convex set K_m in E^m by $K_m = P_m(K)$, where $P_m(w^0, w^1, \cdots, w^m) = (w^1, w^2, \cdots, w^m)$.

If $\partial f(\hat{x})(\tilde{y})/\partial x = 0$, the claim of Theorem 6 is that the prohibited ray P (equation (21)) and the set K (equation (22)) are separated, while if $\partial f(\hat{x})(\tilde{y})/\partial x < 0$, the claim of Theorem 6 is that the origin in E^m is not in the interior of the set K_m. The proof is by contradiction.

Consider first the case $\partial f(\hat{x})(\tilde{y})/\partial x = 0$, and suppose that Theorem 6 is false, i.e., that the ray P and the convex set K are not separated. Then it follows that there exist $m + 1$ vectors $y_1, y_2, \cdots, y_{m+1}$ in $C(\hat{x}, \tilde{y}, \Omega)$, any m of which are linearly independent, such that the ray P has points in the interior of the set

$$co\left\{0, \frac{\partial^2 F}{\partial x^2}(\hat{x})(\tilde{y}, \tilde{y}) + \frac{\partial F}{\partial x}(\hat{x})(y^* + y_1), \cdots, \frac{\partial^2 F}{\partial x^2}(\hat{x})(\tilde{y}, \tilde{y}) + \frac{\partial F}{\partial x}(\hat{x})(y^* + y_{m+1})\right\}$$

(but P does not necessarily have points in the interior of the set K). Using the preceding relation and the map $\zeta_{\tilde{y}}(\cdot, \cdot)$ and α_0 associated with $y_1, y_2, \cdots, y_{m+1}$ (see Definition 5), the following results are readily established:

$$\hat{x} + \zeta_{\tilde{y}}(\alpha, y) \in \Omega \quad \text{for all} \quad \alpha \in (0, \alpha_0] \quad \text{and for all} \quad y \in co\{y_1, y_2, \cdots, y_{m+1}\}; \tag{23}$$

The set

$$\Sigma = co\{w_1, w_2, \cdots, w_m\}, \tag{24}$$

where

$$w_i = \frac{\partial^2 r}{\partial x^2}(\hat{x})(\tilde{y}, \tilde{y}) + \frac{\partial r}{\partial x}(\hat{x})(y^* + y_i)$$

for $i = 1, 2, \cdots, m + 1$ is a simplex[5] in E^m, containing the origin in its interior;

$$f(\hat{x} + \zeta_{\tilde{y}}(\alpha, y)) < f(\hat{x}) \quad \text{for all} \quad \alpha \in (0, \alpha_0'] \quad \text{and} \quad y \in co\{y_1, y_2, \cdots, y_{m+1}\} \quad \text{for some } \alpha_0' \in (0, \alpha_0]. \tag{25}$$

Now suppose that $\partial f(\hat{x})(\tilde{y})/\partial x < 0$, and that the corresponding statement of Theorem 6 is false. Then the origin in E^m is in the interior of the convex set K_m, and hence there exist vectors $y_1, y_2, \cdots, y_{m+1}$ in $C(\hat{x}, \tilde{y}, \Omega)$, any m of which are linearly independent, such that the origin is in the interior of the set Σ, where Σ is constructed as in (24). Thus (24) is again satisfied, and using the $\alpha_0 > 0$ and map $\zeta_{\tilde{y}}(\cdot, \cdot)$ associated with $y_1, y_2, \cdots, y_{m+1}$, (23) is satisfied. But since $\partial f(\hat{x})(\tilde{y})/\partial x < 0$, it follows that there exists an α_0' such that (25) is satisfied; and hence, irrespective of whether $\partial f(\hat{x})(\tilde{y})/\partial x = 0$ or $\partial f(\hat{x})(\tilde{y})/\partial x < 0$, if Theorem 6 is false, then (23), (24) and (25) are satisfied (for appropriate vectors $y_1, y_2, \cdots, y_{m+1}$, $\alpha_0 > 0$, $\alpha_0' > 0$ and map $\zeta_{\tilde{y}}(\cdot, \cdot)$). We shall now use these facts to complete the contradiction.

[5] A simplex in E^m is a convex polyhedron with $m + 1$ vertices which has an interior.

Now, let W be an $m \times m$ matrix whose ith column is $w_i - w_{m+1}$, $i = 1, 2, \cdots, m$, and let Y be a $n \times m$ matrix whose ith column is $y_i - y_{m+1}$, $i = 1, 2, \cdots, m$. Then W is nonsingular since Σ is a simplex. Hence, for every $w \in \Sigma$ and $\alpha \in (0, \alpha_0]$,

$$YW^{-1}(w - w_{m+1}) + y_{m+1} \in co\{y_1, y_2, \cdots, y_{m+1}\}$$

and $\zeta_{\tilde{y}}(\alpha, YW^{-1}(w - w_{m+1}) + y_{m+1}) \in \Omega - \hat{x}$.

For $\alpha \in (0, \alpha_0']$, we now define the continuous map $G_\alpha : \Sigma \to E^m$ by

$$G_\alpha(w) = w - \frac{2}{\alpha^2} r(\hat{x} + \zeta_{\tilde{y}}(\alpha, YW^{-1}(w - w_{m+1}) + y_{m+1})). \tag{26}$$

Then recalling the form of $\zeta_{\tilde{y}}(\cdot, \cdot)$ (see (6)) and noting that $r(\hat{x}) = 0$, $\partial r(\hat{x})(\tilde{y})/\partial x = 0$ and $\partial r(\hat{x})Y/\partial x = W$, we obtain

$$\begin{aligned} G_\alpha(w) = w - \frac{2}{\alpha^2}\Big\{ r(\hat{x}) &+ \alpha \frac{\partial r}{\partial x}(\hat{x})(\tilde{y}) + \frac{\alpha^2}{2}\frac{\partial r}{\partial x}(\hat{x})(YW^{-1}(w - w_{m+1}) + y_{m+1} + y^*) \\ &+ \frac{\alpha^2}{2}\frac{\partial^2 r}{\partial x^2}(\hat{x})(\tilde{y}, \tilde{y}) + o'(\alpha^2, w)\Big\} = -\frac{o'(\alpha^2, w)}{\alpha^2}, \end{aligned} \tag{27}$$

where $\|o'(\alpha^2, w)\|/\alpha^2 \to 0$ as $\alpha \to 0$, uniformly for $w \in \Sigma$. Hence, there exists an $\alpha^* \in (0, \alpha_0']$ such that $G_{\alpha^*}(\cdot)$ maps Σ into Σ, and therefore, by Brouwer's fixed-point theorem, there is a w^* in Σ such that $G_{\alpha^*}(w^*) = w^*$. But now, from (26), we see that the point

$$x^* = \hat{x} + \zeta_{\tilde{y}}(\alpha^*, YW^{-1}(w^* - w_{m+1}) + y_{m+1})$$

satisfies $r(x^*) = 0$, and since from (23), $x^* \in \Omega$, and from (25), $f(x^*) < f(\hat{x})$, we have a contradiction of the optimality of $\hat{x}$.

Thus, if $\partial f(\hat{x})(\tilde{y})/\partial x = 0$, then the set K and the ray P must be separated, while if $\partial f(\hat{x})(\tilde{y})/\partial x < 0$, the origin in E^m is not contained in the interior of the set K_m, and the statements of Theorem 6 follow.

Proof of Theorem 7. Since we need not distinguish in our proof between the cases (i), (ii) and (iii) in the statement of Theorem 7, it is convenient to define an indicator set J taking the form $J = \{0\}$ for case (i), $J = \{1\}$ for case (ii) or $J = \{0, 1\}$ for case (iii). For J taking any one of the preceding forms we define the continuous function $H : E^n \to E^{m+1}$ by

$$H^i(y) = \begin{cases} \dfrac{1}{2}\dfrac{\partial^2 F^i}{\partial x^2}(\hat{x})(y, y) & \text{for } i \in J, \\[2ex] \dfrac{\partial F^i}{\partial x}(\hat{x})(y) & \text{for } i \in \bar{J} = \{0, 1, \cdots, m\} \sim J. \end{cases} \tag{28}$$

If $C(\hat{x}, \Omega)$ is a conical approximation to Ω at x, the claim of Theorem 7 is now seen to be that the prohibited ray P has no points in the interior of the set $H(\text{int}\,(C(\hat{x}, \Omega)))$. The proof is by contradiction and requires a preliminary lemma.

LEMMA 6. *Let A_0 and A_1 be $n \times n$ symmetric matrices, and let $a_0, a_1, \cdots, a_m$ be vectors in E^m. Suppose that for J taking any one of the forms $J = \{0\}$, $J = \{1\}$ or $J = \{0, 1\}$, the function $\tilde{H}: E^n \to E^{m+1}$ is defined by $\tilde{H}^i(x) = \langle x, A_i x\rangle$ for $i \in J$ and $\tilde{H}^i(x) = \langle a_i, x\rangle$ for $i \in \bar{J}$, and suppose that C is a convex cone in E^n.*

If for J taking any one of the preceding forms, the prohibited ray P has points in the interior *of the set $\tilde{H}(C)$, then there exists an $\tilde{x}$ in C such that* (i) $\tilde{H}(\tilde{x}) \in P$ *and* (ii) *the Jacobian matrix $\partial\tilde{H}(\tilde{x})/\partial x$ has rank $m + 1$.*

Proof. We shall prove the lemma only for the case $J = \{0, 1\}$; the proofs for the other two cases are similar. Now, under the hypothesis of Lemma 6, there exists a vector $w_0 \in P$ such that w_0 belongs to the *interior* of $\tilde{H}(C)$. It follows that there is an $\bar{x} \in C$ such that (i) $\tilde{H}(\bar{x}) = w_0$, and (ii) the vectors $A_0\bar{x}, a_2, a_3, \cdots, a_m$ are linearly independent. If all of the vectors $A_0\bar{x}, A_1\bar{x}, a_2, a_3, \cdots, a_m$ are linearly independent, then $\partial\tilde{H}(\bar{x})/\partial x$ has rank $m + 1$, and we are finished. If they are not, then we shall construct a new vector $\tilde{x}$ which satisfies the conclusions of Lemma 6.

Suppose that $A_0\bar{x}, A_1\bar{x}, a_2, a_3, \cdots, a_m$ are linearly dependent. Then, since $\langle \bar{x}, A_0\bar{x}\rangle < 0$, we must have, for some $\beta^2, \beta^3, \cdots, \beta^m$,

$$A_1\bar{x} = \sum_{i=2}^{m} \beta^i a_i. \tag{29}$$

Now, letting $w_0 = (\alpha, 0, 0, \cdots, 0)$, where $\alpha < 0$, it is clear that for some $\gamma > 0$, we may choose x_1 and x_2 in C such that $\tilde{H}(x_1) = (\alpha, -\gamma, 0, 0, \cdots, 0)$ and $\tilde{H}(x_2) = (\alpha, \gamma, 0, 0, \cdots, 0)$. For $\mu \geqq 0$ and $\lambda \in [0, 1]$, let $x(\lambda, \mu) = \bar{x} + \mu(\lambda x_1 + (1 - \lambda)x_2)$, and observe that $x(\lambda, \mu) \in C$ and that $\langle a_i, x(\lambda, \mu)\rangle = 0$ for $i = 2, 3, \cdots, m$. It then follows from the symmetry of A_1 and relation (29) that

$$\langle x(\lambda, \mu), A_1 x(\lambda, \mu)\rangle = \mu^2\langle \lambda x_1 + (1 - \lambda)x_2, A_1(\lambda x_1 + (1 - \lambda)x_2)\rangle.$$

But now, by choice of x_1 and x_2, there is a $\lambda^* \in (0, 1)$ such that $\langle x(\lambda^*, \mu), A_1 x(\lambda^*, \mu)\rangle = 0$ for all $\mu \geqq 0$.

Finally, let $\mu^* > 0$ be chosen so that $\langle x(\lambda^*, \mu^*), A_1 x(\lambda^*, \mu^*\rangle < 0$, and let $\tilde{x} = x(\lambda^*, \mu^*)$. Then $\tilde{H}(\tilde{x}) \in P$, and $A_0\tilde{x}, A_1\tilde{x}, a_2, a_3, \cdots, a_m$ are linearly independent, since otherwise

$$A_1\bar{x} + \mu^*\lambda^* A_1 x_1 + \mu^*(1 - \lambda^*)A_1 x_2 = \sum_{i=2}^{m} \tilde{\beta}^i a_i$$

for some $\tilde{\beta}^2, \cdots, \tilde{\beta}^m$, i.e., from (29),

$$\lambda^* A_1 x_1 + (1 - \lambda^*)A_1 x_2 = \sum_{i=2}^{m} \bar{\beta}^i a_i$$

for some $\bar{\beta}^2, \cdots, \bar{\beta}^m$. But this implies $\langle x_1, A_1 x_2\rangle < 0$ and $\langle x_2, A_1 x_1\rangle > 0$, which is impossible, and hence the lemma is proved (for $J = \{0, 1\}$).

Let us now assume that Theorem 7 is false. Then it follows from Lemma 6 that there is a vector $\tilde{y} \in \operatorname{int}(C(\hat{x}, \Omega))$ such that $H(\tilde{y}) = w_0 \in P$ and such that the Jacobian $\partial H(\tilde{y})/\partial x$ has rank $m + 1$, where $H(\cdot)$ is defined by (28). We may assume without loss of generality that the first $m + 1$ columns of $\partial H(\tilde{y})/\partial x$ are linearly independent, hence, letting $\tilde{y} = (\tilde{y}', \tilde{y}'')$, where $\tilde{y}' = (\tilde{y}^1, \tilde{y}^2, \cdots, \tilde{y}^{m+1})$ and $\tilde{y}'' = (\tilde{y}^{m+2}, \cdots, \tilde{y}^n)$, it follows from the implicit function theorem [15] that there are closed neighborhoods U and V of the origin in E^{m+1} such that $H(\cdot)$ is a homeomorphism from $\{\tilde{y}' + U\} \times \{\tilde{y}''\}$ onto $w_0 + V$. So restricted, we shall denote the continuous inverse of this function by $H^-(\cdot)$. Clearly the set V may be assumed to be convex.

Since we may assume that U is sufficiently small, there is a linearly independent set of vectors $y_1, y_2, \cdots, y_n$ in $\operatorname{int}(C(\hat{x}, \Omega))$, with corresponding map $\zeta(\cdot)$ and $\alpha_0 > 0$, defined as in Definition 1, such that

$$\{\tilde{y}' + U\} \times \{\tilde{y}''\} \subset co\{0, y_1, y_2, \cdots, y_n\}.$$

We now define, for $\alpha \in (0, \alpha_0]$, the uniformly continuous map $G_\alpha : V \to E^{m+1}$ by

$$G_\alpha(w) = w_0 + w - D(\alpha)^{-1}F(\hat{x} + \zeta(\alpha H^-(w_0 + w))), \tag{30}$$

where $D(\alpha) = [d(\alpha)_{ij}]$ is an $(m + 1) \times (m + 1)$ nonsingular diagonal matrix such that for $i, j = 0, 1, \cdots, m$,

$$d(\alpha)_{ij} = \begin{cases} 0, & i \neq j, \\ \dfrac{\alpha^2}{2}, & i = j \text{ and } i \in J, \\ \alpha, & i = j \text{ and } i \in \bar{J}. \end{cases}$$

Expanding (30) and noting that $F(\hat{x}) = 0$, we obtain

$$\begin{aligned} G_\alpha(w) = w_0 + w - D(\alpha)^{-1}\Bigg(&\alpha\frac{\partial F}{\partial x}(\hat{x})(H^-(w_0 + w)) \\ &+ \frac{\alpha^2}{2}\frac{\partial^2 F}{\partial x^2}(\hat{x})(H^-(w_0 + w), H^-(w_0 + w)) + o(\alpha^2, w)\Bigg), \end{aligned} \tag{31}$$

which, recalling (28), may be rearranged to yield

$$G_\alpha(w) = w_0 + w - H(H^-(w_0 + w)) - D(\alpha)^{-1}\tilde{o}(\alpha, w) = -D(\alpha)^{-1}\tilde{o}(\alpha, w),$$

where, for $i = 0, 1, \cdots, m$, $|\tilde{o}^i(\alpha, w)|/d(\alpha)_{ii} \to 0$ as $\alpha \to 0$, uniformly for $w \in V$. Thus, we may choose $\alpha^* \in (0, \alpha_0]$ such that $G_{\alpha^*}(w) \in V$ for all $w \in V$, and, therefore, from Brouwer's fixed-point theorem there is $w^* \in V$ such that $G_{\alpha^*}(w^*) = w^*$. But now, from (30), the point $x^* = \hat{x} + \zeta(\alpha^* H^-(w_0 + w^*)) \in \Omega$ satisfies $F(x^*) = D(\alpha^*)w_0 \in P$, i.e., $f(x^*) < f(\hat{x})$ and $r(x^*) = 0$; this is a contradiction of the optimality of $\hat{x}$. This completes the proof of Theorem 7.

Conclusion. We have constructed in this paper a theory of second order conditions of optimality, which is consistent with the modern approach to first order necessary conditions. Also, we have shown that this theory not only results in a number of new conditions of optimality, but also yields most, if not all, the previously known second order conditions. The application of our results to specific nonlinear programming or optimal control problems is reasonably straightforward and, consequently, was not emphasized in our treatment.

In conclusion, we should like to point out that a number of the results in this paper may be extended to optimization problems in linear topological spaces. These extensions are obtained by stipulating the existence of suitable linear and bilinear functionals to replace the gradients and Hessians used in this paper.

REFERENCES

[1] M. Canon, C. Cullum and E. Polak, *Constrained minimization problems in finite-dimensional spaces*, this Journal, 4 (1966), pp. 528–547.

[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] K. Arrow, L. Hurwicz and H. Uzawa, *Constraint qualifications in maximization problems*, Naval Res. Logist. Quart., 8 (1961), pp. 175–191.

[4] I. M. Gelfand and S. V. Fomin, *Calculus of Variations*, Prentice-Hall, Englewood Cliffs, New Jersey, 1963.

[5] G. A. Bliss, *Lectures on the Calculus of Variations*, University of Chicago Press, Chicago, 1959.

[6] M. R. Hestenes, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.

[7] H. J. Kelley, R. E. Kopp and H. G. Gardner, *Singular extremals*, Topics in Optimization, Academic Press, New York, 1967.

[8] A. ya. Dubovickii and A. A. Milyutin, *Extremum problems with constraints*, Soviet Math. Dokl., 4 (1963), pp. 452–455.

[9] ———, *Second variations in extremal problems with constraints*, Ibid., 6 (1965), pp. 12–16.

[10] G. P. McCormick, *Second order conditions for constrained minima*, SIAM J. Appl. Math., 15 (1967), pp. 641–652.

[11] H. W. Kuhn and A. W. Tucker, *Nonlinear programming*, Proc. Second Berkeley Symp. on Mathematical Statistics and Probability, University of California Press, Berkeley, 1951, pp. 481–492.

[12] E. Polak and E. J. Messerli, *Second order conditions of optimality for constrained optimization problems in finite dimensional spaces*, Memo. ERL-M224, Electronics Research Laboratory, University of California, Berkeley, 1967.

[13] C. Berge, *Topological Spaces*, Macmillan, New York, 1963.

[14] L. M. Graves, *The Theory of Functions of a Real Variable*, 2nd ed., McGraw-Hill, New York, 1956.

[15] J. Dieudonné, *Foundations of Modern Analysis*, Academic Press, New York, 1960.

[16] R. E. Edwards, *Functional Analysis: Theory and Applications*, Holt, Rhinehart and Winston, New York, 1965.

SIAM J. CONTROL
Vol. 7, No. 2, May 1969

ON THE ADMISSIBLE SYNTHESIS IN OPTIMAL CONTROL THEORY AND DIFFERENTIAL GAMES*

STEFAN MIRICĂ†

1. Introduction.

In [4] the notion of a "regular" admissible synthesis is presented for the time optimal control problem.

In [2] Berkovitz uses a strategy inducing a "regular decomposition" of the phase space for a class of differential games. The control problems corresponding to the differential games considered are nonautonomous, with n-dimensional terminal manifold in $(n + 1)$-dimensional (t, x)-space.

In the present paper we combine the ideas from [2] and [4] in order to define an "admissible synthesis" for a nonautonomous control problem, with general terminal manifold.

In the first part of this paper we study the properties of the trajectories generated by the admissible synthesis, dual variables and the value of the functional along these trajectories, using the same methods as in [2].

In the second part, we use the notions introduced in [4] (see also [5]) to prove that the functional equation of dynamic programming and a certain form of the maximum principle are equivalent and represent necessary and sufficient conditions of optimality for the trajectories generated by the synthesis.

Thus, the problem of admissibility is separated from the problem of optimality, and this allows us to see when and how the optimality conditions occur.

By optimal synthesis we do not mean a synthesis which is optimal with respect to other admissible syntheses, but a synthesis generating optimal trajectories with respect to a much wider class of admissible trajectories.

In the final section we present analogous results for a class of differential games.

2. Statement of problem and definition of the admissible synthesis.

We consider a bounded region $\mathscr{G} \subset R \times R^n$ and a bounded region $\mathscr{C} \subset R \times R^n \times R^p$. We let U denote the projection of $\mathscr{C}$ into R^p; U is generally supposed to be a closed set. $\mathscr{G}$ is assumed to be open. The vector-valued (with values in R^n) function $f(t, x, u)$ and the real-valued function $f^0(t, x, u)$ defined on $\mathscr{C}$ are assumed to be of class $C^{(1)}$ with respect to (x, u) and continuous in t.

The differentiable manifold $\mathscr{T} \subset \overline{\mathscr{G}}$ of class $C^{(1)}$ and of dimension k, $0 \leqq k \leqq n$, will be called the *terminal manifold* (for $k = 0$, $\mathscr{T}$ a point of $\overline{\mathscr{G}}$).

A real-valued function g of class $C^{(1)}$ is given on $\mathscr{G}_1$, a subregion containing $\mathscr{T}$ in its interior.

* Received by the editors March 5, 1968, and in final revised form November 4, 1968.

† Facultatea de Matematica-Mecanica, Universitatea Bucuresti, Bucuresti, Romania.

The control system is:

$$\text{(2.1)} \qquad \frac{dx}{dt} = f(t, x, u), \qquad x(\tau) = \xi, \qquad (\tau, \xi) \in \mathscr{G}, \qquad u \in U.$$

DEFINITION 2.1. The vector-valued function $u(t)$ defined on $I \subset pr_R\mathscr{G}$ with values in U is called an *admissible control corresponding to the initial point* $(\tau, \xi) \in \mathscr{G}$ if:

(i) it is piecewise continuous in I;

(ii) the system of differential equations

$$\text{(2.2)} \qquad \frac{dx}{dt} = f(t, x, u(t)), \quad x(\tau) = \xi$$

has the solution $\varphi(t; \tau, \xi)$, with $\varphi(\tau; \tau, \xi) = \xi$, which remains in $\mathscr{G}$ and intersects $\mathscr{T}$ in a finite time (that is, there exists $t_1 > \tau$, so that

$$\{(t, \varphi(t; \tau, \xi)) | \tau \leqq t < t_1\} \subset \mathscr{G} - \mathscr{T}$$

and

$$(t_1, \varphi(t_1; \tau, \xi)) \in \mathscr{T}).$$

We shall call the curve $\{(t, \varphi(t; \tau, \xi)) | \tau \leqq t \leqq t_1\}$ an admissible trajectory of the system (2.1) related to the initial point (τ, ξ).

We let $\mathscr{U}(\tau, \xi)$ denote the class of the admissible controls corresponding to (τ, ξ) and let $\mathscr{U}$ denote the class of admissible controls corresponding to all the points of $\mathscr{G}$.

For every admissible trajectory we define the functional

$$\text{(2.3)} \qquad P(\tau, \xi, u) = g(t_1, x_1) + \int_\tau^{t_1} f^0(t, \varphi(t; \tau, \xi), u(t)) \, dt,$$

where $x_1 = \varphi\,(t_1; \tau, \xi)$ and $(t_1, x_1) \in \mathscr{T}$.

DEFINITION 2.2. The admissible control $\tilde{u} \in \mathscr{U}(\tau, \xi)$ is *optimal* with respect to the initial point $(\tau, \xi) \in \mathscr{G}$ if

$$P(\tau, \xi, u) = V(\tau, \xi) \leqq P(\tau, \xi, u) \quad \text{for any } u \in \mathscr{U}(\tau, \xi).$$

In what follows, the notions of "curvilinear polyhedron" and "piecewise smooth set" are those defined by Boltyanskii in [4], [5].

Let $N, P^k, \cdots, P^n$ be piecewise smooth sets, such that $\mathscr{T} \subset P^k \subset P^{k+1} \subset \cdots \subset P^n \subset \mathscr{G}$, and let $v(t, x)$ be a function defined on $\mathscr{G}$ with values in U.

We denote $P^{k-1} = \mathscr{T}$, $P^{n+1} = \mathscr{G}$.

DEFINITION 2.3 (Boltyanskii [4], [5]). The sets P^k, $P^{k+1}, \cdots, P^{n+1}$ and the function $v(t, x)$ represent an *admissible synthesis* for the control problem if the following requirements are fulfilled:

A. (i) The connected components of the sets $P^i - (P^{i-1} \cup N), i = k, k + 1, \cdots, n + 1$, are differentiable manifolds of class $C^{(1)}$ of dimension i; we call them i-dimensional cells.

(ii) The function $v(t, x)$ is of class $C^{(1)}$ on every cell and can be extended to a function of class $C^{(1)}$ on a neighborhood of the cell.

B. Every cell is either of type I or of type II.

(i) The $(n + 1)$-dimensional cells are of type I, the k-dimensional ones of type II.

(ii) If c is an i-dimensional cell of type I, then through any point $(t, x) \in c$, there passes a unique trajectory of the system:

$$\frac{dx}{dt} = f(t, x, v(t, x)) \tag{2.4}$$

with the following property:

There exists a unique $(i - 1)$-dimensional cell $\Pi(c)$ (of type I or II) such that the trajectory starting from $(t, x) \in c$ leaves c after a finite time and reaches $\Pi(c)$ at a nonzero angle. (In particular, at the point of intersection we have $f(t, x, v(t, x)) \neq 0$.)

(iii) If c is an i-dimensional cell of type II, $i > k$, then there is a unique $(i + 1)$-dimensional cell $\Sigma(c)$ of type I so that from any point $(t, x) \in c$, there starts a unique trajectory of the system (2.4) reaching $\Sigma(c)$ and having only one common point with c. Moreover, the function v is of class $C^{(1)}$ on $c \cup \Sigma(c)$.

It follows that a trajectory of the system may be prolonged from cell to cell as follows: from c to $\Pi(c)$ if $\Pi(c)$ is of type I, and from c to $\Sigma(\Pi(c))$ if $\Pi(c)$ is of type II.

C. (i) Every trajectory of the system (2.4) remains in $\mathscr{G}$, reaches $\mathscr{T}$ in a finite time and is not tangent to $\mathscr{T}$.

(ii) Every trajectory goes through a finite number of cells.

(iii) From a given point in N there need not necessarily emanate a unique trajectory of the system (2.4). Trajectories of (2.4) starting at points in N do not remain in N, but reach a cell of type I. The value (2.3) is the same for all trajectories starting from the same point in N.

D. The value (2.3) along all the trajectories which fulfill the requirements A–C is continuous in $\mathscr{G}$.

As in [4], we shall call these trajectories *marked trajectories* and denote them by $x = \Phi(t; \tau, \xi)$. Denote by t_F the first moment when the trajectory reaches $\mathscr{T}$, that is, t_F is a real number such that

$$\{(t, \Phi(t; \tau, \xi)) | \tau \leqq t < t_F\} \subset \mathscr{G} - \mathscr{T}$$

and $x_F = \Phi(t_F; \tau, \xi), (t_F, x_F) \in \mathscr{T}$. For a marked trajectory, the value of the functional will be

$$W(\tau, \xi) = g(t_F, x_F) + \int_\tau^{t_F} f^0(t, \Phi(t; \tau; \xi), v(t, \Phi(t; \tau, \xi)))\, dt. \tag{2.5}$$

Remark 2.1. It is obvious that the function $\bar{u}(t) = v(t, \Phi(t; \tau, \xi))$ is an admissible control corresponding to the initial point (τ, ξ), and therefore the marked trajectories are admissible according to Definition 2.1. That is, the set of marked trajectories (generated by admissible synthesis) is included in the set of admissible trajectories.

3. Properties of the marked trajectories. Since c_1 is an $(n + 1)$-dimensional cell (hence of type I), for any $(\tau, \xi) \in c_1$, we obtain a unique marked trajectory $x = \Phi(t; \tau, \xi)$ starting from (τ, ξ) and reaching $\mathscr{T}$ in (t_F, x_F).

From the definition of admissible synthesis we deduce the following properties of marked trajectories:

(a) The trajectory passes through a finite number of type I cells $c_1, c_2, \cdots, c_q$.

(b) For any $1 \leqq i \leqq q$, if $\Pi(c_i)$ is of type I, then $c_{i+1} = \Pi(c_i)$ and the trajectory goes from c_i directly to c_{i+1}; if $\Pi(c_i)$ is of type II, then $c_{i+1} = \Sigma(\Pi(c_i))$ and the trajectory goes from c_i to c_{i+1} by crossing the manifold $\Pi(c_i)$ in a single point. In this case, c_{i+1} and c_i have the same dimension. For $i = q$ we let $\Pi(c_q) = \mathscr{T}$.

(c) Let t_i, $i = 1, 2, \cdots, q$, denote the moments at which the trajectory reaches the cell $\Pi(c_i)$, $i = 1, 2, \cdots, q, t_q = t_F$, and let $x_i = \Phi(t_i; \tau, \xi)$. Then $(t_i, x_i) \in \Pi(c_i)$, $i = 1, 2, \cdots, q$, and $(t, \Phi(t; \tau, \xi)) \in c_i$ for $t_{i-1} < t < t_i$ if $\Pi(c_i)$ is a cell of type II, and $(t, \Phi(t; \tau, \xi)) \in c_i$ for $t_{i-1} \leqq t < t_i$ if $\Pi(c_i)$ is of type I.

Now let $\Pi(c_i)$ be m-dimensional, $k \leqq m \leqq n$. Then in a neighborhood of the point $(t_i, x_i) \in \Pi(c_i)$, the points of the manifold $\Pi(c_i)$ are given parametrically by the relations:

$$(3.1)_{(i)} \qquad \begin{aligned} t &= T_{(i)}(\theta^1, \cdots, \theta^m), \\ x &= \chi_{(i)}(\theta^1, \cdots, \theta^m), \end{aligned}$$

where $(\theta^1, \theta^2, \cdots, \theta^m) \in \mathscr{K}^m$, the m-dimensional cube. Then, there exists a point $(\theta_0^1, \cdots, \theta_0^m) \in \mathscr{K}^m$ such that

$$(3.1')_{(i)} \qquad \begin{aligned} t_i &= T_{(i)}(\theta_0^1, \cdots, \theta_0^m), \\ x_i &= \chi_{(i)}(\theta_0^1, \cdots, \theta_0^m). \end{aligned}$$

The functions $T_{(i)}, \chi_{(i)}$ are of class $C^{(1)}$. We shall use the same notation (θ) for the parameters on every manifold $\Pi(c_i)$ and, for $i = q$, we denote

$$\begin{aligned} T_{(q)}(\theta^1, \cdots, \theta^k) &= T(\theta^1, \cdots, \theta^k), \\ \chi_{(q)}(\theta^1, \cdots, \theta^k) &= \chi(\theta^1, \cdots, \theta^k). \end{aligned}$$

The function $\Phi(t; \tau, \xi)$ defined on the interval $[\tau, t_F]$ is the solution of the system:

$$(3.2) \qquad \frac{dx}{dt} = f(t, x, v(t, x)), \quad x(\tau) = \xi,$$

where $f(t, x, v(t, x))$ is a function of class $C^{(1)}$ on every cell. As some of these cells are of smaller dimension than $n + 1$, we cannot apply directly the theorems on continuity and differentiability with respect to the initial conditions for this system.

To obtain some properties of this kind for $\Phi(t;\tau,\xi)$, we shall use the hypothesis A (ii) from Definition 2.3 of the admissible synthesis. The function $v(t,x)$ can be extended to a function $v_{(i)}(t,x)$ of class $C^{(1)}$ on a neighborhood $\tilde{c}_i$ of the manifold c_i. We shall then use the systems

$$(3.3)_{(i)} \qquad \frac{dx}{dt} = f(t,x,v_{(i)}(t,x)), \quad x(\alpha) = \beta,$$

where $(\alpha,\beta) \in \tilde{c}_i$.

LEMMA 3.1. (i) *The points* (t_i, x_i), $i = 1, 2, \cdots, q$, *where the trajectory passes from one cell to another, are functions of class* $C^{(1)}$ *with respect to* $(\tau,\xi) \in c_1$.

(ii) *The function* $\Phi(t;\tau,\xi)$ *is continuous with respect to t on the interval* $[\tau, t_F]$ *and with respect to* $(\tau,\xi) \in c_1$. *The derivatives*

$$\frac{\partial\Phi}{\partial\tau}(t;\tau,\xi), \quad \frac{\partial\Phi}{\partial\xi}(t;\tau,\xi), \quad \frac{d}{dt}\left(\frac{\partial\Phi}{\partial\tau}(t;\tau,\xi)\right), \quad \frac{d}{dt}\left(\frac{\partial\Phi}{\partial\xi}(t;\tau,\xi)\right)$$

are continuous on every interval (t_i, t_{i+1}) *and have one-sided limits at the points* $t = t_i$.

(iii) *The matrices*

$$(3.4)_{(i)} \qquad N_{(i)} = \left(\frac{d\Phi}{dt}(t_i - 0;\tau,\xi)\frac{\partial T_{(i)}}{\partial\theta}(\theta_0) - \frac{\partial\chi}{\partial\theta}(i)(\theta_0)\right), \qquad i = 1, 2, \cdots, q,$$

have maximum rank (*equal to the dimension of the cell* $\Pi(c_i)$).

(iv) *The one-sided limits at the points* $t = t_i$, $i = 1, 2, \cdots, q$, *of the derivatives in* (ii) *satisfy the relations*:

$$(3.5)_{(i)} \qquad \begin{aligned} \frac{\partial\Phi}{\partial\xi}(t_i - 0;\tau,\xi) &= -f(t_i; x_i, v_{(i)}(t_i, x_i))\frac{\partial t_i}{\partial\xi}(\tau,\xi) + \frac{\partial x_i}{\partial\xi}(\tau,\xi), \\ \frac{\partial\Phi}{\partial\xi}(t_i + 0;\tau,\xi) &= -f(t_i; x_i, v_{(i+1)}(t_i, x_i))\frac{\partial t_i}{\partial\xi}(\tau,\xi) + \frac{\partial x_i}{\partial\xi}(\tau,\xi), \end{aligned}$$

$$(3.6)_{(i)} \qquad \begin{aligned} \frac{\partial\Phi}{\partial\tau}(t_i - 0;\tau,\xi) &= -f(t_i; x_i, v_{(i)}(t_i; x_i))\frac{\partial t_i}{\partial\tau}(\tau,\xi) + \frac{\partial x_i}{\partial\tau}(\tau,\xi), \\ \frac{\partial\Phi}{\partial\tau}(t_i + 0;\tau,\xi) &= -f(t_i; x_i, v_{(i+1)}(t_i; x_i))\frac{\partial t_i}{\partial\tau}(\tau,\xi) + \frac{\partial x_i}{\partial\tau}(\tau,\xi). \end{aligned}$$

Proof. We shall prove the statements (i)–(iii) by recurrence.

The function $\Phi(t;\tau,\xi)$, the solution of the system (3.2), is defined on the interval $[\tau, t_1]$. Also the system (3.2) coincides with $(3.3)_{(1)}$ for $(t,x) \in c_1$. On the other hand, for the system of differential equations

$$(3.3)_{(1)} \qquad \frac{dx}{dt} = f(t,x,v_{(1)}(t,x)), \quad x(\alpha) = \beta,$$

we can apply the theorems on continuity and differentiability with respect to the initial conditions (for instance, Theorem 15 from [8]) and so, we can state: there exist $r_1, r_2 > 0$ so that, for $|t - \alpha| < r_1, |\tau - \alpha| < r_2$, $|\xi - \beta| < r_2$, there exists the unique solution $x = \psi_{(1)}(t; \tau, \xi)$ of the system $(3.3)_{(1)}$ with $\psi_{(1)}(\tau; \tau, \xi) = \xi$ continuous with respect to the variables (t, τ, ξ) together with the derivatives

$$\frac{\partial \psi_{(1)}}{\partial \tau}(t; \tau, \xi), \quad \frac{\partial \psi_{(1)}}{\partial \xi}(t; \tau, \xi), \quad \frac{d}{dt}\left(\frac{\partial \psi_{(1)}}{\partial \tau}(t; \tau, \xi)\right), \quad \frac{d}{dt}\left(\frac{\partial \psi_{(1)}}{\partial \xi}(t; \tau, \xi)\right).$$

Because of the uniqueness, the functions $\Phi(t; \tau, \xi)$ and $\psi_{(1)}(t; \tau, \xi)$ will coincide on the common domain of definition for $(\tau, \xi) \in c_1$. Since $\Phi(t; \tau, \xi)$ is defined on $[\tau, t_1]$, the solution $\psi_{(1)}(t; \tau, \xi)$ can be prolonged, and therefore $\psi_{(1)}(t; \tau, \xi)$ is defined on some interval $(\tau - \varepsilon, t_1 + \varepsilon)$, $\varepsilon > 0$. Since $(t_1, x_1) \in \Pi(c_1)$ and $\Pi(c_1)$ is n-dimensional, in a neighborhood of (t_1, x_1) we have the representation

$$(3.1)_{(1)} \qquad \begin{aligned} t &= T_{(1)}(\theta^1, \cdots, \theta^n), \\ x &= \chi_{(1)}(\theta^1, \cdots, \theta^n), \end{aligned}$$

and there exists $(\theta_0^1, \cdots, \theta_0^n)$ so that

$$(3.1')_{(1)} \qquad \begin{aligned} t_1 &= T_{(1)}(\theta_0^1, \cdots, \theta_0^n), \\ x_1 &= \chi_{(1)}(\theta_0^1, \cdots, \theta_0^n). \end{aligned}$$

Since $x_1 = \Phi(t_1; \tau, \xi) = \psi_{(1)}(t; \tau, \xi)$, we have

$$(3.6)_{(1)} \qquad \Phi(T_{(1)}(\theta_0^1, \cdots, \theta_0^n); \tau, \xi) = \chi_{(1)}, (\theta_0^1, \cdots, \theta_0^n).$$

or

$$(3.6)_{(1)} \qquad \psi_{(1)}(T_{(1)}(\theta_0^1, \cdots, \theta_0^n); \tau, \xi) = \chi_{(1)}(\theta_0^1, \cdots, \theta_0^n).$$

If (τ, ξ) ranges over a neighborhood in c_1, then $(\theta^1, \cdots, \theta^n)$ ranges over a neighborhood of $(\theta_0^1, \cdots, \theta_0^n)$ in the n-dimensional cube $\mathscr{K}^n$.

We shall first show that the relations $(3.6)_{(1)}$ give $\theta^1, \cdots, \theta^n$ as implicit functions of class $C^{(1)}$ in (τ, ξ).

The functional matrix of the equations $(3.6)_{(1)}$ is

$$N_{(1)} = \left(\frac{d\psi_{(1)}}{dt}(T_{(1)}(\theta_0); \tau, \xi)\frac{\partial T_{(1)}}{\partial \theta}(\theta_0) - \frac{\partial \chi_{(1)}}{\partial \theta}(\theta_0)\right)$$

$$= \left(\frac{d\Phi}{dt}(t_1 - 0; \tau, \xi)\frac{\partial T_{(1)}}{\partial \theta}(\theta_0) - \frac{\partial \chi_{(1)}}{\partial \theta}(\theta_0)\right).$$

On the other hand, because of the nontangency of the trajectory $\Phi(t; \tau, \xi)$ to the manifold $\Pi(c_1)$, the vectors

$$\gamma^{(1)} = \begin{pmatrix} 1 \\ \\ \dfrac{d\Phi}{dt}(t_1 - 0; \tau, \xi) \end{pmatrix}, \qquad \delta_{(i)}^{(1)} = \begin{pmatrix} \dfrac{\partial T_{(1)}}{\partial \theta^i}(\theta_0) \\ \\ \dfrac{\partial \chi_{(1)}}{\partial \theta^i}(\theta_0) \end{pmatrix}, \qquad i = 1, 2, \cdots, n,$$

are linearly independent; that is, the matrix

$$M_{(1)} = \begin{pmatrix} 1 & \dfrac{\partial T_{(1)}}{\partial \theta}(\theta_0) \\ \\ \dfrac{d\Phi}{dt}(t_1 - 0; \tau, \xi) & \dfrac{\partial \chi_{(1)}}{\partial \theta}(\theta_0) \end{pmatrix}.$$

has the maximum rank, $n + 1$.

Since the rank of a matrix does not change if we apply elementary transformations, the rank of matrix $M_{(1)}$ is the same as that of the matrix $M'_{(1)}$ obtained by multiplying the first column by $-\partial T_{(1)}/\partial \theta^i(\theta_0)$ and adding the result to each of the other columns:

$$M'_{(1)} = \begin{pmatrix} 1 & 0 \\ \dfrac{d\Phi}{dt}(t_1 - 0; \tau, \xi) & -N_{(1)} \end{pmatrix}.$$

It is obvious that the rank of the matrix $M'_{(1)}$ is $n + 1$ if and only if the rank of the matrix $N_{(1)}$ is n.

Hence we can apply the implicit functions theorem to the relations $(3.6)_{(i)}$ and obtain that $\theta^i = \theta^i(\tau, \xi)$, $i = 1, \cdots, n$, are $C^{(1)}$-functions of (τ, ξ) in some neighborhood of c_1.

It follows that the functions

$$t_1(\tau, \xi) = T_{(1)}(\theta(\tau, \xi)),$$

$$x_1(\tau, \xi) = \chi_{(1)}(\theta^1(\tau, \xi), \cdots, \theta^n(\tau, \xi))$$

are of class $C^{(1)}$ in (τ, ξ) on c_1.

To prolong the trajectory, let us consider·the two possible cases:

(a) If $\Pi(c_1)$ is of type I and $c_2 = \Pi(c_1)$ is n-dimensional, then the system $(3.3)_{(2)}$ will coincide with (3.2) for $\alpha_1 = t_1$, $\beta = x_1$.

The solution $\Phi(t; t_1, x_1) = \Phi(t; \tau, \xi)$ of the system

$$\frac{dx}{dt} = f(t, x, v(t, x)), \quad x(t_1) = x_1$$

will be defined on the interval (t_1, t_2) and will coincide with the solution $\psi_{(2)}(t; t_1, x_1)$ of the system

$$(3.3)_{(2)} \qquad \frac{dx}{dt} = f(t, x, v_{(2)}(t, x)), \quad x(t_1) = x_1$$

for $t \in [t_1, t_2)$. As $\psi_{(2)}(t; t_1, x_1)$ may be prolonged further, it follows that it is defined on $(t_1 - \delta, t_2 + \delta)$, $\delta > 0$, and is continuous together with its derivatives of the first order with respect to t, t_1, x_1.

Therefore we have

$$\Phi(t; t_1, x_1) = \psi_{(2)}(t; t_1, x_1) = \Phi(t; \tau, \xi)$$

for $t \in [t_1, t_2]$. In particular,

$$x_2 = \Phi(t_2; \tau, \xi) = \psi_{(2)}(t_2; t_1, x_1).$$

Furthermore, in a neighborhood of the point $(t_2, x_2) \in \Pi(c_2)$ ($\Pi(c_2)$ being an $(n-1)$-dimensional differentiable manifold), the points of $\Pi(c_2)$ are given parametrically by the relations

$$(3.1)_{(2)} \qquad \begin{aligned} t &= T_{(2)}(\theta^1, \cdots, \theta^{n-1}), \\ x &= \chi_{(2)}(\theta^1, \cdots, \theta^{n-1}), \end{aligned}$$

where $(\theta^1, \cdots, \theta^{n-1}) \in \mathscr{K}^{n-1}$, the $(n-1)$-dimensional cube.

Note. θ^i in $(3.1)_{(2)}$ is not the same as θ^i in $(3.1)_{(1)}$.

For fixed (τ, ξ), the point (t_2, x_2) is perfectly determined on the manifold $\Pi(c_2)$. Hence there exists a point $(\theta_0^1, \cdots, \theta_0^{n-1})$ so that

$$(3.1)_{(2)} \qquad \begin{aligned} t_2 &= T_{(2)}(\theta_0^1, \cdots, \theta_0^{n-1}), \\ x_2 &= \chi_{(2)}(\theta_0^1, \cdots, \theta_0^{n-1}). \end{aligned}$$

As above, we shall show that the relations

$$(3.6)_{(2)} \qquad \Phi(T_{(2)}(\theta^1, \cdots, \theta^{n-1}); t_1, x_1) = \chi_{(2)}(\theta^1, \cdots, \theta^{n-1}),$$

$$(3.6')_{(2)} \qquad \psi_{(2)}(T_{(2)}(\theta^1, \cdots, \theta^{n-1}); t_1, x_1) = \chi_{(2)}(\theta^1, \cdots, \theta^{n-1})$$

define $\theta^1, \cdots, \theta^{n-1}$ as implicit functions of class $C^{(1)}$ with respect to (t_1, x_1). Since we have already shown that (t_1, x_1) is a $C^{(1)}$-function of (τ, ξ), it will follow that $\theta^1, \cdots, \theta^{n-1}$ are $C^{(1)}$-functions of (τ, ξ).

The functional matrix of the equations $(3.6')_{(2)}$ is

$$\begin{aligned} N_{(2)} &= \left(\frac{d\psi_{(2)}}{dt}(T_{(2)}(\theta_0); t_1, x_1) \frac{\partial T_{(2)}}{\partial \theta}(\theta_0) - \frac{\partial \chi_{(2)}}{\partial \theta}(\theta_0) \right) \\ &= \left(\frac{d\Phi}{dt}(t_2 - 0 : \tau, \xi) \frac{\partial T_{(2)}}{\partial \theta}(\theta_0) - \frac{\partial \chi_{(2)}}{\partial \theta}(\theta_0) \right). \end{aligned}$$

On the other hand, because of the nontangency of the trajectory $\Phi(t; \tau, \xi)$ to the manifold $\Pi(c_2)$, the matrix

$$M_{(2)} = \begin{pmatrix} 1 & \dfrac{\partial T_{(2)}}{\partial \theta}(\theta_0) \\ \dfrac{d\Phi}{dt}(t_2 - 0; \tau, \xi) & \dfrac{\partial \chi_{(2)}}{\partial \theta}(\theta_0) \end{pmatrix}$$

has maximum rank n, and so does the matrix

$$M'_{(2)} = \begin{pmatrix} 1 & 0 \\ \dfrac{d\Phi}{dt}(t_2 - 0; \tau, \xi) & -N_{(2)} \end{pmatrix}.$$

It follows that the rank of matrix $N_{(2)}$ is $n - 1$, and from the implicit functions theorem it follows that the relations $(3.6')_{(2)}$ define $\theta^1, \cdots, \theta^{n-1}$ as $C^{(1)}$-functions of (t_1, x_1) and therefore of (τ, ξ).

Hence, the functions

$$t_2(\tau, \xi) = T_{(2)}(\theta^1(\tau, \xi), \cdots, \theta^{n-1}(\tau, \xi)), \quad x_2(\tau, \xi) = \chi_{(2)}(\theta^1(\tau, \xi), \cdots, \theta^{n-1}(\tau, \xi))$$

are of class $C^{(1)}$.

(b) If $\Pi(c_1)$ is a cell of type II, then $c_2 = \Sigma(\Pi(c_1))$ is of dimension $n + 1$ and $v(t, x)$ is of class $C^{(1)}$ on $\Pi(c_1) \cup c_2$. Applying the arguments of the case (a) to the manifold $\Pi(c_1) \cup c_2$, we obtain the sentences (i)–(iii) of Lemma 3.1 for $t_1 \leqq t \leqq t_2$.

Continuing in this way, we obtain the proof for the whole interval $[\tau, t_F]$. In particular, $t_q(\tau, \xi) = t_F(\tau, \xi)$, $x_q(x, \xi) = x_F(\tau, \xi)$ are $C^{(1)}$ functions of $(\tau, \xi) \in c_1$ and

$$(3.7) \qquad \begin{aligned} \frac{\partial t_i}{\partial \tau}(\tau, \xi) &= \sum_{l=1}^{m_i} \frac{\partial T_{(i)}}{\partial \theta^l}(\theta_0) \frac{\partial \theta^l}{\partial \tau}(\tau, \xi), \\ \frac{\partial t_i}{\partial \xi}(\tau, \xi) &= \sum_{l=1}^{m_i} \frac{\partial T_{(i)}}{\partial \theta^l}(\theta_0) \frac{\partial \theta^l}{\partial \xi}(\tau, \xi), \end{aligned}$$

where m_i is the dimension of the manifold $\Pi(c_i)$. Analogous formulas hold for $\partial x_i/\partial \tau$, $\partial x_i/\partial \xi$.

(iv) To prove $(3.5)_{(i)}$ and $(3.5')_{(i)}$, we shall use the relations

$$(3.8) \qquad \begin{aligned} x_i = \Phi(t_i; \tau, \xi) &= \psi_{(i)}(t_i; t_{i-1}, x_{i-1}) \\ &= \psi_{(i+1)}(t_i; t_i, x_i), \end{aligned}$$

which we have previously obtained. For $t_{i-1} < t < t_i$, $i = 1, 2, \cdots, q$, we have

$$(3.9) \qquad \Phi(t; \tau, \xi) = \psi_{(i)}(t; t_{i-1}(\tau, \xi), x_{i-1}(\tau, \xi)).$$

Then:

$$(3.10) \qquad \begin{aligned} \frac{\partial \Phi}{\partial \xi}(t; \tau, \xi) &= \frac{\partial \psi_{(i)}}{\partial \alpha}(t; t_{i-1}, x_{i-1}) \frac{\partial t_{i-1}}{\partial \xi}(\tau, \xi) + \frac{\partial \psi_{(i)}}{\partial \beta}(t; t_{i-1}, x_{i-1}) \frac{\partial x_{i-1}}{\partial \xi}(\tau, \xi), \\ \frac{\partial \Phi}{\partial \tau}(t; \tau, \xi) &= \frac{\partial \psi_{(i)}}{\partial \alpha}(t; t_{i-1}, x_{i-1}) \frac{\partial t_{i-1}}{\partial \tau}(\tau, \xi) + \frac{\partial \psi_{(i)}}{\partial \beta}(t; t_{i-1}, x_{i-1}) \frac{\partial x_{i-1}}{\partial \tau}(\tau, \xi). \end{aligned}$$

For $t \to t_i - 0$, we have:

$$(3.11) \qquad \frac{\partial \Phi}{\partial \xi}(t_i - 0; \tau, \xi) = \frac{\partial \psi_{(i)}}{\partial \alpha}(t_i; t_{i-1}, x_{i-1}) \frac{\partial t_{i-1}}{\partial \xi}(\tau, \xi) + \frac{\partial \psi_{(i)}}{\partial \beta}(t_i; t_{i-1}, x_{i-1}) \frac{\partial x_{i-1}}{\partial \xi}(\tau, \xi),$$

and similarly for the derivatives with respect to τ. But (3.8) may be written

$$x_i(\tau, \xi) = \psi_{(i)}(t_i(\tau, \xi); t_{i-1}(\tau, \xi), x_{i-1}(\tau, \xi)). \tag{3.12}$$

Hence

$$\begin{aligned}\frac{\partial x_i}{\partial \xi}(\tau, \xi) &= \frac{d\psi_{(i)}}{dt}(t_i; t_{i-1}, x_{i-1})\frac{\partial t_i}{\partial \xi}(\tau, \xi)\\ &+ \frac{\partial \psi_{(i)}}{\partial \alpha}(t_i; t_{i-1}, x_{i-1})\frac{\partial t_{i-1}}{\partial \xi}(\tau, \xi) + \frac{\partial \psi_{(i)}}{\partial \beta}(t_i; t_{i-1}, x_{i-1})\frac{\partial x_{i-1}}{\partial \xi}(\tau, \xi),\end{aligned} \tag{3.13}$$

and similarly for the derivatives with respect to τ.

As $\psi_{(i)}(t; t_{i-1}, x_{i-1})$ is the solution of the system $(3.3)_{(i)}$, we have:

$$\frac{d\psi_{(i)}}{dt}(t_i; t_{i-1}, x_{i-1}) = f(t_i, x_i, v_{(i)}(t_i, x_i)). \tag{3.14}$$

Hence, according to (3.11) and (3.13), we obtain:

$$\frac{\partial \Phi}{\partial \xi}(t_i - 0; \tau, \xi) = -f(t_i, x_i, v_{(i)}(t_i, x_i))\frac{\partial t_i}{\partial \xi}(\tau, \xi) + \frac{\partial x_i}{\partial \xi}(\tau, \xi),$$

and similarly for $\partial \Phi(t_i - 0; \tau, \xi)/\partial \tau$.

For the right limits in $t = t_i$, $i = 1, 2, \cdots, q - 1$, we remark that for $t_i < t < t_{i+1}$, $i = 1, 2, \cdots, q - 1$, the following relations hold:

$$\Phi(t; \tau, \xi) = \psi_{(i+1)}(t; t_{i+1}, x_{i+1}), \tag{3.15}$$

$$\frac{\partial \Phi}{\partial \xi}(t; \tau, \xi) = \frac{\partial \psi_{(i+1)}}{\partial \alpha}(t; t_i, x_i)\frac{\partial t_i}{\partial \xi}(\tau, \xi) + \frac{\partial \psi_{(i+1)}}{\partial \beta}(t; t_i, x_i)\frac{\partial x_i}{\partial \xi}(\tau, \xi), \tag{3.16}$$

and similarly for the derivative with respect to τ.

For $t \to t_i + 0$,

$$\frac{\partial \Phi}{\partial \xi}(t_i + 0; \tau, \xi) = \frac{\partial \psi_{(i+1)}}{\partial \alpha}(t_i; t_i, x_i)\frac{\partial t_i}{\partial \xi}(\tau, \xi) + \frac{\partial \psi_{(i+1)}}{\partial \beta}(t_i; t_i, x_i)\frac{\partial x_i}{\partial \xi}(\tau, \xi). \tag{3.17}$$

Since $\psi_{(i+1)}(t; t_i, x_i)$ is the solution of the system $(3.3)_{(i+1)}$, its derivatives

$$\frac{\partial \psi_{(i+1)}}{\partial \beta}(t; \alpha, \beta), \quad \frac{\partial \psi_{(i+1)}}{\partial \alpha}(t; \alpha, \beta)$$

are the solutions of the equations of variation with the initial conditions:

$$\frac{\partial \psi_{(i+1)}}{\partial \beta}(t_i; t_i, x_i) = E, \quad \frac{\partial \psi_{(i+1)}}{\partial \alpha}(t_i; t_i, x_i) = -f(t_i, x_i, v_{(i+1)}(t_i, x_i)),$$

respectively.

From (3.17) we have $(3.5)_{(i)}$ and $(3.5')_{(i)}$, and the lemma is completely proved.

We let

$$(3.18) \qquad P_i - 0 = (t_i, x_i, v_{(i)}(t_i, x_i)), \quad P_i + 0 = (t_i, x_i, v_{(i+1)}(t_i, x_i)).$$

The relations (3.5) and $(3.5')_{(i)}$ may now be written:

$$(3.19) \qquad \frac{\partial \Phi}{\partial \xi}(t_i - 0; \tau, \xi) = -f(P_i - 0)\frac{\partial t_i}{\partial \xi}(\tau, \xi) + \frac{\partial x_i}{\partial \xi}(\tau, \xi),$$

$$(3.20) \qquad \frac{\partial \Phi}{\partial \xi}(t_i + 0; \tau, \xi) = -f(P_i + 0)\frac{\partial t_i}{\partial \xi}(\tau, \xi) + \frac{\partial x_i}{\partial \xi}(\tau, \xi),$$

with analogous formulas for

$$\frac{\partial \Phi}{\partial \tau}(t_i - 0; \tau, \xi), \quad \frac{\partial \Phi}{\partial \tau}(t_i + 0; \tau, \xi).$$

Remark 3.1. An admissible synthesis solves the controllability problem for the system (3.2).

Remark 3.2. The examples studied in [4] satisfy the properties stated in this section if we consider them in (t, x)-space.

Remark 3.3. The results of this section hold if the functions $f(t, x, u)$ and $f^0(t, x, u)$ are not of class $C^{(1)}$ in $\mathscr{C}$, but only of class $C^{(1)}$ on every cell, and may be extended to some functions of class $C^{(1)}$ on a neighborhood of the cell. In this case, the systems $(3.3)_{(i)}$ would be replaced by the systems

$$\frac{dx}{dt} = f_{(i)}(t, x, v_{(i)}(t, x)), \quad x(\alpha) = \beta.$$

4. Dual variables for marked trajectories. Since some of the properties we proved, or will prove, hold only for (τ, ξ) in cells of maximum dimension $n + 1$, we let $M = P^n \cup N$. Then $\mathscr{G} - M$ will be the union of all $(n + 1)$-dimensional cells. If $\lambda = (\lambda^1, \lambda^2, \cdots, \lambda^n)$, we set

$$(4.1) \quad \mathscr{H}(t, x, u, \lambda) = f^0(t, x, u) + \lambda f(t, x, u) = f^0(t, x, u) + \sum_{j=1}^{n} \lambda^j f^j(t, x, u),$$

$$(4.2) \qquad H(t, x, \lambda) = \mathscr{H}(t, x, v(t, x), \lambda)$$

for all (t, x) in $\tilde{c}_i$, where c_i is a cell of type I and $\tilde{c}_i$ is the neighborhood where the extension $v_{(i)}$ is defined. Let

$$(4.3) \qquad H_{(i)}(t, x, \lambda) = \mathscr{H}(t, x, v_i(t, x), \lambda).$$

For $(t, x) \in \tilde{c}_i$, the derivative

$$\frac{\partial H_{(i)}}{\partial x}(t, x, \lambda) = \frac{\partial \mathscr{H}}{\partial x}(t, x, v_{(i)}(t, x), \lambda) + \frac{\partial \mathscr{H}}{\partial u}(t, x, v_{(i)}(t, x), \lambda)\frac{\partial v_{(i)}}{\partial x}(t, x)$$

exists and we denote, for $(t, x) \in c_i$,

$$(4.4) \qquad \frac{\partial H}{\partial x}(t, x, \lambda) = \frac{\partial H_{(i)}}{\partial x}(t, x, \lambda).$$

For arbitrary $\lambda_{(i)}^-, \lambda_{(i)}^+$, we denote:

$$\Pi_i^- = (t_i, x_i, v_i(t_i, x_i), \lambda_{(i)}^-), \quad \Pi_i^+ = (t_i, x_i, v_{(i+1)}(t_i, x_i), \lambda_i^+). \tag{4.5}$$

Then with the notations (3.18),

$$\begin{aligned} \mathscr{H}(\Pi_i^-) &= f^0(P_i - 0) + \lambda_{(i)}^- f(P_i - 0), \\ \mathscr{H}(\Pi_i^+) &= f^0(P_i + 0) + \lambda_{(i)}^+ f(P_i + 0). \end{aligned} \tag{4.6}$$

In this section we shall study the "adjoint" systems

$$\frac{d\lambda}{dt} = -\frac{\partial H_{(i)}}{\partial x}(t, x, \lambda) \tag*{(4.7)$_{(i)}$}$$

or

$$\frac{d\lambda}{dt} = -\frac{\partial \mathscr{H}}{\partial x}(t, x, v_{(i)}(t, x), \lambda) - \frac{\partial \mathscr{H}}{\partial u}(t, x, v_{(i)}(t, x), \lambda)\frac{\partial v_{(i)}}{\partial x}(t, x). \tag*{(4.7')$_{(i)}$}$$

For $(t, x) \in \mathscr{G} - M$, the system $(3.3)_{(i)}$ coincides with the system (3.2) and the system $(4.7)_{(i)}$ coincides with the linear system that is adjoint to the equations of variations of (3.2).

For $(t, x) \in M$ we shall use $(3.7)_{(i)}$ without writing subscripts (i) for the function $H_{(i)}(t, x, \lambda)$.

LEMMA 4.1. *For $(\tau, \xi) \in c_1 \subset \mathscr{G} - M$, there exists a nonzero vector-valued function $\lambda(t; \tau, \xi) = (\lambda^1(t; \tau, \xi), \cdots, \lambda^n(t; \tau, \xi))$ defined on the interval $[\tau, t_F]$ so that on every interval (t_{i-1}, t_i), it is a solution of the system*

$$\frac{d\lambda}{dt} = -\frac{\partial H}{\partial x}(t, \Phi(t; \tau, \xi), \lambda), \quad \lambda(t_i) = \lambda_{(i)}^-, \qquad i = 1, 2, \cdots, q, \tag*{(4.8)$_{(i)}$}$$

where the $\lambda_{(i)}^-$ are given by the relations:

$$\begin{aligned} &\frac{\partial g}{\partial t}(T(\theta_0), \chi(\theta_0))\frac{\partial T}{\partial \theta^j}(\theta_0) + \frac{\partial g}{\partial x}(T(\theta_0), \chi(\theta_0))\frac{\partial \chi}{\partial \theta^j}(\theta_0) \\ &\quad + f^0(P_F)\frac{\partial T}{\partial \theta^j}(\theta_0) + \lambda_F\left[f(P_F)\frac{\partial T}{\partial \theta^j}(\theta_0) - \frac{\partial \chi}{\partial \theta^j}(\theta_0)\right] = 0, \qquad j = 1, 2, \cdots, k, \end{aligned} \tag{4.9}$$

$$\begin{aligned} \mathscr{H}(\Pi_i^-)\frac{\partial T_{(i)}}{\partial \theta^j}(\theta^0) - \lambda_{(i)}^- \frac{\partial \chi_{(i)}}{\partial \theta^j}(\theta_0) &= \mathscr{H}(\Pi_i^+)\frac{\partial T_{(i)}}{\partial \theta^j}(\theta_0) - \lambda_{(i)}^+ \frac{\partial \chi_{(i)}}{\partial \theta^j}(\theta_0), \\ & \qquad j = 1, 2, \cdots, m_i, \end{aligned} \tag{4.10}$$

where $\lambda_{(q)}^- = \lambda_F$, $P_q - 0 = P_F$, $\lambda_{(i)}^+ = \lambda(t_i + 0; \tau, \xi)$, $i = 1, 2, \cdots, q - 1$.

Proof. The algebraic linear system (4.9) has k equations for $\lambda_F^1, \cdots, \lambda_F^n$ (n unknowns). The coefficient matrix

$$N_F = N_{(q)} = \left(f(P_F)\frac{\partial T}{\partial \theta}(\theta_0) - \frac{\partial \chi}{\partial \theta}(\theta_0) \right) = \left(\frac{d\Phi}{dt}(t_F; \tau, \xi)\frac{\partial T}{\partial \theta}(\theta_0) - \frac{\partial \chi}{\partial \theta}(\theta_0) \right) \tag{4.11}$$

has rank k according to Lemma 3.1. Then the system (4.9) has infinitely many solutions which depend, generally, on $n - k$ parameters. When (τ, ξ) ranges over a neighborhood in c_1, $\theta = \theta(\tau, \xi)$ is a function of class $C^{(1)}$. Since all functions in (4.9) are continuous in (τ, ξ), we can choose $\lambda_F = \lambda_F(\tau, \xi)$ as a continuous function of $(\tau, \xi) \in c_1$.

Then the system of linear differential equations

$$\frac{d\lambda}{dt} = -\frac{\partial H}{\partial x}(t, \Phi(t; \tau, \xi), \lambda), \quad \lambda(t_F) = \lambda_F,$$

which we can write

$$\begin{aligned}\frac{d\lambda}{dt} = &-\frac{\partial \mathscr{H}}{\partial x}(t, \Phi(t; \tau, \xi), v_{(q)}(t, \Phi(t; \tau, \xi)), \lambda) \\ &-\frac{\partial \mathscr{H}}{\partial u}(t, \Phi(t; \tau, \xi), v_{(q)}(t, \Phi(t; \tau, \xi), \lambda)\frac{\partial v_{(q)}}{\partial x}(t, \Phi(t; \tau, \xi)))\end{aligned}$$

has the solution $\lambda = \lambda(t; t_F, \lambda_F)$ with $\lambda(t_F; t_F, \lambda_F) = \lambda_F$ defined on the interval $(t_{q-1}, t_F]$. We set

$$\lambda(t; \tau, \xi) = \lambda(t; t_F(\tau, \xi), \lambda_F(\tau, \xi)).$$

If $\lambda^+_{(q-1)} = \lambda(t_{q-1} + 0; \tau, \xi)$, then the vector $\lambda^-_{(q-1)}$ is the solution of the linear algebraic system

$$\begin{aligned}&f^0(P_{q-1} - 0)\frac{\partial T_{(q-1)}}{\partial \theta^j}(\theta_0) + \lambda^-_{(q-1)}\left[f(P_{q-1} - 0)\frac{\partial T_{(q-1)}}{\partial \theta^j}(\theta_0) - \frac{\partial \chi_{(q-1)}}{\partial \theta^j}(\theta_0)\right] \\ (4.12)\quad &= f^0(P_{q-1} + 0)\frac{\partial T_{(q-1)}}{\partial \theta^j}(\theta_0) + \lambda^+_{(q-1)}\left[f(P_{q-1} + 0)\frac{\partial T_{(q-1)}}{\partial \theta^j}(\theta_0) - \frac{\partial \chi_{(q-1)}}{\partial \theta^j}(\theta_0)\right], \\ &\qquad\qquad\qquad\qquad\qquad\qquad j = 1, 2, \cdots, m_{q-1}.\end{aligned}$$

The coefficient matrix

$$N_{(q-1)} = \left(f(P_{(q-1)} - 0)\frac{\partial T_{(q-1)}}{\partial \theta}(\theta_0) - \frac{\partial \chi_{(q-1)}}{\partial \theta}(\theta_0)\right)$$

has rank equal to the dimension m_{q-1} of $\Pi(c_{q-1})$. Therefore, the solution $\lambda^-_{(q-1)}$ of the system (4.12) depends generally on $n - m_{(q-1)}$ arbitrary constants.

The linear system

$$\frac{d\lambda}{dt} = -\frac{\partial H}{\partial x}(t, \Phi(t; \tau, \xi), \lambda), \quad \lambda(t_{q-1}) = \lambda^-_{(q-1)},$$

has the solution

$$\lambda(t; \tau, \xi) = \lambda(t; t_{q-1}(\tau, \xi), \lambda^-_{(q-1)}(\tau, \xi))$$

defined on (t_{q-2}, t_{q-1}). Continuing in the same way, we obtain the function $\lambda(t; \tau, \xi)$ defined on the whole interval $[\tau, t_F]$ with the properties from the statement.

Remark 4.1. If $k = n$ as in the case considered in [2], $\lambda(t; \tau, \xi)$ is uniquely determined.

5. The value of the functional for marked trajectories. According to § 2 and § 3, the value of the functional for the marked trajectory $x = \Phi(t; \tau, \xi)$ is

$$(5.1)\qquad \begin{aligned} W(\tau, \xi) &= g(t_F(\tau, \xi), x_F(\tau, \xi)) \\ &\quad + \sum_{i=1}^{q} \int_{t_{i-1}(\tau,\xi)}^{t_i(\tau,\xi)} f^0(t, \Phi(t; \tau, \xi), v(t, \Phi(t; \tau, \xi)))\, dt, \end{aligned}$$

where

$$t_0(\tau, \xi) = \tau, \quad t_q(\tau, \xi) = t_F(\tau, \xi).$$

We can write this as

$$(5.2)\qquad \begin{aligned} W(\tau, \xi) &= g(t_F(\tau, \xi), x_F(\tau, \xi)) \\ &\quad + \sum_{i=1}^{q} \int_{t_{i-1}(\tau,\xi)}^{t_i(\tau,\xi)} f^0(t, \Phi(t; \tau, \xi), v_{(i)}(t, \Phi(t; \tau, \xi)))\, dt, \end{aligned}$$

since, for $t_{i-1} \leqq t < t_i$, $v_{(i)}(t, \Phi(t; \tau, \xi)) = v(t, \Phi(t; \tau, \xi))$.

It is obvious, from Lemma 3.1, that $W(\tau, \xi)$ is a function of class $C^{(1)}$ in (τ, ξ) on $\mathscr{G} - M$.

We shall now prove the following lemma.

LEMMA 5.1. *For any* $(\tau, \xi) \in \mathscr{G} - M$, *we have*

$$(5.3)\qquad \frac{\partial W}{\partial \xi}(\tau, \xi) = \lambda(\tau; \tau, \xi),$$

$$(5.4)\qquad \frac{\partial W}{\partial \tau}(\tau, \xi) = -f^0(\tau, \xi, v(t, \xi)) - \lambda(\tau; \tau, \xi) f(\tau, \xi, v(\tau, \xi)).$$

Proof. We set

$$f^j_{(i)} = f^j(t, \Phi(t; \tau, \xi), v_{(i)}(t, \Phi(t; \tau, \xi))), \qquad j = 0, 1, 2, \cdots, n,$$

$$\bar{v}_{(i)} = v_{(i)}(t, \Phi(t; \tau, \xi)).$$

From (5.2) we obtain:

$$(5.5)\qquad \begin{aligned} \frac{\partial W}{\partial \xi}(\tau, \xi) &= \frac{\partial g}{\partial t}(t_F, x_F)\frac{\partial t_F}{\partial \xi}(\tau, \xi) + \frac{\partial g}{\partial x}(t_F, x_F)\frac{\partial x_F}{\partial \xi}(\tau, \xi) + \sum_{i=1}^{q} f^0(P_i - 0)\frac{\partial t_i}{\partial \xi}(\tau, \xi) \\ &\quad - \sum_{i=1}^{q-1} f^0(P_i + 0)\frac{\partial t_i}{\partial \xi}(\tau, \xi) \\ &\quad + \sum_{i=1}^{q} \int_{t_{i-1}}^{t_i} \left(\frac{\partial f^0_{(i)}}{\partial x} + \frac{\partial f^0_{(i)}}{\partial u}\frac{\partial \bar{v}_{(i)}}{\partial x}\right)\frac{\partial \Phi}{\partial \xi}(t; \tau, \xi)\, dt. \end{aligned}$$

From systems $(4.8)_{(i)}$, we obtain

$$(5.6)_{(i)} \qquad \begin{aligned} \frac{d\lambda}{dt}\frac{\partial\Phi}{\partial\xi}(t;\tau,\xi) = -\Bigg[&\left(\frac{\partial f^0_{(i)}}{\partial x} + \frac{\partial f^0_{(i)}}{\partial u}\cdot\frac{\partial \bar{v}_{(i)}}{\partial x}\right) \\ &+ \lambda\left(\frac{\partial f_{(i)}}{\partial x} + \frac{\partial f}{\partial u}\cdot\frac{\partial \bar{v}_{(i)}}{\partial x}\right)\Bigg]\frac{\partial\Phi}{\partial\xi}(t;\tau,\xi). \end{aligned}$$

On the other hand, according to Lemma 3.1,

$$\begin{aligned} \left[\frac{\partial f_{(i)}}{\partial x} + \frac{\partial f_{(i)}}{\partial u}\cdot\frac{\partial \bar{v}_{(i)}}{\partial x}\right]\frac{\partial\Phi}{\partial\xi}(t;\tau,\xi) &= \frac{\partial}{\partial\xi}[f(t,\Phi(t;\tau,\xi)v_{(i)}(t,\Phi(t;\tau,\xi)))] \\ &= \frac{\partial}{\partial\xi}\left(\frac{d\Phi}{dt}(t;\tau,\xi)\right) = \frac{d}{dt}\left(\frac{\partial\Phi}{\partial\xi}(t;\tau,\xi)\right). \end{aligned}$$

Therefore,

$$\begin{aligned} &\left(\frac{\partial f^0_{(i)}}{\partial x} + \frac{\partial f^0_{(i)}}{\partial u}\frac{\partial \bar{v}_{(i)}}{\partial x}\right)\frac{\partial\Phi}{\partial\xi}(t;\tau,\xi) \\ &\quad = -\frac{d\lambda}{dt}\frac{\partial\Phi}{\partial\xi}(t;\tau,\xi) - \lambda\frac{d}{dt}\left(\frac{\partial\Phi}{\partial\xi}(t;\tau,\xi)\right) \\ &\quad = -\frac{d}{dt}\left(\lambda\frac{\partial\Phi}{\partial\xi}(t;\tau,\xi)\right). \end{aligned}$$

We now have

$$\begin{aligned} &\int_{t_{i-1}}^{t_i}\left(\frac{\partial f^0_{(i)}}{\partial x} + \frac{\partial f^0_{(i)}}{\partial u}\frac{\partial \bar{v}_{(i)}}{\partial x}\right)\frac{\partial\Phi}{\partial\xi}(t;\tau,\xi)\,dt \\ &\quad = \lambda^+_{(i-1)}\frac{\partial\Phi}{\partial\xi}(t_i+0;\tau,\xi) - \lambda^-_{(i)}\frac{\partial\Phi}{\partial\xi}(t_i-0;\tau,\xi) \\ &\quad = \lambda^-_{(i)}f(P_i-0)\frac{\partial t_i}{\partial\xi}(\tau,\xi) \\ &\qquad - \lambda^+_{(i-1)}f(P_i+0)\frac{\partial t_{i-1}}{\partial\xi}(\tau,\xi) \\ &\qquad - \lambda^-_{(i)}\frac{\partial x_i}{\partial\xi}(\tau,\xi) + \lambda^+_{(i-1)}\frac{\partial x_{i-1}}{\partial\xi}(\tau,\xi) \end{aligned}$$

for $i = 2, 3, \cdots, q$, and

$$\int_{\tau}^{t_1}\left(\frac{\partial f^0_{(i)}}{\partial x} + \frac{\partial f^0_{(i)}}{\partial u}\frac{\partial \bar{v}_{(i)}}{\partial x}\right)\frac{\partial\Phi}{\partial\xi}(t;\tau,\xi)\,dt = \lambda(\tau;\tau,\xi)\frac{\partial\Phi}{\partial\xi}(\tau;\tau,\xi) - \lambda^-_{(1)}\frac{\partial\Phi}{\partial\xi}(t_1-0;\tau,\xi)$$

for $i = 1$.

From (5.5), we have

$$\begin{aligned}\frac{\partial W}{\partial \xi}(\tau, \xi) &= \frac{\partial g}{\partial t}(t_F, x_F)\frac{\partial t_F}{\partial \xi}(\tau, \xi) + \frac{\partial g}{\partial x}(t_F, x_F)\frac{\partial x_F}{\partial \xi}(\tau, \xi) \\ &\quad + \mathscr{H}(\Pi_F)\frac{\partial t_F}{\partial \xi}(\tau, \xi) - \lambda_F\frac{\partial x_F}{\partial \xi}(\tau, \xi) + \sum_{i=1}^{q-1} \mathscr{H}(\Pi_i)\frac{\partial t_i}{\partial \xi}(\tau, \xi) \\ &\quad - \sum_{i=1}^{q-1} \lambda_{(i)}^{-}\frac{\partial x_i}{\partial \xi}(\tau, \xi) - \sum_{i=1}^{q-1} \mathscr{H}(\Pi_i^{+})\frac{\partial t_i}{\partial \xi}(\tau, \xi) \\ &\quad + \sum_{i=1}^{q-1} \lambda_{(i)}^{+}\frac{\partial x_i}{\partial \xi}(\tau, \xi) + \lambda(\tau; \tau, \xi)\frac{\partial \Phi}{\partial \xi}(\tau; \tau, \xi).\end{aligned}$$

From (4.9), $(4.10)_{(i)}$, and

$$\frac{\partial T_{(i)}}{\partial \theta}(\theta_0)\frac{\partial \theta}{\partial \xi}(\tau, \xi) = \frac{\partial t_i}{\partial \xi}(\tau, \xi), \quad \frac{\partial \chi_{(i)}}{\partial \theta}(\theta_0)\frac{\partial \theta}{\partial \xi}(\tau, \xi) = \frac{\partial x_i}{\partial \xi}(\tau, \xi),$$

we obtain

$$\begin{aligned}&\frac{\partial g}{\partial t}(t_F, x_F)\frac{\partial t_F}{\partial \xi}(\tau, \xi) + \frac{\partial g}{\partial x}(t_F, x_F)\frac{\partial x_F}{\partial \xi}(\tau, \xi) \\ &\quad + \mathscr{H}(\Pi_F)\frac{\partial t_F}{\partial \xi}(\tau, \xi) - \lambda_F\frac{\partial x_F}{\partial \xi}(\tau, \xi) = 0\end{aligned}$$

and

$$\mathscr{H}(\Pi_i^{-})\frac{\partial t_i}{\partial \xi}(\tau, \xi) - \lambda_{(i)}^{-}\frac{\partial x_i}{\partial \xi}(\tau, \xi) = \mathscr{H}(\Pi_i^{+})\frac{\partial t_i}{\partial \xi}(\tau, \xi) - \lambda_{(i)}^{+}\frac{\partial x_i}{\partial \xi}(\tau, \xi),$$

$$i = 1, 2, \cdots, q - 1.$$

Since

$$\frac{\partial \Phi}{\partial \xi}(\tau; \tau, \xi) = E,$$

it follows that

$$\frac{\partial W}{\partial \xi}(\tau, \xi) = \lambda(\tau; \tau, \xi).$$

Formula (5.4) is obtained in the same way.

Remark 5.1. We note that at every stage we have a great liberty in choosing the conditions that define $\lambda(t; \tau, \xi)$. Nevertheless, for (τ, ξ) in $\mathscr{G} - M$, the value $\lambda(\tau; \tau, \xi)$ is uniquely determined. This fact is in some sense "dual" to the property of the trajectories $\Phi(t; \tau, \xi)$ which start from different points in $\mathscr{G} - M$ and reach the same terminal point in $\mathscr{T}$.

6. Necessary and sufficient conditions for optimality in the form of the dynamic programming equation. Lemmas 6.1–6.4 can be proved as in [4] or [5].

LEMMA 6.1. (i) *Let $V(t, x)$ be a real-valued function defined and continuous in $\mathscr{G}$, satisfying the following: $V(t, x) = g(t, x)$ for $(t, x) \in \mathscr{T}$; V is of class $C^{(1)}$ in $\mathscr{G} - M$, where M is an arbitrary set and such that, for $(t, x) \in \mathscr{G} - M$, the inequality*

$$\frac{\partial V}{\partial t}(t, x) + \frac{\partial V}{\partial x}(t, x) f(t, x, u) + f^0(t, x, u) \geqq 0 \tag{6.1}$$

holds for all $u \in U$.

(ii) *Let $u(t)$ be an admissible control corresponding to $(\tau, \xi) \in \mathscr{G}$ such that the corresponding trajectory $x = \varphi(t; \tau, \xi)$ intersets M in a finite number of points.*

Then

$$P(\tau, \xi, u) \geqq V(\tau, \xi).$$

LEMMA 6.2. *Let condition* (i) *of Lemma* 6.1 *hold, and let $u(t)$ be an admissible control corresponding to (τ, ξ) such that the trajectory $x = \varphi(t; \tau, \xi)$ reaches $\mathscr{T}$ in $(t_1, \varphi(t_1; \tau, \xi))$.*

Then in every neighborhood of (τ, ξ), there exists a point (t_0, x_0) such that the solution of the system

$$\frac{dx}{dt} = f(t, x, u(t)), \quad x(t_0) = x_0$$

remains in $\mathscr{G}$, intersects M in a finite number of points, is defined on $[t_0, t_1]$ and is such that

$$P(\tau, \xi, u) \geqq V(\tau, \xi).$$

LEMMA 6.3. *If M is a piecewise smooth set of dimension $m \leqq n$ in $\mathscr{G}$ and $u(t)$ is an admissible control corresponding to (τ, ξ), then the second condition of Lemma* 6.2 *holds.*

LEMMA 6.4. *If M is a piecewise smooth set in $\mathscr{G}$ of dimension $m \leqq n$, $V(t, x)$ satisfies* (i) *of Lemma* 6.1 *and $u(t)$ is an admissible control corresponding to $(\tau, \xi) \in \mathscr{G}$, then*

$$P(\tau, \xi, u) \geqq V(\tau, \xi).$$

We can now prove the following theorem.

THEOREM 6.1. *If the value $W(\tau, \xi)$ of the functional corresponding to marked trajectories satisfies the inequality*

$$\frac{\partial W}{\partial \tau}(\tau, \xi) + \frac{\partial W}{\partial \xi}(\tau, \xi) f(\tau, \xi, u) + f^0(\tau, \xi, u) \geqq 0$$

for all $u \in U$ and for every $(t, x) \in \mathscr{G} - M$, then the marked trajectories are optimal.

Proof. The set $M = P^n \cup N$ is (by hypothesis) an n-dimensional piecewise smooth set and $W(\tau, \xi)$ satisfies the condition (i) from Lemma 6.1. Applying Lemma 6.4, it is obvious that for any admissible control $u(t)$ related to $(\tau, \xi) \in \mathscr{G}$ we have $P(\tau, \xi, u) \geqq W(\tau, \xi)$. Since for the marked trajectories, $P(\tau, \xi, \bar{u}) = W(\tau, \xi)$, the theorem is proved.

THEOREM 6.2. *If the marked trajectories are optimal, then for every* $(\tau, \xi) \in \mathscr{G} - M$ *and for every* $u \in U$, *we have*

$$\frac{\partial W}{\partial \tau}(\tau, \xi) + \frac{\partial W}{\partial \xi}(\tau, \xi) f(\tau, \xi, u) + f^0(\tau, \xi, u) \geqq 0. \tag{6.2}$$

Proof. The proof will be the same as that given in [2] for the corresponding result.

Let $(\tau, \xi) \in \mathscr{G} - M$ and let $u_0 \in U$ be arbitrary. From the existence theorem it follows that there exist a number $r > 0$ and the unique solution $x = \psi(t)$ of the system

$$\frac{dx}{dt} = f(t, x, u_0), \quad x(\tau) = \xi, \tag{6.3}$$

defined on $|t - \tau| < r$ with $\psi(\tau) = \xi$.

Since M contains no interior points, $\mathscr{G} - M$ is open, and there exists a neighborhood V_0 of (τ, ξ) included in $\mathscr{G} - M$. Let $V_0\{(t, x)| \, |(t, x) - (\tau, \xi)| < \varepsilon < r\}$. On the other hand, $\psi(t)$ is continuous at $t = \tau$. Hence for $\varepsilon > 0$, there exists a $\delta > 0$ such that $|\psi(t) - \psi(\tau)| < \varepsilon$ for $|t - \tau| < \delta$.

Let $\delta < \varepsilon$. Then for $\tau \leqq t \leqq \tau + \delta$ we have $(t, \psi(t)) \in V_0$, with $V_0 \subset \mathscr{G} - M$. For every admissible control $u(t)$ corresponding to $(\tau + \delta, \psi(\tau, \delta))$, we have

$$W(\tau + \delta, \psi(\tau + \delta)) \leqq P(\tau + \delta, \psi(\tau + \delta), u) \tag{6.4}$$

since the marked trajectories are optimal.

On the other hand, the function

$$\hat{u}(t) = \begin{cases} u_0 & \text{if } \tau \leqq t \leqq \tau + \delta, \\ u(t) & \text{if } t > \tau + \delta \end{cases}$$

is an admissible control related to (τ, ξ) according to Definition 2.1. It follows that

$$W(\tau, \xi) \leqq P(\tau, \xi, \hat{u}), \tag{6.5}$$

where

$$P(\tau, \xi, \hat{u}) = g(t_1, x_1) + \int_{\tau}^{\tau+\delta} f^0(t, \psi(t), u_0)\, dt$$

$$+ \int_{\tau+\delta}^{t_1} f^0(t, \varphi(t), u(t))\, dt.$$

If instead of $x = \psi(t)$ we take the marked trajectory starting at $(\tau + \delta, \psi(\tau + \delta))$ (that is, instead of $u(t)$ we take $\bar{u}(t) = v(t, \Phi(t; \tau + \delta, \psi(\tau + \delta))$), we have:

$$W(\tau + \delta, \psi(\tau + \delta)) = g(t_F, x_F) + \int_{\tau+\delta}^{t_F} f^0(t, \Phi; \tau + \delta, \psi(\tau + \delta)), \bar{u}(t))\, dt.$$

Hence (6.5) becomes

$$W(\tau, \xi) \leqq W(\tau + \delta, \psi(\tau + \delta)) + \int_{\tau}^{\tau+\delta} f^0(t, \psi(t), u_0)\, dt \tag{6.6}$$

or

$$W(\tau + \delta, \psi(\tau + \delta)) - W(\tau, \xi) + \int_{\tau}^{\tau+\delta} f^0(t, \psi(t), u_0)\, dt \geqq 0.$$

Multiplying by $1/\delta$, $\delta > 0$, and using

$$\begin{aligned} \lim_{\delta \to 0} \frac{W(\tau + \delta, \psi(\tau + \delta)) - W(\tau, \xi)}{\delta} &= \frac{\partial W}{\partial \tau}(\tau, \xi) + \frac{\partial W}{\partial \xi}(\tau, \xi)\frac{d\psi}{dt}(\tau) \\ &= \frac{\partial W}{\partial \tau}(\tau, \xi) + \frac{\partial W}{\partial \xi}(\tau, \xi) f(\tau, \xi, u_0) \end{aligned}$$

and

$$\lim_{\delta \to 0} \int_{\tau}^{\tau+\delta} f^0(t, \psi(t), u_0)\, dt = f^0(\tau, \psi(\tau), u_0) = f^0(\tau, \xi, u_0),$$

we obtain the inequality in the statement.

Remark 6.1. Theorems 6.1 and 6.2 show that a necessary and sufficient condition for an admissible synthesis to be optimal is that the corresponding $W(\tau, \xi)$ satisfy the dynamic programming equation

$$\frac{\partial W}{\partial \tau}(\tau, \xi) + \min_{u \in U} \left[\frac{\partial W}{\partial \xi}(\tau, \xi) f(\tau, \xi, u) + f^0(\tau, \xi, u) \right] = 0$$

for every $(\tau, \xi) \in \mathscr{G} - M$.

7. Necessary and sufficient conditions for optimality in the form of the maximum principle.

THEOREM 7.1. *If for every $(\tau, \xi) \in \mathscr{G}$ and for every $u \in U$, the following inequality holds*:

$$\mathscr{H}(\tau, \xi, v(\tau, \xi), \lambda(\tau; \tau, \xi)) \leqq \mathscr{H}(\tau, \xi, u, \lambda(\tau; \tau, \xi)), \tag{7.1}$$

then the marked trajectories are optimal.

Proof. Inequality (7.1) may be written

$$\begin{aligned} &f^0(\tau, \xi, v(\tau, \xi)) + \lambda(\tau; \tau, \xi) f(\tau, \xi, v(\tau, \xi)) \\ &\leqq f^0(\tau, \xi, u) + \lambda(\tau; \tau, \xi) f(\tau, \xi, u). \end{aligned} \tag{7.2}$$

Using (5.3) and (5.4) we obtain, for $(\tau, \xi) \in \mathscr{G} - M$,

$$-\frac{\partial W}{\partial \tau}(\tau, \xi) \leqq f^0(\tau, \xi, u) + \frac{\partial W}{\partial \xi}(\tau, \xi) f(\tau, \xi, u)$$

or

$$\frac{\partial W}{\partial \tau}(\tau, \xi) + \frac{\partial E}{\partial \xi}(\tau, \xi) f(\tau, \xi, u) + f^0(\tau, \xi, u) \geqq 0$$

for every $u \in U$.

The conditions of Theorem 6.1 are verified; hence the marked trajectories are optimal.

Remark 7.1. It is obvious that for the optimality of the marked trajectories it is sufficient that (7.1) hold only for $(\tau, \xi) \in \mathscr{G} - M$ and not for every $(\tau, \xi) \in \mathscr{G}$.

THEOREM 7.2. *If the marked trajectories are optimal, then for every* $(\tau, \xi) \in \mathscr{G} - M$, *the inequality* (7.1) *holds.*

Proof. Indeed, in this case, Theorem 6.2 states that $W(\tau, \xi)$ satisfies the inequality

$$\frac{\partial W}{\partial \tau}(\tau, \xi) + \frac{\partial W}{\partial \xi}(\tau, \xi) f(\tau, \xi, u) + f^0(\tau, \xi, u) \geqq 0$$

for every u in U and every (τ, ξ) in $\mathscr{G} - M$. Using (5.3) and (5.4) we obtain (7.1).

Remark 7.2. Theorems 7.1 and 7.2 show that a necessary and sufficient condition for an admissible synthesis to be optimal is that the function $\mathscr{H}(t, x, u, \lambda)$ satisfy the following:

$$\min_{u \in U} \mathscr{H}(\tau, \xi, u, \lambda(\tau; \tau, \xi)) = \mathscr{H}(\tau, \xi, v(\tau, \xi), \lambda(\tau; \tau, \xi)) \tag{7.3}$$

for every $(\tau, \xi) \in \mathscr{G} - M$.

To show that this relation together with (3.2), (4.8)–(4.10) represents a special form of the maximum principle and transversality conditions, we note the following:

(a) If $\tau \leqq t < t_F$ is such that $(t, \Phi(t; \tau, \xi)) \in \mathscr{G} - M$ and if $x = \Phi(t; \tau, \xi)$, then we have $\lambda(t; t, x) = \lambda(t; \tau, \xi)$ (see Remark 5.1) and

$$\frac{\partial W}{\partial x}(t, x) = \lambda(t; t, x) = \lambda(t; \tau, \xi).$$

Hence, for such points (t, x), (7.3) may be written

$$\min_{u \in U} \mathscr{H}(t, \Phi(t; \tau, \xi), u, \lambda(t; \tau, \xi)) = \mathscr{H}(t, \Phi(t; \tau, \xi), v(t, \Phi(t; \tau, \xi))). \tag{7.4}$$

(b) See [10]. If (7.1) (and hence 7.4)) is fulfilled and U has a piecewise smooth boundary, then on every interval (t_{i-1}, t_i) such that $c_i \subset \mathscr{G} - M$ the system (4.8) becomes

$$\frac{d\lambda}{dt} = -\frac{\partial \mathscr{H}}{\partial x}(t; \Phi(t_i, \tau, \xi), v(t, \Phi(t; \tau, \xi)), \lambda) \tag{7.5}$$

since

$$\frac{\partial \mathscr{H}}{\partial u}(t, \Phi(t; \tau, \xi), v(t, \Phi(t; \tau, \xi)), \lambda(t; \tau, \xi)) \frac{\partial v}{\partial x}(t, \Phi(t; \tau, \xi)) = 0. \tag{7.6}$$

Indeed, if the minimum in (7.4) is obtained for $v(t, \Phi(t; \tau, \xi)) \in \operatorname{int} U$, then $\partial \mathscr{H}(\cdot)/\partial u = 0$. If $v(t, \Phi(t; \tau, \xi))$ lies on a smooth region of the boundary of U, then $\partial \mathscr{H}(\cdot)/\partial u$ is normal to the boundary of U. That is, $\partial \mathscr{H}(\cdot)/\partial u$ is normal to the vectors $\partial v(t, \Phi(t; \tau, \xi))/\partial x^i$, $i = 1, 2, \cdots, n$. Hence

$$\frac{\partial \mathscr{H}}{\partial u}(\cdot)\frac{\partial v}{\partial x^i}(\cdot) = 0, \qquad i = 1, 2, \cdots, n.$$

If $v(t, \Phi(t; \tau, \xi))$ lies in a corner or on an edge of the boundary of U, then (7.6) holds by continuity.

On the intervals (t_{i-1}, t_i) for which $c_i \subset M$, we cannot state (7.4) and (7.5). Hence (3.2), (7.4), (7.5) represent a special form of the maximum principle.

If we let $\lambda^0 = -(f^0(P_F) + \lambda_F f(P_F))$ and suppose $g(t, x) \equiv 0$, then the relations (4.9) are the transversality conditions as they are formulated in [4] and [5] in geometrical language.

8. Admissible synthesis for a class of differential games. We consider the following differential games problem:

Let $\mathscr{G} \subset R \times R^n$, and let $\mathscr{C} \subset R \times R^n \times R^r \times R^s$ be a bounded region such that $\mathscr{G} \subset P^r_{R \times R^n}\mathscr{C}$. Let $Y = P^r_{R^r}\mathscr{C}$ and let $Z = P^r_{R^s}\mathscr{C}$; the sets Y and Z will be, generally, closed regions.

The vector-valued function $f(t, x, y, z)$ and the real-valued function $f^0(t, x, y, z)$ defined on $\mathscr{C}$ are of class $C^{(1)}$ with respect to (x, y, z) and are continuous with respect to t.

A k-dimensional differentiable manifold $\mathscr{T} \subset \overline{\mathscr{G}}$ of class $C^{(1)}$ is given and is called the terminal manifold.

In a neighborhood $\mathscr{G}_1$ of $\mathscr{T}$ there is given a real-valued function $g(t, x)$ of class $C^{(1)}$ which will be called terminal payoff.

The state $x(t)$ of the game is determined by the system

$$\text{(8.1)} \qquad \frac{dx}{dt} = f(t, x, y, z), \quad x(\tau) = \xi, \quad (\tau, \xi) \in \mathscr{G}, \quad y \in Y, \quad z \in Z.$$

In the following we use the terminology proposed in [10].

DEFINITION 8.1. The functions $y(t)$ and $z(t)$ defined and piecewise continuous on $I \subset P^r_R\mathscr{G}$ with ranges in Y and Z, respectively, will be called *admissible strategies with respect to the initial point* (τ, ξ) if the solution $x = \varphi(t; \tau, \xi)$ of the system

$$\text{(8.2)} \qquad \frac{dx}{dt} = f(t, x, y(t), z(t)), \quad x(\tau) = \xi$$

remains in $\mathscr{G}$ and reaches $\mathscr{T}$ in a finite time.

Let $t_1 > \tau$ be a real number such that, if $x_1 = \varphi(t_1; \tau, \xi)$, then $(t_1, x_1) \in \mathscr{T}$ and

$$\{(t, \varphi(t; \tau, \xi)) | \tau \leqq t < t_1\} \subset \mathscr{G} - \mathscr{T}.$$

The trajectory $x = \varphi(t; \tau, \xi)$ will be called an admissible trajectory of the game.

For every admissible strategy (trajectory) we may define the payoff of the game:

$$P(\tau, \xi, y, z) = g(t_1, x_1) + \int_\tau^{t_1} f^0(t, \varphi(t; \tau, \xi), y(t), z(t))\, dt. \tag{8.3}$$

Definition 8.2. The admissible strategy with respect to (τ, ξ), $(\tilde{y}(t), \tilde{z}(t))$ is called *optimal*, if for all admissible strategies $(y(t), \tilde{z}(t))$, $(\tilde{y}(t), z(t))$ (with respect to (τ, ξ)) the following inequalities are satisfied:

$$P(\tau, \xi, \tilde{y}, z) \leqq P(\tau, \xi, \tilde{y}, \tilde{z}) \leqq P(\tau, \xi, y, \tilde{z}). \tag{8.4}$$

Definition 8.3. The piecewise smooth sets $N \subset \mathscr{C}$, $P^k = \mathscr{T} \subset P^{k+1} \subset \cdots \subset P^n \subset P^{n+1} = \mathscr{G}$ and the pair of functions $(y(t, x),\ z(t, x))$ defined on $\mathscr{G}$ with values in $Y \times Z$ represent an admissible synthesis for the defined differential game if the conditions A—D of Definition 2.3 are satisfied when the system (2.4) is replaced by the system

$$\frac{dx}{dt} = f(t, x, y(t, x), z(t, x)). \tag{8.5}$$

The trajectories $x = \Phi(t; \tau, \xi)$ generated by such admissible syntheses are also called marked trajectories.

It is obvious that all statements of § 3–§ 5 hold for this problem (see also [2]).

Concerning the optimality conditions for marked trajectories, we prove the following theorem.

Theorem 8.1. *If $W(\tau, \xi)$ is the payoff of the game corresponding to the marked trajectory $x = \Phi(t; \tau, \xi)$, then the trajectory $x = \Phi(t; \tau, \xi)$ is optimal if and only if*

$$\begin{aligned} &\frac{\partial W}{\partial \tau}(\tau, \xi) + \min_{y \in Y} \max_{z \in Z} \left[\frac{\partial W}{\partial \xi}(\tau, \xi) f(\tau, \xi, y, z) + f^0(\tau, \xi, y, z) \right] \\ &= \frac{\partial W}{\partial \tau}(\tau, \xi) + \max_{z \in Z} \min_{y \in Y} \left[\frac{\partial W}{\partial \xi}(\tau, \xi) f(\tau, \xi, y, z) + f^0(\tau, \xi, y, z) \right] \end{aligned} \tag{8.6}$$

for every $(\tau, \xi) \in \mathscr{G} - M$.

Proof. We let, as before, $\bar{y}(t) = y(t, \Phi(t; \tau, \xi))$, $\bar{z}(t) = z(t, \Phi(t; \tau, \xi))$. For the control problem with differential equations

$$\frac{dx}{dt} = f(t, x, y, z(t, x)), \qquad x(\tau) = \xi, \qquad y \in Y, \tag{8.7}$$

the sets $N, P^n, \cdots, P^{n+1}$ and the function $y(t, x)$ represent an admissible synthesis according to Definition 2.3 since the function $F(t, x, y) = f(t, x, y, z(t, x))$ satisfies the condition of Remark 3.3.

The marked trajectories of this problem are also the functions $\Phi(t; \tau, \xi)$ and the value of the functional is $W(\tau, \xi)$. Moreover, $W(\tau, \xi)$ satisfies the dynamic programming equation:

$$(8.8)\quad \frac{\partial W}{\partial \tau}(\tau, \xi) + \min_{y \in Y}\left[\frac{\partial W}{\partial \xi}(\tau, \xi) f(\tau, \xi, y, z(\tau, \xi)) + f^0(\tau, \xi, y, z(\tau, \xi))\right] = 0.$$

Indeed, (5.3) and (5.4) may now be written in the form

$$(8.9)\quad \frac{\partial W}{\partial \tau}(\tau, \xi) + \frac{\partial W}{\partial \xi}(\tau, \xi) f(\tau, \xi, y(\tau, \xi), z(\tau, \xi)) + f^0(\tau, \xi, y(\tau, \xi), z(\tau, \xi)) = 0.$$

From (8.6) and (8.9), we have

$$\begin{aligned}
&\min_{y \in Y} \max_{z \in Z}\left[\frac{\partial W}{\partial \xi}(\tau, \xi) f(\tau, \xi, y, z) + f^0(\tau, \xi, y, z)\right] \\
&= \max_{z \in Z} \min_{y \in Y}\left[\frac{\partial W}{\partial \xi}(\tau, \xi) f(\tau, \xi, y, z) + f^0(\tau, \xi, y, z)\right] \\
&= \frac{\partial W}{\partial \xi}(\tau, \xi) f(\tau, \xi, y(\tau, \xi), z(\tau, \xi)) + f^0(\tau, \xi, y(\tau, \xi), z(\tau, \xi)) \\
&= -\frac{\partial W}{\partial \tau}(\tau, \xi).
\end{aligned}$$

Hence (see [7, Theorem 1.5]),

$$\begin{aligned}
&\min_{y \in Y}\left[\frac{\partial W}{\partial \xi}(\tau, \xi) f(\tau, \xi, y, z(\tau, \xi)) + f^0(\tau, \xi, y, z(\tau, \xi))\right] \\
(8.10)\qquad &= \max_{z \in Z}\left[\frac{\partial W}{\partial \xi}(\tau, \xi) f(\tau, \xi, y(\tau, \xi), z) + f^0(\tau, \xi, y(\tau, \xi), z)\right] \\
&= -\frac{\partial W}{\partial \tau}(\tau, \xi)
\end{aligned}$$

for $(\tau, \xi) \in \mathscr{G} - M$.

Equation (8.8) now follows from (8.10).

Since the control problem (8.7) satisfies the requirements of Theorem 6.1, it follows that for every pair $(y(t), z(t))$ which is admissible with respect to (τ, ξ), we have

$$(8.11)\qquad P(\tau, \xi, y, \bar{z}) \geqq P(\tau, \xi, \bar{y}, \bar{z}) = W(\tau, \xi).$$

(The marked trajectories $x = \Phi(t; \tau, \xi)$ are optimal for the control system (8.7).)

The same arguments, applied to the control system

$$(8.7')\qquad \frac{dx}{dt} = f(t, x, y(t, x), z), \quad x(\tau) = \xi$$

show that the following inequality holds for every admissible strategy $(y(t), z(t))$:

$$(8.11')\qquad P(\tau, \xi, \bar{y}, z) \leqq W(\tau, \xi).$$

From (8.11) and (8.11′) we have that

$$P(\tau, \xi, y, \bar{z}) \leqq W(\tau, \xi) \leqq P(\tau, \xi, \bar{y}, z).$$

We suppose now that the marked trajectories are optimal (hence, the admissible strategy $(\bar{y}(t), \bar{z}(t))$ is optimal).

In this case, the trajectories $x = \Phi(t; \tau, \xi)$ are optimal for the control problems (8.7) and (8.7′). Applying Theorem 6.2 we obtain for every $(\tau, \xi) \in \mathscr{G} - M$:

$$\frac{\partial W}{\partial \tau}(\tau, \xi) + \max_{z \in Z} \left[\frac{\partial W}{\partial \xi}(\tau, \xi) f(\tau, \xi, y(\tau, \xi), z) + f^0(\tau, \xi, y(\tau, \xi), z) \right] = 0$$

and

$$\frac{\partial W}{\partial \tau}(\tau, \xi) + \min_{y \in Y} \left[\frac{\partial W}{\partial \xi}(\tau, \xi) f(\tau, \xi, y, z(\tau, \xi)) + f^0(\tau, \xi, y, z(\tau, \xi)) \right] = 0.$$

Remark 8.1. The sufficiency part of Theorem 8.1 may be proved directly by means of results analogous to Lemmas 6.1–6.4.

Similarly, the necessity part of this theorem may be proved directly proceeding as in Theorem 6.2. Such a direct proof is given in [2].

THEOREM 8.2. *Let $\mathscr{H}(t, x, y, z, \lambda) = f^0(t, x, y, z) + \lambda f(t, x, y, z)$ and let the function $\lambda(t; \tau, \xi)$ be defined as in* § 4. *Then the relation*

$$\begin{aligned} &\min_{y \in Y} \max_{z \in Z} \mathscr{H}(\tau, \xi, y, z, \lambda(\tau; \tau, \xi)) \\ &= \max_{z \in Z} \min_{y \in Y} \mathscr{H}(\tau, \xi, y, z, \lambda(\tau; \tau, \xi)) \\ &= \mathscr{H}(\tau, \xi, y(\tau, \xi), z(\tau, \xi), \lambda(\tau; \tau, \xi)) \end{aligned}$$

for every $(\tau, \xi) \in \mathscr{G} - M$ is a necessary and sufficient condition for optimality of marked trajectories.

Proof. From (5.3) and (5.4) it follows that the requirements of Theorem 8.2 are fulfilled. Moreover, if Y and Z have piecewise smooth boundaries, Remark 7.2 holds also for differential games.

Acknowledgments. The author wishes to express his most sincere thanks to Professor A. Halanay for his constant guidance and encouragement. The author also wishes to thank the reviewer for valuable suggestions.

REFERENCES

[1] R. BELLMAN, *Dynamic Programming*, Princeton University Press, Princeton, 1967.

[2] L. D. BERKOVITZ, *Necessary conditions for optimal strategies in a class of differential games and control problems*, J. SIAM Control, 5 (1967), pp. 1–24.

[3] ———, *Variational methods in problems of control and programming*, J. Math. Anal. Appl., 3 (1961), pp. 145–169.

[4] V. G. BOLTYANSKII, *Mathematical Methods of Optimal Control*, Izdat. Nauka, Moscow, 1966.

[5] ———, *Sufficient conditions for optimality and the justification of the dynamic programming method*, J. SIAM Control, 4 (1966), pp. 326–361.

[6] R. ISAACS, *Differential Games*, John Wiley, New York, 1965.

[7] J. C. MCKINSEY, *Introduction to the Theory of Games*, McGraw-Hill, New York, 1952.

[8] L. S. PONTRYAGIN, *Ordinary Differential Equations*, Addison–Wesley, Reading, Massachusetts, 1962.

[9] ———, *Smooth manifolds and their applications in homotopy theory*, Trudy Mat. Inst. Steklov., no. 45, Izdat. Akad. Nauk SSSR, Moscow, 1955, 139 pp.

[10] M. I. ZELIKIN AND N. T. TINEASCKII, *Deterministic differential games*, Uspehi Mat. Nauk, 20 (1965), no. 4, pp. 151–157.

SIAM J. CONTROL
Vol. 7, No. 2, May 1969

OPTIMUM CONTROL OF NON-GAUSSIAN LINEAR STOCHASTIC SYSTEMS WITH INACCESSIBLE STATE VARIABLES*

JAMES G. ROOT†

Summary. This article presents a new result in the optimum control of linear systems with respect to a quadratic performance criterion. It is assumed that the system is subject to additive random disturbance and that some state variables cannot be measured or can only be measured with additive noise. It is well known that when the disturbances and noise are Gaussian random variables, the optimum controller is a certain linear function of the mean of the posteriori distribution of state variables. It is shown here that this result holds without qualification.

1. Introduction. It is often the case in linear models that full information concerning the state of the system is not available to the decision maker. Rather than observing the state $x(t)$ at each epoch t, he observes a vector $y(t)$ that contains partial information concerning the state of the system. He must then base his control $u(t)$ upon partial information. A popular model for decision processes of this type is given by the matrix equations

$$x(t+1) = A(t)x(t) + B(t)u(t) + d(t) \quad \text{for } t = 0, 1, \cdots, T-1, \tag{1a}$$

$$y(t) = C(t)x(t) \quad \text{for } t = 0, 1, \cdots, T, \tag{1b}$$

where $x(t)$ and $d(t)$ are $m \times 1$ vectors, $u(t)$ is a $q \times 1$ vector, $y(t)$ is an $r \times 1$ vector with $r \leqq m$, the matrices $A(t)$, $B(t)$ and $C(t)$ are of appropriate dimensions and $C(t)$ is of rank r. We shall also assume, without loss of generality, that $C(t)$ consists of the first r rows of the $m \times m$ identity matrix. To require $r \leqq m$ and $C(t)$ to be of rank r would clearly result in no loss of generality. That no loss results from requiring $C(t)$ to be of such a simple form will be shown in § 2. The vector $d(t)$ represents an independent[1] random disturbance. We assume the probability distributions of $d(0), d(1), \cdots, d(T-1)$ are known and $E(d(t)) = 0$.

To start the decision process, $x(0)$ is chosen randomly according to a known probability distribution with mean zero. For $t > 0$, $x(t)$ is not deterministic, owing to the random disturbances. The control $u(t)$ must be based upon the observables, which are the history of observations, $y(t)$, $y(t-1), \cdots, y(0)$, and the history of decisions, $u(t-1), u(t-2), \cdots, u(0)$. All of this information can be summarized in a conditional probability distribution over the state space—this distribution then being updated from epoch to epoch.

* Received by the editors May 16, 1968, and in revised form February 7, 1969.

† System Sciences Department, The RAND Corporation, Santa Monica, California 90406. This research was supported in part by the National Bureau of Standards and in part by the United States Air Force under Project RAND.

[1] That is, $d(t)$ and $d(t')$ are independent if $t \neq t'$. However, $d_i(t)$ and $d_j(t)$ may be dependent.

To complete the specification of the model we select a quadratic performance criterion for the system, namely, to minimize

$$I = E\left(\sum_{t=0}^{T} x(t)'Z(t)x(t) + u(t)'Q(t)u(t) \right), \tag{2}$$

where $Z(t)$ is a symmetric nonnegative matrix and $Q(t)$ is a symmetric positive definite matrix.

The model simplifies in an essential manner if $C(t)$ is square, since then $y(t) = x(t)$. It is widely known (see, for example, [1]) that in this case the optimum control $u^o(t)$ is given by the relation

$$u^o(t) = M(t)x(t), \tag{3}$$

where $M(t)$ is a $q \times m$ matrix that can be computed recursively. The important fact here, of course, is that the *optimum* control is also linear in the state variables.

Returning to our case in which only partial information is available about the state variable $x(t)$ (i.e., when $C(t)$ is not square), let $\bar{x}(t)$, a function of the observables, be the mean of the posteriori distribution of $x(t)$; i.e.,[2]

$$\bar{x}(t) = E[x(t)|y(t), y(t-1), \cdots, y(0), u(t-1), u(t-2), \cdots, u(0)]. \tag{4}$$

It has been shown (see [2] and [3]) that the optimum control $u^o(t)$ is still of the simple form

$$u^o(t) = M(t)\bar{x}(t) \tag{5}$$

when the $d(t)$ are Gaussian distributed. Our contribution is to show that without the qualification optimum control is still given by (5).

The proof of this observation occupies § 3. In § 2, we pause to translate another popular model into the form given by (1) and (2) and to prove our assertion that no generality is lost by assuming the simple form for $C(t)$.

2. Translation of linear models to standard form. The problem described by (1) and (2) is general enough to include in its scope a number of other problems which at first glance may seem more general. We give an important example (see also [2] and [4]) which indicates the translation procedure and as a by-product substantiates our assertion about the form of $C(t)$.

Suppose (1a) is given by

$$y(t) = C_1(t)x(t) + s(t) \qquad \text{for } t = 0, 1, \cdots, T, \tag{6}$$

where $y(t)$ is an $r \times 1$ vector, where r may be greater than m, $C_1(t)$ is an $r \times m$ matrix with no other restrictions and the measurement errors, $s(t)$, $t = 0, 1, \cdots, T$, are independent[3] random $r \times 1$ vectors with $E(s(t)) = 0$. Suppose also there is a sequence $\{r(t)\}$, $r(t) \neq 0$, of reference vectors which, without loss of generality,

[2] The conditioning here on the $u(i)$, $i < t$, can be interpreted literally or as indicating that the expectation is to be computed under the assumption that control *functions* $u(i)$, $i = 0, 1, \cdots, t-1$, were used at epochs 0 through $t - 1$.

[3] Letting $v(t)' = (s(t+1)', d(t)')$, we only require that the variables $s(0)$, $d(T)$ and $v(t)$, $t = 0, 1, \cdots, T-1$, be independent.

we may suppose are generated by the recursion formula

$$r(t+1) = R(t)r(t), \tag{7}$$

where $R(t)$ is a square matrix. Also, suppose the criterion is given by either

$$I' = E\left(\sum_{t=0}^{T} ((x(t) - r(t))'Z(t)(x(t) - r(t)) + u(t)'Q(t)u(t))\right) \tag{8}$$

($r(t)$ is taken to be an $m \times 1$ vector in (8)) or

$$I'' = E\left(\sum_{t=0}^{T} ((y(t) - r(t))'Z(t)(y(t) - r(t)) + u(t)'Q(t)u(t))\right) \tag{9}$$

($r(t)$ is taken to be an $r \times 1$ vector in (9)). We shall show that, in fact, the problem defined by (1a), (6), (7) and (8) or (9) can be put in the form of (1) and (2).

We define

$$g(t) = \begin{pmatrix} r(t) \\ s(t) \\ x(t) \end{pmatrix}, \quad F(t) = \begin{pmatrix} R(t) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & A(t) \end{pmatrix}, \quad H(t) = \begin{pmatrix} 0 \\ 0 \\ B(t) \end{pmatrix},$$

$$e(t) = \begin{pmatrix} 0 \\ s(t+1) \\ d(t) \end{pmatrix}, \quad C_2(t) = \begin{pmatrix} I & 0 & 0 \\ 0 & I & C_1(t) \end{pmatrix}, \tag{10}$$

where the 0's and I's represent zero matrices and identity matrices of appropriate dimensions. Then (1a), (6) and (7) may now be rewritten as

$$g(t+1) = F(t)g(t) + H(t)u(t) + e(t), \tag{11a}$$

$$y(t) = C_2(t)g(t), \tag{11b}$$

where $C_2(t)$ is of full rank.

In the case of (8) we have

$$\begin{aligned} (x(t) - r(t))'Z(t)(x(t) - r(t)) &= g(t)' \begin{pmatrix} -I \\ 0 \\ I \end{pmatrix} Z(t)(-I, 0, I)g(t) \\ &= g(t)'\theta(t)g(t), \end{aligned} \tag{12}$$

where $\theta(t)$ is a nonnegative symmetric matrix, and for (9) we have

$$\begin{aligned} (y(t) - r(t))'Z(t)(y(t) - r(t)) &= (C_1(t)x(t) + s(t) - r(t))'Z(t)(C_1(t)x(t) + s(t) - r(t)) \\ &= g(t)' \begin{pmatrix} -I \\ I \\ C_1(t)' \end{pmatrix} Z(t)(-I, I, C_1(t))g(t) \\ &= g(t)'\phi(t)g(t), \end{aligned} \tag{13}$$

where $\phi(t)$ is a nonnegative symmetric matrix. The translation is complete except that $C_2(t)$ is not of the simple form of $C(t)$ in (1b).

To finish the translation, let $C_3(t)$ be a matrix such that $(C_2(t)', C_3(t)') = C_4(t)'$ is nonsingular. Let $w(t) = C_4(t)g(t)$. Then (11) and (8) (or (9)) become

$$w(t + 1) = C_4(t + 1)F(t)C_4(t)^{-1}w(t) + C_4(t + 1)H(t)u(t) + C_4(t + 1)e(t), \tag{14a}$$

$$y(t) = C_2(t)C_4(t)^{-1}w(t) \tag{14b}$$

and

$$I = E\left(\sum_{t=0}^{T} (w(t)'C_4(t)^{-1\prime}\theta(t)C_4(t)^{-1}w(t) + u(t)'Q(t)u(t))\right) \tag{15}$$

which is of the form of (1) and (2).

3. Derivation of the optimum controller. We begin the proof with some preliminary definitions and calculations which will be used later. It will be convenient to let $x_1(t)$ be the first r components of $x(t)$ which are observable and $x_2(t)$ be the last $m - r$ components of $x(t)$ which are not observable and rewrite (1) as

$$x_1(t + 1) = A_{11}(t)x_1(t) + A_{12}(t)x_2(t) + B_1(t)u(t) + d_1(t), \tag{16a}$$

$$x_2(t + 1) = A_{21}(t)x_1(t) + A_{22}(t)x_2(t) + B_2(t)u(t) + d_2(t), \tag{16b}$$

where $(d_1(t)', d_2(t)') = d(t)'$ and $y(t) = x_1(t)$.

The state (of knowledge) of the system described by (16) at epoch t is completely described by the couple (a, F^t), where $x_1(t) = a$ and F^t is the conditional distribution of $x_2(t)$ given $x_1(0), x_1(1), \cdots, x_1(t)$ and $u(0), u(1), \cdots, u(t - 1)$. We assume the distribution functions of $d(t)$ and $x(0)$ admit ordinary densities and have finite covariance matrices. This assumption implies F^t admits an ordinary density and has finite covariance matrix $V(F^t) = \{v_{ij}(F^t)\}$ and mean vector $m(t)$.[4]

For any vector b let $F^t_b(x) = F^t(x - b)$. Then $V(F^t) = V(F^t_b)$ for any b. Also let[5]

$$\begin{aligned} H^t(\xi, F^t, a, u) &= Pr[x_1(t + 1) \leqq \xi | x_2(t) : F^t, x_1(t) = a, u(t) = u] \\ &= Pr[A_{12}(t)x_2(t) + d_1(t) \leqq \xi - A_{11}(t)a - B_1(t)u | x_2(t) : F^t, x_1(t) = a, \\ &\qquad u(t) = u] \\ &= Pr[A_{12}(t)x_2(t) + d_1(t) \leqq \xi - A_{11}(t)a + A_{12}(t)b - B_1(t)u | x_2(t) : F^t_b, \\ &\qquad x_1(t) = a, u(t) = u] \qquad \text{(for any } b\text{)} \\ &= H^{*t}[\xi - A_{11}(t)a + A_{12}(t)b - B_1(t)u, F^t_b], \end{aligned} \tag{17}$$

[4] If we are in state (a, F^t), then $\bar{x}(t)' = (a', m(t)')$.

[5] The notation $y : F$ means the distribution function of y is F. If F is a distribution function with an ordinary density, we shall designate the density function with the symbol f.

$$\begin{aligned} G^t(\xi, F^t, a, u) &= Pr[x(t+1) \leqq \xi | x_2(t) : F^t, x_1(t) = a, u(t) = u] \\ &= Pr\left[\begin{pmatrix} A_{12}(t) \\ A_{22}(t) \end{pmatrix} x_2(t) + d(t) \leqq \xi - \begin{pmatrix} A_{11}(t) \\ A_{21}(t) \end{pmatrix} a - B(t)u \middle| x_2(t) : F^t, \right. \\ &\qquad \left. x_1(t) = a, u(t) = u \vphantom{\begin{pmatrix} A \\ A \end{pmatrix}} \right] \\ &= Pr\left[\begin{pmatrix} A_{12}(t) \\ A_{22}(t) \end{pmatrix} x_2(t) + d(t) \leqq \xi + \begin{pmatrix} A_{12}(t) \\ A_{22}(t) \end{pmatrix} b - \begin{pmatrix} A_{11}(t) \\ A_{21}(t) \end{pmatrix} a - B(t)u \middle| \right. \\ &\qquad \left. x_2(t) : F^t_b, x(t) = a, u(t) = u \vphantom{\begin{pmatrix} A \\ A \end{pmatrix}} \right] \qquad \text{(for any } b\text{)} \\ &= G^{*t}\left[\xi + \begin{pmatrix} A_{12}(t) \\ A_{22}(t) \end{pmatrix} b - \begin{pmatrix} A_{11}(t) \\ A_{21}(t) \end{pmatrix} a - B(t)u, F^t_b\right], \end{aligned} \tag{18}$$

$$\begin{aligned} J^t(\xi, e, c, F^t, a, u) &= Pr[x_2(t+1) + e \leqq \xi | x_2(t) : F^t, x_1(t) = a, u(t) = u, \\ &\qquad x_1(t+1) = c] \\ &= \int_{-\infty}^{\xi} \frac{g^{*t}\left[\begin{pmatrix} c \\ x - e \end{pmatrix} + \begin{pmatrix} A_{12}(t) \\ A_{22}(t) \end{pmatrix} b - \begin{pmatrix} A_{11}(t) \\ A_{21}(t) \end{pmatrix} a - B(t)u, F^t_b\right]}{h^{*t}[c - A_{11}(t)a + A_{12}(t)b - B_1(t)u, F^t_b]} \, dx \end{aligned} \tag{19}$$

and

$$\begin{aligned} &J^t(\xi, A_{22}(t)b - A_{21}(t)a - B_2(t)u, c, F^t, a, u) \\ &\quad = K^t(\xi, c + A_{12}(t)b - A_{11}(t)a - B_1(t)u, F^t_b). \end{aligned} \tag{20}$$

Our proof will be inductive and use a dynamic programming argument. We assume that the minimum expected loss from epoch $n + 1$ to T starting in state (a, F^{n+1}) at epoch $n + 1$ is

$$\begin{aligned} R^{n+1}(a, F^{n+1}) &= (a', m(n+1)')L(n+1)(a', m(n+1)')' \\ &\quad + \phi^{n+1}(F^{n+1}), \end{aligned} \tag{21}$$

where $\phi^{n+1}(F^{n+1}) = \phi^{n+1}(F^{n+1}_b)$ for any b and F^{n+1}; and $L(n+1) = \{l_{ij}(n+1)\}$ is a nonnegative $m \times m$ symmetric matrix.

We have

$$R^T(a, F^T) = \sum_{i,j \leqq m-r} z_{i+r,j+r}(T) v_{ij}(F^T) + (a', m(T)')Z(T)(a', m(T)')' \tag{22}$$

which is of the form of (21), since $v_{ij}(F^T) = v_{ij}(F^t_b)$ for any b.

We now express $R^n(c, F^n)$ as

$$\begin{aligned} R^n(c, F^n) = \min_u \Bigg\{ &\sum_{i,j \leqq m-r} z_{i+r,j+r}(n) v_{ij}(F^n) + (c', m(n)')Z(n)(c', m(n)')' \\ &+ u'Q(n)u + E[R^{n+1}(a, F^{n+1})] \Bigg\}, \end{aligned} \tag{23}$$

where $F^{n+1}(\cdot) = J^n(\cdot, 0, a, F^n, c, u)$ and the expectation is taken with respect to $H^n(\cdot, F^n, c, u)$.

Since $\phi^{n+1}(F^{n+1}) = \phi^{n+1}(F_e^{n+1})$ for any e, by using (20) the expectation of the last term in (21) can be written as

$$\begin{aligned} &E[\phi^{n+1}(J^n(\cdot, 0, a, F^n, c, u))] \\ &\quad = E[\phi^{n+1}(K^n(\cdot, a + A_{12}(n)b - A_{11}(n)c - B_1(n)u, F_b^n))] \end{aligned} \tag{24}$$

for any b. Making the change of variable $x = a + A_{12}(n)b - A_{11}(n)c - B_1(n)u$, we can express (24) as

$$\int_{-\infty}^{\infty} \phi^{n+1}(K^n(\cdot, x, F_b^n))h^{*t}(x, F_b^n)\,dx = \theta^n(F_b^n), \tag{25}$$

where for any b, $\theta^n(F^n) = \theta^n(F_b^n)$.

Noting that $E(x^2) = (E(x))^2 + E[(x - E(x))^2]$, we may write the first term of (21) as

$$\begin{aligned} &\sum_{i,j \leqq m-r} - (l_{i+r,j+r}(n + 1))v_{ij}(F^{n+1}) \\ &\quad + E(x(n + 1)'L(n + 1)x(n + 1)|x_2(n + 1):F^{n+1}, x_1(n + 1) = a). \end{aligned} \tag{26}$$

Since $E(v_{ij}(F^{n+1})|x_2(n):F_b^n, x_1(n) = c,\ u(n) = u)$ is independent of constants c, u or b, the expected value of the first term in (26) may be written as $D^n(F^n)$, where $D^n(F^n) = D^n(F_b^n)$ for any b.

For the expected value of the second term in (26) we have

$$\begin{aligned} &E[E(x(n + 1)'L(n + 1)x(n + 1)|x_2(n + 1):F^{n+1}, x_1(n + 1) = a)| \\ &\qquad x_2(n):F^n, x_1(n) = c, u(n) = u] \\ &= E[x(n + 1)'L(n + 1)x(n + 1)|x_2(n):F^n, x_1(n) = c, u(n) = u] \quad \text{(footnote 6)} \\ &= E[(A(n)x(n) + B(n)u + d(n))'L(n + 1)(A(n)x(n) + B(n)u + d(n))] \\ &= E[x(n)'(A(n)'L(n + 1)A(n))x(n) + x(n)'(A(n)'L(n + 1)B(n))u \\ &\quad + u'(B(n)'L(n + 1)A(n))x(n) + u'(B(n)'L(n + 1)B(n))u \\ &\quad + d(n)'L(n + 1)d(n)] \\ &= \sum_{i,j \leqq m-r} s_{i+r,j+r}(n)v_{ij}(F^n) + (c', m(n)')S(n)(c', m(n)')' \\ &\quad + (c', m(n)')J(n)'u + u'J(n)(c', m(n)')' + u'K(n)u \\ &\quad + E(d(n)'L(n + 1)d(n)), \end{aligned} \tag{27}$$

where $S(n) = A(n)'L(n + 1)A(n)$, $J(n) = B(n)'L(n + 1)A(n)$ and $K(n) = B(n)'L(n + 1)B(n)$ and $S(n)$ and $K(n)$ are nonnegative.

[6] In the following equations the conditioning is not stated but should be understood as indicated here.

Using the preceding results we may express (23) as

$$
\begin{aligned}
R^n(c, F^n) = \min_u \Bigg\{ \sum_{i,j \leqq m-r} & [z_{i+r,j+r}(n) + s_{i+r,j+r}(n)]v_{ij}(F^n) \\
& + (c', m(n)')(Z(n) + S(n))(c', m(n)')' \\
& + (c', m(n)')J(n)'u + u'J(n)(c', m(n)')' \\
& + u'(K(n) + Q(n))u + E(d(n)'L(n + 1)d(n)) \\
& + \theta^n(F^n) + D^n(F^n)\Bigg\}.
\end{aligned} \tag{28}
$$

Since $(K(n) + Q(n))$ is symmetric, the derivative of the function in the brackets on the right side of (28) is

$$\frac{d\{\cdot\}}{du} = 2(K(n) + Q(n))u + 2J(n)(c', m(n)')'. \tag{29}$$

Since $Q(n)$ is positive definite, so is $(K(n) + Q(n))$, and we find the optimum control function is

$$u^o(n) = -(K(n) + Q(n))^{-1}J(n)(c', m(n)')'. \tag{30}$$

By letting

$$
\begin{aligned}
\phi^n(F^n) = \sum_{i,j \leqq m-r} & [z_{i+r,j+r}(n) + s_{i+r,j+r}(n)]v_{ij}(F^n) \\
& + E(d'(n)L(n + 1)d(n)) + \theta^n(F^n) + D^n(F^n)
\end{aligned} \tag{31}
$$

and substituting $u^o(n)$ into (28), it may be verified that $R^n(c, F^n)$ is of the form of (21), and thus the proof is completed.

REFERENCES

[1] J. T. Tou, *Optimum Design of Digital Control Systems*, Academic Press, New York, 1963.

[2] T. L. Gunckel and G. F. Franklin, *A general solution for linear sampled-data control*, Trans. ASME Ser. D. J. Basic Engrg., 85D (1963), pp. 197–203.

[3] P. D. Joseph and J. T. Tou, *On linear control theory*, Trans. AIEE (Appl. and Indust.), 80 (1961), pp. 193–196.

[4] Masanao, Aoki, *Optimization of Stochastic Systems*, Academic Press, New York, 1967.

[5] R. E. Kalman, *A new approach to linear filtering and prediction problems*, Trans. ASME Ser. D. J. Basic Engrg., 82D (1960), pp. 35–45.

[6] A. R. M. Noton, *The extension of certain results in the treatment of inaccessible state variables*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 669–675.

[7] D. S. Adorno, *Optimum control of certain linear systems with quadratic loss*, Information and Control, 5 (1962), pp. 1–12.

SIAM J. CONTROL
Vol. 7, No. 2, May 1969

OPTIMAL CONTROL OF PROCESSES DESCRIBED BY INTEGRAL EQUATIONS. I*

V. R. VINOKUROV†

1. Let us assume that the behavior of a plant is described in n-dimensional Euclidean space, E_n, by the system of equations

$$x(t) = f(t) + \int_0^t K(t, x(s), u(s), s)\, ds, \tag{1.1}$$

which is regular in a neighborhood of each of its boundary points. That is, we column vector. The vector function $u(t)$, with values in E_r, will be called a *control.*

The allowable controls will be assumed to lie in a certain closed region $U \subset E_r$, which is regular in a neighborhood of each of its boundary points. That is, we assume that for every boundary point u_1 of U, it is possible to find continuously differentiable functions

$$q_i(u), \quad i = 1, 2, 3, \cdots, k, \tag{1.2}$$

such that in a neighborhood of u_1, U is described by $q_i(u) \leqq 0$, with $q_i(u_i) = 0$, $i = 1, 2, \cdots, k$ (see [1]). In the case that u_1 lies inside U, we take $k = 0$ in (1.2). As the class of allowable controls, we take the set of all piecewise continuous and piecewise smooth vector functions $u(t)$ on the segment $[0, T]$ having values in U at every instant. At points of discontinuity, we take $u(t) = u(t - 0)$. The solution of (1.1) for a given $u(t)$ will be called the *trajectory* corresponding to the given control $u(t)$.

Let us assign a region B in the space E_n, containing the point $f(0)$ and having a smooth boundary. This region is defined near the boundary by the inequality $\phi(x) \leqq 0$, where the scalar function $\phi(x)$ is twice continuously differentiable near the boundary $\phi(x) = 0$, and

$$\frac{\partial \phi(x)}{\partial x} = \text{grad } \phi(x)$$

vanishes nowhere on the boundary. Let there also be assigned functionals

$$I_0(x, u) = \int_0^T K^{(0)}(x(s), u(s), s)\, ds, \tag{1.3}$$

$$I_j(x, u) = \int_0^T K^{(n+j)}(x(s), u(s), s)\, ds + \Phi^{(j)}(x(T)), \quad j = 1, 2, \cdots, l. \tag{1.4}$$

* Originally published in Izv. Vyssh. Uchebn. Zaved. Matematika, 1967, no. 7, pp. 21–33, Submitted for publication on April 28, 1966. This translation into English has been prepared by R. N. and N. B. McDonough.

Translated and printed for this Journal under a grant-in-aid from the National Science Foundation.

† Orsk, U.S.S.R.

Let us finally assume that for $0 \leqq s \leqq t \leqq T, x \in B, u \in U$,

$$f^{(i)}(t), \quad \frac{df^{(i)}(t)}{dt}, \quad K^{(j)}(t, x, u, s), \quad \frac{\partial K^{(j)}(t, x, u, s)}{\partial t},$$

$$\frac{\partial K^{(j)}(t, x, u, s)}{\partial x^{(\alpha)}}, \quad \frac{\partial K^{(j)}(t, x, u, s)}{\partial u^{(\beta)}}, \quad \frac{\partial^2 K^{(j)}(t, x, u, s)}{\partial t \partial x^{(\alpha)}},$$

$$\frac{\partial^2 K^{(j)}(t, x, u, s)}{\partial t \partial u^{(\beta)}}, \quad \frac{\partial \Phi^{(j)}(x)}{\partial x^{(\alpha)}},$$

$$i = 1, 2, \cdots, n, \quad j = 0, 1, \cdots, n + l, \quad \alpha = 1, 2 \cdots, n, \quad \beta = 1, 2, \cdots, r,$$

exist and are continuous in all their arguments. We now examine the following problem.

Problem. Find an allowable control $u(t)$ having values in U, such that the corresponding trajectory $x(t)$ lies in the assigned region B, and such that $I_j(x, u) = 0$, $j = 1, 2, \cdots, l$, and $I_0(x, u)$ is minimum. For brevity, this will be called Problem (1.1)–(1.4). The solution functions $x(t)$ and $u(t)$ for this problem will be called the *optimal trajectory* and the *optimal control.*

A problem similar to Problem (1.1)–(1.4) for systems of differential equations was examined in [1, Chap. 6]. Equations and functionals more general than (1.1), (1.3) and (1.4) were considered in [3], but the problem with bounded phase coordinates was not examined. In addition, equations and functionals of the type (1.1), (1.3) and (1.4) lead to results which are more convenient for applications.

2. Let us define

$$g(x) = \begin{cases} \phi(x) & \text{for } x \text{ on the boundary of } B, \\ 0 & \text{for } x \text{ inside } B, \end{cases}$$

$$\tilde{K}^{(0)}(x, u, t) = K^{(0)}(x, u, t) + \sum_{j=1}^{l} \mu_j K^{(n+j)}(x, u, t),$$

$$x^{(0)}(t) = \int_0^t \tilde{K}^{(0)}(x(s), u(s), s)\, ds, \qquad \tilde{x} = \begin{pmatrix} x^{(0)} \\ x \end{pmatrix}, \quad \tilde{K} = \begin{pmatrix} \tilde{K}^{(0)} \\ K \end{pmatrix}, \quad \tilde{f} = \begin{pmatrix} 0 \\ f \end{pmatrix}.$$

Then system (1.1) is obviously contained in the system

$$\tilde{x}(t) = \tilde{f}(t) + \int_0^t \tilde{K}(t, x(s), u(s), s)\, ds. \tag{2.1}$$

We use the convention that if z is a vector and $\lambda(z)$ a scalar, then $\partial\lambda/\partial z = \operatorname{grad} \lambda(z)$ is a row vector, while if $f(z)$ is a vector, then $\partial f/\partial z$ is the matrix with elements

$$\left(\frac{\partial f(z)}{\partial z}\right)_{ij} = \frac{\partial f^{(i)}(z)}{\partial z^{(j)}}.$$

We further define

$$p_1(x,u,t) = \frac{\partial g(x)}{\partial x}K(t,x,u,t),$$

$$p_2(t,x,u,s) = \frac{\partial g(x(t))}{\partial x}\frac{\partial K(t,x(s),u(s),s)}{\partial t},$$

$$\frac{d}{dt}g(x(t)) \equiv p(x,u,t) = \left(\frac{\partial g(x(t))}{\partial x}, f'(t)\right) + p_1(x(t),u(t),t) + \int_0^t p_2(t,x,u,s)\,ds.$$

A trajectory $x(t)$ will be called *regular* relative to the control $u(t)$ if for all $t \in [0, T]$ the following conditions are satisfied:

(i) If $q_i(u(t))$ are the functions (1.2) for the $u(t)$ in question, then the vectors

$$\tag{2.2} \frac{\partial p_1(x(t),u(t),t)}{\partial u}, \quad \frac{\partial q_1(u(t))}{\partial u}, \quad \cdots, \quad \frac{\partial q_k(u(t))}{\partial u}$$

are linearly independent.

(ii) $p(x(t), u(t), t) = 0$. The set of controls u relative to which the trajectory x is regular will be denoted $\omega(x)$.

Let us define

$$\tag{2.3} H(x,z,u,t) = F(x,u,t) + \int_t^T z(s)K(s,x,u,t)\,ds,$$

where

(2.4)
$$F(x,u,t) = K^{(0)}(x,u,t) + \sum_{j=1}^{l}\mu_j\left[K^{(n+j)}(x,u,t) + \frac{\partial\Phi^{(j)}(x(T))}{\partial x}K(T,x,u,t)\right],$$

and $z(t)$ is the row vector which is the solution of the system

$$\tag{2.5} \begin{aligned} z(t) = {} & \frac{\partial F(x(t),u(t),t)}{\partial x} + \lambda(t)\left[x'(t)\frac{\partial^2 g(x(t))}{\partial x^2} + \frac{\partial g(x(t))}{\partial x}\frac{\partial K(t,x(t),u(t),t)}{\partial x}\right] \\ & + \int_t^T \lambda(s)\frac{\partial g(x(s))}{\partial x}\frac{\partial^2 K(s,x(t),u(t),t)}{\partial s\partial x}\,ds + \int_t^T z(s)\frac{\partial K(s,x(t),u(t),t)}{\partial x}\,ds \end{aligned}$$

($x'(t)$ is the row vector obtained by differentiating $x(t)$ in (1.1)). We see that $\lambda(t)$ is uniquely specified by

$$\tag{2.6} \begin{aligned} & \frac{\partial H(x(t),z,u(t),t)}{\partial u} + \lambda(t)\frac{\partial p_1(x(t),u(t),t)}{\partial u} \\ & + \int_t^T \lambda(s)\frac{\partial p_2(s,x,u,t)}{\partial u}\,ds + \sum_{j=1}^{k} v_j(t)\frac{\partial q_j(u(t))}{\partial u} = 0. \end{aligned}$$

The trajectory $x(t)$ and control $u(t)$ are said to satisfy the *maximum principle* if there exist a constant vector $\mu = (\mu_1, \mu_2, \cdots, \mu_l)$ and a piecewise smooth scalar function $\lambda(t)$ such that for the optimal trajectory $x(t)$ and optimal control $u(t)$, for which the functionals (1.4) vanish, (2.1)–(2.6) are satisfied, and for almost all $t \in [0, T]$,

$$H(x, z, u, t) = \min_{v \in \omega(x)} H(x, z, v, t), \tag{2.7}$$

$$\frac{d\lambda(t)}{dt} \leqq 0. \tag{2.8}$$

The unusual form of (2.7) results from the choice of sign for H.

THEOREM 2.1. *In order that $x(t)$, $u(t)$ be an optimal solution to Problem* (1.1)–(1.4), *with $x(t) \in B$, $u(t) \in U$, and $x(t)$ such that it undergoes at most a finite number of transitions from the boundary of B into its interior and back, it is both necessary and sufficient that $x(t)$ and $u(t)$ satisfy the maximum principle.*

Proof. Necessity. Let $\tilde{x}(t)$ be a trajectory of system (2.1) lying in $B \times Ox^{(0)}$, and let $\tilde{x}_i, i = 1, 2, \cdots, k$, be fixed points along the trajectory other than the endpoint $\tilde{x}(T)$. Following [1], we shall construct column vectors N_i such that

$$\left(\frac{\partial g(x)}{\partial x}, \; N_i\right) \geqq c > 0$$

for x lying in a sufficiently small neighborhood of x_i and continuously differentiable scalar functions $a_i(x)$ which are equal to unity in a certain neighborhood of $\tilde{x}_i$ and to zero in a certain other neighborhood of $\tilde{x}_i$. Also following [1], we introduce the functions

$$h(\tilde{x}, \mu) = g\left(x + \mu \sum_{\alpha=1}^{k} a_\alpha(\tilde{x}) N_\alpha\right), \tag{2.9}$$

$$P(\tilde{x}, u, \varepsilon\delta\mu, t) = \frac{dh(\tilde{x}, \varepsilon\delta\mu)}{dt}, \tag{2.10}$$

where the derivative is found using (2.1). Obviously $P(\tilde{x}, u, 0, t) = p(\tilde{x}, u, t)$. We consider the system

$$\begin{gathered} \tilde{y}(t) = \tilde{f}(t) + \int_0^t \tilde{K}(t, y(s), v(s), s)\, ds, \\ P(\tilde{y}, v, \varepsilon\delta\mu, t) = 0. \end{gathered} \tag{2.11}$$

As in [1], if $\tilde{y}(t)$ is sufficiently close to $\tilde{x}(t)$, and if $h(y(t_0), \varepsilon\delta\mu) = 0$, then $\tilde{v}(t) \in B \times Ox^{(0)}$ (and conversely). $\tilde{x}(t)$ and $u(t)$ satisfy (2.11) with $\delta\mu = 0$.

We show that if for $0 \leqq t \leqq \theta$ we have

$$\begin{aligned} \tilde{y}(t) &= \tilde{x}(t) + \varepsilon\delta\tilde{x}(t) + o(\varepsilon), \\ v(t) &= u(t) + \varepsilon\delta u(t) + o(\varepsilon) \end{aligned} \tag{2.12}$$

everywhere except possibly on segments the total length of which is of order ε, then it is possible to construct a solution of the system (2.11) such that (2.12) is also satisfied for $t \in [\theta, T]$, with smooth $\delta\tilde{x}(t)$ and piecewise smooth $\delta u(t)$. To do this, we divide the segment $[0, T]$ into partial segments of sufficiently small length, using points $0 = \tau_0 < \tau_1 < \cdots < \tau_m < \tau_{m+1} = T$, with all jump points of the control included among the points τ_i. Assuming that the solution $\tilde{y}(t)$, $v(t)$ of (2.12) has already been constructed on the segment $[\theta, \tau_i]$, we continue it to the segment $[\tau_i, \tau_{i+1}]$. According to the conditions, the vectors (2.2) are linearly independent at the point $\tau_i + 0$. Let us assume that the following Jacobian does not vanish:

$$(2.13) \qquad \frac{D(p_1(x(\tau_i), v(\tau_i + 0), \tau_i + 0), q_1(v(\tau_i + 0)), \cdots, q_k(v(\tau_i + 0)))}{D(v^{(1)}, v^{(2)}, \cdots, v^{(k+1)})} \neq 0.$$

Let us write the second equation of the system (2.11) in the form

$$(2.14) \quad \frac{\partial h(\tilde{y}(t), \varepsilon\delta\mu)}{\partial \tilde{y}} \tilde{K}(t, y(t), v(t), t) = -\frac{\partial h(\tilde{y}(t), \varepsilon\delta\mu)}{\partial \tilde{y}} \tilde{f}'(t) - \phi(t, \tilde{y}, v) - \psi(t, \tilde{y}, v),$$

where

$$(2.15) \qquad \begin{aligned} \phi(t, \tilde{y}, v) &= \int_0^{\tau_i} \frac{\partial h(\tilde{y}(t), \varepsilon\delta\mu)}{\partial \tilde{y}} \frac{\partial K(t, y(s), v(s), s)}{\partial t} ds, \\ \psi(t, \tilde{y}, v) &= \int_{\tau_i}^{t} \frac{\partial h(\tilde{y}(t), \varepsilon\delta\mu)}{\partial \tilde{y}} \frac{\partial K(t, y(s), v(s), s)}{\partial t} ds, \end{aligned}$$

and adjoin to it the equations

$$(2.16) \qquad q_i(v(t)) = q_i(u(t)), \qquad i = 1, 2, \cdots, k.$$

For t near τ_i and for sufficiently small ε, from (2.14)–(2.16) it is possible to find $v^{(1)}$, $v^{(2)}, \cdots, v^{(k+1)}$ as functions of the form $v^{(j)} = \eta^{(j)}(t, \tilde{y}, v^{(k+2)}, \cdots, v^{(r)}, \varepsilon\delta\mu, \phi, \psi)$, $j = 1, 2, \cdots, k + 1$, where the $\eta^{(j)}$ are differentiable in all arguments.

For $j = k + 2, \cdots, r$, let us take $v^{(j)}(t) = u^{(j)}(t)$, and examine the system

$$(2.17) \qquad \begin{aligned} \tilde{y}(t) &= \tilde{f}(t) + \int_0^{\tau_i} K(t, y(s), v(s), s)\, ds + \int_{\tau_i}^{t} K(t, y(s), v(s), s)\, ds, \\ v^{(j)}(t) &= \eta^{(j)}(t, \tilde{y}(t), u^{(k+2)}(t), \cdots, u^{(r)}(t), \varepsilon\delta\mu, \phi, \psi), \\ &\qquad\qquad j = 1, 2, \cdots, k + 1, \end{aligned}$$

where ϕ and ψ are defined by (2.15). Since the $\eta^{(j)}$ satisfy a Lipschitz condition in the arguments $\tilde{y}(t)$ and ψ, and $\partial\tilde{K}(t, y, v, s)/\partial t$ satisfies a Lipschitz condition in y and v, by the usual method of successive approximations it can be proved that on a small interval $[\tau_i, \tau_{i+1}]$ the system (2.17) has a unique solution. Let us now show that it satisfies (2.12). Assume that $\tilde{y}(t) = \tilde{x}(t) + \tilde{z}(\varepsilon, t)$, $v(t) = u(t) + w(\varepsilon, t)$, where for $j = k + 2, \cdots, r$ we have $w^{(j)}(\varepsilon, t) = 0$. For $\delta\mu = 0$, we have

$$\tilde{x}(t) = \tilde{f}(t) + \int_0^{\tau_i} \tilde{K}(t, x(s), u(s), s)\, ds + \int_{\tau_i}^{t} \tilde{K}(t, x(s), u(s), s)\, ds,$$

$$u^{(j)}(t) = \eta^{(j)}\left(t, \tilde{x}(t), u^{(k+2)}(t), \cdots, u^{(r)}(t), 0, \int_0^{\tau_i} p_2(t, \tilde{x}, u, s)\, ds, \int_{\tau_i}^{t} p_2(t, \tilde{x}, u, s)\, ds\right),$$
$$j = 1, 2, \cdots, k + 1.$$

Since by assumption, for $t \leqq \tau_i$, (2.12) is satisfied everywhere except possibly on intervals of total length of order ε, we obtain

$$\|\tilde{z}(\varepsilon, t)\| \leqq A\varepsilon \max_{0 \leqq s \leqq \tau_i} (\|\delta\tilde{x}(s)\| + \delta u(s)\|) + B \int_{\tau_i}^{t} (\|\tilde{z}(\varepsilon, s)\| + \|\omega(\varepsilon, s)\|)\, ds + K\varepsilon,$$

$$\begin{aligned}\|\omega(\varepsilon, t)\| &\leqq C\varepsilon\delta\mu + D\|\tilde{z}(\varepsilon, t)\| + F \max_{0 \leqq s \leqq \tau_i} (\|\delta\tilde{x}(s)\| + \|\delta u(s)\|)\\ &\quad + G \int_{\tau_i}^{t} (\|\tilde{z}(\varepsilon, s)\| + \|\omega(\varepsilon, s)\|)\, ds + L\varepsilon\\ &\leqq \varepsilon[C\delta\mu + DK + L + (AD + F) \max_{0 \leqq s \leqq \tau_i} (\|\delta\tilde{x}(s)\| + \|\delta u(s)\|)]\\ &\quad + (BD + G) \int_{\tau_i}^{t} (\|\tilde{z}(\varepsilon, s)\| + \|\omega(\varepsilon, s)\|)\, ds.\end{aligned}$$

From this it follows that $\tilde{z}(\varepsilon, t)$ and $w(\varepsilon, t)$ are of order ε.

We shall now prove the existence of piecewise smooth n-dimensional column vector functions $\Delta(t, s)$, $L_\beta(t, s)$, $\beta = 1, 2, \cdots, k$, and piecewise smooth scalar functions $\Delta^{(0)}(t, s)$, $L_\beta^{(0)}(t, s)$, $\beta = 1, 2, \cdots, k$, such that for $0 \leqq s \leqq t \leqq T$, we have

$$\begin{aligned}&\left[\frac{\partial \tilde{K}(t, x(s), u(s), s)}{\partial u} + \tilde{\Delta}(t, s)\frac{\partial p_1(x(s), u(s), s)}{\partial u}\right.\\ &\left.\quad + \int_s^t \Delta(t, \sigma)\frac{\partial p_2(\sigma, x, u, s)}{\partial u}\, d\sigma\right]\delta u(s) = 0,\end{aligned} \tag{2.18}$$

where

$$\tilde{\Delta}(t, s) = \begin{pmatrix}\Delta^{(0)}(t, s)\\ \Delta(t, s)\end{pmatrix}, \qquad \tilde{L}_\beta(t, s) = \begin{pmatrix}L_\beta^{(0)}(t, s)\\ L_\beta(t, s)\end{pmatrix}.$$

If at the point s the Jacobian (2.13) is nonzero, we define $\Delta^{(j)}(t, s)$, $L_\beta^{(j)}(t, s)$ as the solution of the system

$$\begin{aligned}&\frac{\partial \tilde{K}^{(j)}(t, x(s), u(s), s)}{\partial u^{(\alpha)}} + \Delta^{(j)}(t, s)\frac{\partial p_1(x(s), u(s), s)}{\partial u^{(\alpha)}} + \int_s^t \Delta^{(j)}(t, \sigma)\frac{\partial p_2(\sigma, x, u, s)}{\partial u^{(\alpha)}}\, d\sigma\\ &\quad + \sum_{\beta=1}^{k} L_\beta^{(j)}(t, s)\frac{\partial q_\beta(u(s))}{\partial u^{(\alpha)}} = 0, \qquad \alpha = 1, 2, \cdots, k + 1,\end{aligned} \tag{2.19}$$

the existence and uniqueness of which are proved in the same way as for (2.12), with the argument made easier in view of the linearity. Then (2.18) is established in the same way as (6.30) of [1], and we find that $\delta u^{(\alpha)} = 0$, $\alpha = k + 2, \cdots, r$.

Let us now divide the interval $[0, T]$ into parts by points $0 < \tau_1 \leqq \tau_2 < \cdots \leqq \tau_m \leqq \tau = T$, and as in [1] select arbitrary nonnegative numbers $\delta t_1, \delta t_2, \cdots, \delta t_m$, and a number δt, with the points $\tau_1, \tau_2, \cdots, \tau_m$ being continuity points of $u(t)$. Further, let the points $v_1, v_2, \cdots, v_m \in U$ be such that the vectors

$$\frac{\partial p_1(x(\tau_i), v_i, \tau_i)}{\partial u}, \quad \frac{\partial q_j(v_j)}{\partial u}, \qquad j = 1, 2, \cdots, k,$$

are linearly independent. Let

$$l_i = \begin{cases} \delta t - (\delta t_i + \cdots + \delta t_m) & \text{for } \tau_i = \tau, \\ \quad - (\delta t_i + \cdots + \delta t_m) & \text{for } \tau_i = \tau_m < \tau, \\ \quad - (\delta t_i + \cdots + \delta t_j) & \text{for } \tau_j = \tau_{i+1} = \cdots = \tau_j < \tau_{j+1} \end{cases}$$

and

$$I_i = (\tau_i + \varepsilon l_i, \tau_i + \varepsilon(l_i + \delta t_i)), \qquad i = 1, 2, \cdots, m,$$

where ε is sufficiently small. We shall construct a varied trajectory $\tilde{x}^*(t)$ and control $u^*(t)$. For $0 \leqq t \leqq \tau_1 + \varepsilon l_1$, we assume $\tilde{x}^*(t) = \tilde{x}(t)$, $u^*(t) = u(t)$, and continue them on the subinterval I_1 in the following way. Let $q_i(v)$ be the functions (1.2) for the point v_1. Like the system (2.14)–(2.16), the system $P(\tilde{y}, v, \varepsilon\delta\mu, t) = q_1(v) = \cdots = q_k(v) = 0$ is solvable, say, for the first $k + 1$ coordinates of the vector v in the neighborhood of $\tilde{x}(\tau_1)$, v_1, $\varepsilon\delta\mu = 0$, $t = \tau_1$. For $i = k + 2, \cdots, r$, we assume $v^{(i)} = v_1^{(i)}$ on I_1, and then, as in the solution of (2.17), find $\tilde{x}^*(t)$, $u^*(t)$ on I_1. As $\varepsilon \to 0$, $u^*(t)$ will tend uniformly to v_1 on I_1. For $\tau_1 = \tau_2 = \cdots = \tau_j \leqq \tau_{j+1}$, in the same way we construct a varied trajectory and control on $I_2, \cdots, I_j$, taking $v_2, \cdots, v_j$ rather than v_1. Then for $t \leqq t_j$, it is easy to establish that $\tilde{x}^*(t)$ satisfies (2.12) everywhere except on the intervals $I_1, I_2, \cdots, I_j$, and therefore they can be extended to the point τ_{j+1}, using (2.14)–(2.16), in such a way that (2.12) remains true. Continuing, we obtain for points t lying to the right of all intervals I_i:

$$\begin{aligned} \tilde{x}^*(t) &= \tilde{x}(t) + \varepsilon\delta\tilde{x}(t) + o(\varepsilon) \\ &= \tilde{f}(t) + \int_0^{\tau_1 + \varepsilon l_1} \tilde{K}(t, x(s), u(s), s)\, ds \\ &\quad + \sum_{i=1}^{m} \int_{Ii} \tilde{K}(t, x(s) + \varepsilon\delta x(s) + o(\varepsilon), v_i + \phi_i(s, \varepsilon), s)\, ds \\ &\quad + \left(\int_0^{\tau_1 + \varepsilon l_1} - \int_{I_1} \right) \tilde{K}(t, x(s) + \varepsilon\delta x(s) + o(\varepsilon), u(s) + \varepsilon\delta u(s) + o(\varepsilon), s)\, ds, \end{aligned}$$

where the $\phi_i(s, \varepsilon)$ tend uniformly to zero on I_i as $\varepsilon \to 0$. From this we obtain

(2.20)

$$\partial\tilde{x}(t) = \delta\tilde{K}(t, x, u, \tau) + \int_{\tau_1}^{t} \left[\frac{\partial\tilde{K}(t, x(s), u(s), s)}{\partial x} \delta x(s) + \frac{\partial\tilde{K}(t, x(s), u(s), s)}{\partial u} \delta u(s) \right] ds,$$

where

$$\delta\tilde{K}(t, x, u, \tau) = \sum_{i=1}^{m} \delta_i\tilde{K}(t, x(\tau_i), u(\tau_i), \tau_i)\delta t_i,$$

$$\delta_i\tilde{K}(t, x(\tau_i), u(\tau_i), \tau_i) = \tilde{K}(t, x(\tau_i), v_i, \tau_i) - \tilde{K}(t, x(\tau_i), u(\tau_i), \tau_i),$$

and $\tilde{x}^*(t)$ and $u^*(t)$ satisfy (2.11) for $\delta\mu = 0$. According to the conditions, $\tilde{x}^*(t)$, $u^*(t)$ must also satisfy (2.11).

Let us multiply the second of equations (2.11), in the case where $t = s$, by $\tilde{\Delta}(t, s)$, and integrate over s from 0 to t. Then taking into account that I_i has length $\varepsilon\delta t_i$, we obtain

$$\int_0^t \tilde{\Delta}(t, s)\frac{\partial g(x(s))}{\partial x}\left[\tilde{f}'(s) + \tilde{K}(s, x(s), u(s), s) + \int_0^s \frac{\partial \tilde{K}(s, x(\sigma), u(\sigma), \sigma)}{\partial s}\,d\sigma\right] ds = 0,$$

$$\begin{aligned}
&\int_0^{\tau_1} \tilde{\Delta}(t, s)\frac{\partial g(x(s))}{\partial x}\left[\tilde{f}'(s) + \tilde{K}(s, x(s), u(s), s) + \int_0^s \frac{\partial \tilde{K}(s, x(\sigma), u(\sigma), \sigma)}{\partial s}\,d\sigma\right] ds \\
&+ \sum_{i=1}^{m} \int_{I_i} \tilde{\Delta}(t, s)\frac{\partial g(x(s))}{\partial x}\left[\tilde{f}'(s) + \tilde{K}(s, x(s), u(s), s)\right. \\
&\left. + \int_0^s \frac{\partial \tilde{K}(s, x(\sigma), u(\sigma), \sigma)}{\partial s}\,d\sigma\right] ds \\
&+ \left(\int_{\tau_1}^t - \sum_{i=1}^{m} \int_{I_i}\right)\tilde{\Delta}(t, s)\,\frac{\partial h(x(s) + \varepsilon\delta x(s), \varepsilon\delta\mu)}{\partial x} \\
&\cdot\left[\tilde{f}'(s) + \tilde{K}(s, x(s) + \varepsilon\delta x(s), u(s) + \varepsilon\delta u(s), s)\right. \\
&+ \int_0^{\tau_1} \frac{\partial \tilde{K}(s, x(\sigma), u(\sigma), \sigma)}{\partial s}\,d\sigma + \sum_{j=1}^{m} \int_{I_j} \frac{\partial \tilde{K}(s, x(\sigma), v_j, \sigma)}{\partial s}\,d\sigma \\
&\left. + \int_{\tau_1}^s \frac{\partial \tilde{K}(s, x(\sigma) + \varepsilon\delta x(\sigma), u(\sigma) + \varepsilon\delta u(\sigma), \sigma)}{\partial s}\,d\delta\right] ds + o(\varepsilon) = 0.
\end{aligned}$$

Subtracting the first equation from the second, we find

$$\begin{aligned}
&\tilde{\Delta}(t, \tau)\delta p_1(x(\tau), u(\tau), \tau) + \int_{\tau_1}^t \tilde{\Delta}(t, s)\delta p_2(s, x, u, \tau)\,ds + \int_{\tau_1}^t \frac{\partial P(x, u, 0, s)}{\partial \mu}\delta\mu\,ds \\
&\quad + \int_{\tau_1}^t \left[\tilde{\Delta}(t, s)\left(x'(s)\frac{\partial^2 g(x(s))}{\partial x^2} + \frac{\partial g(x(s))}{\partial x}\,\frac{\partial K(s, x(s), u(s), s)}{\partial x}\right)\right. \\
&\quad \left. + \int_s^t \tilde{\Delta}(t, \sigma)\frac{\partial g(x(\sigma))}{\partial x}\,\frac{\partial^2 K(\sigma, x(s), u(s), s)}{\partial\sigma\partial x}\,d\sigma\right] \delta x(s)\,ds \\
&\quad + \int_{\tau_1}^t \left[\tilde{\Delta}(t, s)\frac{\partial p_1(x(s), u(s), s)}{\partial u} + \int_s^t \tilde{\Delta}(t, \sigma)\frac{\partial p_2(\sigma, x, u, s)}{\partial u}\,d\sigma\right]\delta u(s)\,ds = 0,
\end{aligned} \tag{2.21}$$

where δ has the same value as in (2.20), and $x'(s)$ is the row vector obtained by differentiating $x(s)$ using (1.1). Using (2.21) with (2.20), and taking into account

(2.18), we obtain finally:

$$\begin{aligned}\delta\tilde{x}(t) = {}& \delta\tilde{M}(t,x,u,\tau) + \int_{\tau_1}^{t} \Delta(t,s)\frac{\partial P(x,u,0,s)}{\partial\mu}\delta\mu\,ds \\ &+ \int_{\tau_1}^{t}\Bigg[\frac{\partial\tilde{K}(t,x(s),u(s),s)}{\partial x} + \tilde{\Delta}(t,s)\bigg(x'(s)\frac{\partial^2 g(x(s))}{\partial x^2} \\ &+ \frac{\partial g(x(s))}{\partial x}\,\frac{\partial K(s,x(s),u(s),s)}{\partial x}\bigg) \\ &+ \int_{s}^{t}\tilde{\Delta}(t,\sigma)\frac{\partial g(x(\sigma))}{\partial x}\,\frac{\partial^2 K(\sigma,x(s),u(s),s)}{\partial\sigma\partial x}\,d\sigma\Bigg]\delta x(s)\,ds,\end{aligned}\tag{2.22}$$

where

$$\tilde{M}(t,x,u,\tau) = \begin{pmatrix} M^{(0)}(t,x,u,\tau) \\ M(t,x,u,\tau)\end{pmatrix},$$

$$\begin{aligned}M^{(0)}(t,x,u,\tau) &= \tilde{K}^{(0)}(x,u,\tau) + \Delta^{(0)}(t,\tau)p_1(x(\tau),u(\tau),\tau) \\ &\quad + \int_{\tau_1}^{t}\Delta^{(0)}(t,s)p_2(s,x,u,\tau)\,ds, \\ M(t,x,u,\tau) &= K(t,x,u,\tau) + \Delta(t,\tau)p_1(x(\tau),u(\tau),\tau) \\ &\quad + \int_{\tau_1}^{t}\Delta(t,s)p_2(s,x,u,\tau)\,dx.\end{aligned}\tag{2.23}$$

From this, in particular

$$\begin{aligned}\delta x^{(0)}(t) = {}& \delta M^{(0)}(t,x,u,\tau) + \int_{\tau_1}^{t}\Delta^{(0)}(t,s)\frac{\partial P(x,u,0,s)}{\partial\mu}\delta\mu\,ds \\ &+ \int_{\tau_1}^{t} Q(t,x,u,s)\,\delta x(s)\,ds,\end{aligned}\tag{2.24}$$

$$\begin{aligned}\delta x(t) = {}& \delta M(t,x,u,\tau) + \int_{\tau_1}^{t}\Delta(t,s)\frac{\partial P(x,u,0,s)}{\partial\mu}\delta\mu\,ds \\ &+ \int_{\tau_1}^{t} Q(t,x,u,s)\delta x(s)\,ds,\end{aligned}\tag{2.25}$$

where

$$\begin{aligned}Q^{(0)}(t,x,u,s) = {}& \frac{\partial\tilde{K}^{(0)}(x(s),u(s),s)}{\partial x} \\ &+ \Delta^{(0)}(t,s)\left[x'(s)\frac{\partial^2 g(x(s))}{\partial x^2} + \frac{\partial g(x(s))}{\partial x}\,\frac{\partial K(s,x(s),u(s),s)}{\partial x}\right] \\ &+ \int_{s}^{t}\Delta^{(0)}(t,\sigma)\frac{\partial g(x(\sigma))}{\partial x}\,\frac{\partial^2 K(\sigma,x(s),u(s),s)}{\partial\sigma\partial x}\,d\sigma,\end{aligned}\tag{2.26}$$

$$Q(t,x,u,s) = \frac{\partial K(t,x(s),u(s),s)}{\partial x}$$

$$+ \Delta(t,s)\left[x'(s)\frac{\partial^2 g(x(s))}{\partial x^2} + \frac{\partial g(x(s))}{\partial x}\frac{\partial K(s,x(s),u(s),s)}{\partial x}\right] \tag{2.27}$$

$$+ \int_s^t \Delta(t,\sigma)\frac{\partial g(x(\sigma))}{\partial x}\frac{\partial^2 K(\sigma,x(s),u(s),s)}{\partial\sigma\partial x}\,d\sigma.$$

Let $R(t,s)$ be the resolvent of the matrix $Q(t,x,u,s)$. Then

$$\delta x(t) = \delta M(t,x,u,\tau) + \int_{\tau_1}^t R(t,s)\delta M(s,x,u,\tau)\,ds$$

$$+ \int_{\tau_1}^t \left[\Delta(t,s) + \int_s^t R(t,\sigma)\Delta(\sigma,s)\,d\sigma\right]\frac{\partial P(x,u,0,s)}{\partial\mu}\delta\mu\,ds. \tag{2.28}$$

Substituting (2.28) into (2.24), we find for $t = T$:

$$\delta x^{(0)}(T) = \delta M^{(0)}(T,x,u,\tau) + \int_{\tau_1}^T \left[Q^{(0)}(T,x,u,s)\right.$$

$$\left. + \int_s^T Q^{(0)}(T,x,u,\sigma)R(\sigma,s)\,d\sigma\right]\delta M(s,x,u,\tau)\,ds$$

$$+ \int_{\tau_1}^T \left\{\Delta^{(0)}(T,s) + \int_s^T \left[Q^{(0)}(T,x,u,\sigma)\right.\right.$$

$$\left.\left. + \int_\sigma^T Q^{(0)}(T,x,u,\theta)R(\theta,\sigma)\,d\theta\right]\Delta(\sigma,s)\,d\sigma\right\}\frac{\partial P(x,u,0,s)}{\partial\mu}\delta\mu\,ds.$$

If $w(t)$ is the solution of

$$w(t) = Q^{(0)}(T,x,u,t) + \int_t^T w(s)Q(s,x,u,t)\,ds, \tag{2.29}$$

then the preceding equation can be written more simply

$$\delta x^{(0)}(T) = \delta M^{(0)}(T,x,u,\tau) + \int_{\tau_1}^T w(s)\delta M(s,x,u,\tau)\,ds$$

$$+ \int_{\tau_1}^T \left[\Delta^{(0)}(T,s) + \int_s^T w(\sigma)\Delta(\sigma,s)\,d\sigma\right]\frac{\partial P(x,u,0,s)}{\partial\mu}\delta\mu\,ds. \tag{2.30}$$

Let

$$I(x,u) = I_0(x,u) + \sum_{j=1}^{l}\mu_j I_j(x,u). \tag{2.31}$$

Then obviously

$$\delta I(x,u) = \delta x^{(0)}(T) + \sum_{j=1}^{l}\mu_j\frac{\partial\Phi^{(j)}(x(T))}{\partial x}\delta x(T).$$

Taking into account (2.28) and (2.30), we find:

$$
\begin{aligned}
\delta I(x,u) = {} & \delta\biggl\{M^{(0)}(T,x,u,\tau) + \sum_{j=1}^{l} \mu_j \frac{\partial \Phi^{(j)}(x(T))}{\partial x} M(T,x,u,\tau) \\
& + \int_{\tau_1}^{T}\biggl[w(s) + \sum_{j=1}^{l} \mu_j \frac{\partial \Phi^{(j)}(x(T))}{\partial x} R(T,s)\biggr] M(s,x,u,\tau)\,ds\biggr\} \\
& + \int_{\tau_1}^{T}\biggl\{\Delta^{(0)}(T,s) + \int_{s}^{T}\biggl[w(\sigma) + \sum_{j=1}^{l} \mu_j \frac{\partial \Phi^{(j)}(x(T))}{\partial x} R(T,\sigma)\biggr]\Delta(\sigma,s)\,d\sigma \\
& + \sum_{j=1}^{l} \mu_j \frac{\partial \Phi^{(j)}(x(T))}{\partial x} \Delta(T,s)\biggr\} \frac{\partial P(x,u,0,s)}{\partial \mu}\delta\mu\,ds,
\end{aligned}
\tag{2.32}
$$

where δ is applied to the functions $M^{(0)}$ and M. Let us introduce the following:

$$
z(t) = w(t) + \sum_{j=1}^{l} \mu_j \frac{\partial \Phi^{(j)}(x(T))}{\partial x} R(T,t),
\tag{2.33}
$$

$$
\lambda(t) = \Delta^{(0)}(T,t) + \int_{t}^{T} z(s)\Delta(s,t)\,ds + \sum_{i=1}^{l} \mu_j \frac{\partial \Phi^{(j)}(x(T))}{\partial x} \Delta(T,t).
\tag{2.34}
$$

Then by simple calculations it can be shown that $z(t)$ satisfies (2.5), and

$$
\begin{aligned}
\delta I(x,u) = {} & \delta\biggl[F(x(\tau),u(\tau),\tau) + \lambda(\tau)p_1(x(\tau),u(\tau),\tau) \\
& + \int_{\tau_1}^{T} \lambda(s)p_2(s,x,u,\tau)\,ds + \int_{\tau_1}^{T} z(s)K(s,x(\tau),u(\tau),\tau)\,ds\biggr] \\
& + \int_{\tau_1}^{T} \lambda(s)\frac{\partial P(x,u,0,s)}{\partial \mu}\delta\mu\,ds.
\end{aligned}
\tag{2.35}
$$

We shall now prove that $\lambda(t)$ defined by (2.34) satisfies (2.6). From (2.19),

$$
\begin{aligned}
& \frac{\partial \tilde{K}^{(0)}(x(t),u(t),t)}{\partial u^{(\alpha)}} + \Delta^{(0)}(T,t)\frac{\partial p_1(x(t),u(t),t)}{\partial u^{(\alpha)}} \\
& \quad + \int_{t}^{T} \Delta^{(0)}(T,s)\frac{\partial p_2(s,x,u,t)}{\partial u^{(\alpha)}}\,ds + \sum_{\beta=1}^{k} L_\beta^{(0)}(T,t)\frac{\partial q_\beta(u(t))}{\partial u^{(\alpha)}} = 0, \\
& \int_{t}^{T} z(s)\biggl[\frac{\partial K(s,x(t),u(t),t)}{\partial u^{(\alpha)}} + \Delta(s,t)\frac{\partial p_1(x(t),u(t),t)}{\partial u^{(\alpha)}} \\
& \quad + \sum_{\beta=1}^{k} L_\beta(s,t)\frac{\partial q_\beta(u(t))}{\partial u^{(\alpha)}}\biggr]\,ds + \int_{t}^{T}\int_{s}^{T} z(\sigma)\Delta(\sigma,s)\,d\sigma\frac{\partial p_2(s,x,u,t)}{\partial u^{(\alpha)}}\,ds = 0, \\
& \sum_{j=1}^{l} \mu_j \frac{\partial \Phi^{(j)}(x(T))}{\partial x}\biggl[\frac{\partial K(T,x(t),u(t),t)}{\partial u^{(\alpha)}} + \Delta(T,t)\frac{\partial p_1(x(t),u(t),t)}{\partial u^{(\alpha)}} \\
& \quad + \int_{t}^{T} \Delta(T,s)\frac{\partial p_2(s,x,u,t)}{\partial u^{(\alpha)}}\,ds + \sum_{\beta=1}^{k} L_\beta(T,t)\frac{\partial q_\beta(u(t))}{\partial u^{(\alpha)}}\biggr] = 0,
\end{aligned}
$$

$$
\alpha = 1, 2, \cdots, k+1.
$$

Adding these equations, we obtain

$$\frac{\partial H(x(t), z, u(t), t)}{\partial u^{(\alpha)}} + \lambda(t)\frac{\partial p_1(x(t), u(t), t)}{\partial u^{(\alpha)}}$$

$$+ \int_t^T \lambda(s)\frac{\partial p_2(s, x, u, t)}{\partial u^{(\alpha)}} ds + \sum_{\beta=1}^{k} v_\beta(t)\frac{\partial q_\beta(u(t))}{\partial u^{(\alpha)}} = 0,$$

$$\alpha = 1, 2, \cdots, k + 1,$$

with

$$v_\beta(t) = L_\beta^{(0)}(T, t) + \int_t^T z(s)L_\beta(s, t)\, ds + \sum_{j=1}^{l} \mu_j \frac{\partial \Phi^{(j)}(x(T))}{\partial x} L_\beta(T, t).$$

Furthermore, taking a single division point $\tau_1 = t$ on the interval $[0, T]$, and assuming $\delta\mu = 0$, we see that $\delta I(x, u)$ coincides with $\delta H(x, z, u, t)$, under the condition that (2.6) is satisfied. From this follows (2.7). Inequality (2.8) is proved analogously to the corresponding inequality in [1], from the expression for $\delta I(x, u)$ with $\delta\mu > 0$ and considering the empty set of division points of the interval $[0, T]$. In this way, the necessity of the criterion is proved.

Sufficiency. Let us integrate the function $H(x(t), z, u(t), t)$ over t from 0 to T. Taking into account (1.1), (1.3), (2.3), (2.4) and the vanishing of the functionals (1.4), we obtain

$$\begin{aligned}\int_0^T H(x(t), z, u(t), t)\, dt &= \int_0^T K^{(0)}(x(t), u(t), t)\, dt \\ &\quad + \int_0^T \sum_{j=1}^{l} \mu_j \Bigg[K^{(n+j)}(x(t), u(t), t) \\ &\quad + \frac{\partial \Phi^{(j)}(x(T))}{\partial x} K(T, x(t), u(t), t) \Bigg] dt \\ &\quad + \int_0^T z(t) \int_0^t K(t, x(s), u(s), s)\, ds\, dt \\ &= I_0(x, u) + \sum_{j=1}^{l} \mu_j \Bigg\{ - \Phi^{(j)}(x(T)) \\ &\quad + \frac{\partial \Phi^{(j)}(x(T))}{\partial x}[x(T) - f(T)] \Bigg\} \\ &\quad + \int_0^T z(t)[x(t) - f(t)]\, dt.\end{aligned}$$

From this it follows that if $H(x(t), z, u_0(t), t) \leq H(x(t), z, u(t), t)$ for $t \in [0, T]$, then also $I_0(x, u_0) \leqq I_0(x, u)$. The theorem is proved.

Note 1. If the system (1.1) is obtained by integration of a system of differential equations, then $K(t, x, u, s)$ does not depend on t. In this case, $\psi^{(0)}(t) = -1$,

$\psi(t) = -\int_t^T z(s)\,ds$ satisfy the adjoint system of differential equations, but nevertheless, constants requiring determination enter into the function H. If in addition, $K^{(n+j)}(x, u, t) \equiv 0, j = 1, 2, \cdots, l$, then

$$\psi^{(0)}(t) = -1, \qquad \psi(t) = -\sum_{j=1}^{l} \mu_j \frac{\partial\Phi^{(j)}(x(T))}{\partial x} - \int_t^T z(s)\,ds$$

satisfy the adjoint system, and H is the corresponding Pontryagin function (Hamiltonian), taken with opposite sign. It is not necessary to formulate the jump conditions, since they enter into the integral equation for $z(t)$.

Note 2. Let B be the entire space E_n, and let $z(t)$ be the solution of the system

$$z(t) = \frac{\partial F(x(t), u(t), t)}{\partial x} + \int_t^T z(s)\frac{\partial K(s, x(t), u(t), t)}{\partial x}\,ds. \tag{2.36}$$

The next theorem follows from Theorem 2.1.

THEOREM 2.2. *Let $f(t)$, $K(t, x, u, s)$ and $\partial K(t, x, u, s)/\partial x$ be continuous for $0 \leqq s \leqq t \leqq T$, $x \in E_n$ and $u \in U$. Then in order that $x(t)$, $u(t)$ be the optimal solution of Problem* (1.1)–(1.4) *with $B = E_n$, it is necessary and sufficient that there exist a constant vector $\mu = (\mu_1, \mu_2, \cdots, \mu_l)$, such that for the optimal trajectory $x(t)$ and optimal control $u(t)$, for which the functionals* (1.4) *vanish,* (1.1), (2.3), (2.4) *and* (2.36) *are satisfied for almost all $t \in [0, T]$, and*

$$H(x(t), z, u(t), t) = \min_{v \in U} H(x(t), z, v, t). \tag{2.37}$$

This entirely agrees with the results of A. G. Butkovski [3].

REFERENCES

[1] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. S. GAMKRELIDZE AND E. F. MISCHENKO, *The Mathematical Theory of Optimal Processes*, Macmillan, New York, 1964.

[2] R. BELLMAN, *Dynamic Programming*, Princeton University Press, Princeton, 1957.

[3] A. G. BUTKOVSKI, *Theory of Optimal Control of Distributed Parameter Systems*, Izdat. Nauka, Moscow, 1965.

SIAM J. Control
Vol. 7, No. 2, May 1969

OPTIMAL CONTROL OF PROCESSES DESCRIBED BY INTEGRAL EQUATIONS. II*

V. R. VINOKUROV†

3. Let there be given a system of equations

$$x(t) = f(t) + \int_0^t A(t,s)x(s)\,ds + \int_0^t B(t,s)u(s)\,ds, \tag{3.1}$$

where $x(t)$ and $f(t)$ are n-dimensional column vectors, $u(t)$ is an r-dimensional column vector, and $A(t,s)$ and $B(t,s)$ are $n \times n$ and $n \times r$ matrices, respectively. Let there also be given a certain functional $I(x,u)$. We shall consider the following problem.

PROBLEM. Find a piecewise continuous control $u(t)$, having values for every $t \in [0, T]$ which lie in a closed bounded region $U \subset E_r$, such that the functional $I(x,u)$ is minimum, with $x(t)$ being the solution of (3.1) corresponding to this $u(t)$.

This problem has been considered in more general form by A. G. Butkovski. In the present paper, for certain forms of the functional $I(x,u)$ we give criteria for the existence and uniqueness of solutions to this problem.

We shall assume throughout that $A(t,s)$, $B(t,s)$ and $f(t)$ are continuous for $0 \leqq s \leqq t \leqq T$. Since $u(t)$ is bounded, so also is $x(t)$. We denote by G the closed bounded region in which lie the solutions of (3.1) for all possible $u(t) \in U$. The norm of a vector will be taken to be its Euclidean length, and the norm of a matrix will be taken in the sense of the norm of an operator in a Euclidean space.

Let the functional $I(x,u)$ have the form

$$I(x,u) = \int_0^T [(a(s), x(s)) + (b(s), u(s))]\,ds, \tag{3.2}$$

where $a(t)$ and $b(t)$ are respectively n-dimensional and r-dimensional row vectors, continuous for $0 \leqq t \leqq T$. In this case, for brevity, we call the preceding Problem (3.1)–(3.2). Its solution will be called the optimal control. Let $z(t)$ be the row vector which is the solution of the system

$$z(t) = a(t) + \int_t^T z(s)A(s,t)\,ds \tag{3.3}$$

and let

$$H(z,u,t) = \left(b(t) + \int_t^T z(s)B(s,t)\,ds, u\right). \tag{3.4}$$

The next theorem is from [3] and [4].

* Originally published in Izv. Vyssh. Uchebn. Zaved. Matematika, 1967, no. 8, pp. 16–23. Submitted for publication on April 28, 1966. This translation into English has been prepared by R. N. and N. B. McDonough.

Translated and printed for this Journal under a grant-in-aid from the National Science Foundation.

† Orsk, U.S.S.R.

THEOREM 3.1. *In order that $x(t)$ and $u(t)$ be the optimal solution to Problem* (3.1)–(3.2), *it is necessary and sufficient that for almost all* $t \in [0, T]$ *equations* (3.1), (3.3) *and* (3.4) *be satisfied, and*

$$H(z, u(t), t) = \min_{v \in U} H(z, v, t). \tag{3.5}$$

Now let

$$H(t) = b(t) + \int_t^T z(s)B(s, t)\, ds. \tag{3.6}$$

We make the following assumptions:

(A) The region U is a bounded convex polyhedron in E_r.

(B) The hyperplane

$$\sum_{i=1}^{r} H^{(i)}(t)v^{(i)} = 0 \tag{3.7}$$

in E_r is parallel to an edge of the polyhedron U for at most a finite number of values of t.

THEOREM 3.2. *If conditions* (A) *and* (B) *are satisfied, then there exists a unique optimal solution of Problem* (3.1)–(3.2), *for which the optimal control $u(t)$ takes values only at the vertices of the polyhedron U. The optimal control can change value only at points t such that hyperplane* (3.7) *is parallel to some edge of U.*

Proof. We shall prove first that there exists at most one solution of Problem (3.1)–(3.2) and that it necessarily satisfies the conditions stated in the theorem. Because of the linearity in u, the function (3.4) is minimum either at one of the vertices of the polyhedron U or on the entire boundary [1, p. 133]. We show that the latter is possible only for a finite number of values of t. If $H(z, v, t)$ in fact attains a minimum on a portion of the boundary of positive dimension, then it has the same value at at least two vertices of the polyhedron U. Let these vertices be v_1 and v_2. The equation of the edge passing through these vertices is $v = v_1 + (v_2 - v_1)\tau$, $0 \leqq \tau \leqq 1$. From this and from (3.6), $H(z, v, t) = (H(t), v_1) + (H(t), v_2 - v_1)\tau$. But by hypothesis, the coefficient of τ in this latter equation can vanish for at most a finite number of values of t. For the remaining values of t, the minimum on this edge is attained either for $\tau = 0$ or for $\tau = 1$. It is now also clear that switch points can only be points t for which $(H(t), v_2 - v_1) = 0$.

For the proof of the existence of optimal solutions, we first look for an optimal control in the space $L_2(0, T)$. Let $\underline{I} = \inf I(x, u)$ for $u \in U$, $u(t) \in L_2(0, T)$, where $x(t)$ is the trajectory corresponding to $u(t)$, and let $\lim_{k \to \infty} I(x_k, u_k) = I$. Since the sequence $u_k(t) \subset U$, it is weakly compact in $L_2(0, T)$ and thus weakly converges to a certain function $u(t) \in L_2(0, T)$. In view of the linearity of (3.1), $x_k(t)$ strongly converges to a certain function $x(t)$, and thus by considering the linearity of (3.2), $\lim_{k \to \infty} I(x_k, u_k) = I(x, u)$. Thus $x(t), u(t)$ is the optimal solution to the problem. According to the above, in this case $u(t)$ is piecewise constant, and the theorem is proved.

4. We shall now examine the case where the functional $I(x, u)$ is of the form

$$(4.1)\qquad I(x,u) = \int_0^T \left[\frac{1}{2}(C(x(s),s)u(s),u(s)) + (a(x(s),s),u(s)) + b(x(s),s) \right] ds.$$

where $C(x, s)$ is an $r \times r$ matrix, $a(x, s)$ is an r-dimensional vector, and $b(x, s)$ is a scalar. We assume that $C(x, s)$, $a(x, s)$ and $b(x, s)$ are twice continuously differentiable with respect to $x^{(i)}$, $i = 1, 2, \cdots, n$. The problem of determining the functions minimizing (4.1), under the constraint (3.1) and with $u \in U$, will be called Problem (3.1)–(4.1). If we assume

$$(4.2)\qquad \begin{aligned} z^{(i)}(t) &= \frac{1}{2}\left(\frac{\partial C(x(t),t)}{\partial x^{(i)}}u(t), u(t)\right) + \left(\frac{\partial a(x(t),t)}{\partial x^{(i)}}, u(t)\right) \\ &\quad + \frac{\partial b(x(t),t)}{\partial x^{(i)}} + \sum_{j=1}^{n} \int_t^T z^{(j)}(s)A_{ji}(s,t)\,ds, \quad i = 1,2,\cdots,n, \end{aligned}$$

$$(4.3)\qquad H(x,z,u,t) = \frac{1}{2}(C(x,t)u,u) + \left(a(x,t) + \int_t^T z(s)B(s,t)\,ds, u\right).$$

The next theorem is from [3] and [4].

THEOREM 4.1. *In order that* $x(t)$, $u(t)$ *be the optimal solution of Problem* (3.1)–(4.1), *it is necessary and sufficient that for the optimal trajectory* $x(t)$ *and for almost all* $t \in [0, T]$, *equations* (3.1), (4.2), (4.3) *and*

$$(4.4)\qquad H(x(t),z,u(t),t) = \min_{v \in U} H(x(t),z,v,t)$$

all be satisfied.

In the following, the closed bounded region in which $z(t)$ is contained for $u(t) \in U$, $x(t) \in G$ will be denoted G'. Let us introduce the following square matrices:

$$(4.5)\qquad \frac{\partial C(x,t)}{\partial x} = \left(\frac{\partial C(x,t)}{\partial x^{(1)}}, \frac{\partial C(x,t)}{\partial x^{(2)}}, \cdots, \frac{\partial C(x,t)}{\partial x^{(n)}}\right),$$

$$(4.6)\qquad \left(\frac{\partial^2 C(x,t)}{\partial x^2}\right)_{ij} = \frac{\partial^2 C(x,t)}{\partial x^{(i)}\partial x^{(j)}}, \qquad i,j = 1,2,\cdots,n,$$

$$(4.7)\qquad \left(\frac{\partial a(x,t)}{\partial x}\right)_{ij} = a'_{ij}(x,t) = \frac{\partial a^{(j)}(x,t)}{\partial x^{(i)}}, \qquad i = 1,2,\cdots,n, \quad j = 1,2,\cdots,r,$$

$$(4.8)\qquad \frac{\partial^2 a(x,t)}{\partial x^2} = \left(\frac{\partial a'(x,t)}{\partial x^{(1)}}, \frac{\partial a'(x,t)}{\partial x^{(2)}}, \cdots, \frac{\partial a'(x,t)}{\partial x^{(n)}}\right),$$

$$(4.9)\qquad \frac{\partial b(x,t)}{\partial x} = \left(\frac{\partial b(x,t)}{\partial x^{(1)}}, \frac{\partial b(x,t)}{\partial x^{(2)}}, \cdots, \frac{\partial b(x,t)}{\partial x^{(n)}}\right),$$

$$(4.10)\qquad \left(\frac{\partial^2 b(x,t)}{\partial x^2}\right)_{ij} = \frac{\partial^2 b(x,t)}{\partial x^{(i)}\partial x^{(j)}}, \qquad i,j = 1,2,\cdots,n.$$

A matrix A will be called *positive* and written $A \geqq 0$, if for any vector $z, (Az, z) \geqq 0$. Inequalities between matrices will be understood in this sense.

THEOREM 4.2. *If U is a closed bounded region in E, and if there exists a number $\theta > 0$ such that for all $x \in G$ and $0 \leqq t \leqq T$, $C(x, t) \geqq \theta I$, where I is the unit matrix, then there exists an optimal solution to Problem* (3.1)–(4.1) *in the space* $L_2(0, T)$.

Proof. Let $\underline{I} = \inf I(x, u)$ for $u \in U$, $u(t) \in L_2(0, T)$, where $x(t)$ is the corresponding trajectory. We take $u_k(t)$ to be a sequence such that if $x_k(t)$ is the solution of (3.1) for $u(t) = u_k(t)$, then $\lim_{k\to\infty} I(x_k, u_k) = \underline{I}$. In view of the weak compactness of $u_k(t)$, the sequence converges weakly in $L_2(0, T)$ to some function $u^*(t)$. Let us introduce the following:

$$\begin{aligned} x_0(t) &= f(t) + \int_0^t A(t, s)x_0(s)\, ds, \\ \tilde{x}_k(t) &= \int_0^t A(t, s)\tilde{x}_k(s)\, ds + \int_0^t B(t, s)u_k(s)\, ds. \end{aligned} \tag{4.11}$$

Obviously $x_k(t) = x_0(t) + \tilde{x}_k(t)$. In view of the linearity, $\tilde{x}_k(t)$ strongly converges in $L_2(0, T)$ to some function $\tilde{x}(t)$, and $\lim_{k\to\infty} x_k(t) = x_0(t) + \tilde{x}(t) = x^*(t)$. We shall prove that $x^*(t), u^*(t)$ is the optimal solution to Problem (3.1)–(4.1). Obviously if $u(t) = \lambda u_n(t) + \mu u_m(t)$, the corresponding solution of (3.1) will be $x(t) = x_0(t) + \lambda\tilde{x}_n(t) + \mu\tilde{x}_m(t)$. Using this, we have the identity (which can be checked by simple calculations):

$$\begin{aligned} &I(x_0 + \tilde{x}_n, u_n) + I(x_0 + \tilde{x}_m, u_m) - 2I\left(x_0 + \frac{\tilde{x}_n + \tilde{x}_m}{2}, \frac{u_n + u_m}{2}\right) \\ &= \frac{1}{4}\int_0^T (C(x^*(s), s)[u_n(s) - u_m(s)], [u_n(s) - u_m(s)])\, ds \\ &+ \int_0^T \left\{\frac{1}{4}\left(\left[2C(x_0(s) + \tilde{x}_n(s), s) - C\left(x_0(s) + \frac{\tilde{x}_n(s) + \tilde{x}_m(s)}{2}, s\right)\right.\right.\right. \\ &\left.\left. - C(x^*(s), s)\right] u_n(s), u_n(s)\right) \\ &+ \frac{1}{4}\left(\left[2C(x_0(s) + \tilde{x}_m(s), s)\right.\right. \\ &\left.\left. - C\left(x_0(s) + \frac{\tilde{x}_n(s) + \tilde{x}_m(s)}{2}, s\right) - C(x^*(s), s)\right] u_m(s), u_m(s)\right) \\ &+ \frac{1}{4}\left(\left[C(x^*(s), s) - C\left(x_0(s) + \frac{\tilde{x}_n(s) + \tilde{x}_m(s)}{2}, s\right)\right] u_n(s), u_n(s)\right) \\ &+ \frac{1}{4}\left(\left[C(x^*(s), s) - C\left(x_0(s) + \frac{\tilde{x}_n(s) + \tilde{x}_m(s)}{2}, s\right)\right] u_m(s), u_m(s)\right) \\ &+ \left(\left[a(x_0(s) + \tilde{x}_n(s), s) - a\left(x_0(s) + \frac{\tilde{x}_n(s) + \tilde{x}_m(s)}{2}, s\right)\right], u_n(s)\right) \end{aligned} \tag{4.12}$$

$$+\left(\left[a(x_0(s)+\tilde{x}_m(s),s)-a\left(x_0(s)+\frac{\tilde{x}_n(s)+\tilde{x}_m(s)}{2},s\right)\right],u_m(s)\right)$$

$$+\left[b(x_0(s)+\tilde{x}_n(s),s)+b(x_0(s)+\tilde{x}_m(s),s)-2b\left(x_0(s)+\frac{\tilde{x}_n(s)+\tilde{x}_m(s)}{2}\right)\right]\bigg]\Bigg\}\,ds.$$

From this, since $\underline{I}$ is the greatest lower bound of $I(x,u)$ and all terms on the right of (4.12), except the first, tend to zero as $n,m\to\infty$, we have for $n,m>N(\varepsilon)$:

$$\frac{1}{4}\int_0^T (C(x^*(s),s)[u_n(s)-u_m(s)],[u_n(s)-u_m(s)])\,ds$$

$$\leqq I(x_0+\tilde{x}_n,u_n)+I(x_0+\tilde{x}_m,u_m)-2\underline{I}+\varepsilon.$$

From this, for $n,m\to\infty$, we have:

$$\int_0^T (C(x^*(s),s)[u_n(s)-u_m(s)],[u_n(s)-u_m(s)])\,ds\to 0,$$

and thus in view of the positive definiteness of $C(x,t)$, $u_k(t)$ strongly converges in $L_2(0,T)$. This means that $\lim_{k\to\infty} I(x_k,u_k)=I(x^*,u^*)$, and the theorem is proved.

THEOREM 4.3. *For $x\in G$ and $0\leqq t\leqq T$, let there exist a number $\theta>0$ and functions $\alpha(x,t)$, $\beta(x,t)$ and $\gamma(x,t)$ such that $C(x,t)\geqq\theta I$ and*

$$0\leqq\alpha(x,t),\beta(x,t),\gamma(x,t)\leqq 1, \tag{4.13}$$

$$M_1(x,t)\equiv\begin{pmatrix}\frac{1}{2}\alpha(x,t)\frac{\partial^2 C(x,t)}{\partial x^2} & \left[\frac{\partial C(x,t)}{\partial x}\right]^* \\ \frac{\partial C(x,t)}{\partial x} & \beta(x,t)C(x,t)\end{pmatrix}\geqq 0, \tag{4.14}$$

$$M_2(x,t)\equiv\begin{pmatrix}\frac{1}{2}[1-\alpha(x,t)]\frac{\partial^2 C(x,t)}{\partial x^2} & \frac{1}{2}\left[\frac{\partial^2 a(x,t)}{\partial x^2}\right]^* \\ \frac{1}{2}\frac{\partial^2 a(x,t)}{\partial x^2} & \gamma(x,t)\frac{\partial^2 b(x,t)}{\partial x^2}\end{pmatrix}\geqq 0, \tag{4.15}$$

$$M_3(x,t)\equiv\begin{pmatrix}[1-\beta(x,t)]C(x,t) & \left[\frac{\partial a(x,t)}{\partial x}\right]^* \\ \frac{\partial a(x,t)}{\partial x} & [1-\gamma(x,t)]\frac{\partial^2 b(x,t)}{\partial x^2}\end{pmatrix}\geqq 0 \tag{4.16}$$

*(here * means transposition). Then if U is a closed bounded convex region, there exists at most one solution to Problem* (3.1)–(4.1) *in the space* $L_2(0,T)$.

Proof. Let $u_1(t)$, $u_2(t)$ be two different optimal controls in $L_2(0,T)$. In view of the convexity of U, $u(t)=\lambda u_1(t)+(1-\lambda)u_2(t)\in U$ for $0\leqq\lambda\leqq 1$. Let $x_1(t)$, $x_2(t)$ be the solutions of the second of equations (4.11) for $u(t)=u_1(t)$ and $u(t)=u_2(t)$, respectively. Then

$$x(t)=x_0(t)+\lambda x_1(t)+(1-\lambda)x_2(t) \tag{4.17}$$

is the solution of (3.1) for $u = u(t)$. We find

$$I(x, u) = \int_0^T \left\{ \frac{1}{2} \sum_{i,j=1}^{r} C_{ij}(x(s), s)[\lambda u_1^{(i)}(s) + (1 - \lambda)u_2^{(i)}(s)] \right.$$
$$\cdot [\lambda u_1^{(j)}(s) + (1 - \lambda)u_2^{(j)}(s)] + \sum_{i=1}^{r} a^{(i)}(x(s), s)[\lambda u_1^{(i)}(s) + (1 - \lambda)u_2^{(i)}(s)]$$
$$\left. + \, b(x(s), s) \right\} ds,$$

where $x(t)$ is given in (4.17). Using this, we can find the second derivative of $I(x, u)$ with respect to λ:

$$\frac{d^2 I(x, u)}{d\lambda^2} = \int_0^T \left\{ \frac{1}{2}[\alpha(x(s), s) + (1 - \alpha(x(s), s))] \sum_{l,k=1}^{n} \sum_{i,j=1}^{r} \frac{\partial^2 C_{ij}(x(s), s)}{\partial x^{(k)} \partial x^{(l)}} \right.$$
$$\cdot [x_1^{(k)}(s) - x_2^{(k)}(s)][x_1^{(l)}(s) - x_2^{(l)}(s)][\lambda u_1^{(i)}(s) + (1 - \lambda)u_2^{(i)}(s)]$$
$$\cdot [\lambda u_1^{(j)}(s) + (1 - \lambda)u_2^{(j)}(s)]$$
$$+ \, 2 \sum_{k=1}^{n} \sum_{i,j=1}^{r} \frac{\partial C_{ij}(x(s), s)}{\partial x^{(k)}}[x_1^{(k)}(s) - x_2^{(k)}(s)]$$
$$\cdot [u_1^{(i)}(s) - u_2^{(l)}(s)][\lambda u_1^{(j)}(s) + (1 - \lambda)u_2^{(j)}(s)]$$
$$+ \, [\beta(x(s), s) + (1 - \beta(x(s), s))] \sum_{i,j=1}^{r} C_{ij}(x(s), s)[u_1^{(i)}(s) - u_2^{(i)}(s)]$$
$$\cdot [u_1^{(j)}(s) - u_2^{(j)}(s)] + \sum_{k,l=1}^{n} \sum_{i=1}^{r} \frac{\partial^2 a^{(i)}(x(s), s)}{\partial x^{(k)} \partial x^{(l)}}[x_1^{(k)}(s) - x_2^{(k)}(s)] \tag{4.18}$$
$$\cdot [x_1^{(l)}(s) - x_2^{(l)}(s)][\lambda u_1^{(i)}(s) + (1 - \lambda)u_2^{(i)}(s)]$$
$$+ \, 2 \sum_{k=1}^{n} \sum_{i=1}^{r} \frac{\partial a^{(i)}(x(s), s)}{\partial x^{(k)}}[x_1^{(k)}(s) - x_2^{(k)}(s)][u_1^{(i)}(s) - u_2^{(i)}(s)]$$
$$+ \, [\gamma(x(s), s) + (1 - \gamma(x(s), s))]$$
$$\left. \cdot \sum_{k,l=1}^{n} \frac{\partial^2 b(x(s), s)}{\partial x^{(k)} \partial x^{(l)}}[x_1^{(k)}(s) - x_2^{(k)}(s)][x_1^{(l)}(s) - x_2^{(l)}(s)] \right\} ds.$$

Let us introduce the vectors

$$q_1(x, u, \lambda) = \{x_1^{(1)} - x_2^{(1)})(\lambda u_1 + (1 - \lambda)u_2), (x_1^{(1)} - x_2^{(1)})(\lambda u_1 + (1 - \lambda)u_2),$$
$$\cdots, (x_1^{(n)} - x_2^{(n)})(\lambda u_1 + (1 - \lambda)u_2)\};$$
$$q_2(x, u) = \{u_1 - u_2, x_1 - x_2\}; \tag{4.19}$$
$$q_3(x, u, \lambda) = \{q_i(x, u, \lambda), u_1 - u_2\};$$
$$q_4(x, u, \lambda) = \{q_1(x, u, \lambda), x_1 - x_2\}.$$

Using these, we can write (4.18) as

$$\begin{aligned}\frac{d^2I(x,u)}{d\lambda^2} = \int_0^T & [(M_1(x(s),s)q_3(x(s),u(s),\lambda),q_3(x(s),u(s),\lambda)) \\ & + (M_2(x(s),s)q_4(x(s),u(s),\lambda),q_4(x(s),u(s),\lambda)) \\ & + (M_3(x(s),u(s))q_2(x(s),u(s)),q_2(x(s),u(s)))]\,ds. \end{aligned} \tag{4.20}$$

In view of (4.14)–(4.16) and the positive definiteness of $C(x,t)$, $d^2I(x,u)/d\lambda^2 > 0$, and thus $I(x,u)$ cannot simultaneously be minimum for $\lambda = 0$ and $\lambda = 1$. The theorem is proved.

Note. Conditions (4.14)–(4.16) are particularly transparent in the case that the integrand in $I(x,u)$ is a quadratic form in the coordinates of the vectors x and u, with coefficients which are functions of s. In this case, the conditions are equivalent to the positive definiteness of that quadratic form.

We have already established that the solution of Problem (3.1)–(4.1) is obtained by minimization of the function (4.3). Let us write the latter in the form

$$H(\Sigma,v) = \frac{1}{2}(C(\Sigma)v,v) + (h(\Sigma),v), \tag{4.21}$$

where Σ is the collection of variables or parameters entering into the definition of the function H upon which it depends continuously (in particular, x, t, T, etc.). We shall suppose that between various sets Σ there is established a distance, which we designate for the sets Σ_1 and Σ_2 by $\rho(\Sigma_1,\Sigma_2)$. Continuity with respect to Σ will be understood in the sense of this metric.

THEOREM 4.4. *If the conditions of Theorem* 4.2 *are satisfied, then every v for which $H(\Sigma,v)$ in* (4.21) *is minimum is a continuous function of Σ. In particular, the optimal control $u(t)$ is a continuous function of t.*

Proof. We shall prove first that $\min_{v\in U} H(\Sigma,v)$ is a continuous function of Σ. Let $v(\Sigma)$ be that v which minimizes $H(\Sigma,v)$ for a given Σ. Because of the positive definiteness of $C(\Sigma)$, for each set Σ there is a unique $v(\Sigma)$. We have

$$\begin{aligned} & H(\Sigma_2,v(\Sigma_2)) - H(\Sigma_1,v(\Sigma_1)) \\ & = H(\Sigma_2,v(\Sigma_2)) - H(\Sigma_2,v(\Sigma_1)) + H(\Sigma_2,v(\Sigma_1)) - H(\Sigma_1,v(\Sigma_1)) \\ & = H(\Sigma_2,v(\Sigma_2)) - H(\Sigma_1,v(\Sigma_2)) + H(\Sigma_1,v(\Sigma_2)) - H(\Sigma_1,v(\Sigma_1)). \end{aligned} \tag{4.22}$$

From (4.22) and the minimizing property of $v(\Sigma)$, it is apparent that

$$\begin{aligned} H(\Sigma_2,v(\Sigma_2)) - H(\Sigma_1,v(\Sigma_2)) & \leqq H(\Sigma_2,v(\Sigma_2)) - H(\Sigma_1,v(\Sigma_1)) \\ & \leqq H(\Sigma_2,v(\Sigma_1)) - H(\Sigma_1,v(\Sigma_1)). \end{aligned}$$

From this, and the continuity of $H(\Sigma, v)$ with respect to Σ, $H(\Sigma, v(\Sigma))$ is continuous with respect to Σ. Further, from the identity

$$\begin{aligned} & \tfrac{1}{4}(C(\Sigma)[v(\Sigma_n) - v(\Sigma)],[v(\Sigma_n) - v(\Sigma)]) \\ & = H(\Sigma,v(\Sigma)) + H(\Sigma_n,v(\Sigma_n)) - 2H\left(\Sigma,\frac{v(\Sigma) + v(\Sigma_n)}{2}\right) \\ & \quad + \tfrac{1}{2}([C(\Sigma) - C(\Sigma_n)]v(\Sigma_n),v(\Sigma_n)) + (h(\Sigma) - h(\Sigma_n),v(\Sigma_n)), \end{aligned} \tag{4.23}$$

we obtain for $\rho(\Sigma, \Sigma_n) < \delta(\varepsilon)$ that

$$\tfrac{1}{4}(C(\Sigma)[v(\Sigma_n) - v(\Sigma)], [v(\Sigma_n) - v(\Sigma)]) \leqq H(\Sigma_n, v(\Sigma_n)) - H(\Sigma, v(\Sigma)) + \varepsilon.$$

From the continuity of $H(\Sigma, v(\Sigma))$ with respect to Σ and the positive definiteness of $C(\Sigma)$, we have that $\lim_{n\to\infty} \|v(\Sigma_n) - v(\Sigma)\| = 0$ for $\lim_{n\to\infty} \rho(\Sigma_n, \Sigma) = 0$. The theorem is proved.

THEOREM 4.5. *If the conditions of Theorem* 4.2 *are satisfied, if U is convex, and if $u(x, z, t)$ minimizes* (4.3) *for given x, z and t, then for $x_1, x_2 \in G$, $z_1, z_2 \in G'$, we have*

$$(4.24)\quad \|u(x_2, z_2, t) - u(x_1, z_1, t)\| \leqq L\left(\|x_2 - x_1\| + \int_t^T \|z_2(s) - z_1(s)\|\, ds\right).$$

Proof. The proof will be carried out in several stages.

(a) Let U be the segment $a \leqq u^{(1)} \leqq b$, $u^{(i)} = \text{const.}$, $i = 2, 3, \cdots, r$. Then it is sufficient to consider minimization of the function

$$H(x, z, v^{(1)}, t) = \tfrac{1}{2}C_{11}(x, t)[v^{(1)}]^2 + \left[b(x, t) + \int_t^T z(s)B(s, t)\, ds\right]^{(1)} v^{(1)} + \text{const.}$$

Let

$$F(x, z, t) = -\frac{\left[b(x, t) + \int_t^T z(s)B(s, t)\, ds\right]^{(1)}}{C_{11}(x, t)}.$$

The following cases are possible: (i) $F(x, z, t) < a, v^{(1)} = a$; (ii) $a \leqq F(x, z, t) \leqq b$, $v^{(1)} = F(x, z, t)$; (iii) $b < F(x, z, t)$, $v^{(1)} = b$. Examining all possible distributions of the points (x_1, z_1) and (x_2, z_2) in these three regions, we can easily be convinced that in all cases

$$|v^{(1)}(x_2, z_2, t) - v^{(1)}(x_1, z_1, t)| \leqq |F(x_2, z_2, t) - F(x_1, z_1, t)|.$$

Proof. The proof will be carried out in several stages.
Because of the existence of various derivatives and $C_{11}(x, t) \geqq \theta > 0$, $F(x, z, t)$ satisfies a Lipschitz condition, and thus the theorem is proved for this case. It is easily seen that the Lipschitz constant depends only on the norms of $b(x, t)$, $B(s, t)$, $C(x, t)$, of their derivatives with respect to x, and on θ, and therefore does not depend on the segment $[a, b]$.

(b) Let U be an arbitrary segment in E_r. This case can be converted to the preceding case by an orthogonal transformation $u = Tv$, with the matrix $T^{-1}C(x, t)T$ of the quadratic form being as before positive definite. In view of $\|T\| = 1$, the Lipschitz constant does not change.

(c) Let U be an arbitrary convex region. We join the points $u(x_2, z_2, t)$, $u(x_1, z_1, t)$ by a segment Γ. Obviously

$$\min_{v\in U} H(x_i, z_i, t) = \min_{v\in\Gamma} H(x_i, z_i, v, t), \qquad i = 1, 2,$$

and the theorem is proved.

REFERENCES

[1] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. S. GAMKRELIDZE AND E. F. MISCHENKO, *The Mathematical Theory of Optimal Processes*, Macmillan, New York, 1964.

[2] R. BELLMAN, I. GLICKSBERG AND O. GROSS, *Some aspects of the mathematical theory of control processes*, Memo R-313, RAND Corp., Santa Monica, 1958.

[3] A. G. BUTKOVSKI, *Theory of Optimal Control of Distributed Parameter Systems*, Izdat. Nauka, Moscow, 1965.

[4] V. R. VINOKUROV, *Optimal control of processes described by integral equations. I*, this Journal, 7 (1969), pp. 324–336.

SIAM J. Control
Vol. 7, No. 2, May 1969

OPTIMAL CONTROL OF PROCESSES DESCRIBED BY INTEGRAL EQUATIONS, III*

V. R. VINOKUROV†

5. Let there be given a system of equations

$$x(t) = f(t) + \int_0^t K(t, x(s), u(s), s)\, ds, \tag{5.1}$$

where f, x and K are n-dimensional column vectors and $u(t)$ is an r-dimensional column vector, all satisfying the conditions given in [6], and functionals

$$I_0(x, u) = \int_0^T K^{(0)}(x(s), u(s), s)\, ds, \tag{5.2}$$

$$I_j(x, u) = \int_0^T K^{(n+j)}(x(s), u(s), s)\, ds + \Phi^{(j)}(x(T)), \quad j = 1, 2, \cdots, l. \tag{5.3}$$

We shall examine the following problem.

Problem (5.1)–(5.3). We wish to find a piecewise continuous control $u(t)$, the value of which for each $t \in [0, T]$ lies in a closed region $U \subset E_r$, such that, when $x(t)$ is the solution of (5.1) with this $u(t)$, the functionals $I_j(x, u)$ in (5.3) vanish and the functional $I_0(x, u)$ in (5.2) is minimum.

This problem was solved in more general form by A. G. Butkovski [4] and was also examined in [6, Theorem 2.2]. In this paper, its approximate solution is considered. The notation and restrictions will be taken to be the same as in [6], [7]. Let us choose a sequence $0 = t_0 < t_1 < t_2 < \cdots < t_m = T$ such that $t_{p+1} - t_p \leqq h$, $p = 0, 1, \cdots, m - 1$, and replace the system (5.1) by the system

$$y_p = f(t_p) + \sum_{q=0}^{p-1} \gamma_{pq}(h) K(t_p, y_q, v_q, t_q), \tag{5.4}$$

and the functionals (5.2), (5.3) by the functionals

$$I_0(y, v, h) = \sum_{q=0}^{m-1} \gamma_{mq}(h) K^{(0)}(y_q, v_q, t_q), \tag{5.5}$$

$$I_j(y, v, h) = \sum_{q=0}^{m-1} \gamma_{mq}(h) K^{(n+j)}(y_q, v_q, t_q) + \Phi^{(j)}(y_m), \qquad j = 1, 2, \cdots, l, \tag{5.6}$$

where $\gamma_{pq}(h) > 0$ are coefficients depending on the choice of quadrature formula.

* Originally published in Izv. Vyssh. Uchebn. Zaved. Matematika, 1967, no. 9, pp. 16–25. Submitted for publication on April 28, 1966. This translation into English has been prepared by R. N. and N. B. McDonough.

Translated and printed for this Journal under a grant-in-aid by the National Science Foundation.

† Orsk, U.S.S.R.

PROBLEM (5.4)–(5.6). We wish to find a control $v_q \in U$ such that (5.4) is satisfied, $I_j(y, v, h) = 0$, $j = 1, 2, \cdots, l$, and $I_0(y, v, h)$ is minimum. The solution y_p, v_p of this problem will be called an optimal solution.

As shown in [4], the maximum principle in its usual form for discrete systems, generally speaking, is not valid. Nevertheless, with certain restrictions it will hold true. Let us require that the coefficients $\gamma_{pq}(h)$ be such that for $0 \leqq q < p \leqq m$,

$$0 < C_1 h \leqq \gamma_{pq}(h) \leqq C_2 h. \tag{5.7}$$

Let us introduce

$$\begin{aligned} F(y_p, v_p, t_p, h) &= K^{(0)}(y_p, v_p, t_p) \\ &\quad + \sum_{j=1}^{l} \mu_j \left[K^{(n+j)}(y_p, v_p, t_p) + \frac{\partial \Phi^{(j)}(y_m)}{\partial y} K(t_m, y_p, v_p, t_p) \right], \end{aligned} \tag{5.8}$$

$$H(y, w, v, p, h) = \gamma_{mp}(h) F(y, v, t_p, h) + \sum_{q=p+1}^{m-1} w_q \gamma_{qp}(h) K(t_q, y, v, t_p), \tag{5.9}$$

where the row vector w_p is the solution of the system

$$w_p = \gamma_{mp}(h) \frac{\partial F(y_p, v_p, t_p, h)}{\partial y} + \sum_{q=p+1}^{m-1} w_q \gamma_{qp}(h) \frac{\partial K(t_q, y_p, v_p, t_p)}{\partial y}. \tag{5.10}$$

THEOREM 5.1. *If condition* (5.7) *is satisfied and if* $0 < h \leqq h_0$, *where* h_0 *is sufficiently small, then in order that* y_p, v_p *be an optimal solution to Problem* (5.4)–(5.6), *it is necessary and sufficient that there exist a constant vector* $\mu = (\mu_1, \mu_2, \cdots, \mu_l)$ *such that for the optimal trajectory* y_p *and control* v_p, *for which functionals* (5.6) *vanish for* $0 \leqq p \leqq m - 1$, (5.4), (5.8), (5.9) *and* (5.10) *are satisfied, together with*

$$H(y_p, w, v_p, p, h) = \min_{v \in U} H(y_p, w, v, p, h). \tag{5.11}$$

Proof. Necessity. Consider the perturbed control

$$v_p^* = \begin{cases} v_p & \text{for } p \neq r, \\ v & \text{for } p = r. \end{cases} \tag{5.12}$$

If y_p is the corresponding perturbed trajectory (5.4), then for $p \geqq r$

$$\begin{aligned} y_p^* &= f(t_p) + \sum_{q=0}^{r-1} \gamma_{pq}(h) K(t_p, y_q, v_q, t_q) + \gamma_{pr}(h) K(t_p, y_r, v, t_r) \\ &\quad + \sum_{q=r+1}^{p-1} \gamma_{pq}(h) K(t_p, y_q^*, v_q, t_q). \end{aligned} \tag{5.13}$$

From (5.7), $y_p^* - y_p$ is of order $\gamma_{pr}(h)$ as $h \to 0$. Then writing $y_p^* = y_p + \gamma_{pr}(h) \delta y_p + o(h)$, we have

$$\gamma_{pr}(h)\, \delta y_p = \gamma_{pr}(h)\, \delta K(t_p, y_r, v_r, t_r) + \sum_{q=r+1}^{p-1} \gamma_{pq}(h) \frac{\partial K(t_p, y_q, v_q, t_q)}{\partial y} \gamma_{qr}(h)\, \delta y_q, \tag{5.14}$$

where $\delta K(t_p, y_r, v_r, t_r) = K(t_p, y_r, v, t_r) - K(t_p, y_r, v_r, t_r)$. Let R_{pq} be the resolvent of the matrix [5]:

$$\gamma_{pq}(h)\frac{\partial K(t_p, y_q, v_q, t_q)}{\partial y}.$$

Then from (5.14),

$$\gamma_{pr}(h)\,\delta y_p = \gamma_{pr}(h)\,\delta K(t_p, y_r, v_r, t_r) + \sum_{q=r+1}^{p-1} R_{pq}\gamma_{qr}(h)\,\delta K(t_q, y_r, v_r, t_r). \tag{5.15}$$

Now introduce

$$I(y, v, h) = I_0(y, v, h) + \sum_{j=1}^{l} \mu_j I_j(y, v, h). \tag{5.16}$$

Then analogous to (5.14),

$$\begin{aligned}\delta I(y, v, h) &= \gamma_{mr}(h)\,\delta\tilde{K}^{(0)}(y_r, v_r, t_r)\\ &\quad+ \sum_{q=r+1}^{m-1} \gamma_{mq}(h)\frac{\partial\tilde{K}^{(0)}(y_q, v_q, t_q)}{\partial y}\gamma_{qr}(h)\,\delta y_q + \sum_{j=1}^{l}\mu_j\frac{\partial\Phi^{(j)}(y_m)}{\partial y}\gamma_{mr}(h)\,\delta y_m,\end{aligned}$$

where

$$\tilde{K}^{(0)}(y_q, v_q, t_q) = K^{(0)}(y_q, v_q, t_q) + \sum_{j=1}^{l}\mu_j K^{(n+j)}(y_q, v_q, t_q).$$

Taking (5.15) into account, we find from this

$$\begin{aligned}\delta I &= \gamma_{mr}(h)\left[\delta\tilde{K}^{(0)}(y_r, v_r, t_r) + \sum_{j=1}^{l}\mu_j\frac{\partial\Phi^{(j)}(y_m)}{\partial y}\delta K(t_m, y_r, v_r, t_r)\right]\\ &\quad+ \sum_{q=r+1}^{m-1}\left[\gamma_{mq}(h)\frac{\partial\tilde{K}^{(0)}(y_q, v_q, t_q)}{\partial y} + \sum_{j=1}^{l}\mu_j\frac{\partial\Phi^{(j)}(y_m)}{\partial y}R_{mq}\right.\\ &\quad\left.+ \sum_{s=q+1}^{m-1}\gamma_{ms}(h)\frac{\partial\tilde{K}^{(0)}(y_s, v_s, t_s)}{\partial y}R_{sq}\right]\gamma_{qr}(h)\,\delta K(t_q, y_r, v_r, t_r).\end{aligned} \tag{5.17}$$

Let $z_p(h)$ be the row vector which is the solution of the system

$$z_p(h) = \gamma_{mp}(h)\frac{\partial\tilde{K}^{(0)}(y_p, v_p, t_p)}{\partial y} + \sum_{q=p+1}^{m-1} z_q(h)\gamma_{qp}(h)\frac{\partial K(t_q, y_p, v_p, t_p)}{\partial y}, \tag{5.18}$$

and let

$$w_p(h) = z_p(h) + \sum_{j=1}^{l}\mu_j\frac{\partial\Phi^{(j)}(y_m)}{\partial y}R_{mp}. \tag{5.19}$$

Then it follows from the equations for the resolvent R_{pq} (see [5]) and from (5.8), (5.18) and (5.19) that

$$\begin{aligned}&\sum_{q=p+1}^{m-1} w_q(h)\gamma_{qp}(h)\frac{\partial K(t_q, y_p, v_p, t_p)}{\partial y}\\ &= \sum_{q=p+1}^{m-1} z_q(h)\gamma_{qp}(h)\frac{\partial K(t_q,y_p,v_p,t_p)}{\partial y} + \sum_{j=1}^{l} \mu_j \frac{\partial \Phi^{(j)}(y_m)}{\partial y}\left[R_{mp} - \gamma_{mp}(h)\frac{\partial K(t_m,y_p,v_p,t_p)}{\partial y}\right]\\ &= w_p(h) - \gamma_{mp}(h)\frac{\partial F(y_p, v_p, t_p, h)}{\partial y};\end{aligned}$$

that is, $w_p(h)$ indeed satisfies (5.10). At the same time, from (5.17)–(5.19) there follows $\delta I = \delta H(y, w, v, r)$, where H is defined by (5.9). From this follows (5.11), and the necessity is proved.

The sufficiency is proved analogously to the proof of sufficiency in Theorem 2.1 of [6].

THEOREM 5.2. *With $u(t) \in U$, $v_p \in U$, let the solutions of the systems* (5.1) *and* (5.4) *lie in a certain bounded region G, and for $0 \leqq s \leqq t \leqq T$, let $K^{(j)}(t, x, u, s)$, $j = 0, 1, \cdots, n + l$, satisfy a Lipschitz condition in x and u. Further suppose that condition* (5.7) *is satisfied; and for $t_p \leqq t < t_{p+1}$, $p = 0, 1, \cdots, m, j = 0, 1, \cdots, n + l$, $x(s) \in G, u_q \in U$, let*

$$\text{(5.20)}\quad \sum_{q=0}^{p-1}\int_{t_q}^{t_{q+1}}\left|K^{(j)}(t, x(s), u_q, s) - \frac{\gamma_{pq}(h)}{t_{q+1} - t_q}K^{(j)}(t_p, x(s), u_q, t_q)\right| ds \leqq \phi(h),$$

$$\text{(5.21)}\qquad \|f(t) - f(t_p)\| \leqq \phi(h).$$

Then, where $\underline{I} = \inf I_0(x, u)$, $\underline{I}_h = \inf I_0(y, v, h)$, *there exists a constant B such that*

$$\text{(5.22)}\qquad |\underline{I} - \underline{I}_h| \leqq B\phi(h).$$

Proof. We shall first prove that for any step function control $u(t)$, such that $u(t) = u(t_p)$ for $t_p \leqq t < t_{p+1}$, where $x(t)$ is the corresponding trajectory of system (5.1), there exists a constant A such that

$$\text{(5.23)}\qquad \max \|x(t) - y_p\| \leqq A\phi(h), \qquad 0 \leqq t \leqq T, \quad t_p \leqq t < t_{p+1}.$$

In fact, taking $M_p = \max \|x(t) - y_p\|$, $t_p \leqq t < t_{p+1}$, in view of (5.20), (5.21) and the Lipschitz conditions, we have

$$M_p \leqq 2\phi(h) + L\sum_{q=0}^{p-1}\gamma_{pq}(h)M_q.$$

Since (5.7) and the condition $t_{p+1} - t_p \leqq h$ imply the boundedness of

$$\sum_{q=0}^{p-1}\gamma_{pq}(h),$$

(5.23) follows from formulas (6) and (9) of [5]. Hence analogously for this control we easily obtain

$$\text{(5.24)}\qquad |I_0(x, u) - I_0(y, u, h)| < B\phi(h).$$

By definition of the infimum and (5.24), $\underline{I} \leqq I_0(x, u) \leqq I_0(y, u, h) + B\phi(h)$. Since this holds for any step function control, we have

$$(5.25) \qquad \underline{I} \leqq \underline{I}_h + B\phi(h).$$

Let us now choose an $\varepsilon > 0$ and find a trajectory and control $x(t)$, $u(t)$ such that

$$(5.26) \qquad \underline{I} > I_0(x, u) - \varepsilon/2.$$

In view of the piecewise continuity of $u(t)$, it is possible to find a step function control $v(t)$ such that the corresponding trajectory $x^*(t)$ will satisfy $|I_0(x^*, v) - I_0(x, u)| < \varepsilon/2$. Taking into account (5.24) and (5.26), we have for the corresponding trajectory y_p,

$$(5.27) \qquad \underline{I} > I_0(y, v, h) - B\phi(h) - \varepsilon \geqq \underline{I}_h - B\phi(h) - \varepsilon.$$

Comparing (5.25) and (5.27), and taking into account the arbitrariness of ε, we obtain (5.22).

COROLLARY. *Let $y_p(h)$, $v_p(h)$ be an optimal solution of Problem* (5.4)–(5.6). *If the conditions of Theorem* 5.2 *are satisfied, with* $\lim_{h\to 0+} \phi(h) = 0$, *and if there exists a sequence $h_k \to 0$ such that* $\lim_{k\to\infty} y_p(h_k) = x(t)$, $\lim_{k\to\infty} v_p(h_k) = u(t)$ *uniformly on* $0 \leqq t \leqq T$ *then $x(t), u(t)$ is an optimal solution of Problem* (5.1)–(5.3).

6. In the case that the system (5.1) is linear and the functional (5.3) is quadratic, we have

$$(6.1) \qquad x(t) = f(t) + \int_0^t A(t, s)x(s)\,ds + \int_0^t B(t, s)u(s)\,ds,$$

where $A(t, s)$ is an $n \times n$ matrix and $B(t, s)$ is an $n \times r$ matrix, and

$$(6.2) \qquad I(x, u) = \int_0^T [\tfrac{1}{2}(C(x(s), s)u(s), u(s)) + (a(x(s), s), u(s)) + b(x(s), s)]\,ds,$$

where $C(x, s)$ is an $r \times r$ matrix, $a(x, s)$ is an r-vector, and $b(x, s)$ is a scalar. We can then consider minimizing the functional (6.2) under the condition (6.1) and with $u \in U$ (Problem (6.1)–(6.2)). We shall assume that all conditions set forth in § 4 of [7] are satisfied and shall retain the notation used there.

Let us replace Problem (6.1)–(6.2) by a discrete problem. To that end, rather than the system (6.1) we shall consider

$$(6.3) \qquad y_p = f(t_p) + \sum_{q=0}^{p-1} \gamma_{pq}(h)[A(t_p, t_q)y_q + B(t_p, t_q)v_q]$$

and replace the functional (6.2) by

$$(6.4) \qquad I(y, v, h) = \sum_{q=0}^{m-1} \gamma_{mq}(h)[\tfrac{1}{2}(C(y_q, t_q)v_q, v_q) + (a(y_q, t_q), v_q) + b(y_q, v_q)].$$

The minimization of functional (6.4) under the constraint (6.3), with $v_q \in U$, will be called Problem (6.3)–(6.4). Let

$$w_p^{(i)} = \gamma_{mp}(h)\left[\frac{1}{2}\left(\frac{\partial C(y_p, t_p)}{\partial y^{(i)}} v_p, v_p\right) + \left(\frac{\partial a(y_p, t_p)}{\partial y^{(i)}}, v_p\right) + \frac{\partial b(y_p, t_p)}{\partial y^{(i)}}\right] \tag{6.5}$$
$$+ \sum_{j=1}^{n} \sum_{q=p+1}^{m-1} w_q^{(i)} \gamma_{qp}(h) A_{ji}(t_q, t_p), \qquad i = 1, 2, \cdots, n,$$

$$H(y, w, v, p, h) = \tfrac{1}{2}\gamma_{mp}(h)(C(y, t_p)v, v \tag{6.6}$$
$$+ \left(\gamma_{mp}(h)a(y, t_p) + \sum_{q=p+1}^{m-1} w_q \gamma_{qp}(h) B(t_q, t_p), v\right).$$

Then from Theorem 5.1 there follows the next theorem.

THEOREM 6.1. *If conditions* (5.7) *are satisfied and if* $0 < h \leqq h_0$, *where* h_0 is *sufficiently small, then for* y_p, v_p *to be an optimal solution of Problem* (6.3)–(6.4), *it is necessary and sufficient that for* $0 \leqq p \leqq m - 1$ *the optimal trajectory* y_p *satisfy* (6.3), (6.5), (6.6) *and*

$$H(y_p, w, v_p, p, h) = \min_{v \in U} H(y_p, w, v, p, h). \tag{6.7}$$

The optimal control v_p has properties analogous to those for the continuous case presented in Theorems 4.2 to 4.5 of [7].

THEOREM 6.2. *Let the conditions of Theorem* 4.5 *of* [7] *and inequality* (5.7) *be satisfied, and, in addition, for* $t_p \leqq t < t_{p+1}$, $p = 0, 1, \cdots, m$, *let*

$$\sum_{q=0}^{p-1} \int_{t_q}^{t_{q+1}} \left\| A(t, s) - \frac{\gamma_{pq}(h)}{t_{q+1} - t_q} A(t_p, t_q) \right\| ds \leqq \phi(h), \tag{6.8}$$

$$\sum_{q=p+1}^{m-1} \int_{t_q}^{t_{q+1}} \left\| A(s, t) - \frac{\gamma_{mq}(h)}{t_{q+1} - t_q} \frac{\gamma_{qp}(h)}{\gamma_{mp}(h)} A(t_q, t_p) \right\| ds \leqq \phi(h), \tag{6.9}$$

$$\sum_{q=0}^{p-1} \int_{t_q}^{t_{q+1}} \left\| B(t, s) - \frac{\gamma_{pq}(h)}{t_{q+1} - t_q} B(t_p, t_q) \right\| ds \leqq \phi(h), \tag{6.10}$$

$$\sum_{q=p+1}^{m-1} \int_{t_a}^{t_{q+1}} \left\| B(s, t) - \frac{\gamma_{mq}(h)}{t_{q+1} - t_q} \frac{\gamma_{qp}(h)}{\gamma_{mp}(h)} B(t_q, t_p) \right\| ds \leqq \phi(h), \tag{6.11}$$

$$\sum_{q=0}^{m-1} \int_{t_q}^{t_{q+1}} \left\| C(y(s), s) - \frac{\gamma_{mq}(h)}{t_{q+1} - t_q} C(y(s), t_q) \right\| ds \leqq \phi(h), \tag{6.12}$$

$$\sum_{q=0}^{m-1} \int_{t_q}^{t_{q+1}} \left\| a(y(s), s) - \frac{\gamma_{mq}(h)}{t_{q+i} - t_q} a(y(s), t_q) \right\| ds \leqq \phi(h), \tag{6.13}$$

$$\sum_{q=0}^{m-1} \int_{t_q}^{t_{q+1}} \left| b(y(s), s) - \frac{\gamma_{mq}(h)}{t_{q+1} - t_q} b(y(s), t_q) \right| ds \leqq \phi(h), \tag{6.14}$$

where $\lim_{h \to 0+} \phi(h) = 0$. *Then there exists a sequence* h_k *with* $h_k \to 0$ *as* $k \to \infty$, *such that if* $y_p(h_k)$, $v_p(h_k)$ *is the optimal solution of Problem* (6.3)–(6.4) *for* $h = h_k$, *then for* $t_p \leqq t < t_{p+1}, p = 0, 1, \cdots, m,$

$$(6.15a) \qquad \lim_{k\to\infty} \|x(t) - y_p(h_k)\| = 0, \qquad \lim_{k\to\infty} \|u(t) - v_p(h_k)\| = 0,$$

$$(6.15b) \qquad \lim_{k\to\infty} |I(x, u) - I(y(h_k), v(h_k), h_k)| = 0,$$

where $x(t), u(t)$ *is the optimal solution to Problem* (6.1)–(6.2).

Proof. Let $\phi_i(h)$ denote functions tending to zero as $h \to 0$, and let C_i be constants. For $t_q \leqq s < t_{q+1}, t_p \leqq t < t_{p+1}$, let us assume that

$$y_h(t) = y_p(h), \quad f_h(t) = f(t_p), \quad w_h(t) = w_p(h),$$

$$A_h(t, s) = \frac{\gamma_{pq}(h)}{t_{q+1} - t_q} A(t_p, t_q), \quad B_h(t, s) = \frac{\gamma_{pq}(h)}{t_{q+1} - t_q} B(t_p, t_q),$$

$$v_h(t) = v_p(h), \quad \gamma_h(t) = \gamma_{mp}(h).$$

Then (6.3), (6.5) and (6.6) can be rewritten

$$(6.16) \qquad y_h(t) = f_h(t) + \int_0^{t_p} [A_h(t, s)y_h(s) + B_h(t, s)v_h(s)]\, ds,$$

$$(6.17) \qquad \begin{aligned} \frac{w_h(t)}{\gamma_h(t)} = \frac{1}{2}\left(\frac{\partial C(y_h(t), t_p)}{\partial y} v_h(t), v_h(t)\right) + \left(\frac{\partial a(y_h(t), t_p)}{\partial y}, v_h(t)\right) \\ + \frac{\partial b(y_h(t), t_p)}{\partial y} + \int_{t_{p+1}}^T \frac{w_h(s)}{\gamma_h(s)} \frac{\gamma_h(s)}{\gamma_h(t)} A_h(s, t)\, ds, \end{aligned}$$

$$(6.18) \qquad \begin{aligned} \frac{H[y_h, (w_h/\gamma_h), v_h, t_p, h]}{\gamma_h} &= \tfrac{1}{2}(C(y_h, t_p)v_h, v_h) \\ &+ \left(a(y_h, t_p) + \int_{t_{p+1}}^T \frac{w_h(s)}{\gamma_h(s)} \frac{\gamma_h(s)}{\gamma_h(t)} B_h(s, t)\, ds, v_h\right). \end{aligned}$$

In formula (6.18), taken for $h = h_1$, let us replace y_{h_1} and $w_{h_1}(s)/\gamma_{h_1}(s)$ by y_{h_2} and $w_{h_2}(s)/\gamma_{h_2}(s)$, respectively, and let $v(t)$ yield the minimum of the function thus obtained. Then, analogous to Theorem 4.5 of [7], for the discrete case

$$\|v_{h_1}(t) - \tilde{v}(t)\| \leqq L\left(\|y_{h_1} - y_{h_2}\| + \int_{t_{p+1}}^T \left\|\frac{w_{h_1}(s)}{\gamma_{h_1}(s)} - \frac{w_{h_2}(s)}{\gamma_{h_2}(s)}\right\| ds\right),$$

where t_{p+1} is taken to be the smaller of t_{p+1} for $h = h_1, h = h_2$. Because of (6.9) the coefficients of the function thus obtained differ from the coefficients of (6.18) taken for $h = h_2$ by amounts which tend to zero as $\max(h_1, h_2) \to 0$. Thus, in analogy to Theorem 4.4, for the discrete case,

$$(6.19) \qquad \begin{aligned} \|v_{h_1}(t) - v_{h_2}(t)\| &\leqq \|v_{h_1}(t) - \tilde{v}(t)\| + \|\tilde{v}(t) - v_{h_2}(t)\| \\ &\leqq L\left(\|y_{h_1}(t) - y_{h_2}(t)\| + \int_{t_{p+1}}^T \left\|\frac{w_{h_1}(s)}{\gamma_{h_1}(s)} - \frac{w_{h_2}(s)}{\gamma_{h_2}(s)}\right\| ds\right) + \phi_1(h). \end{aligned}$$

Taking into account the existence of the second derivatives of $C(x, t)$, $a(x, t)$ and $b(x, t)$ with respect to x, from (6.17) and (6.9) there follows

$$\left\| \frac{w_{h_1}(t)}{\gamma_{h_1}(t)} - \frac{w_{h_2}(t)}{\gamma_{h_2}(t)} \right\| \leqq C_1 \|y_{h_1}(t) - y_{h_2}(t)\| + C_2 \|v_{h_1}(t) - v_{h_2}(t)\| + C_3 \int_{t_{p+1}}^{T} \left\| \frac{w_{h_1}(s)}{\gamma_{h_1}(s)} - \frac{w_{h_2}(s)}{\gamma_{h_2}(s)} \right\| ds + \phi_2(h).$$

From this

$$\left\| \frac{w_{h_1}(t)}{\gamma_{h_1}(t)} - \frac{w_{h_2}(t)}{\gamma_{h_2}(t)} \right\| \leqq C_4 \|y_{h_1}(t) - y_{h_2}(t)\| + C_5 \|v_{h_1}(t) - v_{h_2}(t)\| + C_6 \int_{t_{p+1}}^{T} (\|y_{h_1}(s) - y_{h_2}(s)\| + \|v_{h_1}(s) - v_{h_2}(s)\|) \, ds + \phi_3(h). \tag{6.20}$$

From (6.19) and (6.20) there follows

$$\|v_{h_1}(t) - v_{h_2}(t)\| \leqq C_7 \left(\|y_{h_1}(t) - y_{h_2}(t)\| + \int_{t_{p+1}}^{T} \|y_{h_1}(s) - y_{h_2}(s)\| \, ds \right) + \phi_4(h). \tag{6.21}$$

Now let $x_h(t)$ be the solution of (6.1) for $u(t) = v_h(t)$. It is easy to show that the set of functions $x_h(t)$ is uniformly bounded and equicontinuous, and thus, according to Arzela's theorem, it is possible to choose a sequence $h_k \to 0$ such that $x_{h_k}(t) \to x(t)$ uniformly on the interval $[0, T]$. In view of (5.23), $y_{h_k}(t)$ converges uniformly to the same function. Thus, according to (6.21), $v_{h_k}(t)$ is a fundamental sequence converging to a certain function $u(t)$. Conditions (6.6), (6.8), (6.10), (6.12) and (6.13) assure the applicability of Theorem 5.2. From the corollary to this theorem, $x(t)$, $u(t)$ is an optimal solution to Problem (6.1)–(6.2), and (6.15) is a corollary to that theorem.

7. In the case that not only the system of equations but the functional, as well, is linear, it is possible to develop an even more general method for finding an approximate solution of the optimization problem. Thus let it be required to minimize the functional

$$I(x, u) = \int_0^T [(a(s), x(s)) + (b(s), u(s))] \, ds, \tag{7.1}$$

where $a(s)$, $b(s)$ are n- and r-dimensional vectors respectively (Problem (6.1)–(7.1)).

Let us choose a sequence $0 = t_0 < t_1 < \cdots < t_m = T$, with $t_{p+1} - t_p \leqq h$, and replace the integrals in (3.1) to (3.4) of [7] by sums derived from some quadrature formulas (not necessarily the same for each integral). We shall obtain as the result of this

$$y_p = f(t_p) + \sum_{q=0}^{p-1} \gamma_{pq}^{(1)}(h) A(t_p, t_q) y_q + \sum_{q=0}^{p-1} \gamma_{pq}^{(2)}(h) B(t_p, t_q) v_q, \tag{7.2}$$

$$I(y, v, h) = \sum_{q=0}^{p-1} \gamma_{mq}^{(3)}(h)(a(t_q), y_q) + \sum_{q=0}^{p-1} \gamma_{mq}^{(4)}(h)(b(t_q), v_q), \tag{7.3}$$

$$w_p = a(t_p) + \sum_{q=p+1}^{m-1} w_q \gamma_{qp}^{(5)}(h) A(t_q, t_p), \tag{7.4}$$

$$H(w, v, t_p, h) = \left(b(t_p) + \sum_{q=p+1}^{m-1} w_q \gamma_{qp}^{(6)}(h) B(t_q, t_p), v_p\right). \tag{7.5}$$

Let the vector v_p satisfy the conditions

$$H(w, v_p, t_p, h) = \min_{v \in U} H(w, v, t_p, h) \tag{7.6}$$

and

$$H(t_p, h) = b(t_p) + \sum_{q=p+1}^{m-1} w_q \gamma_{qp}^{(6)} B(t_q, t_p). \tag{7.7}$$

THEOREM 7.1. *Let conditions* (A) *and* (B) *of* [7] *and equations* (7.2) to (7.6) *all be satisfied. Also, for* $0 \leqq q < p \leqq m$, *let the coefficients* $\gamma_{pq}^{(i)}$ *satisfy* $\gamma_{pq}^{(i)}(h) > 0$, $i = 1, 2, \cdots, 6$, *as well as the conditions*

$$\sum_{q=0}^{p-1} \gamma_{pq}^{(i)}(h) \leqq A, \quad i = 1, 2, 3, 4, \qquad \sum_{q=p+1}^{m-1} \gamma_{qp}^{(i)}(h) \leqq A, \quad i = 5, 6, \tag{7.8}$$

$$\sum_{q=0}^{p-1} \int_{t_q}^{t_{q+1}} \left\| A(t, s) - \frac{\gamma_{pq}^{(1)}(h)}{t_{q+1} - t_q} A(t_p, t_q) \right\| ds \leqq \phi(h), \tag{7.9}$$

$$\sum_{q=0}^{p-1} \int_{t_q}^{t_{q+1}} \left\| B(t, s) - \frac{\gamma_{pq}^{(2)}(h)}{t_{q+1} - t_q} B(t_p, t_q) \right\| ds \leqq \phi(h), \tag{7.10}$$

$$\sum_{q=0}^{m-1} \int_{t_q}^{t_{q+1}} \left\| a(s) - \frac{\gamma_{mq}^{(3)}(h)}{t_{q+1} - t_q} a(t_q) \right\| ds \leqq \phi(h), \tag{7.11}$$

$$\sum_{q=0}^{m-1} \int_{t_q}^{t_{q+1}} \left\| b(s) - \frac{\gamma_{mq}^{(4)}(h)}{t_{q+1} - t_q} b(t_q) \right\| ds \leqq \phi(h), \tag{7.12}$$

$$\sum_{q=p+1}^{m-1} \int_{t_q}^{t_{q+1}} \left\| A(s, t) - \frac{\gamma_{qp}^{(5)}(h)}{t_{q+1} - t_q} A(t_q, t_p) \right\| ds \leqq \phi(h), \tag{7.13}$$

$$\sum_{q=p+1}^{m-1} \int_{t_q}^{t_{q+1}} \left\| B(s, t) - \frac{\gamma_{qp}^{(6)}(h)}{t_{q+1} - t_q} B(t_q, t_p) \right\| ds \leqq \phi(h), \tag{7.14}$$

where $\phi(h) \to 0$ *as* $h \to 0$. *Then if* $x(t), u(t)$ *is the optimal solution to Problem* (6.1)–(7.1), $\max \|x(t) - y_p\| \to 0$ *uniformly as* $h \to 0$ *for*

$$t_p \leqq t < t_{p+1}, \qquad p = 0, 1, \cdots, m-1, \tag{7.15}$$

and

$$\lim_{h \to 0} |I(x, u) - I(y, v, h)| = 0. \tag{7.16}$$

Proof. Let us denote by $\phi_i(h)$ functions which tend to zero as $h \to 0$. Subtracting (7.4) from (3.3) in [7], and taking into account (7.8), (7.13) and the continuity of $a(t)$, we find that for $t_p \leqq t < t_{p+1}$, $p = 0, 1, \cdots, m - 1$, $\max \|z(t) - w_p\| \leqq \phi_1(h)$. From this, in view of the continuity of $b(t)$, taking into account (7.8) and (7.14) we obtain $\max \|H(t) - H(t_p, h)\| \leqq \phi_2(h)$, where the maximum is taken with the same conditions. Let us divide the interval $[0, T]$ into two parts $\Delta_1(\delta)$ and $\Delta_2(\delta)$. In $\Delta_1(\delta)$ we shall include those points at which the minimum angle between the hyperplane (3.7) in [7] and the edges of the polyhedron U is no less than $\delta/2$, with the remaining points being assigned to $\Delta_2(\delta)$. It is easy to prove that for any $\varepsilon > 0$ there can be found $\delta_0 > 0$ such that for $\delta \leqq \delta_0$, the measure of $\Delta_2(\delta)$ is less than ε. For if the contrary were the case, we could let δ tend to zero, obtaining in the limit a set of positive measure at the points of which the hyperplane (3.7) of [7] would be parallel to some edge of U, which would contradict condition (B) of [7]. Now let us choose an $h(\delta)$ such that for $h \leqq h(\delta)$ and $t_p \in \Delta_1(\delta)$, the minimum angle between the hyperplane

$$\sum_{i=1}^{r} H^{(i)}(t_p, h)v^{(i)} = 0$$

and the edges of U is not less than $\delta/2$. Then in view of (3.5) of [7] and (7.6), $u(t) = v_p$ on $\Delta_1(\delta)$, for $t_p \leqq t < t_{p+1}$. From this, since the measure of $\Delta_2(\delta)$ is less than ε and conditions (7.8) to (7.10) are satisfied, (7.15) is proved. Equation (7.16) follows analogously from (7.8), (7.11) and (7.12).

REFERENCES

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

[2] R. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, 1957.

[3] R. Bellman, I. Glicksberg and O. Gross, *Some aspects of the mathematical theory of control processes*, Memo R-313, RAND Corp., Santa Monica, 1958.

[4] A. G. Butkovskii, *The Theory of Optimal Control of Distributed Parameter Systems*, Izdat. Nauka, Moscow, 1965.

[5] V. R. Vinokurov, *On the stability of the solution of an infinite system of algebraic equations arising in the approximation of linear Volterra integral equations*, Izv. Vyssh. Uchebn. Zaved. Matematika, 4 (1963), pp. 33–43.

[6] ———, *Optimal control of processes described by integral equations. I*, this Journal, 7 (1969), pp. 324–336.

[7] ———, *Optimal control of processes described by integral equations. II*, this Journal, 7 (1969), pp. 337–345.

SIAM J. CONTROL
Vol. 7, No. 2, May 1969

L^p-STABILITY $(1 \leqq p \leqq \infty)$ OF NONLINEAR TIME-VARYING FEEDBACK SYSTEMS*

M. Y. WU AND C. A. DESOER†

1. Introduction. In the past few years the L^2-stability [1], [2] and the L^∞-stability [3], [4], [5] of certain classes of nonlinear and time-varying feedback systems have been extensively studied. However, no general L^p-stability results valid for any p in $[1, \infty]$ are known. Recently Desoer and Wu [6], [7], [8] have derived L^p-stability conditions for a very broad class of linear time-invariant systems whose impulse responses may include an integrator and an infinite sequence of impulses. Chen [9] has considered the L^p-stability for a class of linear time-varying systems. In this paper the transformation technique, the small gain theorem and some results of Desoer and Wu [6], [7], [8] are used to derive the L^p-stability for a class of nonlinear time-varying systems. As an application, L^p-stability conditions for the damped Mathieu equation are obtained and compared with recent zero-input results.

To save space, we shall derive only the stability results for the multiple-input, multiple-output case. The results for the scalar case will then be stated as a corollary.

2. Notations. In this paper we shall encounter real numbers, vectors (in R^n) and elements of function spaces. Lower-case boldface (e.g., **e**, **u**) denotes vectors and upper-case boldface (e.g., **K**, **G**) denotes matrices; R_+ denotes the set of nonnegative real numbers. The symbol $|\cdot|$ is used to denote both the magnitude of a real number and the norm of a vector in R^n. For function spaces, we use the following norms: let $\mathbf{x}: R_+ \to R^n$, then, by definition,

$$\|\mathbf{x}\|_p \triangleq \left[\int_0^\infty |\mathbf{x}(t)|^p \, dt\right]^{1/p}, \qquad 1 \leqq p < \infty,$$

and

$$\|\mathbf{x}\|_\infty \triangleq \operatorname*{ess\,sup}_{t \geqq 0} |\mathbf{x}(t)| < \infty, \qquad p = \infty.$$

The resulting normed spaces are denoted by L_n^p, $1 \leqq p \leqq \infty$. If $n = 1$ (scalar case), we write L^p. When the symbols $|\cdot|$ and $\|\cdot\|$ are applied to a matrix or a matrix-valued function, they denote the induced operator norms. Note that the norms

* Received by the editors August 13, 1968, and in revised form November 22, 1968.

† Department of Electrical Engineering and Computer Sciences and the Electronics Research Laboratory, University of California, Berkeley, California 94720. This research was sponsored by the National Aeronautics and Space Administration under Grant NsG-354 (Suppl. 5) and the National Science Foundation under Grant GK-2277.

defined above are valid independently of the choice of norm in R^n because all norms in R^n are equivalent.

Following Zames [2], the space L^p_{ne}, the extension of L^p_n space, is defined as follows:

$$L^p_{ne} \triangleq \left\{ \mathbf{x}(\cdot) \,\middle|\, \int_0^T |\mathbf{x}(t)|^p \, dt < \infty, \quad \forall T \in [0, \infty), \quad 1 \leqq p < \infty \right\}$$

and

$$L^\infty_{ne} \triangleq \{\mathbf{x}(\cdot) | \operatorname*{ess\,sup}_{t \in [0,T]} |\mathbf{x}(t)| < \infty, \quad \forall T \in [0, \infty)\}.$$

Roughly speaking, if $\mathbf{x} \in L^\infty_{ne}$, then $\mathbf{x}$ does not have a finite escape time.

In order to allow us to consider a larger class of linear subsystems whose impulse responses may include an infinite sequence of impulses, we introduce the Banach algebra $\mathscr{A}_n$ (see [6], [7], [8]): Let $\mathbf{f}$ be a distribution whose support is in $[0, \infty)$. We say that $\mathbf{f}$ is an element of $\mathscr{A}_n$ if

$$\mathbf{f}(t) = \mathbf{f}_a(t) + \sum_{i=0}^{\infty} \mathbf{f}_i \delta(t - t_i),$$

where $\mathbf{f}_a\colon [0, \infty) \to R^n$ is in L^1_n, the sequence $\{t_i\}_0^\infty$ is in $[0, \infty)$ with $0 = t_0 < t_1 < t_2 < \cdots$, $\{\mathbf{f}_i\}$ is a sequence of constant vectors in R^n subject to $\sum_{i=0}^\infty |\mathbf{f}_i| < \infty$ and δ is the Dirac "function." The set of all elements in $\mathscr{A}_n$ constitutes a commutative Banach algebra with the usual definition for addition, the product defined by convolution, and the norm defined by

$$\|\mathbf{f}\| \triangleq \int_0^\infty |\mathbf{f}_a(t)| \, dt + \sum_{i=0}^{\infty} |\mathbf{f}_i|.$$

These facts are well known [10], [11]. Similarly, we say that an $n \times n$ matrix-valued distribution $\mathbf{F}$ is in $\mathscr{A}_{n \times n}$ whenever each of its column vectors is in $\mathscr{A}_n$. If $n = 1$, we write $\mathscr{A}$.

The symbol "$\hat{}$" over a function, such as $\hat{\mathbf{f}}$, denotes the Laplace transform of $\mathbf{f}$; it is defined by

$$\hat{\mathbf{f}}(s) \triangleq \int_0^\infty \mathbf{f}(t) \varepsilon^{-st} \, dt.$$

For distributions, it is defined according to L. Schwartz [15] or, by using Stieltjes integrals, according to Widder [16].

The subscript T, as in $\mathbf{f}_T$, denotes the truncation of the function $\mathbf{f}$ at time T, namely,

$$\mathbf{f}_T(t) = \begin{cases} \mathbf{f}(t) & \text{for} \quad 0 \leqq t \leqq T, \\ \mathbf{0} & \text{for} \quad t > T. \end{cases}$$

3. System descriptions. In this paper we shall consider the multiple-input, multiple-output, nonlinear, time-varying system $\mathbf{S}$ as shown in Fig. 1. The vectors

$\mathbf{u}_1(t)$, $\mathbf{u}_2(t)$, $\mathbf{e}_1(t)$, $\mathbf{e}_2(t)$, $\mathbf{y}_1(t)$ and $\mathbf{y}_2(t)$ belong to R^n. The block labeled $\mathbf{G}$ is a linear, time-invariant, nonanticipative subsystem whose input-output relation is defined in terms of its impulse response matrix $\mathbf{G}$ by the convolution integral

$$(1) \qquad \mathbf{y}_1(t) \triangleq (\mathbf{G} * \mathbf{e}_1)(t) = \int_{-\infty}^{\infty} \mathbf{G}(t - \tau)\mathbf{e}_1(\tau)\, d\tau.$$

The block labeled $\mathbf{\Phi}_t$ is a memoryless, time-varying nonlinearity whose input-output relation is defined in terms of a nonlinear function $\boldsymbol{\varphi}: R^n \times R_+ \to R^n$ by

$$(2) \qquad \mathbf{y}_2(t) = \boldsymbol{\varphi}[\mathbf{e}_2(t), t].$$

The system equations (see Fig. 1) are (1), (2) and

$$(3) \qquad \mathbf{e}_1 = \mathbf{u}_1 - \mathbf{y}_2,$$

$$(4) \qquad \mathbf{e}_2 = \mathbf{u}_2 + \mathbf{y}_1.$$

In the analysis we consider only the behavior of the system for $t \geqq 0$; therefore we take $\mathbf{u}_1(t)$, $\mathbf{u}_2(t)$, $\mathbf{e}_1(t)$, $\mathbf{e}_2(t)$, $\mathbf{y}_1(t)$ and $\mathbf{y}_2(t)$ to be zero for $t < 0$. The inputs $\mathbf{u}_1(\cdot)$ and $\mathbf{u}_2(\cdot)$ take into account the effects of the outside disturbances and the effects of the initial condition at $t = 0$.

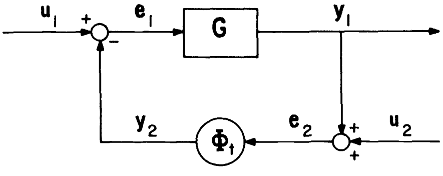


FIG. 1

4. Main results.

THEOREM 1. *Consider the system* S (*Fig.* 1). *Let the input-output relation of the linear time-invariant subsystem* $\mathbf{G}$ *be given by* (1), *where the open-loop impulse response matrix* $\mathbf{G}$ *is of the form*

$$(5) \qquad \mathbf{G}(t) = \begin{cases} \mathbf{R} + \mathbf{G}_\ell(t) & \text{for } t \geqq 0, \\ \mathbf{O} & \text{for } t < 0, \end{cases}$$

with $\mathbf{R}$ *being an* $n \times n$ *constant real matrix and* $\mathbf{G}_\ell \in \mathscr{A}_{n\times n}$. *Let* $\mathbf{\Phi}_t$ *be the time-varying nonlinearity whose characteristic* $\boldsymbol{\varphi}(\cdot, \cdot)$ *has the following properties*:

N1. $\boldsymbol{\varphi}(\cdot, \cdot)$: $R^n \times R_+ \to R^n$ *and* $\boldsymbol{\varphi}$ *is a continuous function with respect to its first argument and is a regulated function*[1] [12] *with respect to its second argument.*

[1] $\Phi(\mathbf{x}, t): R^n \times R_+ \to R^n$ is called a *regulated function* if for fixed $\mathbf{x} \in R^n$, $\varphi(\mathbf{x}, t)$ has (finite) one-sided limits at every $t \in R_+$.

N2. *There is an* $n \times n$ *constant real matrix* $\mathbf{K}$ *and a positive real number* λ *such that*

$$|\boldsymbol{\varphi}(\boldsymbol{\sigma}, t) - \mathbf{K}\boldsymbol{\sigma}| \leqq \lambda|\boldsymbol{\sigma}| \tag{6}$$

for all $t \in R_+$ *and for all* $\boldsymbol{\sigma} \in R^n$.

For some p *in* $[1, \infty]$, *let the inputs* $\mathbf{u}_1$, $\mathbf{u}_2$ *be in* L_n^p *and, for all such inputs, the corresponding* $\mathbf{e}_1$, $\mathbf{e}_2$ *be in* L_{ne}^p. *Under these conditions, if*

$$\inf_{\operatorname{Re} s \geqq 0} |\det[\mathbf{I} + \hat{\mathbf{G}}(s)\mathbf{K}]| > 0 \tag{7}$$

and if either $\mathbf{R} = \mathbf{O}$ *or all the eigenvalues of* $\mathbf{RK}$ *are in the open right half-plane,*

$$\|\mathbf{H_K}\|\lambda < 1, \tag{8}$$

where

$$\mathbf{H_K}(t) \triangleq \mathscr{L}^{-1}\{[\mathbf{I} + \hat{\mathbf{G}}(s)\mathbf{K}]^{-1}\hat{\mathbf{G}}(s)\}, \tag{9}$$

then $\mathbf{e}_1 . \mathbf{e}_2$, $\mathbf{y}_1$ *and* $\mathbf{y}_2$ *are in* L_n^p.

As a special case of Theorem 1, we state the results for the scalar case as Corollary 1.

COROLLARY 1. *Consider the system* S (Fig. 1). *Let the input-output relation of the linear time-invariant subsystem* G *be defined in terms of its open-loop impulse response* g *by*

$$y_1(t) = (g * e_1)(t) \triangleq \int_{-\infty}^{\infty} g(t - \tau)e_1(\tau)\,d\tau, \tag{10}$$

where g *is of the form*

$$g(t) = \begin{cases} r + g_\ell(t) & \text{for } t \geqq 0, \\ 0 & \text{for } t < 0, \end{cases} \tag{11}$$

with r *being a nonnegative constant and* $g_\ell \in \mathscr{A}$. *Let* Φ_t *be the time-varying nonlinearity whose characteristic* $\varphi(\cdot, \cdot)$ *has the following properties*:

N1. $\varphi(\cdot, \cdot): R \times R_+ \to R$ *and* φ *is a continuous function with respect to its first argument and is a regulated function with respect to its second argument.*

N2. *For some finite constants* k_1 *and* k_2

$$k_1\sigma^2 \leqq \sigma\varphi(\sigma, t) \leqq k_2\sigma^2 \tag{12}$$

for all $t \in R_+$ *and* $\sigma \in R$. *For some* p *in* $[1, \infty]$, *let the inputs* u_1, u_2 *be* L^p *and, for all such inputs, the corresponding* e_1, e_2 *be in* L_e^p. *Under these conditions, if for some constant* $k \in [k_1, k_2]$ *with* $kr > 0$

$$\inf_{\operatorname{Re} s \geqq 0} |1 + k\hat{g}(s)| > 0, \tag{13}$$

$$\|h_k\|\lambda < 1, \tag{14}$$

where

$$h_k(t) \triangleq \mathscr{L}^{-1}\{[1 + k\hat{g}(s)]^{-1}\hat{g}(s)\} \tag{15}$$

and

$$(16) \qquad \lambda \triangleq \max\{|k_2 - k|, |k_1 - k|\},$$

then e_1, e_2, y_1 *and* y_2 *are in* L^p.

Remark 1. If, in Corollary 1, we take $k = (k_1 + k_2)/2$, then N2 may be written as $|\varphi(\sigma, t) - k\sigma| \leqq \lambda|\sigma|$, with $\lambda = |k_2 - k_1|/2$, which is the direct specialization of **N2** to the scalar case.

Remark 2. In Corollary 1, since both h_k and λ depend on k, there may exist an optimal choice of k such that (14) will be satisfied for the largest class of the linear subsystems.

Remark 3. If the assumptions of Corollary 1 are specialized to $g \in L^1$ and $p = \infty$, then Corollary 1 reduces to a result of Sandberg [3].

5. Proof. To prove Theorem 1, we need the following two lemmas.

LEMMA 1 (see [7], [8]). *Consider the system* **S** (Fig. 1), *where for all* $\mathbf{e}_2 \in R^n$, *all* $t \in R_+$, $\boldsymbol{\varphi}(\mathbf{e}_2, t) = \mathbf{K}\mathbf{e}_2$ *with* **K** *being an* $n \times n$ *constant real matrix. Let the open-loop impulse response matrix* **G** *be defined by* (5). *Under these conditions, if*

$$\inf_{\mathrm{Re}\, s \geqq 0} |\det[\mathbf{I} + \hat{\mathbf{G}}(s)\mathbf{K}]| > 0$$

and if either $\mathbf{R} = \mathbf{O}$ *or all the eigenvalues of* **RK** *are in the open right half-plane, then* $(\mathbf{I} + \mathbf{GK})^{-1}$ *is a well-defined nonanticipative operator in* $\mathscr{A}_{n \times n}$. *Furthermore, the closed-loop impulse response matrix* $\mathbf{H_K}$ *of the system is also in* $\mathscr{A}_{n \times n}$, *where* $\mathbf{H_K}(\cdot)$ *is defined by* (9).

LEMMA 2 (Small gain theorem). *Consider a more general system than the one shown in Fig.* 1, *in that* **G** *and* $\mathbf{\Phi}_t$ *are replaced by* $\mathbf{H}_1$ *and* $\mathbf{H}_2$ *respectively. Let* $\mathbf{H}_1$ *and* $\mathbf{H}_2$ *be nonanticipative maps of* L^p_{ne} *into* L^p_{ne}, *for some fixed* $p \in [1, \infty]$. *Let* $\mathbf{H}_1$ *be linear. Let* $\mathbf{e}_1$ *and* $\mathbf{e}_2$ *be in* L^p_{ne}, *and* $\mathbf{u}_1, \mathbf{u}_2$ *be defined by the system equations. Under these conditions, if*

(a) *for some* $n \times n$ *constant real matrix* **K**, $(\mathbf{I} + \mathbf{H}_1\mathbf{K})^{-1}$ *maps* L^p_{ne} *into* L^p_{ne} *and is nonanticipative,*

(b) *there exists some positive real number* λ *such that*

$$\|(\mathbf{H}_2\mathbf{e}_2)_T - \mathbf{K}\mathbf{e}_{2T}\|_p \leqq \lambda\|\mathbf{e}_{2T}\|_p \qquad \textit{for all } T \in [0, \infty), \quad \mathbf{e}_2 \in L^p_{ne},$$

$$\text{(c)} \qquad \|(\mathbf{I} + \mathbf{H}_1\mathbf{K})^{-1}\mathbf{H}_1\| < \infty,$$

$$\text{(d)} \qquad \gamma = \|(\mathbf{I} + \mathbf{H}_1\mathbf{K})^{-1}\mathbf{H}_1\|\lambda < 1,$$

then

$$\|\mathbf{e}_{2T}\|_p \leqq (1 - \gamma)^{-1}[\|(\mathbf{I} + \mathbf{H}_1\mathbf{K})^{-1}\mathbf{u}_{2T}\|_p + \|(\mathbf{I} + \mathbf{H}_1\mathbf{K})^{-1}\mathbf{H}_1\mathbf{u}_{1T}\|_p].$$

In particular, if $\mathbf{u}_1, \mathbf{u}_2 \in L^p_n$, *then* $\mathbf{e}_2 \in L^p_n$.

Lemma 2 is a slight modification of the results in [13], [14].

Proof of Theorem 1. Let **K** be an $n \times n$ constant real matrix. Make the system transformation such that the block in the forward path becomes

$$(17) \qquad \mathbf{H_K} \triangleq (\mathbf{I} + \mathbf{GK})^{-1}\mathbf{G}$$

and the block in the feedback path becomes

$$\bar{\mathbf{\Phi}}_t \triangleq \mathbf{\Phi}_t - \mathbf{KI}. \tag{18}$$

Let $\hat{\mathbf{H}}_{\mathbf{K}}(s)$ be the transfer matrix of $\mathbf{H}_{\mathbf{K}}$; then

$$\hat{\mathbf{H}}_{\mathbf{K}}(s) = [\mathbf{I} + \hat{\mathbf{G}}(s)\mathbf{K}]^{-1}\hat{\mathbf{G}}(s). \tag{19}$$

By assumption (7) of the theorem and Lemma 1 we see that $(\mathbf{I} + \mathbf{GK})^{-1}$ is a well-defined nonanticipative operator in $\mathscr{A}_{n\times n}$. Furthermore, $\mathbf{H}_{\mathbf{K}}(\cdot)$, defined by (9), is also in $\mathscr{A}_{n\times n}$. Therefore, $\mathbf{H}_{\mathbf{K}}$ is of the form

$$\mathbf{H}_{\mathbf{K}}(t) = \begin{cases} \mathbf{H}_a(t) + \sum_{j=0}^{\infty} \mathbf{H}_j\delta(t - t_j) & \text{for } t \geqq 0, \\ \mathbf{O} & \text{for } t < 0, \end{cases} \tag{20}$$

where $\mathbf{H}_a$ has all its column vectors in L_n^1 and the $\mathbf{H}_j$'s are constant matrices such that $\sum_{j=0}^{\infty}|\mathbf{H}_j| < \infty$ and $0 = t_0 < t_1 < t_2 < \cdots$. Also $\mathbf{H}_{\mathbf{K}}$ has a well-defined norm

$$\|\mathbf{H}_{\mathbf{K}}\| \triangleq \int_0^{\infty} |\mathbf{H}_a(t)|\,dt + \sum_{j=0}^{\infty} |\mathbf{H}_j|. \tag{21}$$

Note that $\|\mathbf{H}_{\mathbf{K}}\|$ is the induced operator norm when $p = \infty$ and is an upper bound on the induced operator norm when $p \neq \infty$. By assumption **N2**, we have

$$\|\boldsymbol{\varphi}(\mathbf{e}_2, t)_T - \mathbf{Ke}_{2T}\|_p \leqq \lambda\|\mathbf{e}_{2T}\|_p, \qquad \text{for all } T \in [0, \infty), \quad \mathbf{e}_2 \in L_{ne}^p.$$

Finally, by assumption, $\|\mathbf{H}_{\mathbf{K}}\|\lambda < 1$. Thus we see that all conditions of Lemma 2 are satisfied and hence it follows that $\mathbf{e}_2 \in L_n^p$, $1 \leqq p \leqq \infty$. Since $\mathbf{y}_2 = \boldsymbol{\varphi}(\mathbf{e}_2, t)$ and

$$\|\boldsymbol{\varphi}(\mathbf{e}_2, t)\|_p - \|\mathbf{Ke}_2\|_p \leqq \|\boldsymbol{\varphi}(\mathbf{e}_2, t) - \mathbf{Ke}_2\|_p \leqq \lambda\|\mathbf{e}_2\|_p,$$

it follows that $\mathbf{y}_2 \in L_n^p$. Finally $\mathbf{e}_1$ and $\mathbf{y}_1$ are also in L_n^p because $\mathbf{e}_1 = \mathbf{u}_1 - \mathbf{y}_2$ and $\mathbf{y}_1 = \mathbf{e}_2 - \mathbf{u}_2$. This completes the proof.

6. Example. As an application of the results in § 4, we derive an L^p stability ($1 \leqq p \leqq \infty$) criterion for the damped Mathieu equation with a *forcing function.* The stability regions in the parameter-plane are then compared with those obtained by Michael [17] and Parks [18] for the *free* damped Mathieu equation. The result is stated as Theorem 2.

THEOREM 2. *Consider the forced damped Mathieu equation defined by*

$$\ddot{y} + a\dot{y} + (b + \varepsilon\cos nt)y = u(t), \tag{22}$$

where a, b, ε and n are some finite constants with $a > 0$, $b > 0$ and $n > 0$. Let k be a real number such that $b + k > 0$. Let $\lambda = \max\{|\varepsilon - k|, |\varepsilon + k|\}$. If either

(i) $a^2 > 4(b + k)$ *and* $\lambda < \{(b + k)[a^2 - 4(b + k)]^{1/2}\}/a$,

or

(ii) $a^2 < 4(b + k)$ *and* $\lambda < \{a[4(b + k) - a^2]^{1/2}\}/4$, *then* $u \in L^p$, $1 \leqq p \leqq \infty$, *implies* $y \in L^p$. *Furthermore, if* $u \in L^2$, *then* $y \in L^2 \cap L^\infty$ *and* $y(t) \to 0$ *as* $t \to \infty$.

Proof. First note that from the Bellman–Gronwall inequality, if $u \in L_e^p$, then $y \in L_e^p$. Now rewrite (22) as

$$\ddot{y} + a\dot{y} + by + (\varepsilon \cos nt)y = u(t) \tag{23}$$

and observe that (23) is the system equation of the system S (Fig. 1), where the linear, time-invariant, nonanticipative subsystem G has

$$\hat{g}(s) \triangleq \frac{1}{s^2 + as + b} \tag{24}$$

as a transfer function and the block in the feedback path is the memory-less, time-varying gain

$$\psi(t) = \varepsilon \cos nt. \tag{25}$$

Let $g(t) \triangleq \mathscr{L}^{-1}[\hat{g}(s)]$ be the open-loop impulse response of G. Since $a > 0$ and $b > 0$, we see easily from (24) that $g \in L^1 \cap L^\infty$, i.e., $g \in L^p$ for any real $p \in [1, \infty]$. Let k be some real number and make a system transformation as described in the proof of Theorem 1 ; we obtain

$$\hat{h}_k(s) \triangleq \frac{1}{s^2 + as + (b + k)} \tag{26}$$

and

$$\bar{\psi}(t) \triangleq \psi(t) - k. \tag{27}$$

By assumption $a > 0$ and $b + k > 0$; clearly $h_k(t) \triangleq \mathscr{L}^{-1}[\hat{h}_k(s)] \in L^1$. From (26) we obtain

$$\begin{aligned} h_k(t) = \frac{1}{(a^2 - 4(b + k))^{1/2}} \Bigg\{ &\exp\left[-\frac{a}{2} + \frac{(a^2 - 4(b + k))^{1/2}}{2} \right] t \\ &- \exp\left[-\frac{a}{2} - \frac{(a^2 - 4(b + k))^{1/2}}{2} \right] t \Bigg\}. \end{aligned} \tag{28}$$

Now we consider two separate cases: In case assumption (i) holds, i.e., $a^2 > 4(b + k)$, $[a^2 - 4(b + k)]^{1/2}$ is a real number. We obtain from (28)

$$\|h_k\|_1 \leqq \frac{a}{(b + k)(a^2 - 4(b + k))^{1/2}}. \tag{29}$$

From (25) and (27), we note that $|\bar{\psi}(t)| \leqq \lambda$. By assumption (i), $\|h_k\|_1 \lambda < 1$. In case assumption (ii) holds, i.e., $a^2 < 4(b + k)$, $[a^2 - 4(b + k)]^{1/2} = j[4(b + k) - a^2]^{1/2}$. From (28) we obtain

$$h_k(t) = \frac{2}{(4(b + k) - a^2)^{1/2}} \left[\exp\left(-\frac{a}{2} t \right) \right] \sin\left[\frac{(4(b + k) - a^2)^{1/2}}{2} t \right]. \tag{30}$$

Hence

$$\|h_k\|_1 \leqq \frac{4}{a(4(b+k)-a^2)^{1/2}}. \tag{31}$$

Since $|\bar{\psi}(t)| \leqq \lambda$, and by assumption (ii), $\lambda < [4(b+k)-a^2]^{1/2}a/4$, we have $\|h_k\|_1\lambda < 1$. So in both cases, $\|h_k\|_1\lambda < 1$. It then follows from Corollary 1 that $u \in L^p$, $1 \leqq p \leqq \infty$, implies that $y \in L^p$. Hence, so is $e \in L^p$ because $e = u - y$. Furthermore, if $u \in L^2$, then $y \in L^\infty$; indeed $y = g * e$ and both g and e are in L^2, thus $\hat{y} = \hat{g}\hat{e} \in L^1$. Moreover, by the Riemann–Lebesgue lemma, $y(t) \to 0$ as $t \to \infty$. This completes the proof.

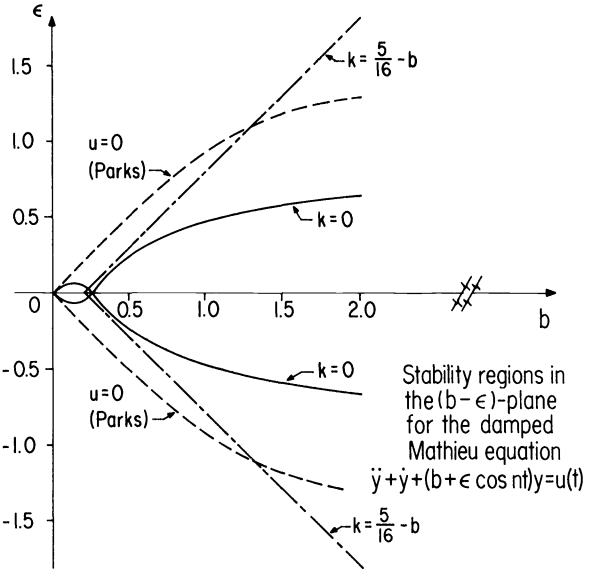


FIG. 2

Remark 4. In Theorem 2, if the linear time-varying gain $\psi(t) = \varepsilon \cos nt$ is replaced by a nonlinear time-varying gain $\varphi(\sigma, t)$ subjected to $|\varphi(\sigma, t)| \leqq |\varepsilon|\ |\sigma|$, then the results of Theorem 2 still hold.

Remark 5. Figure 2 shows various stability regions of (22) in the $(b - \varepsilon)$-plane with $a = 1$ obtained by Theorem 2 and by Parks [18]. With $k = 0$, the L^p-stability region (i.e., the regions enclosed by the solid line) are $|\varepsilon| < [b(a^2 - 4b)^{1/2}]/a$ when $a^2 > 4b$ and $|\varepsilon| < [a(4b - a^2)^{1/2}]/4$ when $a^2 < 4b$. With $k = -b + 5a^2/16$, the L^p-stability region (i.e., the region enclosed by the dash-dot line) is $|\varepsilon| < b - (3a^2)/16$. Note that the region obtained with $k = -b + 5a^2/16$ is much larger than those obtained with $k = 0$. This justifies Remark 2 that an appropriate choice of k will give a better stability result. The stability regions (i.e., the regions enclosed by dashed lines) obtained by Parks [18] for the *free* damped Mathieu equation (i.e., $u(t) \equiv 0$ in (22)) are $\varepsilon^2 < a^2(4b - a^2)/4$ when $a^2 < 4b$ and $\varepsilon^2 < b^2$ when $a^2 > 2b$.

Note that even though there is a forcing function, the stability region obtained by Theorem 2 with $k = -b + 5a^2/16$ is not contained in that obtained by Parks for the zero-input case.

REFERENCES

[1] I. W. Sandberg, *On the L^2-boundedness of solutions of nonlinear functional equations*, Bell System Tech. J., 43 (1964), pp. 1581–1599.

[2] G. Zames, *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.

[3] I. W. Sandberg, *A condition for L^∞-stability of feedback systems containing a single time-varying nonlinear element*, Bell System Tech. J., 43 (1964), pp. 1815–1817.

[4] G. Zames, *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.

[5] A. R. Bergen, R. P. Iwens and A. J. Rault, *On the input-output stability of nonlinear feedback systems*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 742–745.

[6] C. A. Desoer and M. Y. Wu, *Stability of linear time-invariant systems*, IEEE Trans. Circuit Theory, CT-15 (1968), pp. 245–250.

[7] ———, *Stability of multiple-loop feedback linear time-invariant systems*, J. Math. Anal. Appl., 22 (1968), pp. 121–130.

[8] ———, *Stability of linear time-invariant systems*, Proc. 2nd Princeton Conference on Information Sciences and Systems, Princeton, 1968, pp. 12–14.

[9] C. T. Chen, *L^p-stability of linear time-varying feedback systems*, this Journal, 6 (1966), pp. 186–193.

[10] E. Hille and R. S. Phillips, *Functional Analysis and Semi-Groups*, Revised edition, American Mathematical Society, Providence, 1957.

[11] G. Zames and P. L. Falb, *Stability conditions for systems with monotone and slope restricted nonlinearities*, this Journal, 6 (1968), pp. 89–108.

[12] J. Dieudonné, *Foundations of Modern Analysis*, Academic Press, New York, 1960.

[13] I. W. Sandberg, *Some results on the theory of physical systems governed by nonlinear functional equations*, Bell System Tech. J., 44 (1965), pp. 871–898 and Bell System Tech. J. Brief, 44 (1965), pp. 1809–1912.

[14] C. A. Desoer, *Lecture notes for EECS* 290D, University of California, Berkeley, 1968.

[15] L. Schwartz, *Théorie des distributions*, Revised edition, Herman, Paris, 1966.

[16] D. V. Widder, *The Laplace Transform*, Princeton University Press, Princeton, 1946.

[17] G. J. Michael, *Explicit stability criteria for the damped Mathieu equation*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 337–338.

[18] P. C. Parks and G. J. Michael, *Comment on explicit stability criteria for the damped Mathieu equation*, Ibid., AC-13 (1968), p. 129.

SIAM J. CONTROL
Vol. 7, No. 2, May 1969

ERRATUM: ON A MATRIX RICCATI EQUATION OF STOCHASTIC CONTROL*

W. M. WONHAM

In § 5, the symbols $N(s)$, $N(t)$, $N(\sigma)$ should be replaced wherever they occur by the symbol N.

In § 6, the expression (6.1) should read:

$$\Pi - PB(\Gamma + N)^{-1}B'P.$$

* This Journal, 6 (1968), pp. 681–697. Received by the editors January 25, 1969.

SIAM J. CONTROL
Vol. 7, No. 3, August 1969

CONTINUOUS TIME MARKOVIAN SEQUENTIAL CONTROL PROCESSES*

S. S. CHITGOPEKAR†

Abstract. Consider a stochastic system with a finite state space and a finite action space. Between actions, the waiting time to transition is a random variable with a continuous distribution function depending only on the current state and the action taken. There are positive costs of taking actions and the system earns at a rate depending upon the state of the system and the action taken. We allow actions to be taken between transitions. A policy for which there is a positive probability of an action between transitions involves "hesitation." A form of the long range average income is the criterion for comparing different policies. It is shown that there exists a nonrandomized stationary policy that is optimal in the class of all policies for which the actions taken form a sequence. "Hesitation" can be eliminated if the waiting time distributions are exponential. Howard's policy improvement method can be used to obtain an optimal policy.

1. Introduction.

We are interested in a stochastic system which at any time can be in one of a finite number of L states. In each of these L states, we have $K(<\infty)$ alternative actions available to us. In state i, it costs $c_i^k(>0)$ to take action k and then the system earns at the rate a_i^k per unit of time, $k = 1, \cdots, K, i = 1, \cdots, L$. The probability of transition from state i to state j, under action k, *if there is no further action before a transition*, is denoted by p_{ij}^k. We have

$$(1.1)\qquad \begin{aligned} &p_{ij}^k \geqq 0, &&\text{for all } i, j, k\\ &\sum_{j=1}^{L} p_{ij}^k = 1, &&\text{for all } i, k. \end{aligned}$$

We assume that under action k in state i the waiting time to transition, *if there is no further action before a transition*, is a random variable with a continuous distribution function with a finite mean and depending only on the current state of the system (i) and the action taken (k). Unlike the models considered in [4], [5] and [6], we have assumed these "waiting time" distributions to be independent of the state to which the system makes the next transition. Let $F_i^k(x)$ denote the waiting time to transition distribution function if action k is used in state i. Then assume

$$F_i^k(0) = 0; \qquad F_i^k(x) > 0, \qquad \text{for all } x > 0$$

and

$$\left.\frac{d}{dx}F_i^k(x)\right|_{x=0} \quad \text{exists and is finite;} \quad k = 1, \cdots, K; \quad i = 1, \cdots, L.$$

We also assume that in each state $i, i = 1, \cdots, L$, there is at least one action k_i such that the resulting transition probability matrix $P = (p_{ij}^{k_i})$ is the transition probability matrix of an irreducible Markov chain. This assumption will be discussed in the Appendix.

* Received by the editors July 16, 1968, and in revised form January 17, 1969.

† Hyderabad, A. P., India. This research is part of a doctoral dissertation submitted to the Florida State University, Tallahassee, Florida 32306. This research was supported by the Office of Naval Research under Contract NONR 988(13): NR 042-003.

Similar systems have been studied in [4], [5] and [6], where the class of stationary policies taking actions only at the instants of transitions is considered. We are interested in the existence and nature of optimal policies in a wider class of policies that allow actions between transitions.

Any policy S is a rule of action. If a policy S is such that there is a positive probability of an action being taken between transitions, following [1], we shall say that the policy S involves "hesitation." Let $G(\cdot)$ be a distribution function (on $[0, \infty]$) such that $G(t)$ denotes the probability that an action would be taken by time t since the previous action, given that there is no transition before then. We shall refer to such a distribution $G(\cdot)$ as a "hesitation distribution."

We shall restrict ourselves to the class of policies $\mathscr{S}$ where for any policy S in $\mathscr{S}$, actions are necessarily taken at the instants of transitions and the actions to be taken form a sequence with probability one. Thus, any policy S in $\mathscr{S}$ is of the following type:

After any transition and when an action is to be taken before a transition, S specifies

(a) the action to be taken, possibly randomized and depending on the past history of the system; and

(b) a hesitation distribution, possibly depending on the past history of the system and the action taken just now.

For any policy S in $\mathscr{S}$, the criterion of interest is $I(S)$, given by

$$I(S) = \liminf_{N\to\infty} I_N(S) = \liminf_{N\to\infty} \frac{\sum_{n=1}^{N} E(i_n(S))}{\sum_{n=1}^{N} E(T_n(S))}, \tag{1.2}$$

where $i_n(S)$ is the income earned under the nth action and $T_n(S)$ is the time spent by the system under the nth action of the policy S. In general, $I(S)$ depends on the initial state of the system. However, we shall see later that for the policies we need to consider, $I(S)$ will be independent of the initial state of the system. In view of this, this dependence of $I(S)$ on the initial state has been suppressed in the notation throughout. Let

$$I^* = \sup_{S\in\mathscr{S}} I(S). \tag{1.3}$$

We are interested in the existence and nature of S^* such that

$$I^* = I(S^*). \tag{1.4}$$

A policy, S^*, satisfying (1.4) is said to be *optimal.*

Definition. A hesitation distribution $G(\cdot)$ is said to be a *one-point distribution* if there exists an x, $0 \leqq x \leqq \infty$, such that

$$G(t) = \begin{cases} 0 & \text{for } t < x, \\ 1 & \text{for } t \geqq x. \end{cases}$$

We shall denote such a distribution by $G_x(\cdot)$.

Definition. A policy S is said to be *nonrandomized stationary* if for each i, $i = 1, \cdots, L$, S specifies a pair (k_i, G_i) such that, whenever an action is to be taken in state i, S prescribes the action k_i and a one-point hesitation distribution $G_i(\cdot)$.

DEFINITION. If, on some occasion, a policy S involves the hesitation distribution $G_0(\cdot)$, then we say that the policy S involves *instantaneous hesitation* on that particular occasion.

Remark. $G_0(\cdot)$ is included only for mathematical convenience and would never be used as costs of actions are positive.

Let $S_0 \subset S$ be the subclass of nonrandomized stationary policies.

THEOREM 1.1. *There exists a policy S^* in $\mathscr{S}_0$ that is optimal in $\mathscr{S}$.*

COROLLARY 1.2. *If all the waiting time distributions are exponential, hesitation can be eliminated.*

The proofs of Theorem 1.1 and Corollary 1.2 appear in § 5. The remainder of this section gives easily derived properties of an optimal policy. Section 2 develops some of the tools required when evaluating policies having hesitation. Properties of the nonrandomized stationary policies are discussed in § 3. The approximation of the continuous time problem by a discrete time problem is developed in § 4. Section 5 contains the proof of the basic theorem and gives some qualitative properties of the optimal policy, § 6 deals with the equivalence of different criteria for the class of nonrandomized stationary policies and § 7 gives a computational method to obtain an optimal policy. Required results from Markov chain theory are in the Appendix.

Let

$$\theta^* = \max_{i,k} \theta(i,k), \qquad \theta_* = \min_{i,k} \theta(i,k),$$

$$a^* = \max_{i,k} a_i^k, \qquad a_* = \min_{i,k} a_i^k,$$

$$c^* = \max_{i,k} c_i^k, \qquad c_* = \min_{i,k} c_i^k$$

and

$$I^* = I(S^*).$$

LEMMA 1.3. $a_* - c^*/\theta_* \leqq I^* \leqq a^* - c_*/\theta^*$.

Proof. The proof is straightforward.

Let

$$\mu_0 = \frac{c_*}{a^* - a_* + c^*/\theta_*}. \tag{1.5}$$

LEMMA 1.4. *Any policy S for which* $\liminf_{N\to\infty}\{\sum_{n=1}^N ET_n(S)/N\} < \mu_0$ *is not optimal.*

Proof. For any policy S we have

$$I_N(S) \leqq \frac{a^* - c_*}{\sum_{n=1}^{N} ET_n(S)/N}.$$

Let $\mu(S) = \liminf_{N\to\infty}\{\sum_{n=1}^N ET_n(S)/N\}$. Then

$$I(S) \leqq \frac{a^* - c_*}{\mu(S)}. \tag{1.6}$$

If $\mu(S) < \mu_0$, we have

$$I(S) < a^* - \frac{c_*}{\mu_0} \qquad \text{(from (1.6))}$$

$$= a^* - \frac{a^* - a_* + c^*}{\theta_*} \qquad \text{(from (1.5))}$$

$$= a_* - \frac{c^*}{\theta_*}$$

$$\leqq I^* \qquad \text{(from Lemma 1.3).}$$

Hence the Lemma.

In view of Lemma 1.4, we shall only consider policies S for which $\mu(S) \geqq \mu_0$.

2. Some preliminary formulas and theorems. $F(\cdot)$ is a waiting time distribution and $G(\cdot)$ is a hesitation distribution. When $F(\cdot) = F_i^k(\cdot)$, F is replaced by i, k in all expressions. Similarly, when $G(\cdot) = G_t(\cdot)$, a one-point distribution, G is replaced by t in all expressions.

(2.1) $\theta(F) =$ the expected value of a random variable with distribution function $F(\cdot)$, such that $F(0^-) = 0$

$$= \int_0^\infty t\,dF(t) = \int_0^\infty (1 - F(t))\,dt.$$

(2.2) $\eta(F; G) =$ expected time to the next action when the waiting time distribution is $F(\cdot)$ and the hesitation distribution is $G(\cdot)$

$$= \int_0^\infty t(1 - G(t))\,dF(t) + \int_0^\infty t(1 - F(t))\,dG(t)$$

$$= \int_0^\infty t\,dF(t) - \int_0^\infty (1 - F(t))G(t)\,dt$$

$$= \int_0^\infty (1 - F(t))(1 - G(t))\,dt$$

$$= \int_0^\infty \left\{ \int_0^x (1 - F(t))\,dt \right\} dG(x).$$

(2.3) $$\eta(F; t) = \int_0^t (1 - F(x))\,dx.$$

From (2.3) and the last form of (2.2) we obtain

(2.4) $$\eta(F; G) = \int_0^\infty \eta(F; t)\,dG(t).$$

(2.5) $q(F;G) =$ probability of a transition before the next action when the waiting time distribution is $F(\cdot)$ and the hesitation distribution is $G(\cdot)$

$$= \int_0^\infty F(t)\,dG(t).$$

(2.6) $$q(F;t) = F(t).$$

(2.7) $N(F;G) =$ expected number of actions before the next transition when we hesitate repeatedly with the hesitation distribution $G(\cdot)$ while the waiting time distribution is $F(\cdot)$.

$$= 1q(F;G) + 2(1 - q(F;G))q(F;G) + 3(1 - q(F;G))^2q(F;G) + \cdots$$

$$= \frac{1}{q(F;G)}, \qquad \text{if } q(F;G) > 0, \text{ that is } G(\cdot) \neq G_0(\cdot).$$

When $q(F;G) = 0$, we have

(2.8) $$N(F;G) = \infty.$$

(2.9) $\theta(F;G) =$ expected time to transition when the waiting time distribution is $F(\cdot)$ and we hesitate repeatedly with the hesitation distribution $G(\cdot)$ until a transition occurs

$$= \int_0^\infty t(1 - G(t))\,dF(t) + \int_0^\infty (t + \theta(F;G))(1 - F(t))\,dG(t)$$

$$= \eta(F;G) + \theta(F;G)(1 - q(F;G)).$$

Thus, when $q(F;G) > 0$,

$$\theta(F;G) = \frac{\eta(F;G)}{q(F;G)}.$$

(2.10) $$\theta(F;t) = \int_0^t \frac{1 - F(x)\,dx}{F(t)}.$$

Note that $\theta(F) = \theta(F;\infty) = \eta(F;\infty)$. We shall define $\theta(F;0) = \lim_{t\to 0}\theta(F;t)$. Note that $q(F;G) = 0$ if and only if $G(\cdot) = G_0(\cdot)$. Hence, when $q(F;G) = 0$, we have $\theta(F;G) = \theta(F;0)$. Let $\underline{G}(\cdot)$ and $\bar{G}(\cdot)$ be such that

(2.11) $$\theta(F;\underline{G}) = \min_G \theta(F;G)$$

(2.12) $$\theta(F;\bar{G}) = \max_G \theta(F;G).$$

LEMMA 2.1. *$\underline{G}(\cdot)$ and $\bar{G}(\cdot)$ can be taken to be one-point distributions.*

Proof. From (2.9), (2.4) and (2.5), we have

$$\theta(F;G)=\frac{\int_0^\infty \eta(F;t)\,dG(t)}{\int_0^\infty F(t)\,dG(t)}$$

$$=\frac{\int_0^\infty \theta(F;t)F(t)\,dG(t)}{\int_0^\infty F(t)\,dG(t)} \qquad \text{(from (2.3) and (2.10)).}$$

Since $\theta(F;x)$ is a continuous function of x, we can obtain x_0 satisfying

$$\theta(F;x_0)=\min_{x\in[0,\infty]}\theta(F;x). \tag{2.13}$$

Then

$$\theta(F;x_0)-\theta(F;G)=\frac{\int_0^\infty F(t)\{\theta(F;x_0)-\theta(F;t)\}\,dG(t)}{\int_0^\infty F(t)\,dG(t)}$$

$$\leqq 0 \qquad \text{for all } G.$$

Similarly, if y_0 is such that

$$\theta(F;y_0)=\max_{x\in[0,\infty]}\theta(F;x) \tag{2.14}$$

then $\theta(F;y_0)\geqq\theta(F;G)$ for all G. Take $\underline{G}(\cdot)=G_{x_0}(\cdot)$ and $\bar{G}(\cdot)=G_{y_0}(\cdot)$. Hence the Lemma.

THEOREM 2.2. *Let $G(\cdot)$ be a distribution function on $[0,\Delta]$, $\Delta\leqq\infty$, and let $h(\cdot)$ be a monotone continuous function on $[0,\Delta]$. For any given*

$$0=t_0<t_1<\cdots<t_{N-1}=t_N=\Delta, \tag{2.15}$$

there exists a discrete distribution $G_*(\cdot)$ with its mass at the points $t_i, i=0,1,\cdots,N$, and such that

$$\int_0^\Delta h(t)\,dG(t)=\int_0^\Delta h(t)\,dG_*(t). \tag{2.16}$$

Further, if $m(\cdot)$ is a monotone continuous function such that

$$|m(t_i)-m(t_{i-1})|\leqq\varepsilon, \qquad i=1,\cdots,N, \tag{2.17}$$

then

$$\left|\int_0^\Delta m(t)\,dG(t)-\int_0^\Delta m(t)\,dG_*(t)\right|\leqq 2\varepsilon. \tag{2.18}$$

Proof. The proof is standard. It should be observed that $G_*(\cdot)$ does not depend on $m(\cdot)$.

3. The class of policies $\mathscr{S}_0$. Any policy S in $\mathscr{S}_0$ is specified by $\{(k_i, G_i)\}$, where k_i is the action and $G_i(\cdot)$ is the hesitation distribution prescribed by S in state $i, i = 1, \cdots, L$. In this section, we will allow policies in $\mathscr{S}_0$ with $\{G(\cdot)\}$ general, that is, not restricted to one-point distributions. Theorem 3.2 will show that this added generality is not advantageous. In view of Lemma 1.4, we can assume $q(i; G_i) > 0$ for all i. Let $X_n(S)$ denote the state of the system after the nth transition, $n = 0, 1, 2, \cdots$. ($X_0(S)$ is the initial state of the system.) Let $Y_n(S)$ denote the state of the system at the time of the nth action, $n = 1, 2, \cdots$. It is easy to see that $\{X_n(S): n = 0, 1, 2, \cdots\}$ and $\{Y_n(S): n = 1, 2, \cdots\}$ are Markov chains with state space $\{1, \cdots, L\}$ and transition probability matrices $P = (p_{ij}^{k_i})$ and $P' = (p'_{ij})$ respectively, where

$$(3.1) \qquad p'_{ij} = p_{ij}^{k_i} q(i; G_i) + \delta_{ij}(1 - q(i; G_i)),$$

$$\delta_{ij} = \begin{cases} 0, & \text{if } i \neq j, \\ 1, & \text{if } i = j. \end{cases}$$

It can be further seen that both the Markov chains $\{X_n(S)\}$ and $\{Y_n(S)\}$ have the same class structure. From the Appendix it then follows that we need only consider policies S in $\mathscr{S}_0$ such that the Markov chains $\{X_n(S)\}$ and $\{Y_n(S)\}$ have only one positive class.

In § 1 we have defined $I(S)$, our criterion of interest for any policy S, by (1.2). Now define

$$(3.2) \qquad I^{(2)}(S) = \liminf_{N \to \infty} \frac{\sum_{n=1}^{N} E(i_n^{(2)}(S))}{\sum_{n=1}^{N} E(T_n^{(2)}(S))},$$

where $T_n^{(2)}(S)$ is the time from the $(n - 1)$th transition to the nth transition and $i_n^{(2)}(S)$ is the income earned during this period.

THEOREM 3.1. *For any policy S in $\mathscr{S}_0$, $I(S) = I^{(2)}(S)$.*

Proof. Since we shall be considering a particular policy S in $\mathscr{S}_0$, we shall drop the index k from all the parameters. Let $G_i(\cdot)$ denote the hesitation distribution in state i, $i = 1, \cdots, L$. Let P and P' be the transition probability matrices associated with the Markov chains $\{X_n(S)\}$ and $\{Y_n(S)\}$ respectively. We then have, from the Appendix,

$$(3.3) \qquad I(S) = \frac{\sum_{i=1}^{L} \pi'_i(a_i \eta(i; G_i) - c_i)}{\sum_{i=1}^{L} \pi'_i \eta(i; G_i)}$$

and

$$(3.4) \qquad I^{(2)}(S) = \frac{\sum_{i=1}^{L} \pi_i(a_i \theta(i; G_i) - c_i N(i; G_i))}{\sum_{i=1}^{L} \pi_i \theta(i; G_i)},$$

where $\eta(i;G_i)$, $\theta(i;G_i)$ and $N(i;G_i)$ are as defined in § 2; $\{\pi_i\}$ are the stationary probabilities associated with $P=(p_{ij})$; and $\{\pi_i'\}$ are the stationary probabilities associated with $P'=(p_{ij}')$.

$\{\pi_i\}$ are given by the system

$$\begin{aligned} &\pi_i \geqq 0, && i=1,\cdots,L,\\ &\sum_{i=1}^{L}\pi_i = 1, && \\ &\pi_j = \sum_{i=1}^{L}\pi_i p_{ij}, && j=1,\cdots,L. \end{aligned} \tag{3.5}$$

$\{\pi_i'\}$ are given by the system

$$\begin{aligned} &\pi_i' \geqq 0, && i=1,\cdots,L,\\ &\sum_{i=1}^{L}\pi_i' = 1, && \\ &\pi_j' = \sum_{i=1}^{L}\pi_i' p_{ij}', && j=1,\cdots,L. \end{aligned} \tag{3.6}$$

From (3.1), (3.5) and (3.6), we obtain

$$\pi_i' = \frac{\pi_i/q(i;G_i)}{\sum_{i=1}^{L}\pi_i/q(i;G_i)}, \qquad i=1,\cdots,L. \tag{3.7}$$

Thus

$$\begin{aligned} I(S) &= \frac{\sum_{i=1}^{L}\pi_i'(a_i\eta(i;G_i)-c_i)}{\sum_{i=1}^{L}\pi_i'\eta(i;G_i)}\\ &= \frac{\sum_{i=1}^{L}\pi_i(a_i\theta(i;G_i)-c_iN(i;G_i))}{\sum_{i=1}^{L}\pi_i\theta(i;G_i)}\\ &= I^{(2)}(S). \end{aligned}$$

THEOREM 3.2. *In $\mathscr{S}_0$ we need only consider policies involving one-point hesitation distributions.*

Proof. Let S be any policy in $\mathscr{S}_0$ and let $I(S)=I$. Let $G_i(\cdot)$ be the hesitation distribution in state i under the policy S, $i=1,\cdots,L$. We then have, dropping the index k from all the parameters,

$$I(S) = I = \frac{\sum_{i=1}^{L}\pi_i(a_i\theta(i;G_i)-c_iN(i;G_i))}{\sum_{i=1}^{L}\pi_i\theta(i;G_i)}.$$

Let $x_i, i = 1, \cdots, L$, be such that

$$(3.8) \qquad \theta(i; x_i)(a_i - I) - c_i N(i; x_i) = \max_{x \in [0,\infty]} \theta(i; x)(a_i - I) - c_i N(i; x).$$

Let S' be the policy that differs from S only in that the hesitation distributions $G_i(\cdot)$ are replaced by $G_{x_i}(\cdot), i = 1, \cdots, L$. We then have

$$I(S') = \frac{\sum_{i=1}^{L} \pi_i(a_i\theta(i; x_i) - c_i N(i; x_i))}{\sum_{i=1}^{L} \pi_i \theta(i; x_i)}.$$

It can be easily verified that

$$(3.9) \quad I(S') - I = \frac{\sum_{i=1}^{L} \pi_i\{(a_i - I)[\theta(i; x_i) - \theta(i; G_i)] - c_i[N(i; x_i) - N(i; G_i)]\}}{\sum_{i=1}^{L} \pi_i \theta(i; x_i)}.$$

Now for each $i, i = 1, \cdots, L$,

$$(a_i - I)[\theta(i; x_i) - \theta(i; G_i)] - c_i[N(i; x_i) - N(i; G_i)]$$

$$= (a_i - I)\left\{\theta(i; x_i) - \frac{\int_0^\infty \theta(i; t)F_i(t)\, dG_i(t)}{\int_0^\infty F_i(t)\, dG_i(t)}\right\} - c_i\left\{\frac{1}{F_i(x_i)} - \frac{1}{\int_0^\infty F_i(t)\, dG_i(t)}\right\}$$

$$= \frac{\int_0^\infty F_i(t)\{[(a_i - I)\theta(i; x_i) - c_i N(i; x_i)] - [(a_i - I)\theta(i; t) - c_i N(i; t)]\}\, dG_i(t)}{\int_0^\infty F_i(t)\, dG_i(t)}.$$

In view of (3.8) this last expression will be nonnegative if $dG_i(x_i) = 1, i = 1, \cdots, L$. Hence the theorem.

THEOREM 3.3. *There exists an optimal policy in the class $\mathscr{S}_0$.*

Proof. From Theorem 3.2, we can restrict our attention to policies in $\mathscr{S}_0$ involving one-point hesitation distributions. Let S be any policy in $\mathscr{S}_0$ involving hesitation distribution $G_{t_i}(\cdot)$ in state $i, i = 1, \cdots, L$. Let $\bar{S}$ denote only the choice of actions in various states dictated by S. Then, dropping the index k from the parameters, we have

$$I(S) = I(\bar{S}; t_1, \cdots, t_L)$$

$$= \frac{\sum_{i=1}^{L} \pi_i\left[a_i \int_0^{t_i} (1 - F_i(x))\, dx - c_i\right]/F_i(t_i)}{\sum_{i=1}^{L} \left\{\pi_i \int_0^{t_i} (1 - F_i(x))\, dx/F_i(t_i)\right\}}.$$

All the $F(\cdot)$ are assumed to be continuous for each $\bar{S}$. As a function of $t_1, \cdots, t_L$, $I(\bar{S}; t_1, \cdots, t_L)$ is continuous on $(0, \infty] \times \cdots \times (0, \infty]$ and is bounded above. Since, if any $t_i = 0$, $I(S) = -\infty$, $I(\bar{S}; t_1, \cdots, t_L)$, as a function of $t_1, \cdots, t_L$, attains its maximum. Further, since there is only a finite number of different $\bar{S}$, the theorem follows.

Example. The continuity assumption for $F(\cdot)$ used in the proof of Theorem 3.3 cannot be dropped.

Consider a system with two states and one action in each state. Let

$$F_1(x) = 1 - e^{-x}, \quad 0 \leqq x < \infty,$$

$$F_2(x) = \begin{cases} 1 - e^{-x}, & 0 \leqq x < 1, \\ 1 - e^{-1}, & 1 \leqq x < 2, \\ 1, & x \geqq 2, \end{cases}$$

$$a_2 > a_1, \quad c_1 > c_2, \quad p_{12} = 1 = p_{21}.$$

We have $\theta(F_1) = \theta(F_2) = 1$.

Since there is only one action in each state, any policy S in $\mathscr{S}_0$ is specified by (t_1, t_2) if $G_{t_1}(\cdot)$ and $G_{t_2}(\cdot)$ are the hesitation distributions specified by S. Thus $I(S) = I(t_1, t_2)$. Since $F_1(\cdot)$ is exponential, $I(t_1, t_2) \leqq I(\infty; t_2)$ for all t_2. We have $I(\infty, \infty) = \{(a_1 + a_2) - (c_1 + c_2)\}/2 = I$, say. Now

$$\theta(2; t) = \int_0^t (1 - F_2(x))\, dx / F_2(t)$$
$$= \begin{cases} 1, & \text{if } t \leqq 1 \quad \text{or} \quad t \geqq 2, \\ 1 + (t-1)e^{-1}/(1 - e^{-1}), & \text{if } 1 < t < 2. \end{cases}$$

Hence,

$$I(\infty, t) = \begin{cases} I, & \text{if } t \leqq 1 \quad \text{or} \quad t \geqq 2, \\ \dfrac{I + m(a_2(t-1) - c_2)}{1 + m(t-1)}, & \text{if } 1 < t < 2, \end{cases}$$

where $m = [2e(1 - e^{-1})]^{-1}$. Since $a_2 > a_1$, $d\, I(\infty, t)/dt > 0$ for $1 < t < 2$. Thus,

$$\sup_{1 < t < 2} I(\infty, t) = \lim_{t \to 2^-} I(\infty, t) = \frac{I + m(a_2 - c_2)}{1 + m} > I$$

since $c_1 > c_2$. Hence no optimal stationary policy exists.

4. An approximation theorem. For any $\varepsilon > 0$, let N_ε be a positive integer such that $\theta^*/N_\varepsilon < \varepsilon$. Let $t_i^k(r)$ be such that

$$\eta(i, k; t_i^k(r)) = \frac{r}{N_\varepsilon} \theta(i, k), \tag{4.1}$$

$r = 0, \cdots, N_\varepsilon$; $k = 1, \cdots, K$; $i = 1, \cdots, L$. Note that since $\eta(i, k; x)$ is a continuous, increasing function of x, for each (i, k), (4.1) has solutions. If there is more than one solution, we take $t_i^k(r)$ to be the largest solution. Further,

$$(4.2) \qquad \eta(i,k;t_i^k(r)) - \eta(i,k;t_i^k(r-1)) < \varepsilon$$

for all r, for any given (i,k).

Let $\mathscr{G}_\varepsilon$ denote the finite class of one-point hesitation distributions $\{G_x(\cdot)\}$, where $x \in \{t_i^k(r) : r = 0, \cdots, N_\varepsilon; k = 1, \cdots, K; i = 1, \cdots, L\}$.

Suppose under some policy S, at some stage, action k has been taken in state i, and a hesitation distribution $G(\cdot)$ is used. Then the probability of a transition before the next action is given by $\int_0^\infty F_i^k(t)\,dG(t)$ and the expected time before the next action is given by $\int_0^\infty \eta(i,k;t)\,dG(t)$. From (4.2) and Theorem 2.2, there exists a discrete distribution $G_*(\cdot)$ with its mass at the points $t_i^k(r)$, $r = 0, \cdots, N_\varepsilon$, such that replacing the hesitation distribution $G(\cdot)$ by $G_*(\cdot)$ does not change the probability of transition before the next action and the expected time to the next action is changed by at most 2ε.

Since the probability of transition before the next action is not changed by replacing $G(\cdot)$ by $G_*(\cdot)$, the distribution of the state at the time of next action also remains the same when $G(\cdot)$ is replaced by $G_*(\cdot)$. Further, observe that $G_*(\cdot)$ is a randomization over the class $\mathscr{G}_\varepsilon$.

DEFINITION. A policy S_ε will be said to be an *ε-approximation* of a policy S if $|I(S) - I(S_\varepsilon)| \leqq \varepsilon$.

THEOREM 4.1. *For any policy S and any sufficiently small $\varepsilon > 0$, we can find a policy S_ε, an ε-approximation of S, such that S_ε only involves hesitation distributions that are randomizations over the class $\mathscr{G}_{\varepsilon^2}$.*

Proof. Let (n_1)th be the first action at which S involves hesitation with positive probability and let $G(\cdot)$ be the hesitation distribution prescribed by S. Let $G_*(\cdot)$ be a randomization over $\mathscr{G}_{\varepsilon^2}$ such that replacing $G(\cdot)$ by $G_*(\cdot)$ does not change the distribution of states at the $(n_1 + 1)$th action and the time spent by the system under the (n_1)th action is changed by at most $2\varepsilon^2$. Replace $G(\cdot)$ by $G_*(\cdot)$ and after reaching the $(n_1 + 1)$th state, create by randomization the time that would have been spent under the (n_1)th action if the hesitation distribution were $G(\cdot)$ and given the $(n_1 + 1)$th state. Record this time as part of the "history" of the system. With this partially artificial history, take the $(n_1 + 1)$th action as prescribed by S and proceed to follow S. If the (n_2)th action is the next action that involves hesitation with positive probability, again replace the hesitation distribution by a randomization over $\mathscr{G}_{\varepsilon^2}$ as before and proceed similarly.

Repeat the same procedure at every step where hesitation is involved.

Let S' be the resulting modified policy. Note that part of the history maintained is artificial in that we are not recording the actual time spent by the system under any action that involves hesitation but creating by randomization the time that would be spent by the system if the hesitation distribution prescribed by S were to be used—conditioned on the state of the system at the time of the next action. Since replacing $G(\cdot)$ by a proper randomization over $\mathscr{G}_{\varepsilon^2}$ on any occasion does not change the distribution of the state of the system at the next action, the distribution of the histories of the system under S' at the time of any action is the same as that under S. Hence we have

$$ET_n(S') = ET_n(S) + \varepsilon_n,$$

where $-2\varepsilon^2 \leqq \varepsilon_n \leqq 2\varepsilon^2$, $n = 1, 2, \cdots$; and hence

$$Ei_n(S') = Ei_n(S) + a_n\varepsilon_n, \qquad n = 1, 2, \cdots,$$

where a_n is the earning rate of the nth action. Thus

$$\begin{aligned} I_N(S') &= \frac{\sum_{n=1}^{N} Ei_n(S')}{\sum_{n=1}^{N} ET_n(S')} \\ &= \frac{\left(\sum_{n=1}^{N} Ei_n(S) + \sum_{n=1}^{N} Ea_n\varepsilon_n\right)}{\left(\sum_{n=1}^{N} ET_n(S) + \sum_{n=1}^{N} \varepsilon_n\right)} \\ &= \frac{I_N(S) + \left(\sum_{n=1}^{N} Ea_n\varepsilon_n\right) \Big/ \left(\sum_{n=1}^{N} ET_n(S)\right)}{1 + \sum_{n=1}^{N} \varepsilon_n \Big/ \left(\sum_{n=1}^{N} ET_n(S)\right)}. \end{aligned}$$

Therefore,

$$\begin{aligned} |I_N(S') - I_N(S)| &= \left| \frac{\sum_{n=1}^{N} Ea_n\varepsilon_n - I_N(S) \sum_{n=1}^{N} \varepsilon_n}{\sum_{n=1}^{N} ET_n(S) + \sum_{n=1}^{N} \varepsilon_n} \right| \\ &\leqq \left| \frac{2N\bar{a}\varepsilon^2 + 2N\bar{a}\varepsilon^2}{\sum_{n=1}^{N} ET_n(S) + \sum_{n=1}^{N} \varepsilon_n} \right| \\ &= \frac{4\bar{a}\varepsilon^2}{\left| \left(\sum_{n=1}^{N} \frac{ET_n(S)}{N} \right) + \bar{\varepsilon} \right|}, \end{aligned}$$

where $\bar{a} = \max_{i,k} |a_i^k|$ and $\bar{\varepsilon} = \sum_{n=1}^{N} \varepsilon_n/N$. Since $\liminf_{N\to\infty} [\sum_{n=1}^{N} ET_n(S)/N] \geqq \mu_0 > 0$ and $-\varepsilon^2 \leqq \bar{\varepsilon} \leqq \varepsilon^2$, it follows that, for sufficiently small ε,

$$|I(S') - I(S)| \leqq \varepsilon.$$

Take $S_\varepsilon = S'$. Thus, S_ε is an ε-approximation of S.

5. Existence of an optimal policy. Theorem 5.1 is a slight modification of Derman's [3] Theorem 3. His theorem applies to the discrete time situation where for each (i, j) there exists a $k = k(i, j)$ such that $p_{ij}^k > 0$.

THEOREM 5.1 (Derman). *Let* w'_{ik} *and* $w''_{i,k} > 0$ *be two sets of expected rewards under action* k *in state* i, $k = 1, \cdots, K$; $i = 1, \cdots, L$. $(K < \infty, L < \infty)$. *Let* W'_n

and W''_n be the respective expected rewards ascribed to the n-th action for a fixed policy S. Consider the reward criterion

$$\psi_S(i) = \liminf_{N\to\infty} \frac{\sum_{n=1}^{N} W'_n}{\sum_{n=1}^{N} W''_n}, \qquad i = 1, \cdots, L,$$

when the initial state of the system is i. Then there exists a nonrandomized stationary policy S^* such that

$$\psi_{S^*}(i) = \max \psi_S(i), \qquad i = 1, \cdots, L,$$

where the maximum is taken over all policies.

Remark. $\psi_S(i)$ may depend on the initial state i for a general policy S but $\psi_{S^*}(i)$ is independent of i.

Proof of Theorem 1.1. For every $\varepsilon > 0$, let $\mathscr{S}_\varepsilon \subset \mathscr{S}$ denote the class of policies involving hesitation distributions that are randomizations over $\mathscr{G}_\varepsilon$. By virtue of Theorem 5.1, there exists a nonrandomized stationary policy S^*_ε that is optimal in $\mathscr{S}_\varepsilon$. Let S_0 be the optimal policy in $\mathscr{S}_0$ (Theorem 3.3). Note that S_0 is independent of ε and S^*_ε is in $\mathscr{S}_0$. Let $I^* = \sup_{S\in\mathscr{S}} I(S)$. Clearly,

$$I^* \geqq I(S_0) \geqq I(S^*_\varepsilon) \qquad \text{for all } \varepsilon > 0. \tag{5.1}$$

Let $\{S_m : m = 1, 2, \cdots\}$ be a sequence of policies in $\mathscr{S}$ such that

$$I(S_m) \uparrow I^*. \tag{5.2}$$

For a given $\varepsilon > 0$, we can find an M such that for $m \geqq M$, we have

$$I(S_m) \geqq I^* - \varepsilon. \tag{5.3}$$

Let $S'_{m,\varepsilon} \varepsilon \mathscr{S}_{\varepsilon^2}$ be an ε-approximation of S_m (Theorem 4.1). Hence

$$I(S'_{m,\varepsilon}) \geqq I(S_m) - \varepsilon. \tag{5.4}$$

We thus have

$$I(S^*_{\varepsilon^2}) \geqq I(S'_{m,\varepsilon}) \geqq I(S_m) - \varepsilon. \tag{5.5}$$

Now from (5.1), (5.5) and (5.3) we obtain

$$I^* \geqq I(S_0) \geqq I(S^*_{\varepsilon^2}) \geqq I^* - 2\varepsilon. \tag{5.6}$$

Since ε is arbitrary, from (5.6) we obtain

$$I^* = I(S_0). \tag{5.7}$$

Take $S^* = S_0$. Hence the theorem.

Remark. We have seen earlier (§ 3) that if the assumption of continuity of the waiting time distributions is dropped, an optimal stationary policy may not exist, and hence, an optimal policy may not exist. However, part of the inequality (5.6) is still true, i.e.,

$$I^* \geqq I(S^*_{\varepsilon^2}) \geqq I^* - 2\varepsilon. \tag{5.6$'$}$$

Thus, it follows that, even if an optimal policy does not exist, we can find a nonrandomized stationary policy that is "almost optimal."

Proof of Corollary 1.2. The corollary follows immediately from the fact that for exponential distributions, $\theta(i; t) = \theta(i)$ for all t.

We see from Corollary 1.2. that when the waiting time distributions are exponential, we can eliminate from consideration policies that involve hesitation. A question that arises is whether the same remains true when the distributions are not necessarily exponential. The following example answers this question in the negative.

Example. Consider a system with two states, with one action in each state and such that

$$F_1(x) = \begin{cases} x, & 0 \leqq x \leqq 1, \\ 1, & x > 1, \end{cases}$$

$$F_2(x) = 1 - e^{-x}, \quad 0 \leqq x < \infty,$$

$$a_1 = 4, \quad a_2 = 2; \qquad c_1 = c_2 = 1; \qquad p_{12} = 1 = p_{21}.$$

There is only one policy S in $\mathscr{S}_0$ not involving hesitation and

$$I(S) = \frac{\pi_1(a_1\theta(1) - c_1) + \pi_2(a_2\theta(2) - c_2)}{\pi_1\theta(1) + \pi_2\theta(2)}.$$

Since $p_{12} = p_{21} = 1$, we have $\pi_1 = \pi_2 = 1/2$. Also $\theta(1) = 1/2$, $\theta(2) = 1$. Thus $I(S) = 4/3$. Let S' be the policy that involves the hesitation distribution $G_{.8}(\cdot)$ in state 1 and no hesitation in state 2. We have

$$\theta(1; t) = \int_0^t \frac{1 - x}{t} dx = 1 - \frac{t}{2}.$$

Hence, $I(S') = 43/32 > 4/3$.

THEOREM 5.2. *Let S be an optimal nonrandomized stationary policy with hesitation distributions $G_{t_i}(\cdot)$, $i = 1, \cdots, L$, and let $I(S) = I$. Then if $a_i \geqq (\leqq) I$ and $F_i(\cdot)$ is such that the expected time to transition in state i cannot be increased (decreased) by hesitation, then $t_i = \infty$.*

Proof. Let S' be a policy that differs from S only in that it involves no hesitation in state i. Let $I' = I(S')$. Then

$$I - I' = \frac{\pi_i(\theta(i) - \theta(i; t_i))(I - a_i) - c_i\pi_i(N(i; t_i) - 1)}{\pi_i\theta(i) + \sum_{j \neq i} \pi_j\theta(j; t_j)}.$$

Now, if $a_i \geqq (\leqq) I$, $\theta(i; t_i) \leqq (\geqq) \theta(i)$ and $t_i < \infty$, we shall have $I - I' < 0$, a contradiction. Hence the theorem.

Remark. The result of the Theorem 5.2 also follows from the fact that the hesitation distributions involved in the optimal policy are such that they maximize, for each i, $\theta(i; t)(a_i - I) - c_iN(i; t)$ (Theorem 3.2).

THEOREM 5.3. *If $c_i^k = c$, $a_i^k = a$ and $\theta(i, k) = \theta$ for all i and k, hesitation can be eliminated.*

Proof. Let S be an optimal nonrandomized stationary policy. For simplicity, we shall drop the superscript k from all the parameters. Let $G_{x_i}(\cdot)$ be the hesitation distribution involved in state i, $i = 1, 2, \cdots, L$. We then have

$$I(S) = a - c\frac{\sum_{i=1}^{L} \pi_i N(i; x_i)}{\sum_{i=1}^{L} \pi_i \theta(i; x_i)}.$$

Let S' be a policy that differs from S only in that it involves no hesitation. We then have $I(S') = a - c/\theta$ and

$$I(S) - I(S') = c\left(\frac{i}{\theta} - \frac{\sum_{i=1}^{L} \pi_i N(i; x_i)}{\sum_{i=1}^{L} \pi_i \theta(i; x_i)}\right)$$

$$= c\frac{\left[\sum_{i=1}^{L} \pi_i N(i; x_i)(\eta(i; x_i) - \theta)\right]}{\theta \sum_{i=1}^{L} \pi_i \theta(i; x_i)}.$$

Thus, unless $x_i = \infty$ for all i, $I(S) - I(S') < 0$, a contradiction. Hence the theorem.

Note that if $c_i^k = c$ and $\theta(i, k) = \theta$ for all i, k but the a_i^k are not necessarily equal, we have, for any nonrandomized stationary policy S,

$$I(S) = \frac{\sum_{i=1}^{L} \pi_i a_i \theta(i; x_i)}{\sum_{i=1}^{L} \pi_i \theta(i; x_i)} - c\frac{\sum_{i=1}^{L} \pi_i N(i; x_i)}{\sum_{i=1}^{L} \pi_i \theta(i; x_i)}.$$

The first term lies between a_* and a^* and if c is large in comparison with $\max_{i,k} |a_i^k|$, the second term is the dominant term and we have

$$I(S) \doteqdot -c\frac{\sum_{i=1}^{L} \pi_i N(i; x_i)}{\sum_{i=1}^{L} \pi_i \theta(i; x_i)}.$$

Hence we can almost eliminate hesitation.

6. Equivalence of different criteria. $I(S)$ and $I^{(2)}(S)$ have been defined in (1.2) and (3.2). Now denote $I(S)$ by $I^{(1)}(S)$ and define

$$I^{(3)}(S) = \liminf_{N \to \infty} \frac{\sum_{n=1}^{N} i_n(S)}{\sum_{n=1}^{N} T_n(S)},$$

$$I^{(4)}(S) = \liminf_{N \to \infty} \frac{\sum_{n=1}^{N} i_n^{(2)}(S)}{\sum_{n=1}^{N} T_n^{(2)}(S)},$$

$$I^{(5)}(S) = \liminf_{T\to\infty} \frac{EI^T(S)}{T},$$

$$I^{(6)}(S) = \liminf_{T\to\infty} \frac{I^T(S)}{T},$$

where $I^T(S)$ is the income earned up to time T under the policy S. Note that $I^{(3)}$, $I^{(4)}$ and $I^{(6)}$ are random variables. We then have the following theorem.

THEOREM 6.1. *If S is any nonrandomized stationary policy such that the associated Markov chain $\{X_n(S)\}$ has only one positive class, then with probability one, $I^{(1)}(S) = I^{(2)}(S) = I^{(3)}(S) = I^{(4)}(S) = I^{(5)}(S) = I^{(6)}(S)$.*

Proof. From Theorem 3.1, we have

$$I^{(1)}(S) = I^{(2)}(S). \tag{6.1}$$

From Theorem 1 of the Appendix, it follows that

$$\begin{aligned} I^{(1)}(S) &= I^{(3)}(S) \quad \text{with probability 1,} \\ I^{(2)}(S) &= I^{(4)}(S) \quad \text{with probability 1.} \end{aligned} \tag{6.2}$$

Now consider $I^T(S)$. Suppose the system starts in state i which belongs to the positive class of the Markov chain $\{X_n(S)\}$. In terms of transitions, define a cycle as the period between successive returns to state i. Since S is nonrandomized stationary and the associated Markov chain has only one positive class, the successive cycles are independently and identically distributed and with probability one, the number of transitions in a cycle will be finite. Let $v_i(\mu_i)$ denote the expected income (length) of each cycle. Let $N(T)$ denote the number of complete cycles in time T; $V_n(i)$, the income earned in the nth cycle and $W_n(i)$, the income earned during the $(n + 1)$th incomplete cycle. We then have

$$\begin{aligned} I^{(5)}(S) &= \liminf_{T\to\infty} E \frac{\left[\sum_{n=1}^{N(T)} V_n(i) + W_{N(T)}(i)\right]}{T}, \\ I^{(6)}(S) &= \liminf_{T\to\infty} \frac{\left[\sum_{n=1}^{N(T)} V_n(i) + W_{N(T)}(i)\right]}{T}. \end{aligned} \tag{6.3}$$

The length of any transition is finite with probability one and so is the number of actions to a transition. Hence, the income earned in any cycle is finite with probability one. Hence $\lim_{T\to\infty} W_{N(T)}(i)/T = 0$ with probability one. Thus,

$$\begin{aligned} I^{(5)}(S) &= \liminf_{T\to\infty} E \sum_{n=1}^{N(T)} \frac{V_n(i)}{T}, \\ I^{(6)}(S) &= \liminf_{T\to\infty} \sum_{n=1}^{N(T)} \frac{V_n(i)}{T}. \end{aligned} \tag{6.4}$$

It then follows from [7] that

$$\begin{aligned} I^{(5)}(S) &= \frac{v_i}{\mu_i}, \\ I^{(6)}(S) &= \frac{v_i}{\mu_i} \quad \text{with probability one.} \end{aligned} \tag{6.5}$$

If $G_i(\cdot)$, $i = 1, \cdots, L$, are the hesitation distributions prescribed by the policy S, we have

$$\begin{aligned} \mu_i &= \theta(i; G_i) + \sum_{j \neq i} \pi_j \theta(j; G_j)/\pi_i, \\ v_i &= \{a_i\theta(i; G_i) - c_iN(i; G_i)\} + \sum_{j \neq i} \pi_j\{a_j\theta(j; G_j) - c_jN(j; G_j)\}/\pi_i. \end{aligned} \tag{6.6}$$

Thus,

$$I^{(1)}(S) = \frac{v_i}{\mu_i}. \tag{6.7}$$

The theorem follows from (6.1), (6.2), (6.5) and (6.7).

Now suppose the initial state of the system i does not belong to the positive class of the Markov chain $\{X_n(S)\}$. With probability one, the system will enter the positive class in a finite number of transitions. Since the length of a transition and the income earned during a transition are finite with probability one, we still have

$$\begin{aligned} I^{(5)}(S) &= \frac{v_j}{\mu_j}, \\ I^{(6)}(S) &= \frac{v_j}{\mu_j} \quad \text{with probability one} \end{aligned}$$

for any state j in the positive class of the Markov chain $\{X_n(S)\}$. Hence the theorem.

Remark. Although there is an optimal nonrandomized stationary policy for $I^{(1)}$ and the six criteria are equal with probability one for all stationary policies, it has not been shown that there are optimal nonrandomized policies for $I^{(2)}$ through $I^{(6)}$.

7. Determination of an optimal policy. In view of Theorem 1.1 we shall restrict ourselves to the class of nonrandomized stationary policies. Any policy in this class is specified by $\{(k_i, x_i)\}_{i=1}^{L}$, where k_i is the action prescribed by S in state i and $G_{x_i}(\cdot)$ is the hesitation distribution involved in state i, $i = 1, \cdots, L$. If S is such that the Markov chain $\{X_n(S)\}$ has only one positive class, then in view of Theorem 6.1 our system under the policy S is equivalent to the system considered in [4]; the mean waiting time in state i is $\theta(i; k_i, x_i)$; the reward due to transition from state i to state j is $-c_i^{k_i}N(i; k_i, x_i)$; the probability of transition from state i to state j is $p_{ij}^{k_i}$; and the actions are taken only at the instants of transitions. Corresponding to equation (37) of [4], we now have

$$v_i + g\theta(i; k_i, x_i) = a_i^{k_i}\theta(i; k_i, x_i) - c_i^{k_i}N(i; k_i, x_i) + \sum_j p_{ij}^{k_i}v_j, \qquad i = 1, \cdots, L. \tag{7.1}$$

Here $v_i - v_j$ can be interpreted as the limit, as $T \to \infty$, of the difference between the total expected income up to time T given that the system starts in states i and j respectively. Rewriting (7.1) we obtain

$$g = \frac{a_i^{k_i}\theta(i; k_i, x_i) - c_i^{k_i}N(i; k_i, x_i) + \sum_j p_{ij}^{k_i}v_j - v_i}{\theta(i; k_i, x_i)}, \quad i = 1, \cdots, L. \tag{7.2}$$

Note that adding the same constant to all v_i does not alter (7.2). Hence, without any loss of generality, we could take $v_L = 0$.

Now assume that for every nonrandomized stationary policy S, the Markov chain $\{X_n(S)\}$ is irreducible. In view of the preceding identification of our system with the system considered in [4], the iteration scheme described in [4] can be used to obtain an approximation to an optimal nonrandomized stationary policy.

The iteration cycle is illustrated in Table 1.

TABLE 1

Policy evaluation

For the present policy $\{(k_i, x_i)\}$ solve

$$v_i + g\theta(i; k_i, x_i) = a_i^{k_i}\theta(i; k_i, x_i) - c_i^{k_i}N(i; k_i, x_i) + \sum_j p_{ij}^{k_i}v_j,$$

$$i = 1, \cdots, L$$

with $v_L = 0$ for $g, v_1, \cdots, v_{L-1}$.

Policy improvement

Find the alternative (k, x) in state i that maximizes the test quantity $m(i; k, x)$, where

$$m(i; k, x) = \frac{a_i^k\theta(i; k, x) - c_i^kN(i; k, x) + \sum_j p_{ij}^k v_j - v_i}{\theta(i; k, x)}$$

using the values v_i of the previous policy. Make this alternative the new action in state i. Repeat for all $i = 1, \cdots, L$ to find a new policy. Do not change the old action in any state i unless another alternative has a strictly higher value of $m(i; k, x)$.

Let $\{S_n\}$ be a sequence of nonrandomized stationary policies such that S_n prescribes the action $(k_i^{(n)}, x_i^{(n)})$ in state i; $i = 1, \cdots, L$, $n = 1, 2, \cdots$. Let S be a nonrandomized stationary policy that prescribes the action (k_i, x_i) in state i, $i = 1, \cdots, L$.

DEFINITION. The sequence of policies $\{S_n\}$ is said to *converge* to the policy S if there exists a positive integer N such that $k_i^{(n)} = k_i$, $i = 1, \cdots, L$, for all $n \geqq N$ and $x_i^{(n)} \to x_i$ as $n \to \infty$.

THEOREM 7.1. *If for every nonrandomized stationary policy S, the Markov chain $\{X_n(S)\}$ is irreducible, then for any nonrandomized stationary policy S_1, the iteration scheme previously described generates a sequence of nonrandomized stationary policies $\{S_n\}$ converging to a nonrandomized stationary policy S^* that is optimal.*

Proof. For any policy S_n, $n = 1, 2, \cdots$, (7.1) can be rewritten as

$$(7.3)\qquad v_i^{S_n} + g^{S_n}\theta(i; S_n) = a_i^{\bar{S}_n}\theta(i; S_n) - c_i^{\bar{S}_n}N(i; S_n) + \sum p_{ij}^{\bar{S}_n} v_j^{S_n}, \qquad i = 1, \cdots, L,$$

where $g^{S_n} = I(S_n)$ and $\bar{S}_n$ refers only to the $\{k_i\}$ of $\{(k_i, x_i)\}$ of S_n.

As in [4, pp. 638–639], it can be shown that

$$(7.4)\qquad g^{S_{n+1}} \geqq g^{S_n}$$

and

$$(7.5)\qquad g^{S_{n+1}} - g^{S_n} = \frac{\sum_{i=1}^{L} \pi_i^{S_{n+1}}\theta(i; S_{n+1})\varepsilon_i^{S_n}}{\sum_{i=1}^{L} \pi_i^{S_{n+1}}\theta(i; S_{n+1})},$$

where $\{\pi_i^{S_n}\}$ are the stationary probabilities associated with the Markov chain $\{X_m(S_n)\}$ and

$$(7.6)\qquad \varepsilon_i^{S_n} = \max_{k,x}\left\{\frac{a_i^k\theta(i; k, x) - c_i^k N(i; k, x) + \sum p_{ij}^k v_j^{S_n} - v_i^{S_n}}{\theta(i; k, x)}\right\}$$
$$- \left\{\frac{a_i^{\bar{S}_n}\theta(i; S_n) - c_i^{\bar{S}_n}N(i; S_n) + \sum_j p_{ij}^{\bar{S}_n} v_j^{S_n} - v_i^{S_n}}{\theta(i; S_n)}\right\},$$
$$i = 1, \cdots, L, \quad n = 1, 2, \cdots.$$

Since $\{g^{S_n}\}$ is a nondecreasing sequence that is bounded above (an upper bound is $a^* = \max_{i,k} a_i^k$), we have $\lim_{n\to\infty} g^{S_n} = g^*$, say, and $\lim_{n\to\infty} g^{S_{n+1}} - g^{S_n} = 0$. Since $\pi_i^{S_{n+1}} > 0$ for all i and n and $\inf_n \theta(i; S_n) > 0$ for all i, from (7.5) it follows that $\varepsilon_i^{S_n} \to 0$ as $n \to \infty$. There exists a subsequence $\{S_{n_v}\}$ of $\{S_n\}$ such that

$$\lim_{v\to\infty} S_{n_v} = S^*, \quad \lim_{v\to\infty} g^{S_{n_v}} = g^*, \quad \lim_{v\to\infty} v_j^{S_{n_v}} = v_j^*, \qquad j = 1, \cdots, L.$$

For simplicity, we shall drop the subscript v and take the sequence $\{S_n\}$ to satisfy the preceding conditions. We shall show that S^* is optimal. Rewriting (7.3) and taking the limit as $n \to \infty$, we obtain

$$(7.7)\qquad g^* = \frac{a_i^{\bar{S}^*}\theta(i; S^*) - c_i^{\bar{S}^*}N(i; S^*) + \sum_j p_{ij}^{\bar{S}^*} v_j^* - v_i^*}{\theta(i; S^*)}, \qquad i = 1, \cdots, L.$$

We also have

$$(7.8)\qquad g^{S^*} = \frac{a_i^{\bar{S}^*}\theta(i; S^*) - c_i^{\bar{S}^*}N(i; S^*) + \sum_j p_{ij}^{\bar{S}^*} v_j^{S^*} - v_i^{S^*}}{\theta(i; S^*)}, \qquad i = 1, \cdots, L.$$

From (7.7) and (7.8) we conclude that $g^* = g^{S^*}$ and $v_j^* = v_j^{S^*}$, $j = 1, \cdots, L$. From (7.8) we obtain

$$(7.9)\quad g^{S^*} \leqq \max_{k,x}\left\{\frac{a_i^k\theta(i; k, x) - c_i^k N(i; k, x) + \sum_j p_{ij}^k v_j^{S^*} - v_i^{S^*}}{\theta(i; k, x)}\right\}, \qquad i = 1, \cdots, L.$$

By the definition of $\varepsilon_i^{S_n}$ given in (7.6), we have

$$g^{S_n} = \max_{k,x} \left\{ \frac{a_i^k\theta(i;k,x) - c_i^kN(i;k,x) + \sum_j p_{ij}^k v_j^{S_n} - v_i^{S_n}}{\theta(i;k,x)} \right\} - \varepsilon_i^{S_n}; \tag{7.10}$$

$$i = 1, \cdots, L, \quad n = 1, 2, \cdots.$$

Taking the limit as $n \to \infty$, we obtain

$$g^{S^*} = g^* \geqq \max_{k,x} \left\{ \frac{a_i^k\theta(i;k,x) - c_i^kN(i;k,x) + \sum_j p_{ij}^k v_j^{S^*} - v_i^{S^*}}{\theta(i;k,x)} \right\}, \tag{7.11}$$

$$i = 1, \cdots, L,$$

since $\varepsilon_i^{S_n} \to 0$ as $n \to \infty$. From (7.9) and (7.11), we conclude that g^{S^*} and $v_j^{S^*}$, $j = 1, \cdots, L$, are a solution of the system of equations

$$g = \max_{k,x} \left\{ \frac{a_i^k\theta(i;k,x) - c_i^kN(i;k,x) + \sum p_{ij}^k v_j - v_i}{\theta(i;k,x)} \right\}, \qquad i = 1, \cdots, L. \tag{7.12}$$

Further, from (7.8) we see that in each state i, $i = 1, \cdots, L$, S^* prescribes the action (k_i, x_i) that maximizes the right side of (7.12). Hence, as in Theorem 1.2 of [8], it can be shown that S^* is an optimal policy and $I(S^*) = g^{S^*}$. Hence the theorem.

Although we have shown that the sequence $\{S_n\}$ converges to an optimal policy S^*, we have not been able to say anything about the rate of convergence to the optimal policy. Further, we cannot say when the successive policies obtained by the described iteration method will prescribe the same k_i in state i and differ only in the x_i, $i = 1, \cdots, L$.

Theorem 7.1 has been proved under the strong condition that for every nonrandomized stationary policy S, the Markov chain $\{X_n(S)\}$ is irreducible. Suppose this condition is not satisfied. For any nonrandomized stationary policy S_1 such that the Markov chain $\{X_n(S_1)\}$ has only one positive class, consider the following modification of the iteration scheme just described. Let S'_1 be the policy obtained from S_1 by this iteration scheme. If $\{X_n(S'_1)\}$ has only one positive class, take $S_2 = S'_1$. If $\{X_n(S'_1)\}$ has more than one positive class, modify S'_1 as in the Appendix to obtain a policy S_2 such that $\{X_n(S_2)\}$ has only one positive class. Repeat the iteration procedure with S_2. Let $\{S_n\}$ be the sequence of policies generated by this modified iteration scheme. For each n, the Markov chain $\{X_m(S_n)\}$ has only one positive class and it can be shown that $I(S_{n+1}) \geqq I(S_n)$ and $\{S_n\}$ converges to a nonrandomized stationary policy S. We have not been able to show that S is optimal, although we conjecture that this is so.

Appendix. Let $\{X_t, t = 0, 1, \ldots\}$ be a Markov chain with a finite state space I and transition probability matrix $P = (p_{ij})$. We then have the following theorem.

THEOREM A.1 (K. L. Chung). *If the Markov chain $\{X_t\}$ is irreducible and $f(\cdot)$ is a function defined on the states, then*

$$p\text{-}\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} f(X_t) = \sum_{j \in I} \pi_j f(j) = M \quad (say), \tag{A.1}$$

where $\{\pi_j\}$ are the stationary probabilities associated with P, i.e., the $\{\pi_j\}$ are given by

$$\pi_j \geqq 0, \qquad j \in I,$$
$$\sum_{j \in I} \pi_j = 1, \tag{A.2}$$
$$\pi_j = \sum_{i \in I} \pi_i p_{ij}, \quad j \in I.$$

Equation (A.1) implies that

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T} Ef(X_t) = \sum_{j \in I} \pi_j f(j). \tag{A.3}$$

Now consider a Markov chain $\{X_t\}$ which is not irreducible but $A \subset I$ is a class of positive states and the rest of the states are transient. Then we can write

$$\frac{1}{T} \sum_{t=0}^{T} f(X_t) = \frac{1}{T} \sum_{t=0}^{\tau-1} f(X_t) + \frac{T-\tau}{T}\left(\frac{1}{T-\tau} \sum_{t=\tau}^{T} f(X_t)\right), \tag{A.4}$$

where τ is the first time the Markov chain enters the class A. If $\tau > T$, we take the first sum on the right side of the preceding equation to be from 0 to T and the second term to be zero. Now for a given τ, we have for Theorem A.1,

$$p\text{-}\lim_{T \to \infty} \frac{1}{T-\tau} \sum_{t=\tau}^{T} f(X_t) = \sum_{j \in A} \pi_j^* f(j) = M_A \quad \text{(say)},$$

where $\{\pi_j^*\}$ are the stationary probabilities associated with the reduced Markov chain with state space A. Note that M_A is independent of τ.

Further, since τ is finite with probability one,

$$p\text{-}\lim_{T \to \infty} \frac{T-\tau}{T} = 1 \qquad \text{and} \qquad p\text{-}\lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{\tau-1} f(X_t) = 0.$$

Hence we have

$$p\text{-}\lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T} f(X_t) = \sum_{j \in A} \pi_j^* f(j) = M_A. \tag{A.5}$$

Now consider a Markov chain $\{X_t\}$ such that its finite state space I consists of two disjoint positive subclasses A and B and the rest of the states are transient. Then (A.4) is still valid provided τ is now defined as the first time the Markov chain is absorbed either in the class A or the class B. Since τ is finite with probability one, we have

$$p\text{-}\lim_{T \to \infty} \frac{T-\tau}{T} = 1 \qquad \text{and} \qquad p\text{-}\lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{\tau-1} f(X_t) = 0.$$

Now, for a given τ and given that the Markov chain is absorbed in the class A at time τ, we have

$$p\text{-}\lim_{T \to \infty} \frac{1}{T-\tau} \sum_{t=\tau}^{T} f(X_t) = \sum_{j \in A} \pi_j^* f(j) = M_A. \tag{A.6}$$

Similarly, for a given τ and given that the Markov chain is absorbed in class B at time τ, we have

$$\text{(A.7)} \qquad p\text{-}\lim_{T\to\infty} \frac{1}{T-\tau} \sum_{t=\tau}^{T} f(X_t) = \sum_{j\in B} \pi_j^{*\prime} f(j) = M_B \qquad \text{(say)},$$

where $\{\pi_j^{*\prime}\}$ are the stationary probabilities associated with the reduced Markov chain with state space B. Let $p_A(p_B)$ be the probability that the Markov chain is absorbed in class A (B). We then have, from (A.6) and (A.7),

$$\text{(A.8)} \qquad \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} f(X_t) = \begin{cases} M_A \text{ with probability } p_A, \\ M_B \text{ with probability } p_B. \end{cases}$$

Now, for our problem, it has been remarked earlier (§ 3) that with every nonrandomized stationary policy S, there is an associated Markov chain $\{Y_n(S)\}$. $I(S)$, defined in (1.2), can be written as

$$\text{(A.9)} \qquad I(S) = \liminf_{N\to\infty} \frac{\sum_{n=1}^{N} Ef(Y_n(S))/N}{\sum_{n=1}^{N} Eg(Y_n(S))/N}.$$

If $\{Y_n(S)\}$ is an irreducible Markov chain using (A.3) in the numerator and denominator of (A.9), we obtain

$$\text{(A.10)} \qquad I(S) = \frac{\sum_{j\in I} \pi_j f(j)}{\sum_{j\in I} \pi_j g(j)} = M^* \quad \text{(say)}.$$

If $\{Y_n(S)\}$ is a Markov chain such that $A \subset I$ is a class of positive states and the rest of the states are transient, then, as in (A.5), we obtain

$$\text{(A.11)} \qquad I(S) = \frac{\sum_{j\in A} \pi_j^* f(j)}{\sum_{j\in A} \pi_j^* g(j)} = M_A^* \quad \text{(say)}.$$

Next if a stationary policy S is such that the Markov chain $\{Y_n(S)\}$ has two disjoint classes A and B of positive states and the rest of the states are transient, we have, as in (A.8),

$$\text{(A.12)} \qquad I(S) = p_A M_A^* + p_B M_B^* \qquad \text{(where } M_B^* \text{ is analogous to } M_B \text{ in (A.7))}.$$

Hence $I(S) \leqq \max(M_A^*, M_B^*)$. Suppose $M_A^* > M_B^*$.

In view of one of our assumptions (§ 1), there exists at least one nonrandomized stationary policy S_0 such that $\{Y_n(S_0)\}$ is an irreducible Markov chain. Let S' be a policy that is identical with S_0 until the Markov chain enters the class A and then S' is identical with S. For this policy S', we have, as in (A.10),

$$I(S') = M_A^* > I(S).$$

Thus, any nonrandomized stationary policy S which results in the associated Markov chain $\{Y_n(S)\}$ having more than one positive class need not be considered.

Acknowledgment. The author acknowledges with great pleasure the invaluable guidance and constant encouragement of Professor I. R. Savage during the course of this investigation.

REFERENCES

[1] G. R. Antelman, C. B. Russell and I. R. Savage, *Surveillance problems: two-dimensional with continuous surveillance*, this Journal, 5 (1967), pp. 245–267.

[2] Kai Lai Chung, *Markov Chains with Stationary Transition Probabilities*, 2nd ed., Springer, Berlin, 1966.

[3] Cyrus Derman, *On sequential decisions and Markov chains*, Management Sci., 9 (1962), pp. 16–24.

[4] Ronald Howard, *Semi-Markovian decision processes*, Bull. Inst. Internat. Statist., 40 (1963), pp. 625–652.

[5] William Jewell, *Markov-renewal programming. I: Formulation, finite return models*, Operations Res., 11 (1963), pp. 938–948.

[6] ———, *Markov-renewal programming. II: Infinite return models, example*. Ibid., 11 (1963), pp. 949–971.

[7] M. V. Johns, Jr. and R. G. Miller, Jr., *Average renewal loss rates*, Ann. Math. Statist., 34 (1963), pp. 396–401.

[8] S. Ross, *Arbitrary state Markovian decision processes*, Ibid., 39 (1968), pp. 2118–2122.

SIAM J. CONTROL
Vol. 7, No. 3, August 1969

A LYAPUNOV CRITERION FOR THE EXISTENCE OF STATIONARY PROBABILITY DISTRIBUTIONS FOR SYSTEMS PERTURBED BY NOISE*

MOSHE ZAKAI†

1. Introduction. Consider a randomly perturbed dynamical system described by the Itô stochastic differential equation

$$dx(t) = m(x(t))\,dt + G(x(t))\,dw(t), \tag{1}$$

where x and $m(x)$ are vectors in the Euclidean n-space E, $G(x)$ is an $n \times q$ matrix-valued function of x and $w(t)$ is the standard q-dimensional Brownian motion. We will assume throughout this note that m and G satisfy a global Lipschitz condition; namely, for all x, y in E,

$$|m(x) - m(y)| + |G(x) - G(y)| \leqq c|x - y|, \tag{2}$$

where, for vectors $|m| = (\sum_i m_i^2)^{1/2}$, and for matrices $|G| = (\sum_{i,j} G_{ij}^2)^{1/2}$. By $\mathscr{G}$ we will denote the differential operator associated with (1)

$$\mathscr{G} = \sum_1^n m_i(x)\frac{\partial}{\partial x_i} + \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n g_{ij}(x)\frac{\partial^2}{\partial x_i \partial x_j}, \tag{3}$$

where g_{ij} is the ijth element of GG'.

Following results of Kashminskii [1], Wonham established the following results [2], [3]. The results require that for some $c_1 > 0$, $y'G(x)G'(x)y \geqq c_1 y'y$ for all x, y in E; let $V(x)$ be nonnegative and twice continuously differentiable in $D = \{x; |x| > R\}$ for some $R < \infty$ and $V(x) \to \infty$ as $|x| \to \infty$; if $\mathscr{G}V(x) \leqq -1$ for all $x \in D$, then the process possesses an invariant probability measure [2]. In [3] it was shown that under the preceding conditions if, for all x in E, $\mathscr{G}V(x) \leqq k - L(x)$ and $L(x) \geqq 0$ then $E\{L(x(t))\} \leqq k$ where $E\{\cdot\}$ is the expectation with respect to the invariant measure. The results of [1] (and consequently those of [2] and [3]) were based on the assumption that the $x(t)$ process is strongly Feller (namely, for all $t > 0$ and all bounded and measurable functions $f(x)$, $E_x f(x_t)$ is continuous in x). It was shown in [2] that the requirement $y'GG'y \geqq c_1 y'y$ is sufficient to assure that the solution to (1) is strongly Feller. However, simple examples (such as $dx(t) = y(t)\,dt$, $dy(t) = dw(t)$) show that this is not a necessary condition; necessary and sufficient conditions for the solution of (1) to be strongly Feller seem to be unknown at present. Furthermore, the problem of the existence of invariant measures remains meaningful for processes which are not strongly Feller (for example, $dx(t) = 0$). Recently Beneš established several equivalent necessary and sufficient conditions for the existence of regular invariant probability measures for

* Received by the editors October 1, 1968, and in revised form February 24, 1969.

† Faculty of Electrical Engineering, TECHNION—Israel Institute of Technology, Haifa, Israel.

Feller (but not necessarily strongly Feller) processes.[1] Following the results of Beneš and the mean ergodic theorem, we will establish in this note (Theorems 1 and 2) the condition of [2] for the existence of an invariant probability measure[2] and the bound of [3] without the additional restrictions on GG'.

The requirement $y'G(x)G'(x)y \geqq c_1 y'y$ together with the existence of an invariant measure implies ([2], [1]) that the process $x(t)$ has useful ergodic [1, Theorem 3.1] and mixing [1, Theorem 3.4] properties. In many important applications this additional requirement on GG' is not satisfied; Theorem 3 gives other conditions under which $x(t)$ has these properties. Of particular interest in many important control vibrational and oscillation problems are equations of the type

$$\begin{aligned} &dx_1(t) = x_2(t)\,dt, \\ &dx_2(t) = x_3(t)\,dt, \\ &\quad \vdots \\ &dx_n(t) = m_n(x_1(t), x_2(t), \cdots, x_n(t))\,dt + dw_n(t). \end{aligned} \tag{4}$$

It is shown in Theorem 4 that if the process defined by (4) possesses an invariant probability measure, then the conditions of Theorem 3 are satisfied.

2. Finite invariant measure. In [4] Beneš considered Markov processes $x(t)$ taking values in a σ-compact metric space X and derived necessary and sufficient conditions for the existence of regular invariant probability measures under the following assumptions:

(a) $x(t)$ is a Feller process, i.e., the operators T_t defined by

$$T_t f(x) = \int_X f(y)P(t, x, dy)$$

take bounded continuous functions into bounded continuous ones.

(b) For any $t > 0$ and any compact K:

$$\lim_{|x| \to \infty} P(t, x, K) = 0. \tag{5}$$

One of the results of [4] is the following. *Let* $x(t)$ *satisfy* (a) *and* (b). *Then a necessary and sufficient condition for the existence of a regular invariant probability measure is the existence of a finite regular positive measure* μ^+ *on the state space of the process and a compact set K such that*

$$\limsup_{t \to \infty} \frac{1}{t} \int_0^t U_s \mu^+(K)\,ds > 0, \tag{6}$$

[1] We have been informed by a referee that the results of [4] have been improved by F. D. Santilles; Proc. Amer. Math. Soc., to appear.

[2] In a recent report [15] (which came to our attention after this note was written) Kushner replaced the restrictions on GG' by weaker assumptions. However, the main results of [15] are for equations of the particular form $dx = Ax\,dt + Bg(x)\,dt + B\,dw$, where A and B are $n \times n$ and $n \times q$ fixed matrices and $g(x)$ together with its first derivatives are bounded.

where

$$U_s\mu(K) = \int_X \mu(dx)P(s, x, K).$$

A condition analogous to (6) was derived by Oxtoby and Ulam for the nonstochastic case [5]; Foguel derived this condition for time discrete processes [6]. For $x(t)$ satisfying (1), $x(t)$ is a Feller process [7, Theorem 11.4] and in order to apply the criterion (6) to (1) we have to verify (5). Equation (5) implies, in the terminology of Dynkin, that $x(t)$ is a $\hat{C}$ process. It was shown in [7] (final assertion of Theorem 11.4) that for $x(t)$ satisfying (1) with $|m(x)|$ and $|G(x)|$ bounded, (5) is satisfied. We prove now that the same result but without the restriction on $|m|$ and $|G|$.

LEMMA. *Let $P(t, x, A)$ be the transition probability of the process satisfying* (1). *Then for every $t > 0$ and every compact set K in E,* (5) *holds.*

Proof. Obviously,

$$\text{Prob}\,\{|x(t)| < R \,|\, x(0) = x\} \leqq E_x\left\{\frac{2}{1 + |x(t)|^2/R^2}\right\}. \tag{7}$$

Let $\psi(x) = (1 + R^{-2}|x|^2)^{-1}$, applying Itô's formula [7, Theorem 7.2] and the vanishing of the expectation of the stochastic integral (or by Dynkin's formula, [7, Theorem 5.1]):

$$E_x\psi(x(t)) = \psi(x) + E_x \int_0^t \mathscr{G}\psi(x(s))\, ds. \tag{8}$$

It follows from (2) that

$$|m_i(x)| \leqq C_1(1 + |x|^2)^{1/2},$$
$$|g_{ij}(x)| \leqq C_2(1 + |x|^2),$$

and since $(\partial\psi/\partial x_i) = -2x_i\psi^2(x)R^{-2}$ and $(\partial^2\psi/\partial x_i\partial x_j) = 8x_ix_j\psi^3(x)R^{-4} - \delta_{ij}2\psi^2(x)R^{-2}$, we have for $R > 1$

$$|\mathscr{G}\psi(x)| \leqq \gamma\psi(x). \tag{9}$$

From (8) and (9) we have

$$\frac{d}{dt}\left(e^{-\gamma t}\int_0^t E_x\psi(x(s))\, ds\right) \leqq \psi(x)\, e^{-\gamma t}.$$

Therefore,

$$E_x\psi(x(t)) \leqq \psi(x)\, e^{\gamma t}; \tag{10}$$

and since $\psi(x) \to 0$ as $|x| \to \infty$, equation (5) follows from (10) and (7).

We shall consider functions $V(x)$ with the following properties:

(A) $V(x)$ is real-valued, nonnegative and twice continuously differentiable in E.

(B) Let $f(a, t)$ stand for any of the functions $E_aV(x(t))$, $E_a|\mathscr{G}V(x(t))|$, or $E_a|(\partial V(x(t))/\partial x_i)G_{ij}(x(t))|^2$. Then $f(a, t)$ is, for each a in E, bounded in t in any bounded t interval.

It follows directly from [8] that if $x(t)$ satisfies (1), then $E_a|x(t)|^p$, $p > 0$, is bounded in any bounded t interval. Therefore condition (*B*) is satisfied when $V, |\mathscr{G}V|$ and $|(\partial V/\partial x_i)G_{ij}|$ are dominated by polynomials. It should be pointed out that assumption (B) was not made in [2] and [3]; on the other hand the condition $V(x) \to \infty$ as $|x| \to \infty$ was imposed in [2] and is not required here.

THEOREM 1. *Let $x(t)$ satisfy* (1) *and $V(x)$ satisfy* (A) *and* (B); *also, assume that there exist numbers $R_0 < \infty$ and $k > 0$ such that*

$$\mathscr{G}V(x) \leqq -k \tag{11}$$

for all x satisfying $|x| > R_0$. Then the process defined by (1) *admits an invariant probability distribution, and if τ denotes the first passage time from x to the sphere $|x| \leqq R_0$ then for all x with $|x| \geqq R_0$*

$$E_x(\tau) \leqq \frac{V(x)}{k}. \tag{12}$$

Proof. By assumption (A) we may apply Itô's formula; therefore

$$V(x_t) = V(a) + \int_0^t \mathscr{G}V(x(s))\,ds + \int_0^t \left(\frac{\partial V(x(t))}{\partial x}\right)' G(x(t))\,dw(t).$$

By assumption (B), the expectation of the stochastic integral is zero and

$$E_aV(x(t)) = V(a) + \int_0^t E_a\mathscr{G}V(x(s))\,ds. \tag{13}$$

Let M be the maximum of $\mathscr{G}V(x)$ over $|x| \leqq R_0$ then, by (11)

$$\begin{aligned} E_a\mathscr{G}V(x(s)) &\leqq M\,\mathrm{Prob}\{|x(s)| \leqq R_0|x(0) = a\} - k\,\mathrm{Prob}\{|x(s)| > R_0|x(0) = a\} \\ &= -k + (M + k)\,\mathrm{Prob}\{|x(s)| \geqq R_0|x(0) = a\}. \end{aligned} \tag{14}$$

Let μ^+ be any finite Borel probability measure on E with a compact support ($\mu^+(K) = \mu^+(E) = 1$, K compact). Consider (1) with μ^+ as the distribution of $x(0)$, taking expectation over the initial distribution, (13) becomes

$$EV(x(t)) = EV(x(0)) + \int_0^t E\mathscr{G}V(x(s))\,ds,$$

then, since $V(x) \geqq 0$,

$$-\frac{1}{t}EV(x(0)) \leqq \frac{1}{t}\int_0^t E\mathscr{G}V(x(s))\,ds$$

and by (14)

$$-\frac{1}{t}EV(x(0)) + k < (M + k)\frac{1}{t}\int_0^t \mathrm{Prob}\,\{|x(s)| \leqq R\}\,ds.$$

The existence of an invariant probability follows by comparing the last equation with (6).

In order to prove (12) we consider $x(t \wedge \tau)$, the $x(t)$ process stopped at the boundary $|x| = R_0$. Applying Itô's formula to the integral representation of the stopped process [7, Theorem 11.6] we obtain, in analogy to (13),

$$E_xV(x(t \wedge \tau)) = V(x) + E_x \int_0^{t \wedge \tau} \mathscr{G}V(x(s))\,ds$$

$$\leqq V(x) - kE_x(t \wedge \tau).$$

Since $V(\cdot)$ is nonnegative, $E_x(t \wedge \tau) \leqq V(x)/k$ and (12) follows by monotone convergence.

Remarks. (i) It is easy to construct processes possessing an invariant probability measure but which do not satisfy (12) (for example, $dx(t) = 0$); it follows, therefore, that (11) is a sufficient but not necessary condition for the existence of an invariant probability. Note that if $x(t)$ is strongly Feller then (11) *is* necessary and sufficient [1]. (ii) In [9] we considered the condition $\mathscr{G}V \leqq k_1 - k_2V$ and $V(x) \to \infty$ as $|x| \to \infty$; therefore the conclusions of Theorem 1 apply to *all* the results of [9].

3. A Lyapunov method for estimating an expectation.

THEOREM 2. *Let $x(t)$ satisfy* (1). *If there exists a function $V(x)$ satisfying* (A) *and* (B) *and a positive constant k such that for all x in E*

$$\mathscr{G}V(x) \leqq k - L(x) \tag{15}$$

and $L(x) \geqq 0$; then, for any invariant probability measure $\nu(dx)$ of (1), *the stationary expectation $E\{L(x(s))\} = \int_E L(x)\nu(dx)$ satisfies*

$$E\{L(x(s))\} \leqq k. \tag{16}$$

Proof. By (13) and (15)

$$E_x \int_0^t L(x(s))\,ds \leqq V(x) + kt.$$

Let

$$L_R(x) = \begin{cases} L(x), & |x| \leqq R, \\ 0, & |x| < R. \end{cases}$$

Therefore,

$$\frac{1}{t}\int_0^t E_xL_R(x(s))\,ds \leqq \frac{V(x)}{t} + k$$

or

$$\limsup_{t \to \infty} \frac{1}{t}\int_0^t E_xL_R(x(s))\,ds \leqq k. \tag{17}$$

Let

$$f(x) = \int_0^1 \int_E L_R(y)P(t, x, dy)\,dt,$$

then $0 \leqq f(x) \leqq R$; since $E_x L_R(x(s))$ is measurable in x, s

$$\frac{1}{n}\int_0^n E_x L_R(x(s))\,ds = \frac{1}{n}\sum_{r=1}^{n} E_x f(x(r)), \tag{18}$$

and, since $v(dx)$ is an invariant measure, we also have

$$E\{f(x(s))\} = E\{L_R(x(s))\}. \tag{19}$$

By the mean ergodic theorem [10, p. 382]

$$\frac{1}{n}\sum_{r=1}^{n} E_x f(x(r)) \to f^*(x) \tag{20}$$

in the mean $(L_2(E, B, v))$ and

$$\int_E f(x)v(dx) = \int_E f^*(x)v(dx).$$

Considering a subsequence of n for which (20) holds for almost all $x(v(dx))$, it follows from (17), (18) and (20) that $E\{f(x(s))\} \leqq k$ and (16) follows from (19) and monotone convergence.

Remark. The result of Theorem 2 is for $E\{L(x)\}$ only; the following example shows that $E\{V(x)\}$ may be infinite. Let $dx(t) = -2x(t)(1 + x^2(t))^{-1}\,dt + dw(t)$ and $V(x) = x^4$, the invariant density is $\lambda(1 + x^2)^{-2}$; therefore $EV(x) = \infty$.

4. Some consequences of the existence of an invariant measure for a particular class of equations.

THEOREM 3. *Let the process* $x(t)$ *satisfy* (1) *and possess an invariant probability measure. Also assume that the transition probability function of* $x(t)$, $P(t, x, A)$, *is equivalent to the Lebesgue measure of* E *for all* $t > 0$ *and all* x *in* E. *Then,* (a) *the invariant probability measure is equivalent to the Lebesgue measure and is unique.* (b) *Let* $v'(x)$ *denote the density of the invariant measure (with respect to the Lebesgue measure) and let* $f(x)$ *denote a real-valued function integrable with respect to the invariant measure. Then, for all* $x \in E$ *and* $T \to \infty$

$$\text{Prob}\left\{\frac{1}{T}\int_0^T f(x(t))\,dt \to \int_E f(y)v'(y)\,dy \,\middle|\, x(0) = x\right\} = 1. \tag{21}$$

(c) *For any Borel set* A *in* E, *and any* $x \in E$ *as* $t \to \infty$

$$P(t, x, A) \to \int_A v'(x)\,dx. \tag{22}$$

Proof. Let $p(t, x, y)$ denote the density of $P(t, x, A)$ and let $v(A)$ be an invariant measure; then

$$v(A) = \int_E v(dx)\left(\int_A p(t, x, y)\,dy\right). \tag{23}$$

Since for each t, $p(t, x, y)$ can be chosen to be measurable in the pair x, y[11, p. 196 and supplement], Fubini's theorem is applicable to (23) and therefore $v(A)$ is absolutely continuous with respect to the Lebesgue measure. Conversely, if the

Lebesgue measure of A is positive, the $P(t, x, A)$ is positive for all $t > 0$ and $x \in E$, and it follows from (23) that $\nu(A) > 0$. This proves the equivalence between the Lebesgue measure and the invariant measure, and we will use "almost all x" without specifying the measure to mean almost all x with respect to $\nu(dx)$ and the Lebesgue measure. The uniqueness of the invariant measure will follow from (21).

Let A be an invariant set; namely for almost all x in A we have $P(t, x, A) = 1$. We also assume $\nu(A) \neq 0$. Therefore $P(t, x, A^c) = 0$, by the equivalence of the measures $P(t, x, A)$, the Lebesgue measure and $\nu(A)$ it follows that $\nu(A^c) = 0$; hence, $\nu(A) = 1$. Therefore, if A is invariant then $\nu(A) = 0$ or 1 and the stationary random function satisfying (1) with $x(0)$ distributed accordingly to $\nu(dx)$ (and independent of the $w(t)$ process) is metrically transitive, hence, ergodic. Therefore, (21) holds for almost all x in E. Let S be the set of x for which (21) is not true and let $x_0 \in S$; then the Lebesgue measure of S is zero; therefore Prob $\{x(1) \in S^c | x(0) = x_0\}$ $= 1$. But, if $x(1) \in S^c$ then Prob $\left\{(T-1)^{-1} \int_1^T f(x(s))\, ds \to \int f(x)\nu(dx) \middle| x(1) \in S^c\right\}$ $= 1$; hence (21) holds for all x. Equation (22) follows from [12, Theorem 5].

Remark. The assumption that the measures $P(t, x, A)$ are equivalent to the Lebesgue measure can be replaced by the assumption of "asymptotic equivalence," namely, that there exists an increasing sequence t_n such that if $P(t, x, A)$ $= P^{(1)}(t, x, A) + P^{(s)}(t, x, A)$, where $P^{(s)}$ is the singular part of $P(t, x, A)$ with respect to the Lebesgue measure, then $P^{(s)}(t_n, x, E) \to 0$ and similarly if $L^{(s)}(t, x, A)$ is the singular part of the Lebesgue measure with respect to $P(t, x, A)$ then $L^{(s)}(t_n, x, E)$ $\to 0$.

THEOREM 4. *For the process defined by* (4), *for all* $t > 0$, $x \in E$, $P(t, x, A)$ *is equivalent to the Lebesgue measure on E.*

Proof. Consider equation (4) and a modified equation obtained from (4) by setting $m_n(\cdot) = 0$; also assume $x(0) = x$ for both equations. By a result of Skorohod [13, Chap. 4, §4] the measure induced by (4) and the modified equation are equivalent. It follows that, in particular, the transition probabilities corresponding to (4) and to the modified equation are equivalent measures on E. In order to prove the theorem it is, therefore, sufficient to prove it for the special case $m_n(\cdot) = 0$.

Let

$$dx(t) = Cx(t)\, dt + B\, dw(t),$$

where C and B are constant $n \times n$ and $n \times q$ matrices respectively. It is easily verified that

$$x(t) = e^{Ct}x(0) + \int_0^t e^{C(t-s)}B\, dw(s)$$

and, therefore, $P(t, x, A)$ is Gaussian with mean $e^{Ct}x$ and covariance matrix

$$\int_0^t e^{C(t-s)}BB'\, e^{C'(t-s)}\, ds. \tag{24}$$

Therefore, $P(t, x, A)$ is equivalent to the Lebesgue measure on E if and only if (24) is positive definite. Hence, $P(t, x, A)$ is equivalent to the Lebesgue measure if and only if the pair (C, B) is controllable (namely, if the column vectors of the $n \times n^2$

matrix $[B, CB, \cdots, C^{n-1}B]$ span the space E) [14, p. 499]. It follows by a direct calculation that for (4) with $m_n(\cdot) = 0$, the pair (C, B) is controllable which proves the theorem.

Theorem 4 can easily be generalized to a collection of subsystems of the form of (4) and coupled through the $m_n(\cdot)$ term. To avoid clumsy notation we give a representative example of this generalization:

$$\begin{aligned} dx_1(t) &= x_2(t)\,dt, \\ dx_2(t) &= m_2(x_1(t), x_2, x_3, x_4)\,dt + dw_2(t), \\ dx_3(t) &= x_4(t)\,dt, \\ dx_4(t) &= m_4(x_1, x_2, x_3, x_4)\,dt + dw_4(t). \end{aligned}$$

REFERENCES

[1] R. Z. Khasminskii, *Ergodic properties of recurrent diffusion processes and stabilization of the solution of the Cauchy problem for parabolic equations*, Theor. Probability Appl., 5 (1960), pp. 179–196.

[2] W. M. Wonham, *Lyapunov criteria for weak stochastic stability*, J. Differential Equations, 2 (1966), pp. 195–207.

[3]———, *A Lyapunov method for the estimation of statistical averages*, Ibid., 2 (1966), pp. 365–377.

[4] V. E. Beneš, *Finite invariant measures for Feller processes*, J. Appl. Probability, 5 (1968), pp. 203–209.

[5] J. C. Oxtoby and S. M. Ulam, *On the existence of a measure invariant under a transformation*, Ann. of Math., 40 (1939), pp. 560–566.

[6] S. R. Foguel, *Existence of invariant measures for Markov process II*, Proc. Amer. Math. Soc., 17 (1966), pp. 387–389.

[7] E. B. Dynkin, *Markov Processes*, Springer, Berlin, 1965.

[8] M. Zakai, *Some moment inequalities for stochastic integrals and solutions to stochastic differential equations*, Israel J. Math., 5 (1967), pp. 170–176.

[9] ———, *On the ultimate boundedness of moments associated with solutions of stochastic differential equations*, this Journal, 5 (1967), pp. 588–593.

[10] K. Yosida, *Functional Analysis*, Springer, Berlin, 1965.

[11] J. L. Doob, *Stochastic Process*, John Wiley, New York, 1953.

[12] ———, *Asymptotic properties of Markoff transition probabilities*, Trans. Amer. Math. Soc., 63 (1948), pp. 393–421.

[13] A. V. Skorohod, *Studies in the Theory of Random Processes*, Addison-Wesley, New York, 1965.

[14] L. A. Zadeh and C. A. Desoer, *Linear System Theory*, McGraw-Hill, New York, 1963.

[15] H. J. Kushner, *The Cauchy problem for a class of degenerate parabolic equations and asymptotic properties of the related diffusion processes*, Tech. Rep. 68-3, Center for Dynamical Systems, Division of Applied Mathematics, Brown University, 1968.

SIAM J. CONTROL
Vol. 7, No. 3, August 1969

NEW RESULTS IN LINEAR SYSTEM STABILITY*

B. D. O. ANDERSON AND J. B. MOORE†

Abstract. This paper considers connections between bounded-input, bounded-output stability and asymptotic stability in the sense of Lyapunov for linear time-varying systems. By modifying slightly the definition of bounded-input, bounded-output stability, an equivalence between the two types of stability is found for systems which are uniformly completely controllable and observable. The various matrices describing the system need not be bounded. Other results relate to the characterization of uniform complete controllability and the derivation of Lyapunov functions for linear time-varying systems.

1. Introduction. Connections between various types of stability are examined in this paper. More precisely, we study linear, finite-dimensional, dynamical systems which in general are time-varying, and consider descriptions of such systems of the form

$$\frac{d}{dt}x(t) = F(t)x(t) + G(t)u(t), \tag{1a}$$

$$y(t) = H'(t)x(t). \tag{1b}$$

Here x is an n-dimensional real column vector, u is a p-dimensional real control vector, y is an m-dimensional real output vector, and $F(t)$, $G(t)$ and $H(t)$ are matrices of real continuous functions with appropriate dimension. It is also assumed that every component of the control vector function $u(\cdot)$ is piecewise continuous. All these constraints will not be explicitly stated in the sequel, but hold throughout the paper.

Under zero-input conditions, the internal stability of (1) may be examined. This internal, or Lyapunov, stability considers the effect of variations in initial conditions on the subsequent trajectory of the homogeneous system

$$\frac{d}{dt}x(t) = F(t)x(t). \tag{2}$$

Obviously, internal stability properties of (1) do not depend at all on the $G(\cdot)$ and $H(\cdot)$ matrices.

The external stability of (1) may be examined by considering the effect on the output of inputs from some restricted class; commonly we may be interested in whether a bounded input will produce a bounded output when the initial state is taken as zero. We are thus really concerned with properties of the weighting function matrix

$$W(t, \tau) = H'(t)\Phi(t, \tau)G(\tau), \tag{3}$$

where $\Phi(\cdot, \cdot)$ is the transition matrix associated with (2); this is because, under zero initial state conditions, $y(\cdot)$ in (1) is related to $u(\cdot)$, assumed zero prior to

* Received by the editors April 29, 1968, and in final revised form December 26, 1968.

† Department of Electrical Engineering, University of Newcastle, New South Wales, 2308, Australia. This work was supported by the Australian Research Grants Committee.

time t_0, through

$$y(t) = \int_{t_0}^{t} W(t, \tau)u(\tau)\, d\tau. \tag{4}$$

The natural question arises as to whether there are connections between external and internal stability.

Without further constraints on the matrices in (1), the answer is no [1]. This is because knowledge of $W(\cdot, \cdot)$ in (3) conveys no knowledge at all about $\Phi(\cdot, \cdot)$ and thus $F(\cdot)$. In fact, a so-called separable $W(t, \tau)$ may be realized as the impulse response of a system of the form (1), with the F matrix being quite arbitrary, except for a constraint on its order.

In an effort to obtain connections between internal and external stability, various extra constraints can be used. When $W(\cdot, \cdot)$ is time invariant, in the sense that $W(t, \tau) = W(t - \tau)$, the natural constraint to impose on F, G and H is that they may be time invariant. Then it can be shown that if the eigenvalues of F all possess negative real parts, corresponding to exponential asymptotic stability in the sense of Lyapunov (abbreviated EAS), the system (1) is bounded-input, bounded-output stable (abbreviated BIBO). Conversely, if (1) is BIBO and completely controllable and observable, then it is EAS.

For time-varying systems, it is not so clear what constraints should be imposed in order to yield equivalences between the two types of stability. Amongst constraints which have been used, we note those implicit in Perron's work [2]. He was essentially concerned just with (1a) and found conditions such that EAS led to bounded-input, bounded-state stability (BIBS). (A system is BIBS if, with the states as outputs, it is BIBO.) His conclusion was that with the elements of F and G bounded, and with G possessing an inverse with bounded elements, EAS and BIBS were equivalent. The nonsingularity of G constituted a major drawback; in [3], the difficulty was partly removed by showing that with G a column vector, consisting of all zeros save for a one in the last place, and F in companion matrix form, EAS and BIBS were equivalent. These special forms of F and G were shown to arise naturally from the representation of some differential equations in the form of (1).

More significant are the results of [4], which essentially include those of [2] and [3]. The initial restriction is made that the elements of F, G and H are bounded. The following results are then demonstrated:

(a) EAS implies BIBS and BIBO;

(b) BIBS and uniform complete controllability (see [4]) imply EAS;

(c) BIBO and uniform complete observability (see [4]) imply BIBS.

Thus under the boundedness assumption, EAS implies BIBO, and with the additional assumptions of uniform complete controllability and uniform complete observability, BIBO implies EAS.

In this paper we improve on the results of [4]; we are mainly concerned with eliminating the boundedness requirement on the elements of F, G and H. It turns out that to do this at the same time to obtain meaningful results, it is necessary to modify the requirement of boundedness of input and output in the BIBO definition; in this modification, it is required that the "energy content" over a fixed-length interval of the input and output should be bounded, independently of the position of the interval. The principal conclusions are then that internal (EAS) stability and external stability, in an appropriately modified form, are equivalent under uniform complete controllability and observability.

Section 2 is concerned with definitions and a preliminary lemma. Included in this section are precise statements of what we mean by the modified form of boundedness discussed above and a review of the uniform complete controllability and observability concepts.

Section 3 is concerned with (1a); in this section EAS is related to a modified form of BIBS stability. Section 4 examines the system defined by both (1a) and (1b) and achieves results relating EAS to modified BIBO stability.

Finally, two related results are presented in § 5. The first establishes a class of state feedback laws under which uniform complete controllability is invariant; the second result presents a time-varying version of a lemma due to Lyapunov which is well known for time-invariant systems.

2. Definitions and preliminaries. The concepts of uniform complete controllability and uniform complete observability appear to have been introduced originally in [5], in order to guarantee the solution of certain time-variable quadratic variational problems. Equation (1a) is uniformly completely controllable, or the pair $[F(t), G(t)]$ is uniformly completely controllable, if any two of the following three conditions hold for some $\delta_c > 0$ (any two imply the third, see [5]):

$$\alpha_1 I \leqq M(s - \delta_c, s) \leqq \alpha_2 I \quad \text{for all } s, \tag{5}$$

$$\alpha_3 I \leqq \Phi(s - \delta_c, s) M(s - \delta_c, s) \Phi'(s - \delta_c, s) \leqq \alpha_4 I \quad \text{for all } s, \tag{6}$$

$$\|\Phi(t, \tau)\| < \alpha_5(|t - \tau|) \quad \text{for all } t \text{ and } \tau, \tag{7}$$

where

$$M(s - \delta_c, s) = \int_{s-\delta_c}^{s} \Phi(s, t) G(t) G'(t) \Phi'(s, t)\, dt. \tag{8}$$

The quantities α_1, α_2, α_3 and α_4 are positive constants, and $\alpha_5(\cdot)$ maps R into R and is bounded on bounded intervals. The notation $X \geqq Y$ $(X > Y)$ for symmetric matrices X and Y means $X - Y$ is nonnegative (positive) definite. For an n-dimensional vector x, $\|x\|$ is $(\sum x_i^2)^{1/2}$; the usual induced matrix norm applies.

Several points should be noted; first, a sufficient condition for (7) is that $F(\cdot)$ should be bounded; one way to see this is to use the Gronwall-Bellman inequality [6]. Second, if (5), (6) and (7) hold for some δ_c, they hold for all $\delta > \delta_c$ (see [5]). Third, there is a consequence of the right-hand inequality of (5) which will be of use. It is based on the inequality

$$\|Ax\|^2 \leqq \|A'A\|\,\|x\|^2 = \lambda_{\max}(A'A)\|x\|^2 \leqq (\operatorname{tr} A'A)\|x\|^2. \tag{9}$$

This consequence, following from (5), (8) and (9), is

$$\int_{s-\delta_c}^{s} \|\Phi(s, t) G(t)\|^2\, dt \leqq n\alpha_2. \tag{10}$$

Uniform complete observability is defined for the pair of equations (1a) and (1b), or the matrix pair $[F, H]$. The system of equations (1) is uniformly completely observable if any two of the following three conditions hold for some $\delta_0 > 0$ (again, any two imply the third [5]):

$$\alpha_6 I < N(s, s + \delta_0) < \alpha_7 I \quad \text{for all } s, \tag{11}$$

$$\alpha_8 I \leqq \Phi'(s, s + \delta_0) N(s, s + \delta_0) \Phi(s, s + \delta_0) \leqq \alpha_9 I \quad \text{for all } s, \tag{12}$$

$$\|\Phi(t, \tau)\| \leqq \alpha_5(|t - \tau|) \quad \text{for all } t \text{ and } \tau,$$

where

$$N(s, s + \delta_0) = \int_s^{s+\delta_0} \Phi'(t, s) H(t) H'(t) \Phi(t, s)\, dt. \tag{13}$$

The quantities α_6, α_7, α_8 and α_9 are positive constants.

The remarks made above concerning uniform complete controllability carry over mutatis mutandis to uniform complete observability.

One of the consequences of uniform complete controllability is contained in the following lemma, a minor variant on a result of [4].

LEMMA 1. *The realization* (1a) *is uniformly completely controllable if and only if there exist a $\delta_c > 0$ such that for every state ξ and for any time s, there exists a minimal energy input u_1 transforming the system* (1a) *from the zero state at time $s - \delta_c$ to the state ξ at time s, and a minimal energy input u_2 transferring* (1a) *from the state ξ at time $s - \delta_c$ to the zero state at time s, such that for positive constants $\alpha_{10}, \alpha_{11}, \alpha_{12}, \alpha_{13}$,*

$$\alpha_{10}\|\xi\|^2 \leqq \int_{s-\delta_c}^{s} u_1'(t) u_1(t)\, dt \leqq \alpha_{11}\|\xi\|^2, \tag{14a}$$

$$\alpha_{12}\|\xi\|^2 \leqq \int_{s-\delta_c}^{s} u_2'(t) u_2(t)\, dt \leqq \alpha_{13}\|\xi\|^2. \tag{14b}$$

The energy associated with u_1 over $(s - \delta_c, s)$ is the value of the integral appearing in (14a); *u_1 is a minimal energy input if no other input taking the zero state at time $s - \delta_c$ to the state ξ at time s has an associated smaller energy.*

Proof. Suppose the realization is uniformly completely controllable. Now from [7], $M(s - \delta_c, s)$ is nonsingular and there exist minimal energy controls u_1 and u_2 achieving the desired state transfer. The controls u_1 and u_2 are uniquely defined, except for a set of measure zero, by

$$u_1(t) = u_2(t) = 0, \qquad t < s - \delta_c, \quad t > s; \tag{15a}$$

$$u_1(t) = G'(t)\Phi'(s, t) M^{-1}(s - \delta_c, s)\xi, \qquad s - \delta_c \leqq t \leqq s; \tag{15b}$$

$$u_2(t) = -G'(t)\Phi'(s, t) M^{-1}(s - \delta_c, s)\Phi(s, s - \delta_c)\xi, \qquad s - \delta_c \leqq t \leqq s. \tag{15c}$$

The fact that $u_1(\cdot)$ and $u_2(\cdot)$ will effect the transferral is readily established using (15), (8) and the formulas

$$\xi = \int_{s-\delta_c}^{s} \Phi(s, \tau) G(\tau) u_1(\tau)\, d\tau, \tag{16a}$$

$$0 = \Phi(s, s - \delta_c)\xi + \int_{s-\delta_c}^{s} \Phi(s, \tau) G(\tau) u_2(\tau)\, d\tau. \tag{16b}$$

Application of (8), (15) and (16) yields

$$\int_{s-\delta_c}^{s} u_1'(t)u_1(t)\,dt = \xi' M^{-1}(s-\delta_c, s)\xi, \tag{17a}$$

$$\int_{s-\delta_c}^{s} u_2'(t)u_2(t)\,dt = \xi\Phi'(s, s-\delta_c)M^{-1}(s-\delta_c, s)\Phi(s, s-\delta_c)\xi. \tag{17b}$$

Equations (5), (6) and (17) then imply that (14a) and (14b) are satisfied.

Conversely, suppose existence of minimal energy controls satisfying (14). Controllability of the state ξ for arbitrary ξ implies $M(s-\delta_c, s)$ is nonsingular, which in turn implies that the minimal energy controls are unique (except on a set of measure zero) and are given by (15) (see [7]). Equations (14) and (17) now hold simultaneously and imply (5) and (6).

The form of (15b) suggests in contrast to [4] that to hope for the existence of a control bounded only in terms of ξ which effected the state transferral would be too much, at least when G, say, is not assumed bounded. This together with (14) suggests that to discuss the external stability of (1), the normally assumed boundedness of the input or output should be replaced by the following definition.

DEFINITION. The vector function $w(\cdot)$ with piecewise continuous components is termed *bounded** if, for some positive δ_b and all s,

$$\int_{s-\delta_b}^{s} w'(t)w(t)\,dt \leqq \alpha_{14}, \tag{18}$$

where α_{14} is a positive constant.

Of course, if $w(\cdot)$ is bounded in the usual sense, $w(\cdot)$ is then bounded*. It should also be noted that if (18) holds for some δ_b it holds for all positive δ, greater or less than δ_b (with, in the case of $\delta > \delta_b$, α_{14} perhaps being replaced by a greater constant depending on δ).

Analogously to the abbreviation BIBO for bounded-input, bounded-output stability, we shall use the abbreviation B*IB*O to denote bounded*-input, bounded*-output stability. Thus a system is B*IB*O if for all inputs $u(\cdot)$ such that

$$\int_{s-\delta_b}^{s} u'(t)u(t)\,dt \leqq \alpha_{14} \tag{19}$$

for some α_{14}, some δ_b and all s, there exists α_{15}, depending on α_{14} and δ_b, with the zero-state response $y(\cdot)$ satisfying

$$\int_{s-\delta_b}^{s} y'(t)y(t)\,dt \leqq \alpha_{15}(\alpha_{14}, \delta_b) \tag{20}$$

for all s. (Note that earlier stated constraints guarantee that the components of $y(\cdot)$ are piecewise continuous.)

The definition of B*IBS proceeds analogously to the definition of BIBS, modification being made to the class of inputs considered. Thus for all inputs satisfying (19), we require the existence of a constant α_{16}, depending on α_{14} and δ_b, such that

$$\|x(t)\| \leqq \alpha_{16}(\alpha_{14}, \delta_b) \quad \text{for all } t \tag{21}$$

when the system is initially in the zero state.

LEMMA 2. *The system* (1) *is B*IBS if and only if for all bounded* inputs satisfying* (19), *there exists a constant* α_{17}, *depending on* α_{14} *and* δ_b, *for which*

$$(22) \qquad \int_{-\infty}^{t} \|\Phi(t,\tau)G(\tau)u(\tau)\|\,d\tau \leqq \alpha_{17} \quad \textit{for all } t.$$

Proof. We first show that (22) implies B*IBS. Suppose the system is in the zero state at initial time t_0. Then

$$\begin{aligned}\|x(t)\| &\leqq \left\|\int_{t_0}^{t} \Phi(t,\tau)G(\tau)u(\tau)\,d\tau\right\| \\ &\leqq \int_{-\infty}^{t} \|\Phi(t,\tau)G(\tau)u(\tau)\|\,d\tau \\ &\leqq \alpha_{17}.\end{aligned}$$

Now suppose (1a) is B*IBS, with (21) holding; suppose too that (22) fails. Then there exist times t_0 and t_1 and a bounded* control u (satisfying (19)) such that

$$\int_{t_0}^{t_1} \|\Phi(t_1,\tau)G(\tau)u(\tau)\|\,d\tau > \sqrt{n}\alpha_{16};$$

and then for some i, say $i = I$,

$$\int_{t_0}^{t_1} |(\Phi(t_1,\tau)G(\tau)u(\tau))_I|\,d\tau = \int_{t_0}^{t_1} \left|\sum_{k,l} \Phi_{Ik}(t_1,\tau)G_{kl}(\tau)u_l(\tau)\right| d\tau > \alpha_{16}.$$

Now define $\hat{u}_l(\cdot)$ by

$$\hat{u}_l(\tau) = u_l(\tau)\left[\operatorname{sgn}\left\{\sum_k \Phi_{Ik}(t_1,\tau)G_{kl}(\tau)u_l(\tau)\right\}\right],$$

arbitrarily taking sgn $\{0\} = 1$ if required. Then $\hat{u}(\cdot)$, the vector with lth entry $\hat{u}_l(\cdot)$, is bounded* because $\hat{u}(\cdot)$ is, and the-same constants δ_b and α_{14} apply. Also, the response $\hat{x}(\cdot)$ to $\hat{u}(\cdot)$ has

$$\begin{aligned}\hat{x}_I(t_1) &= \int_{t_0}^{t_1} \sum_{k,l} \Phi_{Ik}(t_1,\tau)G_{kl}(\tau)\hat{u}_l(\tau)\,d\tau \\ &= \int_{t_0}^{t_1} \left|\sum_{k,l} \Phi_{Ik}(t_1,\tau)G_{kl}(\tau)u_l(\tau)\right| d\tau \\ &> \alpha_{16}.\end{aligned}$$

This contradicts (21), i.e., the assumption that the system is B*IBS. Thus the lemma is proved.

3. Relations between Lyapunov and bounded*-input, bounded-state stability. In this section, attention is focused on (1a). By analogy with time-invariant systems, we seek relations between external stability and internal *asymptotic* stability; in time-invariant systems the asymptotic stability, because it is uniform, is also exponential. Here also, it is convenient to specialize to exponential asymptotic stability. The main result is contained in the following theorem.

THEOREM 1. *Suppose* (1a) *is uniformly completely controllable. Then it is B*IBS if and only if it is EAS.*

Proof. We show first that EAS implies B*IBS. Suppose the system is excited with a bounded* input commencing at time t_0, being initially in the zero state. Then

$$\begin{aligned} x(t) &= \int_{t_0}^{t} \Phi(t,\tau)G(\tau)u(\tau)\,d\tau \\ &= \Phi(t, t_0 + \delta_c)\int_{t_0}^{t_0+\delta_c} \Phi(t_0 + \delta_c, \tau)G(\tau)u(\tau)\,d\tau \\ &\quad + \Phi(t, t_0 + 2\delta_c)\int_{t_0+\delta_c}^{t_0+2\delta_c} \Phi(t_0 + 2\delta_c, \tau)G(\tau)u(\tau)\,d\tau + \cdots \\ &\quad + \Phi(t, t_0 + k\delta_c)\int_{t_0+(k-1)\delta_c}^{t_0+k\delta_c} \Phi(t_0 + k\delta_c, \tau)G(\tau)u(\tau)\,d\tau \\ &\quad + \int_{t_0+k\delta_c}^{t} \Phi(t,\tau)G(\tau)u(\tau)\,d\tau \end{aligned} \tag{23}$$

with the integer k being chosen so that $0 < t - (t_0 + k\delta_c) \leqq \delta_c$. Consider now the following sequence of inequalities for a typical integral on the right of (23):

$$\begin{aligned} &\left|\int_{t_0+(j-1)\delta_c}^{t_0+j\delta_c} \Phi(t_0 + j\delta_c, \tau)G(\tau)u(\tau)\,d\tau\right| \\ &\leqq \left\{\int_{t_0+(j-1)\delta_c}^{t_0+j\delta_c} \|\Phi(t_0 + j\delta_c, \tau)G(\tau)\|^2\,d\tau\right\}^{1/2} \left\{\int_{t_0+(j-1)\delta_c}^{t_0+j\delta_c} u'(\tau)u(\tau)\,d\tau\right\}^{1/2} \\ &\leqq (n\alpha_2\alpha_{14})^{1/2}. \end{aligned} \tag{24a}$$

Here we have identified, as is legitimate, the δ_b of the bounded* definition with the δ_c of the uniform complete controllability definition; as earlier pointed out, if a vector function satisfies (19) for one pair δ_b, α_{14}, it will satisfy it for arbitrary positive δ_b and some new α_{14}.

Because $t - (t_0 + k\delta_c) \leqq \delta_c$, the same bound exists on the absolute value of the last integral in (23) as on the first k integrals. Also

$$\begin{aligned} \|\Phi(t, t_0 + j\delta_c)\| &\leqq \|\Phi(t, t_0 + k\delta_c)\|\,\|\Phi(t_0 + k\delta_c, t_0 + j\delta_c)\| \\ &\leqq \|\Phi(t, t_0 + k\delta_c)\|\alpha_{18}\exp\{-\alpha_{19}(k - j)\delta_c\} \end{aligned} \tag{24b}$$

for some positive constants α_{18}, α_{19} existing because of the EAS assumption. Using (24a) and (24b) in (23), we then have

$$\begin{aligned} \|x(t)\| &\leqq (n\alpha_2\alpha_{14})^{1/2}\alpha_{18}\|\Phi(t, t_0 + k\delta_c)\| \\ &\quad \cdot[\exp\{-\alpha_{19}(k - 1)\delta_c\} + \exp\{-\alpha_{19}(k - 2)\delta_c\} + \cdots + 1]. \end{aligned} \tag{25}$$

The geometric series has a sum bounded independently of k. Also, because $0 \leqq t - (t_0 + k\delta_c) \leqq \delta_c$, $\|\Phi(t, t_0 + k\delta_c)\|$ is bounded independently of t, t_0 and k; hence $\|x(t)\|$ is bounded, as required.

We now turn to proving that B*IBS implies EAS. Let $\lambda(\cdot)$ be a vector function such that $\lambda(s)$ has unit norm for all s. By uniform complete controllability, there exists a control $u_s(\cdot)$ taking the zero state at time $s - \delta_c$ to the state $\lambda(s)$ at time s. One such control is given by (see (15)):

$$u_s(t) = 0, \qquad t < s - \delta_c, \quad t > s, \tag{26a}$$

$$u_s(t) = G'(t)\Phi'(s,t)M^{-1}(s - \delta_c, s)\lambda(s), \qquad s - \delta_c \leqq t \leqq s. \tag{26b}$$

Then

$$\lambda(s) = \int_{s-\delta_c}^{s} \Phi(s,\tau)G(\tau)u_s(\tau)\,d\tau$$

and thus

$$\|\Phi(t,s)\lambda(s)\| \leqq \int_{s-\delta_c}^{s} \|\Phi(t,\tau)G(\tau)u_s(\tau)\|\,d\tau.$$

Integrating with respect to s, we have

$$\int_{t_0}^{t} \|\Phi(t,s)\lambda(s)\|\,ds \leqq \int_{t_0}^{t} ds \int_{s-\delta_c}^{s} \|\Phi(t,\tau)G(\tau)u_s(\tau)\|\,d\tau.$$

By defining a new variable $r = \tau - s + \delta_c$, it follows that

$$\begin{aligned}\int_{t_0}^{t} \|\Phi(t,s)\lambda(s)\|\,ds &\leqq \int_{t_0}^{t} ds \int_{0}^{\delta_c} \|\Phi(t, r + s - \delta_c)G(r + s - \delta_c)u_s(r + s - \delta_c)\|\,dr \\ &= \int_{0}^{\delta_c} dr \int_{t_0}^{t} \|\Phi(t, r + s - \delta_c)G(r + s - \delta_c)u_s(r + s - \delta_c)\|\,ds.\end{aligned}$$

Now define a new variable again by $\tau = r + s - \delta_c$ to obtain

$$\int_{t_0}^{t} \|\Phi(t,s)\lambda(s)\|\,ds \leqq \int_{0}^{\delta_c} dr \int_{t_0+r-\delta_c}^{t+r-\delta_c} \|\Phi(t,\tau)G(\tau)u_{\tau-r+\delta_c}(\tau)\|\,d\tau. \tag{27}$$

Our aim is to demonstrate that the right-hand side of this inequality is bounded. Note that

$$\int_{t_0+r-\delta_c}^{t+r-\delta_c} \|\Phi(t,\tau)G(\tau)u_{\tau-r+\delta_c}(\tau)\|\,d\tau \leqq \int_{t_0+r-\delta_c}^{t} \|\Phi(t,\tau)G(\tau)u_{\tau-r+\delta_c}(\tau)\|\,d\tau \tag{28}$$

because, as is evident from the interval of integration with respect to r in (27), $0 \leqq r \leqq \delta_c$, and so $t + r - \delta_c \leqq t$.

From Lemma 2, it follows that the right-hand side of (28) is bounded if $v_r(\tau)$, defined by $v_r(\tau) = u_{\tau-r+\delta_c}(\tau)$, is bounded* for fixed arbitrary r. (Note that for fixed s, $u_s(\tau)$ is a bounded* function of τ, but this certainly does not itself imply that $u_{\tau-r+\delta_c}(\tau) = v_r(\tau)$ is bounded*.)

An explicit formula is available for $v_r(\tau)$, following from (26), for all τ:

$$v_r(\tau) = G'(\tau)\Phi'(\tau - r + \delta_c, \tau)M^{-1}(\tau - r, \tau - r + \delta_c)\lambda(\tau - r + \delta_c), \quad 0 \leqq r \leqq \delta_c.$$

Evidently for arbitrary s,

$$\begin{aligned}\int_{s-\delta_c}^{s} v_r'(\tau)v_r(\tau)\,d\tau &= \int_{s-\delta_c}^{s} \{\lambda'(\tau - r + \delta_c)M^{-1}(\tau - r, \tau - r + \delta_c)\\ &\qquad \cdot \Phi(\tau - r + \delta_c, \tau)G(\tau)G'(\tau)\Phi'(\tau - r + \delta_c, \tau)\\ &\qquad\qquad \cdot M^{-1}(\tau - r, \tau - r + \delta_c)\lambda(\tau - r + \delta_c)\}\,d\tau\\ &\leqq \frac{1}{\alpha_1^2}\int_{s-\delta_c}^{s} \|\Phi(\tau - r + \delta_c, s)\Phi(s,\tau)G(\tau)G'(\tau)\Phi'(s,\tau)\\ &\qquad\qquad \cdot \Phi'(\tau - r + \delta_c, s)\|\,d\tau\\ &\leqq \frac{1}{\alpha_1^2}\sup_{\substack{0 \leqq r \leqq \delta_c\\ s-\delta_c \leqq \tau \leqq s}} \{\|\Phi(\tau - r + \delta_c, s)\|^2\}\int_{s-\delta_c}^{s}\|\Phi(s,\tau)G(\tau)\|^2\,d\tau\\ &\leqq \frac{n\alpha_2}{\alpha_1^2}\left\{\sup_{0 \leqq \rho \leqq \delta_c}|\alpha_5(\rho)|\right\}^2.\end{aligned}$$

This bound is evidently independent of s and r.

Hence, by Lemma 2, for some positive constant α_{17}, independent of t, t_0 and r,

$$\int_{t_0+r-\delta_c}^{t+r-\delta_c} \|\Phi(t,\tau)G(\tau)u_{\tau-r+\delta_c}(\tau)\|\,d\tau \leqq \alpha_{17},$$

and thus in (27),

$$\int_{t_0}^{t} \|\Phi(t,s)\lambda(s)\,ds\| \leqq \delta_c\alpha_{17}.$$

Since $\lambda(s)$ in the above derivation has only been restricted to have unit norm, we may at this stage further restrict it so that $\|\Phi(t,s)\lambda(s)\|$ is maximized. Since this maximum is precisely $\|\Phi(t,s)\|$, we then have

$$\int_{t_0}^{t} \|\Phi(t,s)\|\,ds \leqq \delta_c\alpha_{17}. \tag{29}$$

The following bound on $\Phi(\cdot,\cdot)$ is derived below, where α_{20} is a positive constant:

$$\|\Phi(t,t_0)\| \leqq \alpha_{20} \quad \text{for all } t_0,\ t \geqq t_0. \tag{30}$$

The proof of this statement follows by noting from Lemma 1 that there exists a control which is bounded* independently of t_0, taking the zero state at $(t_0 - \delta_c)$ to state $\lambda(t_0)$ at time t_0, where $\lambda(t_0)$ is an arbitrary vector of unit norm. Set the control equal to zero for $t \geqq t_0$. Then over $[t_0 - \delta_c, \infty)$ the control is bounded*, independently of t_0, while for $t \geqq t_0$,

$$x(t) = \Phi(t,t_0)\lambda(t_0).$$

The B*IBS constraint implies $x(t)$ is bounded independently of t_0 and thus yields (30).

Arguments as in [8] then establish that (29) and (30) together imply EAS; thus Theorem 1 is proved.

It is important to note that [8] shows that for a bounded matrix F, (29) alone implies EAS. Actually the boundedness of F is only used to deduce (30); hence the applicability of our proof. It is also interesting to note that, although EAS and boundedness of the matrices F and G imply BIBS (see [4]), EAS does not itself imply B*IBS, but requires the addition of the uniform complete controllability constraint, though to be sure, not all the uniform complete controllability conditions are used. Those conditions which are used amount to a generalization of the boundedness constraints on F and G and do not include the left-hand inequalities of (5) and (6).

4. Relations between Lyapunov and bounded*-input, bounded*-output stability. Hitherto, we have been concerned with relating the control and state variables; in this section, we are concerned with relating the state and output variables. Because the relation (1b) between them is nondynamic, and thus does not involve derivatives or integrals as the relation (1a) does, the results are much simpler to achieve. The key theorem is as follows.

THEOREM 2. *Suppose the system* (1) *is uniformly completely observable. Then it is B*IBS if and only if it is B*IB*O.*

Proof. We prove first that B*IBS implies B*IB*O. Observe that

$$\int_s^{s+\delta_0} y'(t)y(t)\,dt = \int_s^{s+\delta_0} x'(t)H(t)H'(t)x(t)\,dt$$

$$= \int_s^{s+\delta_0} \hat{x}'(t)\Phi'(t,s)H(t)H'(t)\Phi(t,s)\hat{x}(t)\,dt,$$

where $\hat{x}(t)$ is defined in the interval $[s, s+\delta_0]$ by $\hat{x}(t) = \Phi(s,t)x(t)$. (Note: $\hat{x}(t)$ is not a state vector.)

If $x(t)$ is bounded, $\hat{x}(t)$ is bounded as follows:

$$\|\hat{x}(t)\| \leqq \sup_{0 \leqq \rho \leqq \delta_0} \{\alpha_5(\rho)\}\|x(t)\|.$$

Denoting the bound on $\|\hat{x}(t)\|$ by α_{21}, we have that

$$\int_s^{s+\delta_0} y'(t)y(t)\,dt < \alpha_{21}^2 n\alpha_7,$$

i.e., y is bounded*. Since any bounded* input results in a bounded state by assumption, and since bounded states imply bounded* outputs by the above, we have that bounded* inputs imply bounded* outputs.

Now suppose that (1) is known to be B*IB*O. Suppose also it is not B*IBS. Then there exist an input u_1 and constant α_{14} such that

$$\int_s^{s+\delta_0} u_1'(t)u_1(t)\,dt \leqq \alpha_{14} \quad \text{for all } s,$$

such that, with y_1 the corresponding output,

$$\int_s^{s+\delta_0} y_1'(t)y_1(t)\,dt \leqq \alpha_{15}(\alpha_{14}),$$

and such that for some T,

$$\|x_1(T)\| > \sqrt{\frac{\alpha_{15}}{\alpha_6}}.$$

Now replace u_1 by u, where $u(t) = u_1(t)$ for $t \leqq T$, and $u(t) = 0$ for $t > T$. Then

$$\int_s^{s+\delta_0} u'(t)u(t)\,dt \leqq \alpha_{14} \quad \text{for all } s,$$

and thus, with y the corresponding ouput,

$$\int_s^{s+\delta_0} y'(t)y(t)\,dt \leqq \alpha_{15} \quad \text{for all } s. \tag{31}$$

Also, of course, since $x(t) = x_1(t)$ for $t \leqq T$,

$$\|x(T)\| > \sqrt{\frac{\alpha_{15}}{\alpha_6}}. \tag{32}$$

Now use the fact that $u(t)$ is zero for $t > T$ to obtain

$$\begin{aligned} \int_T^{T+\delta_0} y'(t)y(t)\,dt &= x'(T)\int_T^{T+\delta_0} \Phi'(t, T)H(t)H'(t)\Phi(t, T)\,dt\, x(T) \\ &\geqq \alpha_6 x'(T)x(T) \\ &> \alpha_{15}. \end{aligned} \tag{33}$$

The first inequality follows from the uniform complete observability assumption, the second from (32). Equation (33) now is in contradiction to (31). Hence B*IB*O must imply B*IBS. This completes the proof.

The arguments above may be used to conclude a result similar to that of Theorem 2, with bounded inputs replacing bounded* inputs. It is as follows.

COROLLARY 1. *Suppose the system* (1) *is uniformly completely observable. Then it is BIBS if and only if it is BIB*O.*

The connection between internal and external stability for the system (1) is obtained by combining Theorems 1 and 2. The proof of the following result, obtained from these theorems, is trivial.

THEOREM 3. *Consider the system* (1), *assumed uniformly completely controllable and uniformly completely observable. Then it is B*IB*O if and only if it is EAS.*

It is interesting to note that the result of Theorem 2 cannot be improved upon to the extent of deducing that B*IBS implies B*IBO, though of course B*IBO implies B*IBS. Construction of a counterexample is easy. Suppose first that F_1 and H_1 are constant matrices such that $[F_1, H_1]$ is completely observable (and thus uniformly so). Define $F(t)$ and $H(t)$ by

$$\begin{aligned} F(t) &= F_1, \\ H(t) &= H_1 \qquad (t \leqq 0, \quad n - 1 \leqq t \leqq n - 1/n^2) \\ &= nH_1 \qquad (n - 1/n^2 \leqq t \leqq n) \end{aligned}$$

for $n = 1, 2, \cdots$. Then it is readily verified that $F(t)$ and $H(t)$ are a uniformly completely observable pair, while evidently the addition of a G_1 so that

$\dot{x} = F_1x + G_1u$ is B*IBS does not imply that, with $y = H'(t)x$, the mapping from u to y is B*IBO. This mapping is of course B*IB*O.

The uniform complete observability assumption is required in going both ways in Theorem 2; this is in contrast to the result that for a bounded realization, BIBS implies BIBO [4]. The explanation is that in establishing that B*IBS implies B*IB*O, not all the uniform complete observability conditions are used, but only those reflecting a natural generalization of boundedness constraints on F and H.

5. Some additional results. In this section, we present two additional results which generalize material of [4] and [9] and which are in part based on the earlier materials. The first extends a well-known result for time-invariant systems and can be of use in establishing whether a given pair $[F, G]$ is uniformly completely controllable. Hence we include the result here.

THEOREM 4. *Uniform complete controllability in a realization* (1) *is invariant under state variable feedback of the form*

$$u(t) = K(t)x(t) + g(t), \tag{34}$$

where the entries of $K(\cdot)$ *are continuous functions,*

$$\int_{s-\delta_c}^{s} \|K(t)\|^2\, dt \leqq \alpha_{22}(\delta_c) \quad \textit{for all } s \tag{35}$$

and some constant α_{22}, *and* $g(\cdot)$ *is the input to the closed loop system.*

Proof. Let (1) be uniformly completely controllable. Then by Lemma 1 there is a $\delta_c > 0$ and a minimal energy input u_1, which transfers the zero state at time $s - \delta_c$ to the state ξ at time s, such that

$$\int_{s-\delta_c}^{s} u_1'(t)u_1(t)\, dt \leqq \alpha_{11}\|\xi\|^2 \tag{36}$$

for all s. It is readily verified that if

$$g(t) = u_1(t) - K(t)x_1(t) \tag{37}$$

is the input to the closed loop system, where x_1 is the trajectory of the open loop system due to u_1, then $z_1(s - \delta_c) = 0$ and $z_1(s) = \xi$, where z_1 is the trajectory of the closed loop system due to g (in fact, $z_1(t) = x_1(t)$ for all $t \in (s - \delta_c, s)$).

Using (15b) and (37), we have that for all $t \in (s - \delta_c, s)$

$$\begin{aligned} \|g(t)\| &\leqq \|u_1(t)\| + \left\{ \|K(t)\| \left\| \int_{s-\delta_c}^{t} \Phi(t,\tau)G(\tau)G'(\tau)\Phi'(s,\tau)\, d\tau \right\| \right. \\ &\qquad \left. \cdot \|M^{-1}(s - \delta_c, s)\|\, \|\xi\| \right\} \\ &\leqq \|u_1(t)\| + \left\{ \|K(t)\|\, \|\Phi(t,s)\|\, \|M^{-1}(s - \delta_c, s)\|\, \|\xi\| \right. \\ &\qquad \left. \cdot \left\| \int_{s-\delta_c}^{t} \Phi(s,\tau)G(\tau)G'(\tau)\Phi'(s,\tau)\, d\tau \right\| \right\} \cdot \end{aligned}$$

The integral in the above equation has an upper bound of $M(s - \delta_c, s)$, and thus using the bounds on M and M^{-1}, we have

$$\|g(t)\| \leqq \|u_1(t)\| + \frac{\alpha^2}{\alpha_1} \sup_{0 \leqq \rho \leqq \delta_c} \alpha_5(\rho)\|\xi\| \, \|K(t)\|.$$

Thus

$$(38) \quad \int_{s-\delta_c}^{s} \|g(t)\|^2 \, dt \leqq 2\left\{\int_{s-\delta_c}^{s} \|u_1(t)\|^2 \, dt + \frac{\alpha_2^2}{\alpha_1^2} \sup_{0 \leqq \rho \leqq \delta_c} \alpha_5^2(\rho)\|\xi\|^2 \int_{s-\delta_c}^{s} \|K(t)\|^2 \, dt\right\}.$$

This means that, with (35) and (36) satisfied, $\int_{s-\delta_c}^{s} \|g(t)\|^2 \, dt$ is bounded above by a term $\alpha_{23}\|\xi\|^2$. A fortiori, the energy of the minimal energy control transferring the zero state to the state ξ is bounded above by $\alpha_{23}\|\xi\|^2$.

We now show that for the closed loop system,

$$(39) \quad M_{\mathrm{CL}}(s - \delta_c, s) = \int_{s-\delta_c}^{s} \Phi_{\mathrm{CL}}(s, t)G(t)G'(t)\Phi'_{\mathrm{CL}}(s, t) \, dt$$

is bounded above, where $\Phi_{\mathrm{CL}}(\cdot, \cdot)$ is the transition matrix of the closed loop system given by

$$\frac{d}{ds}\Phi_{\mathrm{CL}}(s, t) = [F(s) - G(s)K(s)]\Phi_{\mathrm{CL}}(s, t), \qquad \Phi_{\mathrm{CL}}(t, t) = I.$$

If we define

$$(40) \quad Y(s, t) = \Phi(s, t) - \Phi_{\mathrm{CL}}(s, t) - \int_{t}^{s} \Phi_{\mathrm{CL}}(s, \tau)G(\tau)K(\tau)\Phi(\tau, t) \, d\tau,$$

the differentiation yields

$$\frac{dY(s, t)}{dt} = -Y(s, t)F(t),$$

while inspection of (40) shows that $Y(t, t) = 0$. This means that $Y(s, t) = 0$ for all t and s. Thus

$$\Phi(s, t) - \Phi_{\mathrm{CL}}(s, t) - \int_{t}^{s} \Phi_{\mathrm{CL}}(s, \tau)G(\tau)K(\tau)\Phi(\tau, t) \, d\tau = 0$$

or

$$\Phi_{\mathrm{CL}}(s, t)\Phi(t, s) = I - \int_{t}^{s} \Phi_{\mathrm{CL}}(s, \tau)\Phi(\tau, s)\Phi(s, \tau)G(\tau)K(\tau)\Phi(\tau, s) \, d\tau.$$

This means that

$$\|\Phi_{\mathrm{CL}}(s, t)\Phi(t, s)\| \leqq I + \int_{t}^{s} \|\Phi_{\mathrm{CL}}(s, \tau)\Phi(\tau, s)\| \, \|\Phi(s, \tau)G(\tau)K(\tau)\Phi(\tau, s)\| \, d\tau$$

which implies that

$$(41) \quad \|\Phi_{\mathrm{CL}}(s, t)\Phi(t, s)\| < \exp\left[\int_{t}^{s} \|\Phi(s, \tau)G(\tau)K(\tau)\Phi(\tau, s)\| \, d\tau\right]$$

from a trivial extension of a result in [6, Theorem 2, p. 134]. With t in the range $s - \delta_c \leqq t \leqq s$,

$$\int_t^s \|\Phi(s, \tau)G(\tau)K(\tau)\Phi(\tau, s)\| \, d\tau$$

$$\leqq \sup_{0 \leqq \rho \leqq \delta_c} \alpha_5(\rho) \left[\int_{s-\delta_c}^s \|\Phi(s, \tau)G(\tau)\|^2 \, d\tau\right]^{1/2} \left[\int_{s-\delta_c}^s \|K(\tau)\|^2 \, d\tau\right]^{1/2},$$

which is bounded independently of s and t by the uniform complete controllability of the open loop system and the restriction on K. Then from (41),

$$\|\Phi_{\mathrm{CL}}(s, t)\Phi(t, s)\| \leqq \alpha_{24}(\delta_c)$$

for $s - \delta_c \leqq t \leqq s$ and some positive constant α_{24} independent of s and t. Now observe that in (39), we may rewrite M_{CL} as

$$M_{\mathrm{CL}}(s - \delta_c, s) = \int_{s-\delta_c}^s \Phi_{\mathrm{CL}}(s, \tau)\Phi(\tau, s)\Phi(s, \tau)G(\tau)G'(\tau)\Phi'(s, \tau)\Phi'(\tau, s)\Phi'_{\mathrm{CL}}(s, \tau) \, d\tau$$

and thus

$$\|M_{\mathrm{CL}}(s - \delta_c, s)\| \leqq \alpha_{24}^2(\delta_c) \int_{s-\delta_c}^s \|\Phi(s, \tau)G(\tau)\|^2 \, d\tau.$$

Thus we have established that $M_{\mathrm{CL}}(s - \delta_c, s)$ is bounded above if the open loop system is uniformly completely controllable.

By using (17a), the above result implies that the minimum energy control u_3 taking the closed loop system from the zero state at time $s - \delta_c$ to the state ξ at time s satisfies the inequalities

$$0 < \alpha_{25}\|\xi\|^2 \leqq \int_{s-\delta_c}^s \|u_3(t)\|^2 \, dt,$$

where α_{25} is a positive constant independent of s.

We have shown earlier that the energy associated with the minimum energy control is also bounded above. Upper and lower bounds may also be established in a similar way for the energy of a minimal control taking the state ξ at time $s - \delta_c$ to the zero state at time s. It then follows by Lemma 1 that the closed loop system is uniformly completely controllable, and the theorem is proved.

As a comment on the application of the above theorem we note that the problem of deciding whether a prescribed pair $F(t), G(t)$ is uniformly completely controllable is often difficult; it may require calculation of the transition matrix. However, the fact that $F(t)$ may be replaced by $F(t) - G(t)K(t)$ for a large class of matrices $K(t)$ may reduce the difficulty, as occasionally $K(t)$ could be taken such that $F - GK$ was constant.

We now turn to an extension of the time-varying version of the lemma of Lyapunov as discussed in [9]. In particular, boundedness constraints are removed and appropriate modifications are made.

THEOREM 5. *Consider the system*

$$\frac{d}{dt}x(t) = F(t)x(t) \tag{42}$$

and let $L(\cdot)$ be a matrix such that $[F, L]$ is uniformly completely observable. Both $F(\cdot)$ and $L(\cdot)$ have entries which are continuous. Then

(i) *If F is exponentially asymptotically stable, there exists a matrix P defined by*

$$P(t) = \lim_{T \to \infty} \Pi(t, T), \tag{43}$$

where $\Pi(t, T)$ in turn is defined by

$$-\dot{\Pi} = \Pi F + F'\Pi + L'L, \qquad \Pi(T, T) = 0. \tag{44}$$

Moreover,

$$V(x, t) = x'(t)P(t)x(t)$$

is a Lyapunov function for (42), *and finally P is given by the formula*

$$P(t) = \lim_{T \to \infty} \int_t^T \Phi'(\lambda, t)L'(\lambda)L(\lambda)\Phi(\lambda, t)\, d\lambda. \tag{45}$$

(ii) *If there exists a symmetric matrix $P(\cdot)$ and positive constants β_1 and β_2 such that for all t*

$$0 < \beta_1 I \leqq P(t) \leqq \beta_2 I < \infty \tag{46}$$

and such that

$$-\dot{P} = PF + FP + L'L, \tag{47}$$

then $V = x'Px$ is a Lyapunov function such that for some positive β_3, δ_0 and all t,

$$\Delta V \Big|_t^{t+\delta_0} / V \leqq -\beta_3.$$

(This condition corresponds to EAS.)

Proof of (i). The solution of (44) can readily be verified to be

$$\Pi(t, T) = \int_t^T \Phi'(\lambda, t)L'(\lambda)L(\lambda)\Phi(\lambda, t)\, d\lambda.$$

Since $[F, L]$ is uniformly completely observable,

$$0 < \beta_4 I \leqq \int_t^{t+\delta_0} \Phi'(\lambda, t)L'(\lambda)L(\lambda)\Phi(\lambda, t)\, d\lambda < \beta_5 I < \infty \tag{48}$$

for all t and some positive constants β_4, β_5 and δ_0. This means that

$$\|\Pi(t, T)\| \leqq \beta_5[1 + \|\Phi(t + \delta_0, t)\|^2 + \|\Phi(t + 2\delta_0, t)\|^2 + \cdots + \|\Phi(t + k\delta_0, t)\|^2]$$

for some k using arguments similar to those in the proof of Theorem 1. Since F is EAS, positive constants α_{18} and α_{19} exist such that

$$\|\Pi(t, T)\| \leqq \beta_5\{1 + \alpha_{18}[\exp(-\alpha_{19}2\delta_c) + \exp(-\alpha_{19}4\delta_c) + \cdots + \exp(-\alpha_{19}2k\delta_c)]\}.$$

Using the nontrivial arguments as in the proof of Lemma 1, it becomes apparent that $\|\Pi(t, T)\|$ is bounded above independently of t and T. This result together with the result that $\Pi(t, T)$ monotonically increases as T increases means that the limit (43) exists and is bounded independently of t. The lower bound on $P(t)$ follows using (45), (48) and the observation that

$$\int_t^{t+\delta_0} \Phi'(\lambda, t)L'(\lambda)L(\lambda)\Phi(\lambda, t)\,d\lambda \leqq \int_t^{\infty} \Phi'(\lambda, t)L'(\lambda)L(\lambda)\Phi(\lambda, t)\,d\lambda.$$

With $V = x'Px$, by using (45), $\dot{V} = -x'L'Lx$ which is plainly nonpositive. This proves (i).

Proof of (ii). Equation (46) guarantees that V as distinct from $\dot{V}$ satisfies the necessary requirements for it to be a Lyapunov function. Since using (47) we have $\dot{V} = -x'L'Lx$, it follows that stability, as distinct from EAS, of F is established. To establish EAS we compute the change in V along a length δ_0 of trajectory. Thus

$$\begin{aligned}\Delta V\Big|_t^{t+\delta_0} &= \int_t^{t+\delta_0} \dot{V}\,dt \\ &= -x'(t)\int_t^{t+\delta_0} \Phi'(\lambda, t)L'(\lambda)L(\lambda)\Phi(\lambda, t)\,d\lambda x(t).\end{aligned}$$

Since $[F, L]$ is uniformly completely observable,

$$\Delta V\Big|_t^{t+\delta_0} \leqq -\beta_4 x'(t)x(t)$$

and

$$\Delta V\Big|_t^{t+\delta_0} / V \leqq -\beta_4/\beta_1.$$

Simple arguments may be used to show that EAS is implied, and the proof of part (ii) is completed.

6. Conclusions. This paper has shown that in developing a number of linear time-varying system stability results, the usually imposed boundedness restriction on the elements of the system matrices is not essential. Of particular interest is the result that internal and external stability are equivalent for uniformly completely controllable and observable systems, provided that in defining external stability, modification is made to the usual requirement of boundedness of inputs and outputs.

REFERENCES

[1] R. E. KALMAN, *On the stability of linear time-varying systems*, IEEE Trans. Circuit Theory, CT-9 (1962), pp. 420–422.

[2] O. PERRON, *Die Stabilitätsfrage bei Differentialgleichungen*, Math. Z., 32 (1930), pp. 703–728.

[3] B. D. O. ANDERSON, *Stability properties of linear systems in phase-variable form*, Proc. IEEE, 115 (1968), pp. 340–341.

[4] L. M. SILVERMAN AND B. D. O. ANDERSON, *Controllability, observability and stability of linear systems*, this Journal, 6 (1968), pp. 121–130.

[5] R. E. KALMAN, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana, 5 (1960), no. 2, pp. 102–119.

[6] E. F. Bechenbach and R. Bellman, *Inequalities*, Springer, New York, 1965.

[7] R. E. Kalman, Y. C. Ho and K. S. Narendra, *Controllability of linear dynamical systems*, Contributions to Differential Equations, 1 (1963), pp. 189–213.

[8] 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. 371–400.

[9] B. D. O. Anderson and J. B. Moore, *Time-varying version of the lemma of Lyapunov*, Electronics Letters, 3 (1967), pp. 293–294.

SIAM J. CONTROL
Vol. 7, No. 3, August 1969

AN EFFICIENT COMPUTATIONAL PROCEDURE FOR A GENERALIZED QUADRATIC PROGRAMMING PROBLEM*

ROBERT O. BARR†

1. Introduction to the basic problem (BP). Many problems in optimal control (see [1], [2] and [3]) have as their essence the following basic problem (BP):

Given K a compact, convex set in E^n. Find a point $z^* \in K$ such that $|z^*|^2 = \min_{z \in K}|z|^2$ (here $|\cdot|$ denotes Euclidean norm).

The constraint set K in this quadratic programming problem, in constrast to the quadratic programming problems usually treated in the literature (e.g., [4], [5], [6], [7] and [8]) need not be specified by some set of functional inequalities. It is required only that there be a known *contact function* of K, i.e., a function $s(\cdot)$ from E^n to K such that the scalar product $y \cdot s(y) = \max_{z \in K} y \cdot z$. For the convex sets K which occur in a wide variety of optimal control problems, a contact function is the only available characterization of K. The connection between BP and optimal control problems as well as techniques for evaluating a contact function in specific problems is given in [1], [2] and [3]. Some other optimal control references ([9], [10], [11], [12], [13], [14], [15] and [16]) give computing procedures which also utilize the convexity of K.

This paper presents an iterative procedure for solving BP. The procedure extends the method of E. G. Gilbert [1] so that on each iteration a quadratic minimization problem is solved on a convex polyhedron instead of on a line segment. As with Gilbert's procedure, convergence is guaranteed and computable error bounds are available. However, the results of many computations for a rather general example using both Gilbert's method and its extension indicate that the extended procedure converges much more rapidly. In fact, under certain conditions on K, the convergence is not only rapid but finite.

For these algorithms the importance of a contact function cannot be overemphasized. The availability of a contact function implies that given $\hat{z} \in K$, $\hat{z} \neq 0$, a point $s(-\hat{z})$ on the boundary of K can be determined such that an inward normal to the boundary at that point is parallel to $\hat{z}$. It is the utilization of this new point which yields a smaller value $|z|^2$.

It should be observed that if $\bar{K} \subset E^n$ is closed and convex but not necessarily bounded, then the intersection of $\bar{K}$ with a closed sphere around the origin and containing at least one point of $\bar{K}$ is a set K satisfying the requirements for BP. Any solution z^* of BP satisfies $|z^*|^2 = \min_{z \in \bar{K}}|z|^2$.

The paper is organized as follows: In § 2 the iterative procedure (IP) and convergence theorem are stated; in § 3 certain results and solution techniques are presented for the subproblem which occurs on each iteration of IP; in § 4, § 5 and § 6 the convergence theorem is proved, several variations of IP are discussed, and conditions guaranteeing finite convergence are established; and in § 7 some numerical results are exhibited.

* Received by the editors April 2, 1968, and in revised form March 5, 1969.

† Systems Science, College of Engineering, Michigan State University, East Lansing, Michigan 48823. This research was supported in part by the United States Air Force under Grants AF-AFOSR-814-65 and AF-AFOSR-814-66 and by the National Science Foundation under Grant GK-726.

Some notation useful in the sequel is presented here. For $x, y, z \in E^n$, scalar ω, set X, and compact convex set C define: $N(x;\omega) = \{z:|z - x| < \omega\}$, $\omega > 0$, the open sphere with center at x and radius ω; $\bar{N}(x;\omega) = \{z:|z - x| \leqq \omega\}$, the corresponding closed sphere; $Q(x;y) = \{z:z \cdot y = x \cdot y\}$, $y \neq 0$, the hyperplane (dimension $n - 1$) through x with normal y; $Q^o(x;y) = \{z:z \cdot y < x \cdot y\}$, $y \neq 0$, the open half-space bounded by $Q(x;y)$ with outward normal y; $Q^c(x;y) = \{z:z \cdot y \geqq x \cdot y\}$, $y \neq 0$, the closed half-space bounded by $Q(x;y)$ with inward normal y; ∂X, the boundary of X; dim X, the dimension of X; $P_C(y) = \{z:z \cdot y = \max_{x \in C} x \cdot y\}$, $y \neq 0$, the support hyperplane of C with outward normal y.

2. The iterative procedure (IP) **and convergence theorem.** It should be noted that since $|z|$ is continuous and K is compact, a solution z^* to BP exists. Furthermore, z^* is unique; $z^* = 0$ if and only if $0 \in K$; and for $0 \notin K$, $z^* \in \partial K$. The uniqueness of z^* is an obvious consequence of the strict convexity of $|z|^2$.

The geometric significance of a contact function $s(\cdot)$ of K is illustrated in Fig. 1: for $y \neq 0$, $s(y)$ lies on the intersection of K and the support hyperplane $P_K(y)$. The figure shows that for certain $y \neq 0$, $s(y)$ may not be uniquely determined. However, the iterative procedure requires only that a method be available for computing one value of $s(y)$ for each given y. Such a method is available for many optimal control problems [1], [2], [3], and no assumptions of "normality" or "unique maximum condition" are necessary.

Consider now the iterative procedure for solving BP.

ITERATIVE PROCEDURE (IP). Let $s(\cdot)$ be an arbitrary contact function of the set K. Take $z_0 \in K$ and choose a positive integer p. Then a sequence of vectors $\{z_{k+1}\}$, $k = 0, 1, 2, \cdots$, is generated as follows:

Step 1. Select any p vectors $y_1(k), y_2(k), \cdots, y_p(k)$ in K and let

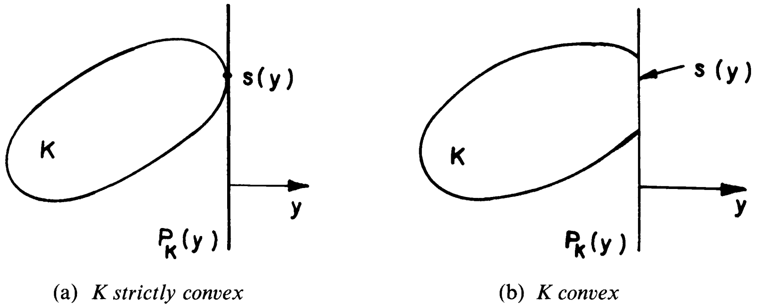
$$H_k = \Delta\{y_1(k), y_2(k), \cdots, y_p(k), s(-z_k), z_k\}$$

(Δ denotes "convex hull of").

Step 2. Find $z_{k+1} \in H_k$ such that

$$|z_{k+1}|^2 = \min_{z \in H_k} |z|^2.$$

Steps 1 and 2 constitute one iteration, called iteration k, of IP.



(a) *K strictly convex* (b) *K convex*

FIG. 1. *The contact function* $s(y)$

Gilbert's procedure [1] (with parameter $\delta = 1$) differs from IP in that z_{k+1} is obtained by minimizing over the line segment $\Delta\{s(-z_k), z_k\}$ instead of over the convex polyhedron H_k. Extensive computational results, some of which are shown in § 7, indicate that for many sets K the rate of convergence for Gilbert's procedure is slow. Propitious choice of the points $y_1(k), y_2(k), \cdots, y_p(k)$ in Step 1 of IP can notably accelerate convergence. Several selection rules for choosing these points are discussed in § 5 and the improved convergence properties of the resulting versions of IP are treated in § 6 and § 7.

It is important to note the distinction between BP and the subproblem in Step 2 of IP, both of which are quadratic programming problems on a compact, convex constraint set. The set K in BP is described only by a contact function $s(\cdot)$ of K, whereas the set H_k in Step 2 is the convex hull of $p + 2$ known points. Thus the subproblem is simpler than BP. It is shown in § 3 that standard quadratic programming techniques can be used to solve the subproblem.

Before stating the convergence theorem for IP, it is convenient to introduce:

$$\text{(1)} \quad \begin{aligned} \gamma(z) &= |z|^{-2}z \cdot s(-z), && |z| > 0, \quad z \cdot s(-z) > 0, \\ &= 0, && z = 0 \quad \text{or} \quad |z| > 0, \quad z \cdot s(-z) \leqq 0, \end{aligned}$$

$$\text{(2)} \quad \begin{aligned} \mu(z) &= |z|^{-1}z \cdot s(-z), && z \neq 0, \\ &= 0, && z = 0. \end{aligned}$$

Thus $\gamma(\cdot)$ and $\mu(\cdot)$ are scalar functions which are defined on K. Figure 2 indicates their geometric significance for the case $|z| > 0$, $z \cdot s(-z) > 0$: $\gamma(z)z$ is the intersection of the line through 0 and z with the support hyperplane $P_K(-z)$ of K having outward normal $-z$; $|\mu(z)|$ is the Euclidean distance from the origin to $P_K(-z)$. The function $\gamma(\cdot)$ is useful in computing error bounds for IP and $\mu(\cdot)$ forms the basis of selection for several versions of IP which exhibit good convergence (see § 5). Gilbert [1] has proved the following properties of $\gamma(z)$, $z \in K$: $0 \leqq \gamma(z) \leqq 1$; if $0 \in K, \gamma(z) \equiv 0$; if $0 \notin K, \gamma(z) = 1$ if and only if $z = z^*$; $\gamma(z)$ is continuous. Now the principal convergence theorem is stated.

THEOREM 1 (Convergence theorem for IP). *Consider the sequence* $\{z_k\}$ *generated by* IP. *For* $k \geqq 0$ *and* $k \to \infty$:

(i) $z_k \in K$;

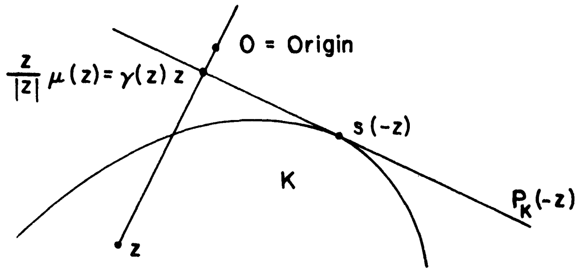


FIG. 2. *Geometric significance of* $\gamma(\cdot)$ *and* $\mu(\cdot)$

(ii) *the sequence* $\{|z_k|\}$ *is decreasing* $(|z_k| \geqq |z_{k+1}|)$, $|z_k| \to |z^*|$, *and* $|z_k| = |z_{k+1}|$ *implies* $z_k = z^*$;

(iii) $z_k \to z^*$;

(iv) $|z_k|\gamma(z_k) \leqq |z^*|$ *and* $|z_k|\gamma|(z_k) \to |z^*|$;

(v) $|z_k - z^*| \leqq \sqrt{1 - \gamma(z_k)|z_k|}$ *and* $\sqrt{1 - \gamma(z_k)|z_k|} \to 0$;

(vi) $|s(-z_k) - z^*| \leqq |s(-z_k) - \gamma(z_k)z_k|$.

The results of the theorem are identical to those stated by Gilbert [1] for his iterative procedure. Since the bounds given in parts (iv), (v) and (vi) are computable as the iterative process proceeds, they may be used to generate stopping criteria for the termination of the iterative process. Example problems show $\{|z_k|\gamma(z_k)\}$ is not necessarily increasing. Thus $|z_k| - \max_{i \leqq k}|z_i|\gamma(z_i)$ is more satisfactory as an upper bound for $|z_k| - |z^*|$ than $|z_k| - |z_k|\gamma(z_k)$.

3. Subproblems 1 and 2. The proof of Theorem 1 is deferred until § 4 so that certain preliminary results for the subproblem in Step 2 of IP can be established. This subproblem will be considered in the following form.

SUBPROBLEM 1. *Given* H, *the convex hull of* m *points* $y_1, y_2, \cdots, y_m$ *from* E^n. *Find a point* $y^* \in H$ *such that* $|y^*|^2 = \min_{y \in H}|y|^2$.

Since H is compact and $|y|$ is a continuous function of y, a solution y^* exists.

THEOREM 2 (Solution properties for Subproblem 1).

(i) y^* *is unique*;

(ii) $|y^*| = 0$ *if and only if* $0 \in H$;

(iii) *for* $|y^*| > 0, y^* \in \partial H$;

(iv) *for* $|y^*| > 0, y = y^*$ *if and only if* $y \in P_H(-y) \cap H$;

(v) *for* $|y^*| > 0$, y^* *has a representation* $y^* = \sum_{i=1}^m \sigma_i y_i$, *where* $\sum_{i=1}^m \sigma_i = 1$, $\sigma_i \geqq 0$ $(i = 1, 2, \cdots, m)$, *and no more than* n *of the* σ_i *are positive.*

Proof. Gilbert [1] has shown that z^*, the solution to BP, has properties similar to (i) through (iv) of Theorem 2. For example, when $|z^*| > 0$, $z = z^*$ if and only if $z \in P_K(-z) \cap K$. The proof of parts (i) through (iv) is omitted since it requires only slight modification of the corresponding proof by Gilbert [1].

Consider part (v). Since $H \subset E^n$, ∂H is the union of a finite number of convex polyhedra each having dimension $\leqq n - 1$. Moreover, the set of extreme points of each of these convex polyhedra is a subset of $\{y_1, y_2, \cdots, y_m\}$. From $|y^*| > 0$, $y^* \in \partial H$. Thus the Caratheodory theorem [17], [18] implies that there exists a subset X of $\{y_1, y_2, \cdots, y_m\}$ containing q points such that $y^* \in \Delta X$ and $1 \leqq q \leqq n$. It follows that y^* has the required representation.

Since the iteration of IP requires the solution of Subproblem 1, it is important that methods exist for readily computing its solution. The standard quadratic programming techniques described in the literature (e.g., [4], [5], [6], [7] and [8]) cannot be directly applied since they begin by assuming the constraint set is described by a set of linear equations and/or inequalities rather than by the points whose convex hull is the constraint polyhedron. Determining from $y_1, y_2, \cdots, y_m$ a description for H in terms of linear equations and/or inequalities presents serious computational difficulties.

There is an alternative method of attacking Subproblem 1 which makes possible the use of the standard algorithms. It will be shown now that the solution to Subproblem 1 is given by the solution to another quadratic programming problem, Subproblem 2, which has a constraint set described by linear equations and inequalities. Subproblem 2 is solvable by any of the well-known quadratic programming techniques such as that due to Frank and Wolfe [4], which was used for the computations in § 7.

Each $y \in H$ has the representation $y = \sum_{i=1}^{m} x^i y_i$, where $\sum_{i=1}^{m} x^i = 1, x^i \geqq 0$, $i = 1, 2, \cdots, m$. Thus

$$(3) \qquad |y|^2 = \left| \sum_{i=1}^{m} x^i y_i \right|^2 = \sum_{i=1}^{m} \sum_{j=1}^{m} x^i x^j y_i \cdot y_j .$$

If x is the m-vector $(x^1, x^2, \cdots, x^m)$ and D is the $m \times m$ symmetric matrix with elements $d_{ij} = y_i \cdot y_j$, then

$$(4) \qquad |y|^2 = x \cdot Dx.$$

Since $|y|^2 \geqq 0$, the quadratic form $x \cdot Dx$ is nonnegative definite, a fact which implies it is a convex function of x on E^m. Consider now Subproblem 2.

SUBPROBLEM 2. *Given D, an $m \times m$ symmetric nonnegative definite matrix, and the constraint set*

$$X = \left\{ x \in E^m : \sum_{i=1}^{m} x^i = 1, x^i \geqq 0, i = 1, 2, \cdots, m \right\}.$$

Find a point $x^* \in X$ such that $x^* \cdot Dx^* = \min_{x \in X} x \cdot Dx$.

If $D = [d_{ij}] = [y_i \cdot y_j]$, Subproblem 2 is said to be associated with Subproblem 1. For this case, (4) implies that minimization of $x \cdot Dx$ on X is equivalent to minimization of $|y|^2$ on H. Thus, if x^* solves Subproblem 2, the solution y^* to Subproblem 1 is given by

$$(5) \qquad y^* = \sum_{i=1}^{m} x^{*i} y_i .$$

4. Proof of Theorem 1. Since $y_1(k), y_2(k), \cdots, y_p(k)$, and $s(-z_k)$ are points in K, it follows from the convexity of K and the definition of convex hull that $z_k \in K$ implies $H_k \subset K$ and $z_{k+1} \in K$. Thus by induction $z_0 \in K$ proves part (i).

The inequalities in (iv), (v) and (vi) follow by the same arguments used in the proof given by Gilbert [1] for the corresponding parts of his convergence theorem. Gilbert's proof also yields the inequality

$$(6) \qquad 0 \leqq |z - z^*|^2 \leqq \Gamma(z), \qquad z \in K,$$

where the function $\Gamma(z)$ is defined by

$$(7) \qquad \Gamma(z) = |z|^2 - |z^*|^2, \qquad z \in K.$$

Now consider part (ii). Given a $z_k \in K$, let $\bar{z}_{k+1} \in \Delta\{s(-z_k), z_k\}$ be such that

$$(8) \qquad |\bar{z}_{k+1}|^2 = \min_{z \in \Delta\{s(-z_k), z_k\}} |z|^2 .$$

Since z_{k+1} is obtained by minimizing over H_k and $\Delta\{s(-z_k), z_k\} \subset H_k$, it follows that

$$(9) \qquad |z_{k+1}|^2 \leqq |\bar{z}_{k+1}|^2.$$

Then (7) and (9) imply

$$(10) \qquad \Gamma(z_k) - \Gamma(z_{k+1}) \geqq \Gamma(z_k) - \Gamma(\bar{z}_{k+1}).$$

But from (4.20) of Gilbert's paper [1]

$$(11) \qquad \Gamma(z_k) - \Gamma(\bar{z}_{k+1}) \geqq \min\{\tfrac{1}{4}\rho^{-2}\Gamma^2(z_k), \tfrac{1}{2}\Gamma(z_k)\} \geqq 0,$$

where $\rho = \max_{z_1, z_2 \in K}|z_1 - z_2|$. Thus (10) and (11) yield

$$(12) \qquad \Gamma(z_k) - \Gamma(z_{k+1}) \geqq \min\{\tfrac{1}{4}\rho^{-2}\Gamma^2(z_k), \tfrac{1}{2}\Gamma(z_k)\} \geqq 0.$$

Therefore the sequence $\{\Gamma(z_k)\}$ is decreasing and, since it is bounded below by zero, has a limit point. Thus passing to the limit on the left side of (12) gives zero and from the right side $\Gamma(z_k) \to 0$. By (6), (7) and (12) this proves (ii) and (iii).

The second result in (iv) and the second result in (v) follow from $\gamma(z^*) = 1$, the continuity of $\gamma(\cdot)$, and (iii).

5. Selection rules for IP. In this section several rules for selecting the vectors $y_1(k), y_2(k), \cdots, y_p(k) \in K$ in Step 1 of iteration k of IP are presented. The symbols Y_k and S_k are used to denote the sets:

$$(13) \qquad Y_k = \{y_1(k), y_2(k), \cdots, y_p(k)\},$$

$$(14) \qquad S_k = \{y_1(k), y_2(k), \cdots, y_p(k), s(-z_k)\}.$$

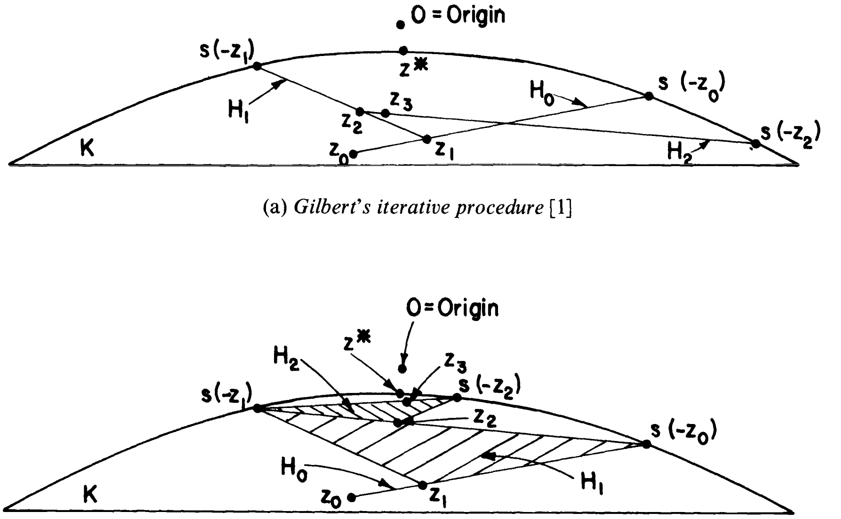
Since the goal is to choose Y_k so that IP will converge more rapidly than Gilbert's procedure [1], some preliminary comments are appropriate. Attention is focused on the case $z^* \in \partial K$ which is most important in applications.

Numerical results are exhibited in § 7 for the set

$$(15) \qquad K = \left\{ z : z^1 \geqq v + \frac{1}{2}\sum_{i=2}^{n} (z^i)^2 \lambda_i^{-1}, z^1 \leqq 10 \right\},$$

where $v, \lambda_2, \lambda_3, \cdots, \lambda_n > 0$ and $z = (z^1, z^2, \cdots, z^n)$. The optimum is $z^* = (v, 0, \cdots, 0)$ and the λ_i are the principal radii of curvature of ∂K at z^*. Since many other convex sets K have a boundary surface which is closely approximated by a similar representation in the neighborhood of z^*, this example is of general interest.

The numerical results indicate that slow convergence is obtained with Gilbert's procedure when the surface ∂K at z^* has at least one principal radius of curvature large compared with $|z^*|$. For problems in which the set $P_K(-z^*) \cap K$, $z^* \neq 0$, contains more than one point (this may occur when K is not strictly convex), convergence is especially poor. If Y_k is chosen so that after a few iterations of IP the surface ∂H_k in the vicinity of z_{k+1} closely approximates ∂K in the vicinity of z^*, then it is likely that IP will exhibit improved convergence. For ∂H_k to approximate ∂K, the dimension of H_k must be sufficiently large, namely n, and Y_k must include boundary points of K. To illustrate these remarks, consider Fig. 3 which is K of (15) with $n = 2$, $v = 1$, $\lambda_2 = 1$, $z^* = (1, 0)$. In Fig. 3a Gilbert's procedure is



(a) *Gilbert's iterative procedure* [1]

(b) IP, *Selection Rule A*, $p = 2$

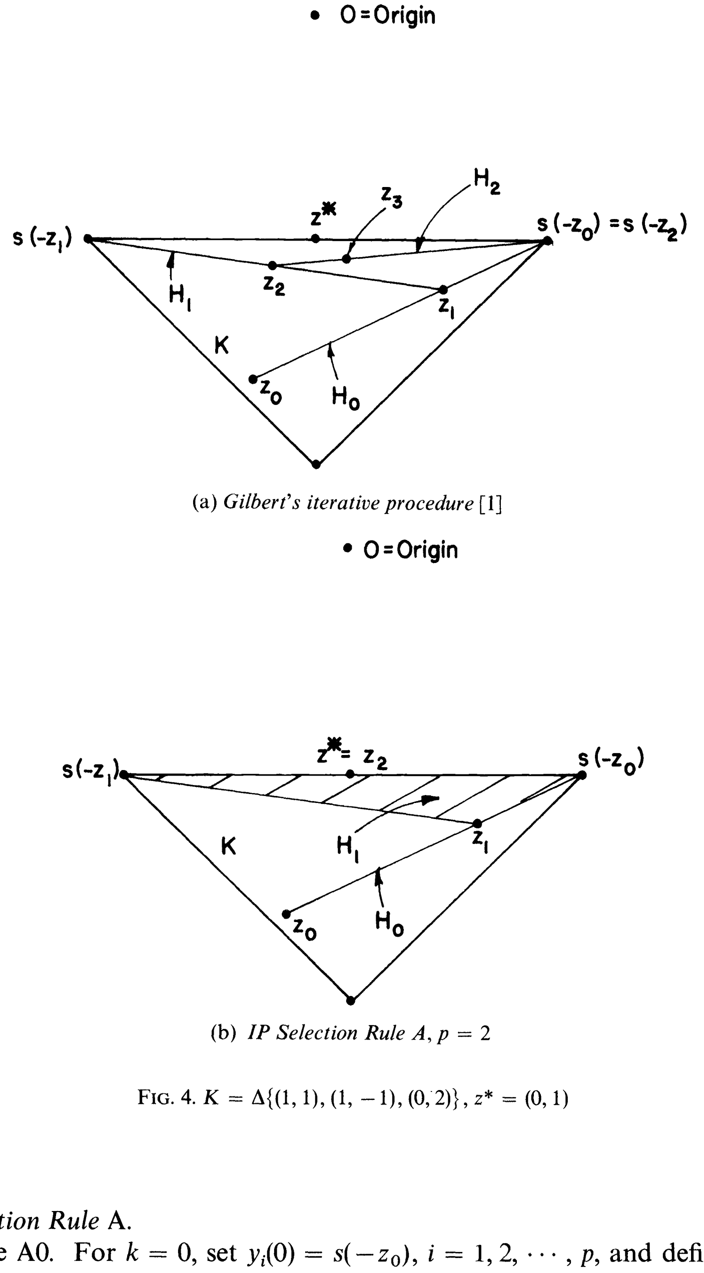
FIG. 3. K *of* (15) *with* $n = 2, v = 1, \lambda_2 = 1, z^* = (1, 0)$

shown, where dim $H_k = 1$ and convergence is slow. In Fig. 3b IP with Selection Rule A (to be described subsequently) is shown, where $\dim H_k \leqq 2$ and convergence is notably improved. An even more startling improvement is exhibited in Fig. 4 for which $K = \Delta\{(1, 1), (-1, 1), (0, 2)\}$, $z^* = (0, 1)$ and $P_K(-z^*) \cap K$ is the line segment $\Delta\{(1, 1), (-1, 1)\}$. Theorem 5 shows that when K is a convex polyhedron, IP (with a suitable selection rule and contact function) converges in a finite number of iterations. Furthermore, the extensive numerical results of § 7 provide strong evidence that IP is far superior to Gilbert's procedure.

Let the p points in Y_k be contact points of K. Observe that ∂H_k is a better local approximation to ∂K for larger values of p. However, the larger p is, the more difficult it is to solve the Subproblem 1 in Step 2 of IP. The computational results of § 7 indicate that convergence is good for $p = n$ and little improvement is obtained for $p > n$. The desirability of choosing $p = n$ is also evident from the finite convergence material in § 6.

In optimal control applications [1], [2], [3] it is advantageous to limit the number of times the contact function is evaluated. Thus in the selction rules which follow Y_{k+1}, $k \geqq 0$, contains every point in Y_k except perhaps one. There remains the question of how to reject one contact point in favor of another. The approach in the selection rules given here is to use $\mu(z)$ as an indication of the quality of the contact point $s(-z)$, $z \in K$. Roughly speaking, contact points corresponding to larger values of $\mu(\cdot)$ are preferred. Other quantities, e.g., $|s(-z)|$, $|z|$, $\gamma(z)$, may be suggested for judging the merit of $s(z)$. However, careful examination of example problems such as K of (15) with $n = 3$, $\lambda_2 \gg \lambda_3$ shows that these quantities are less desirable than $\mu(z)$.

It is convenient to view each selection rule as several phases, which are to be used sequentially.



(a) *Gilbert's iterative procedure* [1]

(b) *IP Selection Rule A*, $p = 2$

FIG. 4. $K = \Delta\{(1,1),(1,-1),(0,2)\}, z^* = (0,1)$

Selection Rule A.

Phase A0. For $k = 0$, set $y_i(0) = s(-z_0)$, $i = 1, 2, \cdots, p$, and define scalars $\mu_1, \mu_2, \cdots, \mu_p$ equal to $\mu(z_0)$.

Phase A1. For $1 \leqq k \leqq p$, set $y_i(k) = y_i(k-1)$, $i = 1, 2, \cdots, p$. Then set $y_k(k) = s(-z_{k-1})$ and $\mu_k = \mu(z_{k-1})$.

Phase A2. For $p + 1 \leqq k$, set $y_i(k) = y_i(k-1)$, $i = 1, 2, \cdots, p$. Then let $\underline{\mu} = \min\{\mu_1, \mu_2, \cdots, \mu_p\}$ and let j be the smallest integer in $[1, p]$ for which $\mu_j = \underline{\mu}$. Whenever $\mu_j \leqq \mu(z_{k-1})$, replace $y_j(k)$ by $s(-z_{k-1})$ and μ_j by $\mu(z_{k-1})$.

In Phase A2 there is nothing crucial about the way of handling the possibility of two or more μ_i, $1 \leqq i \leqq p$, being equal to $\underline{\mu}$. Moreover, $\mu_j < \mu(z_{k-1})$ may be used as the condition for replacement instead of $\mu_j \leqq \mu(z_{k-1})$.

For selection rules such as A in which Y_{k+1} is a subset of S_k for all $k \geqq 0$, it is possible to state additional results. Consider the determination of z_{k+1} in Step 2 of iteration k of IP. Since $z_{k+1} \in H_k$, it has a representation: $z_{k+1} = \sum_{i=1}^{p+2} x^i y_i$, where $y_i = y_i(k)$, $i = 1, 2, \cdots, p$, $y_{p+1} = s(-z_k)$, $y_{p+2} = z_k$, $\sum_{i=1}^{p+2} x^i = 1$, $x^i \geqq 0$, $i = 1, 2, \cdots, p+2$. Let I_k be the set of superscripts i, $1 \leqq i \leqq p+2$, for which $x^i = 0$. Observe that I_k may be empty. The following results hold.

THEOREM 3. *Consider the sequence* $\{z_k\}$ *generated by* IP.

(i) *Suppose the selection rule is such that* $Y_{k+1} \subset S_k$ *for all* $k \geqq 0$. *If in Step* 2 *of iteration* k, $k > 0$, *the integer* $p+1 \in I_k$, *then* $z_k = z_{k+1} = z^*$.

(ii) *Suppose* p *is chosen so that* $q = p + 2 - n \geqq 0$. *Then in Step* 2 *of iteration* k, $k \geqq 0$, *either* $z_{k+1} = 0$ (*which implies* $z_{k+1} = z^* = 0$) *or there exists a vector* $x = (x^1, x^2, \cdots, x^{p+2})$ *such that the corresponding* I_k *contains at least* q *elements.*

Now assume $z_{\bar{k}+1} \neq z^*$ *for some* $\bar{k} \geqq 0$. *Then for all* k, $0 \leqq k \leqq \bar{k}$:

(iii) $Y_k \subset S_k \subset H_k \subset Q^c(z_{k+1}; z_{k+1})$;

(iv) *if* $Y_{k+1} \subset S_k$, $s(-z_{k+1}) \in Q^0(z_{k+1}; z_{k+1})$;

(v) *if* $Y_1 \subset S_0$, S_1 *contains two distinct points*;

(vi) *if* $Y_{k+1} \subset S_k$, Y_{k+1} *contains every point in* S_k *except one, and* S_k *contains* $\bar{p}$ *distinct points* $(1 \leqq \bar{p} \leqq p+1)$, *it follows that there are at least* $\bar{p}$ *distinct points in* S_{k+1};

(vii) *if Phases* A0 *and* A1 *are used and* $z_p \neq z^*$, S_p *contains* $p+1$ *distinct points.*

Proof. Consider (i). Note that $k > 0$ and let $\bar{H} = \Delta\{y_1(k-1), y_2(k-1), \cdots, y_p(k-1), s(-z_{k-1}), z_k\}$. Since $\bar{H} \subset H_{k-1}$ and $z_k \in \bar{H}$, $|z_k| = \min_{z \in \bar{H}} |z|$. But $p+1 \in I_k$ and $Y_k \subset S_{k-1}$ imply $z_{k+1} \in \bar{H}$ and $|z_{k+1}| \geqq \min_{z \in \bar{H}} |z|$. Thus by part (ii) of Theorem 1 $|z_k| = |z_{k+1}|$ and $z_k = z_{k+1} = z^*$. In (ii) suppose $z_{k+1} \neq 0$. Part (iii) of Theorem 2 yields $z_{k+1} \in \partial H_k$. Then part (v) of the same theorem and $p + 2 - n \geqq 0$ imply that a vector x with the desired property exists. Consider (iii). Since $z_{\bar{k}+1} \neq z^*$, part (ii) of Theorem 1 implies $|z_k| > |z^*| \geqq 0$ for all k, $0 \leqq k \leqq \bar{k} + 1$. Hence, for $0 \leqq k \leqq \bar{k}$, the hyperplane $Q(z_{k+1}; z_{k+1})$ is defined. By part (iv) of Theorem 2 it follows that $z_{k+1} \in P_{H_k}(-z_{k+1}) = Q(z_{k+1}; z_{k+1})$. Thus H_k is contained in the closed half-space $Q^c(z_{k+1}; z_{k+1})$ and (iii) is true. In (iv) suppose for some k, $0 \leqq k \leqq \bar{k}$, $s(-z_{k+1}) \notin Q^o(z_{k+1}; z_{k+1})$. Then $s(-z_{k+1}) \in Q^c(z_{k+1}; z_{k+1})$, and by (iii), $Q^c(z_{k+1}; z_{k+1})$ also contains S_k and z_{k+1}. Since $Y_{k+1} \subset S_k$, $H_{k+1} \subset Q^c(z_{k+1}; z_{k+1})$ and $|z_{k+1}| = |z_{k+2}|$. By part (ii) of Theorem 1 this implies $z_{k+1} = z^*$ for some k, $0 \leqq k \leqq \bar{k}$, which contradicts $|z_j| > |z^*|$, $0 \leqq j \leqq \bar{k} + 1$, and thus establishes (iv). From (iii), (iv) and $Y_1 \subset S_0$ it follows that $Y_1 \subset Q^c(z_1; z_1)$ and $s(-z_1) \in Q^o(z_1; z_1)$. Hence S_1 must contain two distinct points. Similarly (iii), (iv) and $Y_{k+1} \subset S_k$ imply $Y_{k+1} \subset Q^c(z_{k+1}; z_{k+1})$ and $s(-z_{k+1}) \in Q^o(z_{k+1}; z_{k+1})$, which means $s(-z_{k+1})$ is distinct from the points in Y_{k+1}. If Y_{k+1} contains every point in S_k except one and S_k contains $\bar{p}$ distinct points, then there are at least $\bar{p} - 1$ distinct points in Y_{k+1}. Since S_{k+1} is the union of Y_{k+1} and $s(-z_{k+1})$, the conclusion in (vi) is true. Consider (vii). From $z_p \neq z^*$, it follows that (iii) and (iv) hold for $0 \leqq k \leqq p - 1$. Thus $Y_{k+1} \subset S_k$ and $0 \leqq k \leqq p - 1$ imply $s(-z_{k+1})$ is distinct from the points in Y_{k+1}. For Phases A0 and A1, Y_1 and S_0 contain only $s(-z_0)$; Y_2 and S_1 contain only $s(-z_0)$ and $s(-z_1)$; $\cdots$; Y_p and S_{p-1} contain only $s(-z_0)$,

$s(-z_1), \cdots, s(-z_{p-1})$. Consequently the fact that $s(-z_{k+1}) \notin Y_{k+1}$ applied successively for $k = 0, 1, \cdots, p-1$ yields the result: $s(-z_0), s(-z_1), \cdots, s(-z_p)$ are distinct. Since S_p is the union of these $p+1$ points, (vii) holds and the proof is complete.

Now consider two additional selection rules.

Selection Rule B. (Assume that $p \geqq n$ and that in Step 2 of every iteration k of IP a vector x is determined such that the corresponding I_k contains at least $p+2-n$ integers. By part (ii) of Theorem 3 such an x must exist or $z_{k+1} = 0$. Terminate IP in Step 2 of iteration k if $z_{k+1} = 0$ (which implies $z_{k+1} = z^* = 0$) or if $p+1 \in I_k$ (which implies $z_k = z_{k+1} = z^*$).)

Phase B0. Use Phase A0.

Phases B1 and B2. Proceed as in Phases A1 and A2 except that if $p+2 \in I_{\hat{k}}$ for some $\hat{k}$, define $k' = \max\{\hat{k}+1, p+1\}$ and for $k \geqq k'$ use Phase B2'.

Phase B2'. Set $y_i(k) = y_i(k-1), i = 1, 2, \cdots, p$. Let $\underline{\mu}' = \min_{i \in I_{k-1}, i \neq p+1, p+2} \mu_i$ and let j be the smallest integer in I_{k-1} for which $\mu_j = \underline{\mu}'$. Replace $y_j(k)$ by $s(-z_{k-1})$ and μ_j by $\mu(z_{k-1})$.

For a particular problem it is possible that the condition $p+2 \in I_{\hat{k}}$ for entering Phase B2' may never be satisfied. In that case Selection Rules A and B are identical. The computational experience with IP indicates, however, that this condition is satisfied after only a few iterations for a broad class of problems.

Selection Rule C. (Assume that $p \geqq n$, that $z_0 = s(-z_{-1})$ for some $z_{-1} \in K$ and that in Step 2 of every iteration k of IP a vector x is determined such that the corresponding I_k contains at least $p+2-n$ integers. By part (ii) of Theorem 3 such an x must exist or $z_{k+1} = 0$. Terminate IP in Step 2 of iteration k if $z_{k+1} = 0$ (which implies $z_{k+1} = z^* = 0$) or if $p+1 \in I_k$ (which implies $z_k = z_{k+1} = z^*$).)

Phase C0. For $k = 0$, set $y_1(0) = z_0 = s(-z_{-1})$, set $y_i(0) = s(-z_0)$, $i = 2, 3, \cdots, p$, and define $\mu_1 = \mu(z_{-1})$, $\mu_i = \mu(z_0)$, $i = 2, 3, \cdots, p$.

Phase C1. For $1 \leqq k \leqq p-1$, set $y_i(k) = y_i(k-1)$, $i = 1, 2, \cdots, p$. Then set $y_{k+1}(k) = s(-z_{k-1})$ and $\mu_{k+1} = \mu(z_{k-1})$.

Phase C2. For $k \geqq p$, use Phase B2'.

The assumption $p \geqq n$ is required for Selection Rule B and Selection Rule C so that the set $\{i : i \in I_{k-1}, i \neq p+1, p+2\}$ which occurs in Phases B2' and C2 is not empty. Since $p+2-n \geqq 2$, I_{k-1} contains at least 2 integers in $[1, p+2]$. Moreover, $p+1 \notin I_{k-1}$ or IP would have terminated in Step 2 of iteration $k-1$.

THEOREM 4. *Consider* IP *and assume that Selection Rule* B *or Selection Rule* C *is used.*

(i) *If* $z_p \neq z^*$, S_p (*with Rule* B) *and* S_{p-1} (*with Rule* C) *contain* $p+1$ *distinct points*;

(ii) *Let* $\hat{k}$ *be the first* $k \geqq 0$ *for which* $p+2 \in I_k$ *if Selection Rule B is used and let* $\hat{k} = 0$ *if Selection Rule C is used. Then in Step* 2 *of iteration* k, *all* $k \geqq \hat{k}$, *Subproblem* 1 *can be solved on* ΔS_k *instead of* H_k. *That is, let* z_{k+1} *satisfy* $z_{k+1} \in \Delta S_k, |z_{k+1}| = \min_{z \in \Delta S_k} |z|$ *and find a vector* $x = (x^1, x^2, \cdots, x^{p+2})$ *such that* $z_{k+1} = \sum_{i=1}^{p+2} x^i y_i$, *where* $y_i = y_i(k)$, $i = 1, 2, \cdots, p$, $y_{p+1} = s(-z_k)$, $y_{p+2} = z_k$, $\sum_{i=1}^{p+2} x^i = 1, x^i \geqq 0, i = 1, 2, \cdots, p+1, x^{p+2} = 0$.

Proof. For (i) note that with Selection Rule B or C, $Y_{k+1} \subset S_k$ and Y_{k+1} contains every point in S_k except one for all $k \geqq 0$. Thus, the argument for part (vii) of Theorem 3 can be essentially repeated to prove (i). Consider (ii). Since

$p + 2 \in I_{\hat{k}}$ (with Rule B) and $y_1(\hat{k}) = z_{\hat{k}}$ (with Rule C), on iteration $\hat{k}$ Subproblem 1 can certainly be solved on $\Delta S_{\hat{k}}$ instead of $H_{\hat{k}}$. Thus $z_{\hat{k}+1} \in \Delta S_{\hat{k}}$. Now $Y_{\hat{k}+1} \subset S_{\hat{k}}$. Furthermore, $Y_{\hat{k}+1}$ contains every point in $S_{\hat{k}}$ except one which has a coefficient of 0 in the convex combination expression for $z_{\hat{k}+1}$. Hence $z_{\hat{k}+1} \in \Delta Y_{\hat{k}+1}$ and $\Delta S_{\hat{k}+1} = H_{\hat{k}+1}$, so that on iteration $\hat{k} + 1$ Subproblem 1 can be solved on $\Delta S_{\hat{k}+1}$ to yield $z_{\hat{k}+2} \in \Delta S_{\hat{k}+1}$. By induction Subproblem 1 can be solved on ΔS_k instead of H_k for all iterations k, $k \geqq \hat{k}$. This completes the proof.

Note that it is simpler to solve Subproblem 1 on ΔS_k rather than on H_k: the constraint set for the quadratic programming problem is the convex hull of only $p + 1$ points instead of $p + 2$. It will henceforth be assumed that whenever Selection Rule B or Selection Rule C is used in IP, Subproblem 1 is solved on ΔS_k for all $k \geqq \hat{k}$.

Section 7 contains computational results for IP with Selection Rules A and B. These results indicate that it is good to choose $p = n$ and that the rate of convergence of IP, which is about the same using Rule A or Rule B, is significantly faster than Gilbert's procedure [1] for a broad class of problems. Selection Rule C is identical with Selection Rule B for $k \geqq \max\{\hat{k} + 1, p + 1\}$, where $\hat{k}$ is the first $k \geqq 0$ for which $p + 2 \in I_k$ when Rule B is used. For all the computations in which Selection Rule B was used (e.g., see Table 5), $\hat{k}$ was observed to be very small. Thus it can be stated that IP with Selection Rule C also converges much more rapidly than Gilbert's procedure.

From part (ii) of Theorem 4 Selection Rules B and C have an advantage over Rule A in that for iterations k, $k \geqq \hat{k}$, Subproblem 1 can be solved on the convex hull of $p + 1$ points instead of $p + 2$. However, the requirement with Rules B and C that I_k contain at least $p + 2 - n$ integers adds complexity to the solution of Subproblem 1 in Step 2 of every iteration k, $k \geqq 0$. Selection Rule C is most desirable for guaranteeing finite convergence in certain problems (see § 6) and Rules B and C are advantageous for certain optimal control applications [1], [2], [3].

6. A finite convergence theorem. The following theorem gives a sufficient condition for IP to exhibit finite convergence.

THEOREM 5. *Let $s(\cdot)$ be an arbitrary contact function of the set K specified in* BP, *choose $p \geqq n$, and consider* IP *with Selection Rule C. Assume that the range of $s(y)$ for $y \in E^n$ is a finite set of points. Then the sequence $\{z_k\}$ generated by* IP *converges in a finite number of iterations.*

Proof. Let $\bar{s}_1, \bar{s}_2, \cdots, \bar{s}_l$ denote the points in the range of $s(y), y \in E^n$ and define $\bar{S} = \{\bar{s}_1, \bar{s}_2, \cdots, \bar{s}_l\}$. Suppose that convergence of IP is not obtained in a finite number of iterations. By part (ii) of Theorem 1, $|z_{k+1}| < |z_k|$ for $k = 0, 1, 2, \cdots$. However, by part (ii) of Theorem 4, $|z_k|^2$ is the value obtained by minimizing $|z|^2$ over the convex hull of $p + 1$ points contained in $\bar{S}$. Since there are only a finite number of ways of choosing $p + 1$ points from $\bar{S}$, there are only a finite number of possible values of $|z_k|^2$ as k runs through the nonnegative integers, contradicting $|z_{k+1}| < |z_k|$ for $k = 0, 1, 2, \cdots$. This completes the proof.

If the range of $s(y)$ for $y \in E^n$ is a finite set of points, then K is a convex polyhedron. However, for a convex polyhedron K in E^n, $n \geqq 2$, it does not necessarily follow that the range of $s(y)$, $y \in E^n$, is finite. For example, in E^3 an entire edge of K could lie in the range of $s(y)$, $y \in E^3$. Nevertheless, if the set of extreme points of a convex polyhedron K is a known set, say $\hat{S}$, then it is possible to choose a contact function $s(y)$ of K whose range for $y \in E^n$ is a finite set of points. This follows from the fact that a contact function of $\hat{S}$ is also a contact function of K. Thus, for many convex polyhedra K such as those which arise in optimization problems for linear sampled-data systems, Theorem 5 can be used to guarantee finite convergence.

Note that Theorem 5 holds for IP with Selection Rule B if there exists $k \geqq 0$ such that the condition $p + 2 \in I_k$ is satisfied in Step 2 of iteration k.

The assumption in Theorem 5 may be replaced by the following less restrictive assumption: there exists $\varepsilon > 0$ such that for $z \in K$ and $|z| - |z^*| < \varepsilon$, the range of $s(-z)$ is a finite set of points. Then the sequence $\{z_k\}$ generated by IP with Selection Rule C converges in a finite number of iterations. This result follows from part (ii) of Theorem 1 and arguments similar to the proof of Theorem 5.

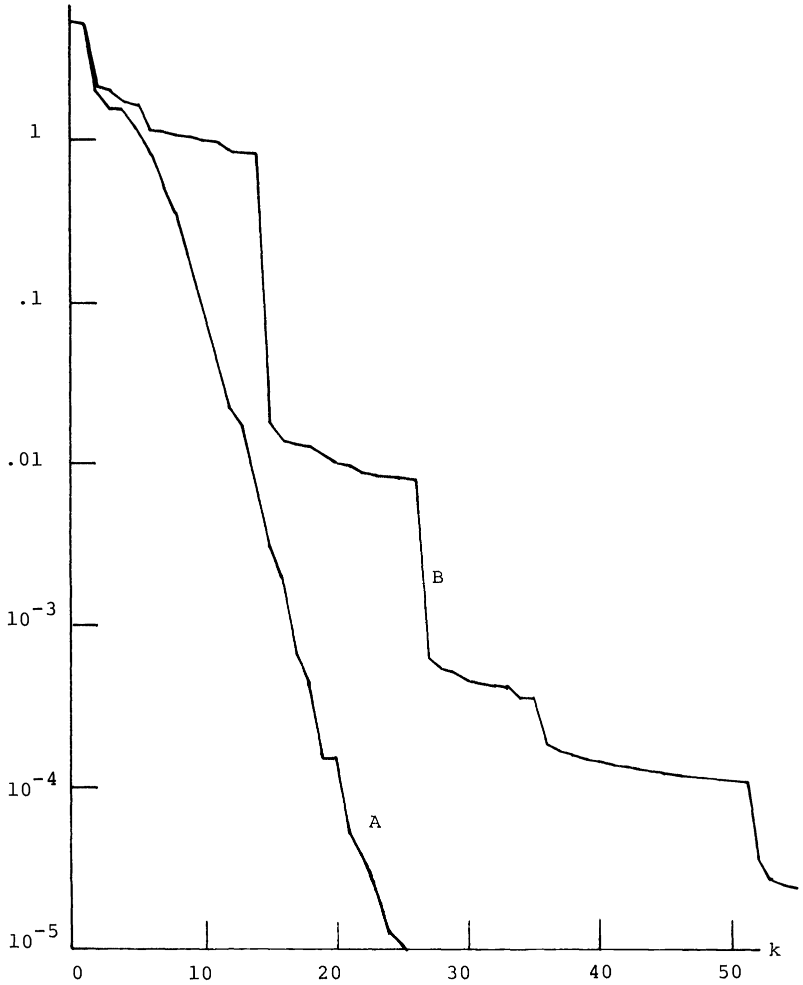


FIG. 5. $|z_k| - |z^*|$ *for* $n = 3$, $z_0 = (6, 2, 2)$, $v = 1$, $\lambda_2 = 100$, $\lambda_3 = 10$: (A) *IP Selection Rule* A, $p = 3$; (B) *Gilbert's procedure* [1]

7. Some numerical results. This section contains some results of numerical computations for the general set K defined by (15). Data for Gilbert's procedure [1] and for IP with Selection Rules A and B are presented.

In optimal control applications the evaluation of a contact function is the most time-consuming part of the iterative procedure. Since such an evaluation is required on each iteration, the number of iterations to satisfy certain error criteria is used as a measure of the speed of convergence.

Figure 5 and Tables 1 and 2 present results for $n = 3$ using Gilbert's procedure and IP with Selection Rule A ($p = 3$). The data indicate that the speed of convergence for Gilbert's procedure depends strongly on the parameter $\bar{\lambda}v^{-1}$, where $\bar{\lambda} = \max\{\lambda_i\}$, convergence being slow when $\bar{\lambda}v^{-1}$ is large. For IP the convergence is much more rapid and shows very little dependence on $\bar{\lambda}v^{-1}$. The behavior of $|z_k| - |z^*|$, $|z_k - z^*|$ and $|z^*| - |z_k|\gamma(z_k)$ is typical: $|z^*| - |z_k|\gamma(z_k)$ decreases most rapidly, followed in order by $|z_k| - |z^*|$ and $|z_k - z^*|$.

TABLE 1
Number of iterations to satisfy $|z_k| - |z^*| \leqq \varepsilon$; $n = 3$, $z_0 = (6, 2, 2)$, $v = 1$

		Gilbert's Procedure			IP Selection Rule A, $p = 3$		
λ_2	ε λ_3	1	10^{-3}	10^{-6}	1	10^{-3}	10^{-6}
10	10	4	28	41	3	7	12
100	100	10	59	111	4	9	13
1,000	1,000	82	216	340	4	10	12
100	10	11	27	81	6	17	32
1,000	1	14	218	321	6	11	17
1,000	10	83	229	359	7	20	29
1,000	100	81	197	358	8	23	32

TABLE 2
Number of iterations to satisfy $|z^*| - |z_k|\gamma(z_k) \leqq \varepsilon$; $n = 3$, $z_0 = (6, 2, 2)$, $v = 1$

		Gilbert's Procedure			IP Selection Rule A, $p = 3$		
λ_2	ε λ_3	1	10^{-3}	10^{-6}	1	10^{-3}	10^{-6}
10	10	0	20	37	0	6	11
100	100	0	18	58	0	8	12
1,000	1,000	0	81	165	0	8	11
100	10	0	14	51	0	16	31
1,000	1	0	146	235	0	10	16
1,000	10	0	88	264	0	14	28
1,000	100	0	80	196	0	16	30

TABLE 3

Number of iterations to satisfy $|z_k| - |z^*| \leqq \varepsilon$; $n = 3$, $z_0 = (5, 4, 2)$, $\lambda_2 = 1{,}000$, $\lambda_3 = 100$, IP *Selection Rule* A *and Gilbert's Procedure*

p \ ε	1	0.1	0.01	10^{-3}	10^{-4}	10^{-5}	10^{-6}	t
0	53	77	138	202	241	298	366	125.3
1	30	52	97	134	171	249	293	319.1
2	9	15	27	36	46	52	58	75.9
3	9	16	23	26	29	33	36	50.1
4	8	14	17	24	27	32	37	77.8
5	8	13	16	23	26	31	35	90.3

Notes. (i) $p = 0$ corresponds to Gilbert's Procedure.
(ii) t = actual computing time (seconds) for IBM 7090.

TABLE 4

Number of iterations to satisfy $|z_k| - |z^*| \leqq \varepsilon$; $n = 4$, $z_0 = (6, 1, 2, 2)$, $\lambda_2 = 1{,}000$, $\lambda_3 = 500$, $\lambda_4 = 100$, IP *Selection Rule* A

p \ ε	1	0.1	0.01	10^{-3}	10^{-4}	10^{-5}	10^{-6}	t
1	23	67	107	152	185	226	264	331.1
2	19	38	79	101	127	141	158	211.4
3	15	27	36	46	62	74	78	123.1
4	12	18	25	29	37	48	53	86.6
5	11	17	22	30	39	49	57	113.3

Note. t = actual computing time (seconds) for IBM 7090.

TABLE 5

Number of iterations to satisfy $|z_k| - |z^*| \leqq \varepsilon$

						IP Selection Rule A			IP Selection Rule B			
$n = p$	λ_2	λ_3	λ_4	λ_5	λ_6 \ ε	1	10^{-3}	10^{-6}	1	10^{-3}	10^{-6}	$\hat{k}$
3	100	50				7	18	30	7	21	31	2
3	100	90				6	17	31	6	19	29	2
3	1,000	100				8	23	32	9	23	32	6
4	100	50	10			7	29	49	8	30	51	2
4	100	90	10			7	28	48	7	28	47	2
4	1,000	500	100			10	31	48	10	35	50	5
5	100	70	50	10		9	39	66	8	36	63	4
5	100	90	80	70		10	38	70	9	37	69	3
5	1,000	900	500	100		17	43	79	15	42	74	8
6	100	90	70	50	10	12	45	91	11	49	92	4
6	1,000	90	70	50	10	14	59	94	15	60	94	6

Notes. (i) $\hat{k}$ = the first k for which $p + 2 \in I_k$ with Rule B.
(ii) For $a = 3$, $z_0 = (6, 2, 2)$; for $n = 4$, $z_0 = (6, 2, 2, 1)$; for $n = 5$, $z_0 = (5, 3, 1, 1.8, 2.6)$; for $n = 6$, $z_0 = (4, 3, 2.6, 2.6, 1.8, 1.8)$.

Tables 3 and 4 show the effects of changing p. Note that convergence is good for $p = n$ and little improvement is obtained for $p > n$. The desirability of choosing $p = n$ is also indicated by the actual computing time required to satisfy $|z_k| - |z^*| \leqq 10^{-6}$.

Table 5 gives results for $n = p = 4, 5, 6$ with IP Selection Rules A and B. Observe that the rate of convergence is about the same for both selection rules. Furthermore, the rate of decrease of $|z_k| - |z^*|$ for IP is, roughly speaking, dependent on n alone. The number of iterations per decade after a few initial iterations is approximately 2 for $n = 2$, 4 for $n = 3$, 6 for $n = 4$, 9 for $n = 5$ and 13 for $n = 6$.

8. Acknowledgment. The author wishes to thank E. G. Gilbert for his many helpful comments and enthusiastic support.

REFERENCES

[1] E. G. GILBERT, *An iterative procedure for computing the minimum of a quadratic form on a convex set*, this Journal, 4 (1966), pp. 61–80.

[2] R. O. BARR, *Computation of optimal controls by quadratic programming on convex reachable sets*, Doctoral thesis, University of Michigan, 1966.

[3] R. O. BARR AND E. G. GILBERT, *Some iterative procedures for computing optimal controls*, Proc. Third Congress of the International Federation of Automatic Control (IFAC), Butterworth, London, 1966.

[4] M. FRANK AND P. WOLFE, *An algorithm for quadratic programming*, Naval Res. Logist. Quart., 3 (1956), pp. 95–110.

[5] G. HADLEY, *Nonlinear and Dynamic Programming*, Addison-Wesley, Reading, Massachusetts, 1964.

[6] H. S. HOUTHAKKER, *The capacity method of quadratic programming*, Econometrica, 28 (1960), pp. 62–87.

[7] S. VAJDA, *Mathematical Programming*, Addison-Wesley, Reading, Massachusetts, 1961.

[8] P. WOLFE, *The simplex methodfor quadratic programming*, Econometrica, 27 (1959), pp. 382–398.

[9] T. FUJISAWA AND Y. YASUDA, *An iterative procedure for solving the time-optimal regulator problem*, this Journal, 5 (1967), pp. 501–512.

[10] L. W. NEUSTADT, *Synthesizing time-optimal control systems*, J. Math. Anal. Appl., 1 (1960), pp. 484–492.

[11] ———, *Minimum effort control systems*, this Journal, 1 (1962), pp. 16–31.

[12] ———, *On synthesizing optimal controls*, Proc. Second Congress of the International Federation of Automatic Control (IFAC), Butterworth, London, 1964.

[13] J. H. EATON, *An iterative solution to time-optimal control*, J. Math. Anal. Appl., 5 (1962), pp. 329–344; 9 (1964), pp. 147–152.

[14] E. J. FADDEN AND E. G. GILBERT, *Computational aspects of the time-optimal control problem*, Computational Methods in Optimization Problems, Academic Press, New York, 1964, pp. 167–192.

[15] T. G. BABUNASHVILI, *The synthesis of linear optimal systems*, this Journal, 2 (1964), pp. 261–265.

[16] V. F. DEM'YANOV, *Determination of the optimum program in a linear system*, Avtomat. i Telemekh., 25 (1964), pp. 3–11.

[17] C. CARATHÉODORY, *Über den variabilitätsbereich der Fourierschen konstanten von positiven harmonischen funktionen*, Rend. Circ. Mat. Palermo, 32 (1911), pp. 193–217.

[18] H. G. EGGLESTON, *Convexity*, Cambridge Tracts in Math. and Math. Physics, no. 47, Cambridge University Press, 1958.

SIAM J. Control
Vol. 7, No. 3, August 1969

SOME PRACTICAL REGULARITY CONDITIONS FOR NONLINEAR PROGRAMS*

DAVID W. WALKUP AND ROGER J.-B. WETS†

1. Introduction. Theorem 1 below gives sufficient conditions for a program with linear constraints and a convex objective with range in the extended real numbers to have well-behaved duality properties. The principal ingredient in the proof of Theorem 1 is Theorem 2, which considers the properties of the optimum value of a (not necessarily convex) program under perturbation of the linear constraints. The terminology used in Theorem 1 is adapted from [1], [2], [3], where the close relationship between properties of the variational function and the duality properties of a program have been discussed at some length. Briefly a program is *solvable* if the value of the infimum is finite and achieved for some value of the variables, it is *dualizable* if there is no duality gap, and it is *stable* if there exist (optimal) nontrivial Lagrange multipliers. The definitions of convexity and lower semicontinuity for functions into the extended real numbers will be reviewed in the next section.

At the end of § 2 an argument will be given suggesting that Theorem 1 is the best possible from the viewpoint of practical applications. Finally, Theorem 1 will be applied in § 4 to show that a broad class of stochastic programs with recourse have desirable duality properties.

THEOREM 1. *Consider the nonlinear program*

$$\inf f(x), \quad Ax = b, \quad x \geqq 0, \tag{1}$$

where the objective f is convex in the sense of functions into the extended real numbers.

(i) *If f is lower semicontinuous in the sense of functions into the extended real numbers and the constraint set $K = \{x|Ax = b, x \geqq 0\}$ is bounded and contains at least one point where f is finite, then* (1) *is solvable and dualizable.*

(ii) *If f is $+\infty$ except on a closed convex polyhedron where it is finite and Lipschitz, and if* (1) *has a finite value, whether achieved or not, then* (1) *is stable.*

THEOREM 2. *Consider the function $\phi(u) = \inf\{f(x)|x \in \kappa(u)\}$, where $\kappa(u) = \{x|Ax = b - u, x \geqq 0\}$ and f is a function with range in the extended real numbers.*

(i) *If f is convex, so is ϕ.*

(ii) *If f is lower semicontinuous and $\kappa(u)$ is compact and nonempty for some u, then ϕ is lower semicontinuous.*

(iii) *If f is $+\infty$ except on some closed convex polyhedron where it is either finite and Lipschitz or identically $-\infty$, then the same holds for ϕ.*

2. Proof of Theorem 2.

DEFINITIONS. Let f be a function with R^n for its domain and the extended real

* Received by the editors December 20, 1968, and in revised form March 10, 1969.

† Boeing Scientific Research Laboratories, Seattle, Washington 98124.

numbers $\bar{R} = R \cup \{+\infty\} \cup \{-\infty\}$ for its range. The set

$$\operatorname{epi} f = \{(z, x) | z \in R,\ x \in R^n,\ z \geqq f(x)\}$$

is called the *epigraph* of f. The function f is said to be *convex* if its epigraph is a convex subset of R^{n+1} or equivalently if

$$f(x_\lambda) = f[(1 - \lambda)x_0 + \lambda x_1] \leqq (1 - \lambda) f(x_0) + \lambda f(x_1)$$

for all $\lambda \in [0, 1]$ and $x_0, x_1 \in R^n$, where the conventions $0 \cdot \infty = 0$ and $(+\infty) + (-\infty) = +\infty$ apply. The function f is said to be *lower semicontinuous* if

$$\liminf f(x_i) \geqq f(\lim x_i)$$

for every convergent sequence (x_i) in R^n, or, equivalently, if epi f is a closed subset of R^{n+1}. If the function f is finite on some subset S of R^n and B is a real number such that

$$|f(x) - f(x')| \leqq B\|x - x'\|$$

for all x, x' in S, where $\|\cdot\|$ is the Euclidean norm in R^n, then f is said to be Lipschitz on S with constant B.

The following two lemmas represent an application of results in [6] to the special situation of Theorem 2. (An inconsequential difference is that in [6], $\kappa(u)$ is a section of P rather than its projection into R^n.)

LEMMA 1. *Let* $P = \{(x, u) | Ax + Du = b, x \geqq 0\}$, $\kappa(u) = \{x | Ax = b - Du, x \geqq 0\}$, $Q = \{u | \kappa(u)$ *is nonempty*$\}$, *and let* $P = P_0 + C$ *be the representation of the polyhedron* P *as the vector sum of a bounded polyhedron* P_0 *and a* (*unique*) *polyhedral cone* C *with apex at the origin. Then* Q *is a polyhedron, and for each* $u \in Q$, $\kappa(u) = \kappa_0(u) + C'$, *where* $\kappa_0(u)$ *is a bounded polyhedron depending on* u *and* C' *is the polyhedral cone* $\{x | (x, 0) \in C\}$. *In particular,* $\kappa(u)$ *is bounded for all* u *in* Q *if it is bounded for some* u *in* Q, *for in this case* $C' = \{0\}$.

LEMMA 2. *Let* κ *and* Q *be defined as in Lemma* 1. *Then there exists a constant* $\bar{B}$ *such that for any two points* u_1 *and* u_2 *in* Q

$$d[\kappa(u_1), \kappa(u_2)] \leqq \bar{B}\|u_1 - u_2\|,$$

where $d[\cdot, \cdot]$ *denotes the Hausdorff distance between sets in* R^n.

Observe that Lemmas 1 and 2 apply directly to Theorem 2 if u is taken to be a vector in R^m and the matrix D is the identity. Accordingly, let P and Q be defined in this manner, and consider part (i) of Theorem 2. Now P is convex, and by assumption epi f is a convex subset of $R \times R^n$. Hence (epi $f \times R^m) \cap R \times P$ is a convex subset of $R \times R^n \times R^m$ and its projection onto the space $R \times R^m$ is a convex set $\mathscr{C}$. But epi ϕ is just the *vertical closure* of $\mathscr{C}$, i.e., the union of $\mathscr{C}$ and any missing endpoints of "vertical" line segments in $\mathscr{C}$. This proves part (i) of Theorem 2. Also, Q is closed and $\phi(u) = +\infty$ if u is not in Q. Thus, in order to prove part (ii) of Theorem 2 it suffices to show

$$\liminf \phi(u_i) \geqq \phi(u_0),$$

where $u_i \in Q$ and $\lim u_i = u_0 \in Q$. By Lemma 1 each $\kappa(u)$ is compact, and hence the lower semicontinuous function f attains a minimum on $\kappa(u)$ at some point, say $x(u)$. Also, it follows from Lemma 2 that $\kappa(u)$ is uniformly bounded for u in some compact neighborhood N of u_0, i.e., $P \cap [R^n \times N]$ is compact. Hence the sequence of points $x(u_i)$ has at least one limit point $x_0 \in \kappa(u_0)$. Then

$$\liminf \phi(u_i) = \liminf f(x(u_i)) \geqq f(x_0) \geqq f(x(u_0)) = \phi(u_0),$$

where the first inequality follows from the lower semicontinuity of f and the second follows from the optimality of $x(u_0)$ on $\kappa(u_0)$. This proves part (ii) of Theorem 2.

We shall begin the proof of part (iii) of Theorem 2 by establishing the following special case.

LEMMA 3. *Under the conditions of Theorem 2, if f is finite and Lipschitz throughout R^m, then either ϕ is identically $-\infty$ on Q or ϕ is finite and Lipschitz on Q.*

Proof. Let u and u' be any two points of Q, let (x_i) be a sequence of points in $\kappa(u)$ such that $\lim f(x_i) = \phi(u)$, and let x_i' be the point of $\kappa(u')$ closest to x_i. Then

$$\begin{aligned} \phi(u') - f(x_i) &\leqq f(x_i') - f(x_i) \\ &\leqq B\|x_i' - x_i\| \\ &\leqq B\bar{B}\|u' - u\|, \end{aligned} \tag{2}$$

where B is the Lipschitz constant for f and $\bar{B}$ is the constant of Lemma 2. Now $\phi(u')$ may be $-\infty$ or finite, and $\lim f(x_i) = \phi(u)$ may be $-\infty$ or finite. But (2) shows that $\phi(u) = -\infty$ implies $\phi(u') = -\infty$; hence ϕ is identically $-\infty$ on Q or ϕ is finite on Q. And if ϕ is finite on Q, then (2) implies $\phi(u') - \phi(u) \leqq B\bar{B}\|u' - u\|$. By the symmetry in u and u' it follows that ϕ is Lipschitz with constant $B\bar{B}$.

Now suppose, as assumed in Theorem 2, that the range of f is contained in the extended real numbers, and the set $K = \{x|f(x) < +\infty\}$ is a closed convex polyhedron. Then ϕ may be defined by the program

$$\begin{gathered} \phi(u) = \inf f(x), \\ Ax = b - u, \quad x \geqq 0, \quad x \in K. \end{gathered} \tag{3}$$

Since K is a polyhedron, (3) may be rewritten

$$\begin{gathered} \phi(u) = \inf f(x) + 0x', \\ Ax = b - u, \quad A'x + x' = b', \quad x \geqq 0, \quad x' \geqq 0. \end{gathered} \tag{4}$$

The set $K_0 = \{u|\phi(u) < +\infty\}$ is exactly the set of u for which the constraints of (3) or (4) are feasible; hence this set is a polyhedron. Moreover, if f is $-\infty$ on K, then ϕ is $-\infty$ on K_0. This establishes a portion of part (iii) of Theorem 2. The remaining possibility is that f is finite and Lipschitz on K. Now f may not be finite and Lipschitz where (4) is infeasible, and u does not perturb all constraints of (4), but clearly the proof of Lemma 3 will apply to (4), with a different choice of D in Lemma 1, and yield the remainder of Theorem 2.

Strictly speaking the requirement that $\{x|f(x) < +\infty\}$ be a polyhedron is not essential to Theorem 2. The proof of Lemma 3 clearly establishes the following more general result.

LEMMA 4. *Suppose f is finite and Lipschitz everywhere, K is a closed convex subset of $R^n \times R^m$, $\phi(u) = \inf\{f(x)|x \in \kappa(u)\}$, $\kappa(u) = \{x|(x, u) \in K\}$, and $Q = \{u|\kappa(u)$*

is nonempty$\}$. Then ϕ is either $-\infty$ on Q or finite and Lipschitz on Q provided there exists a constant $\bar{B}$ such that

$$d[\kappa(u), \kappa(u')] \leqq \bar{B}\|u - u'\| \tag{5}$$

for all u, u' in Q.

However, it is difficult to see how this more general result can be applied to practical problems. For, consider a program of the form

$$\phi(u) = \inf f(x), \quad Ax = b - u, \quad x \in K,$$

where K is a convex set. Surely any practical condition to be imposed on K ought not to depend on the particular form of A. But it is shown in [6] that if K is not a polyhedron there even exist $a_1, \cdots, a_n$ such that

$$\kappa(u_1) = \{x | a_1 x_1 + \cdots + a_n x_n = - u_1, x \in K\}$$

fails to satisfy (5) for any $\bar{B}$.

3. Derivation of Theorem 1. For easy reference, we repeat the convex program of Theorem 1:

$$\inf f(x), \quad Ax = b, \quad x \geqq 0. \tag{1}$$

The basic duality properties of this program are conveniently represented in an equivalent infimum problem

$$\inf \eta, \quad (\eta, u) \in \mathscr{L} \cap \mathscr{C}, \tag{6}$$

where $\mathscr{L}$ is the vertical line $\{(\eta, u) | \eta \in R, u = 0\}$ and $\mathscr{C}$ is the convex set whose vertical closure is the epigraph of ϕ, as described in the proof of part (i) of Theorem 1. The program (1) will be *solvable*, i.e., have a finite optimum value $\phi(0) = f(x^0)$ achieved by some feasible x^0, if and only if $\mathscr{L} \cap \mathscr{C}$ is a closed half-line.

A natural dual to the infimum problem (6) is the supremum problem:

$$\sup_{(\mu,u^*)} \mu, \quad \mu \leqq \eta + u^* u \quad \text{for all } (\eta, u) \in \mathscr{C}.$$

The infimum problem asks for the infimal height η of points on $\mathscr{L}$ inside $\mathscr{C}$. The supremum problem asks for the supremal height μ of points on $\mathscr{L}$ through which can be passed nonvertical hyperplanes (i.e., those not containing any line parallel to $\mathscr{L}$) bounding $\mathscr{C}$. The supremum problem can be recast in the form

$$\sup_{(\mu,u^*)} \mu, \quad \mu \leqq f(x) + u^*(b - Ax) \quad \text{for all } x \geqq 0,$$

where the role of u^* as a row m-vector of Lagrange multipliers is apparent.

It is natural to say that the infimum and supremum problems are *dual* or that (1) is *dualizable* if $\sup \mu = \inf \eta$, i.e., if there is no duality gap. We shall say that the program (1) is *stable* if the supremum problem is solvable, i.e., has a finite optimum which is achieved for some u^*. Thus (1) is stable if there is a nonvertical hyperplane supporting $\mathscr{C}$ at the point $(\phi(0), 0)$ on $\mathscr{L}$. Equivalently, (1) is stable if there exist Lagrange multipliers which convert it into an equivalent unconstrained problem.

The above discussion of duality properties of (1) follows the approach to mathematical programming in abstract spaces given in [3] which is closely related to the more extensive development given by Rockafellar in [1, 2]. Theorem 1 now follows easily from the above discussion, well-known properties of convex sets and their supports, the simple relationship between $\mathscr{C}$ and the epigraph of ϕ, and from Theorem 2.

Remark. It is clear that if $\phi(u)$ is locally Lipschitz at $u = 0$, then the convex program (1) is stable. However, Example 1 in the Appendix shows that part (ii) of Theorem 1 can fail if "Lipschitz" is replaced by "locally Lipschitz."

4. Applications to stochastic programs with recourse. In this section we shall use Theorem 1 to show that fairly broad classes of stochastic programs with recourse have well-behaved duals. The necessary results on the objective functions of such programs have already been proved in earlier papers [4], [5]. We shall make liberal use of the notation and terminology introduced in [4], [5], [8].

THEOREM 3. *If the value of a stochastic program with recourse is finite and the first stage feasibility set $K_1 = \{x | Ax = b,\ x \geqq 0\}$ is bounded, then the equivalent convex program is solvable. Moreover, either the equivalent convex program is dualizable or the objective takes the value $-\infty$ at points belonging to K_1 for arbitrarily small perturbations of b, i.e., there is an infinite duality gap.*

Proof. It is shown in [5] that the objective $z(x)$ of the equivalent convex program is either lower semicontinuous as a function into the extended real numbers or takes the value $-\infty$ at some point. The second part of the theorem follows immediately from this and Theorem 1. Now let K_2^s be the set on which $z(x)$ is less than $+\infty$ and let M be the affine hull of the intersection of K_1 and K_2^s. Since z is convex, so is K_2^s, and hence $K_1 \cap K_2^s$ has an interior with respect to M. Since the restriction z_M of z to M is convex, and since it is finite on $K_1 \cap K_2^s$, it follows that it is nowhere $-\infty$. The results of [5] apply equally well to z_M and show that it is lower semicontinuous. A straightforward application of Theorem 1 completes the proof.

THEOREM 4. *If the value of a stochastic program with recourse is finite, the recourse matrix W is fixed, the second-stage feasibility set K_2 is a polyhedron, and the random variables ξ have finite variance, then the equivalent convex program is stable.*

Proof. Theorem (4.5) of [4] shows that under the second and fourth hypotheses either the objective of the equivalent convex program is $-\infty$ throughout K_2 or it is finite and Lipschitz on K_2. The rest follows from Theorem 1.

It is worth mentioning that Proposition 3.16 of [4] gives practical conditions which insure that K_2 is a polyhedron. In addition, Corollary 4.7 of [4] gives some alternate conditions on the distribution of ξ under which the conclusions of Theorem 4.5, and hence the above theorem, remain valid.

Examples 2 through 5 in the Appendix demonstrate that various qualifying statements in Theorems 3 and 4 cannot be omitted.

Appendix.

Example 1. Let $f(x, y)$ be the function whose epigraph is the closed convex hull of the union of the following two sets:

$$\{(z, x, y) | x = 0, y \geqq 0, z \geqq (y + 2)/(y + 1)\},$$

$$\{(z, x, y) | z = 0, y \geqq 0, x \geqq 2/(y + 1)\}.$$

Then $f(x, y)$ is Lipschitz on any bounded subset of $K = \{(x, y)|x \geqq 0, y \geqq 0\}$ but

$$\phi(u) = \inf_y f(x, y), \quad x + 0y = b - u, \quad x \geqq 0, \quad y \geqq 0,$$

has a finite discontinuity at $u = b$.

Example 2. Consider the special class of stochastic programs considered in detail in [7], for which $W = [I, -I]$ and only the right-hand sides are random. It is easy to construct examples in this class such that the random variables have a distribution possessing all moments, and the value of the infimum if finite, but the program is not solvable. To do so, however, it is necessary to make K_1 and the support of the random variables unbounded. Thus boundedness of K_1 cannot be omitted from the first part of Theorem 3. Note also that in such an example the objective will be finite for every choice of x, and hence, by Theorem 4, such an example will be stable, i.e., dual solvable, even though it is not solvable.

Example 3. Consider the following variational form of a stochastic program with recourse:

$$\begin{array}{lllllll} \phi(u) = \inf\limits_x - x_1 & & + E_\xi\{\min(y_1)\}, & & & & \\ \quad x_1 & -x_2 & & & & & = 0 - u_1, \\ \quad \xi_1 x_1 & & -y_1 & -y_2 & & & = 0, \\ & \xi_2 x_2 & & -y_2 & -y_3 & & = 0, \\ & & & x_1, x_2, y_1, y_2, y_3 \geqq 0, & & & \end{array}$$

where ξ_1 is distributed continuously on $[1, \infty)$ with density ξ_1^{-2} and $\xi_2 = \xi_1 - 1$. It can be verified that K_2 is a polyhedron, namely, $\{x|x_1 \geqq 0, x_2 \geqq 0\}$, and the bracketed term, i.e., $Q(x, \xi)$, is given by

$$Q(x, \xi) = \max\{0, \xi_1 x_1 - \xi_2 x_2\}$$

provided $\xi \in \Xi$ and $x \in K_2$. Making use of the equations $\xi_2 = \xi_1 - 1$ and $x_1 - x_2 = -u_1$, we may write $\phi(u) = \inf_{x_1 \geqq M} f(x_1, u)$, where

$$f(x_1, u) = -x_1 + \int_1^L (-\xi_1 u_1 + x_1 + u_1)\xi_1^{-2}\, d\xi_1,$$

$M = \max\{0, -u_1\}$, and L is $+\infty$ or $(u_1 + x_1)/u_1$ according as $u_1 \leqq 0$ or $u_1 > 0$. Then

$$f(x_1, u) = \begin{cases} +\infty & \text{if } u_1 < 0, x_1 \geqq -u_1, \\ 0 & \text{if } u_1 = 0, x_1 \geqq 0, \\ u_1 \ln[u_1/(u_1 + x_1)] & \text{if } u_1 > 0, x \geqq 0, \end{cases} \tag{7}$$

and

$$\phi(u) = \inf_{x_1 \geqq M} f(x_1, u) = \begin{cases} +\infty & \text{if } u_1 < 0, \\ 0 & \text{if } u_1 = 0, \\ -\infty & \text{if } u_1 > 0. \end{cases}$$

This example shows that the second half of Theorem 3 need not hold if K_1 is unbounded.

Example 4. Consider the problem in Example 3 above with the additional first-stage constraint $x_1 = 1 - u_2$. For this new problem $\phi(u)$ can be derived easily by determining the values of u which yield a feasible stochastic program and substituting $x_1 = 1 - u_2$ into (7). Thus

$$\phi(u) = \begin{cases} u_1 \ln[u_1/(1 + u_1 - u_2)] & \text{if } u_1 > 0, u_2 \leqq 1, \\ 0 & \text{if } u_1 = 0, u_2 \leqq 1, \\ +\infty & \text{otherwise.} \end{cases}$$

It can be verified that the gradient of $\phi(u)$ is unbounded as u_1 decreases to 0 along the line $u_2 = 0$. This example shows that even if K_1 is bounded and $\phi(u) > -\infty$, the assumption that ξ has finite variance cannot be omitted from the statement of Theorem 4.

Example 5. The stochastic program

$$\begin{aligned} \phi(u) = \inf_x \quad & E_\xi\{-y\}, \\ x \quad & = 0 - u, \\ \xi x \quad & -y = 0, \\ x \geqq 0, \quad & y \geqq 0, \end{aligned}$$

where ξ is distributed on $[1, \infty)$ with density ξ^{-2}, shows that the alternative at the end of Theorem 3 cannot be omitted.

Remark. Example 3 shows that a stochastic program with recourse can exhibit an infinite duality gap, even if the objective of the equivalent convex program is nowhere $-\infty$. In addition, Example 1 shows that the more general class of problems of type (1) can easily exhibit finite duality gaps. However, we do not know whether a stochastic program can exhibit a finite duality gap.

REFERENCES

[1] R. T. Rockafellar, *Duality and stability in extremum problems involving convex functions*, Pacific J. Math., 21 (1967), pp. 167–187.

[2] ———, *Duality in nonlinear programming*, Mathematics of the Decision Sciences, vol. 11, Lectures Appl. Math., American Mathematical Society, Providence, Rhode Island, 1968, pp. 401–422.

[3] R. M. Van Slyke and R. J.-B. Wets, *A duality for abstract mathematical programs with applications to optimal control theory*, J. Math. Anal. Appl., 22 (1968), pp. 679–706.

[4] D. W. Walkup and R. J.-B. Wets, *Stochastic programs with recourse*, SIAM J. Appl. Math., 15 (1967), pp. 1299–1314.

[5] ———, *Stochastic programs with recourse: On the continuity of the objective*, Ibid., 17 (1969), pp. 98–103.

[6] ———, *A Lipschitzian characterization of convex polyhedra*, Proc. Amer. Math. Soc. (1969), to appear.

[7] R. J.-B. Wets, *Programming under uncertainty: The complete problem*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 4 (1966), pp. 316–339.

[8] ———, *Programming under uncertainty: The equivalent convex program*, SIAM J. Appl. Math., 14 (1966), pp. 89–105.

SIAM J. Control
Vol. 7, No. 3, August 1969

ON THE STRUCTURE OF MULTIVARIABLE SYSTEMS*

W. A. WOLOVICH† and P. L. FALB‡

1. Introduction. The primary purpose of this paper is to state and prove a structure theorem for time-invariant multivariable linear systems. The theorem can be used for controller design and synthesis and is applied here to the problems of realization [1] and decoupling [2], [3]. These applications are illustrative of the ways in which the structure theorem can be used.

We consider systems of the form

$$\dot{\mathbf{x}} = A\mathbf{x} + B\mathbf{u}, \qquad \mathbf{y} = C\mathbf{x}, \tag{1}$$

where $\mathbf{x}$ is an n-vector, called the state, $\mathbf{u}$ is an m-vector, called the input, $\mathbf{y}$ is a p-vector, called the output, and A, B, C are constant matrices of the appropriate dimension. We assume that the matrices B and C are of full rank. Now, it is well known [4], [5] that if the pair $\{A, B\}$ is controllable, then there is a Lyapunov transformation Q such that the system

$$\dot{\mathbf{z}} = QAQ^{-1}\mathbf{z} + QB\mathbf{u}, \qquad \mathbf{y} = CQ^{-1}\mathbf{z} \tag{2}$$

is in "companion" form. The systems (1) and (2) are equivalent and have the same transfer matrix $T(s)$. In § 2, we shall show that if state variable feedback of the form $\mathbf{u} = F\mathbf{x} + \mathbf{w}$ (or $\mathbf{u} = FQ^{-1}\mathbf{z} + \mathbf{w}$) is applied to (1) (or (2)), then the resulting transfer matrix $T_F(s)$ is of the form $\hat{C}S(s)\boldsymbol{\delta}_F^{-1}(s)\hat{B}_m$, where $\hat{C}, \hat{B}_m$ are constant matrices, $S(s)$ is a matrix of single-termed monic polynomials in s, and $\boldsymbol{\delta}_F(s)$ is a matrix of polynomials in s whose coefficients depend on $A + BF$. This result is generalized to systems which are not completely controllable in § 3 and applied to the problems of realization (§ 4) and decoupling (§ 5).

2. A structure theorem for controllable systems. Suppose that the system (1) is completely controllable. Let $K = [B, AB, \cdots, A^{n-1}B]$. Then the $n \times nm$ matrix K has rank n and it is possible to define a *lexicographic* basis for R_n consisting of the first n linearly independent columns of K possibly reordered (cf. [5]). We let L be the matrix whose columns are the elements of the "lexicographic" basis so that

$$L = [\mathbf{b}_1, A\mathbf{b}_1, \cdots, A^{\sigma_1 - 1}\mathbf{b}_1, \mathbf{b}_2, \cdots, A^{\sigma_2 - 1}\mathbf{b}_2, \cdots, A^{\sigma_m - 1}\mathbf{b}_m], \tag{3}$$

where $\mathbf{b}_1, \cdots, \mathbf{b}_m$ are the columns of B. Setting

$$d_0 = 0, \qquad d_k = \sum_{i=1}^{k} \sigma_i, \qquad k = 1, 2, \cdots, m, \tag{4}$$

* Received by the editors November 12, 1968, and in revised form April 14, 1969.

† NASA Electronics Research Center, Cambridge, Massachusetts 02139.

‡ Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912 and NASA Electronics Research Center, Cambridge, Massachusetts. The work of this author was supported in part by the National Science Foundation under Grant GK-2788.

and letting $\mathbf{l}'_k$ be the d_kth row of L^{-1}, we can see that the matrix Q given by

$$(5) \qquad Q = \begin{bmatrix} \mathbf{l}'_1 \\ \mathbf{l}'_1 A \\ \vdots \\ \mathbf{l}'_1 A^{\sigma_1 - 1} \\ \vdots \\ \mathbf{l}'_m Q^{\sigma_m - 1} \end{bmatrix}$$

generates a Lyapunov transformation for which (2) is in companion form [4], [5]. More precisely, if we let $\hat{A} = QAQ^{-1}$, $\hat{B} = QB$, and $\hat{C} = CQ^{-1}$, then (2) becomes

$$(6) \qquad \dot{\mathbf{z}} = \hat{A}\mathbf{z} + \hat{B}\mathbf{u}, \qquad \mathbf{y} = \hat{C}\mathbf{z},$$

where $\hat{A} = (\hat{a}_{ij})$ is a block matrix of the form

$$(7) \qquad \hat{A} = \begin{bmatrix} \hat{A}_{11} & \cdots & \hat{A}_{1m} \\ \hat{A}_{21} & \cdots & \hat{A}_{2m} \\ \vdots & & \vdots \\ \hat{A}_{m1} & \cdots & \hat{A}_{mm} \end{bmatrix}$$

with $\hat{A}_{ii}$ a $\sigma_i \times \sigma_i$ companion matrix given by

$$(8) \qquad \hat{A}_{ii} = \begin{bmatrix} 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & & 0 & 1 \\ \hat{a}_{d_i, d_{i-1}+1} & \hat{a}_{d_i, d_{i-1}+2} & & \hat{a}_{d_i, d_i - 1} & \hat{a}_{d_i, d_i} \end{bmatrix}$$

and $\hat{A}_{ij}$ a $\sigma_i \times \sigma_j$ matrix given by

$$(9) \qquad \hat{A}_{ij} = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & & 0 \\ \hat{a}_{d_i, d_{j-1}+1} & & \hat{a}_{d_i, d_j} \end{bmatrix}$$

for $i \neq j$ and with $\hat{B} = (\hat{b}_{ij})$ an $n \times m$ matrix given by

$$\hat{B} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \hat{b}_{d_1,2} & \hat{b}_{d_1,3} & & \hat{b}_{d_1,m} \\ 0 & 0 & 0 & & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & \hat{b}_{d_2,3} & & \hat{b}_{d_2,m} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & & 1 \end{bmatrix}. \tag{10}$$

We now have the following proposition.

PROPOSITION 2.1. *Let* $\mathbf{u} = F\mathbf{x} + \mathbf{w} = \hat{F}\mathbf{z} + \mathbf{w}$, *where* $\hat{F} = FQ^{-1}$. *Then the transfer matrices of the systems* $\dot{\mathbf{x}} = (A + BF)\mathbf{x} + B\mathbf{w}$, $\mathbf{y} = C\mathbf{x}$ *and* $\dot{\mathbf{z}} = (\hat{A} + \hat{B}\hat{F})\mathbf{z} + \hat{B}\mathbf{w}$, $\mathbf{y} = \hat{C}\mathbf{z}$ *are the same.*

Proof. Simply note that $C(sI - A - BF)^{-1}B = CQ^{-1}Q(sI - A - BF)^{-1} \times Q^{-1}QB = CQ^{-1}[(sI - QAQ^{-1} - QBFQ^{-1})]^{-1}QB = \hat{C}(sI - \hat{A} - \hat{B}\hat{F})^{-1}\hat{B}$.

Since $\hat{B}$ as given by (10) has zero rows except for the d_1th, d_2th, $\cdots$, d_mth rows, we need only calculate the corresponding columns of $(sI - \hat{A} - \hat{B}\hat{F})^{-1}$ in order to obtain the transfer matrix $T_F(s) = C(sI - A - BF)^{-1}B = \hat{C}(sI - \hat{A} - \hat{B}\hat{F})^{-1}\hat{B}$. Moreover, $\hat{B}\hat{F}$ has zero rows except for the d_1th, d_2th, $\cdots$, d_mth rows and so $\hat{A} + \hat{B}\hat{F}$ is again a block matrix of exactly the same form as $\hat{A}$. In other words, $\hat{A} + \hat{B}\hat{F} = (\phi_{ij})$ is a block matrix of the form

$$\hat{A} + \hat{B}\hat{F} = \begin{bmatrix} \Phi_{11} & \cdots & \Phi_{1m} \\ \Phi_{21} & \cdots & \Phi_{2m} \\ \vdots & & \\ \Phi_{m1} & \cdots & \Phi_{mm} \end{bmatrix}, \tag{11}$$

where Φ_{ii} is a $\sigma_i \times \sigma_i$ companion matrix given by

$$\Phi_{ii} = \begin{bmatrix} 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ \phi_{d_i,d_{i-1}+1} & \phi_{d_i,d_{i-1}+2} & \cdots & \phi_{d_i,d_{i-1}} & \phi_{d_i,d_i} \end{bmatrix} \tag{12}$$

and Φ_{ij} is a $\sigma_i \times \sigma_j$ matrix given by

$$\Phi_{ij} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \\ \phi_{d_i,d_{j-1}+1} & \phi_{d_i,d_{j-1}+2} & \cdots & \phi_{d_i,d_j} \end{bmatrix} \tag{13}$$

for $i \neq j$. These two simple and obvious observations are basic to the structure theorem, Theorem 2.2.

THEOREM 2.2. *Suppose that the pair (A, B) is controllable and let $T_F(s) = C(sI - A - BF)^{-1}B$ be the transfer matrix of the system $\dot{\mathbf{x}} = (A + BF)\mathbf{x} + B\mathbf{w}$, $\mathbf{y} = C\mathbf{x}$. Then*

$$T_F(s) = \hat{C}S(s)\boldsymbol{\delta}_F^{-1}(s)\hat{B}_m, \tag{14}$$

where $\hat{C} = CQ^{-1}$, $S(s)$ is the $n \times m$ matrix given by

$$S(s) = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ s & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ s^{\sigma_1 - 1} & 0 & & 0 \\ 0 & 1 & & 0 \\ \vdots & \vdots & & \vdots \\ 0 & s^{\sigma_2 - 1} & & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & & s^{\sigma_m - 1} \end{bmatrix}, \tag{15}$$

$\boldsymbol{\delta}_F(s)$ is the $m \times m$ matrix $(\delta_{F,ij}(s))$ with entries given by $\delta_{F,ii}(s) = \det(sI_{\sigma_i} - \Phi_{ii})$ and $\delta_{F,ij}(s) = -\phi_{d_i,d_{j-1}+1} - s\phi_{d_i,d_{j-1}+2} - \cdots - s^{\sigma_i - 1}\phi_{d_i,d_j}$ for $i \neq j$, and $\hat{B}_m$ is the $m \times m$ matrix given by

$$\hat{B}_m = \begin{bmatrix} 1 & \hat{b}_{d_1 2} & \cdots & \hat{b}_{d_1 m} \\ 0 & 1 & \cdots & \hat{b}_{d_2 m} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & & 1 \end{bmatrix}, \tag{16}$$

where $\hat{B} = QB = (\hat{b}_{ij})$.

Proof. In view of Proposition 2.1, we need only show that $\hat{C}(sI - \hat{A} - \hat{B}\hat{F})^{-1}\hat{B} = \hat{C}S(s)\boldsymbol{\delta}_F^{-1}(s)\hat{B}_m$. To do this, it will be sufficient to show that

$$(sI - \hat{A} - \hat{B}\hat{F})^{-1}\hat{B} = S(s)\boldsymbol{\delta}_F^{-1}(s)\hat{B}_m \tag{17}$$

or, equivalently, that

$$(sI - \hat{A} - \hat{B}\hat{F})S(s) = \hat{B}\hat{B}_m^{-1}\boldsymbol{\delta}_F(s). \tag{18}$$

But (18) is an immediate consequence of the definitions of $S(s)$ and $\boldsymbol{\delta}_F(s)$. Thus the theorem is established.

This seemingly innocuous and easily proved theorem has, as we shall see, a number of significant consequences. For a beginning, we note that $\hat{C}$, $S(s)$ and $\hat{B}_m$ are invariant under state feedback and that if $p = m$, then the inverse system [6] to (1) exists if and only if $C^*(s) = \hat{C}S(s)$ is nonsingular.

COROLLARY 2.3. *Let $\Delta_F(s) = \det(sI - A - BF)$. Then $\Delta_F(s) = \det(\boldsymbol{\delta}_F(s))$ and if $p = m$, then*

$$\det T_F(s) = \det C^*(s)/\Delta_F(s), \tag{19}$$

where $T_F(s) = N_F(s)/\Delta_F(s)$ (i.e., $N_F(s)$ is the numerator of the transfer matrix).

Proof. By the definition of $T_F(s)$, we have $T_F(s) = N_F(s)/\Delta_F(s)$. It follows from the theorem that

$$\frac{N_F(s)}{\Delta_F(s)} = \frac{C^*(s)D_F(s)\hat{B}_m}{\det(\boldsymbol{\delta}_F(s))}, \tag{20}$$

where $\boldsymbol{\delta}_F^{-1}(s) = D_F(s)/\det(\boldsymbol{\delta}_F(s))$. However $\Delta_F(s)$ and $\det(\boldsymbol{\delta}_F(s))$ are both monic polynomials of degree n and the entries in $N_F(s)$ are polynomials of at most degree $n - 1$. It follows that $\Delta_F(s) = \det(\boldsymbol{\delta}_F(s))$, and hence, that (19) holds (since $\det(\boldsymbol{\delta}_F^{-1}(s)) = 1/\det(\boldsymbol{\delta}_F(s))$ and $\det \hat{B}_m = 1$).

COROLLARY 2.4. $\boldsymbol{\delta}_F(s) = \boldsymbol{\delta}_0(s) - \hat{B}_m\hat{F}S(s)$.

Proof. From (18), it follows that $\hat{B}\hat{B}_m^{-1}\boldsymbol{\delta}_0(s) - \hat{B}\hat{F}S(s) = \hat{B}\hat{B}_m^{-1}\boldsymbol{\delta}_F(s)$. Equating the nonzero rows in this equality gives us the corollary.

We observe that entirely analogous results can be obtained for observable systems by a consideration of the dual system [1], [7]

$$\dot{\mathbf{x}} = A'\mathbf{x} + C'\mathbf{v}, \qquad \mathbf{y} = B'\mathbf{x} \tag{21}$$

which is controllable if and only if (1) is observable. While we shall not derive the results for observable systems here, we shall use them without further ado in the sequel.

3. A general structure theorem. Consider the system (1) and again let $K = [B, AB, \cdots, A^{n-1}B]$. However, we no longer assume that (1) is controllable, and so the $n \times nm$ matrix K has rank r with $r \leqq n$. To obtain a structure theorem in this general context, we shall consider a controllable extension of (1) and apply Theorem 2.2. With this in mind, we let $q = n - r$ and W be the r-dimensional subspace of R_n spanned by the columns of K. Denoting the orthogonal complement of W by $W^\perp$ so that $R_n = W \oplus W^\perp$ and letting $\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_q$ be a basis of $W^\perp$, we consider the system

$$\dot{\mathbf{x}} = A\mathbf{x} + B_e\mathbf{v}, \qquad \mathbf{y} = C\mathbf{x}, \tag{22}$$

where B_e is the $n \times (m + q)$ matrix given by $B_e = [B\, \boldsymbol{\beta}_1 \cdots \boldsymbol{\beta}_q]$. The system (22) is controllable and there is a Lyapunov transformation Q_e which carries (22) into block companion form. We note that Q_e is a nonsingular $n \times n$ matrix. It follows that the system

$$\dot{\mathbf{z}} = \hat{A}\mathbf{z} + \hat{B}\mathbf{u}, \qquad \mathbf{y} = \hat{C}\mathbf{z}, \tag{23}$$

where $\hat{A} = Q_e^{-1}AQ_e$, $\hat{B} = Q_eB$, and $\hat{C} = CQ_e^{-1}$ is equivalent to (1). Moreover, the matrix $\hat{A}$ is in block companion form, the last $n - r$ rows of $\hat{B}$ are 0, and the lower left-hand $(n - r) \times r$ block of $\hat{A}$ is 0. Thus, the last $n - r$ rows of $\hat{A}$ cannot be altered by state variable feedback of the form $\mathbf{u} = \hat{F}\mathbf{z} + \mathbf{w}$. We now have the following theorem.

THEOREM 3.1. *Let $T_F(s) = C(sI - A - BF)^{-1}B$ be the transfer matrix of the system $\dot{\mathbf{x}} = (A + BF)\mathbf{x} + B\mathbf{w}$, $\mathbf{y} = C\mathbf{x}$. Then*

$$T_F(s) = \frac{\hat{C}S(s)\Delta_{F,u}(s)\boldsymbol{\delta}_{F,c}^{-1}(s)\hat{B}_m}{\Delta_{F,u}(s)}, \tag{24}$$

where $\hat{C} = CQ_e^{-1}$, $S(s)$ is the $n \times m$ matrix given by

$$S(s) = \left[\begin{array}{ccc} 1 & 0 & \cdots\ 0 \\ s & 0 & 0 \\ \vdots & \vdots & \vdots \\ s^{\sigma_1 - 1} & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & s^{\sigma_2 - 1} & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & s^{\sigma_m - 1} \\ \hdashline \vdots & \vdots & \vdots \\ 0 & 0 & 0 \end{array}\right] \tag{25}$$

(*with $\mathbf{b}_1, A\mathbf{b}_1, \cdots, A^{\sigma_1 - 1}\mathbf{b}_1, \cdots, A^{\sigma_m - 1}\mathbf{b}_m$ a "lexicographic" basis of the range of K so that $\sum_{i=1}^m \sigma_i = r$*), *$\Delta_{F,u}(s) = \det \boldsymbol{\delta}_{F,u}(s)$, $\boldsymbol{\delta}_F(s)$ is the $(m + q) \times (m + q)$ matrix $(\delta_{F,ij}(s))$ with entries given by $\delta_{F,ii}(s) = \det(sI - \Phi_{ii})$ and $\delta_{F,ij}(s) = -\phi_{d_i,d_{j-1}+1} - \cdots - s^{\sigma_1 - 1}\phi_{d_i,d_j}$ for $i \neq j$, where $d_k = \sum_{i=1}^k \sigma_i$, $\sigma_i = 1$ for $i = m + 1, \cdots, m + q$, and $\hat{A} + \hat{B}\hat{F} = (\phi_{ij}) = [\Phi_{ij}]$ so that*

$$\begin{aligned} \boldsymbol{\delta}_F(s) &= \left[\begin{array}{ccc;{2pt/2pt}ccc} \delta_{F,11}(s) & \cdots & \delta_{F,1m}(s) & \delta_{F,1,m+1}(s) & \cdots & \delta_{F,1,m+q}(s) \\ \vdots & & & \vdots & & \vdots \\ \delta_{F,m1}(s) & \cdots & \delta_{F,mm}(s) & \vdots & & \vdots \\ \hdashline & & & \delta_{F,m+1,m+1}(s) & \cdots & \delta_{F,m+1,m+q}(s) \\ & \mathbf{0} & & \vdots & & \\ & & & \delta_{F,m+q,m+1}(s) & \cdots & \delta_{F,m+q,m+q}(s) \end{array}\right] \\ &= \left[\begin{array}{c;{2pt/2pt}c} \boldsymbol{\delta}_{F,c}(s) & \boldsymbol{\delta}_{F,cu}(s) \\ \hdashline \mathbf{0} & \boldsymbol{\delta}_{F,u}(s) \end{array}\right][1] \end{aligned} \tag{26}$$

and where $\hat{B}_m$ is the $m \times m$ matrix consisting of the nonzero rows of $\hat{B}$.

The proof is a simple application of Theorem 2.2 and is left to the reader.

COROLLARY 3.2. *$\boldsymbol{\Delta}_{F,u}(s)$ is independent of F and the uncontrollable poles of the system $\dot{\mathbf{x}} = (A + BF)\mathbf{x} + B\mathbf{w}$, $\mathbf{y} = C\mathbf{x}$ are the zeros of $\boldsymbol{\Delta}_{F,u}(s)[= \boldsymbol{\Delta}_{0,u}(s)]$.*

[1] $\boldsymbol{\delta}_{F,cu}(s)$ involves only constant terms, and the off-diagonal terms in $\boldsymbol{\delta}_{F,u}(s)$ are constant.

Corollary 3.2 is simply a statement of the fact that the uncontrollable poles cannot be altered by state variable feedback. We also note that the factorization (24) involves the well-known pole-zero cancellation of the uncontrollable portion of the system [8].

COROLLARY 3.3. *The matrices $\hat{C}$, $S(s)$ and $\hat{B}_m$ are invariant under state variable feedback.*

COROLLARY 3.4. *Let $p = m$ and $C^*(s) = \hat{C}S(s)$. Then the inverse system to* (1) *exists if and only if $C^*(s)$ is nonsingular.*

COROLLARY 3.5. *Let $p = m$ and let $\Delta_F(s) = \det \delta_F(s)$. Then* $\det(T_F(s)) = (\det C^*(s))(\Delta_{F,u}(s))/\Delta_F(s)$, *where* $\Delta_F(s) = \Delta_{F,u}(s)\Delta_{F,c}(s)$.

We again observe that entirely analogous results can be obtained for systems which are not observable by a consideration of the dual system (21). We use these results without further ado in the sequel.

4. The problem of realization. We now apply the structure theorem to obtain an algorithm analogous to that of Mayne [10] for solving the problem of realization [1], [9]. More precisely, we consider the following problem.

REALIZATION PROBLEM. Let $T(s)$ be a $p \times m$ matrix whose entries $t_{ij}(s)$ are rational functions of s. Suppose that $t_{ij}(s) = n_{ij}(s)/d_{ij}(s)$, where $n_{ij}(s)$ and $d_{ij}(s)$ are relatively prime and $\deg n_{ij}(s) < \deg d_{ij}(s)$. Then, determine a triple $\{A, B, C\}$ of matrices such that

$$T(s) = C(sI - A)^{-1}B, \tag{27}$$

$\{A, B\}$ is controllable and $\{A, C\}$ is observable. Such a triple is called a minimal realization of $T(s)$ (see [1], [9]).

Kalman and B. L. Ho [9] proved that the realization problem has a solution and provided a constructive procedure for determining a minimal realization. Mayne [10] obtained a constructive algorithm for determining minimal realizations using the ideas of [1]. Here, we present an alternate derivation based on the structure theorem. A computer program has been developed for applying the algorithm.

The basic steps in the algorithm are now given.

Step 1. Calculate the least common multiple of the denominator polynomials $\{d_{1j}(s), \cdots, d_{pj}(s)\}$ in each column of $T(s)$.

Step 2. Construct a standard controllable realization $\{A_c, B_c, C_c\}$ (not necessarily minimal).

Step 3. Construct a minimal realization by applying a suitable transformation to $\{A_c', C_c', B_c'\}$.

We shall examine each of these steps in detail paying particular attention to Step 2.

Now let $g_j(s)$ be the least common multiple of the denominator polynomials $\{d_{1j}(s), \cdots, d_{pj}(s)\}$ (which are assumed, for convenience, to be monic). Let h_j denote the degree of $g_j(s)$ and let $T^*(s)$ be the $p \times m$ matrix given by

$$T^*(s) = \begin{bmatrix} n_{11}^*(s)/g_1(s) & \cdots & n_{1m}^*(s)/g_m(s) \\ \vdots & & \\ n_{p1}^*(s)/g_1(s) & \cdots & n_{pm}^*(s)/g_m(s) \end{bmatrix}, \tag{28}$$

where $n_{ij}^*(s) = n_{ij}(s)g_j(s)/d_{ij}(s)$. In other words, $T^*(s)$ is obtained from $T(s)$ by multiplying each numerator $n_{ij}(s)$ by $g_j(s)/d_{ij}(s)$ and replacing each denominator $d_{ij}(s)$ by $g_j(s)$. The construction of $T^*(s)$ completes Step 1.

Let $n_1 = \sum_{j=1}^m h_j$ and $p_k = \sum_1^k h_j$. Since $g_j(s)$ is the least common multiple of $\{d_{1j}(s), \cdots, d_{pj}(s)\}$ and $\deg n_{ij}(s) < \deg d_{ij}(s)$ and the $d_{ij}(s)$ are assumed monic, we have

$$g_j(s) = s^{h_j} + \gamma_{j1}s^{h_j-1} + \cdots + \gamma_{jh_j}, \tag{29}$$

$$n_{ij}^*(s) = v_{ij1}s^{h_j-1} + v_{ij2}s^{h_j-2} + \cdots + v_{ijh_j} \tag{30}$$

for all i, j and suitable constants γ_{jk}, v_{ijk}. Let $A_{c,j}$ be a companion matrix corresponding to $g_j(s)$ so that

$$A_{c,j} = \begin{bmatrix} 0 & 1 & & 0 \\ 0 & 0 & & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & & 0 \\ -\gamma_{jh_j} & -\gamma_{jh_j-1} & \cdots & -\gamma_{j1} \end{bmatrix}, \tag{31}$$

and let A_c be the $n_1 \times n_1$ block diagonal matrix given by

$$A_c = \begin{bmatrix} A_{c,1} & & \mathbf{0} \\ & A_{c,2} & \\ & & \ddots \\ \mathbf{0} & & A_{c,m} \end{bmatrix}. \tag{32}$$

If B_c is the $n_1 \times m$ matrix with zero entries in all but the p_kth rows, each of which is zero except for a one in the kth column, then the pair $\{A_c, B_c\}$ is controllable. We now have the following proposition.

PROPOSITION 4.1. *Let C_c be the $m \times n_1$ matrix given by*

$$C_c = \left[\begin{array}{cccc|ccc|cc} v_{11h_1} & v_{11h_1-1} & \cdots & v_{111} & v_{12h_2} & \cdots & v_{121} & \cdots & v_{1m1} \\ v_{21h_1} & v_{21h_1-1} & \cdots & v_{211} & v_{22h_2} & \cdots & v_{221} & & v_{2m1} \\ \vdots & \vdots & & \vdots & \vdots & & & & \vdots \\ v_{p1h_1} & v_{p1h_1-1} & & v_{p11} & v_{p2h_2} & \cdots & v_{p21} & \cdots & v_{pm1} \end{array}\right] \tag{33}$$

Then $\{A_c, B_c, C_c\}$ is a controllable realization of $T(s)$.

Proof. Since $\{A_c, B_c\}$ is controllable, it follows from the structure theorem, Theorem 2.2 and the definitions of A_c, B_c, C_c, that

$$C_c(sI - A_c)^{-1}B_c = C_c^*(s)\delta_c^{-1}(s)\hat{B}_{c,m}, \tag{34}$$

where $\hat{B}_{c,m} = I_m$, $\delta_c^{-1}(s) = \operatorname{diag}[1/g_1(s), \cdots, 1/g_m(s)]$, and $C_c^*(s) = (n_{ij}^*(s))$. Since $n_{ij}^*(s)/g_j(s) = n_{ij}(s)/d_{ij}(s)$, we deduce that $C_c(sI - A_c)^{-1}B_c = (n_{ij}(s)/d_{ij}(s)) = T(s)$. Thus, the proposition is established. This proposition completes the description of Step 2.

With regard to Step 3, we consider the triple $\{A'_c, C'_c, B'_c\}$ and apply a Lyapunov transformation Q_e to it of the type used in § 3. Letting n be the rank of $[C'_c A'_c C'_c \cdots A'^{n_1-1}_c C'_c]$ and setting $\hat{A}'_c = Q_e A'_c Q_e^{-1}$, $\hat{C}'_c = Q_e C'_c$, $\hat{B}'_c = B'_c Q_e^{-1}$, we have

$$(35) \qquad \hat{C}'_c = \begin{bmatrix} C' \\ \mathbf{0}_{n_1-n,p} \end{bmatrix}, \qquad A'_c = \left[\begin{array}{c|c} A' & * \\ \mathbf{0}_{n_1-n,n} & * \end{array}\right]$$

and $\hat{B}'_c = [B' \quad *_{m,n_1-n}]$, where C' is $n \times p$, A' is $n \times n$ and B' is $m \times n$. Since $T(s) = C_c(sI - A_c)^{-1}B_c$, it follows that $T'(s) = \hat{B}'_c(sI - \hat{A}'_c)^{-1}\hat{C}'_c = B'(sI - A')^{-1}C'$ or, equivalently, that $T(s) = C(sI - A)^{-1}B$. Thus, $\{A, B, C\}$ is a realization of $T(s)$. But $\{A, B, C\}$ is both controllable and observable, and hence is a minimal realization [9]. The triple $\{A, B, C\}$ is in "observable canonical form." The actual available program also produces a minimal realization in "controllable canonical form" as well as all the relevant Lyapunov transformations. A sample of the computer program print-out for an example by Kalman [1, p. 182] is given in the Appendix. A detailed write-up and listing of the program can be obtained from the authors.

5. The problem of decoupling. We now apply the structure theorem to obtain some results related to the problem of decoupling. This problem has been examined previously by a number of authors (e.g., [2], [3]) and a number of relevant questions have been resolved. Here, our main emphasis will be on the question of pole assignability. More precisely, consider the following problem.

DECOUPLING PROBLEM. Let $\dot{\mathbf{x}} = A\mathbf{x} + B\mathbf{u}$, $\mathbf{y} = C\mathbf{x}$ be an m-input, m-output system. Does there exist a pair of matrices $\{F, G\}$ such that the transfer matrix

$$(36) \qquad C(sI - A - BF)^{-1}BG = T_{F,G}(s)$$

is diagonal and nonsingular? (i.e., does the state variable feedback $\mathbf{u} = F\mathbf{x} + G\mathbf{w}$ "decouple" the system?).

A necessary and sufficient condition for the existence of a decoupling pair was first given in [2]. In particular, it has been shown that the system

$$(37) \qquad \dot{\mathbf{x}} = A\mathbf{x} + B\mathbf{u}, \qquad \mathbf{y} = C\mathbf{x}$$

can be decoupled if and only if B^* is nonsingular, where B^* is the $m \times m$ matrix given by

$$(38) \qquad B^* = \begin{bmatrix} \mathbf{c}_1 A^{f_1} B \\ \vdots \\ \mathbf{c}_m A^{f_m} B \end{bmatrix}$$

with $\mathbf{c}_i$ the ith row of C, and $f_i = \min[\{j : \mathbf{c}_i A^j B \neq 0\}, n - 1]$. B^* and the f_i can also be characterized in the following way (cf. [3]): let $T_{F,G,i}(s)$ be the ith row of the transfer matrix $T_{F,G}(s)$; then $f_i = \min[\{j : \lim_{s\to\infty} s^{j+1} T_{F,G,i}(s) \neq \mathbf{0}\}, n - 1]$ and $B^*G = \lim_{s\to\infty} \Delta(s) T_{F,G}(s)$, where $\Delta(s)$ is a diagonal matrix with entries s^{f_i+1}. It can be shown [2], [3] that B^* and the f_i are invariant under state variable feedback.

Here, we shall use the structure theorem to answer the following questions.

Question 1. Assuming that (37) can be decoupled, what is the maximum number of closed loop poles which can be arbitrarily specified while simultaneously decoupling the system?

Question 2. Assuming that (37) can be decoupled, which closed loop poles are invariant under decoupling state variable feedback?

Question 3. How can a decoupling pair which specifies the maximum number of closed loop poles be implemented?

While these questions are to some degree resolved in [2] and [3], we provide a complete and elementary answer to them here.

Let $T(s)$ be the transfer matrix of (40). Then $T(s) = C^*(s)(\Delta_u(s)/\Delta_u(s))\delta_{\mathbf{0},c}^{-1}(s)\hat{B}_m$, where $C^*(s) = \hat{C}S(s)$ and $\Delta_u(s) = \Delta_{\mathbf{0},u}(s)$ by the structure theorem, Theorem 3.1. We recall that $C^*(s)$ and $\Delta_u(s)$ are invariant under state variable feedback. Now we let $p_i(s)$ be the greatest common divisor of the polynomials which are the entries in the ith row $C_i^*(s)$ of $C^*(s)$. We note that $p_i(s)$ is invariant under state variable feedback. We let r_i be the degree of $p_i(s)$ and we use the notation ∂_p to denote the degree of a polynomial (thus, $r_i = \partial_{p_i}$). We now have the following theorem.

THEOREM 5.1. *Suppose that the system* (37) *can be decoupled. Then* (i) *the maximum number v of closed loop poles which can be arbitrarily specified while decoupling is given by*

$$v = \sum_{i=1}^{m} (r_i + f_i + 1) \tag{39}$$

and (ii) *the invariant poles under decoupling feedback are the zeros of* $\Delta_u(s)$ *and* $\{\det C^*(s)\}/\prod_{i=1}^{m} p_i(s)$.

Proof. Let $\{F, G\}$ be any decoupling pair. Then $T_{F,G}(s) = C(sI - A - BF)^{-1}BG$ is a diagonal matrix with entries $n_{ii}(s)/d_{ii}(s)$, where $n_{ii}(s)$ and $d_{ii}(s)$ are relatively prime. We note that, since $f_i = \min\{j : \lim_{s\to\infty} s^{j+1}T_{F,G,i}(s) \neq \mathbf{0}\}$,[2] $\partial_{n_{ii}} = \partial_{d_{ii}} - f_i - 1$. It follows from Corollary 3.5 and the definition of the $p_i(s)$ that

$$\prod_{i=1}^{m} \frac{n_{ii}(s)}{d_{ii}(s)} = \prod_{i=1}^{m} p_i(s) \det C_{\text{II}}^*(s)\frac{\Delta_u(s)}{\Delta_F(s)} \det G, \tag{40}$$

where $C_{\text{II}}^*(s)$ is the matrix with rows $C_{\text{II},i}^*(s) = (1/p_i(s))C_i^*(s)$. Since $\Delta_F(s) = \Delta_u(s)\Delta_{F,c}(s)$, we have

$$\partial_{F,c} = \sum_{i=1}^{m} (r_i + f_i + 1) + \partial_{\text{II}}^*, \tag{41}$$

where ∂_{II}^* is the degree of $\det C_{\text{II}}^*(s)$ and $\partial_{F,c}$ is the degree of $\Delta_{F,c}(s)$. Now, it is clear from Theorem 3.1 that

$$T_{F,G,i}(s)G^{-1}\hat{B}_m^{-1}\boldsymbol{\delta}_{F,c}(s) = C_i^*(s), \tag{42}$$

and hence that $n_{ii}(s)$ is a common divisor of the entries in $C_i^*(s)$ (since $n_{ii}(s)$ and $d_{ii}(s)$ are relatively prime). In other words, $n_{ii}(s)$ *must divide* $p_i(s)$, and so $\partial_{n_{ii}} \leqq r_i$. Since no more than $\sum_{i=1}^{m} \partial_{d_{ii}}$ poles are assignable through $\{F, G\}$ and $\sum_{i=1}^{m} \partial_{d_{ii}} = \sum_{i=1}^{m} (\partial_{n_{ii}} + f_i + 1)$, we deduce that at most $v = \sum_{i=1}^{m} (r_i + f_i + 1)$ poles are assignable while decoupling.

[2] Note that B^* is nonsingular.

Writing $T_{F,G}(s)$ as a diagonal matrix with entries $q_{ii}(s)/\Delta_F(s) = n_{ii}(s)/d_{ii}(s)$, we deduce that $q_{ii}(s)$ must divide $p_i(s)\Delta_F(s)$ or, equivalently, that

$$\frac{q_{ii}(s)}{\Delta_F(s)} = \frac{p_i(s)}{q_i(s)} \tag{43}$$

for $i = 1, \cdots, m$ and polynomials $q_i(s)$ with $\partial_{qi} = r_i + f_i + 1$. It follows that $\det T_{F,G}(s) = \prod_{i=1}^m p_i(s)/\prod_{i=1}^m q_i(s)$, and hence, from (40), that

$$\begin{aligned}\Delta_F(s) &= \det C_{\text{II}}^*(s)\,\Delta_u(s)\det G \prod_{i=1}^{m} q_i(s)\\ &= \frac{\det C^*(s)}{\prod_{i=1}^m p_i(s)}\Delta_u(s)\det G\prod_{i=1}^{m} q_i(s).\end{aligned} \tag{44}$$

Since $C_{\text{II}}^*(s)$ is invariant under decoupling feedback, it follows that the zeros of $\Delta_u(s)$ and $\det C_{\text{II}}^*(s)$ are invariant poles under decoupling feedback.

Thus, to complete the proof we need only construct a decoupling pair $\{F, G\}$ such that the resulting polynomials $q_i(s)$ are arbitrary polynomials of degree $r_i + f_i + 1$. To begin with, we note that the transfer matrix

$$T(s) = C^*(s)\frac{\Delta_u(s)}{\Delta_u(s)}\boldsymbol{\delta}_{0,c}^{-1}(s)\hat{B}_m = P(s)C_{\text{II}}^*(s)\frac{\Delta_u(s)}{\Delta_u(s)}\boldsymbol{\delta}_{0,c}^{-1}(s)\hat{B}_m,$$

where $P(s)$ is a diagonal matrix with entries $p_i(s)$. Setting

$$T_{\text{II}}(s) = C_{\text{II}}^*(s)\frac{\Delta_u(s)}{\Delta_u(s)}\boldsymbol{\delta}_{0,c}^{-1}(s)\hat{B}_m, \tag{45}$$

we can easily see that $r_i + f_i = \min\{j : \lim_{s\to\infty} s^{j+1}T_{\text{II},i}(s) \neq \mathbf{0}\}$ and that $B_{\text{II}}^* = \lim_{s\to\infty} \boldsymbol{\Delta}_{\text{II}}(s)T_{\text{II}}(s) = B^*$, where $\boldsymbol{\Delta}_{\text{II}}(s)$ is a diagonal matrix with entries $s^{r_i+f_i+1}$. (Note that the $p_i(s)$ are monic.) Moreover, as $C^*(s)$ is given by $\hat{C}S(s)$ and $p_i(s)$ is the greatest common divisor of the entries in $C_i^*(s)$, we can write $C_{\text{II}}^*(s)$ in the form $\hat{C}_{\text{II}}S(s)$ for some constant matrix $\hat{C}_{\text{II}}$ (where $S(s)$ is given by (25)). In other words, $T_{\text{II}}(s)$ is the transfer matrix of the system $\dot{\mathbf{x}} = A\mathbf{x} + B\mathbf{u}$, $\mathbf{v}_{\text{II}} = C_{\text{II}}\mathbf{x}$, where $C_{\text{II}} = \hat{C}_{\text{II}} = \hat{C}_{\text{II}}Q$ (and Q is the Lyapunov transformation corresponding to (37)). Since $P(s)$ is diagonal, it will be sufficient to construct a decoupling pair $\{F, G\}$ for the system

$$\dot{\mathbf{x}} = A\mathbf{x} + B\mathbf{u}, \qquad \mathbf{y}_{\text{II}} = C_{\text{II}}\mathbf{x} \tag{46}$$

such that the closed loop poles are arbitrarily placed. However, letting $d_i = r_i + f_i$ and applying the synthesis procedure of [2, p. 655], we find that (46) can be decoupled and all its closed loop poles assigned. To be more explicit, if $q_i(s) = s^{d_i+1} - \sum_{j=0}^{d_i} m_j^i s^j$,[3] then the decoupling pair is given by

$$F = B^{*^{-1}}\left[\sum_0^d M_k C_{\text{II}}A^k - A^*\right], \qquad G = B^{*^{-1}}, \tag{47}$$

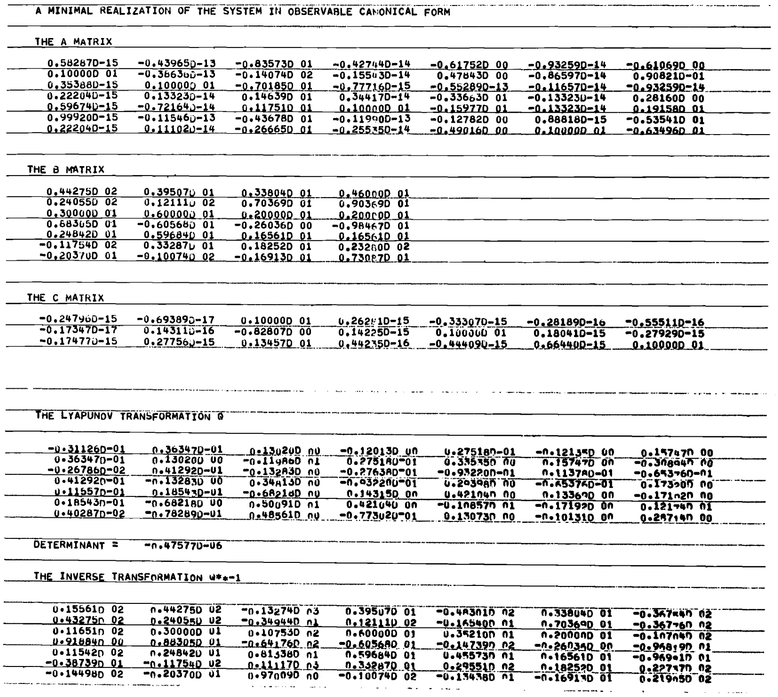
[3] Clearly, it is enough to consider the case of a monic $q_i(s)$.

where $d = \max d_i$, the M_k are diagonal matrices with entries m_k^i (i.e., $M_k = \mathrm{diag}\,[m_k^1, \cdots, m_k^m]$), and $A^* = (C_{\mathrm{II},i}A^{d_i+1})$ (i.e., the ith row of A^* is given by $C_{\mathrm{II},i}A^{d_i+1}$). This completes the proof.

Appendix. A sampleof the computer print-out for an example by Kalman [1, p. 182] is given here. See Table 1. The transfer matrix is given by

$$T(s) = \begin{bmatrix} \dfrac{3(s+3)(s+5)}{(s+1)(s+2)(s+4)} & \dfrac{6(s+1)}{(s+2)(s+4)} & \dfrac{2s+7}{(s+3)(s+4)} & \dfrac{2s+5}{(s+2)(s+3)} \\ \dfrac{2}{(s+3)(s+5)} & \dfrac{1}{s+3} & \dfrac{2(s+5)}{(s+1)(s+2)(s+3)} & \dfrac{8(s+2)}{(s+1)(s+3)(s+5)} \\ \dfrac{2(s^2+7s+18)}{(s+1)(s+3)(s+5)} & \dfrac{-2s}{(s+1)(s+3)} & \dfrac{1}{s+3} & \dfrac{2(5s^2+27s+34)}{(s+1)(s+3)(s+5)} \end{bmatrix}.$$

TABLE 1

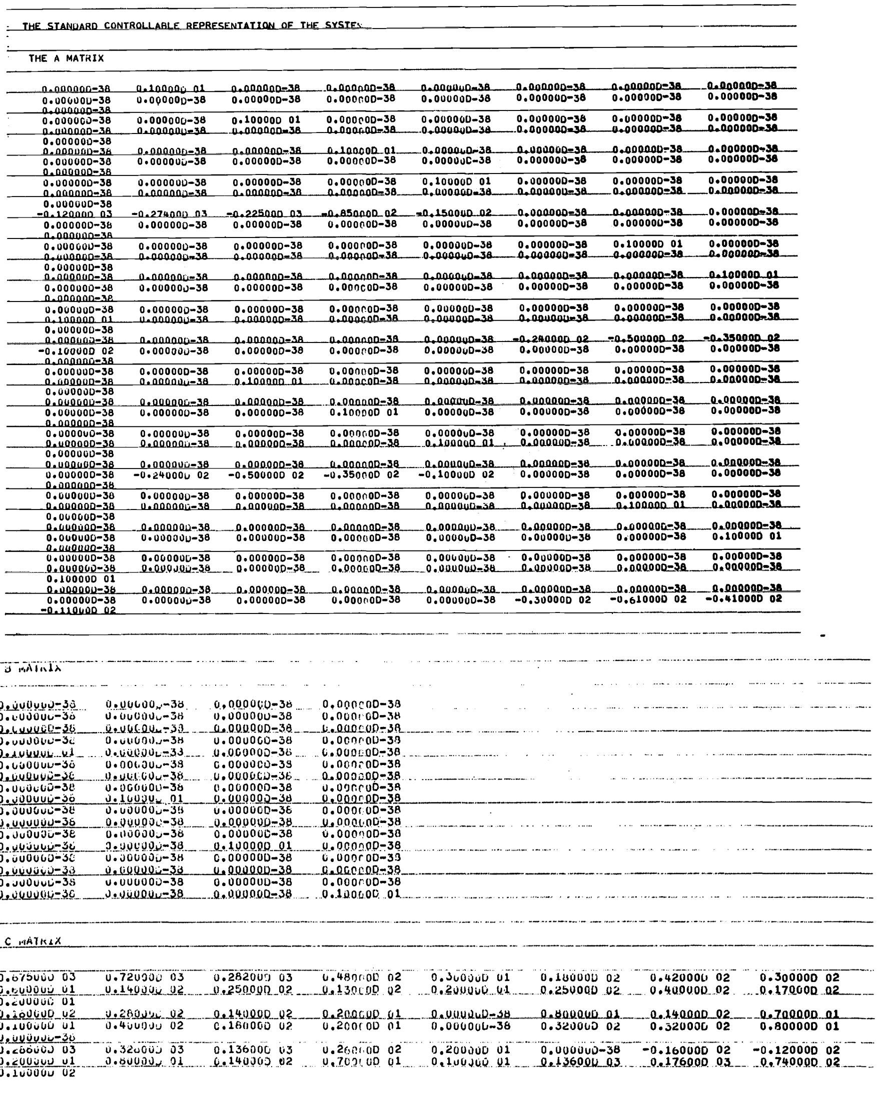


```
A MINIMAL REALIZATION OF THE SYSTEM IN OBSERVABLE CANONICAL FORM

THE A MATRIX

   0.58287D-15  -0.43965D-13  -0.83573D 01  -0.42744D-14  -0.61752D 00  -0.93259D-14  -0.61069D 00
   0.10000D 01  -0.38636D-13  -0.14074D 02  -0.15543D-14   0.47843D 00  -0.86597D-14   0.90821D-01
   0.35388D-15   0.10000D 01  -0.70185D 01  -0.77716D-15  -0.55289D-13  -0.11657D-14  -0.93259D-14
   0.22204D-15   0.13323D-14   0.14639D 01   0.34417D-14  -0.33663D 01  -0.13323D-14   0.28160D 00
   0.59674D-15  -0.72164D-14   0.11751D 01   0.10000D 01  -0.15977D 01  -0.13323D-14   0.19158D 01
   0.99920D-15  -0.11546D-13  -0.43678D 01  -0.11990D-13  -0.12782D 00   0.88818D-15  -0.53541D 01
   0.22204D-15   0.11102D-14  -0.26665D 01  -0.25535D-14  -0.49016D 00   0.10000D 01  -0.63496D 01

THE B MATRIX

   0.44275D 02   0.39507D 01   0.33804D 01   0.46000D 01
   0.24055D 02   0.12111D 02   0.70369D 01   0.90369D 01
   0.30000D 01   0.60000D 01   0.20000D 01   0.20000D 01
   0.68305D 01  -0.60568D 01  -0.26036D 00  -0.98467D 01
   0.24842D 01   0.59684D 01   0.16561D 01   0.16561D 01
  -0.11754D 02   0.33287D 01   0.18252D 01   0.23260D 02
  -0.20370D 01  -0.10074D 02  -0.16913D 01   0.73087D 01

THE C MATRIX

  -0.24796D-15  -0.69389D-17   0.10000D 01   0.26271D-15  -0.33307D-15  -0.28189D-16  -0.55511D-16
  -0.17347D-17   0.14311D-16  -0.82807D 00   0.14225D-15   0.10000D 01   0.18041D-15  -0.27929D-15
  -0.17477D-15   0.27756D-15   0.13457D 01   0.44235D-16  -0.44409D-15   0.66440D-15   0.10000D 01

THE LYAPUNOV TRANSFORMATION Q

    -0.31126D-01   0.36347D-01   0.13020D 00  -0.12013D 00   0.27518D-01  -0.12135D 00   0.15747D 00
     0.36347D-01   0.13020D 00  -0.11986D 01   0.27518D-01   0.33535D 00   0.15747D 00  -0.30894D 00
    -0.26786D-02   0.41292D-01  -0.13283D 00  -0.27638D-01  -0.93220D-01   0.11378D-01  -0.65376D-01
     0.41292D-01  -0.13283D 00   0.34813D 00  -0.93220D-01   0.20398D 00  -0.65376D-01   0.17320D 00
     0.11557D-01   0.18543D-01  -0.68218D 00   0.14315D 00   0.42104D 00   0.13369D 00  -0.17102D 00
     0.18543D-01  -0.68218D 00   0.50091D 01   0.42104D 00  -0.10857D 01  -0.17192D 00   0.12174D 01
     0.40287D-02  -0.78289D-01   0.48561D 00  -0.77302D-01   0.13073D 00  -0.10131D 00   0.28714D 00

DETERMINANT =    -0.47577D-06

THE INVERSE TRANSFORMATION Q**-1

     0.15561D 02   0.44275D 02  -0.13274D 03   0.39507D 01  -0.48301D 02   0.33804D 01  -0.36784D 02
     0.43275D 02   0.24055D 02  -0.34944D 01   0.12111D 02  -0.16540D 01   0.70369D 01  -0.36776D 02
     0.11651D 02   0.30000D 01   0.10753D 02   0.60000D 01   0.35210D 01   0.20000D 01  -0.10704D 02
     0.91884D 00   0.88305D 01  -0.64176D 02  -0.60568D 01  -0.14739D 02  -0.26036D 00  -0.96819D 01
     0.11542D 02   0.24842D 01   0.81338D 01   0.59684D 01   0.45573D 01   0.16561D 01  -0.96961D 01
    -0.38739D 01  -0.11754D 02   0.11117D 03   0.33287D 01   0.29551D 02   0.18252D 01   0.22737D 02
    -0.14498D 02  -0.20370D 01   0.97009D 00  -0.10074D 02  -0.13438D 01  -0.16913D 01   0.21905D 02
```

TABLE 1 (*cont.*)

```
  A MINIMAL REALIZATION OF THE SYSTEM IN CONTROLLABLF CANONICAL FORM

  THE A MATRIX

     0.22204D-14   0.10000D 01   0.61062D-15  -0.22204D-15  -0.49960D-15   0.16653D-15  -0.88818D-15
     0.65812D 00  -0.65420D 00   0.26423D 01   0.35746D 01   0.12810D 01   0.13161D 01  -0.13085D 01
     0.17764D-14   0.77716D-15   0.41633D-14   0.10000D 01   0.83267D-15   0.36082D-15  -0.19984D-14
     0.21867D 01   0.10560D 00   0.44178D 01  -0.22209D 01   0.23761D 01   0.35836D 00   0.29787D 00
    -0.17764D-14   0.88818D-15  -0.11879D-13  -0.17764D-14  -0.17764D-14   0.10000D 01  -0.35527D-14
    -0.14509D 02  -0.36610D 01  -0.45726D 02  -0.19048D 02  -0.15816D 02  -0.83097D 01   0.56365D-01
     0.22956D 01  -0.82157D-14   0.20499D 01  -0.88818D-15   0.11986D 01   0.66613D-15  -0.37719D 01

  THE B MATRIX

     0.16653D-15  -0.11102D-15  -0.55511D-16   0.00000D-38
     0.10000D 01   0.13323D-14   0.66613D-15   0.63892D 00
     0.97145D-16  -0.55511D-16  -0.55511D-16  -0.11102D-15
     0.88818D-15   0.10000D 01   0.22204D-15   0.83450D 00
     0.17208D-14   0.22204D-15  -0.27756D-16  -0.14433D-14
     0.17764D-14  -0.00000D-38   0.10000D 01   0.28901D 01
    -0.11102D-14  -0.66613D-15  -0.16653D-15   0.10000D 01

  THE C MATRIX

     0.11651D 02   0.30000D 01   0.10753D 02   0.60000D 01   0.35210D 01   0.20000D 01  -0.10709D 02
     0.18944D 01  -0.79909D-14  -0.77008D 00   0.10000D 01   0.16416D 01  -0.28621D-14  -0.83450D 00
     0.11806D 01   0.20000D 01   0.15439D 02  -0.20000D 01   0.33942D 01   0.10000D 01   0.75011D 01


  THE NUMERATOR OF THE MATRIX TRANSFER FUNCTION

  S** 2     3.000000   0.000000   0.000000   0.000000
  S** 1    24.000000   6.000000   2.000000   2.000000
  S** 0    45.000000   6.000000   7.000000   5.000000

  S** 1     0.000000   0.000000   2.000000   8.000000
  S** 0     2.000000   1.000000  10.000000  16.000000

  S** 2     2.000000   0.000000   0.000000  10.000000
  S** 1    14.000000  -2.000000   0.000000  54.000000
  S** 0    36.000000   0.000000   1.000000  68.000000


  THE DENOMINATOR OF THE MATRIX TRANSFER FUNCTION

  S** 3     1.000000   0.000000   0.000000   0.000000
  S** 2     7.000000   1.000000   1.000000   1.000000
  S** 1    14.000000   6.000000   7.000000   5.000000
  S** 0     8.000000   8.000000  12.000000   6.000000

  S** 3     0.000000   0.000000   1.000000   1.000000
  S** 2     1.000000   0.000000   6.000000   9.000000
  S** 1     8.000000   1.000000  11.000000  23.000000
  S** 0    15.000000   3.000000   6.000000  15.000000

  S** 3     1.000000   0.000000   0.000000   1.000000
  S** 2     9.000000   1.000000   0.000000   9.000000
  S** 1    23.000000   4.000000   1.000000  23.000000
  S** 0    15.000000   3.000000   3.000000  15.000000
```

TABLE 1 (*cont.*)



```
THE STANDARD CONTROLLABLE REPRESENTATION OF THE SYSTE

THE A MATRIX

  0.000000D-38  0.100000D 01  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38
  0.000000D-38  0.000000D-38  0.100000D 01  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.100000D 01  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.100000D 01  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38
 -0.120000D 03 -0.274000D 03 -0.225000D 03 -0.850000D 02 -0.150000D 02  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.100000D 01  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.100000D 01
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.100000D 01  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38 -0.280000D 02 -0.500000D 02 -0.350000D 02
 -0.100000D 02  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38  0.000000D-38  0.100000D 01  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.100000D 01  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.100000D 01  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38 -0.240000D 02 -0.500000D 02 -0.350000D 02 -0.100000D 02  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.100000D 01  0.000000D-38
  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.100000D 01
  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.100000D 01
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38 -0.300000D 02 -0.610000D 02 -0.410000D 02
 -0.110000D 02

THE B MATRIX

  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.100000D 01  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38  0.100000D 01  0.000000D-38  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38  0.000000D-38  0.100000D 01  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.000000D-38
  0.000000D-38  0.000000D-38  0.000000D-38  0.100000D 01

THE C MATRIX

  0.675000D 03  0.720000D 03  0.282000D 03  0.480000D 02  0.300000D 01  0.180000D 02  0.420000D 02  0.300000D 02
  0.600000D 01  0.140000D 02  0.250000D 02  0.130000D 02  0.200000D 01  0.250000D 02  0.400000D 02  0.170000D 02
  0.200000D 01
  0.160000D 02  0.280000D 02  0.140000D 02  0.200000D 01  0.000000D-38  0.800000D 01  0.140000D 02  0.700000D 01
  0.100000D 01  0.400000D 02  0.160000D 02  0.200000D 01  0.000000D-38  0.320000D 02  0.320000D 02  0.800000D 01
  0.000000D-38
  0.288000D 03  0.320000D 03  0.136000D 03  0.260000D 02  0.200000D 01  0.000000D-38 -0.160000D 02 -0.120000D 02
 -0.200000D 01  0.800000D 01  0.140000D 02  0.700000D 01  0.100000D 01  0.136000D 03  0.176000D 03  0.740000D 02
  0.100000D 02
```

TABLE 1 (*cont.*)

```
THE LYAPUNOV TRANSFORMATION Q

 -0.16169D-02  0.72712D-03  0.16155D-02  0.61147D-03  0.69839D-04  0.40036D-03  0.14590D-02  0.95216D-03
  0.17227D-03 -0.59273D-03 -0.88856D-03 -0.23354D-03 -0.99882D-05 -0.13212D-02  0.74196D-03  0.12405D-02
  0.26430D-03
  0.72712D-03  0.16155D-02  0.61147D-03  0.69839D-04 -0.42172D 00  0.14580D-02  0.95216D-03  0.17227D-03
 -0.11756D 00 -0.88856D-03 -0.23354D-03 -0.99882D-05  0.66927D-01  0.74196D-03  0.12405D-02  0.26430D-03
 -0.59393D-01
  0.16155D-02  0.61147D-03  0.69839D-04 -0.42172D 00  0.56523D 01  0.95216D-03  0.17227D-03 -0.11756D 00
  0.10869D 01 -0.23354D-03 -0.99882D-05  0.66927D-01 -0.63592D 00  0.12405D-02  0.26430D-03 -0.59393D-01
  0.54455D 00
 -0.12132D-02 -0.33720D-03  0.24659D-02  0.22844D-02  0.41642D-03  0.36255D-02 -0.26107D-02 -0.38084D-02
 -0.73479D-03 -0.84228D-02  0.20040D-01  0.98460D-02  0.10771D-02  0.41208D-02 -0.43145D-02 -0.15748D-02
  0.77452D-03
 -0.33720D-03  0.24659D-02  0.22844D-02  0.41642D-03 -0.51728D 00 -0.26107D-02 -0.38084D-02 -0.73479D-03
  0.18417D 00  0.20040D-01  0.98460D-02  0.10771D-02 -0.11552D 01 -0.43145D-02 -0.15748D-02  0.77452D-03
  0.19561D 00
 -0.41453D-03  0.79065D-03 -0.90982D-03 -0.39271D-03 -0.28788D-04  0.48172D-02 -0.14238D-02 -0.39208D-02
 -0.92027D-03 -0.59243D-02  0.85321D-02  0.40155D-02  0.36645D-03  0.10065D-01 -0.56163D-02 -0.74013D-02
 -0.10795D-02
  0.79065D-03 -0.90982D-03 -0.39271D-03 -0.28788D-04  0.71628D-01 -0.14238D-02 -0.39208D-02 -0.92027D-03
  0.10201D 00  0.85321D-02  0.40155D-02  0.36645D-03 -0.42863D 00 -0.56163D-02 -0.74013D-02 -0.10795D-02
  0.35598D 00
  0.35159D 00 -0.58335D 00  0.61770D 00  0.26202D 00  0.27401D-01  0.14946D 00  0.33605D-01 -0.13209D 00
 -0.41515D-01  0.50119D-01 -0.10638D 00 -0.10928D 00 -0.20857D-01  0.10320D 00 -0.62712D-01 -0.41273D-01
  0.81666D-03
  0.30504D-17  0.11597D 00 -0.36359D 00  0.55925D 00  0.85131D-01  0.40028D 00  0.44866D-01 -0.45461D 00
 -0.13684D 00  0.11169D 00 -0.15080D-01 -0.19919D 00 -0.50070D-01 -0.86500D-01 -0.88721D-01  0.25285D 00
  0.11197D 00
 -0.12056D-16  0.19854D-16  0.14684D 00 -0.74329D 00 -0.11666D-01  0.34586D 00  0.31131D-01 -0.40392D 00
 -0.12063D 00  0.52996D-01  0.46684D-01 -0.13406D 00 -0.39171D-01 -0.12012D 00 -0.57204D-01  0.27017D 00
  0.11145D 00
  0.39796D-17 -0.25448D-17 -0.98532D-17  0.71592D-01 -0.99219D 00  0.41767D-01  0.11946D-01 -0.66552D-01
 -0.22825D-01  0.86509D-02 -0.10779D-01  0.39134D-01  0.99154D-02 -0.34256D-01  0.21061D-01  0.45102D-02
 -0.11400D-01
  0.16163D-16  0.12511D-16 -0.91588D-16 -0.19469D-16  0.79827D-01  0.11072D 00  0.11849D 00 -0.34816D 00
 -0.14202D 00  0.91918D-01 -0.24431D 00  0.62013D 00  0.16626D 00 -0.25657D 00  0.31825D 00 -0.30769D 00
 -0.28943D 00
 -0.29532D-16  0.16148D-16 -0.87650D-16  0.16635D-16  0.13584D-16  0.64271D 00 -0.59622D 00  0.40462D 00
  0.16241D 00  0.57647D-01 -0.54691D-01  0.42486D-01  0.19261D-01 -0.14652D 00  0.12332D 00 -0.74164D-01
 -0.50076D-01
 -0.26899D-16  0.31518D-16 -0.47668D-16 -0.18393D-16  0.21101D-18  0.17407D-15  0.99678D-01 -0.28569D 00
  0.66648D 00 -0.89067D-01  0.20921D 00 -0.35528D 00 -0.74387D-01 -0.21400D 00  0.31808D 00 -0.36427D 00
 -0.23626D-01
 -0.36513D-17  0.23252D-16 -0.75060D-16 -0.17360D-16  0.33170D-18  0.18128D-15 -0.61905D-16  0.12383D 00
 -0.68200D 00 -0.10805D 00  0.24048D 00 -0.39981D 00 -0.10386D 00 -0.21050D 00  0.32019D 00 -0.36169D 00
 -0.43676D-01
 -0.66853D-17  0.72750D-17  0.86647D-17  0.32836D-16  0.37222D-17  0.62582D-17  0.86306D-17 -0.20178D-17
  0.27493D-01 -0.38416D-01  0.73053D-01 -0.55538D-01 -0.48730D 00  0.12955D-01 -0.64640D-01  0.24490D 00
 -0.82920D 00
  0.88997D-17 -0.80463D-17 -0.33047D-17 -0.20298D-16 -0.42977D-30  0.43880D-17 -0.43809D-17 -0.29259D-16
  0.10241D-16  0.31235D-01 -0.93704D-01  0.28111D 00 -0.84334D 00 -0.15617D-01  0.46852D-01 -0.14056D 00
  0.42167D 00

DETERMINANT =    -0.23268D-15
```

REFERENCES

[1] R. E. KALMAN, *Mathematical description of linear dynamical systems*, this Journal, 1 (1963), pp. 152–192.

[2] P. L. FALB AND W. A. WOLOVICH, *Decoupling in the design and synthesis of multivariable control systems*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 651–659.

[3] E. G. GILBERT, *The decoupling of multivariable systems by state feedback*, this Journal, 7 (1969), pp. 50–63.

[4] D. G. LUENBERGER, *Canonical forms for linear multivariable systems*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 290–293.

[5] W. A. WOLOVICH, *On the stabilization of controllable systems*, Ibid., AC-13 (1968).

[6] R. W. BROCKETT, *Poles, zeroes, and feedback: state space interpretation*, Ibid., AC-10 (1965), pp. 129–135.

[7] R. E. KALMAN, P. L. FALB AND M. A. ARBIB, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1968.

[8] E. G. GILBERT, *Controllability and observability in multivariable control systems*, this Journal, 1 (1963), pp. 128–151.

[9] B. L. HO AND R. E. KALMAN, *Effective construction of linear state-variable models from input/output functions*, Proc. Third Allerton Conference, 1965, pp. 449–459.

[10] D. Q. MAYNE, *Computational procedure for the minimal realization of transfer function matrices*, Proc. IEEE, 115 (1968), pp. 1363–1368.

SIAM J. CONTROL
Vol. 7, No. 3, August 1969

A GENERALIZED TRANSFORM THEORY FOR CAUSAL OPERATORS*

P. L. FALB† AND M. I. FREEDMAN‡

1. Introduction. Let G be a locally compact Abelian group and let H be a complex Hilbert space. Here we develop a transform theory on $L_2(G, H)$ and $L_1(G, \mathscr{L}(H, H))$ (where $\mathscr{L}(H, H)$ is the space of bounded linear maps of H into itself). We were naturally led to the problem of developing such a theory during the course of obtaining stability theorems for systems defined on locally compact Abelian groups [1].

We shall assume that the reader is familiar with the standard transform theory for $L_1(G, C)$ and $L_2(G, C)$ (see [2]) and is familiar with the theory of integration of Banach-space-valued functions [3]. Moreover, we shall suppose that the reader knows the Gelfand theory for commutative Banach algebras and shall use this theory freely [2]. Let μ denote Haar measure on G and let $\hat{G}$ denote the character group of G. Elements of $\hat{G}$ are usually denoted by γ and their action on G is written as (γ, g). We recall [4] that $\hat{G}$ is also a locally compact Abelian group with respect to the topology of uniform convergence on compact subsets of G. We let m denote Haar measure on $\hat{G}$.

We begin our analysis in the next section with a discussion of the notion of causality. Then, in § 3, we develop the basic transform theory. We next introduce two Banach algebras, B_P and W_P, with B_P the algebra of causal operators in $L_1(G, \mathscr{L}(H, H)) \oplus \{\Delta\}$ (Δ an identity) and W_P an algebra of causal operators from $L_2(G, H)$ into $L_2(G, H)$ (see § 4). In § 5, we prove the following theorem which is our main result.

THEOREM. *Let Φ be an approximable element of B_P. Then*

$$\mathrm{SPEC}_{W_P}\,\Phi \subset \bigcup_{M \in \mathscr{M}} \operatorname{spec} \hat{\Phi}(M) \subset \mathrm{SPEC}_{B_P}\,\Phi,$$

where $\mathscr{M}$ is the maximal ideal space of a suitable Banach algebra, the "$\hat{\ }$" indicates a suitable Gelfand representation, and the "spec's" *indicate suitable spectrums.*

Approximability means, loosely speaking, that Φ is a limit of finite-dimensional maps. The theorem is used to obtain a generalized circle criterion for stability in [1]. Finally, we present some illustrative examples and make some concluding comments in § 6.

2. Causality. We now introduce a generalization of the notions of truncation and causality.

DEFINITION 2.1. Let $P \subset G$ be a semigroup of positive Haar measure and let P' be the negative of P in G (i.e., $P' = -P$). Let $P' + g_0$ be the subset of G given by

$$(1) \qquad P' + g_0 = \{g \in G : g = g_1 + g_0, g_1 \in P'\},$$

* Received by the editors June 6, 1968, and in revised form April 15, 1969.

† Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912 and NASA Electronics Research Center, Cambridge, Massachusetts. The work of this author was supported in part by the National Science Foundation under Grant GK-967 and in part by the United States Air Force under Grant AFOSR 693-67.

‡ NASA Electronics Research Center, Cambridge, Massachusetts 02139.

and let χ_{g_0} be the characteristic function of $P' + g_0$. If f is a measurable map of G into H, then the *truncation of f at g_0*, f_{g_0}, is given by

$$f_{g_0}(g) = \begin{cases} f(g), & g \in P' + g_0, \\ 0, & g \notin P' + g_0; \end{cases} \tag{2}$$

i.e., $f_{g_0} = \chi_{g_0} f$.[1]

With a notion of truncation at hand, we can define causality.

DEFINITION 2.2. Let Φ be a map of $L_2(G, H)$ into $L_2(G, H)$. Then Φ is called *causal with respect to P* (or simply causal) if

$$(\Phi x)_{g_0}(\cdot) = (\Phi(x_{g_0}(\cdot))_{g_0} \tag{3}$$

for all g_0 in G and x in $L_2(G, H)$.[2]

We observe that if $G = R$ (R the real numbers) and $P = [0, \infty)$, then Definition 2.2 coincides with the usual notion of causality or nonanticipativeness [5]. We also observe that if $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are causal with respect to P, then $\mathbf{\Phi\Psi}$ is causal with respect to P as

$$(\mathbf{\Phi\Psi} x)_{g_0}(\cdot) = (\mathbf{\Phi}\{\mathbf{\Psi} x\})_{g_0}(\cdot) = (\mathbf{\Phi}\{\mathbf{\Psi} x_{g_0}(\cdot)\}_{g_0})_{g_0} = (\mathbf{\Phi\Psi} x_{g_0}(\cdot))_{g_0}.$$

Example 2.3. Let N be a map of H into H with $|N(h)| \leqq C|h|$ for all h in H (where $|\cdot|$ denotes the norm on H). Then the map $\mathbf{N}$ of $L_2(G, H)$ into $L_2(G, H)$ given by $(\mathbf{N}x)(g) = N(x(g))$ is causal.

Example 2.4. Let ψ be an element of $L_1(G, \mathscr{L}(H, H))$ with support contained in P.[3] Then the map $\boldsymbol{\psi}$ of $L_2(G, H)$ into $L_2(G, H)$ given by

$$(\boldsymbol{\psi} x)(g) = \int_G \psi(g - g')x(g')\,d\mu$$

is well-defined and causal with respect to P. The integral converges by virtue of Fubini's theorem [3].

3. Transform theory. We now develop the rudimentary transform theory on $L_2(G, H)$ and $L_1(G, \mathscr{L}(H, H))$. We observe, first of all, that if x is an element of $L_2(G, H)$ and h is an element of H, then $x_h(\cdot) = \langle x(\cdot), h\rangle$ is in $L_2(G, C)$ and thus has a transform $\hat{x}_h(\cdot)$ in $L_2(\hat{G}, C)$. This leads us to the following definition.

DEFINITION 3.1. Let $x(\cdot)$ be an element of $L_2(G, H)$. Then the *Fourier transform of* $x(\cdot)$, in symbols, $\hat{x}(\cdot)$, is an element of $L_2(\hat{G}, H)$ such that $\hat{x}_h(\cdot) = \hat{x}(\cdot)_h$ for all h in H (i.e., $\langle \hat{x}(\cdot), h\rangle = \hat{x}_h(\cdot)$).[4]

We then have the following proposition.

PROPOSITION 3.2. *Let $x(\cdot)$ be an element of $L_2(G, H)$. Then* (i) *$\hat{x}(\cdot)$ is defined,* (ii) *$\hat{x}(\cdot)$ is unique and* (iii) $\|x(\cdot)\|^2 = \|\hat{x}(\cdot)\|^2$ (*Parseval*).

Proof. Since $x(\cdot)$ is essentially separably-valued [3], there is a closed separable subspace H_1 of H with H_1 containing the range of x. Let $A_1 = \{e_i : i = 1, \cdots\}$ be an orthonormal basis of H_1 and let H_2 be the orthogonal complement of H_1 in H.

[1] Strictly speaking, we should write $f_{g_0,P}$ since truncation depends on the semigroup P. However, we usually deal with a fixed P and so the distinction is unnecessary.

[2] Note that if x is in $L_2(G, H)$, then x_{g_0} is also in $L_2(G, H)$ for all g_0 in G.

[3] This means that $\mu(P^c \cap \overline{\{g : \psi(g) \neq 0\}}) = 0$, where P^c is the complement of P in G.

[4] This is often called the Plancherel transform.

Then

$$
\begin{aligned}
& x(g) = \sum_1^\infty \langle x(g), e_i\rangle e_i = \sum_1^\infty x_{e_i}(g)e_i, \qquad g \in G,\\
(4) \qquad & |x(g)|^2 = \sum_1^\infty |x_{e_i}(g)|^2, \qquad g \in G,\\
& \|x(\cdot)\|^2 = \int_G |x(g)|^2\, d\mu = \sum_1^\infty \int_{\hat{G}} |x_{e_i}(g)|^2\, d\mu < \infty,
\end{aligned}
$$

since $x(\cdot) \in L_2(G, H)$. Now let $\hat{x}_{e_i}(\cdot)$ denote the transform of $x_{e_i}(\cdot)$ and let us set

$$
(5) \qquad \hat{x}(\gamma) = \sum_1^\infty \hat{x}_{e_i}(\gamma)e_i
$$

for γ in $\hat{G}$. Since $\int_G |x_{e_i}(g)|^2\, d\mu = \int_G |\hat{x}_{e_i}(\gamma)|^2\, dm$ (see [2]), it follows that

$$
(6) \qquad \|\hat{x}(\cdot)\|^2 = \int_G |\hat{x}(\gamma)|^2\, dm = \sum_1^\infty \int_{\hat{G}} |\hat{x}_{e_i}(\gamma)|^2\, dm = \|x(\cdot)\|^2,
$$

and so (5) defines an element of $L_2(\hat{G}, H)$ for which (iii) holds. Now let h be an element of H. Then $h = h_1 + h_2$, $h_1 \in H_1$, $h_2 \in H_2$ and $\langle x(\cdot), h\rangle = \langle x(\cdot), h_1\rangle$ since h_2 is orthogonal to H_1. However, the range of $\hat{x}$ is contained in H_1 and so we need only show that $\hat{x}_{h_1}(\cdot) = \hat{x}(\cdot)_{h_1}$. In view of (4), $\langle x(\cdot), h_1\rangle = \sum_1^\infty x_{e_i}(\cdot) \cdot \langle e_i, h_1\rangle$, and so

$$
(7) \qquad \hat{x}_{h_1}(\cdot) = \sum_1^\infty \hat{x}_{e_i}(\cdot)\langle e_i, h_1\rangle = \left\langle \sum_1^\infty \hat{x}_{e_i}(\cdot)e_i, h_1\right\rangle = \hat{x}(\cdot)_{h_1},
$$

which establishes (i). Since (ii) is obvious, the proof is complete.

COROLLARY 3.3 (Plancherel). *Let x, y be elements of $L_2(G, H)$. Then*

$$
(8) \qquad \langle x, y\rangle = \int_G \langle x(g), y(g)\rangle\, d\mu = \int_G \langle \hat{x}(\gamma), \hat{y}(\gamma)\rangle\, dm = \langle \hat{x}, \hat{y}\rangle,
$$

where $\hat{x}$, $\hat{y}$ are the Fourier transforms of x, y.

Proof. The proof is an immediate consequence of (iii) and the well-known formula $4\langle x, y\rangle = \{\|x + y\|^2 - \|x - y\|^2\} + i\{\|x + iy\|^2 - \|x - iy\|^2\}$.

We note that since $\hat{G} = G$, the Fourier transform of an element of $L_2(\hat{G}, H)$ is an element of $L_2(G, H)$. Moreover, if $x \in L_2(G, H)$, then $x_h(g) = \hat{\hat{x}}_h(-g)$, and so $x(g) = \hat{\hat{x}}(-g)$ from which it follows that the Fourier transform has an inverse. We shall make use of this observation in § 4.

Let $\mathscr{C}(\hat{G}, \mathscr{L}(H, H))$ denote the space of continuous functions from $\hat{G}$ into $\mathscr{L}(H, H)$. We then have the following definition.

DEFINITION 3.4. Let $\Phi(\cdot)$ be an element of $L_1(G, \mathscr{L}(H, H))$. Then the *Fourier transform of* Φ, in symbols, $\hat{\Phi}(\cdot)$, is the map of $\hat{G}$ into $\mathscr{L}(H, H)$ given by

$$
(9) \qquad \hat{\Phi}(\gamma) = \int_G \overline{(\gamma, g)}\Phi(g)\, d\mu,
$$

where (γ, g) denotes the action of γ on g.

PROPOSITION 3.5. *Let* $\Phi(\cdot)$ *be an element of* $L_1(G, \mathscr{L}(H, H))$. *Then* (i) $\hat{\Phi}(\cdot)$ *is in* $\mathscr{C}(\hat{G}, \mathscr{L}(H, H))$, (ii) $\hat{\Phi}(\cdot)$ *is bounded and* (iii) *the bounded bilinear map* T_γ *of* $H \times H$ *into* C *given by*

$$T_\gamma(h_1, h_2) = \int_G \overline{(\gamma, g)} \langle \Phi(g)h_1, h_2 \rangle \, d\mu \tag{10}$$

is uniquely represented by $\hat{\Phi}(\gamma)$ *for all* γ *in* $\hat{G}$, *i.e.*, $T_\gamma(h_1, h_2) = \langle \hat{\Phi}(\gamma)h_1, h_2 \rangle$.

Proof. Since $|(\gamma, g)| = 1$ for all γ and $\Phi(\cdot)$ is in $L_1(G, \mathscr{L}(H, H))$, it is clear that $\hat{\Phi}(\cdot)$ is well-defined and that

$$\|\hat{\Phi}(\gamma)\| \leqq \int_G |\overline{(\gamma, g)}| \, \|\Phi(g)\| \, d\mu = \|\Phi(\cdot)\|_1{}^{5} \tag{11}$$

so that $\hat{\Phi}(\cdot)$ is bounded. Moreover, we have

$$\|\hat{\Phi}(\gamma_1) - \hat{\Phi}(\gamma_2)\| \leqq \sup_{g \in G} |\overline{(\gamma_1, g)} - \overline{(\gamma_2, g)}| \, \|\Phi(\cdot)\|_1. \tag{12}$$

Since $\|\Phi(\cdot)\|$ is in $L_1(G, C)$ and the topology on $\hat{G}$ is that of uniform convergence on compacta, $\hat{\Phi}(\cdot)$ is uniformly continuous. The assertion (iii) is obvious.

We observe that (iii) implies that the mapping φ of G into C given by $\varphi(g) = \langle \Phi(g)h_1, h_2 \rangle$ for fixed h_1, h_2 in H has the Fourier transform $\hat{\varphi}(\gamma) = \langle \hat{\Phi}(\gamma)h_1, h_2 \rangle$ (note that φ is in $L_1(G, C)$). We also note that if G is not discrete and if $\hat{\varphi}(\gamma)$ is a constant, then $\hat{\varphi}(\gamma) = 0$ for all γ in $\hat{G}$ (see [2]) and, hence, that if $\langle \hat{\Phi}(\gamma)h_1, h_2 \rangle = \lambda \langle h_1, h_2 \rangle$ for all h_1, h_2 with λ a constant, then $\hat{\Phi}(\gamma) = 0$ for all γ. This observation will be used in the proof of Lemma 4.1.

Now suppose that x is in $L_2(G, H)$ and that Φ is in $L_1(G, \mathscr{L}(H, H))$. Then the integral

$$\int_G \Phi(g - g_1)x(g_1) \, d\mu = (\Phi * x)(g) \tag{13}$$

is defined almost everywhere and is an element of $L_2(G, H)$ with

$$\|(\Phi * x)(\cdot)\| \leqq \|\Phi(\cdot)\|_1 \|x(\cdot)\| \tag{14}$$

provided that G is σ-finite. Since μ is translation-invariant, we can replace (13) by

$$(\Phi * x)(g) = \int_G \Phi(g_1')x(g - g_1') \, d\mu. \tag{15}$$

We then have the following lemma.

LEMMA 3.6. *If* $x \in L_2(G, H)$, $\Phi \in L_1(G, \mathscr{L}(H, H))$ *and* G *is* σ*-finite, then* $(\Phi * x)(g)$ *is defined almost everywhere and is an element of* $L_2(G, H)$ *which satisfies* (14).

Proof. Formally, we have

$$\begin{aligned} &\int_G \|(\Phi * x)(g)\|^2 \, d\mu \\ &= \int_G \left\langle \int \Phi(g_1)x(g - g_1) \, d\mu(g_1), \int \Phi(g_2)x(g - g_2) \, d\mu(g_2) \right\rangle d\mu, \end{aligned} \tag{16}$$

[5] $\|\Phi(\cdot)\|_1$ is the $L_1(G, \mathscr{L}(H, H))$ norm of $\Phi(\cdot)$.

and so it will be enough to show that the right-hand side of (16) is finite as this will imply that $\Phi * x$ is defined almost everywhere and is in $L_2(G, H)$. Now

$$\begin{aligned}
(17)\quad & \|\Phi(\cdot)\|_1^2\|x(\cdot)\|^2 \\
& \geqq \int_G \int_G \|\Phi(g_1)\| \, \|\Phi(g_2)\| \cdot \|x(\cdot)\|^2 \, d\mu(g_1) \, d\mu(g_2) \\
(18)\quad & \geqq \int_G \int_G \|\Phi(g_1)\| \, \|\Phi(g_2)\| \left\{ \int_G |x(g - g_1)| \, |x(g - g_2)| \, d\mu \right\} d\mu(g_1) \, d\mu(g_2) \\
(19)\quad & \geqq \int_G \int_G \int_G \|\Phi(g_1)\| \, |x(g - g_1)| \, \|\Phi(g_2)\| \, |x(g - g_2)| \, d\mu(g_1) \, d\mu(g_2) \, d\mu \\
(20)\quad & \geqq \int_G \int_G \int_G \langle \Phi(g_1)x(g - g_1), \Phi(g_2)x(g - g_2) \rangle \, d\mu(g_1) \, d\mu(g_2) \, d\mu
\end{aligned}$$

by virtue of Holder's inequality, the Fubini and Tonelli theorems [3] and the Schwarz inequality. But,

$$\begin{aligned}
(21)\quad & \int_G \int_G \int_G \langle \Phi(g_1)x(g - g_1), \Phi(g_2)x(g - g_2) \rangle \, d\mu(g_1) \, d\mu(g_2) \, d\mu \\
& = \int_G \left\langle \int_G \Phi(g_1)x(g - g_1) \, d\mu(g_1), \int_G \Phi(g_2)x(g - g_2) \, d\mu(g_2) \right\rangle d\mu
\end{aligned}$$

since $\int_G \int_G \langle \Phi(g_1)x(g - g_1), \Phi(g_2)x(g - g_2) \rangle \, d\mu(g_1) \, d\mu(g_2)$ exists almost everywhere by the Fubini and Tonelli theorems and satisfies the relation

$$\begin{aligned}
& \int_G \int_G \langle \Phi(g_1)x(g - g_1), \Phi(g_2)x(g - g_2) \rangle \, d\mu(g_1) \, d\mu(g_2) \\
(22)\quad & = \int_G \left\{ \left\langle \int_G \Phi(g_1)x(g - g_1) \, d\mu(g_1), \Phi(g_2)x(g - g_2) \right\rangle \right\} d\mu(g_2) \\
& = \left\langle \int_G \Phi(g_1)x(g - g_1) \, d\mu(g_1), \int_G \Phi(g_2)x(g - g_2) \, d\mu(g_2) \right\rangle .
\end{aligned}$$

Thus, the lemma is established.

In view of the lemma, we have, for G σ-finite, the following definition.

DEFINITION 3.7. If $x \in L_2(G, H)$ and $\Phi \in L_1(G, \mathscr{L}(H, H))$, then $\Phi * x$ is called the *convolution* of Φ and x.

PROPOSITION 3.8. *If* $x \in L_2(G, H)$ *and* $\Phi \in L_1(G, \mathscr{L}(H, H))$, *then* $(\Phi * x)\hat{}(\gamma) = \hat{\Phi}(\gamma)\hat{x}(\gamma)$ *for all* γ *in* $\hat{G}$. (*The hat accent to the right of the parenthesis indicates the Fourier transform of the entire expression.*)

Proof. Since $x(\cdot)$ and $(\Phi * x)(\cdot)$ are in $L_2(G, H)$, the transforms $\hat{x}$ and $(\Phi * x)\hat{}$ are defined and, since $\hat{\Phi}(\cdot)$ is bounded and continuous, $\hat{\Phi}(\cdot)\hat{x}(\cdot)$ is in

$L_2(\hat{G}, H)$. Now let h be an element of H. Then

$$\begin{aligned}\langle(\Phi * x)\hat{}(\cdot), h\rangle &= (\langle(\Phi * x)(\cdot), h\rangle)\hat{} \\ &= \left(\left\langle \int_G \Phi(g - g_1)x(g_1)\,d\mu, h\right\rangle\right)\hat{} \\ &= \left(\int_G \langle\Phi(g - g_1)x(g_1), h\rangle\,d\mu\right)\hat{}\,. \end{aligned} \tag{23}$$

Since $x(\cdot)$ is essentially separably-valued, there is a closed separable subspace H_1 of H with H_1 containing the range of $x(\cdot)$. Let $\{e_i : i = 1, \ldots\}$ be an orthonormal basis of H_1. Then

$$x(g_1) = \sum_{i=1}^{\infty} \langle x(g_1), e_i\rangle e_i = \sum_{i=1}^{\infty} x_{e_i}(g_1)e_i \tag{24}$$

for all g_1 in G. It follows that

$$\begin{aligned}\langle(\Phi * x)\hat{}(\cdot), h\rangle &= \sum_{i=1}^{\infty}\left(\int_G \langle\Phi(g - g_1)e_i, h\rangle x_{e_i}(g_1)\,d\mu\right)\hat{} \\ &= \sum_{i=1}^{\infty} (\langle\Phi(\cdot)e_i, h\rangle)\hat{}\,\hat{x}_{e_i}(\cdot)\end{aligned} \tag{25}$$

since $\int_G \langle\Phi(g - g_1)e_i, h\rangle x_{e_i}(g_1)\,d\mu$ is the convolution of elements of $L_1(G, C)$ and $L_2(G, C)$. But $\langle\hat{\Phi}(\cdot)e_i, h\rangle = (\langle\Phi(\cdot)e_i, h\rangle)\hat{}$ and so,

$$\begin{aligned}\langle(\Phi * x)\hat{}(\cdot), h\rangle &= \left\langle \hat{\Phi}(\cdot)\sum_{i=1}^{\infty} \hat{x}_{e_i}(\cdot)e_i, h\right\rangle \\ &= \langle\hat{\Phi}(\cdot)\hat{x}(\cdot), h\rangle\end{aligned} \tag{26}$$

in view of the proof of Proposition 3.2.

We shall next show that convolution can be defined for elements of $L_1(G, \mathscr{L}(H, H))$. To begin with, we have the following lemma.

LEMMA 3.9. *Let Φ, Ψ be elements of $L_1(G, \mathscr{L}(H, H))$ with G σ-finite. Then the function $\Phi(g - g_1)\Psi(g_1)$ is an element of $L_1(G \times G, \mathscr{L}(H, H))$.*

Proof. Suppose, for the moment, that $\Phi(g - g_1)\Psi(g_1)$ is measurable on $G \times G$. Then, since

$$\int_G \left\{\int_G \|\Phi(g - g_1)\|\,d\mu(g)\right\} \|\Psi(g_1)\|\,d\mu(g_1) \leqq \|\Phi(\cdot)\|_1\|\Psi(\cdot)\|_1, \tag{27}$$

it will follow [3, Corollary 15, p. 194] that $\Phi(g - g_1)\Psi(g_1)$ is integrable on $G \times G$ and that $\|\Phi(\cdot - \cdot)\Psi(\cdot)\|_1 \leqq \|\Phi(\cdot)\|_1\|\Psi(\cdot)\|_1$. Thus, it will be enough to show that $\Phi(g - g_1)\Psi(g_1)$ is measurable.

Since $\Phi(\cdot)$, $\Psi(\cdot)$ are integrable on G, there are sequences of simple functions $\Phi_n(\cdot)$, $\Psi_n(\cdot)$ which converge to $\Phi(\cdot)$, $\Psi(\cdot)$, respectively, almost everywhere [3]. Since $\Phi_n(\cdot)$, $\Psi_n(\cdot)$ are essentially finitely-valued (being simple functions), it is clear that $\Phi_n(g - g_1)\Psi_n(g_1)$ is essentially finitely-valued and is a simple function on

$G \times G$. Thus, it will be enough to show that $\Phi_n(g - g_1)\Psi_n(g_1)$ converges to $\Phi(g - g_1)(g_1)$ almost everywhere on $G \times G$.[6] Let $F = \{g_1 : \Psi_n(g_1) \to \Psi(g_1)\}$, $E = \{g' : \Phi_n(g') \to \Phi(g')\}$ so that $\mu(G - F) = 0$ and $\mu(G - E) = 0$. Let $E' = E \times G$, $E'' = \{(g, g_1) : g - g_1 \in E\}$. Then, $(g, g_1) \in E''$ implies that $\Phi_n(g - g_1)$ converges to $\Phi(g - g_1)$. Moreover, since the homeomorphism $(g', g'') \to (g' + g'', g'')$ of $G \times G$ onto $G \times G$ maps E' onto E'' and since $G \times G - E' = (G - E) \times G$ has measure zero, it follows that $(\mu \times \mu)(G \times G - E'') = 0$. Let $E''' = E'' \cap \{G \times F\}$. Then $G \times G - E''' \subset (G \times G - E'') \cup (G \times \{G - F\})$ so that $(\mu \times \mu)(G \times G - E''') = 0$. However, if $(g, g_1) \in E'''$, then $\Phi_n(g - g_1)$ converges to $\Phi(g - g_1)$, and $\Psi_n(g_1)$ converges to $\Psi(g_1)$. It follows that $\Phi_n(g - g_1)\Psi_n(g_1)$ converges to $\Phi(g - g_1)\Psi(g_1)$ on E''' (a fortiori almost everywhere on $G \times G$). Thus, the lemma is established.

COROLLARY 3.10. *Let Φ, Ψ be elements of $L_1(G, \mathscr{L}(H, H))$ with G σ-finite. Then*

$$(\Phi * \Psi)(g) = \int_G \Phi(g - g_1)\Psi(g_1)\, d\mu \tag{28}$$

is defined almost everywhere on G and is an element of $L_1(G, \mathscr{L}(H, H))$ with $\|(\Phi * \Psi)(\cdot)\|_1 \leqq \|\Phi(\cdot)\|_1 \|\Psi(\cdot)\|_1$.

We call $\Phi * \Psi$ the *convolution* of Φ and Ψ. Moreover, we have the following proposition.

PROPOSITION 3.11. *Let Φ, Ψ be elements of $L_1(G, \mathscr{L}(H, H))$ with G σ-finite. Then*

$$(\Phi * \Psi)\hat{}(\gamma) = \hat{\Phi}(\gamma)\hat{\Psi}(\gamma) \tag{29}$$

for all γ in $\hat{G}$.

Proof. Since $\overline{(\gamma, g)} = \overline{(\gamma, g - g_1)}\,\overline{(\gamma, g_1)}$, we have

$$\begin{aligned}(\Phi * \Psi)\hat{}(\gamma) &= \int_G \overline{(\gamma, g)} \left\{ \int_G \Phi(g - g_1)\Psi(g_1)\, d\mu(g_1) \right\} d\mu(g) \\ &= \int_G \left\{ \int_G \overline{(\gamma, g - g_1)}\Phi(g - g_1)\overline{(\gamma, g_1)}\Psi(g_1)\, d\mu(g_1) \right\} d\mu(g) \\ &= \int_G \left\{ \int_G \overline{(\gamma, g - g_1)}\Phi(g - g_1)\, d\mu(g) \right\} \overline{(\gamma, g_1)}\Psi(g_1)\, d\mu(g_1)^{7} \\ &= \int_G \left\{ \int_G \overline{(\gamma, g)}\Phi(g)\, \mathrm{d}\mu(g) \right\} \overline{(\gamma, g_1)}\Psi(g_1)\, d\mu(g_1) \\ &= \int_G \overline{(\gamma, g)}\Phi(g)\, d\mu(g) \cdot \int_G \overline{(\gamma, g_1)}\Psi(g_1)\, d\mu(g_1) \\ &= \hat{\Phi}(\gamma)\hat{\Psi}(\gamma)\end{aligned} \tag{30}$$

for all γ in $\hat{G}$.

In view of the results of this section, we shall *assume from now on that G is σ-finite.*

[6] By [3, Corollary 14, p. 150].

[7] The interchange of order of integration is justified by Fubini's theorem in view of Lemma 3.9.

4. Some Banach algebras. We now introduce some Banach algebras which play an important role in our main result. To begin with, we observe that $L_1(G, \mathscr{L}(H, H))$ is a noncommutative Banach algebra over C with respect to convolution and $\| \cdot \|_1$ in view of Corollary 3.10. If G is discrete, then $L_1(G, \mathscr{L}(H, H)$ contains an identity Δ. In the case where G is not discrete, it is convenient to adjoin an identity Δ to this algebra so that $\Delta * \Delta = \Delta$ and $\Delta * \Phi = \Phi * \Delta = \Phi$.[8] We set $\|\Delta\|_1 = 1$ and we write $L_1(G, \mathscr{L}(H, H)) \oplus \{\Delta\}$ for this Banach algebra. The elements of $L_1(G, \mathscr{L}(H, H)) \oplus \{\Delta\}$ are pairs $\{\Phi, \lambda\Delta\}$ with Φ in $L_1(G, \mathscr{L}(H, H))$ and λ in C and multiplication is defined by $*$ so that

$$\{\Phi, \lambda\Delta\} * \{\Psi, \nu\Delta\} = \{\Phi * \Psi + \nu\Phi + \lambda\Psi, \lambda\nu\Delta\} \tag{31}$$

for all Φ, Ψ, λ, ν.

For the sake of exposition, we treat the case where G is not discrete. The discrete group case can be dealt with in an analogous manner and is left to the reader. However, the key modifications required will be noted. We point out that all the results of § 5 and § 6 remain true in the discrete group context (with the appropriate definitions).

Now $L_1(G, \mathscr{L}(H, H)) \oplus \{\Delta\}$ can be "realized" as an algebra of bounded linear transformations of $L_2(G, H)$ into $L_2(G, H)$. More precisely, we let B be the set of all linear transformations of $L_2(G, H)$ into $L_2(G, H)$ of the form

$$(\mathbf{\Phi} x)(g) = \int_G \Phi(g - g_1)x(g_1)\, d\mu + \lambda x(g), \tag{32}$$

where $\Phi(\cdot) \in L_1(G, \mathscr{L}(H, H))$ and $\lambda \in C$. It is clear from Lemma 3.6 that $\mathbf{\Phi}$ is an element of $\mathscr{L}(L_2(G, H), L_2(G, H))$. If $\mathbf{\Phi}$ is an element of B, then we shall often write $\mathbf{\Phi} = \Phi + \lambda\Delta$.[9] We note that B is clearly a linear space over C. We shall soon show that B can be viewed as a Banach algebra which is isomorphic to $L_1(G, \mathscr{L}(H, H)) \oplus \{\Delta\}$. (In the discrete group case, B is simply isomorphic to $L_1(G, \mathscr{L}(H, H))$.) We begin by proving the following lemma.

LEMMA 4.1. *Let $\mathbf{\Phi} = \Phi + \lambda\Delta$ be an element of B. Then $\mathbf{\Phi} = \mathbf{0}$ if and only if $\Phi(\cdot) = 0(\cdot)$ and $\lambda = 0$.*

Proof. Clearly, if $\Phi(\cdot) = 0(\cdot)$ and $\lambda = 0$, then $\mathbf{\Phi} = \mathbf{0}$. On the other hand, suppose that $\mathbf{\Phi} = \Phi + \lambda\Delta = \mathbf{0}$. Then $\Phi * x = -\lambda x$ for all x in $L_2(G, H)$, and hence $(\Phi * x)\hat{}(\gamma) = -\lambda\hat{x}(\gamma)$ for all γ in $\hat{G}$ and x in $L_2(G, H)$. Since the Fourier transform is onto,[10] we deduce that $\hat{\Phi}(\gamma)\hat{x}(\gamma) = -\lambda\hat{x}(\gamma)$ for all $\hat{x}$ in $L_2(\hat{G}, H)$. It follows that $\hat{\Phi}(\gamma) = -\lambda I$ for all γ and hence, in view of the remarks following Lemma 3.6, that $\lambda = 0$ and $\Phi(\cdot) = 0(\cdot)$.

COROLLARY 4.2. *If $\mathbf{\Phi} = \Phi + \lambda\Delta$ is an element of B, then Φ and λ uniquely represent $\mathbf{\Phi}$.*

Now, knowing that the representation $\mathbf{\Phi} = \Phi + \lambda\Delta$ is unique, we can see immediately that $\|\mathbf{\Phi}\|_B = \|\Phi(\cdot)\|_1 + |\lambda|$ defines a norm on B. Moreover, since $L_1(G, \mathscr{L}(H, H))$ and C are complete, it is clear that B is complete with respect to $\| \cdot \|_B$. Thus, B is a Banach space. We now have the following lemma.

[8] Δ may be viewed as a generalization of the δ-function.

[9] This use of Δ will be justified shortly.

[10] See remarks following Corollary 3.3.

LEMMA 4.3. *Let* $\boldsymbol{\Phi} = \Phi + \lambda\Delta$ *and* $\boldsymbol{\Psi} = \Psi + \nu\Delta$ *be elements of* B. *Then* $\boldsymbol{\Phi}\boldsymbol{\Psi} = \{\Phi * \Psi + \lambda\Psi + \nu\Phi\} + \lambda\nu\Delta$ *is an element of* B *and* $\|\boldsymbol{\Phi}\boldsymbol{\Psi}\|_B \leqq \|\boldsymbol{\Phi}\|_B\|\boldsymbol{\Psi}\|_B$.

Proof. Let x be an element of $L_2(G, H)$. Then

$$
\begin{aligned}
(\boldsymbol{\Phi}\boldsymbol{\Psi}x)(g) &= \int_G \Phi(g - g_1)(\boldsymbol{\Psi}x)(g_1)\,d\mu + \lambda(\boldsymbol{\Psi}x)(g) \\
&= \int_G \Phi(g - g_1)\left\{\int_G \Psi(g_1 - g_2)x(g_2)\,d\mu\right\} d\mu + \nu(\Phi * x)(g) \\
&\quad + \lambda(\Psi * x)(g) + \lambda\nu x(g) \\
&= \int_G \left\{\int_G \Phi(g - g_2 - g_1)\Psi(g_1)\,d\mu(g_1)\right\} x(g_2)\,d\mu + \lambda(\Psi * x)(g) \\
&\quad + \nu(\Phi * x)(g) + \lambda\nu x(g)
\end{aligned}
$$

and the lemma follows.

Thus, B is a Banach algebra. It is clear from the lemmas that the mapping T of $L_1(G, \mathscr{L}(H, H)) \oplus \{\Delta\}$ *onto* B given by

$$(33) \qquad T\{\Phi, \lambda\Delta\} = \Phi + \lambda\Delta$$

is a continuous, norm preserving, injective linear map which preserves multiplication. In other words, *B and $L_1(G, \mathscr{L}(H, H)) \oplus \{\Delta\}$ are isometrically isomorphic.*[11]

We now have the following definition.

DEFINITION 4.4. Let $\boldsymbol{\Phi} = \Phi + \lambda\Delta$ be an element of B. Then the *Fourier transform of* $\boldsymbol{\Phi}$, in symbols, $\hat{\boldsymbol{\Phi}}(\cdot)$, is the map of $\hat{G}$ into $\mathscr{L}(H, H)$ given by

$$(34) \qquad \hat{\boldsymbol{\Phi}}(\gamma) = \hat{\Phi}(\gamma) + \lambda I,$$

where I is the identity in $\mathscr{L}(H, H)$.

We note that $(\boldsymbol{\Phi}\boldsymbol{\Psi})\hat{} = \hat{\boldsymbol{\Phi}}\hat{\boldsymbol{\Psi}}$ by virtue of Proposition 3.11 and Lemma 4.3, and that, for each γ in $\hat{G}$, the Fourier transform is a continuous homomorphism of B into $\mathscr{L}(H, H)$. We also note that $\hat{\boldsymbol{\Phi}}(\cdot)$ is a uniformly continuous element of $\mathscr{C}(\hat{G}, \mathscr{L}(H, H))$ by virtue of Proposition 3.5.

Let $P \subset G$ be a closed semigroup of positive Haar measure. We then let B_P be the subset of B given by

$$(35) \qquad B_P = \{\boldsymbol{\Phi} = \Phi + \lambda\Delta \in B : \operatorname{supp}\Phi \subset P\},$$[12]

where $\operatorname{supp}\Phi = \overline{\{g : \Phi(g) \neq 0\}}$ is the support of Φ. We then have the following proposition.

PROPOSITION 4.5. *B_P is a closed subalgebra of B.*

Proof. Let $\boldsymbol{\Phi} = \Phi + \lambda\Delta$ and $\boldsymbol{\Psi} = \Psi + \nu\Delta$ be elements of B_P. We claim that $\boldsymbol{\Phi}\boldsymbol{\Psi}$ is in B_P. To verify this claim, it will, by virtue of Lemma 4.3, be enough to show that $\operatorname{supp}(\Phi * \Psi) \subset P$. Now $\Phi(g - g_1)\Psi(g_1) = 0$ unless $g_1 \in \operatorname{supp}\Phi$ and $g - g_1 \in \operatorname{supp}\Psi$, i.e., unless $g \in \operatorname{supp}\Psi + \operatorname{supp}\Phi$. But $\operatorname{supp}\Psi \subset P$ and $\operatorname{supp}\Phi \subset P$ and P is a closed semigroup together imply that

$$\operatorname{supp}(\Phi * \Psi) = \overline{\{g : (\Phi * \Psi)(g) \neq 0\}} \subset \overline{\operatorname{supp}\Psi + \operatorname{supp}\Phi} \subset P.$$

[11] If G is discrete, then B and $L_1(G, \mathscr{L}(H, H))$ are isometric.

[12] In the case G discrete, define B_P by setting $B_P = L_{1P}(G, \mathscr{L}(H, H))$ or $L_{1P}(G, \mathscr{L}(H, H)) \oplus \{\Delta\}$ according as $0 \in P$ or $0 \notin P$, where $L_{1P}(G, \mathscr{L}(H, H)) = \{\Phi \in B : \operatorname{supp}\Phi \subset P\}$.

Thus, B_P is a subalgebra of B. To show that B_P is closed, we need only show that if $\Phi_n(\cdot)$ converges to $\Phi(\cdot)$ in $\|\cdot\|_1$ and $\operatorname{supp}\Phi_n \subset P$, then $\operatorname{supp}\Phi \subset P$. But, if $\Phi_n(\cdot)$ converges to $\Phi(\cdot)$ in $\|\cdot\|_1$, then a subsequence $\Phi_{n_k}(\cdot)$ will converge to $\Phi(\cdot)$ almost everywhere. It follows that, to within a μ-null set,

$$\{g : \Phi(g) \neq 0\} \subset \bigcup \operatorname{supp}\Phi_{n_k}{}^{13} \tag{36}$$

and, hence, that $\overline{\{g : \Phi(g) \neq 0\}} \subset P$ as P is closed.

We note that the elements of B_P are causal with respect to P (cf. Example 2.4) and we call B_P the *causal subalgebra of B with respect to P*. We also note that if $\mathbf{\Phi}$ is in B_P (or B), then $\mathbf{\Phi}(\cdot)$ is a uniformly continuous element of $\mathscr{C}(\hat{G}, \mathscr{L}(H, H))$.

Let W_P be the set of all linear maps $\mathbf{Z}$ of $L_2(G, H)$ into $L_2(G, H)$ such that (i) $\mathbf{Z}$ is causal with respect to P, and (ii) $(\mathbf{Z}x)\hat{}(\gamma) = Z(\gamma)\hat{x}(\gamma)$ for all x in $L_2(G, H)$ and γ in $\hat{G}$, where $Z(\cdot)$ is a bounded uniformly continuous map of $\hat{G}$ into $\mathscr{L}(H, H)$). We immediately see that W_P is a linear space over C and also that $\|\mathbf{Z}\|_{W_P} = \sup_{\gamma\in\hat{G}} \{\|Z(\gamma)\|\}$ is a norm on W_P. We then have the following theorem.

THEOREM 4.6. *W_P is a Banach algebra.*

Proof. Let $\mathbf{Z}_1$ and $\mathbf{Z}_2$ be elements of W_P. Then, $\mathbf{Z}_1\mathbf{Z}_2$ is causal with respect to P and $(\mathbf{Z}_1\mathbf{Z}_2x)\hat{} = Z_1(\gamma)(\mathbf{Z}_2x)\hat{} = Z_1(\gamma)Z_2(\gamma)\hat{x}(\gamma)$. But $\|Z_1(\gamma)Z_2(\gamma)\| \to \|Z_1(\gamma)\| \cdot \|Z_2(\gamma)\|$, and so, $Z_1(\cdot)Z_2(\cdot)$ is a uniformly continuous element of $\mathscr{C}(\hat{G}, \mathscr{L}(H, H))$ and $\|\mathbf{Z}_1\mathbf{Z}_2\|_{W_P} \geqq \|\mathbf{Z}_1\|_{W_P}\|\mathbf{Z}_2\|_{W_P}$. Thus, to complete the proof, we need only show that W_P is complete.

Therefore let Z_n be a Cauchy sequence in W_P. Then $Z_n(\cdot)$ is a Cauchy sequence in $\mathscr{C}(\hat{G}, \mathscr{L}(H, H))$ and so has a limit $Z(\cdot)$. We claim that $Z(\cdot)$ is uniformly continuous. Let $\varepsilon > 0$ be given. Then there is an n_ε such that $n, m \geqq n_\varepsilon$ implies that $\|Z_n(\gamma) - Z_m(\gamma)\| < \varepsilon/9$ for all γ. Let γ' be any element of $\hat{G}$. Then there is an $n_{\gamma'}$ such that $n \geqq n_{\gamma'}$ implies that $\|Z_n(\gamma') - Z(\gamma')\| < \varepsilon/3$; and there is a neighborhood Γ of the identity 0 in $\hat{G}$ (with Γ independent of γ') such that if $\gamma \in \gamma' + \Gamma$, then $\|Z_{n_\varepsilon}(\gamma) - Z_{n_\varepsilon}(\gamma')\| < \varepsilon/9$. Now if $\gamma \in \gamma' + \Gamma$ and if $n \geqq \max(n_{\gamma'}, n_\gamma, n_\varepsilon)$, then

$$\begin{aligned}\|Z(\gamma) - Z(\gamma')\| &\leqq \|Z(\gamma) - Z_n(\gamma)\| + \|Z_n(\gamma) - Z_n(\gamma')\| + \|Z_n(\gamma') - Z(\gamma')\| \\ &< \varepsilon/3 + \varepsilon/3 + \varepsilon/3 = \varepsilon\end{aligned}$$

as

$$\begin{aligned}\|Z_n(\gamma) - Z_n(\gamma')\| &\leqq \|Z_n(\gamma) - Z_{n_\varepsilon}(\gamma)\| + \|Z_{n_\varepsilon}(\gamma) - Z_{n_\varepsilon}(\gamma')\| + Z_{n_\varepsilon}(\gamma') - Z_n(\gamma)\| \\ &< \varepsilon/9 + \varepsilon/9 + \varepsilon/9 = \varepsilon/3.\end{aligned}$$

Since Γ is independent of γ', $Z(\cdot)$ is uniformly continuous. We define a map $\mathbf{Z}$ of $L_2(G, H)$ into itself by setting $(\mathbf{Z}x)(g) = (Z(\cdot)\hat{x}(\cdot))\hat{}(-g)$. It is clear that $\mathbf{Z}$ is linear and that $(\mathbf{Z}x)\hat{} = Z(\cdot)\hat{x}(\cdot)$. Moreover, $\mathbf{Z}_n$ converges to $\mathbf{Z}$ in that $Z_n(\cdot)$ converges to $Z(\cdot)$. Thus, we need only show that $\mathbf{Z}$ is causal with respect to P.

Now we note that, by Proposition 3.2 (iii), $\mathbf{Z}_n y$ converges to $\mathbf{Z}y$ for all y in $L_2(G, H)$. Setting $y = x - x_{g_0}$, we see that $\mathbf{Z}_n(x - x_{g_0}) - \mathbf{Z}(x - x_{g_0})$ tends to zero in $L_2(G, H)$. Since χ_{g_0} is bounded, it follows that $\chi_{g_0}(\mathbf{Z}_n(x - x_{g_0})) - \chi_{g_0}(\mathbf{Z}(x - x_{g_0}))$

[13] Since Φ is actually an equivalence class. we can alter Φ on a μ-null set. Thus, if $\tilde{G} = \{g : \Phi_{n_k}(g) \to \Phi(g)\}$, then we replace Φ by Φ', where $\Phi'(g) = \Phi(g)$ on $\tilde{G}$ and $\Phi'(g) = 0$ on $\tilde{G}^c$. We note also that, strictly speaking, in this context, the support is an "equivalence class of sets."

goes to zero. But $\mathbf{Z}_n$ is causal and linear so that $\chi_{g_0}(\mathbf{Z}_n(x - x_{g_0})) = 0$ for all n. Thus, $(\mathbf{Z}x)_{g_0} = \chi_{g_0}(\mathbf{Z}x) = \chi_{g_0}(\mathbf{Z}x_{g_0}) = (\mathbf{Z}x_{g_0})_{g_0}$ and $\mathbf{Z}$ is causal with respect to P.

We observe that W_P has an identity $\mathbf{\Delta}$ given by $\mathbf{\Delta}x = x$ and that B_P is a subalgebra of W_P. However, B_P is not closed in $\|\cdot\|_{W_P}$. We let $\overline{B_P}$ denote the closure of B_P in W_P.

We let $L_{1P}(G, C) = \{f \in L_1(G, C) : \operatorname{supp} f \subset P\}$. Then $L_{1P}(G, C)$is a commutative Banach algebra under convolution. We adjoin an identity δ with norm 1 to $L_{1P}(G, C)$[14] and we denote this extended Banach algebra by L_P. If we set $\hat{\delta}(\gamma) = 1$ for all γ in $\hat{G}$, then the Fourier transform extends in a natural way to L_P. Since L_P is a commutative Banach algebra with identity, we can apply the Gelfand theory [6]. Thus, we let $\mathscr{M}$ denote the maximal ideal space of L_P and we denote the Gelfand representation of $f \in L_P$ by $\hat{f}(M)$. We note that if $\mathscr{M}$ is given the weakest topology such that the maps $\hat{f}(\cdot)$ are continuous, then $\mathscr{M}$ is compact [6]. We also observe that, for every γ in $\hat{G}$, there is an M in $\mathscr{M}$ such that $\hat{f}(\gamma) = \hat{f}(M)$ for all f in L_P[15] (see [6]). It follows from this observation and the formula $\sup_{M \in \mathscr{M}} |\hat{f}(M)| = \lim_{n \to \infty} \|f^{*m}\|^{1/m}$ (see [6, p. 194]) that

$$\sup_{\gamma \in \hat{G}} |\hat{f}(\gamma)| = \sup_{M \in \mathscr{M}} |\hat{f}(M)| \tag{37}$$

(see [6, pp. 194, 214] or [4, p. 264]). This relation between characters and maximal ideals will be exploited in the sequel.

We can now extend the Gelfand representation of L_P to a continuous homomorphism of B_P into $\mathscr{B}(\mathscr{M}, \mathscr{L}(H, H))$ where $\mathscr{B}(\mathscr{M}, \mathscr{L}(H, H))$ is the space of bounded maps of $\mathscr{M}$ into $\mathscr{L}(H, H)$ and H *is assumed to be separable.* If $\mathbf{\Phi} = \Phi + \lambda\Delta$ is an element of B_P, then the mapping φ of G into C given by

$$\varphi(g) = \langle \Phi(g)h_1, h_2 \rangle \tag{38}$$

for fixed h_1, h_2 in H, is an element of $L_P(G, C)$. Let $\hat{\varphi}(M)$ denote the Gelfand representation of φ so that

$$\hat{\varphi}(M) = (\langle \Phi(\cdot)h_1, h_2 \rangle)\hat{\ }(M) \tag{39}$$

for all M in $\mathscr{M}$. We now have the following lemma.

LEMMA 4.7. *Let* Φ *be an element of* $L_1(G, \mathscr{L}(H, H))$ *with* $\operatorname{supp} \Phi \subset P$ *and let* M *be a fixed element of* $\mathscr{M}$. *Then the map* $T_M(h_1, h_2)$ *given by*

$$T_M(h_1, h_2) = (\langle \Phi(\cdot)h_1, h_2 \rangle)\hat{\ }(M) \tag{40}$$

is a bounded bilinear map of $H \times H$ *into* C.

Proof. The proof is an immediate consequence of the fact that the map φ into $\hat{\varphi}$ is a homomorphism of $L_P(G, C)$ into $\mathscr{C}(\mathscr{M}, C)$ and $\sup_{M \in \mathscr{M}} |\hat{\varphi}(M)| \leqq \|\varphi(\cdot)\|$ (see [4, p. 263]).

It follows from Lemma 4.7 that for each M in $\mathscr{M}$, there is a unique element $\hat{\Phi}(M)$ of $\mathscr{L}(H, H)$ such that $T_M(h_1, h_2) = \langle \hat{\Phi}(M)h_1, h_2 \rangle$. We therefore define a map of B_P by setting $\hat{\mathbf{\Phi}}(M) = \hat{\Phi}(M) + \lambda I$ and we call the map, $\mathbf{\Phi} \to \hat{\mathbf{\Phi}}(M)$, the *extended*

[14] If G is discrete, then the modifications are along the lines of the modifications used for B_P.

[15] Here $\hat{f}(\gamma)$ is the Fourier transform while $\hat{f}(M)$ is the Gelfand representation. We are willing to accept this ambiguity because of the observation.

Gelfand representation. It is clear from standard properties of the Gelfand representation [4, p. 263] that the extended Gelfand representation is linear. We now have the following lemma.

LEMMA 4.8. *The extended Gelfand representation is a continuous homomorphism of B_P into $\mathscr{B}(\mathscr{M}, \mathscr{L}(H, H))$, where $\mathscr{B}(\mathscr{M}, \mathscr{L}(H, H))$ is the space of all bounded maps of $\mathscr{M}$ into $\mathscr{L}(H, H)$ and H is separable.*

Proof. We first observe that

$$\sup_{M \in \mathscr{M}} |T_M(h_1, h_2)| = \sup_{M \in \mathscr{M}} |\hat{\varphi}(M)| \leqq \|\Phi(\cdot)\|_1 |h_1||h_2| \tag{41}$$

for any h_1, h_2 in H and Φ in $L_1(G, \mathscr{L}(H,H))$. It follows that $\sup_{M \in \mathscr{M}} \|\hat{\Phi}(M)\| \leqq \|\Phi(\cdot)\|_1$ for Φ in $L_1(G, \mathscr{L}(H, H))$ and, hence, that the extended Gelfand representation is an element of the space $\mathscr{C}(B_P, \mathscr{B}(\mathscr{M}, \mathscr{L}(H, H)))$.

Now, by virtue of Lemma 4.3 and the fact that all the maps $\varphi \to \hat{\varphi}$ are homomorphisms of $L_{1P}(G, C)$ into $\mathscr{C}(\mathscr{M}, C)$, we need only show that $(\Phi * \Psi)\hat{}(M) = \hat{\Phi}\,(M)\hat{\Psi}(M)$ for $\Phi(\cdot)$ and $\Psi(\cdot)$ in $L_1(G, \mathscr{L}(H, H))$. Since H is separable, we let $\{e_1, \cdots\}$ be an orthonormal basis of H. Then

$$\begin{aligned} \langle(\Phi * \Psi)(g)e_i, e_j\rangle &= \int_G \langle \Phi(g - g_1)\Psi(g_1)e_i, e_j\rangle \, d\mu \\ &= \int_G \left\langle \Phi(g - g_1)\left\{\sum_k \langle \Psi(g_1)e_i, e_k\rangle e_k\right\}, e_j\right\rangle d\mu \\ &= \sum_k \int_G \langle \Phi(g - g_1)e_k, e_j\rangle\langle \Psi(g_1)e_i, e_k\rangle \, d\mu \\ &= \sum_k (\varphi_{k_j} * \psi_{ik})(g), \end{aligned} \tag{42}$$

where $\varphi_{kj}(\cdot) = \langle \Phi(\cdot)e_k, e_j\rangle$ and $\psi_{ik}(\cdot) = \langle \Psi(\cdot)e_i, e_k\rangle$.[16] It follows that

$$(\langle \Phi * \Psi\rangle)\hat{}(M)e_i, e_j\rangle = \sum_k \hat{\varphi}_{kj}(M)\hat{\psi}_{ik}(M) \tag{43}$$

for all M in $\mathscr{M}$. Now we observe that

$$\begin{aligned} \langle \hat{\Phi}(M)\hat{\Psi}(M)e_i, e_j\rangle &= \left\langle \hat{\Phi}(M)\left\{\sum_k \langle \hat{\Psi}(M)e_i, e_k\rangle e_k\right\}, e_j\right\rangle \\ &= \sum_k \langle \hat{\Phi}(M)e_k, e_j\rangle\langle \hat{\Psi}(M)e_i, e_k\rangle \\ &= \sum_k \hat{\varphi}_{kj}(M)\hat{\psi}_{ik}(M) \end{aligned} \tag{44}$$

for all M in $\mathscr{M}$. Since (43) and (44) hold for all e_i and e_j, $(\Phi * \Psi)\hat{}(M) = \hat{\Phi}(M)\hat{\Psi}(M)$ and the lemma is established.

[16] The interchange of summation and integration is justified by the Lebesgue dominated convergence theorem [3].

We note that, in view of (37), we have

$$\tag{45}\begin{aligned}\sup_{\gamma\in\hat{G}}\|\hat{\Phi}(\gamma)\| &= \sup_{\gamma\in\hat{G}}\sup_{\substack{|\xi|=1\\|\eta|=1}}|\langle\hat{\Phi}(\gamma)\xi,\eta\rangle|\\ &= \sup_{\substack{|\xi|=1\\|\eta|=1}}\sup_{\gamma\in\hat{G}}|\langle\hat{\Phi}(\gamma)\xi,\eta\rangle|\\ &= \sup_{\substack{|\xi|=1\\|\eta|=1}}\sup_{M\in\mathscr{M}}|\langle\hat{\Phi}(M)\xi,\eta\rangle|\\ &= \sup_{M\in\mathscr{M}}\sup_{\substack{|\xi|=1\\|\eta|=1}}|\langle\hat{\Phi}(M)\xi,\eta\rangle| = \sup_{M\in\mathscr{M}}\|\hat{\Phi}(M)\|.\end{aligned}$$

We shall use this observation in the next section.

5. The main result. We shall prove our main result in this section. We let $P \subset G$ be a closed semigroup of positive Haar measure and we suppose that the space H is separable. We begin with some definitions.

DEFINITION 5.1. Let $\mathbf{\Phi}$ be an element of B_P. Then the *spectrum of* $\mathbf{\Phi}$, in symbols, $\Sigma_{B_P}(\mathbf{\Phi})$, is the subset of C given by

$$\tag{46}\Sigma_{B_P}(\mathbf{\Phi}) = \{\lambda:\mathbf{\Phi} - \lambda\Delta \text{ does not have an inverse in } B_P\}.$$

Similarly, if $\mathbf{Z}$ is an element of W_P, then the *spectrum of* $\mathbf{Z}$, in symbols, $\Sigma_{W_P}(\mathbf{Z})$, is the subset of C given by

$$\tag{47}\Sigma_{W_P}(\mathbf{Z}) = \{\lambda:\mathbf{Z} - \lambda\Delta \text{ does not have an inverse in } W_P\}.$$

DEFINITION 5.2. Let $\mathbf{\Phi}$ be an element of B_P and let M be an element of $\mathscr{M}$. Then $\hat{\mathbf{\Phi}}(M)$ is an element of $\mathscr{L}(H, H)$ and the *spectrum of* $\hat{\mathbf{\Phi}}(M)$, in symbols, $\sigma(\hat{\mathbf{\Phi}}(M))$, is the subset of C given by

$$\tag{48}\sigma(\hat{\mathbf{\Phi}}(M)) = \{\lambda:\hat{\mathbf{\Phi}}(M) - \lambda I \text{ does not have an inverse in } \mathscr{L}(H, H)\}.$$

We observe that $\Sigma_{B_P}(\mathbf{\Phi})$ and $\Sigma_{W_P}(\mathbf{Z})$ are the usual Banach algebra notions of spectrum and that $\sigma(\hat{\mathbf{\Phi}}(M))$ is the usual operator notion of spectrum. We now have the following proposition.

PROPOSITION 5.3. *If* $\mathbf{\Phi}$ *is an element of* B_P, *then* $\bigcup_{M\in\mathscr{M}}\sigma(\hat{\mathbf{\Phi}}(M)) \subset \Sigma_{B_P}(\mathbf{\Phi})$.

Proof. Suppose that $\lambda \notin \Sigma_{B_P}(\mathbf{\Phi})$. Then there is a $\mathbf{\Psi}$ in B_P such that

$$\tag{49}(\mathbf{\Phi} - \lambda\Delta) * \mathbf{\Psi} = \mathbf{\Psi} * (\mathbf{\Phi} - \lambda\Delta) = \Delta.$$

Since the extended Gelfand representation is a homomorphism, it follows that

$$\tag{50}(\hat{\mathbf{\Phi}}(M) - \lambda I)\hat{\mathbf{\Psi}}(M) = \hat{\mathbf{\Psi}}(M)(\hat{\mathbf{\Phi}}(M) - \lambda I) = I$$

for all M in $\mathscr{M}$. Thus, $\lambda \notin \bigcup_{M\in\mathscr{M}}\sigma(\hat{\mathbf{\Phi}}(M))$ and the proposition is established.

We are now ready to develop our main result. We start with the following lemma.

LEMMA 5.4. *If* H *is finite-dimensional and if* $\mathbf{\Phi}$ *is an element of* B_P, *then* $\bigcup_{M\in\mathscr{M}}\sigma(\hat{\mathbf{\Phi}}(M)) = \Sigma_{B_P}(\mathbf{\Phi})$.

Proof. We must show that $\Sigma_{B_P}(\mathbf{\Phi}) \subset \bigcup_{M\in\mathscr{M}}\sigma(\hat{\mathbf{\Phi}}(M))$ which amounts to showing that if $\hat{\mathbf{\Phi}}(M)$ is invertible for every M in $\mathscr{M}$, then $\mathbf{\Phi}$ has an inverse in B_P. Therefore, let $\{e_1, \cdots, e_n\}$ be an orthonormal basis of H and let $[\mathbf{\Phi}] = [\varphi_{ij}(\cdot)]$ be the matrix of $\mathbf{\Phi}$, i.e., $\varphi_{ij}(\cdot) = \langle\mathbf{\Phi}(\cdot)e_j, e_i\rangle$. We let $*\det[\mathbf{\Phi}]$ be the element of L_P

given by

$$(51)\qquad {}^*\det[\Phi] = \sum_{\theta} (\operatorname{sgn}\theta)\varphi_{1\theta(1)} * \varphi_{2\theta(2)} * \cdots * \varphi_{n\theta(n)},$$

where the summation extends over all elements θ of the symmetric group of order $n!$ and $\operatorname{sgn}\theta$ is the sign of the permutation θ. Similarly, we let ${}^*\operatorname{cof}[\Phi]_{ij}$ be the element of L_P given by

$$(52)\qquad {}^*\operatorname{cof}[\Phi]_{ij} = (-1)^{i+j}\,{}^*\det[\Phi]_{ij},$$

where $[\Phi]_{ij}$ is the $(n-1)\times(n-1)$ matrix obtained from $[\Phi]$ by deleting the ith row and jth column. We observe that, for each M in $\mathscr{M}$,

$$(53)\qquad \begin{aligned} ({}^*\det[\Phi])\hat{}(M) &= \det[\hat{\Phi}(M)], \\ ({}^*\operatorname{cof}[\Phi]_{ij})\hat{}(M) &= \operatorname{cof}[\hat{\Phi}(M)]_{ij}, \end{aligned}$$

where $\det[\,\cdot\,]$ and $\operatorname{cof}[\,\cdot\,]_{ij}$ are the standard determinant and cofactor for the matrix $[\hat{\Phi}(M)] = [\hat{\varphi}_{ij}(M)]$. Now if $\hat{\Phi}(M)$ is invertible, then $\det[\hat{\Phi}(M)] \neq 0$, and so it follows that if $\hat{\Phi}(M)$ is invertible for *every* M in $\mathscr{M}$, then ${}^*\det[\Phi]$ lies in *no* maximal ideal and is, therefore, invertible in L_P. Let $\Psi(\cdot)$ be the element of B_P whose matrix $[\Psi]$ is given by $[\Psi] = [\psi_{ij}(\cdot)]$, where

$$(54)\qquad \psi_{ij}(\cdot) = ({}^*\det[\Phi])^{-1} * ({}^*\operatorname{cof}[\Phi]_{ij}).$$

It is clear that Ψ is an element of B_P and a simple calculation shows that Ψ is a two-sided inverse of Φ with respect to the multiplication in B_P. Thus the lemma is established.

We now introduce the notion of approximability for elements of B_P.

DEFINITION 5.5. Let Φ be an element of B_P and let $\{e_1, \cdots\}$ be an orthonormal basis of H. Let H_n be the span of $\{e_1, \cdots, e_n\}$ and let E_n be the projection of H onto H_n. Let $\Phi_n = E_n\Phi E_n$. We call Φ *approximable* if $\hat{\Phi}_n(M)$ converges to $\hat{\Phi}(M)$ uniformly on $\mathscr{M}$.

Approximability is an intrinsic notion in view of the following proposition.

PROPOSITION 5.6. *An element Φ of B_P is approximable if and only if each $\hat{\Phi}(M)$ is a completely continuous element of $\mathscr{L}(H, H)$ and the map $M \to \hat{\Phi}(M)$ is continuous on $\mathscr{M}$.*

Proof. If Φ is approximable, then the result is an immediate consequence of that fact that each $\hat{\Phi}_n(M)$ is in $\mathscr{C}(\mathscr{M}, \mathscr{L}(H, H))$ (as $\hat{\Phi}_n(M)$ is finite-dimensional) and the well-known properties of operators [7, p. 204]. On the other hand, if each $\hat{\Phi}(M)$ is completely continuous and if $\hat{\Phi}(M)$ is in $\mathscr{C}(\mathscr{M}, \mathscr{L}(H, H))$, then $\hat{\Phi}_n(M)$ converges to $\hat{\Phi}(M)$ for each M [7, p. 204]. But the $\hat{\Phi}_n(M)$ form an equicontinuous family on the compact set $\mathscr{M}$, and hence, $\hat{\Phi}_n(M)$ converges to $\hat{\Phi}(M)$ uniformly on $\mathscr{M}$.

We are now ready for our main result.

THEOREM 5.7. *If Φ is an approximable element of B_P, then*

$$(55)\qquad \Sigma_{W_P}(\Phi) \subset \bigcup_{M\in\mathscr{M}} \sigma(\hat{\Phi}(M)) \subset \Sigma_{B_P}(\Phi).$$

Proof. In view of Proposition 5.3 and the fact that $B_P \subset W_P$, we need only show that $\Sigma_{W_P}(\Phi)$ is contained in $\bigcup_{M\in\mathscr{M}} \sigma(\hat{\Phi}(M))$. So let us suppose that

$\lambda \notin \bigcup_{M \in \mathscr{M}} \sigma(\hat{\Phi}(M))$. Since $\hat{\Phi}(M)$ is completely continuous for each M in $\mathscr{M}$ by Proposition 5.6, $0 \in \sigma(\hat{\Phi}(M))$ for all M, and hence, $\lambda \neq 0$. Moreover, $\hat{\Phi}(M) - \lambda I$ is invertible for every M.

Now let M_0 be a given element of $\mathscr{M}$. Then there is an open subset U_{M_0} of $\mathscr{M}$ with $M_0 \in U_{M_0}$ such that

$$\text{(56)} \qquad \|\hat{\Phi}(M) - \hat{\Phi}(M_0)\| \leqq \tfrac{1}{2}\|(\lambda I - \hat{\Phi}(M_0))^{-1}\|^{-1}$$

for all M in U_{M_0} (as $\hat{\Phi}$ is continuous on $\mathscr{M}$). It follows [3, p. 585] that

$$\text{(57)} \quad (\lambda I - \hat{\Phi}(M))^{-1} = (\lambda I - \hat{\Phi}(M_0))^{-1} \sum_{n=0}^{\infty} [(\hat{\Phi}(M) - \hat{\Phi}(M_0))(\lambda I - \hat{\Phi}(M_0))^{-1}]^n$$

for all M in U_{M_0} and hence, that

$$\text{(58)} \qquad \begin{aligned} \|(\lambda I - \hat{\Phi}(M))^{-1}\| &\leqq \|(\lambda I - \hat{\Phi}(M_0))^{-1}\| \sum_0^{\infty} \frac{1}{2^n} \\ &\leqq 2\|(\lambda I - \hat{\Phi}(M_0))^{-1}\| \end{aligned}$$

for all M in U_{M_0}. Since the collection $\{U_{M_0} : M_0 \in \mathscr{M}\}$ forms an open covering of $\mathscr{M}$ and since $\mathscr{M}$ is compact, a *finite* collection $\{U_{M_1}, \cdots, U_{M_k}\}$ covers $\mathscr{M}$. It follows that if we set $K = 2 \max_{i=1,\cdots,k} \{\|(\lambda I - \hat{\Phi}(M_i))^{-1}\|\}$, then

$$\text{(59)} \qquad \|(\lambda I - \hat{\Phi}(M))^{-1}\| \leqq K$$

for all M in $\mathscr{M}$.

Now let $\{e_1, \cdots\}$ be an orthonormal basis of H, H_n be the span of $\{e_1, \cdots, e_n\}$, and E_n be the projection of H on H_n as in Definition 5.5. Since $\hat{\Phi}$ is approximable, we have, for n sufficiently large,

$$\text{(60)} \qquad \|\hat{\Phi}_n(M) - \hat{\Phi}(M)\| < 1/(2K)$$

for all M in $\mathscr{M}$ (where $\mathbf{\Phi}_n = E_n \mathbf{\Phi} E_n$). It follows that $(\hat{\Phi}_n(M) - \lambda I)^{-1}$ exists and that

$$\text{(61)} \qquad \|(\hat{\Phi}_n(M) - \lambda I)^{-1}\| \leqq 2K$$

for sufficiently large n and all M in $\mathscr{M}$.

Now $H = H_n \oplus H_n'$, where H_n' is the span of $\{e_{n+1}, e_{n+2}, \cdots\}$. We observe that both H_n and H_n' are invariant under $\mathbf{\Phi}_n$ in the sense that, for every g, $\mathbf{\Phi}_n(g) = \Phi_n(g) + 0\Delta_n$, where $\Phi_n(g)$ is an element of $\mathscr{L}(H, H)$ for which both H_n and H_n' are invariant. It follows that $\mathbf{\Phi}_n - \lambda\Delta$ has the representation

$$\text{(62)} \qquad \mathbf{\Phi}_n - \lambda\Delta = (\mathbf{\Phi}_n' - \lambda\Delta_n) \oplus (-\lambda\Delta_n'),$$

where $\mathbf{\Phi}_n' - \lambda\Delta_n$ is an element of $B_{P,n} \approx L_{1P}(G, \mathscr{L}(H_n, H_n)) \oplus \{\Delta_n\}$[17] and $\mathbf{\Phi}_n'(g) = E_n\mathbf{\Phi}(g)$ on H_n. In view of (62), we see that

$$\text{(63)} \qquad \hat{\Phi}_n(M) - \lambda I = (\hat{\Phi}_n'(M) - \lambda I_n) \oplus (-\lambda I_n')$$

for all M in $\mathscr{M}$ (where I_n is the identity on H_n and I_n' is the identity on H_n'). Thus, for n sufficiently large,

$$\text{(64)} \qquad (\hat{\Phi}_n(M) - \lambda I)^{-1} = (\hat{\Phi}_n'(M) - \lambda I_n)^{-1} \oplus (-1/\lambda)I_n'$$

[17] Or $B_{P,n} \approx L_{1P}(G, \mathscr{L}(H_n, H_n))$ if G discrete and $0 \in P$.

and

$$\|(\hat{\mathbf{\Phi}}_n'(M) - \lambda I_n)^{-1}\| \leqq 2K \tag{65}$$

for all M in $\mathscr{M}$ by virtue of (61). Now (65) implies that $\lambda \notin \bigcup_{M\in\mathscr{M}} \sigma(\hat{\mathbf{\Phi}}_n'(M))$ for n sufficiently large. But, $\mathbf{\Phi}_n'$ is an element of $B_{P,n}$ and H_n is finite-dimensional. Thus, it follows from Lemma 5.4 that

$$\lambda \notin \Sigma_{B_{P,n}}(\mathbf{\Phi}_n') \tag{66}$$

for n large. In other words, $\mathbf{\Phi}_n' - \lambda\Delta_n$ is *invertible in* $B_{P,n}$.

Since $(\mathbf{\Phi}_n' - \lambda\Delta_n)^{-1} + (\lambda^{-1}\Delta_n)$ is in $L_{1P}(G, \mathscr{L}(H_n, H_n))$, we let $\mathbf{\Psi}_n$ be the element of B_P given by

$$\mathbf{\Psi}_n = (\mathbf{\Phi}_n' - \lambda\Delta_n)^{-1} \oplus (-\lambda^{-1}\Delta_n') \tag{67}$$

for n large. Then $\mathbf{\Psi}_n$ is the inverse of $\mathbf{\Phi}_n - \lambda\Delta$ so that $\mathbf{\Phi}_n - \lambda\Delta$ is invertible in B_P. Moreover,

$$\sup_{M\in\mathscr{M}} \|\hat{\mathbf{\Psi}}_n(M)\| \leqq \max\,(2K, |\lambda|^{-1}) = \delta \tag{68}$$

in view of (65). Now $\mathbf{\Psi}_n$ is an element of W_P (as $B_P \subset W_P$) and

$$(\mathbf{\Psi}_n x)\hat{}(\gamma) = \psi_n(\gamma)\hat{x}(\gamma) \tag{69}$$

for γ in $\hat{G}$ and x in $L_2(G, H)$. But (45) and (68) together imply that

$$\sup_{\gamma\in\hat{G}} \|\hat{\psi}_n(\gamma)\| = \sup_{M\in\mathscr{M}} \|\hat{\mathbf{\Psi}}_n(M)\| \geqq \delta, \tag{70}$$

and hence, that $\|\mathbf{\Psi}_n\|_{W_P} \leqq \delta$ for n large.

We claim that $\mathbf{\Psi}_n$ is a Cauchy sequence in W_P. Suppose, for the moment, that this claim is valid. Then $\mathbf{\Psi}_n$ converges to an element $\mathbf{\Psi}$ of W_P and $\mathbf{\Psi}$ is an inverse of $\mathbf{\Phi} - \lambda\Delta$ in W_P since

$$\begin{aligned} \mathbf{\Psi}(\mathbf{\Phi} - \lambda\Delta) &= (\mathbf{\Psi} - \mathbf{\Psi}_n)(\mathbf{\Phi} - \lambda\Delta) + \mathbf{\Psi}_n([\mathbf{\Phi} - \mathbf{\Phi}_n]) + \mathbf{\Psi}_n(\mathbf{\Phi}_n - \lambda\Delta) \\ &= (\mathbf{\Psi} - \mathbf{\Psi}_n)(\mathbf{\Phi} - \lambda\Delta) + \mathbf{\Psi}_n([\mathbf{\Phi} - \mathbf{\Phi}_n]) + \Delta \end{aligned} \tag{71}$$

for all large n.[18] Thus we now show that $\mathbf{\Psi}_n$ is a Cauchy sequence in W_P. We have

$$\begin{aligned} \sup_{\gamma\in\hat{G}} \|\hat{\psi}_n(\gamma) - \hat{\psi}_m(\gamma)\| &= \sup_{M\in\mathscr{M}} \|\hat{\mathbf{\Psi}}_n(M) - \hat{\mathbf{\Psi}}_m(M)\| \\ &\leqq \sup_{M\in\mathscr{M}} \|(\hat{\mathbf{\Phi}}_n(M) - \lambda I)^{-1} - (\hat{\mathbf{\Phi}}_m(M) - \lambda I)^{-1}\| \\ &\leqq \sup_{M\in\mathscr{M}} \|\hat{\mathbf{\Phi}}_n(M) - \hat{\mathbf{\Phi}}_m(M)\|\,\|\hat{\mathbf{\Psi}}_n(M)\|\,\|\hat{\mathbf{\Psi}}_m(M)\| \\ &\leqq \sup_{M\in\mathscr{M}} \|\hat{\mathbf{\Phi}}_n(M) - \hat{\mathbf{\Phi}}_m(M)\|\delta^2 \end{aligned} \tag{72}$$

by virtue of (70). Since $\mathbf{\Phi}$ is approximable, the sequence $\hat{\mathbf{\Phi}}_n(M)$ is uniformly Cauchy on $\mathscr{M}$, and thus $\mathbf{\Psi}_n$ is a Cauchy sequence in W_P by virtue of (72). The theorem is now established.

[18] Note that $\|\mathbf{\Psi}_n(\mathbf{\Phi}_n - \mathbf{\Phi})\|_{W_P} \leqq \|\mathbf{\Psi}_n\|_{W_P}\|\mathbf{\Phi}_n - \mathbf{\Phi}\|_{W_P} \leqq \delta\|\mathbf{\Phi}_n - \mathbf{\Phi}\|_{W_P}$ and that $\|\mathbf{\Phi}_n - \mathbf{\Phi}\|_{W_P} = \sup_{\gamma\in\hat{G}} \|\hat{\mathbf{\Phi}}_n(\gamma) - \hat{\mathbf{\Phi}}(\gamma)\| = \sup_{M\in\mathscr{M}} \|\hat{\mathbf{\Phi}}_n(M) - \hat{\mathbf{\Phi}}(M)\|$. Since $\hat{\mathbf{\Phi}}$ is approximable, $\|\mathbf{\Phi}_n - \mathbf{\Phi}\|_{W_P} \to 0$ as $n \to \infty$ in view of Proposition 5.6.

We can extend this theorem to $\bar{B}_P$ (the closure of B_P in W_P). To do this we first observe that if $\mathbf{Z}$ is an element of $\bar{B}_P$ with $\mathbf{Z} = \lim_{n\to\infty} \mathbf{\Phi}^n$, $\mathbf{\Phi}^n \in B_P$, then $(\mathbf{Z}x)\hat{}(\gamma) = Z(\gamma)\hat{x}(\gamma)$, $(\mathbf{\Phi}^n x)\hat{}(\gamma) = \varphi^n(\gamma)\hat{x}(\gamma)$ and $\varphi^n(\gamma)$ converges uniformly to $Z(\gamma)$ on $\hat{G}$. It follows from (45) that $\hat{\mathbf{\Phi}}^n(M)$ is uniformly Cauchy on $\mathscr{M}$ and hence has a limit which we denote by $\hat{\mathbf{Z}}(M)$. In other words, we can extend the Gelfand representation to $\bar{B}_P$ and this extension is a homomorphism of $\bar{B}_P$ into $\mathscr{B}(\mathscr{M}, \mathscr{L}(H, H))$. Similarly, we can define the notion of approximability for elements of $\bar{B}_P$ in exact analogy with Definition 5.5. We note that if $\mathbf{Z}$ is an approximable element of $\bar{B}_P$, we can then assume that $\mathbf{Z} = \lim_{m\to\infty} E_m\mathbf{\Phi}^m E_m$, where the $\mathbf{\Phi}^m$ are in B_P.[19] We then have the following corollary.

COROLLARY 5.8. *If $\mathbf{Z}$ is an approximable element of $\bar{B}_P$, then*

$$\Sigma_{\bar{B}_P}(\mathbf{Z}) = \bigcup_{M\in\mathscr{M}} \sigma(\hat{\mathbf{Z}}(M)). \tag{73}$$

Proof. A direct extension of Proposition 5.3 leads to the inclusion

$$\bigcup_{M\in\mathscr{M}} \sigma(\hat{\mathbf{Z}}(M)) \subset \Sigma_{\bar{B}_P}(\mathbf{Z}).$$

The other inclusion follows from the theorem and the fact that Z is a limit of approximable elements of B_P.

6. Concluding comments. We now make some concluding remarks and present some illustrative examples.

We first note that if $\mathbf{\Phi}$ is an approximable element of B_P, then we can define "analytic functions" of $\mathbf{\Phi}$ as elements of W_P. More precisely, if $f(\xi)$ is any complex-valued function which is analytic on a subdomain of C containing $\bigcup_{M\in\mathscr{M}} \sigma(\hat{\mathbf{\Phi}}(M))$ in its interior, then

$$f(\mathbf{\Phi}) = \frac{1}{2\pi i}\oint_\Gamma f(\xi)(\xi\Delta - \mathbf{\Phi})^{-1}\, d\xi \tag{74}$$

represents an element of W_P for any rectifiable Jordan curve Γ with $\bigcup_{M\in\mathscr{M}}\sigma(\hat{\mathbf{\Phi}}(M))$ inside Γ [6, p. 203]. Moreover, $f(\mathbf{\Phi})$ is independent of Γ (see [6]). Similarly, $f(\hat{\mathbf{\Phi}}(\gamma))$ is defined by an integral of the form

$$f(\hat{\mathbf{\Phi}}(\gamma)) = \frac{1}{2\pi i}\oint_\Gamma f(\xi)(\xi I - \hat{\mathbf{\Phi}}(\gamma))^{-1}\, d\xi \tag{75}$$

for γ in $\hat{G}$. We then have the following proposition.

PROPOSITION 6.1. *If $\mathbf{\Phi}$ is an approximable element of B_P and $f(\xi)$ is analytic on a domain containing $\bigcup_{M\in\mathscr{M}} \sigma(\hat{\mathbf{\Phi}}(M))$ in its interior, then*

$$(f(\mathbf{\Phi})x)\hat{}(\gamma) = f(\hat{\mathbf{\Phi}}(\gamma))\hat{x}(\gamma) \tag{76}$$

for all γ in $\hat{G}$ and x in $L_2(G, H)$.

[19] This follows from the fact that $\mathbf{Z} = \lim_{m\to\infty}E_m Z E_m$ and that $\mathbf{Z} = \lim_{n\to\infty} \mathbf{\Phi}^n$ with $\mathbf{\Phi}^n$ in B_P so that $\lim_{n\to\infty} E_m\mathbf{\Phi}^n E_m = E_m Z E_m$.

Proof. In view of the definition of $f(\Phi)$, we have

$$(77)\qquad \begin{aligned}(f(\Phi)x)\hat{}(\gamma) &= \frac{1}{2\pi i}\int_\Gamma f(\xi)[(\xi\Delta - \Phi)^{-1}x]\hat{}(\gamma)\,d\xi \\ &= \frac{1}{2\pi i}\oint_\Gamma f(\xi)(\xi I - \hat{\Phi}(\gamma))^{-1}\hat{x}(\gamma)\,d\xi\end{aligned}$$

by Proposition 3.8 and the fact that B is isometric to $L_1(G, \mathscr{L}(H, H)) \oplus \{\Delta\}$. The result will then follow from (71) provided that $\bigcup_{\gamma\in\hat{G}} \sigma(\hat{\Phi}(\gamma))$ is inside Γ. However, $\bigcup_{\gamma\in\hat{G}} \sigma(\hat{\Phi}(\gamma)) \subset \bigcup_{M\in\mathscr{M}} \sigma(\hat{\Phi}(M))$,[20] and so the proposition is established. We use this proposition and analytic functions of Φ extensively in [1].

We further note here that the results developed will be used in the derivation of stability theorems for partial differential equations, integro-differential equations, and difference equations [1]. We also observe that the transform theory can be applied to problems of existence as well as stability. We have the following illustrative example.

Example 6.2. Consider the integro-difference equation

$$(78)\qquad P_{k+n}(t) + (a_1 * P_{k+n-1})(t) + \cdots + (a_k * P_n)(t) = 0$$

with $a_1, \cdots, a_k$ in $L_1(R, R)$. This equation may be written in the equivalent vector form

$$(79)\qquad x_{n+1}(t) = (A * x_n)(t),$$

where A is the $k \times k$ matrix given by

$$(80)\qquad A = \begin{bmatrix} -a_1 & -a_2 & \cdots & -a_{k-1} & -a_k \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}.$$

If $x_0(t)$ is given, then the solution of (79) has the form $x_{n+1}(t) = (A * \cdots * A * x_0)(t)$, and thus we wish to determine A^{*n+1}. This can be done by transforming to obtain the equation $\hat{x}_{n+1} = \hat{A}^{n+1}\hat{x}_0$ and making use of Proposition 6.1.

Further results involving the notion of positivity and such things as Bochner's theorem can also be obtained. Some of this is done in [8].

Some typical situations in which the theory applies are illustrated in Table 1. In the table, Z is the integers, $Z_+ = \{0, 1, 2, \cdots\}$, $L_2[0, 1]$ is the set of square integrable functions on $[0, 1]$ and l_2 is the space of square summable sequences. We shall not detail the theory in each of these situations here; however, we do present the following examples.

[20] This follows from the fact that, for every γ in $\hat{G}$, there is an M in $\mathscr{M}$ such that $\hat{q}(\gamma) = \hat{q}(M)$ for *all* q in L_P (see [6]).

TABLE 1

Group G	Semigroup P	Hilbert space H
R	$[0, \infty)$	$L_2[0, 1]$
R	$[0, \infty)$	l_2
R	$[0, \infty)$	C_n
Z	Z_+	$L_2[0, 1]$
Z	Z_+	l_2
Z	Z_+	C_n
$R \times R$	$[0, \infty) \times [0, \infty)$	$L_2[0, 1]$
$R \times R$	$[0, \infty) \times [0, \infty)$	l_2
$R \times R$	$[0, \infty) \times [0, \infty)$	C_n
$R \times R$	$[0, \infty) \times R$	$L_2[0, 1]$
$R \times R$	$[0, \infty) \times R$	l_2
$R \times R$	$[0, \infty) \times R$	C_n

Example 6.3. Let $G = Z$, $P = Z_+$ and $H = C_n$. Then $\hat{G}$ is the circle group, i.e., $\hat{G} = \{e^{i\theta}: 0 \leqq \theta < 2\pi\}$ under multiplication. We observe that $\mathscr{L}(H, H)$ is simply the set of $n \times n$ matrices with complex entries and that

$$L_2(G, H) = \left\{\mathbf{x}(k): \sum_{-\infty}^{\infty} \langle \mathbf{x}(k), \mathbf{x}(k)\rangle \text{ is finite}\right\},$$

$$L_2(\hat{G}, H) = \left\{\hat{\mathbf{y}}(e^{i\theta}): \int_0^{2\pi} \langle \hat{\mathbf{y}}(e^{i\theta}), \hat{\mathbf{y}}(e^{i\theta})\rangle \, d\theta < \infty\right\}$$

in our case. If $\mathbf{x}(\cdot)$ is an element of $L_2(G, H)$, then its Fourier transform $\hat{\mathbf{x}}(e^{i\theta})$ is given by $\hat{\mathbf{x}}(e^{i\theta}) = \sum_{-\infty}^{\infty} \mathbf{x}(k)e^{ik\theta}$. We note also that $\Phi(\cdot)$ is an element of $L_{1P}(G, \mathscr{L}(H, H))$ if and only if $\Sigma_k \|\Phi(k)\| < \infty$ and $\Phi(k) = 0$ for $k < 0$. To determine the Gelfand transform of $\Phi(\cdot)$, we first remark that the maximal ideal space of $L_{1P}(G, C)$ is in one-to-one correspondence with the unit disk in C. Thus, the Gelfand representation corresponds to the Z-transform, and so, if Z is an element of the unit disk, then $\hat{\Phi}(Z)$ is the element of $\mathscr{L}(H, H)$ given by $\sum_0^{\infty} \Phi(k)Z^k$. Noting that every element of $L_{1P}(G, \mathscr{L}(H, H))$ is approximable in our present case, we see that a particular consequence of our main theorem is the following: $\Phi \in B_P$ is invertible if and only if $\det(\hat{\Phi}(Z)) \neq 0$ for all Z with $|Z| \leqq 1$.

Example 6.4. Let $G = R$, $P = [0, \infty)$ and $H = L_2[0, 1]$. Then $\hat{G} = R$ and $L_2(G, H)$ can be identified with $L_2(\hat{G}, H)$. Moreover, we can view $L_2(G, H)$ as the set

$$\left\{x(t, \alpha): \int_{-\infty}^{\infty} \left\{\int_0^1 |x(t, \alpha)|^2 \, d\alpha\right\} dt < \infty\right\},$$

i.e., as $L_2(R \times [0, 1])$. If $x(t, \alpha)$ is an element of $L_2(G, H)$, then its Fourier transform $\hat{x}(\omega, \alpha)$ is given by $\hat{x}(\omega, \alpha) = \int_{-\infty}^{\infty} e^{-i\omega t}x(t, \alpha)\, dt$. A typical approximable element Φ of B_P is given by

$$(\Phi x)(t, \alpha) = \int_{-\infty}^{\infty} \int_0^1 \varphi(t - \tau, \alpha, \beta)x(\tau, \beta)\, d\beta \, d\tau, \tag{81}$$

where the function $\varphi(t, \alpha, \beta)$ satisfies the following conditions:

(i) $\varphi(t, \alpha, \beta) = 0$ if $t < 0$;

(ii) $I(\alpha, \beta) = \int_0^\infty |\varphi(t, \alpha, \beta)|\,dt < \infty$ for almost all (α, β); and

(iii) $$\int_0^1 \int_0^1 |I(\alpha, \beta)|^2\,d\alpha\,d\beta < \infty, \qquad \int_0^\infty \left[\int_0^1 \int_0^1 |\varphi(t, \alpha, \beta)|^2\,d\alpha\,d\beta\right]^{1/2} dt < \infty.$$

To be somewhat more precise, we define $\boldsymbol{\Phi}(t)$ by setting $\boldsymbol{\Phi}(t)x = (\Phi x)(t, \alpha)$ so that $\boldsymbol{\Phi}(t)$ is an element of $B_P \approx L_{1P}(G, \mathscr{L}(H, H)) \oplus \{\Delta\}$. With a view toward determining the Gelfand representation of $\boldsymbol{\Phi}$, we note that $L_{1P}(G, C) \oplus \{\delta\}$ has, in our case, a maximal ideal space which corresponds to the right half-plane "compactified." Thus, the Gelfand representation corresponds to the Laplace transform and so, if Re $\{s\} \geqq 0$ and $\boldsymbol{\Phi}(\,\cdot\,)$ is defined via (81), then$\hat{\boldsymbol{\Phi}}(s)$ is the element of $\mathscr{L}(H, H)$ given by

$$\hat{\boldsymbol{\Phi}}(s) = \int_0^\infty e^{-st}\boldsymbol{\Phi}(t)\,dt.$$

A particular consequence of our main theorem is the following: $\boldsymbol{\Phi} = \Phi + \lambda\Delta$ is invertible in W_P (or $\bar{B}_P$) but not necessarily in B_P, provided that $\hat{\boldsymbol{\Phi}}(s) = \hat{\Phi}(s) + \lambda I$ is invertible for every s with Re $\{s\} \geqq 0$ (including ∞).

REFERENCES

[1] M. I. Freedman, P. L. Falb and G. Zames, *A Hilbert space stability theory over locally compact Abelian groups*, this Journal, 7 (1969), pp. 479–495.

[2] L. Loomis, *An Introduction to Abstract Harmonic Analysis*, Van Nostrand, Princeton, 1953.

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

[4] W. Rudin, *Fourier Analysis on Groups*, Interscience, New York, 1962.

[5] G. Zames and P. L. Falb, *Stability conditions for systems with monotone and slope-restricted nonlinearities*, this Journal, 6 (1968), pp. 89–108.

[6] M. Naimark, *Normed Rings*, P. Noordhoff, Groningen, the Netherlands, 1959.

[7] F. Riesz and B. Sz.-Nagy, *Functional Analysis*, Frederick Ungar, New York, 1955.

[8] P. L. Falb, *On a theorem of Bochner*, Inst. Hautes Études Sci. Publ. Math., 1969.

SIAM J. CONTROL
Vol. 7, No. 3, August 1969

A NOTE ON CAUSALITY AND ANALYTICITY*

M. I. FREEDMAN†, P. L. FALB‡ AND J. ANTON

1. Introduction. A very thorough study of the relationship between causality and analyticity in a group context appears in [1]. In this short note, we examine a much more specific problem, namely: the characterization of a certain subalgebra $\bar{B}_{[0,\infty)}(H)$ of the algebra of causal maps of $L_2(R, H)$ into $L_2(R, H)$, where H is a separable complex Hilbert space (see § 4). We combine the notions of analyticity and complete continuity to obtain the characterization. The algebra $\bar{B}_{[0,\infty)}(H)$ plays a critical role in the development of the frequency domain stability results for parabolic partial differential equations given in [3]. In essence, $\bar{B}_{[0,\infty)}(H)$ provides a key link between the general theory of [4] and the results of [3].

We begin with some basic definitions and propositions in the next section. Then, in § 3, we prove a theorem essentially due to Foures and Segal [1] by different means. This theorem relates causality and analyticity in the case where $H = C$ is simply the complex numbers. Finally, we characterize $\bar{B}_{[0,\infty)}(H)$ in § 4.

2. Preliminaries. First recall the following definitions (e.g., [2]).

DEFINITION 2.1. Let $x(\cdot)$ be a map of R into H and let t be an element of R. Then the *truncation of* $x(\cdot)$ *at* t, in symbols $x_t(\cdot)$, is the map of R into H given by

$$x_t(\tau) = \begin{cases} x(\tau), & \tau \leqq t, \\ 0, & \tau > t, \end{cases} \tag{1}$$

for all τ in R.

DEFINITION 2.2. Let Z be a map of $L_2(R, H)$ into itself. Then Z is called *causal* (or nonanticipative) if

$$(Zx)_t(\cdot) = (Zx_t(\cdot))_t \tag{2}$$

for all $x(\cdot)$ in $L_2(R, H)$ and t in R.

We then have the following proposition.

PROPOSITION 2.3. *Let $Z, Z_n, n = 1, 2, \cdots$ be elements of $\mathscr{L}(L_2(R, H), L_2(R, H))$, the space of bounded linear maps of $L_2(R, H)$ into $L_2(R, H)$. Suppose that the Z_n are causal and that Z_n converges strongly to Z. Then Z is causal.*

Proof. Let t be any element of R and let $\chi_t(\cdot)$ denote the characteristic function of the set $(-\infty, t]$. To show that Z is causal, we must check that $\chi_t[Z(y_t - y)] = 0$ for all y in $L_2(R, H)$. Now $\chi_t[Z_n(y_t - y)] = 0$ for all y in $L_2(R, H)$ and $n = 1, 2, \cdots$ since each Z_n is causal. But $\|\chi_t[(Z_n - Z)(y_t - y)]\| \leqq \|(Z_n - Z)(y_t - y)\| \leqq \|Z_n - Z\| \, \|y_t - y\|$ and so, $\chi_t[Z(y_t - y)] = \lim_{n\to\infty} \chi_t[Z_n(y_t - y)] = 0$. In other words, Z is causal.

* Received by the editors September 10, 1968, and in revised form April 15, 1969.

† NASA Electronics Research Center, Cambridge, Massachusetts 02139.

‡ Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912 and NASA Electronics Research Center, Cambridge, Massachusetts. The work of this author was supported by the National Science Foundation under Grant GK-2788.

¶ Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. The work of this author was sponsored by the National Science Foundation under a research fellowship.

PROPOSITION 2.4. *Let Z be a bounded linear map of $L_2(R, H)$ into itself and let $\{e_i\}$ be an orthonormal basis of H. Define the maps Z_{ij} of $L_2(R, C)$ into $L_2(R, C)$ by*

$$(Z_{ij}y)(\cdot) = \langle Z(ye_j)(\cdot), e_i\rangle \tag{3}$$

for y in $L_2(R, C)$ and $1 \leqq i, j < \infty$.[1] *Then Z is causal if and only if all the Z_{ij} are causal.*

Proof. Suppose first that Z is causal. Then, $(Z_{ij}y_t)_t(\cdot) = \langle Z(y_te_j)(\cdot), e_i\rangle_t = \langle (Z(ye_j)_t)_t(\cdot), e_i\rangle = \langle [Z(ye_j)]_t(\cdot), e_i\rangle = \langle Z(ye_j)(\cdot), e_i\rangle_t = (Z_{ij}y)_t(\cdot)$ for any i, j, any t in R and any y in $L_2(R, C)$. Thus, all the Z_{ij} are causal.

On the other hand, suppose that all the Z_{ij} are causal. Then, to show that Z is causal, it will be enough to show that $\langle (Zx_t)_t(\cdot), e_i\rangle = \langle (Zx)_t(\cdot), e_i\rangle$ for all i, all t in R and all x in $L_2(R, H)$. However,

$$\langle (Zx_t)_t, e_i\rangle = \langle (Zx_t), e_i\rangle_t = \sum_{j=1}^{\infty} \langle [Z(x_t^je_j)], e_i\rangle_t = \sum_{j=1}^{\infty} (Z_{ij}x_t^j)_t = \sum_{j=1}^{\infty} (Z_{ij}x^j)_t$$

$$= \sum_{j=1}^{\infty} \langle [Z(x^je_j)]_t, e_i\rangle = \langle Z(x)_t, e_i\rangle,$$

where $x^j(\cdot) = \langle x(\cdot), e_j\rangle$. Thus, the proof of the proposition is complete.

We observe that Propositions 2.3 and 2.4 are also valid in the general context of locally compact Abelian groups developed in [2] and [4].

Now, recall that $B(H)$ is the set of all linear transformations of $L_2(R, H)$ into itself of the form

$$(\boldsymbol{\Phi}x)(t) = \int_{-\infty}^{\infty} \boldsymbol{\Phi}(t - \tau)x(\tau)\,d\tau + \lambda x(t), \tag{4}$$

where $\Phi \in L_1(R, \mathscr{L}(H, H))$ and $\lambda \in C$ (cf. [2], [4]). It is shown in [2] that $B(H)$ is a Banach algebra with respect to composition and the norm, $\|\boldsymbol{\Phi}\|_B = \|\Phi(\cdot)\|_1 + |\lambda|$, and that $B(H)$ is isometrically isomorphic with $L_1(R, \mathscr{L}(H, H)) \oplus \{\Delta\}$ where Δ is a unit. Writing elements $\boldsymbol{\Phi}$ of $B(H)$ in the form $\boldsymbol{\Phi} = \Phi + \lambda\Delta$, we let $B_{[0,\infty)}(H)$ be the subset of $B(H)$ given by

$$B_{[0,\infty)}(H) = \{\boldsymbol{\Phi} = \Phi + \lambda\Delta \in B(H) : \operatorname{supp}\Phi \subset [0, \infty)\}, \tag{5}$$

where supp Φ is the support of Φ. $B_{[0,\infty)}(H)$ is a closed subalgebra of $B(H)$. For our purposes, we need to view $B_{[0,\infty)}(H)$ as a subalgebra of an algebra rather different from $B(H)$. With this in mind, let $W_{[0,\infty)}(H)$ be the set of all linear maps Z of $L_2(R, H)$ into $L_2(R, H)$ such that (i) Z is causal, and (ii)[2] $(Zx)\hat{}(\omega) = z(\omega)\hat{x}(\omega)$ for all x in $L_2(R, H)$ and ω in R where $z(\cdot)$ is a bounded uniformly continuous element of $\mathscr{C}(R, \mathscr{L}(H, H))$ and $\hat{x}(\omega)$ is the Fourier transform of $x(\cdot)$ (see [2]). $W_{[0,\infty)}(H)$ is a Banach algebra with respect to composition and the norm, $\|Z\|_W = \sup_{\omega \in R}\{\|z(\omega)\|\}$, and has an identity Δ given by $\Delta x = x$. Moreover, $B_{[0,\infty)}(H)$ is a subalgebra of $W_{[0,\infty)}(H)$ which is not closed in $\|\cdot\|_W$. We let $\bar{B}_{[0,\infty)}(H)$ denote the closure of $B_{[0,\infty)}(H)$ in $W_{[0,\infty)}(H)$, (see [2]).

[1] The Z_{ij} are called the components of Z with respect to the basis $\{e_i\}$.

[2] Hat accent to right of parenthesis indicates Fourier transform of expression within parentheses.

3. The case $H = C$. Suppose, for the moment, that $H = C$, the complex numbers. We then have the following theorem.

THEOREM 3.1. (cf., [1]). *Let Z be a bounded linear map of $L_2(R, C)$ into itself such that*

$$(Zx)\hat{}(\omega) = f(\omega)\hat{x}(\omega) \tag{6}$$

for x in $L_2(R, C)$, where $f(\cdot)$ is a continuous complex-valued function on $(-\infty, \infty)$ which tends to a limit γ as $|\omega| \to \infty$. Then the following three statements are equivalent:

(a) *there is a bounded continuous complex-valued function $\psi(\cdot)$ on* $\operatorname{Re}\{s\} \geqq 0$ *such that $\psi(\cdot)$ is analytic on* $\operatorname{Re}\{s\} > 0$ *and $\psi(i\omega) = f(\omega)$ for all real ω;*

(b) *Z is causal; and,*

(c) *Z is an element of $\bar{B}_{[0,\infty)}(C)$.*

Proof. We can assume without loss of generality that $\gamma = 0$ throughout as $([Z - \gamma I]x(\omega))\hat{} = [f(\omega) - \gamma 1]\hat{x}(\omega)$.

We first prove that (a) implies (b). Let $\varphi(t)$ be any fixed nonnegative C^∞-function with compact support contained in $[0, \infty)$ and with $\int_0^\infty \varphi(t)\,dt = 1$. Let $\{\varphi_\varepsilon(t)\}$ be the associated approximate identity so that $\varphi_\varepsilon(t) = \varphi(t/\varepsilon)/\varepsilon$ for $\varepsilon > 0$. If $x(\cdot)$ is an element of $L_2(R, C)$, then $\varphi_\varepsilon * x$ is in $L_2(R, C)$ and we can define the linear maps Z_ε of $L_2(R, C)$ into itself by setting

$$Z_\varepsilon x = Z(\varphi_\varepsilon * x). \tag{7}$$

It follows that $(Z_\varepsilon x)\hat{}(\omega) = f(\omega)\hat{\varphi}(\varepsilon\omega)\hat{x}(\omega)$ and hence, that Z_ε converges strongly to Z as $\varepsilon \to 0$ (i.e., $\sup_{\|x\|_2 = 1}\|Z_\varepsilon x - Zx\|_2 \to 0$ as $\varepsilon \to 0$).[3] In view of Proposition 2.3, we need only show that each Z_ε is causal.

For any fixed $\varepsilon > 0$, $f(\omega)\hat{\varphi}(\varepsilon\omega)$ is in $L_2(R, C)$ as $f(\cdot)$ is bounded and $\hat{\varphi}(\cdot)$ is in $L_2(R, C)$. It follows that there is a $g(\cdot)$ in $L_2(R, C)$ with[4] $\hat{g}(\omega) = f(\omega)\hat{\varphi}(\varepsilon\omega)$ and hence, that $Z_\varepsilon x = g * x$. Thus, Z_ε will be causal if $\operatorname{supp} g \subset [0, \infty)$.

Now, letting $\hat{\Phi}_\varepsilon(s)$ be the Laplace transform of $\varphi_\varepsilon(t)$ so that

$$\hat{\Phi}_\varepsilon(s) = \int_0^\infty e^{-st}\varphi_\varepsilon(t)\,dt \tag{8}$$

for $\operatorname{Re}\{s\} \geqq 0$, we consider the function $\psi(s)\hat{\Phi}_\varepsilon(s)$ which is analytic on $\operatorname{Re}\{s\} > 0$ and which takes the values $f(\omega)\hat{\varphi}(\varepsilon\omega)$ when $s = i\omega$. Since $\psi(s)$ is bounded on $\operatorname{Re}\{s\} \geqq 0$, we have

$$\begin{aligned} \int_{-\infty}^{x} |\psi(\sigma + i\omega)\hat{\Phi}_\varepsilon(\sigma + i\omega)|^2\,d\omega &\leqq K \int_{-\infty}^{\infty} |\hat{\Phi}_\varepsilon(\sigma + i\omega)|^2\,d\omega \\ &\leqq K \int_{-\infty}^{\infty} |\varphi_\epsilon(t)|^2\,dt < \infty \end{aligned} \tag{9}$$

[3] By Proposition 3.2 of [2]. Note also that $\hat{\varphi}_\varepsilon(\omega) = \hat{\varphi}(\varepsilon\omega)$.

[4] Note that $\hat{g}(\cdot)\hat{x}(\cdot) = (Z_\varepsilon x)\hat{}(\cdot)$ is in $L_1(R, C) \cap L_2(R, C)$.

for $\sigma > 0$, where K is a fixed constant which is independent of σ. It follows from a theorem of Paley–Wiener [7] that the inverse transform $g(t)$ of $\psi(i\omega)\hat{\Phi}_\varepsilon(i\omega)$ has support contained in $[0, \infty)$. Thus, Z_ε is causal for any $\varepsilon > 0$, and so (a) implies (b).

We now show that (b) implies (c). Again, we let Z_ε be given by (7). Here, however, Z_ε is causal since Z is causal and supp $\varphi_\varepsilon \subset [0, \infty)$.[5] Moreover, $\| Z_\varepsilon - Z \|_W = \sup_{\omega\in R}|f(\omega)\hat{\varphi}(\varepsilon\omega) - f(\omega)|$ tends to 0 as $\varepsilon \to 0$. To see this, we note that if $\delta > 0$ is given, then there is an $\alpha > 0$ such that $|\omega| > \alpha$ implies that $|f(\omega)| < \delta/2$ since $\lim_{|\omega|\to\infty} f(\omega) = 0$. Thus,

$$\begin{aligned}\sup_{\omega\in R} &|f(\omega)\varphi(\varepsilon\omega) - f(\omega)|\\ &\leqq \max\{\sup_{\omega\in[-\alpha,\alpha]} |f(\omega)\hat{\varphi}(\varepsilon\omega) - f(\omega)|, \sup_{|\omega|>\alpha} |f(\omega)\hat{\varphi}(\varepsilon\omega) - f(\omega)|\}\\ &\leqq \max\{\sup_{\omega\in[-\alpha,\alpha]} |f(\omega)\hat{\varphi}(\varepsilon\omega) - f(\omega)|, \delta\}\end{aligned}$$

since

$$\sup_{|\omega|>\alpha} |f(\omega)\hat{\varphi}(\varepsilon\omega) - f(\omega)| \leqq \sup_{|\omega|>\alpha} 2|f(\omega)| < \delta.$$

But

$$\begin{aligned}|f(\omega)\hat{\varphi}(\varepsilon\omega) - f(\omega)| &\leqq |f(\omega)|\left|\int_0^\infty [e^{-i\varepsilon\omega t} - 1]\varphi(t)\,dt\right|\\ &\leqq |f(\omega)| \int_0^\infty |[e^{-i\varepsilon\omega t} - 1]|\varphi(t)\,dt.\end{aligned}$$

Since $f(\omega)$ is continuous, $|f(\omega)|$ is bounded on $[-\alpha, \alpha]$ and so, for suitably small ε, we have $\|Z_\varepsilon - Z\|_W < \delta$.

Now, just as in the proof that (a) implies (b), we have $Z_\varepsilon x = g * x$, where $g(\cdot)$ is an element of $L_2(R, C)$ and $\hat{g}(\omega) = f(\omega)\hat{\varphi}(\varepsilon\omega)$. Here, however, supp $g \subset [0, \infty)$ as Z_ε is causal. Since $\|Z_\varepsilon - Z\|_W \to \infty$ as $\varepsilon \to 0$, we need only show that Z_ε is in $\bar{B}_{[0,\infty)}(C)$ for every ε.

With this in mind, we let

$$g_\sigma(t) = e^{-\sigma|t|}g(t) \tag{10}$$

for $\sigma > 0$ and

$$Z_{\varepsilon,\sigma}x = g_\sigma * x \tag{11}$$

for x in $L_2(R, C)$, $\sigma > 0$ and a fixed $\varepsilon > 0$. Since $g_\sigma(\cdot)$ is in $L_1(R, C)$ and supp $g_\sigma \subset [0, \infty)$, $Z_{\varepsilon,\sigma}$ is an element of $B_{[0,\infty)}(C)$. To complete the proof, we shall show that $\|Z_{\varepsilon,\sigma} - Z_\varepsilon\|_W$ tends to 0 as $\sigma \to 0$, i.e., that $\sup_{\omega\in R} |\hat{g}_\sigma(\omega) - \hat{g}(\omega)| \to 0$ as $\sigma \to 0$. However,

$$(e^{(-\sigma|\cdot|)}g(\cdot))\hat{} = \frac{1}{2\pi} e^{(-\sigma|\cdot|)\hat{}} * \hat{g}(\cdot), \tag{12}$$

and so, we have

$$\hat{g}_\sigma(\omega) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\sigma}{\sigma^2 + (\omega - v)^2}\hat{g}(v)\,dv \tag{13}$$

[5] Note that the composition of causal maps is causal.

for $\sigma > 0$. The result is now an immediate consequence of the fact that $\hat{g}(\cdot)$ is uniformly continuous[6] and of Lemma 3.2 which follows the theorem.

We now prove that (c) implies (a). If Z is an element of $\bar{B}_{[0,\infty)}(C)$, then there is a sequence of elements $g_n(\cdot)$ of $L_1(R, C)$ with $\operatorname{supp} g_n \subseteq [0, \infty)$ such that $\hat{g}_n(\omega)$ converges to $f(\omega)$ uniformly in ω. Therefore, $\hat{g}_n(\omega)$ is uniformly Cauchy in ω. Now, let $G_n(s)$ be given by

$$(14) \qquad G_n(s) = \int_0^\infty e^{-st} g_n(t)\,dt$$

for $\operatorname{Re}\{s\} \geqq 0$. Since

$$\sup_{\operatorname{Re}\{s\}\geqq 0} |G_n(s) - G_m(s)| = \sup_{\omega\in R} |G_n(i\omega) - G_m(i\omega)|^7 = \sup_{\omega\in R} |\hat{g}_n(\omega) - \hat{g}_m(\omega)|,$$

the sequence $G_n(s)$ is uniformly Cauchy on $\operatorname{Re}\{s\} \geqq 0$. But the $G_n(s)$ are analytic on $\operatorname{Re}\{s\} > 0$ and continuous on $\operatorname{Re}\{s\} \geqq 0$ and have the limit 0 at ∞. It follows that $G_n(s)$ converges uniformly to a function $G(s)$ which is analytic on $\operatorname{Re}\{s\} > 0$, continuous on $\operatorname{Re}\{s\} \geqq 0$, and has limit 0 at ∞. Moreover, $G(i\omega) = f(\omega)$. Thus, (a) holds and the proof is complete.

Now let H again be any separable complex Hilbert space. We then have the following lemma.

LEMMA 3.2 (cf. Titchmarsh [5]). *If $h(\cdot)$ is an element of $L_2(R, \mathscr{L}(H, H))$ which is uniformly continuous on R, and if $K(w, v, \sigma)$ is given by*

$$(15) \qquad K(w, v, \sigma) = \frac{1}{\pi}\frac{\sigma}{\sigma^2 + (w - v)^2},$$

then $\int_{-\infty}^{\infty} K(w, v, \sigma)h(v)\,dv$ converges to $h(w)$ uniformly in w as σ tends to 0.

Proof. We first observe that

$$(16) \qquad \int_w^\infty K(w, v, \sigma)\,dv = \int_{-\infty}^w K(w, v, \sigma)\,dv = \tfrac{1}{2},$$

$$(17) \qquad K(w, v, \sigma) = \begin{cases} 0(1/\sigma), & |w - v| \leqq \sigma, \\ 0(\sigma/|w - v|^2), & |w - v| > \sigma. \end{cases}$$

Thus, it will be sufficient to show that $\int_w^\infty K(w, v, \sigma)[h(v) - h(w)]\,dv \to 0$ uniformly in w as $\sigma \to 0+$. In view of (17), this will be true if (a) $(1/\sigma)\int_w^{w+\sigma} \|h(v) - h(w)\|\,dv \to 0$ uniformly in w as $\sigma \to 0+$, and (b) $\sigma\int_{w+\sigma}^\infty (\|h(v) - h(w)\|/(v - w)^2)\,dv \to 0$ uniformly

[6] This follows from the fact that $f(\cdot)$ has a limit as $|\omega| \to \infty$, the continuity of $f(\cdot)$, and the properties of $\hat{\varphi}(\varepsilon\omega)$.

[7] By the maximum modulus theorem.

in w as $\sigma \to 0+$. Since (a) is an immediate consequence of the uniform continuity of $h(\cdot)$, we need only establish (b). Now let

$$\psi(T) = \int_0^T \|h(w + t) - h(w)\| \, dt. \tag{18}$$

Then, given $\varepsilon > 0$, there is an $\eta > 0$ such that

$$\frac{\psi(T)}{T} < \varepsilon, \qquad 0 < T \leqq \eta, \tag{19}$$

for all w in R. It follows that

$$\begin{aligned} \sigma \int_{w+\sigma}^{w+\eta} \frac{\|h(v) - h(w)\|}{(v - w)^2} dv &= \sigma \int_{\sigma}^{\eta} \frac{\|h(w + t) - h(w)\|}{t^2} dt \\ &= \sigma \left[\frac{\psi(t)}{t^2} \right]_{\sigma}^{\eta} + 2\sigma \int_{\sigma}^{\eta} \frac{\psi(t)}{t^3} dt \\ &\leqq \varepsilon + 2\sigma\varepsilon \int_{\sigma}^{\eta} \frac{dt}{t^2} \leqq 3\varepsilon, \end{aligned} \tag{20}$$

uniformly in w. Now, with η fixed, it is clear that $\sigma \int_{w+\eta}^{\infty} (\|h(v) - h(w)\|/(v - w)^2) \, dv \to 0$ uniformly in w as $\sigma \to 0+$. The lemma is now established.

4. The general case. Suppose now that H is a separable complex Hilbert space. We then have the following theorem.

THEOREM 4.1. *Let Z be a bounded linear map of $L_2(R, H)$ into itself such that*

$$(Zx)\hat{}(\omega) = f(\omega)\hat{x}(\omega) \tag{21}$$

for x in $L_2(R,H)$, where $f(\cdot)$ is a continuous map of R into $\mathscr{L}(H, H)$ which tends to the limit γI as $|\omega| \to \infty$, $\gamma \in C$. Then (i) *if there is a bounded continuous $\mathscr{L}(H, H)$-valued function $\psi(\cdot)$ on* Re $\{s\} \geqq 0$ *such that $\psi(\cdot)$ is analytic on* Re $\{s\} > 0$ *and $\psi(i\omega) = f(\omega)$ for all real ω, then Z is causal (i.e., $Z \in W_{[0,\infty)}(H)$),* and (ii) *if Z is causal and if $z(\omega) = f(\omega) - \gamma I$ is completely continuous for all real ω, then Z is in $\bar{B}_{[0,\infty)}(H)$.*

Proof. We first prove (i). Let $\{e_i\}$ be an orthonormal basis of H and let Z_{ij} be the components of Z with respect to $\{e_i\}$ (i.e., Z_{ij} is the map of $L_2(R, C)$ into itself given by (3)). Then, letting $\psi_{ij}(s) = \langle \psi(s)e_j, e_i \rangle$, we can see that $(Z_{ij}y)\hat{}(\omega) = \psi_{ij}(\omega)\hat{y}(\omega)$ for all y in $L_2(R, C)$ and, hence, that Z_{ij} satisfies (a) of Theorem 3.1. Thus, each Z_{ij} is causal and (i) follows immediately from Proposition 2.4.

We now prove (ii). We assume without loss of generality that $\gamma = 0$ and we let Z_{ij} be the components of Z with respect to the basis $\{e_i\}$. Then each Z_{ij} is causal by virtue of Proposition 2.4 and, hence, is an element of $\bar{B}_{[0,\infty)}(C)$ in view of Theorem 3.1. Now let H_n be the span of $\{e_1, \cdots, e_n\}$ and E_n be the projection of H onto H_n. Defining the map Z_n of $L_2(R, H)$ into itself by $Z_n = E_n Z E_n$, we have

$$(Z_n x)\hat{}(\omega) = E_n f(\omega) E_n \hat{x}(\omega) \tag{22}$$

for x in $L_2(R, H)$. However, (22) implies that

$$(23) \qquad (Z_{n,ij}y)\hat{}(\omega) = \begin{cases} f_{ij}(\omega)\hat{y}(\omega), & 1 \leqq i \leqq n, \quad 1 \leqq j \leqq n, \\ 0, & i > n \text{ or } j > n, \end{cases}$$

for y in $L_2(R, C)$, where $f_{ij}(\omega) = \langle f(\omega)e_j, e_i \rangle$. It follows that Z_n is in $\bar{B}_{[0,\infty)}(H)$.

To complete the proof, we shall show that Z_n converges to Z in $\|\cdot\|_W$. We first note that $\lim_{n\to\infty} \|f(\omega) - E_n f(\omega)E_n\| = 0$ pointwise in ω as $f(\omega)$ is completely continuous for each ω. Let $f_n(\omega) = E_n f(\omega)E_n$. Then $\{f_n(\cdot)\}$ is an equicontinuous family since $\|f_n(\omega_1) - f_n(\omega_0)\| \leqq \|f(\omega_1) - f(\omega_0)\|$ for all ω_1, ω_0 and $f(\cdot)$ is continuous. Now let $\varepsilon > 0$ be given. Then there is a $\delta > 0$ such that $|\omega| > \delta$ implies that $\|f(\omega)\| < \varepsilon/4$ since $\lim_{|\omega|\to\infty} f(\omega) = 0$. But $\|f_n(\omega)\| \leqq \|f(\omega)\|$, and so $\|f(\omega) - f_n(\omega)\| < \varepsilon/2$ for all ω with $|\omega| > \delta$ and *all n*. Since $\{f_n(\cdot)\}$ is an equicontinuous family which converges pointwise to the continuous function $f(\cdot)$ on the compact set $|\omega| \leqq \delta$, $f_n(\cdot)$ converges to $f(\cdot)$ *uniformly* on $|\omega| \leqq \delta$. It follows that, for all ω in $|\omega| \leqq \delta$, there is an n (independent of ε) such that $\|f_n(\omega) - f(\omega)\| < \varepsilon/2$. Thus, Z_n converges to Z in $\|\cdot\|_W$, and the proof is complete.

COROLLARY 4.2. *Suppose that the hypotheses of* (i) *are satisfied and that* $f(\omega) - \gamma I$ *is completely continuous for all* ω. *Then* Z *is an element of* $\bar{B}_{[0,\infty)}(H)$.

REFERENCES

[1] Y. FOURES AND I. E. SEGAL, *Causality and analyticity*, Bull. Amer. Math. Soc., 61 (1955), pp. 385–405.

[2] P. L. FALB AND M. I. FREEDMAN, *A generalized transform theory for causal operators*, this Journal, 7 (1969), pp. 452–471.

[3] M. I. FREEDMAN, P. L. FALB AND J. ANTON, *Transform theoretic approach to the stability of a class of nonlinear partial differential equations*, in preparation.

[4] M. I. FREEDMAN, P. L. FALB AND G. ZAMES, *A Hilbert space stability theory over locally compact Abelian groups*, this Journal, 7 (1969), pp. 479–495.

[5] E. C. TITCHMARSH, *Theory of Fourier Integrals*, 2nd ed., Oxford University Press, New York, 1948.

[6] N. DUNFORD AND J. SCHWARTZ, *Linear Operators. Part I: General Theory*, Interscience, New York, 1958.

[7] R. E. A. C. PALEY AND N. WIENER, *Fourier Transforms in the Complex Domain*, American Mathematical Society, Providence, Rhode Island, 1934.

SIAM J. CONTROL
Vol. 7, No. 3, August 1969

A HILBERT SPACE STABILITY THEORY OVER LOCALLY COMPACT ABELIAN GROUPS*

M. I. FREEDMAN†, P. L. FALB‡ AND G. ZAMES†

1. Introduction. Let G be a separable locally compact Abelian group and let H be a separable real Hilbert space. Here we develop a very general stability theory for systems defined on G and taking values in H. Our theory draws heavily on the generalized transforms developed in [1] and culminates in a generalization of the circle criterion [2] which is applicable to a very wide class of systems.

We shall assume that the reader is familiar with the theory of integration of Banach-space-valued functions [3] and has perused the work on generalized transforms in [1]. Let μ denote Haar measure on G and let $\hat{G}$ denote the character group of G. Elements of $\hat{G}$ are usually denoted by γ and their action on G is written as (γ, g). We recall [4] that $\hat{G}$ is also a locally compact Abelian group (with respect to the topology of uniform convergence on compact subsets of G). Let m denote the Haar measure on $\hat{G}$. We deal quite freely with spaces of the form $L_2(G, H)$, $L_2(\hat{G}, H)$, $L_1(G, \mathscr{L}(H, H))$, etc., where, for example, $L_2(G, H)$ is the space of maps f of G into H for which $\|f\|^2$ is integrable with respect to μ and $\mathscr{L}(H, H)$ is the space of bounded linear maps of H into itself [3], [1].

We begin our development with some basic definitions in the next section. Then, in §3, we prove several positivity lemmas which play an important role in our treatment of the circle criterion. We next review (briefly) the main theorem of [1] and use it to obtain a spectral theory and positivity conditions (§4). In §5, we state and prove our main result, the generalized circle criterion for stability. Finally, we present some illustrative examples and make some concluding comments in §6.

2. Basic definitions. Following [1], we introduce a generalization of the notions of truncation and causality. We also define the notions of finite gain, positivity and stability (cf. [2]). All of these ideas play a significant role in the sequel.

DEFINITION 2.1. Let $P \subset G$ be a closed semigroup of positive Haar measure and let $P' = \{g : -g \in P\}$. Let $P' + g_0$ be the subset of G given by

$$P' + g_0 = \{g \in G : g = g_1 + g_0, \quad g_1 \in P'\} \tag{1}$$

and let χ_{g_0} be the characteristic function of $P' + g_0$. If f is a measurable map of G into H, then the *truncation* of f at g_0, f_{g_0}, is given by

$$f_{g_0}(g) = \begin{cases} f(g), & g \in P' + g_0, \\ 0, & g \notin P' + g_0, \end{cases} \tag{2}$$

i.e., $f_{g_0} = \chi_{g_0} f$.[1]

* Received by the editors August 13, 1968, and in revised form April 15, 1969.

† NASA Electronics Research Center, Cambridge, Massachusetts 02139.

‡ Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912 and NASA Electronics Research Center, Cambridge, Massachusetts. The work of this author was supported by the National Science Foundation under Grant GK-2788.

[1] Strictly speaking, we should write $f_{g_0,P}$ since truncation depends on the choice of P. However, we usually consider a fixed P and so this distinction is unnecessary.

If P is a closed semigroup of positive Haar measure, then we introduce the space

$$L_{2P}(G, H) = \{f : f \text{ is measurable and } f_{g_0} \in L_2(G, H) \text{ for all } g_0\},$$

and we call $L_{2P}(G, H)$ the *extension of* $L_2(G, H)$ (*relative to* P). We shall assume from now on that there are elements $\{g_1, \cdots\}$ in P such that $\bigcup_{n=1}^{\infty} (P' + g_n) = G$. We now have the following proposition.

PROPOSITION 2.2. *If $f(\cdot)$ is an element of $L_{2P}(G, H)$ and if $\|f_{g_0}\|_2 \leqq K$ for all g_0 in G and a fixed constant K, then $f(\cdot)$ is an element of $L_2(G, H)$ and $\|f\|_2 \leqq K$.*

Proof. There are elements $g_1, \cdots$ in P such that $\bigcup_{n=1}^{\infty} (P' + g_n) = G$. Let $h_n = \sum_1^n g_i$. Then $h_n \in P$, $\bigcup_{n=1}^{\infty} (P' + h_n) = G$ and $P' + h_n \subset P' + h_k$ if $k > n$. We let[2] $f_N(g) = \sup\{|f_{h_1}(g)|^2, \cdots, |f_{h_N}(g)|^2\}$. Then $f_N(\cdot)$ is a monotone sequence of integrable functions and $\int f_N(g)\, d\mu \leqq K^2$ since $f_{h_i} = f_{h_j}$ on $(P' + h_i) \cap (P' + h_j)$. It follows from the Beppo Levi theorem on monotone convergence [5] that $\int |f(g)|^2\, d\mu \leqq K^2$ and hence that $f(\cdot)$ is an element of $L_2(G, H)$ with $\|f\|_2 \leqq K$.

PROPOSITION 2.3. *If $f_1(\cdot)$ and $f_2(\cdot)$ are elements of $L_{2P}(G, H)$ and if $f_{1_g} = f_{2_g}$ for all g in G, then $f_1 \equiv f_2$.*

PROPOSITION 2.4. *Let $g_\alpha, \alpha \in I$, be any ordering of G, where I is an index set. Suppose that $\{f_\alpha(\cdot) : \alpha \in I\}$ is a collection of elements of $L_2(G, H)$ such that $f_\alpha(g) = 0$, for almost all g with $g \notin P' + g_\alpha$, and $\chi_{g_\beta} f_\alpha = \chi_{g_\alpha} f_\beta$ for all α, β in I. Then $\{f_\alpha\}$ defines a unique element f of $L_{2P}(G, H)$; namely, $f(g) = f_\alpha(g)$ if $g \in P' + g_\alpha$.*

Proof. Since $G = \bigcup (P' + g_\alpha)$, $f(g)$ is defined for every g in G. Moreover, if g is an element of $(P' + g_\alpha) \cap (P' + g_\beta)$, then $f(g) = f_\alpha(g) = \chi_{g_\beta}(g) f_\alpha(g) = \chi_{g_\alpha}(g) f_\beta(g) = f_\beta(g)$ so that f is well defined. There are elements $g_1, \cdots$ of P such that $G = \bigcup_{n=1}^{\infty} (P' + g_n)$, and so f will be defined by a countable set $\{f_{\alpha_i} : i = 1, \cdots\}$. Since the f_{α_i} are measurable, f will also be measurable. In view of the fact that $f_\alpha(g) = 0$ for almost all g with $g \notin P' + g_\alpha$, we can see that f_g will be in $L_2(G, H)$ for all g in G. Thus, $f \in L_{2P}(G, H)$. The uniqueness of f is then an immediate consequence of Prcposition 2.3.

With a suitable notion of truncation at hand, we can define causality as follows:

DEFINITION 2.5. Let Φ be a map of $L_{2P}(G, H)$ into $L_{2P}(G, H)$ (or of $L_2(G, H)$ into $L_2(G, H)$). Then Φ is called *causal with respect to P* if

$$(\Phi x)_{g_0}(\cdot) = (\Phi(x_{g_0}(\cdot)))_{g_0} \tag{3}$$

for all g_0 in G and x in $L_{2P}(G, H)$ (or $L_2(G, H)$).

We observe that if $G = R$ and $P = [0, \infty)$, then Definition 2.5 coincides with the usual notion of causality [2]. We also note that if Φ and Ψ are causal with respect to P, then $\Phi\Psi$ is causal with respect to P as $(\Phi\Psi x)_{g_0}(\cdot) = (\Phi\{\Psi x\})_{g_0}(\cdot) = (\Phi\{\Psi x_{g_0}(\cdot)\}_{g_0})_{g_0} = (\Phi\Psi x_{g_0}(\cdot))_{g_0}$. Some basic examples of causal maps are as follows.

[2] $|\cdot|$ denotes the norm on H.

Example 2.6. Let N be a map of H into H with $|N(h)| \leqq c|h|$ for all h in H. Then the map $\mathbf{N}$ of $L_{2P}(G, H)$ into $L_{2P}(G, H)$ given by $(\mathbf{N}x)(g) = N(x(g))$ is causal.

Example 2.7. Let ψ be an element of $L_1(G, \mathscr{L}(H, H))$ with support contained in P. Then the map $\boldsymbol{\psi}$ of $L_{2P}(G, H)$ into $L_{2P}(G, H)$ given by

$$(\boldsymbol{\psi}x)(g) = \int_G \psi(g - g')x(g')\,d\mu$$

is well-defined and causal with respect to P (see, [1, Lemma 3.6]).

We now have the following definition.

DEFINITION 2.8. Let Φ be a map of $L_{2P}(G, H)$ into $L_{2P}(G, H)$ (or of $L_2(G, H)$ into $L_2(G, H)$). Then the *gain of* Φ, $\nu(\Phi)$, is given by

$$\nu(\Phi) = \sup\left\{\frac{\|(\Phi x)_{g_0}(\cdot)\|_2}{\|x_{g_0}(\cdot)\|_2} : x \in L_{2P}(G, H), \quad g_0 \in G, \quad x_{g_0} \neq 0\right\}$$

$$\left(\text{or by} \quad \nu(\Phi) = \sup\left\{\frac{\|(\Phi x)(\cdot)\|_2}{\|x(\cdot)\|_2} : x \in L_2(G, H), \quad x(\cdot) \neq 0\right\}\right),$$

and Φ is said to be of *finite gain* if $\nu(\Phi) < \infty$.

We observe that the map $\mathbf{N}$ of Example 2.6 is of finite gain with $\nu(\mathbf{N}) \leqq c$ and that the map $\boldsymbol{\psi}$ of Example 2.7 is also of finite gain with $\nu(\boldsymbol{\psi}) \leqq \|\psi\|_1$.

PROPOSITION 2.9. *If Φ is a causal map of $L_2(G, H)$ into itself, then there is a unique causal extension Φ' of Φ with Φ' mapping $L_{2P}(G, H)$ into itself. Moreover, if $\nu(\Phi) < \infty$, then $\nu(\Phi') < \infty$.*

Proof. Let $g_\alpha, \alpha \in I$, be any ordering of G, where I is an index set. If f is an element of $L_{2P}(G, H)$, then we define a collection $\{\phi_\alpha(\cdot)\}$ of elements of $L_2(G, H)$ by setting $\phi_\alpha = (\Phi f_{g_\alpha})_{g_\alpha}$. Then

$$\chi_{g_\beta}\phi_\alpha = \chi_{g_\beta}\chi_{g_\alpha}\Phi(f_{g_\alpha}) = \chi_{g_\alpha}\chi_{g_\beta}\Phi(\chi_{g_\beta}\chi_{g_\alpha}f) = \chi_{g_\alpha}\chi_{g_\beta}\phi(f_{g_\beta}) = \chi_{g_\alpha}\phi_\beta$$

since Φ is causal over $L_2(G, H)$. By Proposition 2.4, $\{\phi_\alpha\}$ represents a unique element ϕ of $L_{2P}(G, H)$. We let $\phi = \Phi'f$. Then $(\Phi'f)_{g_\alpha} = \phi_\alpha = (\Phi f_{g_\alpha})_{g_\alpha}$ and $(\Phi'f_{g_\alpha})_{g_\alpha} = (\Phi f_{g_\alpha})_{g_\alpha}$ so that Φ' is causal. The final assertion of the proposition is obvious.

Now define positivity and stability as follows.

DEFINITION 2.10. Let Φ be a map of $L_{2P}(G, H)$ into itself. Then Φ is *positive with respect to P* if

$$\langle x_{g_0}(\cdot), (\Phi x)_{g_0}(\cdot)\rangle = \int_G \langle x_{g_0}(g), (\Phi x)_{g_0}(g)\rangle\,d\mu \geqq 0 \tag{4}$$

for all g_0 in G and x in $L_{2P}(G, H)$. Similarly, Φ is *strongly positive with respect to P* if there is a $\delta > 0$ such that

$$\langle x_{g_0}(\cdot), (\Phi x)_{g_0}(\cdot)\rangle \geqq \delta\|x_{g_0}(\cdot)\|_2^2 \tag{5}$$

for all g_0 in G and x in $L_{2P}(G, H)$.

DEFINITION 2.11. Let Φ and Ψ be maps of $L_{2P}(G, H)$ into itself and let S be a subset of $L_2(G, H) \times L_2(G, H)$. Then the system

$$e(\cdot) + \Phi[(\Psi e)(\cdot) + x_1(\cdot)] = x_2(\cdot) \tag{6}$$

is called *stable over S* (or *L_2-stable over S*) if there is a continuous nonnegative function $K_s(\cdot)$ on $[0, \infty)$ with $K_s(0) = 0$ such that $(x_1(\cdot), x_2(\cdot)) \in S$ and $\|x_1\|_2 \leqq A$, $\|x_2\|_2 \leqq A$ together imply that $\|e(\cdot)\|_2 \leqq K_s(A)$ for every solution $e(\cdot)$ of (6) in $L_{2P}(G, H)$ (a fortiori, $e(\cdot) \in L_2(G, H)$).

We shall, on occasion, deal with relations rather than mappings and so, make the following definitions.

DEFINITION 2.12. A *relation* Φ on $L_{2P}(G, H)$ is a subset of the product $L_{2P}(G, H) \times L_{2P}(G, H)$. If x is an element of the domain of Φ, then the *image* of x, $\Phi[x]$, is the subset of $L_{2P}(G, H)$ given by $\Phi[x] = \{y \in L_{2P}(G, H) : (x, y) \in \Phi\}$. The *inverse* of Φ, Φ^{-1}, is the relation $\{(y, x) : (x, y) \in \Phi\}$. If Φ is a relation on $L_{2P}(G, H)$, then the *gain* of Φ, $\nu(\Phi)$, is given by

$$\nu(\Phi) = \sup\left\{\frac{\|y_{g_0}(\cdot)\|_2}{\|x_{g_0}(\cdot)\|_2} : x \in L_{2P}(G, H), y \in \Phi[x], g_0 \in G, x_{g_0} \neq 0\right\},$$

and Φ is said to be *finite gain* if $\nu(\Phi) < \infty$.

DEFINITION 2.13. Let Φ be a relation on $L_{2P}(G, H)$. Then Φ is *positive with respect to P if*

$$\langle x_{g_0}(\cdot), y_{g_0}(\cdot)\rangle = \int_G \langle x_{g_0}(g), y_{g_0}(g)\rangle \, d\mu \geqq 0 \tag{7}$$

for all g_0 in G, x in $L_{2P}(G, H)$ and $y \in \Phi[x]$. Similarly, Φ is *strongly positive with respect to P* if there is a $\delta > 0$ such that $\langle x_{g_0}(\cdot), y_{g_0}(\cdot)\rangle \geqq \delta\|x_{g_0}(\cdot)\|_2^2$ for all g_0 in G, x in $L_{2P}(G, H)$ and $y \in \Phi[x]$.

DEFINITION 2.14. Let Φ and Ψ be relations on $L_{2P}(G, H)$ and let S be a subset of $L_2(G, H) \times L_2(G, H)$. Then the *inclusion equation*

$$x_2(\cdot) - e(\cdot) \in \Phi[\Psi[e(\cdot)] + x_1(\cdot)] \tag{8}$$

is called *stable over S* (or *L_2-stable over S*) if there is a continuous nonnegative function $K_s(\cdot)$ on $[0, \infty)$ with $K_s(0) = 0$, such that $(x_1(\cdot), x_2(\cdot)) \in S$ and $\|x_1\|_2 \leqq A$, $\|x_2\|_2 \leqq A$ together imply that $\|e(\cdot)\|_2 \leqq K_s(A)$ for every solution $e(\cdot)$ of (8) in $L_{2P}(G, H)$.

3. Some positivity lemmas.

We now state and prove several basic lemmas which play a key role in our treatment of the circle criterion. We begin with the following lemma.

LEMMA 3.1. *Let Φ_1 and Φ_2 be positive maps of $L_{2P}(G, H)$ into $L_{2P}(G, H)$. Suppose that (say) Φ_2 is strongly positive with respect to P and has finite gain. Then the systems*

$$e(\cdot) + \Phi_1((\Phi_2 e)(\cdot) + x_1(\cdot)) = x_2(\cdot), \tag{9}$$

$$e(\cdot) + \Phi_2((\Phi_1 e)(\cdot) + x_1(\cdot)) = x_2(\cdot) \tag{10}$$

are L_2-stable over S for every $S \subset L_2(G, H) \times L_2(G, H)$.

Proof. We consider the case of (9) only since the proof in the other case is similar and is, therefore, omitted. Let S be any subset of $L_2(G, H) \times L_2(G, H)$. Now, for any g_0 in G and $(x_1, x_2) \in S$ and e in $L_{2P}(G, H)$ satisfying (9),

$$\langle(\Phi_1(\Phi_2 e + x_1))_{g_0}, (\Phi_2 e + x_1)_{g_0}\rangle \geqq 0, \tag{11}$$

by virtue of the positivity of Φ_1. Since $(\Phi_1(\Phi_2 e + x_1))_{g_0} = (x_2 - e)_{g_0}$, it follows that

$$(12)\quad \langle e_{g_0}, (\Phi_2 e)_{g_0}\rangle \leqq \|x_{1g_0}\|_2\|x_{2g_0}\|_2 + \|x_{1g_0}\|_2\|e_{g_0}\|_2 + \nu(\Phi_2)\|x_{2g_0}\|_2\|e_{g_0}\|_2.$$

However, Φ_2 is strongly positive with respect to P, and so, $\langle e_{g_0}, (\Phi_2 e)_{g_0}\rangle \geqq \delta\|e_{g_0}\|_2^2$ for some $\delta > 0$ and all $g_0 \in G$. Thus, for any $A > 0$, $\|x_1\|_2 \leqq A$ and $\|x_2\|_2 \leqq A$ imply that

$$(13)\qquad \delta\|e_{g_0}\|_2^2 \leqq A^2 + A(1 + \nu(\Phi_2))\|e_{g_0}\|_2,$$

and hence, that $\|e_{g_0}\|_2 \leqq K_s(A)$, where, for example, $K_s(A) = A(1 + \nu(\Phi_2) + \delta)/\delta$ and for *all* g_0 in G. By virtue of Proposition 2.2, $\|e(\cdot)\|_2 \leqq K_s(A)$ and the lemma is established.

An analogous result holds for relations. In other words, we have the following lemma.

LEMMA 3.2. *Let Φ_1 and Φ_2 be relations on $L_{2P}(G, H)$ (with domains all of $L_{2P}(G, H)$). Suppose that Φ_1 and Φ_2 are positive with respect to P and that (say) Φ_2 is strongly positive with respect to P and has finite gain. Then the systems*

$$(14)\qquad x_2(\cdot) - e(\cdot) \in \Phi_1[\Phi_2[e(\cdot)] + x_1(\cdot)],$$

$$(15)\qquad x_2(\cdot) - e(\cdot) \in \Phi_2[\Phi_1[e(\cdot)] + x_1(\cdot)]$$

are L_2-stable over S for every $S \subset L_2(G, H) \times L_2(G, H)$,

Proof. The proof is a direct analogue of the proof of Lemma 3.1.

We now use Lemma 3.2 to prove Lemma 3.3.

LEMMA 3.3. (Skeleton circle criterion). *Let Φ and Ψ be maps of $L_{2P}(G, H)$ into itself and let $\mathbf{A}$ and $\mathbf{B}$ be elements of $\mathscr{L}(H, H)$. Let A and B be the maps of $L_{2P}(G, H)$ into itself given by $(Ax)(g) = \mathbf{A}x(g)$ and $(Bx)(g) = \mathbf{B}x(g)$, respectively. Suppose that* (i) *the relation $\Phi_1 = (A\Phi + I)(B\Phi + I)^{-1}$ on $L_{2P}(G, H)$ is strongly positive with respect to P and has finite gain*; (ii) *the relation $(B\Phi + I)^{-1}$ on $L_{2P}(G, H)$ has finite gain*; *and* (iii) *the relation $\Phi_2 = (B - \Psi)(\Psi - A)^{-1}$ on $L_{2P}(G, H)$ is positive with respect to P. Then the system $e(\cdot) + \Psi((\Phi e)(\cdot) + x_1(\cdot)) = x_2(\cdot)$ is L_2-stable over every $S \subset L_2(G, H) \times L_2(G, H)$.*

Proof. We observe that $\Phi_1' = -(A\Phi + I)(-B\Phi - I)^{-1}$ is strongly positive and of finite gain since Φ_1 is. Thus, by virtue of Lemma 3.2, it will be sufficient to establish the following claim.

CLAIM. *If the inclusion equation*

$$(16)\qquad x_2' - e' \in \Phi_2[\Phi_1'[e'] + x_1']$$

is L_2-stable over every $S' \subset L_2(G, H) \times L_2(G, H)$, then the system

$$(17)\qquad e(\cdot) + \Psi((\Phi e)(\cdot) + x_1(\cdot)) = x_2(\cdot)$$

is L_2-stable over every $S \subset L_2(G, H) \times L_2(G, H)$.

Now, to verify the claim, we first observe that if x_1, x_2, e satisfy (16), then x_1', x_2', e' satisfy (15), where $x_1' = x_2 - Ax_1$, $x_2' = Bx_1 - x_2$ and $e' = -(B\Phi + I)e$. Moreover, if S is a subset of $L_2(G, H) \times L_2(G, H)$, then $S' = \{(x_1', x_2') : x_1' = x_2 - Ax_1, x_2' = Bx_1 - x_2\}$ is also a subset of $L_2(G, H) \times L_2(G, H)$. Since (15) is L_2-stable over S', it follows that there is a $K_{s'}(\cdot)$ such that $\|x_1'\|_2 \leqq a'$, $\|x_2'\| \leqq a'$

and $(x'_1, x'_2) \in S'$ together imply that $\|e'\|_2 \leqq K_{s'}(a')$. However, given $k > 0$, $\|x_1\|_2 \leqq k$, $\|x_2\|_2 \leqq k$ and $(x_1, x_2) \in S$ together imply that $\|x'_1\|_2 \leqq k'$, $\|x'_2\|_2 \leqq k'$ and $(x'_1, x'_2) \in S'$, where $k' = k \max(1 + \|\mathbf{A}\|, 1 + \|\mathbf{B}\|)$, and hence, that $\|e'\|_2 \leqq K_{s'}(k')$. But $\|e\|_2 \leqq \nu((B\Phi + I)^{-1})\|e'\|_2$, and so, letting

$$K_s(k) = \nu((B\Phi + I)^{-1})K_{s'}(k'),$$

we find that (16) is L_2-stable over S.

COROLLARY 3.4. *Suppose that, in addition to the assumptions of Lemma* 3.3, Φ *has finite gain. Then the system* $e(\cdot) + \Phi((\Psi e)(\cdot) + x_1(\cdot)) = x_2(\cdot)$ *is* L_2*-stable over every* $S \subset L_2(G, H) \times L_2(G, H)$.

Proof. Let Φ' be the map of $L_{2P}(G, H)$ into itself given by $\Phi'x = -\Phi(-x)$. Then $\nu(\Phi') = \nu(\Phi) < \infty$ and the system $e(\cdot) + \Phi((\Psi e)(\cdot) + x_1(\cdot)) = x_2(\cdot)$ is the same as the system $e(\cdot) - \Phi'(-(\Psi e)(\cdot) - x_1(\cdot)) = x_2(\cdot)$. Letting $e' = -\Psi e - x_1$, we deduce that $e'(\cdot) + \Psi((\Phi' e')(\cdot) + x_2(\cdot)) = -x_1(\cdot)$, whenever $e(\cdot)$ is a solution of our original system. But Lemma 3.3 applies to the system $e'(\cdot) + \Psi((\Phi' e')(\cdot) + x_2(\cdot)) = -x_1(\cdot)$ since $(A\Phi' + I)(B\Phi' + I)^{-1}$ is strongly positive with respect to P and has finite gain (as $(A\Phi + I)(B\Phi + I)^{-1}$ does) and since $\nu((B\Phi' + I)^{-1}) = \nu((B\Phi + I)^{-1})$. Thus, given $S \subset L_2(G, H) \times L_2(G, H)$, there is a $K'_s(\cdot)$ such that $\|x_1\|_2 \leqq k$, $\|x_2\|_2 \leqq k$ and $(x_1, x_2) \in S$ together imply that $\|e'\|_2 \leqq K'_s(k)$. It follows that $\|\Psi e\|_2 \leqq \|e'\|_2 + \|x_1\|_2 \leqq k + K'_s(k)$ and that

$$\|e\|_2 \leqq \|x_2\|_2 + \nu(\Phi')(\|\Psi e\|_2 + \|x_1\|_2) \leqq k + \nu(\Phi')(2k + K'_s(k)) = K_s(k).$$

The corollary is now established.

4. A spectral theory. We shall combine Lemma 3.3 with several "frequency domain" conditions for positivity in § 5. Here, we use the generalized transform theory of [1] to obtain the requisite positivity conditions. We begin with a brief review of the relevant results of [1]. We let K be a separable complex Hilbert space (which will, in much of the sequel, represent the complexification H^c of H).

Now, we recall that $B(K)$ is the set of all linear transformations of $L_2(G, K)$ into $L_2(G, K)$ of the form

$$(\mathbf{\Phi} x)(g) = \int_G \Phi(g - g_1)x(g_1)\,d\mu + \lambda x(g), \tag{18}$$

where $\Phi(\cdot) \in L_1(G, \mathscr{L}(K, K))$ and $\lambda \in C$ (cf. [1]). It is shown in [1] that $B(K)$ is a Banach algebra with respect to composition and the norm, $\|\mathbf{\Phi}\|_B = \|\Phi(\cdot)\|_1 + |\lambda|$. Moreover, $B(K)$ and $L_1(G, \mathscr{L}(K, K)) \oplus \{\Delta\}$ (Δ a unit) are isometrically isomorphic, and so we write elements $\mathbf{\Phi}$ of $B(K)$ in the form $\mathbf{\Phi} = \Phi + \lambda\Delta$. We also note that if $\mathbf{\Phi} = \Phi + \lambda\Delta$ is an element of $B(K)$, then the *Fourier transform of* $\mathbf{\Phi}$, $\mathbf{\Phi}(\cdot)$, is the map of $\hat{G}$ into $\mathscr{L}(K, K)$ given by $\mathbf{\Phi}(\gamma) = \hat{\Phi}(\gamma) + \lambda I$, where

$$\hat{\Phi}(\gamma) = \int_G \overline{(\gamma, g)}\Phi(g)\,d\mu \tag{19}$$

and (γ, g) denotes the action of γ on G. The Fourier transform is a uniformly continuous element of $\mathscr{C}(\hat{G}, \mathscr{L}(K, K))$ and $(\mathbf{\Phi\Psi})\hat{} = \hat{\mathbf{\Phi}}\hat{\mathbf{\Psi}}$(A hat accent to the right indicates the Fourier transform of the entire enclosed expression.)

We let $B_P(K)$ be the subset of $B(K)$ given by

$$(20) \qquad B_P(K) = \{\boldsymbol{\Phi} = \Phi + \lambda\Delta \in B(K) : \operatorname{supp}\Phi \subset P\},$$

where $\operatorname{supp}\Phi$ is the support of Φ, i.e., $\operatorname{supp}\Phi = \overline{\{g : \Phi(g) \neq 0\}}$, and P is our closed semigroup of positive Haar measure. $B_P(K)$ is a closed subalgebra of $B(K)$ and is called the *causal subalgebra of* $B(K)$ *with respect to* P. For our purposes, we need to view $B_P(K)$ as a subalgebra of an algebra rather different from $B(K)$. With this in mind, we let $W_P(K)$ be the set of linear maps $\mathbf{Z}$ of $L_2(G, K)$ into $L_2(G, K)$ such that (i) $\mathbf{Z}$ is causal with respect to P, and (ii) $(\widehat{(\mathbf{Z}x)})(\gamma) = z(\gamma)\hat{x}(\gamma)$ for all x in $L_2(G, K)$ and γ in $\hat{G}$, where $z(\cdot)$ is a bounded uniformly continuous element of $\mathscr{C}(\hat{G}, \mathscr{L}(K, K))$. $W_P(K)$ is a Banach algebra with respect to composition and the norm, $\|\mathbf{Z}\|_{W_P} = \sup_{\gamma \in \hat{G}} \{\|z(\gamma)\|\}$. We observe (cf. [1]) that $W_P(K)$ has an identity $\boldsymbol{\Delta}$ given by $\boldsymbol{\Delta}x = x$ and that $B_P(K)$ is a subalgebra of $W_P(K)$. However, $B_P(K)$ is not closed in $\|\cdot\|_{W_P}$ and so, we let $\overline{B_P(K)}$ denote the closure of $B_P(K)$ in $W_P(K)$.

We let $L_1(G, P; C) = \{f \in L_1(G, C) : \operatorname{supp} f \subset P\}$. Then $L_1(G, P; C)$ is, as is well known, a commutative Banach algebra under convolution. We let $L_P = L_1(G, P; C) \oplus \{\delta\}$, where δ is an identity. Then the Gelfand theory applies to L_P. So, we let $\mathscr{M}$ be the maximal ideal space of L_P and we denote the Gelfand representation of $f \in L_P$ by $\hat{f}(M)$. In [1], we extended the Gelfand representation of L_P to a continuous homorphism of $B_P(K)$ (in fact, $\overline{B_P(K)}$) into $\mathscr{B}(\mathscr{M}, \mathscr{L}(K, K))$, where $\mathscr{B}(\mathscr{M}, \mathscr{L}(K, K))$ is the space of bounded maps of $\mathscr{M}$ into $\mathscr{L}(K, K)$. The extended Gelfand representation was the map $\boldsymbol{\Phi} \to \hat{\boldsymbol{\Phi}}(M) = \hat{\Phi}(M) + \lambda I$, where the definition of $\hat{\Phi}(M)$ involved a consideration of the bilinear map $T_M(k_1, k_2) = \langle \Phi(\cdot)k_1, k_2 \rangle (M)$ (see [1] for details).

For the sake of exposition, we considered the case where $L_1(G, \mathscr{L}(K, K))$ does not contain an identity for convolution, i.e., G is not discrete. In other words, we treated the case where it is necessary to adjoin the unit Δ. If G is discrete so that $L_1(G, \mathscr{L}(K, K))$ already contains an identity, then $B(K)$ consists of the convolution maps $\boldsymbol{\Phi}x = \Phi * x$. $B(K)$ is then a Banach algebra with respect to composition and the norm, $\|\boldsymbol{\Phi}\|_B = \|\Phi(\cdot)\|_1$, which is isometrically isomorphic to $L_1(G, \mathscr{L}(K, K))$. Letting $L_{1P}(G, \mathscr{L}(K, K)) = \{\Phi(\cdot) \in L_1(G, \mathscr{L}(K, K)) : \operatorname{supp}\Phi \subset P\}$, we then define $B_P(K)$ by setting $B_P(K) = L_{1P}(G, \mathscr{L}(K, K))$ or $B_P(K) = L_{1P}(G, \mathscr{L}(K, K)) \oplus \{\Delta\}$ according as $0 \in P$ or $0 \notin P$. Similar remarks apply to L_P. We leave the details of the case G discrete to the reader noting that all the subsequent results remain true verbatim (with the proper definitions).

Now recall [1], the notion of approximability and its basic property.

DEFINITION 4.1.[3] Let $\boldsymbol{\Phi}$ be an element of $B_P(K)$ (or $\overline{B_P(K)}$) and let $\{e_1, \cdots\}$ be an orthonormal basis of K. Let K_n be the span of $\{e_1, \cdots, e_n\}$ and let E_n be the projection of K onto K_n. Then $\boldsymbol{\Phi}_n = E_n\boldsymbol{\Phi}E_n$ is an element of $B_P(K)$ (or $\overline{B_P(K)}$) and $\boldsymbol{\Phi}$ is called *approximable* if $\hat{\boldsymbol{\Phi}}_n(M)$ converges to $\hat{\boldsymbol{\Phi}}(M)$ uniformly on $\mathscr{M}$.

PROPOSITION 4.2.[4] *An element* $\boldsymbol{\Phi}$ *of* $B_P(K)$ (*or* $\overline{B_P(K)}$) *is approximable if and only if each* $\hat{\boldsymbol{\Phi}}(M)$ *is a completely continuous element of* $\mathscr{L}(K, K)$ *and the map* $M \to \hat{\boldsymbol{\Phi}}(M)$ *is continuous on* $\mathscr{M}$.

[3] See Definition 5.5 of [1].

[4] See Proposition 5.6 of [1]. In view of the proposition, the notion of approximability is intrinsic.

We used the notion of approximability and the algebras $B_P(K)$, $\overline{B_P(K)}$ and $W_P(K)$ in the following spectral theorem which was proved in [1].

THEOREM 4.3. *If* $\mathbf{Z}$ *is an approximable element of* $\overline{B_P(K)}$, *then*

$$\mathrm{SPEC}_{W_P}\mathbf{Z} \subset \bigcup_{M\in\mathcal{M}} \operatorname{spec}\hat{\mathbf{Z}}(M) = \mathrm{SPEC}_{\bar{B}_P}\mathbf{Z} \subset \mathrm{SPEC}_{B_P}\mathbf{Z}, \tag{21}$$

and[5]

$$\{0\}\cup \bigcup_{\gamma\in\hat{G}} \operatorname{spec} z(\gamma) = \mathrm{SPEC}_{\bar{B}}\mathbf{Z} \subset \bigcup_{M\in\mathcal{M}} \operatorname{spec}\hat{\mathbf{Z}}(M),$$

where the "SPEC's" *and* "spec's" *are suitable spectrums and* $z(\cdot)$ *is the element of* $\mathscr{C}(\hat{G}, \mathscr{L}(K, K))$ *corresponding to* $\mathbf{Z}$.

We then employed this theorem to show that analytic functions $f(\cdot)$ of elements of $\overline{B_P(K)}$ may be defined and that $(f(\mathbf{Z})x)\hat{}(\gamma) = f(z(\gamma))\hat{x}(\gamma)$ for all x in $L_2(G, K)$. More precisely, we actually showed that if $\mathbf{\Phi} = \mathbf{Z} + \lambda\Delta$, where $\mathbf{Z}$ is an approximable element of $\overline{B_P(K)}$ and $f(\xi)$ is analytic on a domain $\mathscr{D}$ containing $\bigcup_{M\in\mathcal{M}} \operatorname{spec}\hat{\mathbf{\Phi}}(M)$ in its interior, then

$$f(\mathbf{\Phi}) = \frac{1}{2\pi i}\int_{\Gamma} f(\xi)(\xi\Delta - \mathbf{\Phi})^{-1}\, d\xi \tag{22}$$

represents a well-defined element of $W_P(K)$ (in fact of $\overline{B_P(K)}$) for any Jordan curve Γ with $\bigcup_{M\in\mathcal{M}} \operatorname{spec}\hat{\mathbf{\Phi}}(M)$ inside Γ. Such analytic functions $f(\mathbf{\Phi})$ play an important role in the sequel. Thus our review of [1] is complete.

Now suppose that T is an element of $\mathscr{L}(K, K)$ and that T is normal. Then, $f(T)$, as given by

$$f(T) = \int_{\operatorname{spec} T} f(v)\, dE(v), \tag{23}$$

where $E(v)$ is the spectral measure corresponding to T, is a well-defined element of $\mathscr{L}(K, K)$ for any function f which is continuous on a domain containing $\operatorname{spec} T$ (see [6]). We then have the following lemma.

LEMMA 4.4. *Suppose that* $f_1(v) = 2Re\{f(v)\} = f(v) + \overline{f(v)} \geqq \delta > 0$ *for all* v *in* $\operatorname{spec} T$. *Then*

$$\langle\{f(T) + f(T)^*\}k, k\rangle \geqq \delta|k|^2 \tag{24}$$

for all k *in* K.

Proof. Since $f(T)^* = \int_{\mathrm{spec}} \overline{f(v)}\, dE(v)$, (see [6]), we have $f(T) + f(T)^* = \int_{\operatorname{spec} T} f_1(v) dE(v) = f_1(T)$. It follows that $f_1(T)$ is self-adjoint and also that $\operatorname{spec} f_1(T) \subset [\delta, \infty)$ since $\operatorname{spec} f_1(T) = f_1(\operatorname{spec} T)$, (see [6]). Hence, $\langle f_1(T)k, k\rangle \geqq \delta|k|^2$ for all k in K which establishes the lemma.

[5] B is the "completion" of $B(K)$ with respect to the norm, $\sup_{\gamma\in\hat{G}}\{\|z(\gamma)\|\}$. The equality follows from the fact that, for every γ in $\hat{G}$, there is a regular maximal ideal N in $L_1(G, C)$ such that $\hat{q}(\gamma) = \hat{q}(N)$ for all q in $L_1(G, C)$ and conversely. The inclusion follows from the fact that, for every γ in $\hat{G}$, there is an $M \in \mathcal{M}$ such that $\hat{r}(\gamma) = \hat{r}(M)$ for all r in L_P (see [4]).

We shall suppose from now on that $K = H^c$ is the complexification of H and consequently, we write B_P, B, W_P in place of $B_P(H^c)$, $B(H^c)$, $W_P(H^c)$, respectively. We also note that if x is an element of $L_2(G, H)$, then x is an element of $L_2(G, H^c)$ with $x(g) = \bar{x}(g)$ for all g (and conversely). Thus we view $L_2(G, H)$ as the set of "real" elements of $L_2(G, H^c)$. We are now ready to derive the requisite positivity conditions. In particular, we have the next lemma.

LEMMA 4.5. *Let* $\mathbf{Z}$ *be an approximable element of* $\bar{\bar{B}}_P$ *and let* f *be any function which is analytic in a symmetric domain* $\mathscr{D}$ (i.e., $\mathscr{D} = \overline{\mathscr{D}}$) *containing* $\bigcup_{M \in \mathscr{M}} \operatorname{spec} \hat{\mathbf{Z}}(M)$ *in its interior. Suppose that* (i) $z(\gamma)$ *is normal for all* γ *in* $\hat{G}$, *where* $z(\cdot)$ *is the element of* $\mathscr{C}(\hat{G}, \mathscr{L}(H^c, H^c))$ *corresponding to* $\mathbf{Z}$; (ii) $\mathbf{Z}$ *is real* (*i.e.*, $(\mathbf{Z}x)(g) = \overline{(\mathbf{Z}x)(g)}$ *for all* g *in* G *and* x *in* $L_2(G, H)$); (iii) $f(\overline{\xi}) = \overline{f(\xi)}$ *for all* ξ *in* $\mathscr{D}$ *so that* $f(\mathbf{Z}) = \overline{f(\mathbf{Z})}$, *where* $f(\mathbf{Z})$ *is given by* (22); *and* (iv) $\operatorname{Re}\{f\} > 0$ *on* $\{0\} \cup \bigcup_{\gamma \in \hat{G}} \operatorname{spec} z(\gamma)$. *Then* $f(\mathbf{Z})$ *is strongly positive on* $L_2(G, H)$.

Proof. Since $\{0\} \cup \bigcup_{\gamma \in \hat{G}} \operatorname{spec} z(\gamma) = \mathrm{SPEC}_{\bar{B}}\mathbf{Z}$ (cf. (21)) is compact, there is a $\delta > 0$ such that $2\operatorname{Re}\{f(v)\} \geqq \delta > 0$ for $v \in \{0\} \cup \bigcup_{\gamma \in \hat{G}} \operatorname{spec} z(\gamma)$. It then follows from the normality of $z(\gamma)$ and Lemma 4.4 that

$$\langle\{f(z(\gamma)) + f(z(\gamma))^*\}k, k\rangle \geqq \delta|k|^2 \tag{25}$$

for all k in H^c. Now, by virtue of the Plancherel theorem [1, Corollary 3.3], we have

$$\int_G \langle f(\mathbf{Z})x(g), x(g)\rangle \, d\mu = \int_{\hat{G}} \langle (f(\mathbf{Z})x)\hat{}(\gamma), \hat{x}(\gamma)\rangle \, dm, \tag{26}$$

and, in view of Proposition 6.1 of [1],

$$\int_G \langle f(\mathbf{Z})x(g), x(g)\rangle \, d\mu = \int_{\hat{G}} \langle f(z(\gamma))\hat{x}(\gamma), \hat{x}(\gamma)\rangle \, \mathrm{dm}. \tag{27}$$

Now, $z(-\gamma) = \overline{z(-\gamma)}$ in view of the fact that $\mathbf{Z}$ is real and the fact that $\hat{x}(\gamma) = \overline{\hat{x}(-\gamma)}$ for x in $L_2(G, H)$. Since the right-hand side of (27) is invariant under the substitution $\gamma \to -\gamma$, we have

$$\int_{\hat{G}} \langle f(z(\gamma))\hat{x}(\gamma), \hat{x}(\gamma)\rangle \, dm = \int_{\hat{G}} \langle f(\overline{z(\gamma)})\overline{\hat{x}(\gamma)}, \overline{\hat{x}(\gamma)}\rangle \, dm \tag{28}$$

for x in $L_2(G, H)$. But $f(\xi) = \overline{f(\xi)}$, so that

$$\begin{aligned} \int_{\hat{G}} \langle f(\overline{z(\gamma)}\,\overline{\hat{x}(\gamma)}, \overline{\hat{x}(\gamma)}\rangle dm &= \int_{\hat{G}} \langle \hat{x}(\gamma), f(z(\gamma))\hat{x}(\gamma)\rangle \, dm \\ &= \int_{\hat{G}} \langle f(z(\gamma)) * \hat{x}(\gamma), \hat{x}(\gamma)\rangle \, dm \end{aligned} \tag{29}$$

for all x in $L_2(G, H)$. It follows from (27) and (29) that

$$\tag{30} \int_G \langle f(\mathbf{Z})x(g), x(g)\rangle\, d\mu = \frac{1}{2}\int_{\hat{G}} \langle\{f(z(\gamma)) + f(z(\gamma))^*\}\hat{x}(\gamma)\rangle\, dm$$
$$\geqq \frac{\delta}{2}\int_{\hat{G}} |\hat{x}(\gamma)|^2\, dm = \frac{\delta}{2}\|x(\cdot)\|_2^2,^{6}$$

and, hence, the lemma is established.

COROLLARY 4.6. *Let* $\mathbf{\Phi} = \mathbf{Z} + \lambda\Delta$, *where* $\mathbf{Z}$ *is an approximable element of* $\bar{B}_P$, *and let* f *be any function which is analytic in a symmetric domain* $\mathscr{D}$ *containing* $\bigcup_{M\in\mathscr{M}} \operatorname{spec} \hat{\mathbf{\Phi}}(M)$ *in its interior. Suppose that* $\mathbf{\Phi}$ *and* f *satisfy the* (*analogue of*) *conditions* (i), (ii) *and* (iii) *of the lemma and that* $\operatorname{Re}\{f\} > 0$ *on* $\{+\lambda\} \cup \bigcup_{\gamma\in\hat{G}} \operatorname{spec} \phi(\gamma)$ *where* ϕ *is the element of* $\mathscr{C}(\hat{G}, \mathscr{L}(H^c, H^c))$ *corresponding to* $\mathbf{\Phi}$. *Then* $f(\mathbf{\Phi})$ *is strongly positive on* $L_2(G, H)$.

COROLLARY 4.7. *Let* $\mathbf{Z}$ *be an approximable element of* $\bar{B}_P$. *Suppose that conditions* (i) *and* (ii) *of the lemma are satisfied and that a and b are positive real numbers with* $0 < a < b$. *If* $-b^{-1} \notin \bigcup_{M\in\mathscr{M}} \operatorname{spec} \hat{\mathbf{Z}}(M)$ *and if the set* $\{0\} \cup \bigcup_{\gamma\in\hat{G}} \operatorname{spec} z(\gamma)$ *does not intersect the circle with center* $-\frac{1}{2}(1/a + 1/b)$ *and radius* $\frac{1}{2}(1/a - 1/b)$,[7] *then* $f(\mathbf{Z}) = (a\mathbf{Z} + \Delta)(b\mathbf{Z} + \Delta)^{-1}$ *is a well-defined element of* $\bar{B}_P$ *such that* $f(\mathbf{Z}) = \overline{f(\mathbf{Z})}$ *and* $f(\mathbf{Z})$ *is strongly positive on* $L_2(G, H)$.

Proof. Let $f(\xi) = (a\xi + 1)/(b\xi + 1)$ and let $\mathscr{D}$ be a symmetric domain such that $-b^{-1} \notin \mathscr{D}$ and such that the *compact set* $\{0\} \cup \bigcup_{\gamma\in\hat{G}} \operatorname{spec} z(\gamma) = \operatorname{SPEC}_{\bar{B}} \mathbf{Z}$ is contained in the interior of $\mathscr{D}$ ($\mathscr{D}$ exists since $-b^{-1} \notin \bigcup_{M\in\mathscr{M}} \operatorname{spec} \hat{\mathbf{Z}}(M)$). Then f is analytic on $\mathscr{D}$ and $f(\xi) = \overline{f(\xi)}$ for all ξ in $\mathscr{D}$. Moreover, we can see immediately that the hypotheses insure that $\operatorname{Re}\{f(v)\} > 0$ for all v in $\{0\} \cup \bigcup_{\gamma\in\hat{G}} \operatorname{spec} z(\gamma)$. Thus the corollary follows from the lemma.

COROLLARY 4.8. *Let* $\mathbf{\Phi} = \mathbf{Z} + \lambda\Delta$, *where* $\mathbf{Z}$ *is an approximable element of* $\bar{B}_P$. *Suppose that* $\mathbf{\Phi}$ *satisfies* (i) *and* (ii) *of the lemma and that a and b are positive real numbers with* $0 < a < b$. *If* $-b^{-1} \notin \bigcup_{M\in\mathscr{M}} \operatorname{spec} \hat{\mathbf{\Phi}}(M)$ *and if the set* $\{+\lambda\} \cup \bigcup_{\gamma\in\hat{G}} \operatorname{spec} \phi(\gamma)$ (*where* ϕ *is the element of* $\mathscr{C}(\hat{G}, \mathscr{L}(H^c, H^c))$ *corresponding to* $\mathbf{\Phi}$) *does not intersect the circle with center* $-\frac{1}{2}(1/a + 1/b)$ *and radius* $\frac{1}{2}(1/a - 1/b)$, *then* $f(\mathbf{\Phi}) = (a\mathbf{\Phi} + \Delta)(b\mathbf{\Phi} + \Delta)^{-1}$ *is a well-defined element of* $\bar{B}_P$ *such that* $f(\mathbf{\Phi}) = \overline{f(\mathbf{\Phi})}$ *and* $f(\mathbf{\Phi})$ *is strongly positive on* $L_2(G, H)$.

5. A generalized circle criterion. We state and prove our main result, the generalized circle criterion for stability, in this section. We begin with some simple propositions.

PROPOSITION 5.1. *Let* N *be a map of* H *into* H *and let* a *and* b *be positive real numbers with* $0 < a < b$. *Suppose that* $\langle bh - N(h), N(h) - ah\rangle \geqq 0$ *for all* h *in* H. *Then* N *is bounded, i.e., there is a* c *with* $|N(h)| \leqq c|h|$ *for all* h *in* H.

Proof. Since $\langle bh - N(h), N(h) - ah\rangle \geqq 0$, we have $\langle (b + a)h, N(h)\rangle \geqq \langle bh, ah\rangle + \langle N(h), N(h)\rangle \geqq 0$ so that $(b + a)|h| \cdot |N(h)| \geqq |N(h)|^2$.

PROPOSITION 5.2. *Let* $\mathbf{Z}$ *be an element of* W_P. *Then* $\mathbf{Z}$ *is of finite gain and has a unique causal extension* $\mathbf{Z}'$ *to* $L_{2P}(G, H)$.

[6] See [1, Proposition 3.2].

[7] As regards $\{0\}$, this is automatically true.

Proof. Simply observe that $\|(\mathbf{Z}x)(\cdot)\|_2 = \|(\mathbf{Z}x)\hat{\ }(\cdot)\|_2$, by virtue of [1] Corollary 3.3, which implies that $\|(\mathbf{Z}x)(\cdot)\|_2 \leqq \|\mathbf{Z}\|_{W_P}\|\hat{x}(\cdot)\|_2 = \|\mathbf{Z}\|_{W_P}\|x(\cdot)\|_2$ for all x in $L_2(G, H^c)$. Then apply Proposition 2.9.

We now have the next theorem.

THEOREM 5.3. *Let* $\mathbf{\Phi} = \mathbf{Z} + \lambda\Delta$ *where* $\mathbf{Z}$ *is an approximable element of* $\bar{B}_P$. *Suppose that* (i) $\phi(\gamma)$ *is normal for all* γ *in* $\hat{G}$; (ii) $\mathbf{\Phi}$ *is real*; (iii) *a and b are real numbers with* $0 < a < b$; (iv) *N is a map of H into H such that* $\langle bh - N(h), N(h) - ah\rangle \geqq 0$ *for all h*; (v) $-b^{-1} \notin \bigcup_{M \in \mathscr{M}} \operatorname{spec} \hat{\mathbf{\Phi}}(M)$; *and* (vi) *the set* $\{+\lambda\} \cup \bigcup_{\gamma \in \hat{G}} \operatorname{spec} \phi(\gamma)$ *does not intersect the circle with center* $-\frac{1}{2}(1/a + 1/b)$ *and radius* $\frac{1}{2}(1/a - 1/b)$ (*where* ϕ *is the element of* $\mathscr{C}(\hat{G}, \mathscr{L}(H^c, H^c))$ *corresponding to* $\mathbf{\Phi}$). *Then the systems*

$$e(\cdot) + \mathbf{\Phi}'(\mathbf{N}e(\cdot) + x_1(\cdot)) = x_2(\cdot), \tag{31}$$

$$e(\cdot) + \mathbf{N}(\mathbf{\Phi}'e(\cdot) + x_1(\cdot)) = x_2(\cdot) \tag{32}$$

are both L_2-*stable over any* $S \subset L_2(G, H) \times L_2(G, H)$. (*Here* $(\mathbf{N}x)(g) = N(x(g))$.)

Proof. Let $\mathbf{\Phi}_1 = (a\mathbf{\Phi} + \Delta)(b\mathbf{\Phi} + \Delta)^{-1}$. Then $\mathbf{\Phi}_1$ is an element of $\bar{B}_P$ and $\mathbf{\Phi}_1$ is real by virtue of (v) and Corollary 4.8. It follows from Propositions 2.9 and 5.2 that $\mathbf{\Phi}_1$ has a unique causal extension $\mathbf{\Phi}_1'$ mapping $L_{2P}(G, H)$ into $L_{2P}(G, H)$ and that $\mathbf{\Phi}_1'$ has finite gain.[8]

Now we claim that $\mathbf{\Phi}_1'$ is strongly positive. To verify this claim, we observe that $\mathbf{\Phi}_1$ is strongly positive in view of Corollary 4.8 and that $\langle x_{g_0}(\cdot), (\mathbf{\Phi}_1'x)_{g_0}(\cdot)\rangle = \langle x_{g_0}(\cdot), (\mathbf{\Phi}_1'x_{g_0}(\cdot))_{g_0}\rangle = \langle x_{g_0}(\cdot), \mathbf{\Phi}_1'x_{g_0}(\cdot)\rangle = \langle x_{g_0}(\cdot), \mathbf{\Phi}_1 x_{g_0}(\cdot)\rangle$ for all x in $L_{2P}(G, H)$ and g_0 in G. Thus, $\mathbf{\Phi}_1'$ satisfies condition (i) of Lemma 3.3.

Since $(b\mathbf{\Phi} + \Delta)^{-1}$ is an element of $\bar{B}_P$ by virtue of (v), we deduce that the relation $(b\mathbf{\Phi}' + \Delta)^{-1}$ has finite gain, and so condition (ii) of Lemma 3.3 is satisfied.

As for condition (iii) of Lemma 3.3, we note that $(bI - \mathbf{N})(\mathbf{N} - aI)^{-1}$ is a positive relation on $L_2(G, H)$ by virtue of (iv) and that, therefore, $(bI - \mathbf{N}')(\mathbf{N}' - aI)^{-1}$ is positive where $\mathbf{N}'$ is the unique causal extension of $\mathbf{N}$. Thus, the system (32) is L_2-stable over every $S \subset L_2(G, H) \times L_2(G, H)$ by Lemma 3.3. Since $\mathbf{\Phi}'$ is of finite gain, the system (31) is also L_2-stable over every $S \subset L_2(G, H) \times L_2(G, H)$ (Corollary (3.4), and the theorem is established.

6. Examples and comments. We now focus our attention on a number of examples which illustrate the theory that we have developed.

Example 6.1. Let $G = R$ be the real numbers and let $H = R_2$ be two-dimensional Euclidean space. We consider the system of nonlinear differential equations

$$\begin{aligned} \ddot{y}_1 + 6\dot{y}_1 + 5y_1 + 2\dot{y}_2 + 2y_2 &= y_1(d + \cos y_2), \\ 2\dot{y}_1 + 2y_1 + \ddot{y}_2 + 3\dot{y}_2 + 2y_2 &= cy_2, \end{aligned} \tag{33}$$

where c and d are positive constants. We suppose that the initial conditions for (33) are given by

$$y_1(0) = \alpha, \quad \dot{y}_1(0) = \alpha', \quad y_2(0) = \beta, \quad \dot{y}_2(0) = \beta', \tag{34}$$

where α, α', β, β' are appropriate constants. We shall find sufficient conditions for $y_1(t)$ and $y_2(t)$ to be elements of $L_2[0, \infty)$.

[8] It is easy to see that $\mathbf{\Phi}_1' = (a\mathbf{\Phi}' + \Delta)(b\mathbf{\Phi}' + \Delta)^{-1}$.

Letting $\mathbf{y}(t) = (y_1(t), y_2(t))'$ and letting N be the map of R_2 into itself given by

$$N\left(\begin{bmatrix} r_1 \\ r_2 \end{bmatrix}\right) = \begin{bmatrix} r_1(d + \cos r_2) \\ cr_2 \end{bmatrix}, \tag{35}$$

we may rewrite (33) as a nonlinear integral equation of the form

$$\mathbf{y}(t) = \mathbf{y}_0(t) + \int_0^t \boldsymbol{\varphi}(t - \tau)N(\mathbf{y}(\tau))\,d\tau, \qquad t \geqq 0, \tag{36}$$

where $\mathbf{y}_0(t)$ depends only on the initial data (34) and where $\boldsymbol{\varphi}(t)$ is the appropriate Green's function for (33). We note that $\mathbf{y}_0(t)$ is an element of $L_2[0, \infty)$ and that the map Φ which corresponds to $\boldsymbol{\varphi}$ has the "complex Laplace transform"

$$\hat{\Phi}(\sigma + i\gamma) = \frac{1}{(\sigma + i\gamma + 1)^2(\sigma + i\gamma + 6)}\begin{bmatrix} \sigma + i\gamma + 2 & -2 \\ -2 & \sigma + i\gamma + 5 \end{bmatrix}, \tag{37}$$

where $\sigma + i\gamma \in C$ and $\sigma \geqq 0$. Now, letting $P = [0, \infty)$, we can see that $\Phi(t)$ is an element of $B_{[0,\infty)}$ and that $\hat{\Phi}(\sigma + i\gamma)$ is normal for all $\sigma + i\gamma$ in C with $\sigma \geqq 0$. We also note that Φ is real and that Φ is automatically approximable since the Hilbert space H $(= R_2)$ is finite-dimensional. Thus, conditions (i) and (ii) of Theorem 5.3 are satisfied.

We can easily see that conditions (iii) and (iv) of Theorem 5.3 will be satisfied if a and b are any positive numbers for which $a \leqq c \leqq b$, $d \geqq 1 + a$, and $b \geqq 1 + d$.

Now we note that $L_P = L_{1P}(R, C) \oplus \{\delta\}$ has a maximal ideal space which corresponds to the right half-plane "compactified" and that the Gelfand representation corresponds to the Laplace transform $\hat{\Phi}(s)$ with $\operatorname{Re}\{s\} \geqq 0$. The spectrum of $\hat{\Phi}(s)$ consists of the eigenvalues

$$\lambda_1(s) = \frac{1}{(s + 6)(s + 1)}, \quad \lambda_2(s) = \frac{1}{(s + 1)^2}, \tag{38}$$

for $\operatorname{Re}\{s\} \geqq 0$. It is clear that the equations $\lambda_1(s) = -b^{-1}$, $\lambda_2(s) = -b^{-1}$ have no solutions for $b > 0$, and so condition (v) of Theorem 5.3 holds.

Finally, condition (vi) will be satisfied provided that $\lambda_1(i\omega)$ and $\lambda_2(i\omega)$ do not intersect (and remain outside of) the circle with center $-\frac{1}{2}(1/a + 1/b)$ and radius $\frac{1}{2}(1/a - 1/b)$ since $\{0\}$ does not meet the circle. It follows that, under these conditions, the system (33) will be L_2-stable in the sense that given $A > 0$, there is a $K(A) > 0$ such that $\|\mathbf{y}_0(\cdot)\|_2 \leqq A$ implies that $\|\mathbf{y}(\cdot)\|_2 \leqq K(A)$. From this, we may conclude that $\lim_{t\to\infty} \mathbf{y}(t) = \mathbf{0}$.

Example 6.2. Again let G be the real numbers and let $P = [0, \infty)$. Let $H = L_2[0, 1]$ so that H is a separable infinite-dimensional Hilbert space. We consider the nonlinear integral equation

$$u(x, t) = u_0(x, t) - \int_0^t \int_0^1 \varphi(x, y, t - \tau)N(u(y, \tau))\,dy\,d\tau, \tag{39}$$

where N is any bounded map of $L_2[0, 1]$ into itself. We shall find sufficient conditions for the stability of (39) and shall examine a particular example involving a nonlinear partial differential equation.

We assume that the function $\varphi(x, y, t)$ has the following properties:

$$(40)\qquad \begin{gathered} \int_{-\infty}^{\infty}\left[\int_0^1\int_0^1 |\varphi(x,y,t)|^2 dx\,dy\right]^{1/2} dt < \infty, \\ \int_0^1\int_0^1\left(\int_0^{\infty} |\varphi(x,y,t)| dt\right)^2 dx\,dy < \infty, \end{gathered}$$

$$(41)\qquad \varphi(x,y,t) = \varphi(y,x,t),$$

$$(42)\qquad \varphi(x,y,t) = 0 \quad \text{for t} < 0.$$

We then define a map Φ by setting

$$(43)\qquad (\Phi v)(x,t) = \int_0^t\int_0^1 \varphi(x,y,t-\tau)v(y,\tau)\,dy\,d\tau$$

for v in $L_{2P}(R, L_2[0, 1])$. It is clear that Φ is in $B_{[0,\infty)}$. Moreover, the "complex Laplace transform" $\hat{\Phi}(s)$ of Φ is, for $\operatorname{Re}\{s\} \geqq 0$, given by

$$(44)\qquad \hat{\Phi}(s)w(x) = \int_0^1 \hat{\Phi}(x,y,s)w(y)\,dy,$$

where $\hat{\varphi}(x, y, s) = \int_0^{\infty} e^{-st}\varphi(x, y, t)\,dt$. It follows that the Fourier transform $\hat{\Phi}(i\omega)$ is normal (by (41)) and that $\hat{\Phi}(s)$ is completely continuous for each s with $\operatorname{Re}\{s\} \geqq 0$. As Φ is real, conditions (i) and (ii) of Theorem 5.3 hold and Φ is approximable. Thus, (39) will be L_2-stable if the remaining conditions of Theorem 5.3 are satisfied.

Let us now consider the nonlinear partial differential equation

$$(45)\qquad \frac{\partial^4 u}{\partial t^2 \partial x^2} + 3\frac{\partial^3 u}{\partial t \partial x^2} + 2\frac{\partial^2 u}{\partial x^2} = N(u),$$

with the auxiliary data

$$(46)\qquad \begin{gathered} u(0,t) = u(1,t) = 0, \\ u(x,0) = f_1(x), \qquad \frac{\partial u}{\partial t}(x,0) = f_2(x), \end{gathered}$$

where f_1 and f_2 are elements of $L_2[0, 1]$ and N is a bounded map of $L_2[0, 1]$ into itself with $N(0) = 0$. We reformulate (45) as an integral equation of the form (39). To do this, we let $\Gamma(x, y)$ be the Green's function for the Sturm–Liouville problem on $[0, 1]$ given by

$$(47)\qquad -\frac{d^2 q(x)}{dx^2} = f(x), \qquad q(0) = q(1) = 0,$$

so that

$$(48)\qquad \Gamma(x,y) = \begin{cases} x(1-y), & x < y, \\ (1-x)y, & x > y, \end{cases}$$

for x, y in $[0, 1]$. We also let $\psi(t)$ be the impulse response for the operator $D_t^2 + 3D_t + 2$ so that

$$\psi(t) = \begin{cases} e^{-t} - e^{-2t}, & t \geqq 0, \\ \quad 0 \quad , & t < 0. \end{cases} \tag{49}$$

Then (45) has the equivalent integral form

$$u(x,t) = u_0(x,t) - \int_0^t \int_0^1 \Gamma(x,y)\psi(t-\tau)N(u(y,\tau))\,\mathrm{d}y\mathrm{d}\tau, \tag{50}$$

where $u_0(x, t)$ is the solution of the equation $\{D_t^2 + 3D_t + 2\}u_0(x, t) = 0$ satisfying the initial conditions $u_0(x, 0) = f_1(x)$, $du_0(x, 0)/dt = f_2(x)$. We note that $\int_0^\infty \int_0^1 |u_0(x, t)|^2\, dx\, dt < \infty$. Clearly, (50) has the required form with $\varphi(x, y, t) = \Gamma(x, y)\psi(t)$ and $\varphi(x, y, t)$ satisfies the conditions (40), (41) and (42).

In order to apply Theorem 5.3, we must compute spec $\hat{\Phi}$(s) for Re $\{s\} \geqq 0$ as the Gelfand representation corresponds to the Laplace transform in the case at hand (cf. Example 6.1). Now, $\hat{\varphi}(x, y, s) = \Gamma(x, y)/((s + 1)(s + 2))$, and so

$$\hat{\Phi}(s)w(x) = \frac{1}{(s+1)(s+2)}\int_0^1 \Gamma(x,y)w(y)\,dy$$

for w in $L_2[0, 1]$. Since the operator T given by $(Tw)(x) = \int_0^1 \Gamma(x, y)w(y)\, dy$ is well known to have the spectrum $\{0,\ 1/(n^2\pi^2)\colon\ n = 1,\ 2, \cdots\}$ (as $\Gamma(x, y) = -\ 2\sum_1^\infty(1/(n^2\pi^2)) \sin n\pi x \sin n\pi y$). It follows that

$$\operatorname{spec}\hat{\Phi}(s) = \{0, 1|((s+1)(s+2)n^2\pi^2)\colon n = 1, 2, \cdots\}.$$

Let a and b be positive numbers with $a < b$. Then $-b^{-1} \notin \operatorname{spec} \hat{\Phi}(s)$ for any s with $\operatorname{Re}\{s\} \geqq 0$, and so condition (v) is satisfied. If, in addition, the set $\{0, 1/((i\omega + 2)(i\omega + 2)n^2\pi^2)\colon n = 1, 2, \cdots\}$ does not meet the proper circle, then (vi) will also hold. Thus, the system (45) will be L_2-stable provided that the nonlinearity N satisfies the condition

$$\int_0^1 (bw(x) - N(w(x)))(N(w(x)) - aw(x))\,dx \geqq 0 \tag{51}$$

for all $w(\cdot)$ in $L_2[0, 1]$.

Example 6.3. Let $G = J$ be the integers and let $H = L_2[0, 1]$. Let $P = J^+ = \{0, 1, 2, \cdots\}$ and consider the nonlinear differential difference equation

$$\frac{d^2u(x, n+2)}{dx^2} - \frac{1}{4}\frac{d^2u(x,n)}{dx^2} = N(u(x,n)), \tag{52}$$

with the auxiliary data

$$\begin{gathered} u(0,n) = \frac{du}{dx}(1,n) = 0, \\ u(x,0) = f_0(x), \qquad u(x,1) = f_1(x), \end{gathered} \tag{53}$$

where f_0, and f_1 are elements of $L_2[0, 1]$ and N is a map of $L_2[0, 1]$ into itself with $N(0) = 0$. We wish to determine conditions which insure that $\sum_{n=0}^{\infty} \int_0^1 |u(x, n)|^2 \, dx < \infty$ for all f_0 and f_1.

We let $\gamma(x, n)$ be the solution of the homogeneous version of (52) with the auxiliary data (53). It is easy to check that $\sum_{n=0}^{\infty} \int_0^1 |\gamma(x, n)|^2 dx < \infty$. We reformulate (52) as an operator equation. To do this, we let $\Gamma(x, y)$ be the Green's function for the Sturm-Liouville problem on [0, 1] given by

$$-\frac{d^2q(x)}{dx^2} = f(x), \qquad q(0) = \frac{dq(1)}{dx} = 0, \tag{54}$$

so that

$$\Gamma(x, y) = \begin{cases} x, & x < y, \\ y, & x > y, \end{cases} \tag{55}$$

for x, y in $[0, 1]$. We also let $\{\psi(n)\}$ be the "impulse response" for the operator $E^2 - (\frac{1}{4})I$ on l_2, where E is given by $E[\{a_n\}] = \{a_{n+1}\}$ (i.e., is a shift). Then

$$\psi(n) = \begin{cases} 0, & n \text{ odd, zero or negative,} \\ (\frac{1}{2})^{n-2}, & n \text{ even,} \end{cases} \tag{56}$$

and it follows that (52), (53) has the equivalent representation

$$u(x, n) = \gamma(x, n) - \sum_{k=0}^{n} \int_0^1 \psi(n - k)\Gamma(x, y)N(u(y, k)) \, dy. \tag{57}$$

Now (57) has the desired form and we can use Theorem 5.3 to establish stability.

We let Φ be the element of B_{J^+} given by

$$(\Phi v)(x, n) = \sum_{k=0}^{n} \psi(n - k) \int_0^1 \Gamma(x, y)v(y, k) \, dy, \tag{58}$$

where v is any element of $L_{2P}(J, L_2[0, 1])$. Now the character group $\hat{J}$ of J is the circle group $\{e^{i\theta}: 0 \leqq \theta < 2\pi\}$ under multiplication and it can be shown that the Gelfand representation corresponds, in this case, to the z-transform. Thus, we let $\hat{\Phi}(z)$ denote the z-transform Φ so that

$$(\hat{\Phi}(z)w)(x) = \frac{z^2}{1 - z^2/4} \int_0^1 \Gamma(x, y)w(y) \, dy, \tag{59}$$

for w in $L_2[0, 1]$ and $z \in C$ with $|z| \leqq 1$. Since $\Gamma(x, y) = \Gamma(y, x)$ and $\int_0^1 \int_0^1 |\Gamma(x, y)| \cdot dx \, dy < \infty$, we deduce that $\hat{\Phi}(z)$ is normal and that $\hat{\Phi}(z)$ is completely continuous on $|z| \leqq 1$. It follows that Φ is approximable and that conditions (i) and (ii) of Theorem 5.3 are satisfied.

Now, to determine the spectrum of $\hat{\Phi}(z)$, it will be sufficient to determine the spectrum of the operator T given by $(Tw)(x) = \int_0^1 \Gamma(x, y)w(y) \, dy$ and then to multiply by $z^2/(1 - z^2/4)$. Since the spectrum of T is the set $\{0, 1((n + \frac{1}{2})^2\pi^2):n = 1, 2, \cdots\}$, we have

$$\text{(60)} \qquad \operatorname{spec} \hat{\Phi}(z) = \left\{0, \frac{z^2}{(1 - z^2/4)} \frac{1}{(n + \frac{1}{2})^2\pi^2} : n = 1, 2, \cdots\right\}$$

for $|z| \leqq 1$. Thus, if we suppose that $0 < a < b$, that $-b^{-1} \notin \operatorname{spec} \hat{\Phi}(z)$ and that the set $\bigcup_{0 \leqq \theta < 2\pi} \operatorname{spec} \hat{\Phi}(e^{i\theta})$ does not meet the proper circle, then the system (52), (53) will be L_2-stable provided that

$$\text{(61)} \qquad \int_0^1 [bw(x) - N(w(x))][N(w(x)) - aw(x)]\, dx \geqq 0$$

for all $w(\cdot)$ in $L_2[0, 1]$.

Example 6.4. Let $G = R \oplus R, H = R$ and $P = [0, \infty) \times [0, \infty)$ (the first quadrant). We consider the partial differential equation

$$\text{(62)} \qquad \frac{\partial^2 u}{\partial x \partial y} + c\frac{\partial u}{\partial x} + d\frac{\partial u}{\partial y} + cdu + N(u(x, y)) = 0,$$

with the auxiliary data

$$\text{(63)} \qquad \begin{aligned} u(x, 0) &= f_1(x), & \frac{\partial u}{\partial y}(x, 0) &= f_2(x), \\ u(0, y) &= h_1(y), & \frac{\partial u}{\partial x}(0, y) &= h_2(y), \end{aligned}$$

where $f_1(0) = h_1(0)$ and f_1, f_2, h_1, h_2 are elements of $L_2[0, \infty)$ and N is a map of R into R with $N(0) = 0$. We wish to determine conditions which ensure the stability of (62), (63).

Letting $\Gamma(x, y)$ be the function given by

$$\text{(64)} \qquad \Gamma(x, y) = \begin{cases} e^{-(dx + cy)}, & \\ 0, & x < 0 \text{ or } y < 0, \end{cases}$$

we can rewrite (62) in the form

$$\text{(65)} \quad u(x, y) = u_0(x, y) - \int_0^x \int_0^y \Gamma(x - \xi_1, y - \xi_2) N(u(\xi_1, \xi_2))\, d\xi_1\, d\xi_2,$$

where $u_0(x, y)$ is in $L_2(R \oplus R)$ for $x \geqq 0, y \geqq 0$ and depends on the auxiliary data (63). We define a map Φ in B_P by setting

$$\text{(66)} \qquad (\Phi v)(x, y) = \int_0^x \int_0^y \Gamma(x - \xi_1, y - \xi_2) v(\xi_1, \xi_2)\, d\xi_1\, d\xi_2$$

for v in $L_2(R \oplus R)$. The Fourier transform $\hat{\Phi}(\gamma_1, \gamma_2)$ is then the element of $\mathscr{L}(C, C)$ given by

$$\text{(67)} \quad \hat{\Phi}(\gamma_1, \gamma_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-i(\gamma_1 x + \gamma_2 y)} \Gamma(x, y)\, dx\, dy = \frac{1}{(i\gamma_1 + c)(i\gamma_2 + d)}$$

for $(\gamma_1, \gamma_2) \in \hat{G} = R \oplus R$. Since the maximal ideal space of L_P here corresponds to the set $\{(s_1, s_2) \in C \oplus C : \mathrm{Re}\{s_1\} \geqq 0, \mathrm{Re}\{s_2\} \geqq 0\}$ "compactified," the Gelfand representation corresponds to the Laplace transform and we have

$$\hat{\Phi}(s_1, s_2) = \frac{1}{(s_1 + c)(s_2 + d)} \tag{68}$$

for $\mathrm{Re}\{s_1\} \geqq 0$, $\mathrm{Re}\{s_2\} \geqq 0$. Since our Hilbert space is 1-dimensional, Φ is clearly approximable. Moreover, Φ is real and $\hat{\Phi}(\gamma_1, \gamma_2)$ is normal. Thus, conditions (i) and (ii) of Theorem 5.3 are satisfied.

Now if a and b are positive numbers with $0 < a < b$, then $-b^{-1} \neq \hat{\Phi}(s_1, s_2)$ for all s_1, s_2 with $\mathrm{Re}\{s_1\} \geqq 0$ and $\mathrm{Re}\{s_2\} \geqq 0$ so that condition (v) is satisfied. Thus, the system (62), (63) will be L_2-stable provided that the set $\{\hat{\Phi}(\gamma_1, \gamma_2) : (\gamma_1, \gamma_2) \in R \oplus R\}$ does not intersect the circle with center $-\frac{1}{2}(1/a + 1/b)$ and radius $\frac{1}{2}(1/a - 1/b)$ and that the nonlinearity N satisfies the condition $a \leqq N(\alpha)/\alpha \leqq b$ for all α in R with $\alpha \neq 0$.

These examples serve to illustrate the wide range of applicability of the theory.

Our results involve a generalization of the circle criterion for stability (see, for example, [2]), and thus represent readily usable frequency domain criteria. We also note that a number of other "functional" stability results (for example, see [2], [9]) can be derived in the general context developed here. This will be done in a later paper.

REFERENCES

[1] P. L. Falb and M. I. Freedman, *A generalized transform theory for causal operators*, this Journal, 7 (1969), pp. 452–471.

[2] G. Zames, *On the input-output stability of time-varying nonlinear feedback systems. I, II*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 228–238, 465–476.

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

[4] W. Rudin, *Fourier Analysis on Groups*, Interscience, New York, 1962.

[5] F. Riesz and B. Ss.-Nagy, *Functional Analysis*, Frederick Ungar, New York, 1955.

[6] P. R. Halmos, *Introduction to Hilbert Space and the Theory of Spectral Multiplicity*, Chelsea, New York, 1951.

[7] M. Naimark, *Normed Rings*, P. Noordhoff, Groningen, Netherlands, 1959.

[8] L. Loomis, *An Introduction to Abstract Harmonic Analysis*, Van Nostrand, Princeton, 1953.

[9] G. Zames and P. L. Falb, *Stability conditions for systems with monotone and slope-restricted nonlinearities*, this Journal, 6 (1968), pp. 89–108.

SIAM J. Control
Vol. 7, No. 3, August 1969

GENERALIZED PREDICTION-CORRECTION ESTIMATION*

J. R. KRASNAKEVICH† and R. A. HADDAD‡

1. Introduction. Estimation of the state of a dynamic system in the presence of noise has received considerable attention in recent years. The pioneering works of Swerling [1], [2] and Kalman [3] in recursive smoothing were extended to continuous-time dynamic systems by Kalman [4], [5] and Bucy [5]. This area attracted the attention of Lee [6], [7], Ho [7], Rauch, Tung and Striebel [8] and others who derived solutions to the filtering problem employing a Bayesian approach, least squares and a maximum likelihood formulation. What emerged from all of these efforts were useful but alternate derivations of what has become to be known as the Kalman filter.

Subsequent work by Smith, Schmidt and McGee [9], Mowery [10] and Cox [11] extended the linear Kalman filter results to message processes with nonlinear dynamics by a process of linearization. The resulting filters were dubbed "extended" Kalman filters or "differential-correction" estimators. Common to all these schemes is a two step prediction-correction method. The prediction is formed by updating the most recent estimate through the system dynamics (differential or difference equation); then a linear correction is added to this prediction to form the new estimate of the state.

Recent efforts in nonlinear filtering by Stratonovich [13], Kushner [14], [15], Wonham [16] and Bucy [17] led to the formal derivation of the stochastic partial differential equation for the conditional probability density function. One approximate implementation of this result has been proposed by Jazwinski [18], [19].[1]

In contrast with the preceding methods, we postulate a predictor-corrector filter structure and then determine the optimal corrector to minimize the mean-squared error. In this paper, the conditions for only the discrete-time optimal mean-square corrector for the predictor-corrector structure are obtained. An orthogonality condition is derived which gives necessary and sufficient conditions on the weighting to minimize the mean-squared error. The resultant corrector generally provides a nonlinear weighting of the residuals. The extended Kalman filter can be derived from our results by postulating a priori a linear weighting scheme. A comparison of the continuous-time version [21] of the presented technique for a specified quasilinear weighting with that proposed by Jazwinski [18], [19] shows that the two algorithms are virtually identical. The results reported herein represent a generalized approach to predictor-corrector estimation.

* Received by the editors August 21, 1968, and in revised form January 30, 1969. This work was supported in part by the National Science Foundation. This paper is based on a dissertation submitted by J. R. Krasnakevich to the faculty of the Polytechnic Institute of Brooklyn in partial fulfillment of the requirements for the Ph.D. degree in System Science.

† Bell Telephone Laboratories, Whippany, New Jersey 07981.

‡ Department of Electrical Engineering, Polytechnic Institute of Brooklyn, Brooklyn, New York 11201.

[1] This and other continuous-time filtering techniques are discussed by Schwartz and Stear [20].

2. Problem formulation. The problem under consideration is the estimation of the state of a vector process described by the nonlinear equation

$$z(k) = f(z(k-1); k) + G(k)u(k) \tag{1}$$

and the observed data vector given by

$$x(k) = h(z(k); k) + v(k), \tag{2}$$

where $z(k)$ is the state, an n-vector; f and h are nonlinear vector functions of their arguments, respectively n and m vectors. The random input $u(k)$ and the noise $v(k)$ are l and m vectors respectively. $G(k)$ is a real-valued $n \times l$ matrix. The initial state $z(0)$ is a vector random variable with known statistics. Finally, $x(k)$ is the observed data vector, an m-vector.

We postulate the filter structure described by[2]

$$\hat{z}(k/k) = f(\hat{z}(k-1/k-1); k) + H[\varepsilon(k); k], \tag{3}$$

$$\hat{z}(0/0) = E\{z(0)\}, \tag{4}$$

where the residual, $\varepsilon(k)$, is

$$\varepsilon(k) = x(k) - \hat{x}(k) \tag{5}$$

and

$$\hat{x}(k) = h(\hat{z}(k/k-1); k). \tag{6}$$

The one step prediction $\hat{z}(k/k-1)$ (the prediction of $z(k)$ using the data up to and including $k-1$) is chosen to be

$$\hat{z}(k/k-1) = f(\hat{z}(k-1/k-1); k). \tag{7}$$

Equation (3) embodies the philosophy of our approach. The first term on the right-hand side of (3) is the prediction of the state valid at k using data up to $k-1$; the second term, $H[\varepsilon(k); k]$ (the weighting of the residuals) is a correction term. H is restricted to be a member of a class of weights $\tilde{\mathscr{H}}$ with the property of additive closure. That is, if $H_\alpha \in \tilde{\mathscr{H}}$, $H_\beta \in \tilde{\mathscr{H}}$, and A, B are real $n \times n$ matrices, then $H_u \in \tilde{\mathscr{H}}$, where

$$H_u[\varepsilon(k); k] = AH_\alpha[\varepsilon(k); k] + BH_\beta[\varepsilon(k); k]. \tag{8}$$

The proposed filter is recursive and only the corrector H is unknown. Our problem then is to choose H from some class of correctors $\tilde{\mathscr{H}}$ with additive closure, such that the mean-squared estimation error is minimized.

3. The generalized orthogonality condition. Let us assume that at $k-1$ an optimal estimate, $\hat{z}(k-1/k-1)$, has been obtained. We wish to determine the weighting, $H \in \tilde{\mathscr{H}}$, so that the m.s.e.[3] is minimized at time k. Let $\hat{z}(k/k)$ and $H[\varepsilon(k); k]$ denote the best estimate and the optimal weight, respectively. Similarly, let $\hat{z}_\alpha(k/k)$ and $H_\alpha[\varepsilon(k); k]$ represent some other estimate and the associated corrector, respectively. We define $e(k)$ and $e_\alpha(k)$ as

$$e(k/k) \triangleq z(k) - \hat{z}(k/k), \tag{9}$$

[2] The orthogonality condition derived herein applies to any filter of the form $\hat{z}(k/k) = g(z(k-1/k-1); k) + H[\varepsilon(k); k]$. The choice of $g = f$ is a convenient one and intuitively appealing.

[3] m.s.e. is an abbreviated notation for the mean-squared error.

$$e_\alpha(k/k) \triangleq z(k) - \hat{z}_\alpha(k/k). \tag{10}$$

The corresponding mean-squared errors are given by[4]

$$V(k) = E\{e'(k/k)\,e(k/k)\}, \tag{11}$$

$$V_\alpha(k) = E\{e'_\alpha(k/k)\,e_\alpha(k/k)\}. \tag{12}$$

The estimator is recursive so both $\hat{z}(k/k)$ and $\hat{z}_\alpha(k/k)$ employ the best previous estimate, $\hat{z}(k-1/k-1)$. Thus,

$$\hat{z}(k/k) = f(\hat{z}(k-1/k-1);k) + H[\varepsilon(k);k], \tag{13}$$

$$\hat{z}_\alpha(k/k) = f(\hat{z}(k-1/k-1);k) + H_\alpha[\varepsilon(k);k]. \tag{14}$$

The optimal weighting H is obtained from the following orthogonality condition (a generalization of the familiar theorems of orthogonal projections as employed by Kalman and Bucy [5] and Pugachev [12].

THEOREM 1 (The generalized orthogonality condition). *For all $H_\beta \in \tilde{\mathscr{H}}$, the necessary and sufficient condition for minimization of the mean-squared error for the predictor-corrector structure is that all the H_β satisfy*

$$E\{e'(k/k)H_\beta[\varepsilon(k);k]\} = 0, \tag{15}$$

i.e., the estimation error must be orthogonal to all corrections, the weighted residuals, in the class $\tilde{\mathscr{H}}$.

Proof. Sufficiency. For some estimate $\hat{z}_\alpha(k/k)$ with a weight $H_\alpha \in \tilde{\mathscr{H}}$, the m.s.e. is expanded as

$$\begin{aligned} V_\alpha(k) = V(k) &+ 2E\{e'(k/k)[\hat{z}(k/k) - \hat{z}_\alpha(k/k)]\} \\ &+ E\{[\hat{z}(k/k) - \hat{z}_\alpha(k/k)]'[\hat{z}(k/k) - \hat{z}_\alpha(k/k)]\}. \end{aligned} \tag{16}$$

From (13) and (14), we observe that

$$\hat{z}(k/k) - \hat{z}_\alpha(k/k) = H[\varepsilon(k);k] - H_\alpha[\varepsilon(k);k]. \tag{17}$$

Because of additive closure, the difference between H and H_α is some weight $H_\beta \in \tilde{\mathscr{H}}$. That is,

$$H_\beta[\varepsilon(k);k] = H[\varepsilon(k);k] - H_\alpha[\varepsilon(k);k] \in \tilde{\mathscr{H}}. \tag{18}$$

When the orthogonality condition (15) is imposed then

$$V_\alpha(k) = V(k) + E\{H'_\beta[\varepsilon(k);k]H_\beta[\varepsilon(k);k]\}. \tag{19}$$

Since the second term on the right-hand side of (19) is nonnegative, then

$$V_\alpha(k) \geqq V(k). \tag{20}$$

Thus the generalized orthogonality condition is sufficient to minimize the mean-squared error.

Necessity. The proof is by contradiction. It is assumed that an estimate, $\hat{z}(k/k)$, and its associated weight, $H \in \tilde{\mathscr{H}}$, can be found which minimizes the m.s.e. with

$$E\{e'(k/k)H_\beta[\varepsilon(k);k]\} = D \neq 0. \tag{21}$$

[4] $(\cdot)'$ denotes the transpose.

Consider an $H_\alpha \in \tilde{\mathscr{H}}$ given by

$$H_\alpha[\varepsilon(k); k] = H[\varepsilon(k); k] - \gamma H_\beta[\varepsilon(k); k], \tag{22}$$

where γ is an arbitrary real scalar. Then from (13) and (14)

$$\hat{z}(k/k) - \hat{z}_\alpha(k/k) = \gamma H_\beta[\varepsilon(k); k]. \tag{23}$$

For the weight H_α, the m.s.e., $V_\alpha(k)$, can be expanded as

$$\begin{aligned} V_\alpha(k) = V(k) &+ 2\gamma E\{e'(k/k)H_\beta[\varepsilon(k); k]\} \\ &+ \gamma^2 E\{H'_\beta[\varepsilon(k); k]H_\beta[\varepsilon(k); k]\}. \end{aligned} \tag{24}$$

For the inequality, $V_\alpha(k) \geqq V(k)$ to hold, it is required that

$$2\gamma E\{e'(k/k)H_\beta[\varepsilon(k); k]\} + \gamma^2 E\{H'_\beta[\varepsilon(k); k]H_\beta[\varepsilon(k); k]\} \geqq 0 \qquad \text{for all } \gamma. \tag{25}$$

But (21) implies that we can always find a real scalar γ which violates the required inequality. For example, when $E\{H'_\beta[\varepsilon/k); k]H_\beta[\varepsilon(k); k]\} > 0$, the following γ would violate the inequality

$$\begin{aligned} 0 > \gamma > \frac{-2D}{E\{H'_\beta[\varepsilon(k); k]H_\beta[\varepsilon(k); k]\}}, \quad D > 0 \\ 0 < \gamma < \frac{-2D}{E\{H'_\beta[\varepsilon(k); k]H_\beta[\varepsilon(k); k]\}}, \quad D > 0. \end{aligned} \tag{26}$$

If $E\{H'_\beta[\varepsilon(k); k]H_\beta[\varepsilon(k); k]\} = 0$, then any nonvanishing γ with the opposite sign of D would yield the desired result. Hence, we have a contradiction to the premise that a minimum m.s.e. can be obtained with $D \neq 0$.

COROLLARY 1. *The mean-squared difference between any two estimates $\hat{z}(k/k)$ and $\hat{z}_0(k/k)$ vanishes if both estimates satisfy the generalized orthogonality condition.*

Proof. The mean-squared errors $V_0(k)$ and $V(k)$ associated with $\hat{z}_0(k)$ and $\hat{z}(k)$, respectively, are expanded as

$$\begin{aligned} V_0(k) = V(k) &+ 2E\{e'(k/k)(\hat{z}(k/k) - \hat{z}_0(k/k))\} \\ &+ E\{[\hat{z}(k/k) - \hat{z}_0(k/k)]'[\hat{z}(k/k) - \hat{z}_0(k/k)]\} \end{aligned} \tag{27}$$

$$\begin{aligned} V(k) = V_0(k) &+ 2E\{e'_0(k/k)[\hat{z}_0(k/k) - \hat{z}(k/k)]\} \\ &+ E\{[\hat{z}(k/k) - \hat{z}_0(k/k)]'[\hat{z}(k/k) - \hat{z}_0(k/k)]\}, \end{aligned} \tag{28}$$

where

$$e_0(k/k) = z(k) - \hat{z}_0(k/k) \tag{29}$$

$$e(k/k) = z(k) - \hat{z}(k/k). \tag{30}$$

The quantities $\hat{z}(k/k)$ and $\hat{z}_0(k/k)$ are both obtained from the best previous estimate. Hence, the difference is

$$\hat{z}(k/k) - \hat{z}_0(k/k) = H[\varepsilon(k); k] - H_0[\varepsilon(k); k]. \tag{31}$$

This difference can be expressed as

$$\hat{z}(k/k) - \hat{z}_0(k/k) = H_\lambda[\varepsilon(k); k]. \tag{32}$$

We now substitute (32) into (27) and (28). When the generalized orthogonality condition is invoked, then the middle terms in (27) and (28) vanish, leaving

$$V_0(k) = V(k) + E\{H'_\lambda[\varepsilon(k);k]H_\lambda[\varepsilon(k);k]\}, \tag{33}$$

$$V(k) = V_0(k) + E\{H'_\lambda[\varepsilon(k);k]H_\lambda[\varepsilon(k);k]\}. \tag{34}$$

Equations (33) and (34) require that

$$E\{H'_\lambda[\varepsilon(k);k]H_\lambda[\varepsilon(k);k]\} = E\{[\hat{z}(k/k) - \hat{z}_0(k/k)]'[\hat{z}(k/k) - \hat{z}_0(k/k)]\} = 0. \tag{35}$$

This result completes the proof.

COROLLARY 2. *Any two estimates, $\hat{z}(k/k)$ and $\hat{z}_0(k/k)$, which satisfy the generalized orthogonality condition give identical mean-squared errors.*

Proof. The proof follows directly from Corollary 1 ((33), (34) and (35)).

These two corollaries demonstrate the mean-square uniqueness property of the proposed predictor-corrector estimation scheme.

The following corollary, a sufficient realization of the orthogonality principle, leads to a direct evaluation of the optimal series weighting.

COROLLARY 3. *The generalized orthogonality condition is satisfied if each component of the error vector is orthogonal to each component of the weighted residuals, for all $H_\alpha \in \tilde{\mathscr{H}}$. That is,*

$$E\{e'(k/k)H_\alpha[\varepsilon(k);k]\} = 0, \qquad \textit{for all } H_\alpha \in \tilde{\mathscr{H}} \tag{36}$$

if

$$E\{e(k/k)H'_\alpha[\varepsilon(k);k]\} = 0, \qquad \textit{for all } H_\alpha \in \tilde{\mathscr{H}}. \tag{37}$$

Proof. If

$$E\{e(k/k)H'_\alpha[\varepsilon(k);k]\} = 0 \tag{38}$$

then

$$\text{TRACE}\left[E\{e(k/k)H'_\alpha[\varepsilon(k);k]\}\right] = 0. \tag{39}$$

But,

$$\text{TRACE}\left[E\{e(k/k)H'_\alpha[\varepsilon(k);k]\}\right] = E\{e'(k/k)H_\alpha[\varepsilon(k);k]\}, \tag{40}$$

thus,

$$E\{e'(k/k)H_\alpha[\varepsilon(k);k]\} = 0. \tag{41}$$

4. Series weighting of residuals. In this section, we investigate the properties of correctors having the series form

$$H_\alpha[\varepsilon(k);k] = J_\alpha(k) + K_\alpha(k)\varepsilon(k) + [\varepsilon'(k)L_{\alpha i}(k)\varepsilon(k)] + \cdots, \tag{42}$$

where $J_\alpha(k)$, $K_\alpha(k)$, $L_{\alpha i}(k)$, etc. are real vectors and matrices, $\varepsilon(k)$ is the residual vector

$$\varepsilon(k) = x(k) - \hat{x}(k) \tag{43}$$

and

$$[\varepsilon'(k)L_{\alpha i}(k)\varepsilon(k)] = \begin{bmatrix} \varepsilon'(k)L_{\alpha 1}\varepsilon(k) \\ \varepsilon'(k)L_{\alpha 2}\varepsilon(k) \\ \vdots \\ \varepsilon'(k)L_{\alpha n}\varepsilon(k) \end{bmatrix}. \tag{44}$$

Next, a concept of length is established. We say H_α is of length "0" if

$$H_\alpha[\varepsilon(k); k] = J_\alpha(k), \qquad \text{all } \alpha,$$

of length "1" if

$$H_\alpha[\varepsilon(k); k] = J_\alpha(k) + K_\alpha(k)\varepsilon(k), \qquad \text{all } \alpha,$$

of length "2" if

$$H_\alpha[\varepsilon(k); k] = J_\alpha(k) + K_\alpha(k)\varepsilon(k) + [\varepsilon'(k)L_{\alpha i}(k)\varepsilon(k)], \qquad \text{all } \alpha$$

and of length j if all the $(j + 1)$st partials of H_α with respect to all components of the residuals vanish.

Next, all series weights of the residuals are classified according to length. All weights of length j or less are in a class designated by $\tilde{\mathscr{H}}^{(j)}$. Thus

$$J_\alpha(k) \in \tilde{\mathscr{H}}^{(0)} \qquad \text{for all } \alpha,$$

$$J_\alpha(k) + K_\alpha(k)\varepsilon(k) \in \tilde{\mathscr{H}}^{(1)} \qquad \text{for all } \alpha, \text{ etc.}$$

It is of interest to consider a class of correctors $\tilde{\mathscr{H}}^{(j)*}$, a subclass of $\tilde{\mathscr{H}}^{(j)}$, formed by restricting all the $J_\alpha(k)$ terms to be zero vectors. The subclass $\tilde{\mathscr{H}}^{(j)*}$, similarly, has the required additive closure property.

THEOREM 2. *For $\tilde{\mathscr{H}}^{(l)}$, the series class of weights of length l, the generalized orthogonality condition is satisfied if and only if*

$$E\{e_i(k/k)\} = 0, \qquad \textit{for all } i, l \tag{45}$$

and

$$\begin{aligned} &E\{e_i(k/k)\varepsilon_{j1}(k)\} = 0, && \textit{for all } i, j1 \textit{ and } l > 0, \\ &E\{e_i(k/k)\varepsilon_{j1}(k)\varepsilon_{j2}(k)\} = 0, && \textit{for all } i, j1, j2 \textit{ and } l > 0, \\ &\vdots \\ &E\{e_i(k/k)\varepsilon_{j1}(k)\varepsilon_{j2}(k)\cdots\varepsilon_{jl}(k)\} = 0, && \textit{for all } i, j1, j2, \cdots, jl \textit{ and } l > 0. \end{aligned} \tag{46}$$

Proof. Sufficiency. Observe that if (45) and (46) are satisfied, then

$$E\{e(k/k)H'_\alpha[\varepsilon(k); k]\} = 0 \qquad \text{for all } H_\alpha \in \tilde{\mathscr{H}}^{(l)}. \tag{47}$$

Hence,

$$E\{e'(k/k)H_\alpha[\varepsilon(k); k]\} = 0 \qquad \text{for all } H_\alpha \in \tilde{\mathscr{H}}^{(l)}. \tag{48}$$

Necessity. The necessity is demonstrated by assuming that the orthogonality condition is satisfied when any of the conditions given by (45) and (46) are not true and show that a contradiction results.

First it is shown that (45) is necessary. Assume that

$$E\{e'(k/k)H_\alpha[\varepsilon(k);k]\} = 0 \qquad \text{for all } H_\alpha \in \tilde{\mathscr{H}}^{(l)} \tag{49}$$

when

$$E(e_i(k/k)\} \neq 0 \qquad \text{for all } i. \tag{50}$$

Consider a specific weight in the class $\tilde{\mathscr{H}}^{(l)}$, given by

$$J\gamma(k) = (0 \cdots 0\, d 0 \cdots 0)', \tag{51}$$

where d, an arbitrary value, is in the γth row of the corrector. It is known that $J_\alpha(k) \in \tilde{\mathscr{H}}^{(l)}$ for all α; therefore

$$E\{e'(k/k)J_\alpha(k)\} = 0, \qquad \text{all } J_\alpha(k). \tag{52}$$

For $J_\gamma(k) \in \tilde{\mathscr{H}}^{(l)}$, it is necessary that

$$E\{e_\gamma(k/k)\} = 0. \tag{53}$$

Since γ is arbitrary, it is required to have

$$E\{e_\gamma(k/k)\} = 0, \qquad \text{all } \gamma, \tag{54}$$

which demonstrates the contradiction. A similar argument is used to establish the necessity of (46).

COROLLARY 4. *The optimal estimate from the class $\tilde{\mathscr{H}}^{(l)}$ is unbiased.* (*The first moment of the estimate error is zero.*)

Proof. The proof follows immediately from Theorem 2, equation (45).

THEOREM 3. *The optimal corrector of length j has a mean-squared error which is less than or equal to the mean-squared error of the optimal corrector of length $j - 1$ if the mean-squared errors exist. That is*

$$V^{(0)} \geqq V^{(1)} \geqq V^{(2)} \cdots \geqq V^{(j)} \geqq \cdots, \tag{55}$$

where $V^{(j)}$ is the minimum m.s.e. corresponding to the optimal corrector (*weight*) *from $\tilde{\mathscr{H}}^{(j)}$.*

Proof. Since

$$\tilde{\mathscr{H}}^{(0)} \subseteq \tilde{\mathscr{H}}^{(1)} \subseteq \tilde{\mathscr{H}}^{(2)} \cdots, \tag{56}$$

it follows immediately that the optimal weighting from $\tilde{\mathscr{H}}^{(j)}$ yields a m.s.e. which is not greater than the m.s.e. for the optimal weighting from $\tilde{\mathscr{H}}^{(j-1)}$.

Theorem 3 illustrates that the performance of the estimator would generally improve as the length of the corrector is increased. The equalities hold for some special cases. For example, it is known that for a linear signal model the best corrector is of the form $K(k)\varepsilon(k)$. Thus,

$$V^{(0)} \geqq V^{(1)} = V^{(2)} = V^{(3)} \cdots. \tag{57}$$

THEOREM 4.

$$V^{(j)} \leqq V^{(j)*} \tag{58}$$

if both $V^{(j)}$ and $V^{(j)*}$ exist and if they are the minimum mean-squared errors from the classes $\tilde{\mathscr{H}}^{(j)}$ and $\tilde{\mathscr{H}}^{(j)*}$, respectively.

Proof. Because $\tilde{\mathscr{H}}^{(j)*} \subseteq \tilde{\mathscr{H}}^{(j)}$, the proof follows directly.

Theorem 4 states that the best corrector from the class $\tilde{\mathscr{H}}^{(j)*}$ yields a m.s.e. which is never better than that obtained by the optimal corrector from the class $\tilde{\mathscr{H}}^{(j)}$.

THEOREM 5. *Equation* (46) *states a necessary and sufficient condition to satisfy the generalized orthogonality condition for the class of correctors* $\tilde{\mathscr{H}}^{(l)*}$, *for* $l \geqq 1$.

Proof. The proof is similar to that for Theorem 2.

Because the best weight from the class $\tilde{\mathscr{H}}^{(l)*}$ need not satisfy (45), the estimates obtained by using this corrector can be biased.[5] A very important difference between the classes $\tilde{\mathscr{H}}^{(l)}$ and $\tilde{\mathscr{H}}^{(l)*}$ is that the best weight from $\tilde{\mathscr{H}}^{(l)}$ yields an unbiased estimate, while the estimate from $\tilde{\mathscr{H}}^{(l)*}$ can be biased. The extended Kalman filter as employed for nonlinear dynamics has a corrector which is a member of $\tilde{\mathscr{H}}^{(1)*}$. Thus, the estimate obtained by using this weight $(K(k)\varepsilon(k))$ can be biased.

From the results of Theorem 2 the next corollary can now be stated.

COROLLARY 5.

$$V^{(j)} < \infty \quad if \quad V^{(0)} < \infty. \tag{59}$$

THEOREM 6 (Boundedness condition). *If the signal process is bounded in the sense that*

$$E\{z'(k)z(k)\} < \infty \tag{60}$$

then all the minimum m.s.e. associated with the weights from $\tilde{\mathscr{H}}^{(j)}$ *for all j are bounded, that is*

$$V^{(j)}(k) < \infty, \qquad \textit{for all } j. \tag{61}$$

Proof. Consider a corrector from the class of length "0",

$$J_\alpha(k) = -E\{f(\hat{z}(k-1/k-1);k)\}, \tag{62}$$

where

$$J_\alpha(k) = H_\alpha[\varepsilon(k);k] \in \tilde{\mathscr{H}}^{(0)}. \tag{63}$$

This corrector yields a *m.s.e.* given by

$$V_\alpha^{(0)}(k) = E\{z'(k)z(k)\}. \tag{64}$$

Since $J_\alpha(k)$ is not necessarily optimal, then

$$V_\alpha^{(0)}(k) \geqq V^{(0)}(k), \tag{65}$$

where $V^{(0)}(k)$ is the minimum m.s.e. for the class $\tilde{\mathscr{H}}^{(0)}$.

From Theorems 6 and 3 it follows that

$$V^{(j)}(k) \leqq V^{(0)}(k) \leqq V_\alpha^{(0)}(k) = E\{z'(k)z(k)\} < \infty \quad \text{all } j = 0, 1, 2 \cdots. \tag{66}$$

[5] It is conjectured that unbiased estimates result for correctors from both $\tilde{\mathscr{H}}^{(l)}$ and $\tilde{\mathscr{H}}^{(l)*}$ when the dynamics $(f(\cdot))$ and the measurements $(h(\cdot))$ possess odd symmetry, i.e., $f(x) = -f(-x)$.

5. Applications and examples. Three examples are presented to illustrate the proposed method of obtaining a corrector. The first illustrates the link between our approach and the so-called "extended" Kalman filter [9], [10], [11]. The second demonstrates the use of a quasilinear weighting scheme which gives an unbiased estimate. The third example extends the procedure to a scalar weighting scheme.

In all cases an estimator is sought for the system described by

$$z(k) = f(z(k-1); k) + G(k)u(k). \tag{67}$$

The observed data vector is

$$x(k) = M(k)z(k) + v(k). \tag{68}$$

For convenience we assume:

(i) The initial state vector, $z(0)$, is Gaussian with a known mean and covariance matrix.

(ii) The driving noise, $u(k)$, is a sequence of uncorrelated Gaussian random vectors with zero mean and a known covariance, $U(k)$.

(iii) The corrupting noise, $v(k)$, is a sequence of uncorrelated Gaussian random vectors with zero mean and known covariance, $N(k)$.

(iv) The initial state is independent of $u(k)$ and $v(k)$.

(v) $u(k)$ and $v(k)$ are correlated.

That is,

$$E\{u(k)\} = 0, \tag{69}$$

$$E\{v(k)\} = 0, \tag{70}$$

$$E\{u(k)u'(j)\} = U(k)\delta_{kj}, \tag{71}$$

$$E\{v(k)v'(j)\} = N(k)\delta_{kj}, \tag{72}$$

$$E\{u(k)v'(j)\} = C(k)\delta_{kj}, \tag{73}$$

where

$$\delta_{kj} = \begin{cases} 0, & k \neq j, \\ 1, & k = j, \end{cases} \tag{74}$$

and

$$E\{u(k)z'(0)\} = 0 \quad \text{all } k, \tag{75}$$

$$E\{v(k)z'(0)\} = 0 \quad \text{all } k. \tag{76}$$

Example 1 (Linear weighting of residuals). An optimal weight from the class $\mathscr{H}^{(1)*}$ is sought for the estimator given by

$$\hat{z}(k/k) = f(\hat{z}(k-1/k-1); k) + K(k)\varepsilon(k) . \tag{77}$$

The initial state of this estimator is taken as the mean of $z(0)$,

$$\hat{z}(0/0) = E\{z(0)\} \tag{78}$$

and

(79) $$\varepsilon(k) = x(k) - \hat{x}(k),$$

(80) $$\hat{x}(k) = M(k)\hat{z}(k/k-1),$$

(81) $$\hat{z}(k/k-1) = f(\hat{z}(k-1/k-1); k).$$

Equations (77) to (81) have the general form of the "extended" Kalman filter. This estimator is completely specified once $K(k)$ is determined.

From Theorem 5, the general orthogonality condition is satisfied when

(82) $$E\{e_i(k/k)\varepsilon_j(k)\} = 0 \qquad \text{all } i, j.$$

From this condition, the equation for $K(k)$ is obtained directly as

(83) $$\begin{aligned} K(k) = {} & [P(k)M'(k) + G(k)C(k)] \\ & \cdot [M(k)P(k)M'(k) + M(k)G(k)C(k) + C'(k)G'(k)M'(k) + N(k)]^{-1}, \end{aligned}$$

where

(84) $$P(k) = E\{e(k/k-1)e'(k/k-1)\}.$$

$C(k)$ vanishes when $u(k)$ and $v(k)$ are uncorrelated and the weight matrix is reduced to the form

(85) $$K(k) = P(k)M'(k)[M(k)P(k)M'(k) + N(k)]^{-1}.$$

Either (83) or (85) is identical in form to that employed in the "extended" Kalman filter. The differences in interpretation are twofold:

(i) $P(k)$ as used in this presentation is given explicitly by (84) and assumes the use of the exact prediction error statistics. Furthermore, $P(k)$ is not necessarily a covariance matrix, because $e(k/k-1)$ can have a nonvanishing mean $(E\{e(k/k-1)\} \neq 0)$.

(ii) The preceding form of weighting is obtained directly from the generalized orthogonality condition. The extended Kalman filter as described in the literature uses either linearization arguments or else the form is postulated a priori as an extension of the linear theory.

Practical implementation of (84), however, involves determining $P(k)$ to whatever accuracy is required. A first order approximation utilizing linearization of the nonlinear dynamics about a nominal trajectory yields results identical to those of the extended Kalman filter. This first order approximation requires knowledge of the mean and covariance of $z(0)$.

Example 2 (Quasi-linear weighting of residuals). For the signal process and observed data vector as given by (67) and (68), an optimal weight from the class $\mathscr{H}^{(1)}$ is found. The generalized orthogonality condition is satisfied by

(86) $$E\{e(k/k)\} = 0,$$

(87) $$E\{e(k/k)\varepsilon'(k)\} = 0.$$

The estimator with a weight from the class $\mathscr{H}^{(1)}$ is of the form

(88) $$\hat{z}(k/k) = f(\hat{z}(k-1/k-1); k) + J(k) + K(k)\varepsilon(k).$$

The initial estimate is given by (78) and $\varepsilon(k)$, $\hat{x}(k)$, $\hat{z}(k/k-1)$ are specified by (79), (80) and (81). With

$$e(k/k-1) = z(k) - \hat{z}(k/k-1) \tag{89}$$

the error is given by

$$e(k/k) = \Gamma(k)e(k/k-1) - K(k)v(k) - J(k), \tag{90}$$

where

$$\Gamma(k) = I - K(k)M(k). \tag{91}$$

Thus, from (86) and (90), we find that

$$J(k) = \Gamma(k)E\{e(k/k-1)\}. \tag{92}$$

Although (86) requires that the error $(e(k))$ has a zero mean, it does not guarantee that the first moment of the prediction error vector vanishes.[6]

The weighting, $K(k)$, of the residual vector, $\varepsilon(k)$, is found from (87) by an approach which parallels that employed in Example 1. It is found that

$$\begin{aligned} K(k) = {} & [\mathrm{Cov}\,(e(k/k-1))M'(k) + G(k)C(k)] \\ & \cdot [N(k) + M(k)\,\mathrm{Cov}\,(e(k/k-1))M'(k) \\ & + M(k)G(k)C(k) + C'(k)G'(k)M'(k)]^{-1}, \end{aligned} \tag{93}$$

where

$$\mathrm{Cov}\,(e(k/k-1)) \triangleq E\{e(k/k-1)e'(k/k-1)\} - E\{e(k/k-1)\}E\{e'(k/k-1)\} \tag{94}$$

or

$$\mathrm{Cov}\,\{e(k/k-1)\} = P(k) - E\{e(k/k-1)\}E\{e'(k/k-1)\}. \tag{95}$$

Again the corrector is obtained by using the true prediction error statistics. If these exact statistics are employed, the generalized orthogonality condition is satisfied and the estimate is unbiased.

An approximate evaluation of $E\{e(k/k-1)\}$ and $\mathrm{Cov}\,(e(k/k-1))$ can be found from the series expansion of f about some nominal trajectory $z^*(k)$.

$$\begin{aligned} f(z(k-1);k) = {} & f(z^*(k-1);k) + A(k)[z(k-1) - z^*(k-1)] \\ & + \big[[z(k-1) - z^*(k-1)]'B_i(k)[z(k-1) - z^*(k-1)]\big] \\ & + R(k) \end{aligned} \tag{96}$$

and

$$\begin{aligned} f(\hat{z}(k-1/k-1);k) = {} & f(z^*(k-1);k) + A(k)[\hat{z}(k-1/k-1) - z^*(k-1)] \\ & + [\hat{z}(k-1/k-1) - z^*(k-1)]'B_i(k) \\ & \cdot\ [z(k-1/k-1) - z^*(k-1)] + \hat{R}(k), \end{aligned} \tag{97}$$

[6] $E\{e(k/k-1)\} = E\{f(z(k-1);k) - f(z(k-1/k-1);k)\}$. If f is an arbitrary nonlinear vector function, this expectation is generally nonzero.

where

$$[A(k)]_{ij} = \left.\frac{\partial f_i(z(k-1);k)}{\partial z_j(k-1)}\right|_{z(k-1)=z^*(k-1)}, \tag{98}$$

$$[B_i(k)]_{lp} = \frac{1}{2}\left.\frac{\partial^2 f_i(z(k-1);k)}{\partial z_l(k-1)\partial z_p(k-1)}\right|_{z(k-1)=z^*(k-1)} \tag{99}$$

and

$$\begin{aligned}&\big[[z(k-1)-z^*(k-1)]'B_i(k)[z(k-1)-z^*(k-1)]\big]\\ &\qquad = \begin{bmatrix}[z(k-1)-z^*(k-1)]'B_1(k)[z(k-1)-z^*(k-1)]\\ [z(k-1)-z^*(k-1)]'B_2(k)[z(k-1)-z^*(k-1)]\\ \vdots\end{bmatrix}.\end{aligned} \tag{100}$$

$R(k)$ and $\hat{R}(k)$ denote the higher order terms in the expansions.

The prediction error, given by

$$e(k/k-1) = f(z(k-1);k) - f(\hat{z}(k-1/k-1);k) + G(k)u(k) \tag{101}$$

is expanded as

$$\begin{aligned}e(k/k-1) &= A(k)e(k-1/k-1)\\ &\quad + \big[[e(k-1/k-1)]'B_i(k)[e(k-1/k-1)]\big] + \lambda(k),\end{aligned} \tag{102}$$

where

$$\begin{aligned}\lambda(k) &= R(k) - \hat{R}(k) + G(k)u(k)\\ &\quad + \big[[\hat{z}(k-1/k-1) - z^*(k-1)]'B_i(k)[z(k-1)-z^*(k-1)]\big]\\ &\quad + \big[[z(k-1)-z^*(k-1)]'B_i(k)[\hat{z}(k-1/k-1)-z^*(k-1)]\big].\end{aligned} \tag{103}$$

If the nominal trajectory is taken as the estimated trajectory ($\hat{z}(k/k) = z^*(k)$), then

$$\lambda(k) = R(k) - \hat{R}(k) + G(k)u(k). \tag{104}$$

If we further assume that all moments of the error component products having three or more terms are negligible, then with

$$E\{e(k/k)\} = 0, \tag{105}$$

it follows that

$$E\{e(k/k-1)\} \doteq E\{[[e(k-1/k-1)]'B_i(k)[e(k-1/k-1)]]\}, \tag{106}$$

where, of course, B_i is deterministic, and

$$P(k) \doteq A(k)Q(k-1)A'(k) + G(k)U(k)G'(k), \tag{107}$$

where

$$Q(k) \doteq \Gamma(k)\,\mathrm{Cov}\,(e(k/k-1)) \tag{108}$$

and

$$\Gamma(k) = I - K(k)M(k), \tag{109}$$

$$Q(0) = E\{z(0)z'(0)\} - E\{z(0)\}E\{z'(0)\}. \tag{110}$$

With these approximate error statistics a recursive procedure for evaluating the weighting parameters can be employed. The corrector cannot be precomputed since its value is dependent on the estimated trajectory.

If this approximate approach is not adequate for a specific application then the prediction error statistics will have to be obtained with greater accuracy. The price, of course, is a greater computational effort.

Example 3 (Scalar weighting of residuals). For convenience we consider a linear signal process

$$z(k) = A(k)z(k-1) + G(k)u(k) \tag{111}$$

and observation vector

$$x(k) = M(k)z(k) + v(k). \tag{112}$$

The estimator chosen is

$$\hat{z}(k/k) = A(k)\hat{z}(k-1/k-1) + \beta(k)D(k)\varepsilon(k), \tag{113}$$

where

$$\hat{z}(0/0) = E\{z(0)\}. \tag{114}$$

The $n \times m$ matrix, $D(k)$, is preassigned arbitrarily. The weight $\beta(k)$ is chosen from the class of scalars, $\mathscr{A}$, to minimize the mean-squared error. As in the earlier considerations the class $\mathscr{A}$ has the additive closure property. From the generalized orthogonality condition it is known that for any class $\tilde{\mathscr{H}}$ with additive closure it is required to have

$$E\{e'(k/k)H_\alpha(\varepsilon(k); k)\} = 0 \qquad \text{for all } H_\alpha \in \tilde{\mathscr{H}}. \tag{115}$$

Thus,

$$E\{e'(k/k)\beta_\alpha(k)D(k)\varepsilon(k)\} = 0 \qquad \text{for all } \beta_\alpha(k) \in \mathscr{A}. \tag{116}$$

Equation (116) is satisfied if

$$\text{TRACE } E\{e(k/k)\varepsilon'(k)D'(k)\} = 0. \tag{117}$$

For this linear system, the residual is

$$\varepsilon(k) = M(k)e(k/k-1) + v(k) \tag{118}$$

and the error is

$$e(k/k) = \Gamma^*(k)e(k/k-1) - \beta(k)D(k)v(k), \tag{119}$$

where

$$e(k/k-1) = z(k) - A(k)\hat{z}(k-1/k-1), \tag{120}$$

$$\Gamma^*(k) \triangleq I - \beta(k)D(k)M(k). \tag{121}$$

Thus,

$$\begin{aligned} E\{e(k/k)\varepsilon'(k)D'(k)\} = {} & \Gamma^*(k)P(k)M'(k)D'(k) \\ & + \Gamma^*(k)G(k)C(k)D'(k) - \beta(k)D(k)N(k)D'(k) \\ & - \beta(k)D(k)C'(k)G'(k)M'(k)D'(k). \end{aligned} \tag{122}$$

From the trace of (122), $\beta(k)$ is found to be

$$\beta(k) = \frac{\text{TRACE}\,[P(k)M'(k)D'(k) + G(k)C(k)D'(k)]}{\text{TRACE}\,[F(k)]}, \tag{123}$$

where

$$\begin{aligned} F(k) = {} & D(k)M(k)P(k)M'(k)D'(k) \\ & + D(k)M(k)G(k)C(k)D'(k) \\ & + D(k)C'(k)G'(k)M'(k)D'(k) \\ & + D(k)N(k)D'(k). \end{aligned} \tag{124}$$

If $u(k)$ and $v(k)$ are uncorrelated, then

$$\beta(k) = \frac{\text{TRACE}\,[P(k)M'(k)D'(k)]}{\text{TRACE}\,[D(k)M(k)P(k)M'(k)D'(k)] + \text{TRACE}\,[D(k)N(k)D'(k)]}. \tag{125}$$

The prediction error covariance matrix is given by

$$P(k) = A(k)Q(k-1)A'(k) + G(k)U(k)G'(k) \tag{126}$$

and the error covariance matrix is

$$\begin{aligned} Q(k) = {} & \Gamma^*(k)P(k)(\Gamma^*(k))' + \beta(k)[\Gamma^*(k)G(k)C(k)D'(k)] \\ & + \beta(k)D(k)C'(k)G'(k)(\Gamma^*(k))' + \beta^2(k)D(k)N(k)D'(k), \end{aligned} \tag{127}$$

where

$$Q(0) = E\{z(0)z'(0)\} - E\{z(0)\}E\{z'(0)\}. \tag{128}$$

Although this example was limited to linear systems, similar arguments can be employed for nonlinear dynamics. As in Example 2, one could easily employ a weight given by

$$H_\alpha(\varepsilon(k); k) = J(k) + \beta(k)D(k)\varepsilon(k). \tag{129}$$

This latter approach would yield unbiased estimates. The distinct advantage afforded by the formulation in (113) and (129) is that the weights can be evaluated without a single matrix inversion.

For a linear system the scalar weighting can be readily precomputed. The storage requirements would be greatly reduced when the scalar weights are employed in place of the complete Kalman weighting matrix. When the steady-state solution of the Kalman weight is known, the $D(k)$ matrix could be replaced by the steady-state solution. Then $\beta(k)$ would be used to handle the transient behavior. A suitable form for $\beta(k)$ is that $\lim_{k\to\infty} \beta(k) = 1$.

This scalar approach can, of course, be extended to $(n \times l)$ matrices, $B(k)$, with arbitrary $l \times m$ matrices, $D(k)$. For this case, the sufficient condition to satisfy the generalized orthogonality condition would be

$$E\{e(k/k)\varepsilon'(k)D'(k)B'(k)\} = 0. \tag{130}$$

The best weight matrix, $B(k)$, is found to be

$$B(k) = P(k)M'(k)D'(k)[D(k)M(k)P(k)M'(k)D'(k) + D(k)N(k)D'(k)]^{-1}, \tag{131}$$

where it is assumed that $u(k)$ and $v(k)$ are uncorrelated. $P(k)$ is given by (126) and the error covariance satisfies

$$Q(k) = \Gamma^*(k)P(k), \tag{132}$$

where

$$\Gamma^*(k) = I - B(k)D(k)M(k). \tag{133}$$

This result reduces to the normal Kalman filter when $D(k)$ is an $m \times m$ identity matrix.

6. Conclusions. A general approach for obtaining an optimal corrector in the mean-squared sense for the predictor corrector structure has been developed. It was shown that the optimal corrector must satisfy the generalized orthogonality condition. The theory was specifically applied to correctors of series form, but the general approach is not restricted to this type of weighting.

Examples of signal models with nonlinear and with linear dynamics were presented to illustrate applications of the proposed technique. It was shown that for the class $\mathscr{H}^{(1)*}$ (linear weighting of the residuals) the optimal weight was identical in form to that obtained for the extended Kalman filter. Furthermore, it was illustrated that unbiased estimates can be obtained by use of the expected value of the prediction error.

In the examples, sequences of independent random vectors with zero mean and known covariance were assumed for the corrupting noise. This assumption was employed for convenience. Correlated noise can be handled by treating the correlated portion of the noise vector as though it were included in the signal process. Estimates of both the signal process and of the correlated noise are then obtained but only estimates of the desired signal process retained.

Although a direct extension of these discrete-time results to continuous-time systems$^{(21)}$ presents some difficulty, the continuous-time weight is obtained from

$$E\{e'(t)H_\alpha[\varepsilon(t);t]\} = 0, \tag{134}$$

for all $H_\alpha \in \mathscr{H}$, where $e(t)$ is the estimation error and $\varepsilon(t)$ is the residual. The resulting continuous-time version of the presented filter, with correctors from the $\tilde{\mathscr{H}}^{(1)*}$ class,[7] are identical to the extended continuous-time Kalman filter. Furthermore, the $\tilde{\mathscr{H}}^{(1)}$ class of correctors yields the filtering algorithm presented by Jazwinski [18], [19].

[7] The class $\tilde{\mathscr{H}}^{(1)*}$ has correctors of the form $K(t)\varepsilon(t)$ and the $\tilde{\mathscr{H}}^{(1)}$ class has correctors of the form $J(t) + K(t)\varepsilon(t)$.

7. Acknowledgments. The authors are deeply indebted to the National Science Foundation, the Polytechnic Institute of Brooklyn and Bell Telephone Laboratories for making this paper possible. The comments and suggestions presented by Professor L. G. Shaw of the Polytechnic Institute of Brooklyn and R. T. Ollman of Bell Telephone Laboratories have been extremely useful and rewarding. We wish to acknowledge the skill and effort of Miss Norma Lockwood in typing this paper.

REFERENCES

[1] P. Swerling, *First order error propagation in a stagewise smoothing procedure for satellite observations*, J. Astronaut. Sci., 6 (1959), pp. 46–52.

[2] ———, *Topics in generalized least squares signal estimation*, SIAM J. Appl. Math., 14 (1966), pp. 998–1031.

[3] R. E. Kalman, *A new approach to linear filtering and prediction problems*, Trans. ASME Ser. B. J. of Engrg. Indust., 82D (1960), pp. 35–45.

[4] ———, *New methods and results in linear prediction and filtering theory*, Engineering Applications of Random Function Theory and Probability, J. L. Boganoff and F. Kozin, eds., John Wiley, New York, 1963.

[5] R. E. Kalman and R. S. Bucy, *New results in linear filtering and prediction theory*, Trans. ASME Ser. B. J. of Engrg. Indust., 83D (1961), pp. 95–108.

[6] R. C. K. Lee, *Optimal estimation identification, and control*, Research Monograph No. 28, MIT Press, Cambridge, Massachusetts, 1964.

[7] Y. C. Ho and R. C. K. Lee, *A Bayesian approach to problems in stochastic estimation and control*, IEEE Trans. Automatic Control, AC-9 (1964), pp. 333–339.

[8] H. E. Rauch, F. Tung and C. T. Striebel, *Maximum likelihood estimates of linear dynamic systems*, AIAA J., 3 (1965), pp. 1445–1450.

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

[10] V. O. Mowery, *Least squares recursive differential correction estimation in non-linear problems*, IEEE Trans. Automatic Control, AC-10 (1965), pp. 399–407.

[11] H. Cox, *On the estimation of state variables and parameters for noisy dynamic systems*, Ibid., AC-9 (1964), pp. 5–12.

[12] V. S. Pugachev, *Theory of Random Functions and its Application to Control Problems*, Pergamon, Oxford, (distributed by) Addison-Wesley, Reading, Massachusetts, 1965.

[13] R. L. Stratonovich, *Conditional Markov processes*, Theor. Probability Appl., V (1960), pp. 156–178.

[14] H. J. Kushner, *On the dynamical equations of the probability density functions, with applications to optimal stochastic control theory*, J. Math. Anal. Appl., 8 (1964), pp. 332–344.

[15] ———, *On the differential equations satisfied by conditional probability densities of Markov processes*, this Journal, 2 (1964), pp. 106–119.

[16] R. S. Bucy, *Nonlinear filtering theory*, IEEE Trans. Automatic Control, AC-10 (1965), p. 198.

[17] W. M. Wonham, *Some applications of stochastic differential equations to optimal nonlinear filtering*, this Journal, 2 (1965), pp. 347–369.

[18] A. H. Jazwinski, *Filtering of nonlinear dynamical systems*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 765–766.

[19] ———, *Nonlinear filtering with discrete observations*, AIAA Paper 66–38, 1966 AIAA Aerospace Science Meeting, New York.

[20] L. Schwartz and E. B. Stear, *A computational comparison of several nonlinear filters*, IEEE Trans. Automatic Control, AC-13 (1968), pp. 83–86.

[21] J. R. Krasnakevich, *A class of nonlinear unbiased estimators*, Doctoral thesis, System Science, Polytechnic Institute of Brooklyn, Brooklyn, 1968.

SIAM J. Control
Vol. 7, No. 3, August 1969

ON SENSITIVITY OF CLOSED LOOP NONLINEAR OPTIMAL CONTROL SYSTEMS*

ELIEZER KREINDLER†

Abstract. A fundamental question in feedback control is: How does the feedback affect the sensitivity of the system's motion to small variations of the system's parameters? This question is investigated for a sufficiently smooth Bolza optimization problem. A necessary and sufficient condition is derived for closed loop sensitivity reduction according to an inequality involving a particular integral-square sensitivity measure that is closely related to Bode's classical sensitivity function. This condition, under a mild assumption, holds for problems that are separable in the state and control variables. The results are specialized for linear problems.

1. Introduction. Modern optimal control problems are usually formulated and solved as open loop problems, i.e., the control is a function of time, which in actual implementation is oblivious to errors in the initial state and disturbances of the system's motion. To account automatically for these, the optimal control is expressed and implemented as a function of the current state, or a feedback correction is added to the open loop control. In addition to providing automatic control, however, feedback affects such dynamic properties of the plant as stability and sensitivity to plant parameter variations and disturbances. This paper is therefore addressed to the question: Is the closed loop optimal system less sensitive than the open loop one, according to some reasonable criterion? One such criterion is derived, and it represents a generalization of earlier results for linear optimal systems [1], [2], which in turn were motivated by results in [3]–[7].

2. Problem formulation. Consider a plant described by the vector differential equation

$$\dot{x} = f(t, x, u, \mu), \qquad x(t_0) = x_0, \tag{2.1}$$

where the scalar t is the time, x is the n-dimensional state, u the r-dimensional control function, and μ a p-dimensional continuously time-varying parameter. The function f is assumed to be continuous in t, and twice continuously differentiable in x, u and μ. The nominal value of $\mu(t)$ is denoted by $\mu_*(t)$ and a solution of (2.1) corresponding to $u_*(t)$ and $\mu_*(t)$ is denoted by $x_*(t)$.

The objective of control is to transfer the state x from the initial point $x(t_0) = x_0$ to some point $x(t_1) = x_1$ in a smooth terminal manifold in (t, x)-space while minimizing the performance index

$$I = g(t_1, x(t_1)) + \int_{t_0}^{t_1} h(t, x, u)\, dt, \tag{2.2}$$

where g and h are scalar functions, continuous in t and twice continuously differentiable in x and u. The control vector is confined to a region U in r-space which

* Received by the editors September 3, 1968, and in revised form March 19, 1969.

† Research Department, Grumman Aircraft Engineering Corporation, Bethpage, New York 11714.

may depend on t,

$$u \in U(t), \tag{2.3}$$

and has a smooth boundary in (t, u)-space.

It is assumed that for $\mu(t) = \mu_*(t)$ there exists a unique continuous optimal solution denoted by $\{x_*(t), u_*(t)\}$. It is further assumed that there exists a single-valued optimal feedback control law $k(t, x)$, continuous in t and continuously differentiable in x. The function k is defined in some $(n + 1)$-dimensional neighborhood of $x_*(t), t_0 \leqq t < t_1$. By definition of an optimal feedback control law,

$$u_*(t) = -k(t, x_*(t)), \tag{2.4}$$

where the minus sign is a notational convention. Furthermore, for all (t, x) for which $k(t, x)$ is defined, the control

$$u = -k(t, x) \tag{2.5}$$

is the unique optimal value of control. The assumptions on k imply that the solution of the problem is normal and that there are no conjugate points.

The assumptions made so far in this section will be referred to as the *smoothness assumptions*. While they exclude many problems of great interest, notably those involving discontinuous controls, they are standard for the derivation of local optimal feedback control based on the theory of neighboring optimal trajectories [8]–[11].

Consider continuous parameter variations $\varepsilon\delta\mu(t)$ defined by

$$\varepsilon\delta\mu(t) = \mu(t) - \mu_*(t) \tag{2.6}$$

and the corresponding state variations $\varepsilon\delta x(t)$. In the limit of $\varepsilon \to 0$, δx is given by the well-known variational equation

$$\delta\dot{x} = f_x\delta x + f_\mu\delta\mu + f\delta u, \qquad \delta x(t_0) = 0, \tag{2.7}$$

where the matrices of partial derivatives, f_x, f_u and f_μ, are understood to be evaluated along $\{u_*(t), x_*(t), \mu_*(t)\}$. It follows from the assumption made as to the existence of a solution of (2.1) that a solution of (2.7) exists on $t_0 \leqq t \leqq t_1$ (see [12]). Norms of $\delta x(t)$ are measures of *trajectory sensitivity*.

The purpose of this paper is to compare, with respect to trajectory sensitivity, the open loop system where

$$u(t) = u_*(t; t_0, x_0, \mu_*(t)) \tag{2.8}$$

with the *nominally equivalent* closed loop system where

$$u(t, x) = -k(t, x). \tag{2.9}$$

Specifically, the objective is to find conditions for closed loop sensitivity reduction according to the criterion

$$(2.10)\qquad \int_{t_0}^{t'} \delta x_c^T(t)Z(t)\delta x(t)\,dt \leqq \int_{t_0}^{t'} \delta x_o^T(t)Z(t)\delta x_o(t)\,dt \quad \text{for all } t', t_0 < t' \leqq t_1,$$

where the subscripts c and o (to be distinguished from subscript zero) refer to a pair of nominally equivalent closed loop and open loop systems, respectively; superscript T denotes matrix transposition; and $Z(t)$ is a symmetric nonnegative definite matrix. Since (2.7) for δx is linear, this is an appropriate criterion. Furthermore, as discussed in [6], [7], [13]–[18] it is closely related to the classical Bode sensitivity function.

3. A basic lemma. For the open loop system where $\delta u \equiv 0$, (2.7) becomes

$$(3.1)\qquad \delta\dot{x}_o = f_x\delta x_o + f_\mu\delta\mu, \qquad \delta x_o(t_0) = 0,$$

and for the closed loop system, since $\delta u = -k_x\delta x_c$,

$$(3.2)\qquad \delta\dot{x}_c = f_x\delta x_c + f_\mu\delta\mu - f_u k_x\delta x_c, \qquad \delta x_c(t_0) = 0,$$

where the matrices f_x, f_u, f_μ and k_x are understood to be evaluated along $\{x_*(t), u_*(t), \mu_*(t)\}$. Let $\Phi(t, \tau)$ denote the transition matrix corresponding to f_x, i.e.,

$$(3.3)\qquad \frac{\partial}{\partial t}\Phi(t,\tau) = f_x\Phi(t,\tau), \qquad \Phi(\tau,\tau) = I.$$

Then

$$(3.4)\qquad \delta x_o(t) = \int_{t_0}^{t} \Phi(t,\tau)f_\mu\delta\mu(\tau)\,d\tau,$$

and

$$(3.5)\qquad \delta x_c(t) = \int_{t_0}^{t} \Phi(t,\tau)f_\mu\delta\mu(\tau)\,d\tau - \int_{t_0}^{t} \Phi(t,\tau)f_u k_x\delta x_c(\tau)\,d\tau.$$

By denoting

$$(3.6)\qquad v(t) = \int_{t_0}^{t} \Phi(t,\tau)f_u k_x\delta x_c(\tau)\,d\tau,$$

(3.5) becomes

$$(3.7)\qquad \delta x_o(t) = \delta x_c(t) + v(t),$$

and hence

$$(3.8)\qquad \delta x_o^T(t)Z(t)\delta x_o(t) - \delta x_c^T(t)Z(t)\delta x_c(t) = 2\delta x_c^T(t)Z(t)v(t) + v^T(t)Z(t)v(t).$$

Clearly, for (2.10) to hold it is necessary and sufficient that

$$(3.9)\qquad \int_{t_0}^{t'} (2\delta x_c^T Zv + v^T Zv)\,dt \geqq 0 \quad \text{for all } t', t_0 < t' \leqq t_1.$$

It is noteworthy that (3.7) is independent of $\delta\mu$ and of f_μ, and it is therefore possible to relate δx_c to δx_o abstractly by

$$(3.10)\qquad \delta x_c = S\delta x_o.$$

The linear operator S, called the *comparison sensitivity operator*, is a generalization of Bode's sensitivity function (of a complex variable), and the condition (3.9) can easily be shown to be a generalization of the classical criterion that for closed loop sensitivity reduction the *return difference* must be larger than unity (in the frequency domain). Such generalizations are discussed in [6], [7], [13]–[18].

The next step is to derive an expression of the form (3.9), using the conditions for optimality. The Hamiltonian function is

$$H(t, x, u, \lambda) = h(t, x, u) + \lambda^T f(t, x, u). \tag{3.11}$$

At a point (t, x, λ), H is minimized, subject to (2.3), by a unique value of u

$$u = c(t, x, \lambda), \tag{3.12}$$

which is assumed to be continuous in t and continuously differentiable in x and λ, and to provide an absolute minimum of H that will be denoted by

$$\mathscr{H}(t, x, \lambda) = H(t, x, c(t, x, \lambda), \lambda). \tag{3.13}$$

For sensitivity analysis, we are interested only in k_x which is given by

$$k_x = -c_x - c_\lambda P, \tag{3.14}$$

where the matrix P is given by the Riccati equation [11]

$$-\dot{P} = P\mathscr{H}_{\lambda x} + \mathscr{H}_{x\lambda}P + P\mathscr{H}_{\lambda\lambda}P + \mathscr{H}_{xx}. \tag{3.15}$$

If the control is unconstrained, then [10]

$$k_x = H_{uu}^{-1}(H_{ux} + f_u^T P), \tag{3.16}$$

and

$$-\dot{P} = Pf_x + f_x^T P - (Pf_u + H_{xu})H_{uu}^{-1}(H_{ux} + f_u^T P) + H_{xx}. \tag{3.17}$$

The terminal condition on $P(t_1)$ is related to the transversality condition of the original optimization problem, and depends on whether t_1 is specified or is free. If t_1 and a terminal manifold are specified, some of the elements of $P(t_1)$ are infinite and $\|k_x(t_1)\| \to \infty$ as $t \to t_1$. Since we are not concerned here with a numerical problem, we need not elaborate (see [10]). The existence of a finite solution $P(t), t_0 \leqq t < t_1$, of (3.15) and (3.17) is necessary for the assumed existence of k_x and is equivalent to the nonexistence of conjugate points [10], [11]. Furthermore, since $\{x_*(t), u_*(t)\}$ provides a minimum of the performance index (2.2), it follows that $P(t), t_0 \leqq t \leqq t_1$, is at least nonnegative definite [11], [19].

We can now prove the following result.

LEMMA. *A necessary and sufficient condition for the optimal system ofplant* (2.1), *performance index* (2.2), *constraint* (2.3) *on u, and satisfying the smoothness assumptions, to exhibit a closed loop sensitivity reduction to continuous first order parameter variations according to the criterion* (2.10), *with $Z(t)$ a nonnegative definite matrix given by*

$$Z(t) = -P(t)\mathscr{H}_{\lambda\lambda}(t)P(t), \tag{3.18}$$

is that

$$(3.19)\quad v^T(t')P(t')v(t') + \int_{t_0}^{t'} [v^T(\mathscr{H}_{xx} + 2c_x^T f_u^T P)v + 2\delta x_c^T c_x^T f_u^T Pv]\, dt \geqq 0,$$

where $v(t)$ *is given by* (3.6). *If there is no constraint on* u, (3.18) *and* (3.19) *are replaced by, respectively,*

$$(3.20)\qquad Z(t) = k_x^T(t, x_*(t))H_{uu}(t)k_x(t, x_*(t))$$

and

$$(3.21)\qquad v^T(t')P(t')v(t') + \int_{t_0}^{t'} [v^T H_{xx} v + 2\delta x_c^T k_x H_{ux} v]\, dt \geqq 0.$$

Proof. Substituting into (3.15) the expression

$$\mathscr{H}_{\lambda x} = f_x + f_u c_x = \mathscr{H}_{x\lambda}^T,$$

premultiplying (3.15) by v^T and postmultiplying by v, where v is given by (3.6), and noting that

$$\dot v = f_x v + f_u k_x \delta x_c, \qquad v(t_0) = 0,$$

we have

$$(3.22)\quad -v^T\dot P v = 2v^T P\dot v - 2\delta x_c^T k_x^T f_u^T Pv + 2v^T P f_u c_x v + v^T P\mathscr{H}_{\lambda\lambda} Pv + v^T \mathscr{H}_{xx} v.$$

We now use

$$\mathscr{H}_{\lambda\lambda} = f_u c_\lambda$$

together with (3.14) to obtain

$$(3.23)\qquad k_x^T f_u^T P = -c_x^T f_u^T P - P\mathscr{H}_{\lambda\lambda} P.$$

Substituting (3.23) into (3.22) and rearranging, we have

$$(3.24)\quad 2\delta x_c^T Zv + v^T Zv = \frac{d}{dt}(v^T Pv) + v^T \mathscr{H}_{xx} v + 2\delta x_c^T c_x f_u Pv + 2v^T P f_u c_x v,$$

where Z is given by (3.18). It is shown in [11] that $\mathscr{H}_{\lambda\lambda}$ is nonpositive definite, and hence, $Z(t)$ is nonnegative definite. By integrating (3.24) between t_0 and t' and noting that $v(t_0) = 0$, the right side of (3.24) becomes equal to the left side of the condition (3.9), which has already been shown to be necessary and sufficient for (2.10) to hold. This proves the lemma. for the constrained case.

The unconstrained case can be proved similarly by using (3.16) and (3.17), or by identifying the expressions (3.14) and (3.16) for k_x. It is noted that by optimality, H_{uu} is nonnegative definite (the Lagendre-Clebsch necessary condition) and hence $Z(t)$ is nonnegative definite. (In fact, H_{uu} is positive definite. This is necessary for the assumed existence of k_x where, by (3.16), the existence of H_{uu}^{-1} is needed.) This completes the proof.

The chief obstacles to demonstrating that (3.19) and (3.21) hold are the terms containing c_x and H_{ux}. These terms vanish for the class of problems, considered next, where the equations are separable in x and u.

4. A class of optimal systems. Consider now the case where a plant is given by

$$\dot{x} = a(t, x, \mu) + b(t, u, \mu), \qquad x(t_0) = x_0, \tag{4.1}$$

where a and b are n-vectors continuous in t, continuously differentiable in μ, and twice continuously differentiable in x and u. The performance index is

$$I = g(x(t_1), t_1) + \int_{t_0}^{t_1} [q(t, x) + r(t, u)]\, dt, \tag{4.2}$$

where g, q and r are scalar functions continuous in t and twice continuously differentiable in x and u. The rest of the optimization problem is as in § 2 and § 3.

In addition to the smoothness assumptions, which are, as already noted, quite standard, it will be assumed that H_{xx} is nonnegative definite along the nominal optimal solution:

$$H_{xx} \geqq 0. \tag{4.3}$$

This assumption is sufficient, but not quite necessary, to guarantee that the Riccati equations for this class of problems, corresponding to (3.15) and (3.17) for the general problem, have finite solutions on $t_0 \leqq t < t_1$, i.e., that there will be no conjugate points [11], [19]. This assumption is routinely made for the linear problem, where H_{xx} corresponds to the matrix Q of (4.8) below.

Except for the not unreasonable assumption (4.3), the problem is a special case of the problem of § 2, and as a direct consequence of the lemma of § 3 we have the following main result.

THEOREM. *For an optimal system, where* (4.2) *is minimized subject to* (4.1) *and* (2.3), *and which satisfies the smoothness assumptions and assumption* (4.3), *the nominally equivalent closed loop system is less sensitive than the open loop system to continuous first order parameter variations according to the criterion* (2.10), *with Z a nonnegative definite matrix given by* (3.18), *or in the case where u is unconstrained, by* (3.20). *The equality sign in* (2.10) *occurs if and only if*

$$v^T(t')P(t')v(t') + \int_{t_0}^{t'} v^T H_{xx} v\, dt = 0, \tag{4.4}$$

where $v(t)$ is given by (3.6) *with $f_u = b_u$. If H_{xx} is positive definite along the nominal optimal solution, then the equality sign in* (2.10) *occurs if and only if*

$$\delta x_c(t) \equiv \delta x_o(t) \quad \text{on } t_0 \leqq t \leqq t'. \tag{4.5}$$

Proof. First it is noted that for the class of problems on hand

$$H_{ux} = c_x = 0 \quad \text{and} \quad \mathscr{H}_{xx} = H_{xx},$$

so that both conditions (3.19) and (3.21) become

$$v^T(t')P(t)v(t') + \int_{t_0}^{t'} v^T H_{xx} v\, dt \geqq 0. \tag{4.6}$$

By the lemma, the inequality (2.10) holds if and only if (4.6) holds, and the equality sign in (2.10) occurs if and only if (4.6) is satisfied with an equality. Since $P(t)$ and $H_{xx}(t)$ are nonnegative definite on $t_0 \leqq t \leqq t_1$, inequality (4.6) holds. If H_{xx} is positive definite, then an equality in (4.6) can occur if and only if $v(t) \equiv 0$ on $t_0 \leqq t \leqq t'$, which by (3.7) implies (4.5). This completes the proof.

A special case of (4.1) and (4.2) is represented by the linear equation

$$\dot{x} = A(t,\mu)x + B(t,\mu)u, \qquad x(t_0) = x_0 \tag{4.7}$$

and quadratic performance index

$$I = \frac{1}{2}x^T(t_1)Dx(t_1) + \frac{1}{2}\int_{t_0}^{t_1} [x^TQ(t)x + u^TR(t)u]\,dt, \tag{4.8}$$

where t_1 is specified, $Q(t)$ and D are nonnegative definite, and $R(t)$ is positive definite. With an unconstrained u and a free endpoint $x(t_1)$, this problem is the standard linear optimal regulator problem [19]. In contrast to the nonlinear problem, the existence of a unique optimal control and the existence of a linear feedback function

$$u(t,x) = K(t)x, \tag{4.9}$$

defined for all x and all $t \leqq t_1$, do not have to be assumed, but are guaranteed for this problem [19]. Since here

$$H_{xx} = Q, \quad H_{uu} = R, \quad k_x = K, \tag{4.10}$$

there is no need to assume (4.3), and the matrix $Z(t)$ in (2.10) is given by

$$Z(t) = K^T(t)RK(t). \tag{4.11}$$

This is directly proved in [2].

It is worth pointing out that in the time-invariant case (when all the matrices are time-invariant, and in (4.8), $t_0 = 0$, $t_1 = \infty$ and $D = 0$), if the completely controllable plant is in the phase-variable canonical form

$$\begin{aligned} \dot{x}^i &= x^{i+1}, \qquad\qquad\qquad\qquad i = 1,2,\cdots,n-1, \\ \dot{x}^n &= a_1x^1 + a_2x^2 + \cdots + a_nx^n + u, \end{aligned} \tag{4.12}$$

where u is scalar and x^i is the ith component of x, then, as shown in [1], (2.10) reduces to the strong and desirable form

$$\int_0^{t'} [\delta x_c^i(t)]^2\,dt < \int_0^{t'} [\delta x_o^i(t)]^2\,dt \quad \text{for all } t' > 0, i = 1,2,\cdots,n, \tag{4.13}$$

for first order continuous time-varying parameter variations.

In closing, it is noted that the symbol μ in (2.1) can represent external disturbances as well as internal parameters; thus the results of this paper apply to sensitivity of first order disturbance variations. A difference exists, however, for linear systems when the disturbances enter the plant in a linear additive fashion as an extra term in (4.7), say $+C(t)w$, where C is a matrix and w is a vector disturbance; then the results apply for finite, not necessarily small, disturbance variations $\Delta w(t)$, or if $w_*(t) \equiv 0$, for a finite disturbance $w(t)$.

5. Discussion. Concerning the results of the paper, several critical points can be made: The results are qualitative rather than quantitative; they are limited to a special class of problems; and the sensitivity measure (2.19) with $Z(t)$ given by (3.19) is ad hoc—perhaps a different measure would yield different results.

These points will be discussed in reverse order. There is no universally accepted measure of sensitivity. Indeed, in the writer's opinion, there is no need for one as each application suggests its own measure. The writer takes the position that any sensitivity measure is valid if it has practical significance and analytical usefulness. The sensitivity measure in (2.10) is natural for linearized systems and, as pointed out, is in the classical tradition. Since the upper limit of the integral is arbitrary, the criterion (2.10) is quite strong. It is conceivable, although it does not appear likely, that (2.10) with $Z(t)$ different from (3.18) and (3.20) may be discovered for the smooth Bolza optimization problem formulated here. A different weighting matrix Z may apply to a particular problem or some subclass of problems. For example, the criterion (2.10) with $Z = I$ is shown to be violated for one example of linear optimal systems and to hold for a second example [1]. The numerical experience of [1] suggests that Z given by (4.11) weights the components of δx approximately as the performance index (4.2) weights those of x. In this respect, Z given by (4.11) is better than the choice $Z = I$. A further vindication of the particular Z discovered in this investigation may be seen in the fact that for an important class of linear, time-invariant systems, (2.10) with Z given by (3.20) reduces to (4.13); again, it is conceivable, but unlikely, that a different Z will do the same. An example of a different measure of sensitivity is the magnitude of the first order variation δI of the performance index I. This is a logical measure of sensitivity and it does indeed lead to a different result: Witsenhausen [20] has shown that for sufficiently smooth systems,δI is the same for nominally equivalent open loop and closed loop optimal systems, provided the endpoint $x(t_1)$ is free.

The optimization problem considered in this paper is actually quite general because, except for certain special problems and relatively simple bang-bang systems, it is the only general problem that admits even a local feedback solution. The fact that, to demonstrate closed loop sensitivity reduction, it appears necessary to further limit consideration to problems separable in control and state is an interesting consequence of this investigation. For problems with a discontinuous control, an entirely different sensitivity analysis must be employed [21].

As to the quantitative aspect of the closed loop sensitivity reduction, the numerical experience of [1] suggests that the larger the weight of x in the performance index, the smaller the magnitude of δx_c. The theoretical possibility of having (2.10) with an equality suggests that for some $\delta\mu(t)$ the closed loop sensitivity reduction may be negligible. If this is a serious problem, then one can resort to special devices that take sensitivity into account by introducing sensitivity terms involving δx into the performance index [22]. It would be interesting to compare the closed loop sensitivity reduction for systems with and without the constraint (2.3) on u, but this appears possible only via numerical examples. Indications are that the more severe the constraint, the less is the closed loop sensitivity reduction [17].

REFERENCES

[1] E. Kreindler, *Closed loop sensitivity reduction of linear optimal control systems*, IEEE Trans. Automatic Control, AC-13 (1968), pp. 254–262.

[2] ———, *Sensitivity of time-varying linear optimal control systems*, J. Optimization Theory and Applications, 3 (1969), pp. 98–106.

[3] R. E. Kalman, *When is a linear control system optimal?* Trans. ASME Ser. D J. Basic Engrg., 86 (1964), pp. 51–60.

[4] B. D. O. Anderson, *The inverse problem of optimal control*, Proc. 4th Annual Allerton Conference, 1966.

[5] ———, *Sensitivity improvement using optimal control*, Proc. IEE, 113 (1966), pp. 1084–1086.

[6] J. B. Cruz, Jr. and W. R. Perkins, *A new approach to the sensitivity problem in multivariable feedback system design*, IEEE Trans. Automatic Control, AC-9 (1964), pp. 216–223.

[7] W. R. Perkins and J. B. Cruz, Jr., *The parameter variation problem in state feedback control systems*, Trans. ASME Ser. D.J. Basic Engrg., 87 (1965), pp. 120–124.

[8] J. V. Breakwell, J. L. Speyer and A. E. Bryson, *Optimization and control of nonlinear systems using the second variation*, this Journal, 1 (1963), pp. 193–223.

[9] H. J. Kelley, *Guidance theory and extremal fields*, IRE Trans. Automatic Control, AC-7 (1962), pp. 75–82.

[10] A. E. Bryson and Y. C. Ho, *Optimization, estimation and control*, Applied Optimal Control, Blaisdell, Waltham, Massachusetts, 1969.

[11] R. E. Kalman and T. S. Englar, *A user's manual for the automatic synthesis program*, Rep. CR-475, NASA, 1966, Chap. VII.

[12] L. S. Pontryagin, *Ordinary Differential Equations*, Addison-Wesley, Reading, Massachusetts, 1962.

[13] E. Kreindler, *On the definition and application of the sensitivity function*, J. Franklin Inst., 285 (1968), pp. 26–36.

[14] W. R. Perkins and J. B. Cruz, Jr., *Sensitivity operators for linear time-varying systems*, Sensitivity Methods in Control Theory, L. Radanović, ed., Pergamon, New York, 1966.

[15] B. D. O. Anderson and R. W. Newcomb, *An approach to the time-varying sensitivity problem*, SIAM J. Appl. Math., 15 (1967), pp. 1001–1010.

[16] J. B. Cruz, Jr., and W. R. Perkins, *Criteria for system sensitivity to parameter variations*, Proc. Third Congress of International Federation for Automatic Control, London, 1966, pp. 18C.1–18C.7.

[17] E. Kreindler, *On the closed-loop sensitivity reduction of nonlinear systems*, Internat. J. Control, 6 (1967), pp. 171–178.

[18] W. A. Porter, *Sensitivity problems in distributive systems*, Ibid, 5 (1967), pp. 393–412.

[19] R. E. Kalman, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana, 5 (1960), no. 2, pp. 102–119.

[20] H. S. Witsenhausen, *On the sensitivity of optimal control systems*, IEEE Trans. Automatic Control, AC-10 (1965), pp. 495–496.

[21] E. Kreindler, *On sensitivity of time-optimal systems*, Ibid, AC-14 (1969).

[22] ———, *On minimization of trajectory sensitivity*, Internat. J. Control, 8 (1968), pp. 89–96.

SIAM J. CONTROL
Vol. 7, No. 4, November 1969

REACHABILITY OF PERTURBED SYSTEMS AND MIN SUP PROBLEMS*

M. C. DELFOUR AND S. K. MITTER†

1. Introduction. The problem of reachability for control processes, that is, the problem of finding an admissible control which steers the system into some target set is a preliminary one in the study of optimal control problems. For linear control processes in both finite and infinite dimensions, reachability has been discussed by several authors [1], [2], [3], [4], [5], [6], [7], [8]. The reachability problem for nonlinear differential control processes has also recently been investigated [9].

Control problems in the presence of disturbances have usually been treated as stochastic control problems. However, in many control problems statistics of the disturbance are not available. For such problems a natural way to model the disturbance is to assume that it belongs to some fixed bounded set in the space of disturbances.

The objective of this paper is to study such control processes in the presence of additive disturbances. We introduce the concept of strong reachability. A control process is said to be strongly reachable if there exists an admissible control which steers the system to the given target set in the presence of the worst disturbance. We show that the problem of finding the best open loop control in the presence of the worst disturbance is related to the concept of strong reachability. For the problem studied in this paper, the operations of finding infimum and supremum are not interchangeable and hence game-theoretical techniques used for example in [10], [11], [12], [13] are not applicable. We present a geometrical necessary and sufficient condition for strong reachability. For linear control processes with closed, bounded, convex control constraint and disturbance sets and a closed, convex target set, the geometrical condition can be translated into analytical form by using separation and embedding theorems for convex sets. By specializing to the case where the control constraint set, the disturbance set and the target set are balls, we can obtain an analytical necessary and sufficient condition in explicit form and also obtain expressions for the minimum norm control, maximum norm disturbance and the minimum target set radius in terms of the control process data. The final section considers applications to control processes described by differential equations. For some work related to this paper see [14].

Notation. For a map $f:X \to Y$, if $A \subset X$, $f(A) = \{f(x)|x \in A\}$. $\{x\}$ is the set consisting of the single element x. For two sets A, B which are subsets of a Banach Space X, $A + B = \{a + b|a \in A, b \in B\}$.

Let X be a Banach space and let X^* be its topological dual space. We define the symbol $\langle x, x^*\rangle$ by $\langle x, x^*\rangle = x^*(x)$, where the right-hand side is the value of the linear form x^* at the point x. The map $(x, x^*) \mapsto \langle x, x^*\rangle$ is a bilinear form on $X \times X^*$. $\mathscr{L}(X, Y)$ denotes the space of continuous linear maps from X into Y. For $S \in \mathscr{L}(X, Y)$, S^* is the adjoint linear map and $S^* \in \mathscr{L}(Y^*, X^*)$.

* Received by the editors August 21, 1968, and in revised form March 11, 1969.

† Systems Research Center, Case Western Reserve University, Cleveland, Ohio 44106. This work was supported by the National Science Foundation under Grants GK-600 and GK-3714.

2. Basic definitions.

2.1. Mathematical definition of the system. Let X_1, X_2, be reflexive Banach spaces and let X_3 be a Banach space. X_1 is to be thought of as the *control space*, X_2 the *disturbance space* and X_3 the *state space* of a control system. Let U, W and B be subsets of X_1, X_2 and X_3 respectively. R denotes the real line.

Let $S:X_1 \to X_3$ (or $u \mapsto Su$), and $T:X_2 \to X_3$ (or $w \mapsto Tw$) be continuous (not necessarily linear) maps and let s be a given element in X_3. We shall be concerned with the *abstract control process*, (C), defined by the equation,

$$x = s + Su + Tw. \qquad \text{(C)}$$

A control $u \in U$ will be called an *admissible control* and a disturbance $w \in W$ will be called an *allowable disturbance*. The set B will be referred to as the *target set*.

2.2. Definition of strong reachability and reachability. For the system (C),

(i) the target set B is said to be *strongly reachable* from s if there exists a $\bar{u} \in U$ such that $s + S\bar{u} + Tw \in B$ for all w in W;

(ii) the target set B is said to be *reachable* from s if for any w in W, there exists a $\bar{u} \in U$ such that $s + S\bar{u} + Tw \in B$.

We shall refer to the system (C) as being strongly reachable (reachable) when we mean B is strongly reachable (reachable) from s.

2.3. Definition of the min sup problem. For the function $f:X_3 \to R$ let $q:X_1 \times X_2 \to R$ be the mapping defined by $(u, w) \mapsto f(s + Su + Tw)$.

The min sup problem can now be formulated as follows: Given the function f, does there exist a $\hat{u} \in U$ such that

$$\sup\{q(\hat{u}, w):w \in W\} = \inf\{\sup[q(u, w):w \in W]:u \in U\}.$$

3. Geometrical necessary and sufficient condition for strong reachability and reachability. To obtain a necessary and sufficient condition for strong reachability we introduce two sets.

DEFINITION 3.1. For the control system (C), the *unperturbed attainable set* is defined as the set

$$A = \{s\} + S(U) \qquad (3.1)$$

and the *modified target set* is defined as the set

$$M = \{x \in X_3 | \{x\} + T(W) \subset B\}. \qquad (3.2)$$

Our first theorem identifies in geometrical form the necessary and sufficient conditions for strong reachability.

THEOREM 3.2. *The system* (C) *is strongly reachable if and only if* $A \cap M$ *is not empty.*

Proof. If (C) is strongly reachable, there exists an admissible control $\bar{u}$ such that

$$\{s\} + \{S\bar{u}\} + T(W) \subset B$$

and hence $s + S\bar{u} \in M$. However, $\bar{u}$ being admissible implies $s + S\bar{u} \in A$ and hence $A \cap M \neq \emptyset$. Conversely $A \cap M \neq \emptyset$ implies that there exists an $x \in X_3$ such that $x \in A$ and $x \in M$. Since $x \in A$, there exists a $\bar{u}$ admissible such that $x = s + S\bar{u}$. Likewise $x \in M$ implies $\{x\} + T(W) \subset B$ and hence $s + S\bar{u} + Tw \in B$, for every $w \in W$.

COROLLARY 3.3. *The system* (C) *is reachable if and only if for every* $w \in W$ $A \cap (B - \{Tw\}) \neq \emptyset$.

Remark. Theorem 3.2 does not make use of any topological properties and hence is true for a control process defined on linear spaces.

4. The min sup problem.

This section contains two theorems. The first theorem exhibits the relationship between the problem of existence of a solution to the min sup problem with that of strong reachability with respect to an appropriately constructed target set. The second theorem shows that under certain assumptions on the function f and the unperturbed attainable set A, the min sup problem has a solution.

THEOREM 4.1. *Let,*

$$\varepsilon^* = \inf\{\sup[f(s + Su + Tw):w \in W]:u \in U\} \tag{4.1}$$

and for any $\varepsilon \in R$, *define* $B(\varepsilon) = f^{-1}((-\infty, \varepsilon])$. *Then*

(i) *for every* $\varepsilon \in R$, *the control process* (C) *is strongly reachable with respect to the target* $B(\varepsilon)$ *if and only if* $\varepsilon^* = -\infty$.

(ii) *There exists no* $\varepsilon \in R$ *such that the control process* (C) *is strongly reachable with respect to the target* $B(\varepsilon)$ *if and only if* $\varepsilon^* = +\infty$.

(iii) *If there exists an* $\bar{\varepsilon} \in R$, *such that for any* $\varepsilon \in R$, (C) *is strongly reachable for the target* $B(\varepsilon)$ *if and only if* $\varepsilon \geqq \bar{\varepsilon}$, *then there exists a* $\hat{u} \in U$ *such that*

$$\sup[f(s + S\hat{u} + Tw):w \in W] = \varepsilon^* = \bar{\varepsilon}. \tag{4.2}$$

(iv) *Conversely, if there exists a* $\hat{u} \in U$ *such that*

$$\sup[f(s + S\hat{u} + Tw):w \in W] = \varepsilon^*$$

and $\varepsilon^* \in R$, *then for any* $\varepsilon \in R$, (C) *is strongly reachable for the target* $B(\varepsilon)$ *if and only if* $\varepsilon \geqq \varepsilon^*$.

Proof. (i) If $\varepsilon^* = -\infty$, then for every $\varepsilon \in R$, there exists a $u \in U$ such that $\sup\{f(s + Su + Tw):w \in W\} \leqq \varepsilon$; that is, for every $\varepsilon \in R$, the control process (C) is strongly reachable for the target $B(\varepsilon)$. Conversely, for an arbitrary $\varepsilon \in R$, $A \cap M(\varepsilon) \neq \emptyset$ implies the existence of a $u \in U$ such that $\{s\} + \{Su\} + T(W) \subset B(\varepsilon)$; that is $\sup\{f(s + Su + Tw):\ w \in W\} \leqq \varepsilon$. Since ε is arbitrary, take $\varepsilon = -n$, where n is a positive integer. For each n, there exists a $u_n \in U$ such that $\sup\{f(s + Su_n + Tw):w \in W\} \leqq -n$ and hence

$$\varepsilon^* \leqq \lim_{n \to \infty} \{\sup[f(s + Su_n + Tw):w \in W]\} \leqq \lim_{n \to \infty} (-n) = -\infty$$

which implies that $\varepsilon^* = -\infty$.

(ii) The following chain of statements are equivalent to $\varepsilon^* = +\infty$:
for every $u \in U$, $\sup\{f(s + Su + Tw):w \in W\} = +\infty$,
for every $u \in U$, $\varepsilon \in R$, $\{s\} + \{Su\} + T(W)$ is not a subset of $B(\varepsilon)$,
for every $\varepsilon \in R$, there exists no $u \in U$ such that $\{s\} + \{Su\} + T(W) \subset B(\varepsilon)$,
for every $\varepsilon \in R$, $A \cap M(\varepsilon) = \phi$; and hence the desired conclusion.

(iii) For $\varepsilon < \bar{\varepsilon}$, there exists no $u \in U$ such that $M(\varepsilon) \cap A \neq \emptyset$; that is for every $u \in U$, $\sup\{f(s + Su + Tw) : w \in W\} \geqq \bar{\varepsilon}$, and then $\varepsilon^* \geqq \bar{\varepsilon}$. However, $M(\bar{\varepsilon}) \cap A \neq \emptyset$ guarantees the existence of a $u^* \in U$ such that $\sup\{f(s + Su^* + Tw) : w \in W\} \leqq \bar{\varepsilon}$. Hence for this particular $u^* \in U$, $\sup\{f(s + Su^* + Tw) : w \in W\} = \bar{\varepsilon} = \varepsilon^*$.

(iv) Since $A \cap M(\varepsilon^*) \neq \emptyset$, for every $\varepsilon \geqq \varepsilon^*$, $B(\varepsilon) \supset B(\varepsilon^*)$ and $M(\varepsilon) \supset M(\varepsilon^*)$ and finally $M(\varepsilon) \cap A \supset M(\varepsilon^*) \cap A \neq \emptyset$. So (C) is strongly reachable for $\varepsilon \geqq \varepsilon^*$. Now if $\varepsilon < \varepsilon^*$, there exists no $u \in U$ such that $\sup\{f(s + Su + Tw) : w \in W\} \leqq \varepsilon$; that is $\{s\} + \{Su\} + T(W)$ is not a subset of $B(\varepsilon)$ and $M(\varepsilon) \cap A = \emptyset$ for $\varepsilon < \varepsilon^*$.

THEOREM 4.2. *Let $f : X_3 \to R$ be lower semicontinuous and let the unperturbed attainable set A be compact* (*in an appropriate topology of X_3*). *Then there exists a $\hat{u} \in U$ such that*

$$(4.3) \quad \sup[f(s + S\hat{u} + Tw) : w \in W] = \inf[\sup\{f(s + Su + Tw) : w \in W\} : u \in U].$$

Proof.

$$\inf\{\sup[f(s + Su + Tw) : w \in W] : u \in U\} = \inf\{\sup[f(x + y) : y \in T(W)] : x \in A\}.$$

Let $q(x, y) = f(x + y)$. Since f is lower semicontinuous, for fixed x the function $q_x(y) = q(x, y)$ is lower semicontinuous. Since the upper envelope of a family of lower semicontinuous functions is lower semicontinuous [15, p. 362, Theorem 4] and A is compact, there exists an $\hat{x} \in A$ such that

$$\sup[q(\hat{x}, y) : y \in T(W)] = \inf\{\sup[q(x, y) : y \in T(W)] : x \in A\}.$$

Hence from the definition of A there exists a $\hat{u} \in U$ such that (4.3) is true.

COROLLARY 4.3. *Let $f : X_3 \to R$ be continuous and let the sets A and $T(W)$ be compact* (*in an appropriate topology of X_3*). *Then there exists a $\hat{u} \in U$ and a $\hat{w} \in W$ such that*

$$f(s + S\hat{u} + T\hat{w}) = \inf\{\sup[f(x + Su + Tw) : w \in W] : u \in U\}.$$

Remark. Theorems 4.1 and 4.2 and Corollary 4.3 are true, for example in linear topological spaces which are Hausdorff.

COROLLARY 4.4. *Let $f : X_3 \to R$ be convex and strongly lower semicontinuous and A be weakly compact. Then Theorem* 4.2 *holds.*

Proof. The proof follows from the fact that a lower semicontinuous convex function defined on a Banach space is weakly lower semicontinuous.

5. Strong reachability of linear control processes. In this section we investigate strong reachability for linear control processes in a Banach space setting.

Let the assumptions of § 2.1 on the control system (C) hold. Further, let U be a closed, bounded, convex subset of X_1, W a closed, bounded, convex subset of X_2 and B a closed, convex subset of X_3. The maps S and T defined in § 2.1 are now assumed to belong to the spaces $\mathscr{L}(X_1, X_3)$ and $\mathscr{L}(X_2, X_3)$ respectively. The linear control system so obtained will now be referred to as (L).

We shall use the separation theorem and embedding theorem for convex sets to translate Theorem 3.2 into analytical form. The following *separation theorem* is an immediate consequence of the strong separation theorem for convex sets in locally convex topological vector spaces [16, p. 119, Corollary 14.4 and p. 23, Theorem 3.9 and p. 14].

THEOREM 5.1. *Let X be a Banach space, let A be a weakly compact, convex subset of X and let B be a closed, convex subset of X. Then*

$$A \cap B \neq \varnothing$$

if and only if

$$\sup\{\inf[\langle x, x^*\rangle : x \in B] - \sup[\langle x, x^*\rangle : x \in A] : \|x^*\|_{X^*} = 1\} \leqq 0.$$

We also quote the following embedding theorem due to Hörmander [17].

DEFINITION 5.2. Let X be a locally convex topological vector space and let K be a nonempty, closed, convex subset of X. The *support functional* $H(x^*)$ of K, $x^* \in X^*$, is defined by

$$H(x^*) = \sup[\langle x, x^*\rangle : x \in K].$$

THEOREM 5.3. *Let K_1 and K_2 be two closed convex sets, and let $H_1(x^*)$ and $H_2(x^*)$ be their corresponding support functionals. Then*

(i) $K_1 \subseteq K_2$ *if and only if* $H_1(x^*) \leqq H_2(x^*)$, *for every* $x^* \in X^*$

(ii) $K_1 = K_2$ *if and only if* $H_1(x^*) = H_2(x^*)$, *for every* $x^* \in X^*$.

5.1. Analytical necessary and sufficient conditions. In order to invoke the above theorems, it is necessary to establish some topological properties for the sets A and M.

PROPOSITION 5.4. *For the control system* (L), *the unperturbed attainable set A and the set $\{x\} + T(W)$ are convex and weakly compact.*

Proof. The proof of this proposition is an immediate consequence of the linearity and weak continuity of the maps S and T[18, p. 422, Theorem 15] and the weak-compactness of the sets U and W [18, p. 425, Corollary 8].

PROPOSITION 5.5. *The target set B and the modified target set M are weakly closed and convex.*

Proof. Since B is convex and closed, it is weakly closed. The convexity of M is obvious. We shall show M is a strongly closed subset of X_3. Consider a strong Cauchy sequence $\{x_n\}$ in M. Since $M \subset X$, $x_n \to x$, where $x \in X$. For any $w \in W$, the translated sequence $\{x_n + Tw\}$ is Cauchy and $x_n + Tw \to x + Tw$. However, as points of M the x_n's are such that $x_n + Tw \in B$, for every $w \in W$. But since B is strongly closed, $x + Tw \in B$, for every $w \in W$, and $x + T(W) \subset B$ which implies $x \in M$. Hence M is strongly closed and being convex is thus weakly closed.

PROPOSITION 5.6. *Given the system* (L), *the set M is given by*

$$M = \{x \in X_3 : h(x) \leqq 0\},$$

where

$$h(x) = \sup[\langle x, x^*\rangle + \sup(\langle w, T^*x^*\rangle : w \in W) - \sup(\langle y, x^*\rangle : y \in B) : \|x^*\|_{X_3^*} = 1]. \tag{5.1}$$

Further, M is nonempty if and only if

$$\inf[h(x) : x \in X_3] \leqq 0. \tag{5.2}$$

Proof. From Theorem 5.3, $\{x\} + T(W) \subset B$ if and only if $H_{\{x\}+T(W)}(x^*) \leqq H_B(x^*)$, for every $x^* \in X_3^*$. That is,

$$\sup[\langle x, x^*\rangle + H_{T(W)}(x^*) - H_B(x^*) : \|x^*\|_{X_3^*} = 1] \leqq 0, \tag{5.3}$$

and hence from the definition of $H_{T(W)}(x^*)$ and $H_B(x^*)$ we obtain (5.1) and the definition of the elements of M.

If M is nonempty, (5.2) clearly holds. To prove the proposition in the other direction we use the fact that $h(x)$ satisfies $|h(x_2) - h(x_1)| \leqq \|x_2 - x_1\|_{X_3}$, for every $x_1, x_2 \in X_3$ and hence $h(x)$ is continuous.

There are two cases to consider. If $\inf[h(x): x \in X_3] < 0$, then clearly there exists $x \in X_3$ such that $x + T(W) \subset B$, and M is not empty. If $\inf[h(x): x \in X_3] = 0$, then there exists a sequence $\{x_n\}$ such that $h(x_n) \to 0$; since h is continuous on X_3, there exists an $x \in X_3$ such that $x_n \to x$ and $h(x) = 0$. So again M is nonempty.

THEOREM 5.7. *The system* (L) *is strongly reachable if and only if*

$$(5.4)\quad \begin{aligned} &\inf\{\sup[\langle x, x^*\rangle + \sup(\langle w, T^*x^*\rangle : w \in W) \\ &\qquad - \sup(\langle y, x^*\rangle : y \in B) : \|x^*\|_{X_3^*} = 1] : x \in X_3\} \leqq 0 \end{aligned}$$

and

$$(5.5)\quad \begin{aligned} &\sup(\langle s, x^*\rangle + \inf(\langle u, S^*x^* \rangle : u \in U) \\ &\qquad - \sup(\langle y, x^*\rangle : y \in M) : \|x^*\|_{X_3^*} = 1) \leqq 0, \end{aligned}$$

where

$$(5.6)\quad \begin{aligned} M = \{x \in X_3 : &\sup[\langle x, x^*\rangle + \sup(\langle w, T^*x^*\rangle : w \in W) \\ &- \sup(\langle y, x^*\rangle : y \in B) : \|x^*\|_{X_3^*} = 1] \leqq 0\}. \end{aligned}$$

Proof. From Theorem 3.2, (L) is strongly reachable if and only if $A \cap M \neq \emptyset$. Using Propositions 5.4 and 5.5 the proof now follows directly from Theorem 5.1 and Proposition 5.6.

COROLLARY 5.8. *The system* (L) *is reachable if and only if*

$$(5.7)\quad \begin{aligned} &\sup\{\langle s, x^*\rangle + \inf(\langle u, S^*x^*\rangle : u \in U) + \sup(\langle w, T^*x^*\rangle : w \in W) \\ &\qquad - \sup(\langle y, x^*\rangle : y \in B) : \|x^*\|_{X_3^*} = 1\} \leqq 0. \end{aligned}$$

6. Specialization of the results of § 5. The results of the previous section will now be specialized to the case where

$$(6.1)\quad \begin{aligned} U &= \{u \in X_1 : \|u\|_{X_1} \leqq \rho\}, && 0 \leqq \rho < \infty, \\ W &= \{w \in X_2 : \|w\|_{X_2} \leqq \beta\}, && 0 \leqq \beta < \infty, \\ B &= \{x \in X_3 : \|x - x_d\|_{X_3} \leqq \varepsilon, x_d \in X_3 \text{ given}\}, && 0 \leqq \varepsilon < \infty. \end{aligned}$$

6.1. Strong reachability and reachability.

THEOREM 6.1. *The system* (L) *with* U, W *and* B *as defined in* (6.1) *is strongly reachable if and only if*

$$(6.2)\quad \beta\|T\| \leqq \varepsilon,$$

$$(6.3)\quad \sup\{\langle s - x_d, x^*\rangle - \rho\|S^*x^*\|_{X_1^*} - \sup[\langle x, x^*\rangle : x \in M_{x_d}] : \|x^*\|_{X_3^*} = 1\} \leqq 0,$$

where $M_{x_d} = M - \{x_d\}$ *and*

$$(6.4)\quad M_{x_d} = \{x \in X_3 : \sup[\langle x, x^*\rangle + \beta\|T^*x^*\|_{X_2^*} : \|x^*\|_{X_3^*} = 1] \leqq \varepsilon\}.$$

Proof. The proof follows from Theorem 5.7 by performing the necessary computations. In particular (5.4) becomes (6.2) as shown below. Equation (5.4) reduces to

$$\inf\{\sup[\langle x, x^*\rangle + \beta\|T^*x^*\|_{X_2^*} - \varepsilon : \|x^*\|_{X_3^*} = 1] : x \in X_3\} \leqq 0.$$

But $k(x) = \sup[\langle x, x^*\rangle + \beta\|T^*x^*\|_{X_2^*} : \|x^*\|_{X_3^*} = 1]$ is an even function of x. Thus

$$k(x) \geqq \sup[\beta\|T^*x^*\|_{X_2^*} : \|x^*\|_{X_3^*} = 1] = \beta\|T^*\| = \beta\|T\|.$$

Hence $\beta\|T\| \leqq \varepsilon$.

Theorem 6.1 can be sharpened somewhat in the sense that in calculating $\sup[\langle x, x^*\rangle : x \in M_{x_d}]$ we may restrict ourselves to x's which belong to the boundary of M_{x_d}. Moreover we can find an analytical expression for the boundary of M_{x_d}. This is done in the following two propositions.

PROPOSITION 6.2. *If $M \neq \varnothing$, its boundary ∂M (in the norm topology) is defined by*

$$\text{(6.5)} \quad \partial M = \{x \in X_3 : \sup[\langle x - x_d, x^*\rangle + \beta\|T^*x^*\|_{X_2^*} : \|x^*\|_{X_3^*} = 1] = \varepsilon\}.$$

Proof. Consider the function

$$f : X_3 \to R : x \mapsto \sup\{\langle x - x_d, x^*\rangle + \beta\|T^*x^*\|_{X_2^*} : \|x^*\|_{X_3^*} = 1\}.$$

From Theorem 5.3,

$$\text{(6.6)} \qquad M = \{x \in X_3 : f(x) \leqq \varepsilon\}.$$

(a) Let $x_0 \in M$ such that $f(x_0) = \varepsilon$ and assume x_0 is an interior point of M. Then there exists an open ball $B(x_0; \delta)$ with center at x_0 and radius δ such $B(x_0; \delta) \subset M$. However, $\sup\{f(x) : x \in B(x_0; \delta)\} \geqq \varepsilon + \delta$ which shows that for some $x \in B(x_0; \delta)$, $f(x) > \varepsilon$ which contradicts (6.6).

(b) Now let $x_0 \in \partial M$ and assume $\alpha = f(x_0) < \varepsilon$. Let $\delta = (\varepsilon - \alpha)/2$ and consider the open ball $B(x_0; \delta)$. For every $y \in B(x_0; \delta)$,

$$f(y) \leqq \alpha + \|y - x_0\|_{X_3} \leqq \alpha + \frac{\varepsilon - \alpha}{2} < \varepsilon,$$

which contradicts that $x_0 \in \partial M$.

PROPOSITION 6.3. *Let $M_{x_d} \neq \varnothing$. Then*

$$\sup\{\langle x, x^*\rangle : x \in M_{x_d}\} = \sup\{\langle x, x^*\rangle : x \in \partial M_{x_d}\}.$$

Proof. By Theorem 5.6, for any $x \in M_{x_d}$

$$\alpha = \sup\{\langle x, x^*\rangle + \beta\|T^*x^*\|_{X_2^*} : \|x^*\|_{X_3^*} = 1\} \leqq \varepsilon.$$

Let $x \in M_{x_d}$ and assume $x \neq 0$ (otherwise $M_{x_d} = \partial M_{x_d} = \{0\}$). Consider the real valued function f on $[0, \infty)$, defined by

$$f(c) = \sup\{c\langle x, x^*\rangle + \beta\|T^*x^*\|_{X_2^*} : \|x^*\|_{X_3^*} = 1\};$$

it is monotone increasing, convex and continuous on $[0, \infty)$. Moreover $f(1) \leqq \varepsilon$ and for $c_0 = (\varepsilon + \beta\|T\| + 1)/\|x\|_{X_3}$, $f(c_0) > \varepsilon$. Hence there exists a unique $\bar{c}$ in

$[1, \infty)$ such that $f(\bar{c}) = \varepsilon$. By Proposition 6.2, $\bar{c}x \in \partial M_{x_d}$. We then have for any x in M_{x_d}

$$|\langle x, x^*\rangle| \leqq |\langle \bar{c}x, x^*\rangle|,$$

$$\sup\{|\langle x, x^*\rangle| : x \in M_{x_d}\} \leqq \sup\{|\langle x, x^*\rangle| : x \in \partial M_{x_d}\}.$$

The theorem now follows from the linearity of the functional $\langle x, x^*\rangle$ and from the fact that M_{x_d} is symmetrical about 0 in X_3.

COROLLARY 6.4. *If X_3 is a reflexive Banach space, Theorem* 6.1 *holds with M_{x_d} in* (6.3) *replaced by the set of extreme points of M_{x_d}.*

Proof. The proof follows from the Krein–Millman theorem [16, p. 131, Theorem 15.1] and from a proposition in Bourbaki [19, p. 106, Proposition 1].

Remark. It might be useful to find a representation for the extreme points of M_{x_d}.

6.2. The characterization problem. If the system (L) is strongly reachable, then it is useful to characterize the minimum values of ρ and ε and the maximum value of β for which the system remains strongly reachable. This is done in the following theorems.

THEOREM 6.5 (minimum norm control). *For given β and ε assume that* (L) *is strongly reachable for some $\tilde{\rho}, 0 \leqq \tilde{\rho} < \infty$. Then there exists a minimum bound ρ^* for which* (L) *is strongly reachable. Moreover ρ^* is given by*

(6.7) (i) $\rho^* = 0$, *if* $g(0) \leqq 0$,

(ii) ρ^* *is the unique solution in* $[0, \tilde{\rho}]$ *of the equation* $g(\rho) = 0$ *if* $g(0) > 0$, *where*

$$g(\rho) = \sup\{\langle s - x_d, x^*\rangle - \rho\|S^*x^*\|_{X_1^*} - \sup[\langle x, x^*\rangle : x \in M_{x_d}] : \|x^*\|_{X_3^*} = 1\}.$$

Proof. Consider the function

$$f : (R^+ \cup \{0\}) \times X_3^* \to R : (\rho, x^*) \mapsto \langle s - x_d, x^*\rangle - \rho\|S^*x^*\|_{X_1^*}$$
$$- \sup[\langle x, x^*\rangle : x \in M_{x_d}]$$

and the function

$$g : R^+ \cup \{0\} \to R : \rho \mapsto \sup\{f(\rho, x^*) : \|x^*\|_{X_3^*} = 1\}.$$

We show that g is a monotonically decreasing, continuous convex function of ρ. For $\rho_2 \geqq \rho_1 \geqq 0, f(\rho_2, x^*) \leqq f(\rho_1, x^*)$, for every $x^* \in X_3^*$ and hence $g(\rho_2) \leqq g(\rho_1)$ showing that g is monotonically decreasing. For $\rho_2 \neq \rho_1$ and $\lambda \in [0, 1]$,

$$f(\lambda\rho_1 + (1 - \lambda)\rho_2, x^*) = \lambda f(\rho_1, x^*) + (1 - \lambda) f(\rho_2, x^*) \qquad \text{for every } x^* \in X_3^*,$$

which implies that g is convex.

Finally for $\rho_2 \geqq \rho_1 \geqq 0$,

$$|g(\rho_2) - g(\rho_1)| = g(\rho_1) - g(\rho_2) \leqq \sup[(\rho_2 - \rho_1)\|S^*x^*\|_{X_1^*} : \|x^*\|_{X_3^*} = 1]$$
$$\leqq |\rho_2 - \rho_1| \cdot \|S\|.$$

Since S is linear and continuous $\|S\| < \infty$. This shows that g is a continuous function of ρ.

Since (L) is strongly reachable for $\tilde{\rho} \geqq 0$, we have $g(\tilde{\rho}) \leqq 0$. There are two cases to consider:

(a) $g(0) \leqq 0$. Then the minimum bound $\rho^* = 0$.

(b) $g(0) > 0$. Then by virtue of the properties of the function $g(\rho)$, ρ^* is given by the unique solution in $[0, \tilde{\rho}]$ of $g(\rho) = 0$.

THEOREM 6.6 (maximum norm disturbance). *Given the bounds* ε *and* ρ, *assume that* (L) ($T \not\equiv 0$) *is strongly reachable for* $\beta = 0$. *Then there exists a maximum bound* β^* *such that* (L) *is strongly reachable if and only if* $\beta \leqq \beta^*$. *Moreover defining*

$$\bar{\beta} = \frac{\varepsilon}{\|T\|},$$

(i) $\beta^* = \bar{\beta}$, *if* $f(\bar{\beta}) \leqq 0$

(ii) $\beta^* = \hat{\beta}$, *if* $f(\bar{\beta}) > 0$, *where* $\hat{\beta}$ *is the unique solution of* $f(\beta) = 0$ *in* $[0, \bar{\beta})$. $f(\beta)$ *is defined by*

$$f(\beta) = \sup\{\langle s - x_d, x^*\rangle - \rho\|S^*x^*\|_{X_1^*} - \sup[\langle x, x^*\rangle : x \in M_{x_d}(\beta)] : \|x^*\|_{X_3^*} = 1\}.$$

Proof. A necessary condition for strong reachability of (L) is $M \neq \varnothing$ which implies $\beta \leqq \bar{\beta}$. We shall show that f is monotonically increasing, convex and continuous on $[0, \bar{\beta})$. Hence there are two cases:

(i) $f(\bar{\beta}) \leqq 0$. In that case (L) is strongly reachable and $\beta^* = \bar{\beta}$.

(ii) $f(\bar{\beta}) > 0$. Since $f(0) \leqq 0$ by hypothesis and $f(\bar{\beta}) > 0$, if we prove the asserted properties of the function $f(\beta)$, $f(\beta)$ has a unique solution $\hat{\beta}$ on $[0, \bar{\beta})$ and $\beta^* = \hat{\beta}$.

Now if $0 \leqq \beta_1 \leqq \beta_2 \leqq \bar{\beta}$, $\beta_1\|T\| \leqq \beta_2\|T\|$, which implies $M(\beta_1) \supset M(\beta_2)$ and using Theorem 5.3, $f(\beta_1) \leqq f(\beta_2)$ which shows that f is monotonically increasing.

Since $\lambda M(\beta_1) + (1 - \lambda)M(\beta_2) \subset M(\lambda\beta_1 + (1 - \lambda)\beta_2)$, for every $\lambda \in [0, 1]$, it follows from Theorem 5.3 that

$$f(\lambda\beta_1 + (1 - \lambda)\beta_2) \leqq \lambda f(\beta_1) + (1 - \lambda)f(\beta_2) \qquad \text{for every } \lambda \in [0, 1],$$

which shows that f is convex on $[0, \bar{\beta}]$.

The convexity of f implies continuity on $(0, \bar{\beta})$. Continuity at 0^+ can be demonstrated in a manner analogous to Theorem 6.5.

THEOREM 6.7 (minimum miss distance). *Given the bounds* ρ *and* β *there exists a minimum bound* ε^* *such that* (L) *is strongly reachable if and only if* $\varepsilon \geqq \varepsilon^*$. *Moreover defining* $\bar{\varepsilon} = \beta\|T\|$,

(i) $\varepsilon^* = \bar{\varepsilon}$ *if* $f(\bar{\varepsilon}) \leqq 0$,

(ii) $\varepsilon^* = \hat{\varepsilon}$, *if* $f(\bar{\varepsilon}) > 0$, *where* $\hat{\varepsilon}$ *is the unique solution of* $f(\varepsilon) = 0$ *in* $(\bar{\varepsilon}, \infty)$. $f(\varepsilon)$ *is defined by*

$$f(\varepsilon) = \sup\{\langle s - x_d, x^*\rangle - \rho\|S^*x^*\|_{X_1^*} - \sup[\langle x, x^*\rangle : x \in M_{x_d}(\varepsilon)] : \|x^*\|_{X_3^*} = 1\}.$$

Proof. We first show that for some $\varepsilon_\alpha \geqq 0$, the system (L) is strongly reachable. Let α and ε_α be defined as

$$\alpha = \sup\{\langle s - x_d, x^*\rangle - \rho\|S^*x^*\|_{X_1^*} : \|x^*\|_{X_3^*} = 1\},$$

$$\varepsilon_\alpha = |\alpha| + \bar{\varepsilon}.$$

Clearly $B(|\alpha|) = \{x \in X_3 : \|x\|_{X_3} \leqq |\alpha|\} \subset M_{x_d}(\varepsilon_\alpha)$ and for any $x^* \in X_3^*$, such that $\|x^*\| = 1$,

$$|\alpha| = \sup\{\langle x, x^*\rangle : x \in B(|\alpha|)\} \leqq \sup\{\langle x, x^*\rangle : x \in M_{x_d}(\varepsilon_\alpha)\}$$

which implies that (L) is strongly reachable for $\varepsilon = \varepsilon_\alpha$ since $f(\varepsilon_\alpha) \leqq 0$. In a manner analogous to the previous two theorems, it may be shown that f is monotonically decreasing, convex on $(\bar{\varepsilon}, \infty)$ and continuous on $(\bar{\varepsilon}, \varepsilon_\alpha]$. A necessary cóndition for strong reachability of (L) is $M \neq \varnothing$ which implies $\varepsilon \geqq \bar{\varepsilon}$. There are two cases to consider:

(i) $f(\bar{\varepsilon}) \leqq 0$. In that case (L) is strongly reachable and hence $\varepsilon^* = \bar{\varepsilon}$.

(ii) $f(\bar{\varepsilon}) > 0$. Since $f(\varepsilon_\alpha) \leqq 0$ and $f(\bar{\varepsilon}) > 0$, in view of the properties of f, $f(\varepsilon) = 0$ has a unique solution $\hat{\varepsilon}$ on $(\bar{\varepsilon}, \varepsilon_\alpha]$ and $\varepsilon^* = \hat{\varepsilon}$.

Remark. Theorems 6.5, 6.6 and 6.7 have obvious corollaries when strong reachability is replaced by reachability.

7. Applications to control processes described by differential equations. We shall illustrate the theory presented in the previous sections by considering its application to control processes described by differential equations.

7.1. Existence theorem for min sup problem. We consider an existence theorem for a min sup problem analogous to the existence theorem for optimal control problems.

Consider the perturbed control process in R^n

$$\frac{dx(t)}{dt} = A(t)x(t) + f(t, u(t)) + g(t, w(t)), \qquad t \in [0, t_1], \tag{7.1}$$

where $A(t)$ is a $n \times n$ measurable and bounded matrix on $[0, t_1]$, f is in C^1 in R^{1+m} and g is in C^1 in R^{1+p}, (n, m and p are integers $\geqq 1$). Furthermore,

(i) The initial state x_0 at time 0 is given.

(ii) The admissible controllers $\mathscr{F}$ consist of all Lebesgue measurable functions $t \mapsto u(t)$ on the compact interval $[0, t_1]$ such that $u(t) \in U$, (almost everywhere on $[0, t_1]$), where U is a compact set in R^m.

(iii) The admissible disturbances $\mathscr{G}$ consist of all Lebesgue measurable functions $t \mapsto w(t)$ on the compact interval $[0, t_1]$ such that $w(t) \in W$, almost everywhere on $[0, t_1]$, where W is a compact set in R^p.

(iv) The cost function for each admissible u and w is given by $C(u, w) = g(x(t_1))$, where g is a continuous function in R^n.

THEOREM 7.1. *For the above system, there exists a $\hat{u} \in \mathscr{F}$ and a $\hat{w} \in \mathscr{G}$ such that*

$$C(\hat{u}, \hat{w}) = \inf[\sup\{C(u, w) : w \in \mathscr{G}\} : u \in \mathscr{F}].$$

Proof. Since the differential equation (7.1) is linear in x, there exists an absolutely continuous function $t \mapsto x(t)$ defined on $[0, t_1]$ which satisfies (7.1) almost everywhere. Moreover by using the variation of parameters formula, the solution x of (7.1) at time t_1 may be written as,

$$\begin{aligned} x(t_1) = \phi(t_1)x_0 + \phi(t_1)\int_0^{t_1} \phi^{-1}(s)f(s, u(s))\,ds \\ + \phi(t_1)\int_0^{t_1} \phi^{-1}(s)g(s, w(s))\,ds, \end{aligned} \tag{7.2}$$

where ϕ is the usual transition matrix associated with (7.1).

Let $s \in R^n$ be defined by, $s = \phi(t_1)x_0$, and define the continuous nonlinear operators

$$S: L_1(R^m; 0, t_1) \to R^n : u \mapsto \phi(t_1) \int_0^{t_1} \phi^{-1}(s) f(s, u(s))\, ds,$$

$$T: L_1(R^p; 0, t_1) \to R^n : w \mapsto \phi(t_1) \int_0^{t_1} \phi^{-1}(s) g(s, w(s))\, ds,$$

where $L_1(X; 0, t_1)$ is the space of all integrable functions $t \mapsto x(t)$ with values in X. Equation (7.2) may then be written as:

$$x(t_1) = s + Su + Tw. \tag{7.3}$$

It follows from a result of Neustadt [20] that for the system (7.3) the unperturbed attainable set $\{s\} + S(U)$ and the set $T(W)$ are compact and hence the theorem follows from Corollary 4.3.

7.2. Strong functional reachability. Consider the linear differential control process

$$\frac{dx(t)}{dt} = A(t)x(t) + B(t)u(t) + C(t)w(t), \tag{7.4}$$

where $x(t) \in R^n$, $u(t) \in R^m$, $w(t) \in R^k$ (n, m, k are integers $\geqq 1$) and $A(t)$, $B(t)$, $C(t)$ are matrices of appropriate order which are measurable and bounded on the given compact interval $[0, t_1]$.

Let $1 < p < \infty$ and let $L^p(R^m; 0, t_1)$ be the reflexive Banach space of R^m-valued measurable functions such that

$$\int_0^{t_1} \|u(t)\|_{R^m}^p \, dt < \infty.$$

The Banach space $L^p(R^m; 0, t_1)$ is normed by

$$\|u\|_p = \left(\int_0^{t_1} \|u(t)\|_{R^m}^p \, dt \right)^{1/p}.$$

In a similar manner define $L^q(R^k; 0, t_1)$, $1 < q < \infty$, as the reflexive Banach space of all R^k-valued measurable functions with norm

$$\|w\|_q = \left(\int_0^{t_1} \|w(t)\|_{R^k}^q \, dt \right)^{1/q}$$

and $L^r(R^n; 0, t_1)$, $1 \leqq r \leqq \infty$, as the Banach space of all R^n-valued measurable functions with norm

$$\|x\|_r = \left(\int_0^{t_1} \|x(t)\|_{R^n}^r \, dt \right)^{1/r}.$$

Let the control restraint set Ω_u, the disturbance set Ω_w and the target set B be defined by

$$\begin{aligned} \Omega_u &= \{u : \|u\|_p \leqq \rho\}, \\ \Omega_w &= \{w : \|w\|_q \leqq \beta\}, \\ B &= \{x : \|x - x_d\|_r \leqq \varepsilon, x_d \text{ given element in } L^r(R^n; 0, t_1)\}. \end{aligned} \tag{7.5}$$

Since the differential equation is linear for given $x(0) \in R^n$, $u \in L^p(R^m; 0, t_1)$ and $w \in L^q(R^k; 0, t_1)$, there exists an absolutely continuous function $t \mapsto x(t)$ defined on the compact interval $[0, t_1]$ which satisfies (7.4) almost everywhere. Since $t \mapsto x(t)$ is absolutely continuous, $x \in L^r(R^n; 0, t_1)$. The solution of (7.4) is given by

$$\begin{aligned} x(t) = \phi(t)x(0) &+ \phi(t)\int_0^t \phi^{-1}(s)B(s)u(s)\,ds \\ &+ \phi(t)\int_0^t \phi^{-1}(s)C(s)w(s)\,ds, \qquad t \in [0, t_1]. \end{aligned} \tag{7.6}$$

Let $S: L^p(R^m; 0, t_1) \to L^r(R^n; 0, t_1)$ be the linear bounded transformation defined by

$$(Su)(t) = \phi(t)\int_0^t \phi^{-1}(s)B(s)u(s)\,ds, \qquad 0 \leqq t \leqq t_1,$$

and let $T: L^q(R^k; 0, t_1) \to L^r(R^n; 0, t_1)$ be the linear bounded transformation defined by

$$(Tw)(t) = \phi(t)\int_0^t \phi^{-1}(s)C(s)w(s)\,ds, \qquad 0 \leqq t \leqq t_1,$$

and let $s \in L^r(R^n; 0, t_1)$ be defined by $s(t) = \phi(t)x(0), 0 \leqq t \leqq t_1$.

Then (7.6) may be written as the operator equation

$$x = s + Su + Tw. \tag{7.7}$$

DEFINITION 7.2. The control process (7.7) is *strongly functionally reachable* with respect to $(s, \Omega_u, \Omega_w, B, t_1)$ if there exists a $\bar{u} \in \Omega_u$ such that $x(\cdot\,; \bar{u}, w) \in B$, for every $w \in \Omega_w$.

Necessary and sufficient conditions for strong functional reachability can now be obtained using the theory developed in previous sections.

REFERENCES

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

[2] R. CONTI, *Contributions to linear control theory*, J. Differential Equations, 1 (1965), pp. 427–445.

[3] ———, *On some aspects of linear control theory*, Mathematical Theory of Control, A. V. Balakrishnan and L. Neustadt, eds., Academic Press, New York, 1967.

[4] F. M. KIRILLOVA, *Applications of functional analysis to the theory of optimal processes*, this Journal, 5 (1967), pp. 25–50.

[5] S. K. MITTER, *Theory of inequalities and the controllability of linear systems*, Mathematical Theory of Control, A. V. Balakrishnan and L. Neustadt, eds., Academic Press, New York, 1967.

[6] H. A. ANTOSIEWICZ, *Linear control systems—controllability*, Functional Analysis and Optimization, E. R. Caianiello, ed., Academic Press, New York, 1966.

[7] R. CONTI, *On linear controllability*, Ibid.

[8] L. MARKUS, *Controllability and observability*, Ibid.

[9] I. TARNOVE, *A controllability problem for nonlinear systems*, Mathematical Theory of Control, A. V. Balakrishnan and L. Neustadt, eds., Academic Press, New York, 1967.

[10] D. O. NORRIS, *Lagrangian saddle-points and optimal control*, this Journal, 5 (1967), pp. 594–599.

[11] V. F. DEM'YANOV AND A. M. RUBINOV, *Minimization of Functionals in Normed Spaces*, this Journal, 6 (1968), pp. 73–88.

[12] W. A. PORTER, *On function space pursuit-evasion games*, this Journal, 5 (1967), pp. 555–574.

[13] ———, *A minimization problem and its application to optimal control and system sensitivity*, this Journal, 6 (1968), pp. 303–311.

[14] H. S. WITSENHAUSEN, *A minimax control problem for sampled linear systems*, IEEE Trans. Automatic Control, AC-13 (1968), pp. 5–21.

[15] N. BOURBAKI, *Elements of Mathematics—General Topology I*, Addison-Wesley, Reading, Massachusetts, 1966.

[16] J. L. KELLEY AND I. NAMIOKA, *Linear Topological Spaces*, Van Nostrand, New York, 1963.

[17] L. HÖRMANDER, *Sur la fonction d'appui des ensembles convexes dans un espace localement convexe*, Ark. Mat., 3 (1955), nr. 12, pp. 181–186. (In French.)

[18] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators. Part I: General Theory*, Interscience, New York, 1967.

[19] N. BOURBAKI, *Espaces Vectoriels Topologiques*, Hermann, Paris, 1966.

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

SIAM J. Control
Vol. 7, No. 4, November 1969

LAGRANGE MULTIPLIERS AND NONCONVEX PROGRAMS*

JAMES E. FALK†

1. Introduction. Lagrange multipliers have been used by several authors to solve convex programming problems [5], [7] and this procedure has been especially worthwhile when the problem has a "decomposible" structure. As opposed to most other methods, the Lagrangian technique preserves problem structure and yields a sequence of fairly tractable subproblems in many instances. On the other hand, the Lagrangian technique fails for a number of simple nonconvex problems which can be solved using other methods.

In order to investigate the reasons for the failure of the Lagrangian technique on some nonconvex problems, it is necessary to study the "convex envelope" of a function. This notion is defined in § 2. Also in this section we generalize the notion of the "conjugate" of a function to apply to nonconvex functions and show that the conjugate of the conjugate of a function yields the convex envelope of that function under rather mild hypothesis. This latter characterization is useful since it allows one to compute the convex envelope of separable functions over certain rectangular domains.

In § 3 we investigate the use of the Lagrangian technique in minimizing nonconvex functions over constraint sets of the form $G \cap C$ where $G = \{x | Ax \geqq b\}$ and where C is an arbitrary compact subset of E^n. Under proper assumptions on ϕ, it is shown that if Lagrange multipliers are associated with the constraints $Ax \geqq b$, the solution of the associated "dual" problem yields the minimum over $G \cap C$ of the convex envelope of ϕ taken over C. This minimum may not coincide with the solution of the desired problem but an algorithm has been developed elsewhere [2] which could use this information to advantage in finding the desired solution.

The algorithm developed in [2] requires the minimum of the convex envelope of a function over a certain set. In § 4 we point out that this minimum may be obtained without explicitly calculating the convex envelope. This may be quite useful since the calculation of the convex envelope is often computationally infeasible.

2. Convex envelopes and conjugate functions. In this section we define the "convex envelope" (function) of a given function ϕ over a (nonempty) set C and derive some of the properties of such functions. In addition, the notion of the conjugate of a function as defined by Fenchel (see [3] or [4]) is generalized to apply to a wide class of nonconvex functions. Heretofore conjugacy has been applied only to convex functions. It is then shown that the conjugate of the conjugate of a general function yields the convex envelope of that function and hence a second (and more useful) characterization of the convex envelope of a function is obtained.

The definition of the convex envelope is based on the notion of the "convex hull" of a set. Given a set S in E^n, the intersection of all convex sets which contain S is called the *convex hull* of S and denoted by S^c. Thus S^c is the smallest convex

* Received by the editors October 29, 1968.

† Research Analysis Corporation, McLean, Virginia.

set which contains S. The set S^c may also be characterized as

$$S^c = \{x : x = \sum_{i=1}^{n+1} \alpha_i x^i, \sum_{i=1}^{n+1} \alpha_i = 1, \alpha_i \geqq 0, x^i \in S\},$$

i.e., S^c is the set of all convex combinations of all $(n + 1)$-tuples of points taken from S (see [4] for a proof of this assertion).

Suppose now, that ϕ is any function defined over some set C in E^n. The set

$$[\phi, C] = \{(\xi, x) : \xi \geqq \phi(x), x \in C\}$$

consists of all points in E^{n+1} which lie on or above the graph of ϕ. Thus $[\phi, C]^c$ is the smallest convex set in E^{n+1} which contains $[\phi, C]$. The "lower" portion of this set forms the graph of ϕ^c, the *convex envelope* of ϕ:

$$\phi^c(x) = \inf\{\xi : (\xi, x) \in [\phi, C]^c\} \tag{2.1}$$

with domain

$$D[\phi^c] = \{x : (\xi, x) \in [\phi, C]^c \text{ for some } \xi\}. \tag{2.2}$$

Figure 2.1 illustrates this definition for a function of a single variable.

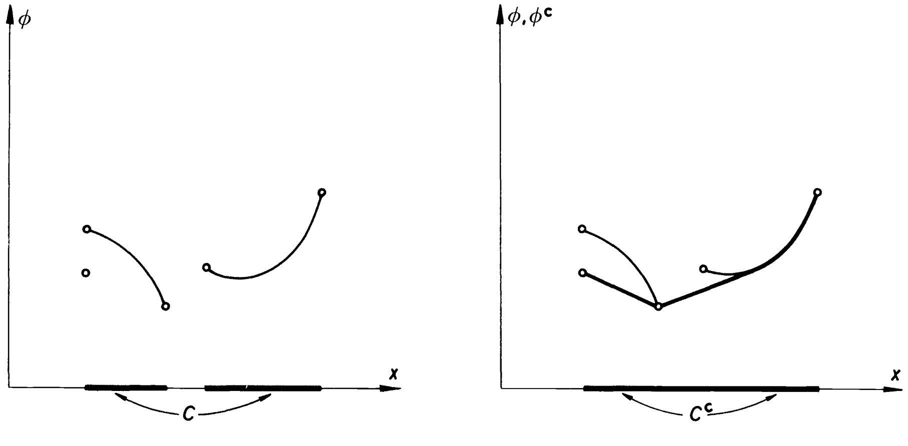


FIG. 2.1. *Convex envelope*

While this definition of the convex envelope is geometrically appealing, in order to derive some of the properties of ϕ^c it is convenient to give an alternative characterization of this function. This will be done in terms of "conjugate" functions as defined by Fenchel ([3], [4]). The notion of the conjugate of ϕ taken over C has heretofore been applied in the case where ϕ was a convex function and C a convex set. In the material which follows, however, we require only that *C be compact and ϕ be lower semicontinuous over C.*

The conjugate of ϕ over C is denoted by ϕ^* and its domain is denoted by $D[\phi^*]$. These quantities are defined by the relations

$$D[\phi^*] = \{t: \sup_{x \in C} (\langle x, t\rangle - \phi(x)) < \infty\}, \tag{2.3}$$

$$\phi^*(t) = \sup_{x \in C} (\langle x, t\rangle - \phi(x)). \tag{2.4}$$

Note that the sup operator could be replaced by the max operator in definitions (2.3) and (2.4) since ϕ is lower semicontinuous over a compact set. We do not do this because these definitions are applied to more general functions in the next paragraph.

It is easily shown that ϕ^* is a convex function and $D[\phi^*]$ is a convex set. In fact, the above assumptions on ϕ and C guarantee that $D[\phi^*] = E^n$. The same operation can be performed on the pair ϕ^* and $D[\phi^*]$ to yield a new (convex) function ϕ^{**} with (convex) domain $D[\phi^{**}]$. It has been shown [4] that if ϕ and C are convex and ϕ is continuous along the boundary of C, then the pair ϕ and C is the same as the pair ϕ^{**} and $D[\phi^{**}]$. In the absence of convexity, this result clearly is not true. It is true, however, that ϕ^{**} and $D[\phi^{**}]$ give the convex envelope of ϕ over C which the following two theorems verify.

THEOREM 2.1. *If C is compact and ϕ is lower semicontinuous over C, then*

$$C^c = D[\phi^c] = D[\phi^{**}].$$

Proof. If $x^0 \in C$, then $x^0 \in D[\phi^c]$ so that $C^c \subset D[\phi^c]$. Now let $x^0 \in D[\phi^c]$. Then there is a pair $(\xi^0, x^0) \in [\phi, C]^c$. But this implies that there are $n + 2$ points $(\xi^j, x^j) \in [\phi, C]$ such that

$$(\xi^0, x^0) = \sum_{j=1}^{n+2} \alpha_j(\xi^j, x^j)$$

for some set of α_j's where $\sum_{j=1}^{n+2} \alpha_j = 1$ and $\alpha_j \geqq 0$. Hence

$$x^0 = \sum_{j=1}^{n+2} \alpha_j x^j$$

with $x^j \in C$ so that $x^0 \in C^c$ (i.e., $D[\phi^c] \subset C^c$) and the first equality of the theorem is established.

Now let $x^0 \in C$. Then

$$\langle x^0, t\rangle - \phi(x^0) \leqq \phi^*(t)$$

for all $t \in E^n$. Hence

$$\langle x^0, t\rangle - \phi^*(t) \leqq \phi(x^0),$$

so that

$$\sup_{t \in E^n} \{\langle x^0, t\rangle - \phi^*(t)\} \leqq \phi(x^0) < \infty,$$

i.e.,

$$x^0 \in D[\phi^{**}],$$

which implies that C and hence C^c are subsets of the convex set $D[\phi^{**}]$.

To prove that $D[\phi^{**}] \subset C^c$, we assume that this is not true. Then there is a point $x^0 \in D[\phi^{**}]$ such that $x^0 \notin C^c$. The set C^c is compact since C was assumed to be compact. Thus there is a vector s such that

$$\langle x, s\rangle < \langle x^0, s\rangle$$

for all $x \in C^c$ (see [4, p. 397]). Let x^* maximize $\langle x, s\rangle$ over C^c. Then

$$\langle x^*, ks\rangle < \langle x^0, ks\rangle, \qquad k = 1, 2, 3, \cdots,$$

from above, i.e.,

$$\langle x^0 - x^*, ks\rangle > 0.$$

Hence

$$\langle x^0 - x^*, ks\rangle \to \infty$$

as $k \to \infty$. Now

$$\langle x^0, ks\rangle - \phi^*(ks) \leqq \phi^{**}(x^0)$$

from the definition of ϕ^{**} and

$$\phi^*(ks) = \max_{x \in C}\{\langle x, ks\rangle - \phi(x)\} \leqq \langle x^*, ks\rangle + \max_{x \in C}\{-\phi(x)\}.$$

(These maxima exist since ϕ is lower semicontinuous and C is compact.) Combining these last two inequalities,

$$\langle x^0 - x^*, ks\rangle - \max_{x \in C}\{-\phi(x)\} \leqq \phi^{**}(x^0).$$

But the left-hand side of this relation tends to $+\infty$ as $k \to +\infty$ so that $\phi^{**}(x^0)$ is not finite, i.e., $x^0 \notin D[\phi^{**}]$, contrary to assumption.

Thus the domains of the functions ϕ^c and ϕ^{**} are identical and, in fact, equal to C^c. The next theorem establishes the identity $\phi^c(x) = \phi^{**}(x)$ for any $x \in C^c$. It is convenient to prove the next lemma before presenting the theorem.

LEMMA. *If $x^0 \in C$, then $\phi^{**}(x^0) \leqq \phi(x^0)$.*

Proof. By definition

$$\langle x^0, t\rangle - \phi(x^0) \leqq \phi^*(t)$$

for any $t \in D[\phi^*]$. Thus

$$\langle x^0, t\rangle - \phi^*(t) \leqq \phi(x^0)$$

so that

$$\phi^{**}(x^0) = \sup_{t \in D[\phi^*]}\{\langle x^0, t\rangle - \phi^*(t)\} \leqq \phi(x^0)$$

and the lemma is proved.

THEOREM 2.2. *If C is compact and ϕ is lower semicontinuous over C, then*

$$\phi^c(x) = \phi^{**}(x)$$

for all $x \in C^c$.

Proof. Since ϕ^{**} is convex over C^c, and since $\phi^{**}(x) \leqq \phi(x)$ over C, it follows from the definition of ϕ^c that

$$\phi^{**}(x) \leqq \phi^c(x)$$

for all $x \in C^c$.

Now we assume that

$$\phi^{**}(x^0) < \phi^c(x^0)$$

for some $x^0 \in C^c$. It follows that the point $(\phi^{**}(x^0), x^0)$ is not a member of the closed convex set

$$[\phi^c, C^c] = \{(\xi, x) : \xi \geqq \phi^c(x), x \in C^c\}.$$

Thus ther˒ is a vector $(\sigma^0, s^0) \in E^{n+1}$ which strictly separates the point $(\phi^{**}(x^0), x^0)$ from $[\phi^c, C^c]$. If $\sigma^0 = 0$, then

$$\langle x, s^0 \rangle < \langle x^0, s^0 \rangle$$

for all $x \in C^c$. Hence $x^0 \notin C^c$ which is a contradiction.

Since $\sigma^0 \neq 0$, we may assume that $\sigma^0 = -1$. Then either

$$-\phi^c(x) + \langle x, s^0 \rangle > -\phi^{**}(x^0) + \langle x^0, s^0 \rangle \tag{2.5}$$

or

$$-\phi^c(x) + \langle x, s^0 \rangle < -\phi^{**}(x^0) + \langle x^0, s^0 \rangle \tag{2.6}$$

for all $x \in C^c$. If (2.5) occurs, it must be true for $x = x^0$ so that

$$-\phi^c(x^0) > -\phi^{**}(x^0),$$

which contradicts the previous assumption that $\phi^{**}(x^0) < \phi^c(x^0)$. If (2.6) holds, it must be true for $x \in C$ where $\phi^c(x) \leqq \phi(x)$ so that

$$-\phi(x) + \langle x, s^0 \rangle < -\phi^{**}(x^0) + \langle x^0, s^0 \rangle$$

for all $x \in C$. However,

$$\phi^{**}(x^0) \geqq -\phi^*(s^0) + \langle x^0, s^0 \rangle$$

so that

$$-\phi(x) + \langle x, s^0 \rangle < \phi^*(s^0)$$

for all $x \in C$. But C is compact so that the supremum over C of the left-hand side of this expression is attained at some point and equals $\phi^*(s^0)$ by definition. This implies the contradiction $\phi^*(s^0) < \phi^*(s^0)$ and the proof is complete.

The next corollary follows immediately from the preceding theorem.

COROLLARY. *If C is compact and ϕ is lower semicontinuous over C, then the conjugate of ϕ over C equals the conjugate of ϕ^c over C^c (i.e., $\phi^* = \phi^{c*}$).*

Proof. ϕ^* is a convex function defined over $D[\phi^*] = E^n$ so that

$$\phi^{***} = \phi^*$$

(see Fenchel [3] or Karlin [4]). By the preceding theorem

$$\phi^{**} = \phi^c$$

so that

$$\phi^{***} = (\phi^c)^*,$$

and the proof is complete.

The second characterization of the convex envelope represented by Theorems 2.1 and 2.2 is most convenient from a computational point of view since it furnishes a method for actually constructing ϕ^c in a number of cases. For example, when ϕ is separable and C is "rectangular," the computational effort required to yield ϕ^c is greatly reduced as is indicated in the next theorem.

THEOREM 2.3. *If*

$$\phi(x) = \sum_{i=1}^{n} \phi_i(x_i)$$

and

$$C = \{x : l \leqq x \leqq L\},$$

where each ϕ_i is lower semicontinuous and l, $L \in E^n$, then

$$\phi^c(x) = \sum_{i=1}^{n} \phi_i^c(x_i),$$

where ϕ_i^c is the convex envelope of ϕ_i taken over $[l_i, L_i]$.

Proof.

$$\begin{aligned}
\phi^*(t) &= \max_{x \in C} \{\langle x, t\rangle - \phi(x)\} \\
&= \sum_{i=1}^{n} \max_{l_i \leqq x_i \leqq L_i} \{x_i t_i - \phi_i(x_i)\} \\
&= \sum_{i=1}^{n} \phi_i^*(t_i); \\
\phi^c(x) = \phi^{**}(x) &= \sup_{t \in E^n} \{\langle x, t\rangle - \phi^*(t)\} \\
&= \sum_{i=1}^{n} \sup_{-\infty < t_i < \infty} \{x_i t_i - \phi_i^*(t_i)\} \\
&= \sum_{i=1}^{n} \phi_i^{**}(x_i) \\
&= \sum_{i=1}^{n} \phi_i^c(x_i);
\end{aligned}$$

and the proof is complete.

Thus, to calculate the convex envelope of a separable function over a rectangular set, it is sufficient to calculate the convex envelopes of the constituents of the function over their respective domains. In particular, if ϕ_i is concave, its convex envelope is simply that linear function whose graph joins the endpoints of the graph of ϕ_i.

The separability of ϕ is crucial in the above proof since it is not true in general that the convex envelope of the sum of two functions is the sum of the convex envelopes. For example, the convex envelope of the function $\phi(x) = x^2 - x^2$ over the interval [0, 1] is the zero function while the sum of the convex envelopes is $x^2 - x$.

The results of this section will now be used in the investigation of the applicability of the Lagrangian technique in nonconvex programming. The characterization of the convex envelope of a function is especially useful in this regard.

3. Lagrange multipliers and nonconvex programming. Lagrange multipliers have proven useful in solving convex programming problems (see [5] and [7]), especially when these problems possess a "decomposible structure." Essentially, the use of multipliers results in the formulation of a "dual" problem which is then solved in lieu of the given primal problem. This duality theory and its connection with the notion of conjugate functions has been examined extensively by Rockafellar [6]. The structure of the dual problem for a particular class of convex programs was investigated by Falk [1].

While Lagrangian methods are well suited for the solution of structured convex programs, the methods may fail for nonconvex problems where other methods (e.g., penalty function techniques) succeed. The reason for this failure will be investigated in this section for a particular class of nonconvex problems. The results contained in this section suggest an application of Lagrange multipliers to a class of nonconvex problems.

The problem we wish to investigate has the form

$$\begin{aligned} &\text{minimize} && \phi(x) \\ &\text{subject to} && Ax \geqq b, \quad x \in C, \end{aligned} \tag{3.1}$$

where ϕ is lower semicontinuous (possibly nonconvex), A is an $m \times n$ matrix, b is a constant vector and C is a compact set.

The *auxiliary function* of problem (3.1) is defined by:

$$\gamma(u) = \min_{x \in C} \{\phi(x) - \langle u, Ax - b\rangle\}. \tag{3.2}$$

Since C is compact and ϕ lower semicontinuous, the function γ is defined over E^m although we will only be interested in its behavior over $(E^m)^+ = \{u \in E^m : u \geqq 0\}$. It is easily shown that γ is a concave function of u.

Associated with each u, we define the set

$$X(u) = \{x \in C : x \text{ minimizes } \phi(x) - \langle u, Ax - b\rangle\}. \tag{3.3}$$

The *auxiliary problem* of (3.1) is defined as:

$$\begin{aligned} &\text{maximize} && \gamma(u) \\ &\text{subject to} && u \geqq 0. \end{aligned} \tag{3.4}$$

When the constituents of problem (3.1) are convex, there is a close connection between problems (3.1) and (3.4) as is illustrated in the following two "duality" theorems. The proof of Theorem 3.1 may be found in [1, p. 154] while the proof of Theorem 3.2 is new, although a version of Theorem 3.2 is proved in [1] under different assumptions.

THEOREM 3.1. *Suppose ϕ is convex and C is a closed convex set. If problem* (3.1) *has a solution x^*, then there is a point u^* which is a solution of the auxiliary problem. Moreover, $x^* \in X(u^*)$ and $\phi(x^*) = \gamma(u^*)$.*

THEOREM 3.2. *Suppose ϕ is convex and C is a compact convex set. If u^* is a solution of the auxiliary problem of* (3.1), *then problem* (3.1) *is feasible and hence has at least one solution x^*. Moreover, $x^* \in X(u^*)$ and $\phi(x^*) = \gamma(u^*)$.*

Proof. Assume (3.1) is not feasible. Then for every $x \in C$ we have

$$g(x) = Ax - b \ngeqq 0.$$

The set $F = \{v : v = g(x),\ x \in C\}$ is compact and convex and does not intersect $(E^m)^+$. Thus there is a vector $t \in E^m$ such that

$$\langle t, g(x)\rangle < 0$$

for all $x \in C$ and

$$\langle t, u\rangle \geqq 0$$

for all $u \in (E^m)^+$. Let

$$\mu(t) = \max_{x \in C} \langle t, g(x)\rangle$$

and

$$\mu = \min_{x \in C} \phi(x).$$

Note that the sequence $\{\mu(kt)\}$ tends to $-\infty$ as $k \to \infty$.

Since $\langle t, u\rangle \geqq 0$ for all $u \in (E^m)^+$, it follows that $t \geqq 0$. Then

$$\begin{aligned}\gamma(kt) &= \min_{x \in C} \{\phi(x) - \langle kt, g(x)\rangle\} \\ &\geqq \min_{u \in C} \phi(x) - \max_{x \in C} \langle kt, g(x)\rangle \\ &= \mu - \mu(kt)\end{aligned}$$

so that $\gamma(kt) \to \infty$ as $k \to \infty$. Hence γ cannot attain its maximum over $(E^m)^+$ which is a contradiction.

It follows that problem (3.1) must have a solution. Let X^+ denote the set of all solutions of problem (3.1) and let $x^+ \in X^+$. By Theorem 3.1, there is a vector $u^+ \in (E^m)^+$ such that $x^+ \in X(u^+)$ and $\gamma(u^+) = \phi(x^+)$. If $x^+ \notin X(u^*)$, we must have

$$\phi(x^+) - \langle u^*, g(x^+)\rangle > \phi(x^*) - \langle u^*, g(x^*)\rangle,$$

where x^* is any vector in $X(u^*)$. But the right-hand side of this expression equals $\gamma(u^*)$. Since both u^* and u^+ maximize γ, we have

$$\gamma(u^*) = \gamma(u^+) = \phi(x^+).$$

Thus

$$\phi(x^+) - \langle u^*, g(x^+)\rangle > \phi(x^+),$$

i.e.,

$$\langle u^*, g(x^+)\rangle < 0,$$

which is impossible since $u^* \geqq 0$ and $g(x^+) \geqq 0$. Hence $x^+ \in X(u^*)$ and the proof is complete.

We now wish to investigate the case where ϕ is lower semicontinuous but not necessarily convex and in this connection we compare the auxiliary function of problem (3.1) with the auxiliary function of the problem

$$\begin{aligned} &\text{minimize} \quad \phi^c(x) \\ &\text{subject to} \quad Ax \geqq b, \quad x \in C, \end{aligned} \tag{3.5}$$

where ϕ^c is the convex envelope of ϕ *taken over* C.

Let

$$\gamma^c(u) = \min_{x \in C^c} \{\phi^c(x) - \langle u, Ax - b\rangle\}$$

i.e., γ^c is the auxiliary function of problem (3.5). We now show that the auxiliary functions of problems (3.1) and (3.5) are identical.

THEOREM 3.3. *If ϕ is lower semicontinuous and C is compact in problem* (3.1), *then*

$$\gamma(u) = \gamma^c(u)$$

for each $u \in (E^m)^+$.

Proof.

$$\begin{aligned} \gamma(u) &= \min_{x \in C} \{\phi(x) - \langle u, Ax - b\rangle\} \\ &= \min_{x \in C} \{\phi(x) - \langle A^T u, x\rangle\} + \langle u, b\rangle \\ &= -\phi^*(A^T u) + \langle u, b\rangle. \end{aligned}$$

However, by the corollary of Theorem 2.2, we have

$$\begin{aligned} \phi^*(A^T u) &= \phi^{c*}(A^T u) \\ &= -\min_{x \in C^c} \{\phi^c(x) - \langle A^T u, x\rangle\} \\ &\quad -\min_{x \in C^c} \{\phi^c(x) - \langle u, Ax\rangle\} \\ &= -(\gamma^c(u) - \langle u, b\rangle) \end{aligned}$$

so that $\gamma(u) = \gamma^c(u)$ and the proof is complete.

Before discussing some implications of this theorem, we will first prove a related result concerning the minimizing sets $X(u)$ of problems (3.1) and (3.5). Let

$$X^c(u) = \{x : x \text{ minimizes } \phi^c(x) - \langle u, Ax - b\rangle \text{ over } C^c\}.$$

The symbol $(X(u))^c$ denotes the convex hull of the set $X(u)$.

THEOREM 3.4. *If ϕ is lower semicontinuous and C is compact in problem* (3.1), *then*

$$X^c(u) = (X(u))^c.$$

Proof. Let $x^0 \in X(u)$. Then, since $\phi^c(x^0) \leqq \phi(x^0)$, we have

$$\phi^c(x^0) - \langle u, Ax^0 - b\rangle \leqq \phi(x^0) - \langle u, Ax^0 - b\rangle = \gamma(u) = \gamma^c(u)$$

so that $x^0 \in X^c(u)$. The convexity of $X^c(u)$ implies that

$$(X(u))^c \subset X^c(u).$$

Now assume that there is a point $x^0 \in X^c(u)$ such that $x^0 \notin (X(u))^c$. Since $x^0 \in X^c(u) \subset C^c$, we can write

$$(\phi^c(x^0), x^0) = \sum_{i=1}^{n+2} \alpha_i(\phi(x^i), x^i),$$

where $x^i \in C$ and $\alpha_i \geqq 0$, $\sum_{i=1}^{n+2} \alpha_i = 1$ from the definition of the convex envelope of ϕ. The representation of $(\phi^c(x^0), x^0)$ may not be unique. However, since $x^0 \notin (X(u))^c$, it is impossible to find a representation of $(\phi^c(x^0), x^0)$ where $x^i \in X(u)$, $i = 1, \cdots, n + 2$. From above, we have

$$\phi^c(x^0) = \sum_{i=1}^{n+2} \alpha_i \phi(x^i).$$

It follows that

$$\phi^c(x^0) - \langle u, Ax^0 - b\rangle = \sum_{i=1}^{n+2} \alpha_i(\phi(x^i) - \langle u, Ax^i - b\rangle) > \sum_{i=1}^{n+2} \alpha_i \gamma(u)$$

since not all x^i can lie in $X(u)$. But $\sum_{i=1}^{n+2} \alpha_i \gamma(u) = \gamma(u) = \gamma^c(u)$ which yields the contradiction

$$\gamma^c(u) = \phi^c(x^0) - \langle u, Ax^0 - b\rangle > \gamma^c(u),$$

and the proof is complete.

The following example illustrates that Theorems (3.3) and (3.4) cannot be extended to the case where the linear constraints $Ax \geqq b$ are replaced by nonlinear constraints $g(x) \geqq 0$, even if the components of g are assumed to be concave.

Example.

$$\begin{aligned} &\text{minimize} && \phi(x) = 1 - x^2 \\ &\text{subject to} && g(x) = 1 - x^2 \geqq 0, \\ & && C = \{x : -1 \leqq x \leqq 1\} \end{aligned}$$

If $u = 1$, we have

$$\gamma(1) = \min_{-1 \leqq x \leqq 1} \{(1 - x^2) - (1 - x^2)\} = 0$$

and

$$X(1) = C.$$

On the other hand

$$\gamma^c(1) = \min_{-1 \leqq x \leqq 1} \{0 - (1 - x^2)\} = -1$$

and

$$X^c(1) = \{0\}.$$

Since the last two theorems show that the auxiliary problems of programs (3.1) and (3.5) are identical, and since the duality theorems (3.1) and (3.2) apply to the convex program (3.5), it is clear that this latter problem rather than the desired problem is solved using Lagrange multipliers. In applications, this event is recognized when

$$\phi(x) > \gamma(u^*)$$

for all $x \in X(u^*)$, where u^* maximizes γ over $(E^m)^+$. We summarize these comments in the next theorem.

THEOREM 3.5. *Assume that ϕ is lower semicontinuous and C is compact in problem* (3.1). *Let u^* maximize the auxiliary function of problem* (3.1). *Then any point $x \in X(u^*)$ minimizes the convex envelope of ϕ taken over C subject to the restrictions of problem* (3.1). *A point $x^* \in X(u^*)$ is a solution of* (3.1) *if and only if* $\phi(x^*) = \gamma(u^*)$.

A related result appears in [1] where it is shown that if γ is differentiable at the point u^* of Theorem (3.5), then any point $x \in X(u^*)$ is a solution of problem (3.1). In view of the preceding theorems, in such a case the convex envelope of ϕ is minimized at the same points as is ϕ.

4. An application. In the preceding section it was shown that the Lagrange multiplier technique may yield false solutions to problems of the form (3.1) since the convex envelope of ϕ taken over C is actually minimized by this method. This fact may be exploited in the method detailed in [2] where branch and bound is used to subdivide the set C into successively smaller subsets. The minimum of the convex envelope of ϕ taken over each of those subsets subject to the constraints $Ax \geqq b$ is required. Thus at each stage of this algorithm one must solve problems of the form

$$\begin{aligned} &\text{minimize} && \psi^j(x) \\ &\text{subject to} && Ax \geqq b, \quad x \in C^j, \end{aligned} \tag{4.1}$$

where ψ^j is the convex envelope of ϕ taken over C^j. According to the theory of the preceding section, it is sufficient to use the Lagrange multiplier technique on the problem

$$\begin{aligned} &\text{minimize} && \phi(x) \\ &\text{subject to} && Ax \geqq b, \quad x \in C^j \end{aligned} \tag{4.2}$$

directly and thus eliminate the need to calculate ψ^j explicitly. If ϕ is separable, and $C^j = \{x : l^j \leqq x \leqq L^j\}$, the calculation of $\gamma(u)(= \gamma^c(u))$ involves finding the minimum of n functions each of a single variable over their intervals of definition:

$$\begin{aligned} \gamma(u) &= \min_{l^j \leqq x \leqq L^j} \left\{ \sum_{i=1}^{n} \phi_i(x_i) - \langle u, Ax\rangle \right\} + \langle u, b\rangle \\ &= \sum_{i=1}^{n} \min_{l_i^j \leqq x_i \leqq L_i^j} \{\phi_i(x_i) - \langle a^i, u\rangle x_i\} + \langle u, b\rangle, \end{aligned}$$

where a^i is the ith column of A. Of course, these minimizations may have to be performed for a number of u values before the maximum of γ is attained.

REFERENCES

[1] J. E. FALK, *Lagrange multipliers and nonlinear programming*, J. Math. Anal. Appl., 19 (1967), pp. 141–159.

[2] J. E. FALK AND R. M. SOLAND, *An algorithm for separable nonconvex programming*, Internal Rep., Research Analysis Corporation, McLean, Virginia, 1968.

[3] W. FENCHEL, *Convex Cones, Sets and Functions*, Princeton University Press, Princeton, New Jersey, 1953.

[4] S. KARLIN, *Mathematical Methods and Theory in Games, Programming and Economics*, Addison-Wesley, Reading, Massachusetts, 1959.

[5] L. LASDON, *Duality and decomposition in mathematical programming*, Internal Rep., Systems Research Center, Case Institute of Technology, Cleveland, Ohio, 1967.

[6] R. F. ROCKAFELLAR, *Convex functions and dual extremum problems*, Doctoral thesis, Harvard, Cambridge, 1963.

[7] I. TAKAHASHI, *Variable separation principle for mathematical programming*, J. Operations Res. Soc. Japan, 6 (1964), pp. 82–105.

SIAM J. Control
Vol. 7, No. 4, November 1969

OPTIMAL CONTROL OF SYSTEMS DESCRIBED BY PARABOLIC EQUATIONS*

L. I. GAL'CHUK†

In this paper the possibility is studied of translating a system governed by a parabolic equation into a stationary regime. This problem is equivalent to a certain problem of moments. Conditions for the attainability of stationary states will be given below, that is, conditions for the solvability of the problem of moments.

1. Statement of the problem. Let us consider the following problem:

$$\frac{\partial f(t,x)}{\partial t} = Lf(t,x) + u(t)\psi(x), \tag{1}$$

$$f(0,x) = 0, \qquad f(t,x)|_{\Gamma} = 0, \tag{2}$$

where $0 \leqq t \leqq T$, $x \in \Omega$, Γ is the boundary of Ω, Ω is a region in Euclidean space E^n, L is a linear elliptic operator with a system of eigenfunctions complete in $L_2(\Omega)$, $u(t)$ is a control, a measurable function with $|u(t)| \leqq 1$, and $\psi(x)$ is a given function. A control $u(t)$ is to be found which transfers system (1), (2) in minimum time into a stationary regime, i.e., such that

$$Lf(T,x) = v\psi(x), \tag{3}$$

where v is a constant with $|v| \leqq 1$, and T is the minimum time.

We shall show that problem (1)–(3) reduces to a problem of moments. The eigenvalues of the operator L will be denoted by λ_k, and we shall assume that $0 < \lambda_1 \leqq \lambda_2 \leqq \cdots \leqq \lambda_k \leqq \cdots$. The corresponding eigenfunctions of L will be denoted by $\varphi_k(x)$, i.e.,

$$L\varphi_k(x) = -\lambda_k\varphi_k(x), \qquad \varphi_k(x)|_{\Gamma} = 0.$$

Let

$$\psi(x) = \sum_{k=1}^{\infty} c_k\varphi_k(x).$$

The solution to problem (1)–(2) can be written in the form

$$f(t,x) = \sum_{k=1}^{\infty} c_k\varphi_k(x) \int_0^t u(s)e^{-\lambda_k(t-s)}\,ds.$$

To satisfy (3), we must have

$$Lf(T,x) = -\sum_{k=1}^{\infty} \lambda_k c_k\varphi_k(x) \int_0^T u(s)e^{-\lambda_k(T-s)}\,ds = v\psi(x) = v\sum_{k=1}^{\infty} c_k\varphi_k(x)$$

* Originally published in *Vestnik Moskovskogo Universiteta, Matematika, Mekhanika*, 1968, no. 3, pp. 21–33. Submitted on May 23, 1967. This translation into English has been prepared by R. N. and N. B. McDonough.

Translated and printed for this Journal under a grant-in-aid by the National Science Foundation.

† Department of Probability Theory and Mathematical Statistics, Moscow University, Moscow.

or

$$-\lambda_k \int_0^T u(s)e^{-\lambda_k(T-s)}\,ds = v, \qquad k = 1,2,\cdots.$$

Thus, solving problem (1)–(3) is equivalent to solving the problem of moments:

$$\int_0^T u(s)e^{-\lambda_k(T-s)}\,ds = -\frac{v}{\lambda_k}, \qquad k = 1,2,\cdots. \tag{4}$$

The problem of precise end-point attainment considered by Egorov [1] reduces to problem (4). Let us consider this.

In the region $Q = \{0 \leqq t \leqq T, 0 \leqq x \leqq 1\}$, the function $f(t,x)$ satisfies

$$\frac{\partial f(t,x)}{\partial t} = \frac{\partial^2 f(t,x)}{\partial x^2}, \tag{5'}$$

and on the boundary of Q,

$$f(0,x) = 0, \quad \frac{\partial f(t,0)}{\partial x} = 0, \quad \frac{\partial f(t,1)}{\partial x} = \alpha[u(t) - f(t,1)], \quad \alpha = \text{const.} > 0, \tag{6'}$$

where the control $u(t)$ is a measurable function with $|u(t)| \leqq 1$. A control $u(t)$ is to be found such that $f(T,x) = v$, where v is a constant and $|v| \leqq 1$. Let $\{\mu_k\}$ be the sequence of positive roots of the equation $\mu \tan \mu = \alpha$. Let us multiply (5') by $\exp(\mu_k^2 t)\cos \mu_k x$ and integrate over Q. Taking into account the boundary and initial conditions, we obtain

$$v\int_0^1 \cos \mu_k x\,dx = \alpha \cos \mu_k \int_0^T u(t)e^{-\mu_k^2(T-t)}\,dt, \qquad k = 1,2,\cdots.$$

Since $\mu_k \tan \mu_k = \alpha$, this can be written as

$$\frac{v}{\mu_k^2} = \int_0^T u(t)e^{-\mu_k^2(T-t)}\,dt, \qquad k = 1,2,\cdots.$$

Letting $\mu_k^2 = \lambda_k$, we then have

$$\int_0^T u(t)e^{-\lambda_k(T-t)}\,dt = \frac{v}{\lambda_k}, \qquad k = 1,2,\cdots.$$

Thus the problem of reaching a constant end-point for (5'), (6') reduces to the moment problem (4). In [1] it is proved that it is possible to hit $f(T,x) = 0$ and a small neighborhood of zero. In this paper, we shall prove that it is possible to hit $f(T,x) = v$, with $|v| < 1$.

Remark. In problem (1)–(3), L can be an arbitrary linear operator, provided its eigenfunctions are complete in $L_2(\Omega)$. For example, we can take $L = \partial^2/\partial x^2$ with $\Omega = [0,1]$. Then the eigenfunctions are $\varphi_k(x) = \sin \pi k x$.

2. Solution of the problem of moments. Let us assume in (4) that $0 < \lambda_1 \leqq \lambda_2 \leqq \cdots \leqq \lambda_k \leqq \cdots$, and $\sum_{i=1}^{\infty} 1/\lambda_i < \infty$. We shall examine the system of vectors $a_1, a_2, \cdots, a_k, \cdots$ in l_2, with

$$a_1 = \left\{-\frac{1}{\lambda_1}, -\frac{e^{-(\lambda_2-\lambda_1)T}}{\lambda_2}, -\frac{e^{-(\lambda_3-\lambda_1)T}}{\lambda_3}, -\frac{e^{-(\lambda_4-\lambda_1)T}}{\lambda_4}, \right.$$
$$\left. \cdots, -\frac{e^{-(\lambda_{k-1}-\lambda_1)T}}{\lambda_{k-1}}, -\frac{e^{-(\lambda_k-\lambda_1)T}}{\lambda_k}, -\frac{e^{-(\lambda_{k+1}-\lambda_1)T}}{\lambda_{k+1}}, \cdots\right\},$$
$$(5) \qquad a_2 = \left\{\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \frac{1}{\lambda_3}, \frac{1}{\lambda_4}, \cdots, \frac{1}{\lambda_{k-1}}, \frac{1}{\lambda_k}, \frac{1}{\lambda_{k+1}}, \cdots\right\},$$
$$a_3 = \{a_{31}, a_{32}, -1, 0, \cdots, 0, 0, 0, \cdots\},$$
$$\vdots$$
$$a_k = \{a_{k1}, a_{k2}, a_{k3}, a_{k4}, \cdots, a_{kk-1}, -1, 0, \cdots\},$$
$$\vdots$$

The system (5) is such that, for $k \geqq 3$, the vector a_k is orthogonal to $a_1, a_2, \cdots, a_{k-1}$. Let us estimate the coordinates of a_k.

LEMMA. $|a_{ki}| \leqq c/\lambda_k$ *for* $1 \leqq i \leqq k-1$, $k \geqq 3$.

Proof. Since for $k \geqq 3$, the vector a_k is orthogonal to $a_1, a_2, \cdots, a_{k-1}$, we have the equations

$$(a_k, a_1) = 0, \quad (a_k, a_2) = 0, \cdots, (a_{k-1}, a_k) = 0$$

for its coordinates. This is a system of $k-1$ equations in $k-1$ unknowns. Taking (5) into account, we can rewrite this system in matrix form as

$$A_k\tilde{a}_k = b_k,$$

where

$$A_k = \begin{pmatrix} -\frac{1}{\lambda_1} & -\frac{e^{-(\lambda_2-\lambda_1)T}}{\lambda_2} & -\frac{e^{-(\lambda_3-\lambda_1)T}}{\lambda_3} & \cdots & -\frac{e^{-(\lambda_{k-1}-\lambda_1)T}}{\lambda_{k-1}} \\ \frac{1}{\lambda_1} & \frac{1}{\lambda_2} & \frac{1}{\lambda_3} & \cdots & \frac{1}{\lambda_{k-1}} \\ a_{31} & a_{32} & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{k-1,1} & a_{k-1,2} & a_{k-1,3} & \cdots & -1 \end{pmatrix},$$

$$\tilde{a}_k = \begin{pmatrix} a_{k1} \\ a_{k2} \\ a_{k3} \\ \vdots \\ a_{k,k-1} \end{pmatrix}, \quad b_k = \begin{pmatrix} -\frac{e^{-(\lambda_k-\lambda_1)T}}{\lambda_k} \\ \frac{1}{\lambda_k} \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

All the rows of the matrix A_k, aside from the first two, are orthogonal to one another. Thus if we orthogonalize these two rows, and then divide each row of the matrix by the length of the row vector formed from its terms, the resulting matrix will be orthogonal. We shall denote by c_k the scalar product of the first two rows of A_k:

$$c_k = (a_1^k, a_2^k) = -\frac{1}{\lambda_1^2} - \sum_{i=2}^{k-1} \frac{e^{-(\lambda_i - \lambda_1)T}}{\lambda_i^2},$$

where a_1^k, a_2^k are the first two rows of A_k. We shall denote by d_k the length of the vector

$$a_1^k - \frac{c_k}{\|a_2^k\|^2} a_2^k,$$

i.e., the length of the vector obtained from the first row after orthogonalization. The system of equations obtained in this way is:

$$\tilde{A}_k \tilde{a}_k = \tilde{b}_k, \tag{6}$$

$$\tilde{A}_k = \begin{pmatrix} -\dfrac{1 + c_k/\|a_2^k\|^2}{\lambda_1 d_k} & -\dfrac{e^{-(\lambda_2 - \lambda_1)T} + c_k/\|a_2^k\|^2}{\lambda_2 d_k} & \cdots & -\dfrac{e^{-(\lambda_{k-1} - \lambda_1)T} + c_k/\|a_2^k\|^2}{\lambda_{k-1} d_k} \\ \dfrac{1}{\lambda_1 \|a_2^k\|} & \dfrac{1}{\lambda_2 \|a_2^k\|} & \cdots & \dfrac{1}{\lambda_{k-1} \|a_2^k\|} \\ \vdots & \vdots & \vdots & \vdots \\ \dfrac{a_{k-1,1}}{\|a_{k-1}^k\|} & \dfrac{a_{k-1,2}}{\|a_{k-1}^k\|} & \cdots & -\dfrac{1}{\|a_{k-1}^k\|} \end{pmatrix},$$

$$\tilde{b}_k = \begin{pmatrix} -\dfrac{e^{-(\lambda_k - \lambda_1)T} + c_k/\|a_2^k\|^2}{\lambda_k d_k} \\ \dfrac{1}{\lambda_k \|a_2^k\|} \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

From (6) we have that $\tilde{a}_k = \tilde{A}_k^{-1} \tilde{b}_k$, and from the orthogonality of $\tilde{A}_k$ we obtain

$$\|\tilde{a}_k\|^2 = (\tilde{A}_k^{-1} \tilde{b}_k, \tilde{A}_k^{-1} \tilde{b}_k) = (\tilde{A}_k^{-1} \tilde{b}_k, \tilde{A}_k' \tilde{b}_k) = (\tilde{A}_k \tilde{A}_k^{-1} \tilde{b}_k, \tilde{b}_k) = (\tilde{b}_k, \tilde{b}_k) = \|\tilde{b}_k\|^2,$$

i.e.,

$$\|\tilde{a}_k\|^2 = \|\tilde{b}_k\|^2.$$

Substituting the coordinates of the vector $\tilde{b}_k$ into this last, we obtain

$$\|\tilde{a}_k\|^2 = \frac{1}{\lambda_k^2}\left[\frac{(e^{-(\lambda_k-\lambda_1)T} + c_k/\|a_2^k\|^2)^2}{d_k^2} + \frac{1}{\|a_2^k\|^2}\right], \tag{7}$$

where

$$\|a_2^k\|^2 = \sum_{i=1}^{k-1}\frac{1}{\lambda_i^2}, \qquad c_k = -\frac{1}{\lambda_1^2} - \sum_{i=2}^{k-1}\frac{e^{-(\lambda_i-\lambda_1)T}}{\lambda_i^2},$$

$$d_k^2 = \left\|a_1^k - \frac{(a_1^k, a_2^k)}{\|a_2^k\|^2}a_2^k\right\|^2 = \|a_1^k\|^2 - \frac{(a_1^k, a_2^k)^2}{\|a_2^k\|^2}.$$

From these we can deduce the following estimates:

$$\left(e^{-(\lambda_k-\lambda_1)T} + \frac{c_k}{\|a_2^k\|^2}\right)^2 = \left[e^{-(\lambda_k-\lambda_1)T} - \frac{1}{\sum_{i=1}^{k-1}(1/\lambda_i^2)}\left(\frac{1}{\lambda_1^2} + \sum_{i=2}^{k-1}\frac{e^{-(\lambda_i-\lambda_1)T}}{\lambda_i^2}\right)\right]^2$$

$$\leqq c_1\frac{1}{\lambda_1^4[\sum_{i=1}^{k-1}(1/\lambda_i^2)]^2} < c_1,$$

$$d_k^2 = \frac{1}{\lambda_1^2} + \sum_{i=2}^{k-1}\frac{e^{-2(\lambda_i-\lambda_1)T}}{\lambda_i^2} - \frac{1}{\sum_{i=1}^{k-1}(1/\lambda_i^2)}\left[\frac{1}{\lambda_1^2} + \sum_{i=2}^{k-1}\frac{e^{-(\lambda_i-\lambda_1)T}}{\lambda_i^2}\right]^2$$

$$\geqq c_2\left(\frac{1}{\lambda_1^2} - \frac{1}{\lambda_1^4\sum_{i=1}^{k-1}(1/\lambda_i^2)}\right) \geqq c_3,$$

$$\frac{1}{\|a_2^k\|^2} = \frac{1}{\sum_{i=1}^{k-1}(1/\lambda_i^2)} \leqq c_4.$$

Substituting these into (7) we find

$$\|\tilde{a}_k\|^2 \leqq \frac{1}{\lambda_k^2}\left[\frac{c_1}{c_3} + c_4\right] = \frac{c}{\lambda_k^2}.$$

Since

$$\|\tilde{a}_k\|^2 = \sum_{i=1}^{k-1} a_{ki}^2,$$

we have

$$|a_{ki}| \leqq \frac{c}{\lambda_k}, \qquad 1 \leqq i \leqq k-1, \qquad k \geqq 3,$$

and the lemma is proved.

Let us orthonormalize the system (5). To do this, we need to orthogonalize the vectors a_1 and a_2, and then normalize the whole system. We shall denote by $\tilde{a}_1$ the vector obtained from a_1 after orthogonalization:

$$\tilde{a}_1 = a_1 - \left(a_1, \frac{a_2}{\|a_2\|}\right)\frac{a_2\cdot}{\|a_2\|},$$

where

$$\|a_2\|^2 = \sum_{i=1}^{\infty} \frac{1}{\lambda_i^2}, \qquad (a_1, a_2) = -\frac{1}{\lambda_1^2} - \sum_{i=2}^{\infty} \frac{e^{-(\lambda_i - \lambda_1)T}}{\lambda_i^2}.$$

After normalization, we shall obtain in l_2 the complete orthogonal system:

$$\frac{\tilde{a}_1}{\|\tilde{a}_1\|}, \frac{a_2}{\|a_2\|}, \ldots, \frac{a_k}{\|a_k\|}, \ldots . \tag{8}$$

We shall now prove the solvability of the problem of moments (4).

THEOREM 1. *If in* (4), $|v| < 1$ *and* $\sum_{i=1}^{\infty} (1/\lambda_i) < \infty$, *then there exists a time* $T\,(T < \infty)$ *and a measurable function* $u(s)$ ($|u(s)| \leqq 1$) *such that the moment problem* (4) *is solved.*

Proof. In l_2 let us consider a vector

$$k(s) = \{e^{-\lambda_1(T-s)}, e^{-\lambda_2(T-s)}, \ldots, e^{-\lambda_k(T-s)}, \ldots\}, \qquad s < T,$$

and expand it in the basis (8):

$$k(s) = \left(k(s), \frac{\tilde{a}_1}{\|\tilde{a}_1\|}\right)\frac{\tilde{a}_1}{\|\tilde{a}_1\|} + \left(k(s), \frac{a_2}{\|a_2\|}\right)\frac{a_2}{\|a_2\|} + \sum_{i=3}^{\infty}\left(k(s), \frac{a_i}{\|a_i\|}\right)\frac{a_i}{\|a_i\|}.$$

This expansion can be multiplied by an arbitrary measurable scalar function $u(s)$ ($|u(s)| \leqq 1$) and integrated from 0 to $T - \varepsilon$ to obtain

$$\int_0^{T-\varepsilon} k(s)u(s)\,ds = \frac{\tilde{a}_1}{\|\tilde{a}_1\|^2}\int_0^{T-\varepsilon} (k(s), \tilde{a}_1)u(s)\,ds + \frac{a_2}{\|a_2\|^2}\int_0^{T-\varepsilon} (k(s), a_2)u(s)\,ds$$
$$+ \int_0^{T-\varepsilon} u(s) \sum_{i=3}^{\infty} (k(s), a_i)\frac{a_i}{\|a_i\|^2}\,ds.$$

The series on the right can be integrated term by term since it converges in mean square. In fact, the remainder term of the series can be estimated as

$$\left|\int_0^{T-\varepsilon} u(s) \sum_{i=n}^{\infty}\left(k(s), \frac{a_i}{\|a_i\|}\right)\frac{a_i}{\|a_i\|}\,ds\right|$$
$$\leqq \int_0^{T-\varepsilon}\left|\sum_{i=n}^{\infty}\left(k(s), \frac{a_i}{\|a_i\|}\right)\frac{a_i}{\|a_i\|}\right| ds$$
$$\leqq (T-\varepsilon)^{1/2}\left\{\int_0^{T-\varepsilon}\left[\sum_{i=n}^{\infty}\left(k(s), \frac{a_i}{\|a_i\|}\right)\frac{a_i}{\|a_i\|}\right]^2 ds\right\}^{1/2}$$
$$= (T-\varepsilon)^{1/2}\left(\int_0^{T-\varepsilon}\sum_{i=n}^{\infty}\left(k(s), \frac{a_i}{\|a_i\|}\right)^2 ds\right)^{1/2}.$$

Since $T - \varepsilon$ is fixed, for any $\delta > 0$ there will be a number $n_0 = n_0(\delta)$ such that, for $n > n_0$,

$$(T-\varepsilon)^{1/2}\left(\int_0^{T-\varepsilon}\sum_{i=n}^{\infty}\left(k(s), \frac{a_i}{\|a_i\|}\right)^2 ds\right)^{1/2} < \delta.$$

Thus,

$$(9)\quad \begin{aligned}\int_0^{T-\varepsilon} k(s)u(s)\,ds = \frac{\tilde{a}_1}{\|\tilde{a}_1\|^2}\int_0^{T-\varepsilon}(k(s),\tilde{a}_1)u(s)\,ds + \frac{a_2}{\|a_2\|^2}\int_0^{T-\varepsilon}(k(s),a_2)u(s)\,ds \\ + \sum_{i=3}^{\infty}\frac{a_i}{\|a_i\|^2}\int_0^{T-\varepsilon}(k(s),a_i)u(s)\,ds.\end{aligned}$$

We shall show that the series in (9) converges uniformly both in the coordinates and in ε for $\varepsilon \geqq 0$, and thus it is possible to pass to the limit as $\varepsilon \to 0$. We shall denote by A_j the jth coordinate of the series in (9). From (5) we have

$$(10)\qquad A_j = \begin{cases} \displaystyle\sum_{i=3}^{\infty}\frac{a_{ij}}{\|a_i\|^2}\int_0^{T-\varepsilon}(k(s),a_i)u(s)\,ds, & j = 1,2, \\ \displaystyle -\frac{1}{\|a_j\|^2}\int_0^{T-\varepsilon}(k(s),a_j)u(s)\,ds + \sum_{i=j+1}^{\infty}\frac{a_{ij}}{\|a_i\|^2}\int_0^{T-\varepsilon}(k(s),a_i)u(s)\,ds, & \\ & j = 3,4,\cdots. \end{cases}$$

We shall estimate

$$\int_0^{T-\varepsilon}(k(s),a_j)u(s)\,ds.$$

From (5) we have

$$\begin{aligned}\left|\int_0^{T-\varepsilon}(k(s),a_i)u(s)\,ds\right| &\leqq \int_0^{T-\varepsilon}|(k(s),a_i)|\,ds \\ &= \int_0^{T-\varepsilon}\left|\sum_{j=1}^{i-1}a_{ij}e^{-\lambda_j(T-s)} - e^{-\lambda_i(T-s)}\right|ds \\ &\leqq \sum_{j=1}^{i-1}|a_{ij}|\left(\frac{e^{-\lambda_j\varepsilon}}{\lambda_j} - \frac{e^{-\lambda_j T}}{\lambda_j}\right) + \left(\frac{e^{-\lambda_i\varepsilon}}{\lambda_i} - \frac{e^{-\lambda_i T}}{\lambda_i}\right) \\ &< \sum_{j=1}^{i-1}|a_{ij}|\frac{1}{\lambda_j} + \frac{1}{\lambda_i} \\ &\leqq \|a_i\|\left(\sum_{j=1}^{i}\frac{1}{\lambda_j^2}\right)^{1/2} < \|a_i\|\left(\sum_{j=1}^{\infty}\frac{1}{\lambda_j^2}\right)^{1/2}.\end{aligned}$$

Thus

$$(11)\qquad \left|\int_0^{T-\varepsilon}(k(s),a_i)u(s)\,ds\right| < \|a_i\|\left(\sum_{j=1}^{\infty}\frac{1}{\lambda_j^2}\right)^{1/2}.$$

Let us estimate A_j. Taking into account (11) and $\|a_i\| > 1$, we have from (10) that

$$|A_j| \leqq \sum_{i=3}^{\infty} \frac{|a_{ij}|}{\|a_i\|^2} \left| \int_0^{T-\varepsilon} (k(s), a_i) u(s)\, ds \right| \leqq \left(\sum_{j=1}^{\infty} \frac{1}{\lambda_j^2} \right)^{1/2} \sum_{i=3}^{\infty} |a_{ij}|, \qquad j = 1, 2;$$

$$|A_j| \leqq \frac{1}{\|a_j\|^2} \left| \int_0^{T-\varepsilon} (k(s), a_j) u(s)\, ds \right| + \sum_{i=j+1}^{\infty} \frac{|a_{ij}|}{\|a_i\|^2} \left| \int_0^{T-\varepsilon} (k(s), a_i) u(s)\, ds \right|$$
$$\leqq \left(\sum_{j=1}^{\infty} \frac{1}{\lambda_j^2} \right)^{1/2} \left[1 + \sum_{i=j+1}^{\infty} |a_{ij}| \right], \qquad j = 3, 4, \cdots.$$

In the lemma it was proved that $|a_{ij}| \leqq c/\lambda_i$; thus

$$|A_j| \leqq c \left(\sum_{i=1}^{\infty} \frac{1}{\lambda_i^2} \right)^{1/2} \sum_{i=3}^{\infty} \frac{1}{\lambda_i} < c \left(\sum_{i=1}^{\infty} \frac{1}{\lambda_i^2} \right)^{1/2} \sum_{i=1}^{\infty} \frac{1}{\lambda_i}, \qquad j = 1, 2;$$

$$|A_j| \leqq \left(\sum_{i=1}^{\infty} \frac{1}{\lambda_i^2} \right)^{1/2} \left[1 + c \sum_{i=j+1}^{\infty} \frac{1}{\lambda_i} \right] < c_1 \left(\sum_{i=1}^{\infty} \frac{1}{\lambda_i^2} \right)^{1/2} \sum_{i=1}^{\infty} \frac{1}{\lambda_i}, \qquad j = 3, 4, \cdots.$$

The series on the right converges and thus series (9) converges uniformly relative to the coordinates and to ε for $\varepsilon \geqq 0$. Thus in (9) it is possible to pass to the limit as $\varepsilon \to 0$, to obtain

$$\begin{aligned} \int_0^T k(s) u(s)\, ds &= \frac{\tilde{a}_1}{\|\tilde{a}_1\|^2} \int_0^T (k(s), \tilde{a}_1) u(s)\, ds + \frac{a_2}{\|a_2\|^2} \int_0^T (k(s), a_2) u(s)\, ds \\ &\quad + \sum_{i=3}^{\infty} \frac{a_i}{\|a_i\|^2} \int_0^T (k(s), a_i) u(s)\, ds. \end{aligned} \tag{12}$$

The moment problem (4) can be rewritten in vector form as

$$\int_0^T k(s) u(s)\, ds = -v a_2.$$

Comparing this last with (12), we find that solving the moment problem is equivalent to solving the following equations:

$$\int_0^T (k(s), \tilde{a}_1) u(s)\, ds = 0, \tag{13}$$

$$\frac{1}{\|a_2\|^2} \int_0^T (k(s), a_2) u(s)\, ds = -v, \tag{14}$$

$$\int_0^T (k(s), a_i) u(s)\, ds = 0, \qquad i \geqq 3. \tag{15}$$

We shall construct a solution to (13)–(15). In (15) let us take $u(s) \equiv 1$. From the conditions of construction of (5) we have

$$\begin{aligned}
\int_0^T (k(s), a_i)\, ds &= \int_0^T \left[\sum_{j=1}^{i-1} a_{ij} e^{-\lambda_j (T-s)} - e^{-\lambda_i (T-s)} \right] ds \\
&= \left(\sum_{j=1}^{i-1} a_{ij} \frac{1}{\lambda_j} - \frac{1}{\lambda_i} \right) + \left(- \sum_{j=1}^{i-1} a_{ij} \frac{e^{-\lambda_j T}}{\lambda_j} + \frac{e^{-\lambda_i T}}{\lambda_i} \right) \\
&= (a_i, a_2) + e^{-\lambda_1 T} \left[-\frac{a_{i1}}{\lambda_1} - \sum_{j=2}^{i-1} a_{ij} \frac{e^{-(\lambda_j - \lambda_i)T}}{\lambda_j} + \frac{e^{-(\lambda_i - \lambda_1)T}}{\lambda_i} \right] \\
&= (a_i, a_2) + e^{-\lambda_1 T} (a_i, a_1) = 0.
\end{aligned}$$

Thus,

$$\int_0^T (k(s), a_i)\, ds = 0, \qquad i \geqq 3. \tag{16}$$

Thus, for any $T > 0$, (15) is satisfied for $u(s) \equiv 1$. Let us examine (13) with $u(s) \equiv 1$. Taking into account that

$$\tilde{a}_1 = a_1 - \frac{(a_1, a_2)}{\|a_2\|^2} a_2,$$

we have

$$\begin{aligned}
\int_0^T (k(s), \tilde{a}_1)\, ds &= \int_0^T (k(s), a_1)\, ds - \frac{(a_1, a_2)}{\|a_2\|^2} \int_0^T (k(s), a_2)\, ds \\
&= - \int_0^T \left[\frac{e^{-\lambda_1 (T-s)}}{\lambda_1} + \sum_{i=2}^{\infty} \frac{e^{-(\lambda_i - \lambda_1)T}}{\lambda_i} e^{-\lambda_i (T-s)} \right] ds \\
&\quad - \frac{(a_1, a_2)}{\|a_2\|^2} \int_0^T \sum_{i=1}^{\infty} \frac{e^{-\lambda_i (T-s)}}{\lambda_i}\, ds \\
&= - \left[\frac{1}{\lambda_1^2} + \sum_{i=2}^{\infty} \frac{e^{-(\lambda_i - \lambda_1)T}}{\lambda_i^2} \right] + \left[\frac{e^{-\lambda_1 T}}{\lambda_1^2} + \sum_{i=2}^{\infty} \frac{e^{-(\lambda_i - \lambda_1)T}}{\lambda_i^2} e^{-\lambda_i T} \right] \\
&\quad - \frac{(a_1, a_2)}{\|a_2\|^2} \left[\sum_{i=1}^{\infty} \frac{1}{\lambda_i^2} - \sum_{i=1}^{\infty} \frac{e^{-\lambda_i T}}{\lambda_i^2} \right] \\
&= (a_1, a_2) + e^{-\lambda_1 T}(a_1, a_1) - \frac{(a_1, a_2)}{\|a_2\|^2} [(a_2, a_2) + e^{-\lambda_1 T}(a_1, a_2)] \\
&= e^{-\lambda_1 T} \left[(a_1, a_1) - \frac{(a_1, a_2)^2}{(a_2, a_2)} \right].
\end{aligned}$$

Thus,

$$\int_0^T (k(s), \tilde{a}_1)\, ds = e^{-\lambda_1 T} \left[(a_1, a_1) - \frac{(a_1, a_2)^2}{(a_2, a_2)} \right]. \tag{17}$$

We shall show that (17) is valid not because of the smallness of the function $(k(s), \tilde{a}_1)$ for large T, but rather because $u(s) \equiv 1$ is "almost" orthogonal to it. To that end, we shall find the norm in $L_2[0, T]$ of the function $(k(s), \tilde{a}_1)$:

$$\|(k(s), \tilde{a}_1)\|_{L_2}^2 = \int_0^T (k(s), \tilde{a}_1)^2 \, ds$$

$$= \int_0^T \left[\frac{e^{-\lambda_1(T-s)}}{\lambda_1} + \sum_{i=2}^{\infty} \frac{e^{-(\lambda_i - \lambda_1)T}}{\lambda_i} e^{-\lambda_i(T-s)} + \frac{(a_1, a_2)}{\|a_2\|^2} \sum_{i=1}^{\infty} \frac{e^{-\lambda_i(T-s)}}{\lambda_i} \right]^2 ds.$$

Let us assume that $\lambda_1 < \lambda_2 \leqq \lambda_3 \leqq \cdots \leqq \lambda_k \leqq \cdots$. If this is not so, then we combine with $e^{-\lambda_1(T-s)}$ all those coinciding with it (a finite number of them). Let us use the notation

$$\alpha = \frac{1}{\lambda_1} + \frac{(a_1, a_2)}{\lambda_1 \|a_2\|^2}, \qquad \beta_i = -\frac{1}{\lambda_i \alpha} \left[e^{-(\lambda_i - \lambda_1)T} + \frac{(a_1, a_2)}{\|a_2\|^2} \right].$$

Then

$$\|k(s), \tilde{a}_1\|_{L_2}^2 = \int_0^T \left[\alpha e^{-\lambda_1(T-s)} - \alpha \sum_{i=2}^{\infty} \beta_i e^{-\lambda_i(T-s)} \right]^2 ds$$

$$= \alpha^2 \int_0^T \left[e^{-\lambda_1(T-s)} - \sum_{i=2}^{\infty} \beta_i e^{-\lambda_i(T-s)} \right]^2 ds.$$

In the last integral, we make the change of variable $e^{-(T-s)} = t$. Then

$$\|(k(s), \tilde{a}_1)\|^2 = \alpha^2 \int_{e^{-T}}^1 \left[t^{\lambda_1 - 1/2} - \sum_{i=2}^{\infty} \beta_i t^{\lambda_i - 1/2} \right]^2 dt. \tag{18}$$

From a theorem of Münz [2] it follows that the system of functions $\{t^{\lambda_i - 1/2}\}$, $i = 2, \cdots$, is not complete in $L_2[0, 1]$ and the function $t^{\lambda_1 - 1/2}$ is thus a positive distance from the subspace spanned by this system, i.e.,

$$\int_0^1 \left[t^{\lambda_1 - 1/2} - \sum_{i=2}^{\infty} \beta_i t^{\lambda_i - 1/2} \right]^2 dt \geqq \rho > 0 \tag{19}$$

for all $\beta_2, \beta_3, \cdots$. Since $|\beta_i| \leqq c/\lambda_i$, the function $t^{\lambda_1 - 1/2} - \sum_{i=2}^{\infty} \beta_i t^{\lambda_i - 1/2}$ is bounded on $[0, 1]$, and thus

$$\int_0^{e^{-T}} \left[t^{\lambda_1 - 1/2} - \sum_{i=2}^{\infty} \beta_i t^{\lambda_i - 1/2} \right]^2 dt \leqq c e^{-T}. \tag{20}$$

From (18)–(20) we know that, for any sufficiently large T,

$$\|(k(s), \tilde{a}_1)\|^2 = \alpha^2 \int_{e^{-T}}^1 \left[t^{\lambda_1 - 1/2} - \sum_{i=2}^{\infty} \beta_i t^{\lambda_i - 1/2} \right]^2 dt \geqq \rho_1 > 0. \tag{21}$$

We shall examine in $L_2[0, T]$ the system of functions

$$\{\varphi(s) = (k(s), \tilde{a}_1), (k(s), a_i), i \geqq 3\}$$

and expand $u(s) \equiv 1$ in this system and its orthogonal complement, which is not empty ([2, Münz's theorem]):

$$1 = \psi(s) + \frac{(1,\varphi)}{\|\varphi\|^2}\varphi(s), \tag{22}$$

where $\psi(s)$ is in the orthogonal complement of the system

$$\{\varphi(s), (k(s), a_i), i \geqq 3\}.$$

Components involving the functions $(k(s), a_i)$, $i \geqq 3$, are absent from (22) because from (16), $u(s) \equiv 1$ is orthogonal to the system $\{(k(s), a_i), i \geqq 3\}$. The function $\varphi(s) = (k(s), \tilde{a}_1)$ is bounded. From (21), $\|\varphi(s)\|^2 \geqq \rho_1 > 0$, and from (17), $(1, \varphi) = ce^{-\lambda_1 T}$. Thus in (22) the component $(1, \varphi)\varphi(s)/\|\varphi\|^2$ is arbitrarily small for sufficiently large T. Since the expansion (22) is true for all $s \in [0, T]$, the function $\psi(s)$ can be made arbitrarily close to unity. Two cases are possible: either $\psi(s) < 1$, $s \in [0, T]$, or somewhere $\psi(s) > 1$. In the second case, we shall divide (22) by $r = \max \psi(s)$, where r is close to unity. We shall use the notation $\psi_1(s) = \psi(s)/r$. From (22) we have

$$\psi_1(s) = \frac{1}{r} - \frac{(1,\varphi)}{r\|\varphi\|^2}\varphi(s).$$

The function $\psi_1(s)$ does not exceed unity for $s \in [0, T]$. Since $\psi_1(s)$ belongs to the orthogonal complement of the system $\{(k(s), \tilde{a}_1), (k(s), a_i), i \geqq 3\}$, (13) and (15) are satisfied with $u(s) = \psi_1(s)$. With $u(s) = \psi_1(s)$ the left side of (14) has the form

$$\begin{aligned} &\frac{1}{\|a_2\|^2}\int_0^T (k(s), a_2)\psi_1(s)\,ds \\ &\quad = \frac{1}{r\|a_2\|^2}\int_0^T (k(s), a_2)\,ds - \frac{(1,\varphi)}{r\|\varphi\|^2\|a_2\|^2}\int_0^T (k(s), a_2)\varphi(s)\,ds. \end{aligned} \tag{23}$$

We estimate the second term on the right:

$$\begin{aligned} &\frac{(1,\varphi)}{r\|\varphi\|^2\|a_2\|^2}\int_0^T (k(s), a_2)(k(s), \tilde{a}_1)\,ds \\ &\quad = \frac{(1,\varphi)}{r\|\varphi\|^2\|a_2\|^2}\left[\int_0^T (k(s), a_2)(k(s), a_1)\,ds - \frac{(a_1, a_2)}{\|a_2\|^2}\int_0^T (k(s), a_2)^2\,ds\right]. \end{aligned} \tag{24}$$

The integrals in the brackets can be estimated as

$$\begin{aligned} &\left|\int_0^T (k(s), a_2)(k(s), a_1)\,ds\right| \\ &\quad \leqq \int_0^T \left[\sum_{i=1}^{\infty} \frac{e^{-\lambda_i(T-s)}}{\lambda_i}\right]\left[\frac{e^{-\lambda_1(T-s)}}{\lambda_1} + \sum_{i=2}^{\infty} \frac{e^{-(\lambda_i-\lambda_1)T}}{\lambda_i} e^{-\lambda_i(T-s)}\right] ds \\ &\quad \leqq \left[\frac{1}{\lambda_1} + \sum_{i=2}^{\infty} \frac{e^{-(\lambda_i-\lambda_1)T}}{\lambda_i}\right]\int_0^T \sum_{i=1}^{\infty} \frac{e^{-\lambda_i(T-s)}}{\lambda_i}\,ds < T\left(\sum_{i=1}^{\infty}\frac{1}{\lambda_i}\right)^2, \end{aligned}$$

$$\int_0^T (k(s), a_2)^2\,ds = \int_0^T \left(\sum_{i=1}^{\infty} \frac{e^{-\lambda_i(T-s)}}{\lambda_i}\right)^2 ds < T\left(\sum_{i=1}^{\infty}\frac{1}{\lambda_i}\right)^2.$$

Substituting these into (24) and taking account of (17), we have

$$(25)\qquad \left|\frac{(1,\varphi)}{r\|\varphi\|^2\|a_2\|^2}\int_0^T (k(s),a_2)(k(s),\tilde{a}_1)\,ds\right| < cTe^{-\lambda_1 T},$$

and thus the second term on the right in (23) is arbitrarily small for large T. Let us consider the first term on the right in (23):

$$(26)\ \frac{1}{r\|a_2\|^2}\int_0^T (k(s),a_2)\,ds = \frac{1}{r\|a_2\|^2}\int_0^T \sum_{i=1}^{\infty}\frac{e^{-\lambda_i(T-s)}}{\lambda_i}\,ds = \frac{1}{r} + e^{-\lambda_1 T}\frac{(a_1,a_2)}{r\|a_2\|^2}.$$

Since $r = \max \psi(s)$ is arbitrarily near to unity for large T, so also is the first term on the right of (23). We shall denote the left side of (23) by w.

From (26), (25) and (23), we have

$$(27)\qquad w = \frac{1}{r} + e^{-\lambda_1 T}\frac{(a_1,a_2)}{r\|a_2\|^2} - c_2Te^{-\lambda_1 T}.$$

Equations (13), (15) will be satisfied for $u(s) = -\psi_1(s)v/w$, since $\psi_1(s)$ is orthogonal to the functions $(k(s),\tilde{a}_1)$, $(k(s),a_i)$, $i \geqq 3$. Equation (14) is also satisfied, since

$$\frac{1}{\|a_2\|^2}\int_0^T (k(s),a_2)u(s)\,ds = \frac{-v}{w\|a_2\|^2}\int_0^T (k(s),a_2)\psi_1(s)\,ds = -v.$$

It only remains to verify that $|u(s)| \leqq 1$. But since $|v| < 1$ (from the hypotheses of the theorem), $|\psi_1(s)| \leqq 1$, and since from (27) it follows that $w \to 1$ as $T \to \infty$, then, for sufficiently large T, we have $|\psi_1(s)v/w| \leqq 1$. Thus, the problem of moments (4) is solved.

Remark 1. From the form of the construction, it is apparent that the control is a differentiable function.

Remark 2. We shall show that if $|v| = 1$ in (4), then for no finite T does there exist a solution of (4).

Let $v = 1$. We take $u(s) \equiv 1$. Then the left side of (4) has the form

$$\int_0^T e^{-\lambda_i(T-s)}\,ds = \frac{1}{\lambda_i} - \frac{e^{-\lambda_i T}}{\lambda_i}.$$

But for any finite T, this is less than the right side of (4), which is $1/\lambda_i$; thus for $v = 1$ there does not exist a $T < \infty$ for which (4) would have a solution. Analogously, if $v = -1$, for $u(s) \equiv -1$ we obtain the same result.

3. The existence and uniqueness of an optimal control. From Theorem 1 it follows that for $|v| < 1$ there exists a $T < \infty$ and a measurable function $u(s)$, with $|u(s)| \leqq 1$, for which problem (1)–(3) is solvable.

THEOREM 2. *Let* $|v| < 1$ *and*

$$\sum_{i=1}^{\infty}\frac{1}{\lambda_i} < \infty.$$

Then for problem (1)–(3) *there exist* $u^*(s)$ *and* T^*, *where* T^* *is the minimum time. The control* $u^*(s)$ *is unique, and* $|u^*(s)| = 1$ *almost everywhere on* $[0, T]$.

Proof. Existence. We shall denote by $\{T\}$ and $\{u(s)\}$ the times and corresponding controls for which problem (1)–(3) or the moment problem (4) is solvable. In Theorem 1 it is proved that these sets are not empty. Let us define $T^* = \inf\{T\}$. If $T^* \in \{T\}$, then the corresponding $u^*(s)$ is an optimal control. Suppose $T^* \notin \{T\}$. We shall rewrite (4) thus:

$$\int_0^T \chi_T(s)u_T(s)e^{-\lambda_k(T-s)}\,ds = -\frac{v}{\lambda_k}, \qquad k = 1, 2, \cdots,$$

where $\chi_T(s)$ is the characteristic function of the segment $[0, T]$, and T and $u_T(s)$ are the time and control constructed in Theorem 1. Since $T^* = \inf\{T\}$, there exists a sequence $T_n \to T^*$, with a corresponding sequence $\chi_{T_n}(s)u_{T_n}(s)$. With this,

$$\int_0^T \chi_{T_n}(s)u_{T_n}(s)e^{-\lambda_k(T_n-s)}\,ds = -\frac{v}{\lambda_k}.$$

Since the set $|u(s)| \leqq 1$ is weakly compact in $L_2[0, T]$, then for $\chi_{T_n}(s)u_{T_n}(s)$ there exists a weak limit, i.e.,

$$\lim_{n\to\infty}\int_0^T \chi_{T_n}(s)u_{T_n}(s)e^{-\lambda_k(T_n-s)}\,ds = \int_0^T \chi_{T^*}(s)u_{T^*}(s)e^{-\lambda_k(T^*-s)}\,ds$$
$$= \int_0^{T^*} u_{T^*}(s)e^{-\lambda_k(T^*-s)}\,ds = -\frac{v}{\lambda_k}.$$

The last equality can be written because it holds for all pre-limit integrals. The resulting $u_{T^*}(s) = u^*(s)$ and is an optimal control.

The proof that $u^*(s)$ is unique, and that $|u^*(s)| = 1$ almost everywhere on $[0, T]$, follows from Theorem 5 proved by Yu. V. Egorov in [1].

Acknowledgment. The author is deeply grateful to I. V. Girsanov for posing the problem, and for his great attention during its solution.

REFERENCES

[1] Yu. V. Egorov, *Some problems in the theory of optimal control*, Zh. Vychisl. Mat. i Mat. Fiz., 3 (1963), no. 5, pp. 887–904.

[2] N. I. Akhiezer, *Lectures on Approximation Theory*, Nauka, Moscow, 1965.

SIAM J. CONTROL
Vol. 7, No. 4, November 1969

A PARAMETRIC METHOD FOR SEMIDEFINITE QUADRATIC PROGRAMS*

M. D. GRIGORIADIS† AND K. RITTER‡

Abstract. A parametric method for solving semidefinite quadratic programs with a large number of constraints is described. All computations are performed by pivotal operations on a tableau, or more efficiently on an inverse which is considerably smaller than that used by other methods. This inverse is updated by elementary row and column operations. Programming of the algorithm is facilitated by its efficient use of the product form of the inverse mechanism in most commercially available linear programming systems. An existing solution to a slightly perturbed problem, if available, may be used as a starting solution for a new problem, with a possible substantial reduction of the required computational effort. Finally, an obvious but rather important advantage of the method is its direct use in post-optimality studies involving the requirements vector and/or the linear part of the objective function.

1. Introduction. In recent years, quadratic programming has found several important applications in business, science and engineering. These include problems in portfolio selection, linear regression analysis with inequality contraints on the coefficients, maximization of consumer's utility in the framework of classical consumption theory, profit maximization under resource constraints, quadratic approximation of general convex programs, pattern recognition and others. This paper, however, was motivated by the application of quadratic programming as a computational method for discrete optimal control problems (see [21]).

Being a special case of convex programs, the quadratic programming problem presents a most desirable feature: the linearity of the objective function gradient. The resulting Kuhn-Tucker optimality conditions [9] are linear with the exception of the "complementarity" condition imposed on certain pairs of variables.

Algorithms for the solution of quadratic programs are abundant in the recent literature (see, e.g., [1], [3], [6], [10], [14], [15], [19], [20], [23], [25]). In general, these methods have the use of the Kuhn-Tucker optimality conditions as an auxiliary problem in common. This is usually solved by modified simplex operations on tableaux of row size $(m + n)$. Excellent reviews of these methods may be found in [2], [3], [8] and computational experiments on their relative efficiencies for small problems found in [13].

Recently, general treatment of this auxiliary problem was effected by the so-called "complementary pivoting theory" (see, e.g., [5], [12], [22]). It provides a unified approach for solving quadratic programs and determining equilibrium points of bimatrix games, and forms the basis for identifying classes of problems which have a complementary solution (see, e.g., [4], [7], [11], [16]). The resulting computational scheme, however, also requires tableaux of row size $(m + n)$.

The parametric method outlined in this paper is a generalization of the basic idea in [17] to the semidefinite case. Nonnegativity restrictions are represented

* Received by the editors May 3, 1968, and in revised form February 17, 1969. This work was supported in part by the National Science Foundation under Grant GP6070 and in part by the Mathematics Research Center, U.S. Army, Madison, Wisconsin, under Contract DA 31-124-ARO-D-462.

† Department of Computer Sciences, University of Wisconsin, Madison, Wisconsin 53706 and New York Scientific Center, IBM Corporation, New York, New York 10021.

‡ Department of Computer Sciences, University of Wisconsin, Madison, Wisconsin 53706 and Mathematics Research Center, United States Army, Madison, Wisconsin.

and handled implicitly. For problems with $m \leqq n$ a constant tableau size of $m + n$, and for problems with $m > n$ a constant tableau size of $2n + 1$ is used for all computations. The method is particularly suited for use with the product form of the inverse mechanism of existing LP codes after minor modifications. The inverse is updated by elementary row and column operations. The most important and obvious advantages of this method are its use for postoptimality studies, i.e., varying the requirement vector b and/or the linear part of the objective function, and its ability to utilize an existing solution to a slightly perturbed problem as a starting solution.

2. The problem. The Quadratic Programming Problem (QP) is defined as:

$$\max_x \{Q(x) = c'x - \tfrac{1}{2}x'Cx | Ax \leqq b, x \geqq 0\}, \tag{2.1}$$

where c and x are n-vectors, C is an (n, n)-symmetric positive semidefinite (p.s.d.) matrix, b is an m-vector and A an (m, n)-matrix[1]. Prime denotes the transpose.

In this paper, QP will be treated by considering the Parametric Quadratic Program (PQP): For each θ in a given real closed interval, say $[0, \theta_0]$, find

$$\max_x \{Q(x, \theta) = (c + \theta d)'x - \tfrac{1}{2}x'Cx | Ax \leqq b + \theta f, x \geqq 0\}, \tag{2.2}$$

where d and f are given n- and m-vectors respectively. Clearly, PQP reduces to QP for $\theta = 0$.

The Kuhn-Tucker necessary optimality conditions [9] for PQP state that if $x = x_0$ solves PQP for some $\theta = \theta_0$, then there exist Lagrange mutipliers u and v, associated with the respective constraints of (2.2), such that

$$Cx + A'u - v = c + \theta_0 d, \tag{2.3.1}$$

$$Ax + y = b + \theta_0 f, \tag{2.3.2}$$

$$v'x + u'y = 0, \tag{2.3.3}$$

$$x, y, u, v \geqq 0, \tag{2.3.4}$$

where u, v are m- and n-vectors and y is an m-vector of slack variables. These conditions are also sufficient when C is p.s.d. We note that the relations (2.3), to be referred to later as the Auxiliary Problem (AP), are linear except for the "complementary slackness" conditions (2.3.3.).

An important aspect of the proposed algorithm is its ability to determine the nature of solutions to QP or PQP from information available in the simplex tableau of AP. For example if, for some $\theta = \theta_k$, PQP (or QP) have no feasible solution, it may be determined from the corresponding AP by a simple test. Similarly, if for some $\theta = \theta_k$, PQP (or QP) have no optimal solution, or have an unbounded solution (provided that they have nonempty feasible domains), this may be detected from the corresponding AP.

[1] Constraints of the form $Ax = b$ may be represented by $Ax \leqq b$ and the additional constraint $-a'_{m+1}x \leqq -\beta$ where $a_{m+1} = \sum_{i=1}^{m} a_i$ and $\beta = \sum_{i=1}^{m} (b)_i$.

3. The algorithm. The first step toward obtaining an optimal solution to a given QP, is to choose an x_0 and determine d, f and θ so that x_0 is an optimal solution to PQP. The selection of:

$$(d)_i = \begin{cases} -1 & \text{if } (c)_i \geqq 0 \\ 0 & \text{otherwise} \end{cases}, \qquad i = 1, \cdots, n, \tag{3.1.1}$$

$$(f)_j = \begin{cases} 1 & \text{if } (b)_j \leqq 0 \\ 0 & \text{otherwise} \end{cases}, \qquad j = 1, \cdots, m, \tag{3.1.2}$$

$$\theta_0 = \max_{i,j} \{0, (c)_i, -(b)_j\}, \quad i = 1, \cdots, n, j = 1, \cdots, m, \tag{3.1.3}$$

insures that $x_0 = 0$ is an optimal solution to PQP for $\theta = \theta_0$. If $\theta_0 = 0$, then $x_0 = 0$ solves QP as well as PQP. Assuming $\theta_0 > 0$, we intend to solve QP by parametrically solving PQP for values of $\theta < \theta_0$ until a solution, if it exists, is obtained for $\theta = 0$. Such successive solutions are obtained by using simplex tableaux of the corresponding AP. Initially, for $\theta = \theta_0$, (2.3.2) and (3.1.2) give $y_0 = b + \theta_0 f$ and $y_0 \geqq 0$. Similarly, (2.3.1) implies $v_0 = -c - \theta_0 d$ and (3.1.1), (2.3.1) imply $v_0 \geqq 0$. Therefore ($x_0 = 0$, $u_0 = 0$, y_0, v_0) solves AP for $\theta = \theta_0$. In later cycles the conditions (2.2.3) are handled by the proper choice of pivot. A "pivot operation" is regarded as an exchange of an active constraint and an inactive one, or as it is frequently portrayed, as an exchange of a basic slack variable (corresponding to the inactive constraint) and a nonbasic one (corresponding to the active constraint).

Since PQP has n variables, at most n constraints are needed to determine the optimal solution of PQP for some $\theta \leqq \theta_0$. This implies that at most n components of u can be basic in AP. Hence, at least $(m - n)$ components of y, corresponding to the inactive constraints, are basic in AP. In degenerate cases, some of these may be at zero level.

This observation suggests partitioning the rows of A as follows. Assume that for $\theta = \theta_k \leqq \theta_0$, we have $p(\leqq n)$ active constraints. Denote by A^c the $(n + 1, n)$-submatrix of A which includes the p active rows and $(n - p + 1)$ of the inactive ones. These will be referred to as the "current" rows of A. Similarly, let A^s denote the $(m - n - 1, n)$-submatrix of A, consisting of the remaining inactive rows of A, to be referred to as the "stand by" rows of A. The corresponding partitioning of u and y gives (u^c, u^s) and (y^c, y^s) respectively. This is shown in Fig. 1 where v, y^c and y^s have been included for exposition purposes.

$$\begin{array}{r|cccc|cc|l}
 & x, & u^c, & v, & y^c & u^s, & y^s & \\
\hline
n\ \{ & C & A^{c'} & -I & 0 & A^{s'} & 0 & = c + \theta d \\
n+1\ \{ & A^c & 0 & 0 & I & 0 & 0 & = b^c + \theta f^c \\
\cline{2-5}
m-n-1\ \{ & A^s & 0 & 0 & 0 & 0 & I & = b^s + \theta f^s \\
 & n & n+1 & n & n+1 & m-n-1 & m-n-1 &
\end{array}$$

FIG. 1. *Matrix structure of Auxiliary Problem* (*AP*)

We further observe that since y^s is basic and $u^s = 0$, all information concerning the solution of AP may be obtained from the $(2n + 1, 4n + 2)$-submatrix within dashed lines in Fig. 1. Thus, the necessary variable exchange (pivot) operations may be performed on the Reduced Auxiliary Problem (RAP):

$$Cx + A^{c'}u^c - v = c + \theta d, \tag{3.2.1}$$

$$A^c x + y^c = b^c + \theta f^c, \tag{3.2.2}$$

$$x, u^c, v, y^c \geqq 0, \tag{3.2.3}$$

always taking into account the side conditions $v'x + u^{c'}y^c = 0$ (and $u^{s'}y^s = 0$).

Since initially $x_0 = 0$, $u_0 = 0$, the variables v and y are basic in the tableau of RAP. If, during the course of the algorithm, a constraint in A^s becomes active, the current partitioning of A will be updated to include this constraint in A^c, in exchange for a currently inactive one in A^c, which is brought into A^s. This operation, also handled by pivoting, requires less effort than an ordinary pivot step.

At any parametric step ($\theta = \theta_k \leqq \theta_0$), the $(2n + 1)$ order working basis B_k induces the following partitioning on the RAP matrix

$$(3.3)\ (B_k, B_{2k}) = \begin{array}{c} \begin{array}{cccccccc} x_{1k}, & v_{1k}, & u^c_{1k}, & y^c_{1k}, & x_{2k}, & v_{2k}, & u^c_{2k}, & y^c_{2k} \end{array} \\ \left[\begin{array}{cccccccc} C_{11} & 0 & A^{c'}_{11} & 0 & C_{12} & 0 & A^{c'}_{21} & 0 \\ C_{21} & -I & A^{c'}_{12} & 0 & C_{22} & -I & A^{c'}_{22} & 0 \\ A^c_{11} & 0 & 0 & 0 & A^c_{12} & 0 & 0 & 0 \\ A^c_{21} & 0 & 0 & I & A^c_{22} & 0 & 0 & I \end{array}\right] \end{array} \begin{array}{l} \\ \left.\begin{array}{l} \}\, p \\ \}\, n - p \end{array}\right\} n \\ \left.\begin{array}{l} \}\, q \\ \}\, n + 1 - q \end{array}\right\} m \end{array}$$

where $x_{1k}, v_{1k}, u^c_{1k}, y^c_{1k}$ denote the basic RAP variables and where the partitions

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}, \quad A^c = \begin{pmatrix} A^c_{11} & A^c_{12} \\ A^c_{21} & A^c_{22} \end{pmatrix}, \quad \begin{array}{ll} x = (x_1, x_2), & v = (v_1, v_2), \\ y^c = (y^c_1, y^c_2), & u^c = (u^c_1, u^c_2), \end{array}$$

have been used. Corresponding partitions are: $(c + \theta d) = (c_1 + \theta d_1, c_2 + \theta d_2)$, $(b^c_1 + \theta f^c_1, b^c_2 + \theta f^c_2)$ and $A^s = (A^s_1, A^s_2)$.

For convenience let $M = \{1, \cdots, m\}$, $N = \{1, \cdots, n\}$ and $I \subseteq M$, $I^s = M - I$ be the index sets corresponding to the rows of A^c and A^s respectively. Let $J_B \subseteq N$ contain all indices of the components of the p-vector x_1 and $I_B \subseteq I$ contain all indices of the active constraint rows (or the components of the q-vector u^c_1.) Similarly, let $J^*_B = N - J_B$ be the indices corresponding to the $(n - p)$-vector v_1 and $I^*_B = I - I_B$ be those corresponding to the $(n + 1 - q)$-vector y^c_1. The basic variables (with respect to B_k; $\theta = \theta_k$) will be denoted by $(x_{1k})_j$ for $j \in J_{B_k}$; $(v_{1k})_j$ for $j \in J^*_{B_k}$; $(u^c_{1k})_j$ for $j \in I_{B_k}$ and $(y^c_{1k})_j$ for $j \in I^*_{B_k}$. We consider $(y^s_k)_j$ for $j \in I^s_{B_k}$ as "implicitly basic" since they are basic in the tableau of AP but do not appear in the RAP basis.

The current basic solution to RAP, as a function of the parameter θ is given by:

$$\begin{array}{l} (3.4.1) \\ (3.4.2) \\ (3.4.3) \\ (3.4.4) \end{array} \qquad \begin{pmatrix} x_{1k}(\theta) \\ v_{1k}(\theta) \\ u^c_{1k}(\theta) \\ y^c_{1k}(\theta) \end{pmatrix} = B_k^{-1} \begin{pmatrix} c_1 + \theta d_1 \\ c_2 + \theta d_2 \\ b^c_1 + \theta f^c_1 \\ b^c_2 + \theta f^c_2 \end{pmatrix} = \begin{pmatrix} p_1 + \theta p_2 \\ p_3 + \theta p_4 \\ p_5 + \theta p_6 \\ p_7 + \theta p_8 \end{pmatrix}, \qquad \theta = \theta_k,$$

and the implicitly basic slacks:

$$y_k^s(\theta) = b^s + \theta f^s - A_1^s x_{1k}(\theta),$$

or:

$$(3.4.5) \qquad y_k^s(\theta) = (b^s - A_1^s p_1) + \theta(f^s - A_1^s p_2) = p_9 + \theta p_{10},$$

where the $p_i (i = 1, \cdots, 10)$ have been introduced for notational convenience.

Now assume that for $\theta = \theta_k$ an optimal solution $x_k(\theta_k)$ has been obtained. We wish to obtain the smallest value of the parameter, say $\theta = \theta_l \leqq \theta_k$, such that $x_k(\theta), \theta_l \leqq \theta \leqq \theta_k$ remains both optimal and feasible.

First, we determine the smallest θ, say $\theta = \theta_l^* \leqq \theta_k$, such that the optimality conditions $v_{1k}(\theta) \geqq 0$ and $u_{1k}^c(\theta) \geqq 0$ are satisfied. From $v_{1k}(\theta) \geqq 0$ and (3.4.2) we have

$$(3.5.1) \qquad \theta_{l1}^* = -(p_3)_{\mu_1}/(p_4)_{\mu_1} = \max_j \{-(p_3)_j/(p_4)_j | (p_4)_j > 0 \text{ and } j \in J_{Bk}^*\}.$$

If $(p_4)_j \leqq 0$ for all $j \in J_{B_k}^*$, let $\theta_{l1}^* = 0$. From $u_{1k}^c \geqq 0$ and (3.4.3)

$$(3.5.2) \qquad \theta_{l2}^* = -(p_5)_{\mu_2}/(p_6)_{\mu_2} = \max_j \{-(p_5)_j/(p_6)_j | (p_6)_j > 0 \text{ and } j \in I_{B_k}\}.$$

If $(p_6)_j \leqq 0$ for all $j \in I_{B_k}$, let $\theta_{l2}^* = 0$. Thus $\theta_l^* = \max\{\theta_{l1}^*, \theta_{l2}^*\}$ and the limiting variable index is $\mu = \mu_i$ if $\theta_l^* = \theta_{li}^*$, $i = 1, 2$.

Second, we determine the smallest θ, say $\theta = \theta_l^{**} \leqq \theta_k$ such that the feasibility conditions $x_{1k}(\theta) \geqq 0$, $y_{1k}^c(\theta) \geqq 0$, $y_k^s(\theta) \geqq 0$ are satisfied. These conditions, in the given order, with (3.4.1), (3.4.4) and (3.4.5), give

$$(3.6.1) \qquad \theta_{l1}^{**} = -(p_1)_{\rho_1}/(p_2)_{\rho_1} = \max\{-(p_1)_j/(p_2)_j | (p_2)_j > 0 \text{ and } j \in J_{B_k}\}.$$

If $(p_2)_j \leqq 0$ for all $j \in J_{B_k}$, let $\theta_{l1}^{**} = 0$.

$$(3.6.2) \qquad \theta_{l2}^{**} = -(p_7)_{\rho_2}/(p_8)_{\rho_2} = \max\{-(p_7)_j/(p_8)_j | (p_8)_j > 0 \text{ and } j \in I_{B_k}^*\}.$$

If $(p_8)_j \leqq 0$ for all $j \in I_{B_k}^*$, let $\theta_{l2}^{**} = 0$.

$$(3.6.3) \qquad \theta_{l3}^{**} = -(p_9)_{\rho_3}/(p_{10})_{\rho_3} = \max\{-(p_9)_j/(p_{10})_j | (p_{10})_j > 0 \text{ and } j \in I_k^s\}.$$

If $(p_{10})_j \leqq 0$ for all $j \in I_k^s$, let $\theta_{l2}^{**} = 0$. Therefore, $\theta_l^{**} = \max\{\theta_{l1}^{**} | i = 1, 2, 3\}$ and the limiting variable index is $\rho = \rho_i$ if $\theta_l^{**} = \theta_{li}^{**}$; $i = 1, 2, 3$. Finally,

$$(3.7) \qquad \theta_l = \max\{\theta_l^*, \theta_l^{**}\} = \max\{\theta_{l1}^*, \theta_{l2}^*, \theta_{l1}^{**}, \theta_{l2}^{**}, \theta_{l3}^{**}\}.$$

Next, we investigate the nature of the basic solution to AP for $\theta = \theta_l - \varepsilon > 0$ for a small $\varepsilon > 0$. Depending on the limiting value of θ_l in (3.7), an appropriate basis change must be performed in order to restore optimality or feasibility. The corresponding variable exchange, performed by a pivot operation, is equivalent to updating the status of a constraint, be it a nonnegativity restriction or an ordinary constraint, from "active" to "inactive" and vice versa. In the absence of degeneracy, once the exiting variable is defined by (3.5) or (3.6), the entering variable is uniquely determined by (2.3.3). The value of the objective function is not altered by such exchanges.

We now examine the types of pivot operations which will restore optimality (Case 1) or feasibility (Case 2) for $\theta = \theta_l - \varepsilon$. Characteristically, in Case 1 a currently (i.e., for $\theta = \theta_l$) active constraint will be reclassified as inactive, and in Case 2 a currently inactive one as active. Thus, one of the following operations must be performed for $\theta = \theta_l$:

Case 1 ($\theta_l = \theta_l^*$). The constraint to become *inactive* is a:

(i) *nonnegativity restriction* ($\theta_l = \theta_l^* = \theta_{l1}^* \leqq \theta_k$). From (3.5.1), $(v_{1k})_\mu = 0$ for at least one $\mu \in J_{B_k}^*$. In view of (2.3.3) replace $(v_{1k})_\mu$ by $(x_{2k})_\mu$.

(ii) *ordinary constraint* ($\theta_l - \theta_l^* = \theta_{l2}^* \leqq \theta_k$). From (3.5.2), $(u_{1k}^c)_\mu = 0$ for at least one $\mu \in I_{B_k}$. In view of (2.3.3), replace $(u_{1k}^c)_\mu$ by $(y_{2k}^c)_\mu$.

Case 2 ($\theta_l = \theta_l^{**}$). The constraint to become *active* is a:

(i) *nonnegativity restriction* ($\theta_l = \theta_l^{**} = \theta_{l1}^{**} \leqq \theta_k$). From (3.6.1), $(x_{1k})_\rho = 0$ for at least one $\rho \in J_{B_k}$. In view of (2.3.3) replace $(x_{1k})_\rho$ by $(v_{2k})_\rho$.

(ii) *ordinary constraint in* A^c ($\theta_l = \theta_l^* = \theta_{l2}^{**} \leqq \theta_k$). From (3.6.2), $(y_{1k}^c)_\rho = 0$ for at least one $\rho \in I_{B_k}^*$. In view of (2.3.3), replace $(y_{1k}^c)_\rho$ by $(u_{1k}^c)_\rho$.

(iii) *ordinary constraint in* A^s ($\theta_l = \theta_l^{**} = \theta_{l3}^{**} \leqq \theta_k$). From (3.6.3), $(y_k^s)_\rho = 0$ for at least one $\rho \in I_k^s$. By definition, $(y_k^s)_\rho$ is implicitly basic. Now that the ρth constraint in A^s will become active, we must:

(a) Update the definition of the current and stand-by constraints, i.e., update the sets I and I^s, by defining the ρth constraint presently in A^s as current, in exchange for an inactive constraint presently in A^c, say the τth. Such a constraint will always be present in A^c since not more than n of the $(n + 1)$ constraints in A^c may be active. This updated partitioning of A defines a new RAP. It is obtained from the current RAP by updating B_k^{-1} as shown in § 4. This has the effect of replacing the elements of the row corresponding to the τth constraint by the elements of the ρth constraint (which is to be made active) and furthermore, of replacing the basic variable $(y_{1k}^c)_\tau$ by the implicitly basic variable $(y_k^s)_\rho$. (Note that the updated partitioning (y^c, y^s) requires that $(y_k^s)_\rho$ be denoted by $(y_{1k}^c)_\tau$. This notational liberty should not cause confusion). In the new RAP, $(y_{1k}^c)_\tau$ is at zero level and must be removed from the basis by (b).

(b) Replace the currently basic $(y_{1k}^c)_\tau$ by its complementary variable $(u_{1k}^c)_\rho$. This step is the same as (ii) above.

Lemma 2 guarantees the nonpositivity of the pivot element. If it is negative, we perform one pivot step (see § 4) and return to (3.4).

If the pivot element is zero, the sought exchange of variables requires a pair of pivot steps. Their validity and existence is demonstrated in Theorem 2. The question of a zero pivot (in Cases 1 and 2) and its remedy will now be discussed in detail.

Case 1. If the pivot element is zero, it can be shown (Lemma 3) that PQP has an infinite number of optimal solutions for $\theta = \theta_l$. However, through simple examples it may be shown that not all elements x_k^* of this infinite set of optimal solutions need have the property that a function $x(\theta)$ exists, such that $x_k^* = x_k(\theta_l)$ and such that $x(\theta)$ solve PQP for some interval $\theta_p \leqq \theta \leqq \theta_l$ with $\theta_p < \theta_l$. A zero pivot indicates that x_k^* does not have this property. In order to continue our parametric procedure we must obtain an optimal solution x_k^{**}, if it exists, for which a function $x(\theta)$ exists with the previous property. This is accomplished by a simplified Search Procedure (SP) outlined below.

Define the directions:

$$(3.8) \qquad s_k^* = \begin{cases} s_{1k} & \text{for Case 1.i,} \\ s_{1k}^{**} & \text{for Case 1.ii,} \end{cases} \qquad s_k = \begin{cases} s_{2k} & \text{for Case 1.i,} \\ s_{2k}^{**} & \text{for Case 1.ii,} \end{cases}$$

where $(s'_{1k}, t'_1, t'_2, s'_{2k})' = -B_k^{-1}a$, $(s_{1k}^{**'},\ t'_3, t'_4, s_{2k}^{**'})' = B^{-1}e_{n+\mu}$, a' is the μth row of $(C'_{12}, C'_{22}, A^{c'}_{12}, A^{c'}_{22})$, $e_{n+\mu}$ is the $(2n + 1)$ order $(n + \mu)$th unit vector, s_{1k}, s_{1k}^{**} are p-vectors and s_{2k}, s_{2k}^{**} are $(n + 1 - q)$-vectors.

By Lemma 3, $x(\lambda) = x_k^* + \lambda s_k^*$ is optimal for all λ for which it remains feasible. Furthermore, by construction of s_k^*, all constraints active at x_k^* are satisfied by $x(\lambda)$; $\lambda \leqq 0$. Thus, the smallest λ for which $x(\lambda)$ is feasible is determined by satisfying:

(a) the inactive constraints in A^c, $y_{1k}^c(\theta_l) + \lambda s_k \leqq 0$, which gives

$$(3.9.1) \qquad \lambda_1 = -(y_{1k}^c)_{\sigma_1}/(s_k)_{\sigma_1} = \max_j \{-(y_{1k}^c)_j/(s_k)_j | (s_k)_j > 0 \text{ and } j \in I_{B_k}^*\}.$$

If $s_k \leqq 0$ for all $j \in I_{B_k}^*$, then $\lambda_1 = -\infty$.

(b) the inactive nonnegativity restrictions, $(x_k^* + \lambda s_k^*)_j \geqq 0$ for all $j \in J_{B_k}$ which gives

$$(3.9.2) \qquad \lambda_2 = -(x_k^*)_{\sigma_2}/(s_k)_{\sigma_2} = \max_j \{-(x_k^*)_j/(s_k^*)_j | (s_k^*)_j > 0 \text{ and } j \in J_{B_k}\}.$$

If $(s_k^*)_j \leqq 0$ for all $j \in J_{B_k}$, then $\lambda_2 = -\infty$.

(c) the (inactive) constraints A^s, $A^s(x_k^* + \lambda s_k^*) \leqq b_s + \theta_l f^s$, which gives

$$(3.9.3) \qquad \lambda_3 = -(y_k^s)_{\sigma_3}/(q)_{\sigma_3} = \max_j \{-(y_k^s)_j/(q)_j | (q)_j < 0 \text{ and } j \in I_k^s\},$$

where $q = A^s s_k^*$. If $(q)_j \geqq 0$ for all $j \in I_k^s$, then $\lambda_3 = -\infty$. We let

$$(3.10) \qquad \lambda_* = \max\{\lambda_1, \lambda_2, \lambda_3\} \quad \text{and} \quad \sigma = \sigma_i \quad \text{if } \lambda_* = \lambda_i; \quad i = 1, 2, 3.$$

If $\lambda_* = -\infty$, then QP has no optimal solution (Theorem 1). If a feasible solution to QP exists, then for $\theta < \theta_l \geqq 0$, PQP (and QP) has an unbounded solution. (Remark 5.2). If $-\infty < \lambda_* \leqq 0$, then $x_k^{**} = x(\lambda_*) = x_k^* + \lambda_* s_k^*$ is an optimal solution to PQP for $\theta = \theta_l$, which can be used to continue the parametric procedure. This completes SP.

The transition from x_k^* to x_k^{**}, accomplished by a pair of pivots explained below, causes the σth constraint to become active instead of the μth.

In terms of variable exchanges, in Case 1.i, $(v_{1k})_\mu$ and $(x_{1k})_\mu$ could not be exchanged due to a zero pivot element. Now we can replace

$$(3.11.1) \qquad (v_{1k})_\mu \text{ by } (u_{2k}^c)_\sigma \quad \text{and} \quad (y_{1k}^c)_\sigma \text{ by } (x_{2k})_\mu \quad \text{if } \sigma = \sigma_1,$$

$$(3.11.2) \qquad (v_{1k})_\mu \text{ by } (v_{2k})_\sigma \quad \text{and} \quad (x_{1k})_\sigma \text{ by } (x_{2k})_\mu \quad \text{if } \sigma = \sigma_2.$$

For $\sigma = \sigma_3$, we must first bring the σth constraint from A^s into A^c in exchange for an inactive constraint, say the τth, in A^c. This is accomplished in a manner similar to Case 2.iii.a. Then, we replace

$$(3.11.3) \qquad (v_{1k})_\mu \text{ by } (u_k^s)_\sigma \quad \text{and} \quad (y_{1k}^c)_\tau \text{ by } (x_{2k})_\mu.$$

In Case 2.ii, $(u^c_{1k})_\mu$ and $(y^c_{2k})_\mu$ could not be exchanged due to a zero pivot. Now we can replace

(3.12.1) $(u^c_{1k})_\mu$ by $(u^c_{2k})_\sigma$ and $(y^c_{1k})_\sigma$ by $(y^c_{2k})_\mu$ if $\sigma = \sigma_1$,

(3.12.2) $(u^c_{1k})_\mu$ by $(v_{2k})_\sigma$ and $(x_{1k})_\sigma$ by $(y^c_{2k})_\mu$ if $\sigma = \sigma_2$.

For $\sigma = \sigma_3$, we apply the procedure preceding (3.11.3). Then we replace

(3.12.3) $(u^c_{1k})_\mu$ by $(u^s_k)_\sigma$ and $(y^c_{1k})_\tau$ by $(x_{2k})_\mu$.

Case 2. Here a zero pivot implies (Lemma 2) that the ρth constraint, which is to become active for $\theta \leqq \theta_l$, is linearly dependent on the constraints already active at $x^*_k = x_k(\theta_l)$. The Constraint Replacement Procedure (CRP) outlined below, identifies an active constraint, say the μth one, which if replaced by the ρth constraint, insures that the so altered set of active constraints is linearly independent and $x_{k+1}(\theta) = x_k(\theta_l)$ solves RAP for some interval $\theta_q \leqq \theta \leqq \theta_l, \theta_q < \theta_l$.

Since (by Lemma 2) the ρth constraint

(3.13) $$a'x = a'_1x_1 + a'_2x_2 \leqq \beta$$

is linearly dependent on the active constraints at x^*_k, there exists an $(n-p)$-vector z_1 and a q-vector z_2 such that

(3.14.1) $$\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 0 & A^{c'}_{11} \\ -I & A^{c'}_{12} \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}.$$

Using (3.3), it is easily verified that

(3.14.2) $$z' = (0, z'_1, z'_2, 0) = (a'_1, a'_2, 0, 0)B_k^{-1}.$$

As shown in Lemma 4, if

(3.15) $$(z)_j \leqq 0 \quad \text{for all } j \in I_{B_k},$$

then PQP has no feasible solution for $\theta < \theta_l$. However, if z has at least one positive component,

(3.16.1) $$\nu^1_0 = (v_{1k})_{\mu_1}/(z_1)_{\mu_1} = \min_j \{(v_{1k})_j/(z_1)_j | (z_1)_j > 0, \text{and } j \in J^*_{Bk}\},$$

(3.16.2) $$\nu^2_0 = (u^c_{1k})_{\mu_2}/(z_1)_{\mu_2} = \min_j \{(u^c_{1k})_j/(z_2)_j | (z_2)_j > 0, \text{and } j \in I_{B_k}\},$$

(3.16.3) $$\nu_0 = \min\{\nu^1_0, \nu^2_0\} \quad \text{and } \mu = \mu_i \quad \text{if } \nu_0 = \nu^i_0, \quad i = 1, 2,$$

identify the μth constraint to be made inactive and terminate CRP.

In Case 2.i the ρth nonnegativity restriction, represented by letting $a_1 = 0$, $a_2 = e_\rho$ in (3.13) and (3.14.2), is to become active. However, $(x_{1k})_\rho$ and $(v_{2k})_\rho$ could not be exchanged due to a zero pivot. N , provided (3.15) is not satisfied, we can replace (Theorem 2)

(3.17.1) $(x_{1k})_\rho$ by $(v_{2k})_\mu$ and $(x_{1k})_\mu$ by $(v_{2k})_\rho$ if $\mu = \mu_1$,

(3.17.2) $(x_{1k})_\rho$ by $(y^c_{2k})_\mu$ and $(u^c_{1k})_\mu$ by $(v_{2k})_\rho$ if $\mu = \mu_2$.

In Case 2.ii, the ρth ordinary constraint, represented by letting $a_1 = a_{1\rho}$ and $a_2 = a_{2\rho}$ in (3.13) and (3.14.2), is to become active. However, $(u^c_{1k})_\rho$ and

$(y_{2k}^c)_\rho$ could not be exchanged due to a zero pivot. Now, provided (3.15) is not satisfied, we can replace (Theorem 2)

$$(3.18.1) \qquad (u_{1k}^c)_\rho \text{ by } (v_{2k})_\mu \text{ and } (x_{1k})_\mu \text{ by } (y_{2k}^c)_\rho \quad \text{if } \mu = \mu_1,$$

$$(3.18.2) \qquad (u_{1k}^c)_\rho \text{ by } (y_{2k}^c)_\mu \text{ and } (u_{1k}^c)_\mu \text{ by } (y_{2k}^c)_\rho \quad \text{if } \mu = \mu_2.$$

Thus the appropriate pivot or pair of pivot steps, specified by Case 1 or 2 above, is executed by updating the current inverse (see § 4) which is then used to continue the algorithm from (3.4).

A summary of the algorithm follows.

Step 1. Use (3.1) to construct vectors d and f such that for $\theta = \theta_0$ the point $x_0(\theta_0) = 0$ solves PQP. Construct B_0 by (3.3), define the sets $J_{B_0}, J_{B_0}^*, I_{B_0}, I_{B_0}^*, I_0^s$ and obtain B_0^{-1}. Let $\theta_k = \theta_0, B_k = B_0$.

Step 2. If $\theta_k = 0$, then B_k is optimal and $x_k(0)$ solves QP. Terminate. If $\theta_k > 0$, compute vectors $p_j, j = 1, \cdots, 10$, using (3.4). Apply (3.5)–(3.7) to obtain $\theta_l \leqq \theta_k$. If $\theta_l = 0$ then $x_k(0)$ solves QP. Terminate. If $\theta_l > 0$, go to the appropriate case of Step 3.

Step 3.

$\theta_l = \theta_{l1}^*$ Case 1.i	$\theta_l = \theta_{l2}^*$ Case 1.ii	$\theta_l = \theta_{l1}^{**}$ Case 2.i	$\theta_l = \theta_{l2}^{**}$ Case 2.ii	$\theta_l = \theta_{l3}^{**}$ Case 2.iii
				Exchange the ρth constraint in A^s and the τth constraint in A^c by a special pivot step (see (4.1.2)) and update B_k^{-1}. (Case 2.iii.a)
Try to exchange:				
$(v_{1k})_\mu$ and $(x_{2k})_\mu$	$(u_{1k}^c)_\mu$ and $(y_{2k}^c)_\mu$	$(x_{1k})_\rho$ and $(v_{2k})_\rho$	$(y_{1k}^c)_\rho$ and $(u_{2k}^c)_\rho$	$(y_{1k}^c)_\tau$ and $(u_{2k}^c)_\rho$
If pivot < 0, perform one pivot step to obtain B_l^{-1}, go to Step 6.				
If pivot $= 0$, go to Step 4. Upon return perform the pair of pivots:		If pivot $= 0$, go to Step 5. Upon return perform the pair of pivots:		
(3.11.1) or (3.11.2) or after exchanging a constraint (see (2.iii.a)) perform (3.11.3) to obtain B_l^{-1}.	(3.12.1) or (3.12.2) or after exchanging a constraint (see (2.iii.a)) perform (3.12.3) to obtain B_l^{-1}.	(3.17.1) or (3.17.2) to obtain B_l^{-1}.	(3.18.1) or (3.18.2) to obtain B_l^{-1}.	(3.18.1) or (3.18.2) to obtain B_l^{-1}.
Go to Step 6				

Step 4 (SP). Compute s_k^*, s_k by (3.8) for the current case (either 1.i or 1.ii) of Step 3, and obtain λ_*, σ by (3.9), (3.10). If $\lambda_* = -\infty$, no optimal solution to QP exists. (If QP has a nonempty feasible domain, then $\lambda_* = -\infty$ implies that QP has an unbounded solution). Terminate. If $-\infty < \lambda_* \leqq 0$, return.

Step 5 (CRP). If entering from Case 2.i, let $a_1 = 0, a_2 = e_\rho$. If entering from Cases 2.ii or 2.iii, let $a_1 = a_{1\rho}, a_2 = a_{2\rho}$. Compute $z' = (a_1', a_2', 0, 0)B_k^{-1}$. If $(z)_j \leqq 0$

for all $j \in I_{B_k}$, then QP has no feasible solution. Terminate. If not, compute μ by (3.16), return.

Step 6. Let $\theta_{k+1} = \theta_l$, $B_{k+1}^{-1} = B_l^{-1}$ and update the basic index sets to obtain $J_{B_{k+1}}$, $J_{B_{k+1}}^*$, $I_{B_{k+1}}$, $I_{B_{k+1}}^*$, I_{k+1}^s. Let $k + 1 \rightarrow k$ and start the next parametric cycle at Step 2.

Remark 3.1. Let exactly one of the AP variables be zero for $\theta = \theta_l$ in (3.7). A basic solution, valid for $\theta_{l+1} \leqq \theta \leqq \theta_l$, $\theta_{l+1} < \theta_l$, is obtained after one pivot operation, provided the pivot element is negative. If the pivot is zero, a basic solution is obtained after a pair of pivots provided that SP or CRP determine a unique constraint.

The degenerate case where several basic variables of the current AP vanish for $\theta = \theta_l$, may be reduced to the above situation by an appropriate choice of the components of d and f. For instance, if several components of y_{1k}^c and/or y_k^s are zero, we can increase all but one of the corresponding components of f by a finite amount and obtain a case with exactly one vanishing component of y_{1k}^c or y_k^s. Similarly, if several components of v_{1k} are zero, we can decrease all but one of the corresponding components of d. If several, say n_1, components of x_{1k} vanish, we can choose one of them, perform the pivot operation dictated by the algorithm, and then decrease the component of d corresponding to the new basic variable which then becomes positive. This results in a case with $(n_1 - 1)$ vanishing basic variables. Finally, if several components of u_{1k} are zero, we can choose one of them, perform the necessary pivot step and then increase the component of f corresponding to the new basic variable. Clearly, any ambiguity caused by combinations of the above cases or by the failure of SP or CRP to determine a unique constraint can be resolved in a similar manner.

Without loss of generality we may therefore assume that Step 3 gives a basic solution, the x part of which is an optimal solution to PQP for some interval $\theta_{l+1} \leqq \theta \leqq \theta_l$ with $\theta_{l+1} < \theta_l$.

4. Computational aspects. For problems with $m \gg n$, for which this method is primarily intended, the $(2n + 1)$-order working basis is decisively smaller than the $(m + n)$-order basis commonly used by other algorithms. This reduction is enjoyed at the expense of CRP and its relative merit depends on the particular structure of the constraint matrix. It is clear however, that for problems with $m \leqq n$ the definition of AP and RAP coincide, the working basis is of order $(m + n)$, and Case 2.iii of the algorithm does not arise.

The main computational tools required for efficient programming of the proposed algorithm are similar to those used by most commercially available large linear programming codes. (See e.g., [24].) The following remarks on the nature of these operations will clarify the necessary revisions to the components of these codes.

We consider the usual elementary transformation:

$$B_{k+1}^{-1} = E_{k+1} B_k^{-1}, \tag{4.1.1}$$

where E_{k+1} is an "elementary matrix". Let (s, p) be the designated pivot position and g_p the "pivot column". A variable exchange operation consists of constructing an "elementary column matrix" E_{k+1}^c as an identity matrix with its sth column replaced by

$$\eta^c_{k+1} = \left(-\frac{(h)_1}{(h)_s}, \cdots, -\frac{(h)_{s-1}}{(h)_s}, \frac{1}{(h)_s}, -\frac{(h)_{s+1}}{(h)_s}, \cdots, -\frac{(h)_{2n+1}}{(h)_s}\right),$$

where $h = B_k^{-1} g_p$ is the updated g_p, and later applying (4.1.1) with $E_{k+1} = E^c_{k+1}$.

The reclassification of a stand-by constraint as current is accomplished by first constructing an elementary row matrix E^r_{k+1}, which is an identity matrix with its σth row replaced by

$$\eta^{r'}_{k+1} = (g'_\tau, e'_\tau), \tag{4.1.2}$$

where the p-row vector $g'_\tau = (a^{c'}_\tau - a^{s'}_\rho)$, e'_τ is the $(2n - p + 1)$-unit row vector with 1 in the τth position and $a^{c'}_\tau$, $a^{s'}_\rho$ are the τth and ρth rows of A^c_{21} and A^s_1 respectively. Then, (4.1.1) is applied with $E_{k+1} = E^r_{k+1}$. This transformation on B_k^{-1} has the effect of replacing $a^{c'}_\tau$ by $a^{s'}_\rho$ in B_k.

In the course of the algorithm, B_k^{-1} is used to update a column or row vector (e.g., (3.4), (3.8), (3.14.2)). These operations may obviously be performed using the explicit form of B_k^{-1} by direct multiplication. However, when A is sparse, it is preferable to keep B_k^{-1} in the "product form": $B_k^{-1} = E_k E_{k-1} \cdots E_1$ where E_j may be a column or row elementary matrix. Updating operations are then handled by special formulas:

The transformation of a column vector g by a single E_j is given by

$$\begin{aligned}
(E_j g)_i &= (g)_i + (\eta^c_j)_i (g)_s && \text{for } i \neq s; \quad E_j = E^c_j, \\
&= (\eta^c_j)_i (g)_s && \text{for } i = s; \quad E_j = E^c_j, \\
&= (g)_i && \text{for } i \neq \tau; \quad E_j = E^r_j, \\
&= \sum_{\mu=1}^{2n+1} (\eta^{r'}_j)_\mu (g)_\mu && \text{for } i = \tau; \quad E_j = E^r_j,
\end{aligned} \tag{4.2}$$

while the transformation of a row vector g' by a single E_j is given by

$$\begin{aligned}
(g' E_j)_i &= (g)_i && \text{for } i \neq s; \quad E_j = E^c_j, \\
&= \sum_{\mu=1}^{2n+1} (g)_\mu (\eta^c_j)_\mu && \text{for } i = s; \quad E_j = E^c_j, \\
&= (g)_i + (\eta^c_j)_i (g)_\tau && \text{for } i = \tau; \quad E_j = E^r_j, \\
&= (\eta^c_j)_i (g)_\tau && \text{for } i = \tau; \quad E_j = E^r_j.
\end{aligned} \tag{4.3}$$

The updating of g requires the repeated use of (4.2) to form

$$h = (E^c_k(E^c_{k-1}(\cdots (E^c_{l+1}(E^r_l(E^c_{l-1}(\cdots (E^c_1 g) \cdots)))) \cdots))), \tag{4.4}$$

computed in the order the E_j were generated. The operation of updating a row vector g' by the current inverse is

$$h' = (\ldots((((\ldots((g' E^c_k) E^c_{k-1}) \ldots) E^c_{l+1}) E^r_l) E_{l-1}) \ldots) E^c_1, \tag{4.5}$$

computed in the reverse order from that in which they were generated.

The transformations (4.2)–(4.5) and reinversion algorithms are integral parts of most large linear programming codes. Slight modifications may be required to handle the case $E_j = E_j^r$ for some of these codes.

5. Validity of the algorithm. In this section, discussions will be based on a simplified RAP basis:

$$B = \begin{array}{c} \begin{array}{cc} x, & w_1 \end{array} \\ \begin{pmatrix} C & A_1' \\ A_1 & 0 \end{pmatrix} \end{array} \quad \text{with } A_1 = \begin{pmatrix} A_{11}^c & A_{12}^c \\ 0 & -I \end{pmatrix}, \tag{5.1}$$

where A_1 denotes the matrix of active constraint rows and $w_1 = (u_1^c, v_1)$. The basis B may be identified within dashed lines in the partitioned RAP matrix

$$\begin{array}{ccccccccc} x_1 & x_2 & u_1^c & v_1 & u_2^c & v_2 & y_2^c & y_2^n & y_1^c \\ \hline C_{11} & C_{12} & A_{11}^{c\prime} & 0 & A_{21}^{c\prime} & -I & 0 & 0 & 0 \\ C_{21} & C_{22} & A_{12}^{c\prime} & -I & A_{22}^{c\prime} & 0 & 0 & 0 & 0 \\ A_{11}^c & A_{12}^c & 0 & 0 & 0 & 0 & I & 0 & 0 \\ 0 & -I & 0 & 0 & 0 & 0 & 0 & I & 0 \\ A_{21}^c & A_{22}^c & 0 & 0 & 0 & 0 & 0 & 0 & I \end{array} \tag{5.2}$$

where the nonnegativity restrictions for x_2 and their slack variables y_2^n have been shown explicitly. For a basic solution $x_2 = 0$, $y_2^n = 0$, the columns of x_2 can be removed, the slacks y_1^c and the inactive rows A_{21}^c may be appended with no effect. Thus, B corresponds to B_k in (3.3). In the following paragraphs the term "constraint" will refer to nonnegativity restrictions and ordinary constraints alike. We will also let $b_1 = (b_1^c, 0)$ and $f_1 = (f_1^c, 0)$.

LEMMA 1. *Let $Q(x(\theta)) = (c^* + \theta d^*)'x - \frac{1}{2}x'Cx$. Suppose B is nonsingular and*

$$x(\theta) = d_1 + \theta d_2, \qquad w_1(\theta) = d_3 + \theta d_4 \tag{5.3.1}$$

is the solution of

$$\begin{aligned} Cx + A_1'w_1 &= c^* + \theta d^*, \\ A_1 x &= b_1^* + \theta f_1^*. \end{aligned} \tag{5.3.2}$$

Then $\partial Q(x(\theta))/\partial\theta|_{\theta=\theta_0} = d^{*\prime}x_0 + f_1^{*\prime}w_{1,0}$ and $\partial^2 Q(x(\theta))/\partial\theta^2|_{\theta=\theta_0} = d^{*\prime}d_2 + f_1^{*\prime}d_4$, where $x_0 = d_1 + \theta_0 d_2$ and $w_{1,0} = d_3 + \theta_0 d_4$. Furthermore, if C is p.s.d., $d^{*\prime}d_2 + f_1^{*\prime}d_4 \leqq 0$.

Proof. Substituting (5.3.1) into (5.3.2) gives

$$\begin{aligned} Cd_1 + \theta Cd_2 + A_1'd_3 + \theta A_1'd_4 &= c^* + \theta d^*, \\ A_1 d_1 + \theta A_1 d_2 &= b_1^* + \theta f_1^* \end{aligned}$$

for all θ. For $\theta = 0$, $Cd_1 + A_1'd_3 = c^*$ and $A_1 d_1 = b_1^*$. Hence

$$\theta(Cd_2 + A_1'd_4) = d^*, \qquad A_1 d_2 = f_1^*. \tag{5.3.3}$$

Now, using $x(\theta)$ from (5.3.1),

$$Q(x(\theta)) = c^*d_1 - \tfrac{1}{2}d_1'Cd_1 + (c'd_2 + d^{*\prime}d_1 - d_1'Cd_2)\theta + (d^{*\prime}d_2 - \tfrac{1}{2}d_2'Cd_2)\theta^2.$$

Differentiating with respect to θ and using (5.3.1)–(5.3.3),

$$\begin{aligned}\partial Q(x(\theta))/\partial\theta|_{\theta=\theta_0} &= c^{*\prime}d_2 + d^{*\prime}d_1 + d_1'Cd_2 + (2d^{*\prime}d_2 - d_2'Cd_2)\theta_0\\ &= d^{*\prime}x_0 + f_1^{*\prime}w_{1,0}.\end{aligned}$$

Next, differentiating $d^{*\prime}x(\theta) + f_1^{*\prime}w_1(\theta)$ and using (5.3.1) gives

$$\partial^2 Q(x(\theta))/\partial\theta^2|_{\theta=\theta_0} = d^{*\prime}d_2 + f_1^{*\prime}d_4.$$

Since C is p.s.d., $Q(x(\theta))$ is concave and therefore $d^{*\prime}d_2 + f_1^{*\prime}d_4 \leqq 0$.

Remark 5.1. In order to put the above result in a form useful for proving Lemma 2 we consider the two cases of pivoting below.

Case 1. For $\theta = \theta_l$, the μth constraint is to become inactive, i.e., $(w_1)_\mu$ and $(y_2)_\mu$ are to be exchanged in the (RAP) system:

$$Cx + A_1'w_1 + (w_1)_\mu a_\mu = c + \theta_l d, \tag{5.4.1}$$

$$A_1x = b_1 + \theta_l f_1, \tag{5.4.2}$$

$$a_\mu' x + (y_2)_\mu = (b_1)_\mu + \theta_l(f_1)_\mu, \tag{5.4.3}$$

where presently $(y_2)_\mu = 0$ is nonbasic and $(w_1)_\mu = 0$ is a basic variable.[2] The current basic solution of (5.4.1)–(5.4.3) is given by

$$\begin{pmatrix} x \\ w_1 \\ (w_1)_\mu \end{pmatrix} = B_I^{-1}\begin{pmatrix} c + \theta_l d \\ b_1 + \theta_l f_1 \\ (b_1)_\mu + \theta_l(f_1)_\mu \end{pmatrix} - B_I^{-1}e_{n+\mu}(y_2)_\mu = g_1 - h_1(y_2)_\mu, \tag{5.4.4}$$

where the vectors g_1, h_1 have been introduced for convenience, $(y_2)_\mu = 0$ and

$$B_I = \begin{pmatrix} C & A_1' & a_\mu \\ A_1 & 0 & 0 \\ a_\mu' & 0 & 0 \end{pmatrix}.$$

In particular,

$$(w_1)_\mu = (g_1)_\mu - (h_1)_\mu(y_2)_\mu, \tag{5.4.5}$$

where $(h_1)_\mu$, the μth component of h_1, is the "pivot element" in Case 1. Now consider (5.4.1)–(5.4.2) and $a_\mu' x = (b_1)_\mu + \theta_l(f_1)_\mu - (y_2)_\mu$. If we let $c^* = c + \theta_l d$, $d^* = 0$, $b_1^* = (b_1 + \theta_l f_1)$, $(b_1)_\mu + \theta_l(f_1)_\mu)$, $f_1^* = e_\mu$ and $\theta = (y_2)_\mu$, (5.4.5) and Lemma 1 give

$$\partial^2 Q(x(\theta))/\partial\theta^2 = (h_1)_\mu \leqq 0. \tag{5.4.6}$$

Case 2. For $\theta = \theta_l$, the ρth constraint is to become active, i.e., $(y_1)_\rho$ and $(w_2)_\rho$ are to be exchanged in the (RAP) system:

$$Cx + A_1'w_1 + (w_2)_\rho a_\rho = c + \theta_l d, \tag{5.4.7}$$

$$A_1x = b_1 + \theta_l f_1, \tag{5.4.8}$$

$$a_\rho' x + (y_1)_\rho = (b)_\rho + \theta_l(f)_\rho, \tag{5.4.9}$$

where presently $(w_2)_\rho = 0$ is nonbasic and $(y_1)_\rho = 0$ is a basic variable. The current

[2] This system representation is valid for both Cases 1.i and 1.ii. In Case 1.i the exchange of $(v_1)_\mu$ by $(x_2)_\mu$ in the basis B_k (3.3) is clearly equivalent to the exchange of $(v_1)_\mu$ by $(y_2'')_\mu$ in the basis B (5.2).

basic solution of (5.4.7)–(5.4.9) is given by:

$$(5.4.10)\quad \begin{pmatrix} x \\ w_1 \\ (y_1)_\rho \end{pmatrix} = B_{II}^{-1} \begin{pmatrix} c + \theta_l d \\ b_1 + \theta_l f_1 \\ (b_2)_\rho + \theta_l (f_2)_\rho \end{pmatrix} - B_{II}^{-1} \begin{pmatrix} a_\rho \\ 0 \\ 0 \end{pmatrix} (w_2)_\rho = g_2 - h_2(w)_\rho,$$

where the vectors g_2, h_2 have been introduced for convenience, $(w_2)_\rho = 0$ and B_{II} is obtained by replacing the last column of B_I by the column vector $(0, 0, 1)$. In particular,

$$(5.4.11)\qquad (y_1)_\rho = (g_2)_\rho - (h_2)_\rho (w_2)_\rho,$$

where $(h_2)_\rho$, the ρth component of h_2, is the pivot element in Case 2.

Lemma 2. *Consider Cases* 1 *and* 2 *of pivotal operations and the corresponding systems* (5.4.1)–(5.4.3) *and* (5.4.7)–(5.4.9). *Then*

A. *The pivot element in both Cases* 1 *and* 2 *is nonpositive.*

B. *Let* s_1 *and* s_2 *denote the first* n *components of* g_1 *and* h_1 *respectively as defined by* (5.4.4). *If, in Case* 1 *the pivot element is zero, then,* $Q(s_1 - \lambda s_2)$ *is either constant or a linear function of* λ.

If, in Case 2, *the pivot element is zero, then, the constraint* $a'_\rho x = (b_2)_\rho + \theta_l (f_2)_\rho$ *is linearly dependent on the constraints* (5.4.8).

Proof. For Case 1 the assertions follow immediately from (5.4.6). For Case 2, from the definition of B_{II} and (5.4.10)–(5.4.11), it follows that

$$(5.5.1)\qquad (h_2)_\rho = -(a'_\rho, 0) \begin{pmatrix} C & A'_1 \\ A_1 & 0 \end{pmatrix}^{-1} \begin{pmatrix} a_\rho \\ 0 \end{pmatrix}.$$

Let (v^*, z^*) be the unique solution of the system

$$\begin{matrix} (5.5.2) \\ (5.5.3) \end{matrix}\qquad \begin{pmatrix} C & A'_1 \\ A_1 & 0 \end{pmatrix} \begin{pmatrix} v \\ z \end{pmatrix} = \begin{pmatrix} a_\rho \\ 0 \end{pmatrix}.$$

Substituting (v^*, z^*) into (5.5.1), premultiplying (5.5.2) by $v^{*\prime}$ and using (5.5.3) gives:

$$(h_2)_\rho = -(a'_\rho, 0) \begin{pmatrix} v^* \\ z^* \end{pmatrix} = -v^{*\prime} C v^*.$$

Since C is p.s.d. this implies $(h_2)_\rho \leqq 0$. If $v^{*\prime} C v^* = 0$, then $C v^* = 0$ (see [22, Lemma 1]) which, by (5.5.2) gives $A'_1 z^* = a_\rho$.

Lemma 3. *Suppose the pivot element in Case* 1 *is zero, i.e.,* $(h_1)_\mu = 0$, *for* $\theta = \theta_l$. *Let* $(x_k^*, w_{1k}^*, (w_{1k}^*)_\mu)$ *be the solution of RAP for* $\theta = \theta_l$ *and* $Q(x_k^*)$ *the corresponding value of the objective function of PQP. Furthermore, let* s_k^* *be the vector defined by SP. Then,*

A. $Q(x_k^* + \lambda s_k^*) = Q(x_k^*)$ *for all* $\lambda \in E^1$.

B. $w_k^*(\lambda) = w_k^*$ *for all* $\lambda \in E^1$.

Proof.

A. Reviewing the definition (3.8.1) of s_k^* and comparing the RAP bases B_k (3.3) and B_I (5.4.4.) it can easily be seen that $(x_k^* + \lambda s_k^*)$ is equivalent to $(s_1 - \lambda s_2)$. Hence, by Lemma 2, $Q(x_k^* + \lambda s_k^*)$ is either constant or a linear function of λ.

$A_1(x_k^* + \lambda s_k^*) = b_1 + \theta_l f_1$ for all $\lambda \in E^1$ implies $A_1 s_k^* = 0$. Hence, using RAP (in particular (5.4.1)),

$$(c + \theta_l d - Cx_k^*)' s_k^* = (A'_1 w_{1k})' s_k^* + (w_{1k})_\mu a'_\mu s_k^* = 0$$

since $(w_{1k})_\mu = 0$. This proves the first assertion.

B. Since $Q(x_k^* + \lambda s_k^*) = Q(x_k^*)$, we have $s_k^{*\prime} C s_k^* = 0$ which implies $Cs_k^* = 0$. Hence $C(x_k^* + \lambda s_k^*) = Cx_k^*$ for all λ, and since the columns of A'_1 are linearly independent, it follows from RAP that w_{1k}^* is independent of λ.

THEOREM 1. *Let $x_k^* = x_k(\theta_l)$ be an optimal solution to PQP for $\theta_l > 0$. Suppose, for $\theta = \theta_l$, Case* 1 *applies and the pivot element is zero.*

If SP yields $\lambda_* = -\infty$, then QP has no optimal solution.

Proof. The assertion is clear, if QP has no feasible solution. Let x_0 be any feasible solution to QP. By assumption, $A(x_k^* + \lambda s_k^*) \leqq b + \theta_l f$ and $x_k^* + \lambda s_k^* \geqq 0$ hold for all $\lambda \leqq 0$. Therefore, $As_k^* \geqq 0$ and $s_k^* \leqq 0$. Thus, $Ax_0 \leqq b$ and $x_0 \geqq 0$ imply

$$A(x_0 + \lambda s_k^*) \leqq b, \qquad x_0 + \lambda s_k^* \geqq 0 \qquad \text{for all } \lambda \leqq 0. \tag{5.6.1}$$

From the definition of s_k^* and from Lemma 1 it follows that, for $\theta = \theta_l$,

$$\partial Q(x_k^* + \lambda s_k^*)/\partial\lambda|_{\lambda=0} = (c + \theta_l d - Cx_k^*)' s_k^* = (w_1(\theta_l))_\mu = 0.$$

It is clear that in Case 1, $(w_1)_\mu < 0$ for $\theta < \theta_l$. Hence,

$$(c - Cx_k(0))' s_k^* = (w_1(0))_\mu < 0. \tag{5.6.2}$$

By Lemma 3, for $\theta = \theta_l$, $\partial^2 Q(x_k^* + \lambda s_k^*)/\partial\lambda^2 = 0$, thus $s^{*\prime} C s^* = 0$. Hence, for $\theta = 0$,

$$\begin{aligned} Q(x_0 + \lambda s_k^*) &= c'x_0 - \tfrac{1}{2}x'_0 C x_0 + (c - Cx_0)' s_k^* \lambda \\ &= c'x_0 - \tfrac{1}{2}x'_0 C x_0 + c' s_k^* \lambda \end{aligned} \tag{5.6.3}$$

since $s_k^{*\prime} C s_k^* = 0$ implies $Cs_k^* = 0$ (see [22, Lemma 1]). The latter with (5.6.2) gives $c's_k^* < 0$. But then (5.6.1) and (5.6.3) show that the objective function of QP can be made arbitrarily large over the feasible domain.

THEOREM 2. *Let the designated pivot element be zero for some $\theta = \theta_l$. Suppose that in Case* 1, *SP terminates with $-\infty < \lambda < 0$ corresponding to a constraint to become active or in Case* 2, *CRP yields a currently active constraint to become inactive.*

Then, the exchanges prescribed by SP or CRP can be accomplished by a pair of pivot steps. The new basic variables are nonnegative for $\theta < \theta_l$.

Proof.

Case 1. Consider the current form of RAP:

$$Cx + A'_1 w_1 + a_\mu (w_1)_\mu + a_\rho (w_2)_\rho = c + \theta_l d, \tag{5.7.1}$$

$$A_1 x = b_1 + \theta_l f_1, \tag{5.7.2}$$

$$a_\mu x + (y_2)_\mu = \alpha_1 + \theta_l \alpha_2, \tag{5.7.3}$$

$$a_\rho x + (y_1)_\rho = \beta_1 + \theta_l \beta_2, \tag{5.7.4}$$

$$(w_1)_\mu (y_2)_\mu = 0, \qquad (w_2)_\rho (y_1)_\rho = 0, \tag{5.7.5}$$

where (5.7.3) is to become inactive and (5.7.4) is found by SP to become active.

Since the pivot is zero, $(w_1)_\mu$ and $(y_2)_\mu$ could not be exchanged. Application of SP indicated that $(y_1)_\rho$ should leave and $(w_2)_\rho$ should enter the basis. We will show that it is always possible to exchange $(w_1)_\mu$ and $(w_2)_\rho$ *and then* $(y_1)_\rho$ and $(y_2)_\mu$.

Consider the current nonsingular basis

$$B_I^1 = \begin{array}{c} \begin{array}{cccc} x, & w, & (w_1)_\mu, & (y_1)_\rho \end{array} \\ \left(\begin{array}{cccc} C & A_1' & a_\mu & 0 \\ A_1 & 0 & 0 & 0 \\ \hdashline a_\mu' & 0 & 0 & 0 \\ a_\rho' & 0 & 0 & 1 \end{array}\right) \end{array}$$

and the matrix B_I^2 obtained by replacing $(w_1)_\mu$ by $(w_2)_\rho$, i.e., by replacing column $(a_\mu, 0, 0, 0)$ in B_I^1 by column $(a_\rho, 0, 0, 0)$. We recall that according to SP

$$\begin{pmatrix} C & A_1' \\ A_1 & 0 \end{pmatrix} \begin{pmatrix} x_k - \lambda s_k^* \\ w_k - \lambda s_k^{**} \end{pmatrix} = \begin{pmatrix} c + \theta_l d \\ b_1 + \theta_l f_1 \end{pmatrix} \qquad \text{for all } \lambda < 0,$$

where $s_k' = (s_k^{*\prime}, s_k^{**\prime})$ consists of the appropriate components of h_1 in (5.4.4). Obviously $Bs_k = 0$. Since by assumption, SP terminated finitely with $a_\rho' x$ as the first constraint encountered, $a_\rho' s_k^* \neq 0$. Clearly then column $(a_\rho, 0)$ cannot be a linear combination of the columns of B. The same is true for column $(a_\rho, 0, 0, 0)$ and the rest of the columns of B_I^2 which are linearly independent since they also belong to B_I^1. Therefore, B_I^2 is nonsingular, the pivot step exchanging $(w_1)_\mu$ and $(w_2)_\rho$ is valid and $(w_2)_\rho$ enters the basis at zero level.

To show that the second pivot step is possible, consider the matrix B_I^3 obtained from B_I^2 by replacing $(y_1)_\rho$ by $(y_2)_\mu$, i.e., by replacing column $(0, 0, 0, 1)$ in B_I^2 by $(0, 0, 1, 0)$. Since column $(a_\rho, 0)$ is linearly independent of the columns of B, column $(0, 0, 1, 0)$ cannot be a linear combination of the remaining ones in B_I^3. Thus, B_I^3 is nonsingular, the pivot step exchanging $(y_1)_\rho$ and $(y_2)_\mu$ is valid and $(y_2)_\mu$ enters the basis at a positive level.

Now we show that $(w_2)_\rho > 0$ for $\theta < \theta_l$. Let $(x^*(\theta), w_1^*(\theta), (w_1(\theta))_\mu, (y_1(\theta))_\rho)$ and $(x^{**}(\theta), w_1^{**}(\theta), (w_2(\theta))_\rho, (y_2(\theta))_\mu)$ denote the basic solution of (5.7) before and after the pair of pivot operations, respectively. Let s_k^* be the vector determined by SP. Then, considering (5.7) with $x = x_k^* + \lambda s_k^*$ and $\theta = \theta_l$ it can easily be shown that $A_1 s_k^* = 0$, $a_\mu' s_k^* > 0$ and $a_\rho' s_k^* < 0$. For the first basic solution, (5.7.1) post multiplied by s_k^* gives $(c + \theta d - Cx^*(\theta))' s_k^* = (A_1' w_1^*(\theta))' s_k^* + (w_1(\theta))_\mu a_\mu' s_k^*$. This implies $(c + \theta d)' s_k^* < 0$ for $\theta < \theta_l$ since $Cs_k^* = 0$ (proof of Theorem 1), and since in Case 1, $(w_1(\theta))_\mu < 0$ for $\theta < \theta_l$. Similarly, for the second basic solution

$$(c + \theta d - Cx^{**}(\theta))' s_k^* = (A_1' w_1^{**}(\theta))' s_k^* + (w_2(\theta))_\rho a_\rho' s_k^*.$$

Therefore, for $\theta < \theta_l$, $(w_2(\theta))_\rho a_\rho' s_k^* = (c + \theta d)' s_k^* < 0$. But since $a_\rho' s_k^* < 0$ we must have $(w_2(\theta))_\rho > 0$ for $\theta < \theta_l$.

Case 2. Let (5.7.4) be the constraint to become active. Since the pivot is zero, $(y_1)_\rho$ and $(w_2)_\rho$ could not be exchanged. Suppose, CRP determined (5.7.3) to become inactive. Applying arguments similar to the above it can be shown that first $(w_1)_\mu$ and $(w_2)_\rho$ and then $(y_1)_\rho$ and $(y_2)_\mu$ can be exchanged by a pair of pivot operations. Lemma 4 below assures that for $\theta < \theta_l$ the new basic variables $(w_2)_\rho$ and $(y_2)_\mu$ are nonnegative.

LEMMA 4. *Let A_1 in* (5.1) *be an* (m, n)*-matrix of full row rank and consider the equations*

$$B\begin{pmatrix} x \\ w \end{pmatrix} = \begin{pmatrix} C & A_1' \\ A_1 & 0 \end{pmatrix}\begin{pmatrix} x \\ w \end{pmatrix} = \begin{pmatrix} c \\ b + \theta f \end{pmatrix}, \tag{5.8.1}$$

where c is an n-vector, b and f are m-vectors and $\theta \geqq 0$. Let

$$B^{-1} = \begin{pmatrix} M_1 & M_2 \\ M_2' & M_4 \end{pmatrix}$$

and suppose, for $\theta_0 > 0$, (5.8.1) *has a solution $(x_0(\theta), u_0(\theta))$ for which $u_0(\theta_0) \geqq 0$, $d'x_0(\theta_0)\ \alpha + \theta_0\beta$, and $d'x_0(\theta) < \alpha + \theta\beta$ for $\theta > \theta_0$, where α, β are scalars and the n-vector d is a linear combination of the columns of A_1', i.e., $d = A_1'\,z$ for some z.*

(i) *If $M_2'd \leqq 0$, the inequalities $A_1x \leqq b + \theta f$, $d'x \leqq \alpha + \theta\beta$ are inconsistent for $\theta < \theta_0$.*

(ii) *Suppose $M_2'd$ has at least one positive component. Let, for $\theta = \theta_0$,*

$$\mu_0 = (u_0)_k/(M_2'\mathrm{d})_k = \min\{(u_0)_j/(M_2'\mathrm{d})_j > 0; \mathit{all}\ j\}.$$

Replace the k-th row of A_1 by d' and denote the new matrix by A_1^*. Furthermore, replace the k-th components of b and f by α and β, respectively and denote the new vectors by b^* and f^*. Then, the columns of A_1^* are linearly independent and the system

$$\begin{aligned} Cx + A_1^{*\prime}u &= c, \\ A_1^*x &= b^* + \theta f^* \end{aligned} \tag{5.8.2}$$

has a solution $(x_1(\theta),\ u_1(\theta))$ for which $u_1(\theta_0) \geqq 0$, $x_1(\theta_0) = x_0(\theta_0)$ and $a_k'x_1(\theta) < (b)_k + \theta(f)_k$ for $\theta < \theta_0$ where a_k' denotes the k-th row of A_1.[3]

Remark 5.2. Theorem 1 and Lemma 4 state an important property of the algorithm which allows the detection of abnormal termination conditions for QP or PQP by simple tests to be performed on RAP.

If, for some $\theta = \theta_l$, the pivot in Case 1 is zero and $\lambda_* = -\infty$, it is immediately concluded that no optimal solution to PQP, for $\theta < \theta_l$, (or to QP) exists. If, however, it is known a priori that QP has a nonempty feasible domain, then the conclusion is that PQP, for $\theta < \theta_l$, and QP have unbounded solutions over their respective feasible domains. As a further clarification, consider the case of an unbounded feasible domain and an objective function whose value is bounded from above on this domain. Then, the SP will yield a finite λ_* and the normal course of the algorithm is followed.

An empty feasible domain in PQP for $\theta < \theta_l$, and hence for QP, may be detected from RAP by a zero pivot element in Case 2 and the failure of CRP evidenced by (3.15) or by part (i) of Lemma 4.

THEOREM 3. *The algorithm outlined in* § 3 *gives one of the following alternatives after a finite number of steps*:

(a) *an $x_k^* = x_k(0)$ which is an optimal solution to QP,*

(b) *the information that QP has no optimal solution.*

[3] The proof of this Lemma is given in [18].

Proof. In view of Remark 3.1 we may assume that at each parametric step, we obtain an interval $[\theta_{k+1}, \theta_k]$ with $\theta_{k+1} < \theta_k$, such that $x_k(\theta) = (x_{1k}, 0)$ with x_{1k} given by (3.4.2) is the optimal solution of PQP for $\theta \in [\theta_{k+1}, \theta_k]$. The representation of $x_k(\theta)$ as function of θ is uniquely determined by the set of active constraints. Therefore, there is only a finite number of different functions $x_j(\theta)$ which represent the optimal solution to PQP for some interval $[\theta_{j+1}, \theta_j]$. Hence, the algorithm described in § 3 gives, after a finite number of steps, an interval $[\theta_{k+1}, \theta_k]$ and a corresponding optimal solution $x_k(\theta)$ to PQP, such that either $0 \in [\theta_{k+1}, \theta_k]$ or the algorithm fails for $\theta < \theta_{k-1}$. In the first case $x_k(0)$ is an optimal solution to QP. The second case occurs if the pivot element is zero. Then, depending on which case applies, either SP gives $\lambda_* = -\infty$ or CRP is unable to determine a constraint to become inactive. Both alternatives indicate that QP has no optimal solution (Theorem 1, Lemma 4).

REFERENCES

[1] E. M. L. BEALE, *On quadratic programming*, Naval Res. Logist. Quart., 6 (1959), pp. 227–244.

[2] ———, *Numerical methods*, Non-Linear Programming, J. Abadie, ed., North-Holland, Amsterdam, 1967, pp. 143–170.

[3] J. C. G. BOOT, *Quadratic Programming, Algorithms-Anomalies-Application*, North Holland, Amsterdam, 1964.

[4] R. W. COTTLE, *The principal pivoting method of quadratic programming*, Lecture notes, AMS Summer Seminar on the Mathematics of the Decision Sciences, Stanford University, Stanford, California, 1967.

[5] R. W. COTTLE AND G. B. DANTZIG, *Complementary pivot theory of mathematical programming*, Tech. Rep. 67-2, OR House, Stanford University, Stanford, California, 1967.

[6] G. B. DANTZIG, *Quadratic programming, a variant of the Wolfe-Markowitz algorithms*, Operations Research Center Rep. 2, University of California, Berkeley, California, 1961.

[7] G. B. DANTZIG AND R. W. COTTLE, *Positive semidefinite programming*, Non-Linear Programming, J. Abadie, ed., North-Holland, Amsterdam, 1967, pp. 55–73.

[8] W. S. DORN, *Non-linear programming—a survey*, Management Sci., 9 (1963), pp. 171–208.

[9] H. W. KUHN AND A. W. TUCKER, *Non-linear programming*, Proc. 2nd Berkeley Symp. on Mathematical Statistics and Probability, University of California Press, Berkeley and Los Angeles, 1951, pp. 481–492.

[10] C. E. LEMKE, *A method of solution for quadratic programs*, Management Sci., 8 (1962), pp. 442–453.

[11] ———, *Bimatrix equilibrium points and mathematical programming*, Ibid., 11 (1965), pp. 681–689.

[12] ———, *On complementary pivot theory*, Rep. 76, Rennselaer Polytechnic Institute, Troy, New York, 1967.

[13] J. H. MOORE AND A. B. WHINSTON, *Experimental methods in quadratic programming*, Management Sci., 13 (1966), pp. 58–76.

[14] C. VAN DE PANNE AND A. WHINSTON, *The simplex and the dual method for quadratic programming*, Economic Institute of the Netherlands School of Economics, Rep. 6314, Rotterdam, 1962.

[15] ———, *A comparison of two methods for quadratic programming*, Operations Res., 14 (1966), pp. 422–441.

[16] T. D. PARSONS, *A combinatorial approach to convex quadratic programming*. Doctoral thesis, Department of Mathematics, Princeton University, Princeton, 1966.

[17] K. RITTER, *Ein Verfahren zur Lösung parameter-abhängiger, nichtlinearer Maximum-Probleme*, Unternehmensforschung, 6 (1962), pp. 149–166; English transl., Naval Res. Logist. Quart., 14 (1967), pp. 147–162.

[18] ———, *A decomposition method for structured quadratic programming problems*, J. Comp. System Sci., 1 (1967), pp. 241–260.

[19] J. B. ROSEN. *The gradient projection method for nonlinear programming. Part I*, J. Soc. Indust. Appl. Math., 8 (1960), pp. 181–217.

[20] H. THEIL AND C. VAN DE PANNE, *Quadratic programming as an extension of conventional quadratic maximization*, Management Sci., 7 (1960), pp. 1–20.

[21] J. B. ROSEN, *Optimal Control and Convex Programming*, Proc. IBM Scientific Computing Symposium on Control Theory and Applications, Publication 320–1939, IBM Corporation, White Plains, New York, 1966, pp. 223–237.

[22] A. W. TUCKER, *Pivotal algebra*, Lecture notes, T. D. Parsons, ed., Princeton University, Princeton. 1965.

[23] P. WOLFE, *The simplex method for quadratic programming*, Econometrica, 27 (1959), pp. 382–398.

[24] ———, *The product form of the simplex method*, Non-Linear Programming, J. Abadie, ed., North-Holland, Amsterdam, 1967, pp. 305–309.

[25] G. ZOUTENDIJK, *Methods of Feasible Directions*, Elsavier, Amsterdam, 1960.

SIAM J. CONTROL
Vol. 7, No. 4, November 1969

A NEW NECESSARY CONDITION OF OPTIMALITY FOR SINGULAR CONTROL PROBLEMS*

D. H. JACOBSON†

1. Introduction. Necessary conditions of optimality for nonsingular, unconstrained, control problems are well known. When control and state variable constraints are present, the situation is more complex, but recent research [1]–[7] indicates that many of the subtleties of this class of problems are now uncovered. In the classical calculus of variations literature, little space is devoted to the analysis of singular variational problems. Recently, interest has been aroused in singular optimal control problems [8]–[19], owing to the appearance of such problems in, for example, the aerospace field and the chemical industry.[1] Kelley discovered [20], and Robbins [21], Tait [22] and Kelley et al. [23] generalized, a new necessary condition of optimality for singular arcs. The condition, known as the generalized Legendre-Clebsch condition, has, in a number of cases, proved useful [16], [18], [23] in eliminating some stationary arcs from the class of candidate arcs for minimizing solutions. The generalized Legendre-Clebsch condition is proved using special control variations. In this paper, by the use of a different special control variation, an additional necessary condition of optimality is derived.[2] The differential dynamic programming approach, outlined in [7], [24]–[26], is used to calculate the expression for the change in cost produced by the introduction of the special variation. The new necessary condition is deduced from this expression. Control problems without terminal constraints are considered first. For this class of problems, the special control variation is a rectangular pulse. With terminal constraints present, the rectangular pulse is followed by a control variation which is designed to keep the terminal constraints satisfied to first order.

2. Preliminaries. Consider the class of control problems where the dynamical system is described by the differential equations:

$$\dot{x} = f(x, u, t), \qquad x(t_0) = x_0, \tag{1}$$

where

$$f(x, u, t) \equiv f_1(x, t) + f_u(x, t)u. \tag{2}$$

The performance of the system is measured by the cost functional

$$V(x_0, t_0) = \int_{t_0}^{t_f} L(x, t)\, dt + F(x(t_f), t_f) \tag{3}$$

* Received by the editors November 27, 1968.

† Division of Engineering and Applied Physics, Harvard University, Cambridge, Massachusetts 02138. This work was supported through the Division of Engineering and Applied Physics, Harvard University by the U.S. Army Research Office, the U.S. Air Force Office of Scientific Research and the U.S. Office of Naval Research under the Joint Services Electronics Program by Contracts N00014-67-A-0298-0006, 0005 and 0008 and by the National Aeronautics and Space Administration under Grant NGR 22-007-068.

[1] Many additional references are given in [11], [21] and [23].

[2] Some control problems are described which illustrate the necessity of the new condition in cases where the generalized Legendre–Clebsch condition is satisfied.

and the terminal state must satisfy

$$\psi(x(t_f), t_f) = 0. \tag{4}$$

The control u is required to satisfy the constraint

$$|u(t)| \leqq 1, \qquad t \in [t_0, t_f]. \tag{5}$$

Here, x is an n-dimensional state vector, and u is a scalar control. f_1 and f_u are n-dimensional vector functions of x at time t, and L and F are scalar functions. ψ is an s-dimensional column vector function of $x(t_f)$ at t_f. The final time is assumed to be given explicitly. The functions f, L and F are assumed to be three times continuously differentiable in each argument.

The control problem is: determine the control function $u(\cdot)$ to satisfy (5) and (4) and minimize the cost $V(x_0, t_0)$.

3. Necessary conditions of optimality. It can be shown, for the case where terminal constraints are absent, that the following necessary conditions of optimality hold:

$$-\dot{\bar{V}}_x = H_x(\bar{x}, \bar{u}, \bar{V}_x, t), \qquad \bar{V}_x(t_f) = F_x(\bar{x}(t_f), t_f), \tag{6}$$

where

$$\bar{u} = \arg \min_{\substack{u \\ \{|u| \leqq 1\}}} H(\bar{x}, u, \bar{V}_x, t) \tag{7a}$$

and

$$H(x, u, V_x, t) = L(x, t) + \langle V_x, f(x, u, t) \rangle. \tag{7b}$$

Here, $\bar{x}(\cdot)$, $\bar{u}(\cdot)$ denote the candidate state and control functions. The derivative $\bar{V}_x(\bar{x}, t)$ is the partial derivative[3] of $\bar{V}$—the cost produced by the control function $\bar{u}(\cdot)$. Here, $\bar{V}_x$ can be identified with Pontryagin's adjoint variable. Note that $\bar{V}_x$ may not be equal to the first partial derivative V_x^0 of the optimal cost V^0 which is obtained when optimal feedback control is used.

In general the optimal control function (for the class of problems formulated in § 2) will consist of bang-bang subarcs and singular subarcs.[4] A bang-bang arc is one along which strict equality holds in (5), except at a finite number of 'switch times' where the control $\bar{u}$ changes sign. A singular arc [15] is one along which

$$H_u(\bar{x}, \bar{V}_x, t) = 0 \tag{8}$$

for a finite time interval. Note that this implies that, on a singular arc, H is independent of the control u.

Along a singular arc, Kelley et al. [20], Robbins [21] and Tait [22] prove that an additional necessary condition of optimality is as follows:

$$(-1)^p \frac{\partial}{\partial u}\left[\frac{d^{2p}}{dt^{2p}} H_u(\bar{x}, \bar{V}_x, t)\right] \geqq 0, \tag{9}$$

[3] $\bar{V}_x$ is determined by *changing* x but *keeping* the control function *unchanged* at $\bar{u}(\cdot)$.

[4] From this point on, "arc" and "subarc" are used synonymously.

where the $2p$th time derivative of H_u is the first to contain explicitly the control u. Inequality (9) is known as the generalized Legendre–Clebsch condition.

4. Expression for change in cost when control variation is present: terminal state unconstrained. If a control function $\bar{u}(\cdot) + \delta u(\cdot)$ is applied to the system, then a trajectory $\bar{x}(\cdot) + \delta x(\cdot)$ is produced. At time t, $V(\bar{x} + \delta x, t)$ is the cost to go, from t to the final time t_f, when starting in state $\bar{x}(t) + \delta x(t)$ and using controls $\bar{u}(\cdot) + \delta u(\cdot)$. Let us assume that the cost can be expanded in a Taylor series about $\bar{x}$, t:

$$V(\bar{x} + \delta x, t) = V(\bar{x}, t) + \langle V_x(\bar{x}, t), \delta x\rangle + \tfrac{1}{2}\langle \delta x, V_{xx}(\bar{x}, t)\delta x\rangle \\ + \text{higher order terms.} \tag{10}$$

The partial derivatives in (10) are obtained by changing x but keeping the control function fixed[5] at $\bar{u}(\cdot) + \delta u(\cdot)$. $V(\bar{x}, t)$, the cost to go from t to t_f when starting in state $\bar{x}(t)$ and using controls $\bar{u}(\cdot) + \delta u(\cdot)$, can be written as

$$V(\bar{x}, t) = \bar{V}(\bar{x}, t) + a(\bar{x}, t), \tag{11}$$

where $a(\bar{x}, t)$ is the change in cost, when starting at time t in state $\bar{x}(t)$, produced by the variation[6] $\delta u(\tau)$, $\tau \in [t, t_f]$.

Using (11) in (10):

$$V(\bar{x} + \delta x, t) = \bar{V}(\bar{x}, t) + a(\bar{x}, t) + \langle V_x(\bar{x}, t), \delta x\rangle + \tfrac{1}{2}\langle \delta x, V_{xx}(\bar{x}, t)\delta x\rangle \\ + \text{higher order terms.} \tag{12}$$

From (3) it is clear that

$$\dot{V}(\bar{x} + \delta x, t) = -L(\bar{x} + \delta x, t), \tag{13}$$

whence,

$$-\frac{\partial V}{\partial t}(\bar{x} + \delta x, t) = L(\bar{x} + \delta x, t) + \langle V_x(\bar{x} + \delta x, t), f(\bar{x} + \delta x, \bar{u} + \delta u, t)\rangle. \tag{14}$$

Substituting (12) into (14) and expanding L and f in Taylor series about $\bar{x}$, we obtain

$$\begin{aligned} -\frac{\partial \bar{V}}{\partial t} - \frac{\partial a}{\partial t} - \left\langle \frac{\partial V_x}{\partial t}, \delta x\right\rangle - \frac{1}{2}\left\langle \delta x, \frac{\partial V_{xx}}{\partial t}\delta x\right\rangle &+ \text{higher order terms} \\ = H(\bar{x}, \bar{u} + \delta u, V_x, t) &+ \langle H_x + V_{xx}f, \delta x\rangle \\ &+ \tfrac{1}{2}\langle \delta x, (H_{xx} + f_x^T V_{xx} + V_{xx}f_x + \tfrac{1}{2}f^T V_{xxx} + \tfrac{1}{2}V_{xxx}f)\delta x\rangle \\ &+ \text{higher order terms.} \end{aligned} \tag{15}$$

All derivatives in (15) are evaluated at $\bar{x}, \bar{u} + \delta u, V_x, t$.

[5] Cf. § 3.

[6] We shall in this section obtain an expression for $\dot{a}(\bar{x}, t)$; this will allow us to compute the change in cost produced by the control change $\delta u(\cdot)$.

Since equality holds for all δx, we equate coefficients to obtain

$$\begin{aligned}
-\frac{\partial \bar{V}}{\partial t} - \frac{\partial a}{\partial t} &= H(\bar{x}, \bar{u} + \delta u, V_x, t), \\
-\frac{\partial V_x}{\partial t} &= H_x(\bar{x}, \bar{u} + \delta u, V_x, t) + V_{xx} f(\bar{x}, \bar{u} + \delta u, t), \\
-\frac{\partial V_{xx}}{\partial t} &= H_{xx}(\bar{x}, \bar{u} + \delta u, V_x, t) + f_x^T(\bar{x}, \bar{u} + \delta u, t) V_{xx} \\
&\quad + V_{xx} f_x(\bar{x}, \bar{u} + \delta u, t) + \tfrac{1}{2} V_{xxx} f(\bar{x}, \bar{u} + \delta u, t) \\
&\quad + \tfrac{1}{2} f^T(\bar{x}, \bar{u} + \delta u, t) V_{xxx}.
\end{aligned} \tag{16}$$

The higher order equations are not presented.

Now,

$$\frac{d}{dt}(\bar{V} + a) = \frac{d}{dt} V = \frac{\partial V}{\partial t} + \langle V_x, f(\bar{x}, \bar{u}, t)\rangle .$$

Therefore,

$$\frac{d}{dt}(\bar{V} + a) = \frac{\partial}{\partial t}(\bar{V} + a) + \langle V_x, f(\bar{x}, \bar{u}, t)\rangle \tag{17a}$$

and

$$\begin{aligned}
\dot{V}_x &= \frac{\partial V_x}{\partial t} + V_{xx} f(\bar{x}, \bar{u}, t), \\
\dot{V}_{xx} &= \frac{\partial V_{xx}}{\partial t} + \frac{1}{2} V_{xxx} f(\bar{x}, \bar{u}, t) + \frac{1}{2} f^T(\bar{x}, \bar{u}, t) V_{xxx}.
\end{aligned} \tag{17b}$$

Using (17) in (16), the following equations result:

$$\begin{aligned}
-\dot{a} &= H - H(\bar{x}, \bar{u}, V_x, t), \\
-\dot{V}_x &= H_x + V_{xx}(f - f(\bar{x}, \bar{u}, t)), \\
-\dot{V}_{xx} &= H_{xx} + f_x^T V_{xx} + V_{xx} f_x + \tfrac{1}{2} V_{xxx}(f - f(\bar{x}, \bar{u}, t)) \\
&\quad + \tfrac{1}{2}(f - f(\bar{x}, \bar{u}, t))^T V_{xxx},
\end{aligned} \tag{18}$$

where, unless otherwise specified, all quantities are evaluated at $\bar{x}, \bar{u} + \delta u$, V_x, t. Using the special structure of f, equation (2), equations (18) become:

$$\begin{aligned}
-\dot{a} &= H_u \delta u, \\
-\dot{V}_x &= H_x + (H_{xu} + V_{xx} f_u)\, \delta u, \\
-\dot{V}_{xx} &= H_{xx} + f_x^T V_{xx} + V_{xx} f_x + (H_{xxu} + f_{xu}^T V_{xx} + V_{xx} f_{xu} \\
&\qquad + \tfrac{1}{2} V_{xxx} f_u + \tfrac{1}{2} f_u^T V_{xxx})\, \delta u .
\end{aligned} \tag{19}$$

In (19), all quantities are now evaluated at $\bar{x}, \bar{u}, V_x, t$. Boundary conditions for

(19) are, clearly,

$$\begin{aligned} a(t_f) &= 0, \\ V_x(t_f) &= F_x(\bar{x}, t_f), \qquad (20) \\ V_{xx}(t_f) &= F_{xx}(\bar{x}, t_f). \end{aligned}$$

The change in cost owing to the presence of a control variation $\delta u(\tau)$; $\tau \in [t_1, t_2]$, $t_2 > t_1$, is given by

$$a(t_1) = a(t_2) + \int_{t_2}^{t_1} \dot{a}(t)\, dt. \qquad (21)$$

5. New necessary condition: unconstrained terminal state. A singular arc is assumed to lie in an interval $[t_a, t_b]$. A control variation in the form of a rectangular pulse[7] of height η and duration T is introduced in an interval $[t_1, t_2]$ where

$$t_a < t_i < t_b, \quad i = 1, 2, \quad t_2 > t_1. \qquad (22)$$

See Fig. 1.

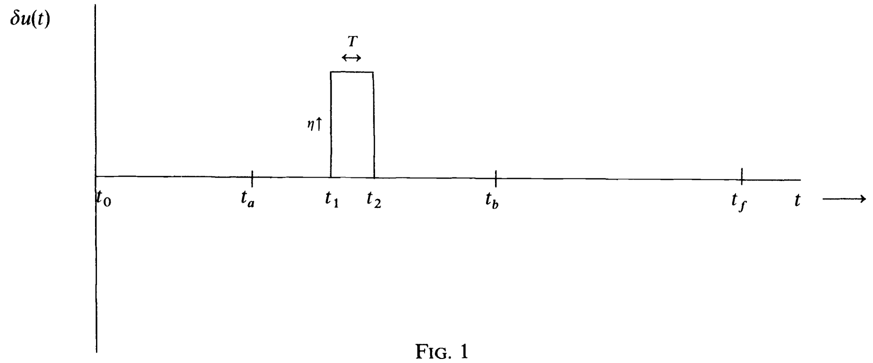


FIG. 1

The change in cost produced by this variation is given by

$$a(t_1) = \int_{t_2}^{t_1} \dot{a}\, dt + a(t_2) = \int_{t_2}^{t_1} -H_u\, \delta u\, dt + a(t_2), \qquad (23)$$

where H_u is evaluated at $\bar{x}$, V_x, t. Expanding the integral in a Taylor series in T, the expression for the change in cost becomes[8]

$$a(t_1) = H_u\, \delta u|_{t_2} T - \frac{1}{2}\frac{d}{dt}[H_u\, \delta u]\bigg|_{t_2} T^2 + \cdots + a(t_2). \qquad (24)$$

At time t_2, one has

$$\begin{aligned} a(t_2) &= 0, \\ V_x(t_2) &= \bar{V}_x(t_2), \qquad (25) \\ V_{xx}(t_2) &= \bar{V}_{xx}(t_2), \end{aligned}$$

where $\bar{V}_x(t_2)$ and $\bar{V}_{xx}(t_2)$ are computed using (19) and (20) with $\delta u(t) = 0, t \in (t_2, t_f]$.

[7] η can always be chosen so that the control constraint (5) remains satisfied.

[8] Note that in (24) quantities are evaluated at the time instant immediately prior to time t_2.

Since $\bar{x}(t_2)$ is on the singular arc, $H_u(\bar{x}, \bar{V}_x, t_2) = 0$. Thus, the first nonzero term in expansion (24) is the T^2 one. We have that

$$\frac{d}{dt}[H_u \, \delta u]\bigg|_{t_2} = \dot{H}_u \, \delta u|_{t_2} + H_u \, \delta \dot{u}|_{t_2} = \dot{H}_u(\bar{x}, \bar{V}_x, t)|_{t_2}\eta. \tag{26}$$

From (19), (20),

$$\dot{H}_u(\bar{x}, V_x, t)|_{t_2} = \{\dot{f}_u^T V_x + f_u^T[-H_x - (H_{xu} + V_{xx}f_u)\eta]\}|_{t_2}. \tag{27}$$

The first two terms in (27) sum to zero.[9] Using (27) and (26) in (24), the change in cost is

$$a(t_1) = \tfrac{1}{2} f_u^T(\bar{x}, t_2)[H_{xu}(\bar{x}, \bar{V}_x, t_2) + \bar{V}_{xx}f_u(\bar{x}, t_2)]\eta^2 T^2 + \text{higher order terms.} \tag{28}$$

For the singular arc to be a candidate as a minimizing arc, it is necessary that the change in cost, owing to the presence of the control variation, be nonnegative. From (28) this implies that

$$f_u^T(\bar{x}, t)[H_{xu}(\bar{x}, \bar{u}, \bar{V}_x, t) + \bar{V}_{xx}f_u(\bar{x}, t)] \geqq 0, \tag{29}$$

where

$$\begin{aligned} -\dot{\bar{V}}_x &= H_x(\bar{x}, \bar{u}, \bar{V}_x, t), \\ -\dot{\bar{V}}_{xx} &= H_{xx}(\bar{x}, \bar{u}, \bar{V}_x, t) + f_x^T(\bar{x}, \bar{u}, t)\bar{V}_{xx} + \bar{V}_{xx}f_x(\bar{x}, \bar{u}, t) \end{aligned} \tag{30}$$

and

$$\begin{aligned} \bar{V}_x(t_f) &= F_x(\bar{x}(t_f), t_f), \\ \bar{V}_{xx}(t_f) &= F_{xx}(\bar{x}(t_f), t_f). \end{aligned} \tag{31}$$

Inequality (29) is the new necessary condition of optimality for singular control problems with unconstrained terminal states.

6. Examples.

Example 1. Consider the following scalar control problem:

$$\dot{x} = u, \qquad x(0) = 1, \tag{32}$$

$$V(1, 0) = \int_0^2 x^2 \, dt, \tag{33}$$

$$|u| \leqq 1.$$

The optimal control is

$$\begin{aligned} \bar{u}(t) &= -1, \qquad t \in [0, 1], \\ \bar{u}(t) &= 0, \qquad t \in (1, 2]. \end{aligned} \tag{34}$$

The arc in x, t space along which $u(t)$ is zero, is singular.

[9] $\dot{H}_u(\bar{x}, \bar{V}_x, t_2) = 0 = \dot{f}_u^T \bar{V}_x - f_u^T H_x(\bar{x}, \bar{u}, \bar{V}_x, t_2)$.

For the above problem we have that

(35) $$H(x, u, V_x, t) = x^2 + V_x u,$$
$$H_u(x, V_x, t) = V_x,$$

(36) $$-\dot{\bar{V}}_x = 2\bar{x}, \qquad \bar{V}_x(t_f) = 0,$$
$$-\dot{\bar{V}}_{xx} = 2, \qquad \bar{V}_{xx}(t_f) = 0,$$

(37) $$\dot{H}_u = -2x \quad \text{and} \quad \ddot{H}_u = -2u,$$

whence

(38) $$\frac{\partial}{\partial u}\ddot{H}_u = -2,$$

so that the generalized Legendre–Clebsch condition is satisfied.

It is clear that

(39) $$f_u^T(H_{xu} + \bar{V}_{xx} f_u) = \bar{V}_{xx},$$

and from (36)

(40) $$\bar{V}_{xx}(\tau) = 2\tau,$$

where

(41) $$\tau = 2 - t.$$

From (40),

(42) $$\bar{V}_{xx}(\tau) \geqq 0 \qquad \text{for every } \tau \in [0, 1],$$

so that the new necessary condition is satisfied.

Let us consider now the following cost functional

(43) $$V(1, 0) = \int_0^2 x^2\, dt - \tfrac{1}{2} S x^2(t_f),$$

where S is positive. The control program (34) is a stationary solution for this cost functional because the first order necessary conditions of optimality are satisfied. Moreover, the generalized Legendre–Clebsch condition is satisfied. However, the differential equation for $\bar{V}_{xx}$ is

(44) $$-\dot{\bar{V}}_{xx} = 2, \qquad \bar{V}_{xx}(t_f) = -S,$$

and hence

(45) $$\bar{V}_{xx}(\tau) = -S + 2\tau.$$

Since S is positive, the new necessary condition is violated for τ sufficiently small. It can be verified directly that, for $S > 1/3$, a control function

(46) $$u(t) = -1, \qquad t \in [0, 1],$$
$$u(t) = \varepsilon, \qquad t \in [1, 2],$$

produces a cost lower than that resulting from the use of the control program (34), confirming the nonoptimality of that control program.

Example 2. Consider the following second order control problem:

$$\begin{aligned} \dot{x}_1 &= x_2, \quad & x_1(0) &= 0, \\ \dot{x}_2 &= u, \quad & x_2(0) &= 1, \end{aligned} \tag{47}$$

$$V(x_0, 0) = \tfrac{1}{2}\int_0^{3\pi/2} (x_1^2 + x_2^2)\,dt, \tag{48}$$

$$|u| \leqq 1. \tag{49}$$

Here,

$$\begin{aligned} -\dot{\bar{V}}_{x_1} &= \bar{x}_1, \quad & \bar{V}_{x_1}(t_f) &= 0, \\ -\dot{\bar{V}}_{x_2} &= \bar{x}_2 + \bar{V}_{x_1}, \quad & \bar{V}_{x_2}(t_f) &= 0 \end{aligned} \tag{50}$$

and

$$\begin{aligned} H_u &= V_{x_2}, \quad & \ddot{H}_u &= -u + x_1, \\ \dot{H}_u &= -x_2 - V_{x_1}, \end{aligned} \tag{51}$$

so that the generalized Legendre–Clebsch condition is satisfied. The expression $f_u^T(H_{xu} + \bar{V}_{xx}f_u)$ is equal to $\bar{V}_{x_2x_2}$, and

$$\bar{V}_{x_2x_2}(\tau) = \tau + \tfrac{1}{3}\tau^3, \tag{52}$$

so that the new necessary condition is satisfied. It can be verified [9] that this problem has a stationary solution which exhibits a singular arc. Moreover, the stationary solution is minimizing.

Consider now the following cost functional:

$$V(x_0, 0) = \tfrac{1}{2}\int_0^{3\pi/2} (-x_1^2 + x_2^2)\,dt. \tag{53}$$

Here,

$$\begin{aligned} -\dot{V}_{x_1} &= -x_1, \\ -\dot{V}_{x_2} &= x_2 + V_{x_1} \end{aligned} \tag{54}$$

and

$$\begin{aligned} H_u &= V_{x_2}, \quad & \ddot{H}_u &= -u - x_1, \\ \dot{H}_u &= -x_2 - V_{x_1}, \end{aligned} \tag{55}$$

so that the generalized Legendre–Clebsch condition is satisfied.

It is easy to see that

$$\bar{x}_1(t) = \sin t, \qquad \bar{x}_2(t) = \cos t \tag{56}$$

is a singular solution, and the cost functional value corresponding to this trajectory is zero.

One can verify that the equation for $\bar{V}_{x_2x_2}(\tau)$ is

$$\bar{V}_{x_2x_2}(\tau) = \tau - \tfrac{1}{3}\tau^3, \tag{57}$$

which is negative for $\tau > \sqrt{3}$: that is, the new necessary condition is violated for $\tau > \sqrt{3}$. The control function corresponding to the trajectory (56) is,

$$\bar{u}(t) = -\sin t. \tag{58}$$

Consider the control function

$$u(t) = \bar{u}(t) + \varepsilon, \tag{59}$$

where ε is a constant. Then, the resulting trajectory is

$$\begin{aligned} \bar{x}_2(t) + \delta x_2(t) &= \cos t + \varepsilon t, \\ \bar{x}_1(t) + \delta x_1(t) &= \sin t + \tfrac{1}{2}\varepsilon t^2. \end{aligned} \tag{60}$$

The integrand in (53), corresponding to trajectory (60) is:

$$-\left(\sin^2 t + \varepsilon t^2 \sin t + \frac{\varepsilon^2}{4} t^4\right) + (\cos^2 t + 2\varepsilon t \cos t + \varepsilon^2 t^2). \tag{61}$$

Using (61) in (53),

$$\begin{aligned} V(x_0, 0) &= \frac{1}{2}\int_0^{3\pi/2} \left[\cos^2 t - \sin^2 t + \varepsilon(2t\cos t - t^2 \sin t) + \varepsilon^2\left(-\frac{t^4}{4} + t^2\right)\right] dt \\ &= \frac{1}{2}\left[\frac{1}{2}\sin 2t + \varepsilon t^2 \cos t + \varepsilon^2\left(-\frac{t^5}{20} + \frac{t^3}{3}\right)\right]_0^{3\pi/2} \qquad (62) \\ &= -40.7\varepsilon^2, \qquad (63) \end{aligned}$$

which is negative, confirming the nonoptimality of the singular trajectory.

The above examples illustrate the necessity of the new condition of optimality. Also demonstrated is the nonequivalence of the new condition and the generalized Legendre–Clebsch condition.

7. Adjoining terminal constraint. Here we consider the case where equality (4) is present. The equality constraint can be adjoined to the cost functional by a vector b of Lagrange multipliers, in the following way.

$$V^*(x_0, b, t_0) = \int_{t_0}^{t_f} L(x, t)\, dt + F(x(t_f), t_f) + \langle b, \psi(x(t_f), t_f)\rangle. \tag{64}$$

Assume that $\bar{x}(\cdot)$, $\bar{b}$ and $\bar{u}(\cdot)$ are stationary solutions of (64); the following necessary conditions are satisfied along a singular arc:

$$\begin{aligned} -\dot{\bar{V}}_x^* = H_x(\bar{x}, \bar{u}, \bar{V}_x^*, t), \qquad & \bar{V}_x^*(t_f) = F_x(\bar{x}, t_f) + \psi_x^T(\bar{x}, t_f)\bar{b}, \\ H_u(\bar{x}, \bar{V}_x^*, t) &= 0. \end{aligned} \tag{65}$$

If $V^*(x_0, \bar{b}, t_0)$ has an unconstrained minimum with respect to $u(\cdot)$ at $\bar{u}(\cdot)$, the following condition must hold along the singular arc:

$$f_u^T(\bar{x}, t)[H_{xu}(\bar{x}, \bar{V}_x^*, t) + \bar{V}_{xx}^* f_u(\bar{x}, t)] \geqq 0, \tag{66}$$

where

$$-\dot{\bar{V}}_{xx}^* = H_{xx} + f_x^T \bar{V}_{xx}^* + \bar{V}_{xx} f_x, \qquad \bar{V}_{xx}^*(t_f) = F_{xx}(\bar{x}, t_f) + \bar{b}\psi_{xx}(\bar{x}, t_f). \tag{67}$$

Condition (66) follows from § 5.

However, failure of condition (66) does *not* imply that $\bar{x}(\cdot)$, $\bar{u}(\cdot)$ is not a minimizing solution for the constrained problem where equality (4) is enforced. This is so because a minimizing solution of the original constrained problem need only be a stationary solution of (64) for fixed $b = \bar{b}$. In order to determine whether $\bar{x}(\cdot)$, $\bar{u}(\cdot)$ is a possible minimizing solution, one has to ensure that, on the introduction of a control variation, equality (4) *remains* satisfied.

8. New necessary condition: constrained terminal state. Let us assume, as in § 5, that a control variation consisting of a rectangular pulse of duration $t_2 - t_1 \equiv T$ and height η is introduced in the singular control interval $[t_a, t_b]$. A further control variation is now introduced in the interval $(t_2, t_b]$ in order to force equality (4) to remain satisfied. We shall assume the following form for the control variation in the interval $(t_2, t_b]$:

$$\delta u(\tau) = \beta(\tau)\sigma, \qquad \tau \in (t_2, t_b]. \tag{68}$$

Here, $\beta(\tau)$ is a time varying, s-dimensional row vector and σ is a constant s-dimensional vector. For δx and δu sufficiently small, the following equations are valid:

$$\begin{aligned} \delta\dot{x}(\tau) &= f_x(\bar{x}, \bar{u}, \tau)\,\delta x(\tau) + f_u(\bar{x}, \tau)\beta(\tau)\sigma, && \tau \in (t_2, t_b], \\ \delta\dot{x}(\tau) &= f_x(\bar{x}, \bar{u}, \tau)\,\delta x(\tau), && \tau \in (t_b, t_f], \end{aligned} \tag{69}$$

where $\delta x(t_2) \neq 0$ owing to the rectangular pulse variation prior to t_2. In order that the control constraints (5) remain satisfied when (68) is used, it is assumed that $\bar{u}(\tau)$, $\tau \in (t_a, t_b)$ is in the interior of the control constraint set.[10]

The solution of (69) is:

$$\delta x(t) = \phi(t, t_2)\,\delta x(t_2) + \int_{t_2}^{t} \phi(t, \tau) f_u(\tau)\beta(\tau)\sigma\, d\tau, \tag{70a}$$

with

$$\beta(\tau)\sigma \equiv 0, \qquad \tau \in (t_b, t_f], \tag{70b}$$

where $\phi(t, \tau)$ satisfies the differential equation

$$\frac{d}{dt}\phi(t, \tau) = f_x(\bar{x}, \bar{u}, t)\phi(t, \tau), \qquad \phi(\tau, \tau) = I. \tag{71}$$

At $t = t_f$, we require, for $\delta x(t_f)$ sufficiently small, that

$$\psi_x(\bar{x}, t_f)\,\delta x(t_f) = 0. \tag{72}$$

Setting $t = t_f$ in (70), and using (72), we obtain

$$0 = \psi_x(\bar{x}, t_f)\phi(t_f, t_2)\,\delta x(t_2) + \psi_x(\bar{x}, t_f)\int_{t_2}^{t_f} \phi(t_f, \tau) f_u(\tau)\beta(\tau)\sigma\, d\tau, \tag{73}$$

which, by (70), is equivalent to

$$0 = \psi_x(\bar{x}, t_f)\phi(t_f, t_2)\,\delta x(t_2) + \psi_x(\bar{x}, t_f)\int_{t_2}^{t_b} \phi(t_f, \tau) f_u(\tau)\beta(\tau)\sigma\, d\tau. \tag{74}$$

[10] If the singular control and the nonsingular "bang" control are continuous at t_b, then (68) is used up until $t_b - \varepsilon$, $\varepsilon > 0$ to ensure that the control constraints remain satisfied in the interval (t_2, t_b).

Let us choose

$$\beta(\tau) = f_u^T(\tau)\phi^T(t_f, \tau)\psi_x^T(\bar{x}, t_f). \tag{75}$$

Using (75) in (74), we obtain

$$\begin{aligned}\psi_x(\bar{x}, t_f)\left[\int_{t_2}^{t_b} \phi(t_f, \tau) f_u(\tau) f_u^T(\tau)\phi^T(t_f, \tau)\, d\tau\right]\psi_x^T(\bar{x}, t_f)\sigma \\ = -\psi_x(\bar{x}, t_f)\phi(t_f, t_2)\,\delta x(t_2).\end{aligned} \tag{76}$$

Denoting the contents of the square brackets on the left-hand side of (76) by

$$W(t_2, t_b) \tag{77}$$

we obtain[11]

$$\begin{aligned}\sigma &= -[\psi_x(\bar{x}, t_f)W(t_2, t_b)\psi_x^T(\bar{x}, t_f)]^{-1}\psi_x(\bar{x}, t_f)\phi(t_f, t_2)\,\delta x(t_2) \\ &\equiv \gamma\,\delta x(t_2).\end{aligned} \tag{78}$$

We have, for $\delta x(t_2)$ sufficiently small, i.e., for η (or T) sufficiently small, that if expressions (75) and (78) are used in (68), then equality (4) is maintained to first order. That is, the change in $\delta x(t_f)$ is at most of order $[\delta x(t_2)]^2$.

For $\tau \in (t_b, t_f]$ we have the same equations for $\bar{V}_x^*$ and $\bar{V}_{xx}^*$, namely

$$\begin{aligned}&-\dot{\bar{V}}_x^* = H_x(\bar{x}, \bar{u}, \bar{V}_x^*, t), \qquad \bar{V}_x^*(t_f) = F_x + \psi_x^T\bar{b}|_{t_f}, \\ &-\dot{\bar{V}}_{xx}^* = H_{xx}(\bar{x}, \bar{u}, \bar{V}_x^*, t) + f_x^T(\bar{x}, \bar{u}, t)\bar{V}_{xx}^* + \bar{V}_{xx}^* f_x(\bar{x}, \bar{u}, t), \\ &\bar{V}_{xx}^*(t_f) = F_{xx} + \bar{b}\psi_{xx}|_{t_f}.\end{aligned} \tag{79}$$

For $\tau \in (t_2, t_b]$ the dynamical equation is

$$(\bar{x} + \delta x)^{\cdot} = f(\bar{x} + \delta x, \bar{u} + \beta\sigma, \tau) \tag{80}$$

and the cost functional is

$$V^*(\bar{x} + \delta x, \bar{b}, \tau) = \int_\tau^{t_b} L(\bar{x} + \delta x, t)\, dt + \bar{V}^*(\bar{x} + \delta x, \bar{b}, t_b). \tag{81}$$

Since the cost $V^*(\bar{x} + \delta x, \bar{b}, \tau)$, $\tau \in [t_2, t_b]$ depends on σ, let us make this dependence explicit by defining

$$J(\bar{x} + \delta x, \bar{b}, \sigma, \tau) \equiv V^*(\bar{x} + \delta x, \bar{b}, \tau) \tag{82}$$

so that

$$J(\bar{x} + \delta x, \bar{b}, \sigma, \tau) = \int_\tau^{t_b} L(\bar{x} + \delta x, t)\, dt + \bar{V}^*(\bar{x} + \delta x, \bar{b}, t_b). \tag{83}$$

In a similar way to that demonstrated in § 5, the following equations can be

[11] It is easy to show that, if the linear system $\delta\dot{x} = f_x\,\delta x + f_u\,\delta u$ is completely controllable, and if $\psi_x^T(\bar{x}, t_f)$ has full rank s, the inverse in (78) exists.

obtained:

$$
\begin{aligned}
&-\dot{J}_x = H_x, && J_x(t_b) = \bar{V}_x^*(t_b),\\
&-\dot{J}_{xx} = H_{xx} + f_x^T J_{xx} + J_{xx} f_x, && J_{xx}(t_b) = \bar{V}_{xx}^*(t_b),\\
(84)\qquad &-\dot{J}_\sigma = \beta^T H_u = 0, && J_\sigma(t_b) = 0,\\
&-\dot{J}_{x\sigma} = f_x^T J_{x\sigma} + (H_{xu} + J_{xx} f_u)\beta, && J_{x\sigma}(t_b) = 0,\\
&-\dot{J}_{\sigma\sigma} = J_{\sigma x} f_u \beta + f_u^T \beta J_{x\sigma}, && J_{\sigma\sigma}(t_b) = 0,
\end{aligned}
$$

where all quantities in (84) are evaluated at $\bar{x}, \bar{u}$. These equations can be integrated backwards from t_b until t_2 is reached. At t_2, σ is given by (78), and the expansion for $J(\bar{x} + \delta x, \bar{b}, \sigma, t_2)$ to second order in $\delta x(t_2)$ and σ is

$$
\begin{aligned}
J(\bar{x} + \delta x, \bar{b}, \sigma, t_2) = J(\bar{x}, \bar{b}, 0, t_2) &+ \langle J_x, \delta x\rangle + \langle J_\sigma, \sigma\rangle + \langle \delta x, J_{x\sigma}\sigma\rangle\\
&+ \tfrac{1}{2}\langle \delta x, J_{xx}\delta x\rangle + \tfrac{1}{2}\langle \sigma, J_{\sigma\sigma}\sigma\rangle .
\end{aligned}
$$

Substituting into (84) the value of σ, we obtain

$$
\begin{aligned}
(85)\quad J(\bar{x} + \delta x, \bar{b}, \gamma\,\delta x, t_2) = J(\bar{x}, \bar{b}, 0, t_2) &+ \langle J_x, \delta x\rangle + \langle J_\sigma, \gamma\,\delta x\rangle + \langle \delta x, J_{x\sigma}\gamma\,\delta x\rangle\\
&+ \tfrac{1}{2}\langle \delta x, J_{xx}\,\delta x\rangle + \tfrac{1}{2}\langle \delta x, \gamma^T J_{\sigma\sigma}\gamma\,\delta x\rangle .
\end{aligned}
$$

Renaming the left-hand side of (85) as $\hat{J}(\bar{x} + \delta x, \bar{b}, t_2)$ we obtain

$$
(86)\qquad \hat{J}_{xx} = J_{xx} + \gamma^T J_{\sigma\sigma}\gamma + J_{x\sigma}\gamma + \gamma^T J_{\sigma x}
$$

and

$$
(87)\qquad \hat{J}_x = J_x, \qquad \text{since } J_\sigma = 0.
$$

Equations (86) and (87) are the second and first partial derivatives of the cost at $t = t_2$, given that the terminal constraints (4) are satisfied to first order.

From § 5, the change in cost, owing to the presence of the rectangular pulse in the interval $[t_1, t_2]$ is

$$
(88)\qquad \tfrac{1}{2} f_u^T(\bar{x}, t_2)[H_{xu}(\bar{x}, \hat{J}_x, t_2) + \hat{J}_{xx} f_u(\bar{x}, t_2)]\eta^2 T^2 + \text{higher order terms}
$$

Thus the new necessary condition of optimality for singular problems with terminal constraints is

$$
(89)\qquad f_u^T(\bar{x}, t)[H_{xu}(\bar{x}, \hat{J}_x, t) + \hat{J}_{xx} f_u(\bar{x}, t)] \geqq 0.
$$

As mentioned earlier, the control $\beta(\tau)\sigma$ only ensures that the terminal constraints are satisfied to first order. In the Appendix it is demonstrated that if the terminal constraints are satisfied to second order, conclusion (89) is unaffected. This is true also if the terminal constraints are satisfied to higher order, or satisfied exactly.

9. Example. Consider the following scalar control problem:

$$
(90)\qquad \dot{x} = u, \qquad x(0) = 1,
$$

$$
(91)\qquad V(1, 0) = \int_0^2 x^2\, dt - \tfrac{1}{2} S x^2(2),
$$

with the terminal constraint that

$$
(92)\qquad x(2) = 0
$$

and the control constraint

$$|u| \leqq 1. \tag{93}$$

In § 6 it was demonstrated that, in the absence of equality (92), and for $S > 0$, the following control program is a stationary, nonminimizing solution:

$$\begin{aligned} u(\tau) &= -1, \qquad \tau \in [0, 1], \\ u(\tau) &= 0, \qquad \tau \in (1, 2]. \end{aligned} \tag{94}$$

We shall demonstrate now that with equality (92) present, the new necessary condition (89) is satisfied (by inspection the control program (94) is optimal for all S).

Since the singular arc extends from $t = 1$ to $t = 2$, we have $t_b = t_f = 2$. For the above problem, equations (84) become

$$\begin{aligned} -\dot{J}_x &= 2x, & -\dot{J}_{x\sigma} &= J_{xx}, \\ -\dot{J}_{xx} &= 2, & -\dot{J}_{\sigma\sigma} &= 2J_{\sigma x}. \\ -\dot{J}_\sigma &= 0, \end{aligned} \tag{95}$$

Boundary conditions for (95) are zero at $t = 2$, except for

$$J_{xx}(2) = -S. \tag{96}$$

From (95) and (96), we obtain the solutions

$$\begin{aligned} J_x(\tau) &= 0, & J_{x\sigma}(\tau) &= -S\tau + \tau^2, \\ J_{xx}(\tau) &= -S + 2\tau, & J_{\sigma\sigma}(\tau) &= -S\tau^2 + \tfrac{2}{3}\tau^3 \\ J_\sigma(\tau) &= 0, \end{aligned} \tag{97}$$

along the singular arc.

In addition,

$$\gamma = -\tau^{-1}, \tag{98}$$

where $\tau = 2 - t$, $t \geqq 1$. From (86), (97) and (98),

$$\hat{J}_{xx} = -S + 2\tau + 2(-S\tau + \tau^2)(-\tau^{-1}) + (\tau^{-1})^2(-S\tau^2 + \tfrac{2}{3}\tau^3) \tag{99}$$

and we have that

$$f_u^T(H_{xu} + \hat{J}_{xx} f_u) = \hat{J}_{xx} = \tfrac{2}{3}\tau \tag{100}$$

so that the new necessary condition is satisfied for all $\tau \geqq 0$, *independent* of S; this is the desired result.

10. Generalized Legendre–Clebsch necessary condition. In [23], Kelley et al. used a special control variation of the form shown in Fig. 2 to derive the first generalization of the Legendre–Clebsch condition. They gave an heuristic argument to demonstrate that, if the control problem is normal, then a control variation can be found such that the terminal constraints (4) are met, at least to first order, and the resulting change in cost owing to this added variation is negligible compared to that caused by the variation shown in Fig. 2.

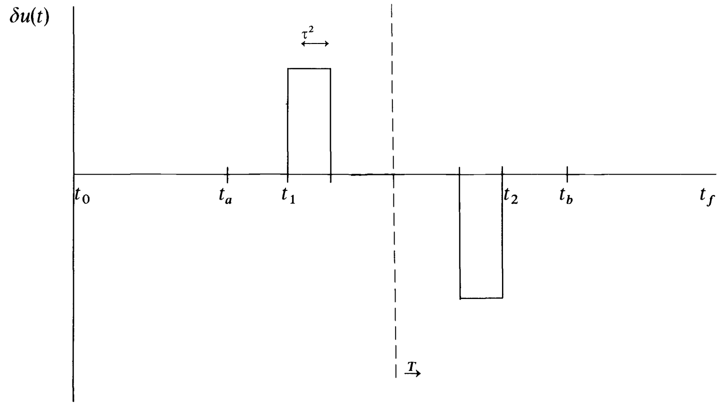


FIG. 2

If our rectangular pulse is replaced by Kelley's special variation, and if (68) is used to maintain the terminal equality (4) to first order, then expansion of (21) yields—upon requiring $a(t_1)$ to be greater than or equal to zero—the first generalization of the Legendre–Clebsch condition

$$(101) \qquad \frac{\partial}{\partial u}\left[\frac{d^2}{dt^2}H_u(\bar{x}, \hat{J}_x, t_2 - \tau)\right] \leqq 0.$$

The normality assumption of Kelley and Robbins is the same as our assumption of controllability of

$$(102) \qquad \delta\dot{x} = f_x\,\delta x + f_u\,\delta u$$

and the maximal rank of ψ_x^T required to ensure the existence of the control variation $\beta(\tau)\sigma$ which maintains satisfaction of terminal constraints (4) to first order. The complete generalized Legendre–Clebsch condition can be derived by using Kelley's generalized special variation.

11. Conclusion. In this paper we have derived a new necessary condition of optimality for singular control problems. The control problem without terminal constraints was treated first. With terminal constraints present, a special admissible control variation has to be constructed; this requires that the control problem be normal.[12]

The differential dynamic programming technique was used to obtain an expression for the change in cost produced by the control variation. For the singular arc to be minimizing it is necessary that this change in cost be nonnegative; from this requirement the new necessary conditions were deduced. Simple examples were used to illustrate the nonequivalence of the new conditions and the generalized Legendre–Clebsch condition. Finally, it was remarked that the generalized Legendre–Clebsch condition can be obtained by expanding (21) and using Kelley's special variation followed by the variation (68) which maintains satisfaction of the terminal constraints, (4), to first order.

[12] Note that if the linearized system is completely controllable then condition (89) applies equally well to the control problem without terminal constraints; that is, (89) must be true for *all* matrices ψ_x^T of full rank.

It should be noted that the new necessary conditions are derived for the case of u a scalar and the final time t_f fixed. There appear to be no conceptual difficulties in extending the derivations to include vector controls and implicitly given final time; however, the algebraic manipulations become involved if t_f is not given explicitly.

A comment on the complexity of the new necessary conditions is in order. For the free endpoint problem, $\bar{V}_{xx}$ has to be obtained from the backward equation (30); however, this is a linear differential equation whose fundamental (transition) matrix is the same as that of the $\dot{\bar{V}}_x$ equation, so that computation of $\bar{V}_{xx}$ is straightforward. Note that $\bar{V}_{xx}$ is an $n \times n$ symmetric matrix. The fixed endpoint problem requires the integration of the extra $\dot{J}_{x\sigma}$ and $\dot{J}_{\sigma\sigma}$ differential equations and γ has to be calculated (78). The $\dot{J}_{x\sigma}$ equation is linear and its fundamental matrix is the same as that of the $\bar{V}_x$ equations. Note that $J_{x\sigma}$ is an $n \times s$ matrix and that, like $\bar{V}_{xx}$, J_{xx} is an $n \times n$ matrix. The simple scalar example of § 9 required the integration of four linear differential equations to obtain J_x, J_{xx}, $J_{x\sigma}$ and $J_{\sigma\sigma}$.

In some aerospace problems, stationary control functions have been determined which pass the generalized Legendre–Clebsch test, but whose optimality remains in doubt. The new necessary condition of optimality should prove useful in ascertaining whether indeed these control functions are extremal or not.

Further, it is hoped that a useful sufficiency condition of optimality will evolve from the type of arguments presented in this paper, and that this will lead to the development of numerical techniques for solving singular optimal control problems.

Appendix.

A1. Satisfaction of terminal constraints to second order. Expansion of equality (4) to second order in $\delta x(t_f)$ about $\bar{x}(t_f)$ yields[13]

$$\psi_x(\bar{x}, t_f)\,\delta x + \tfrac{1}{2}\psi_{xx}(\bar{x}, t_f)\,\delta x\,\delta x = 0, \tag{A.1}$$

and the expansion of (1) to second order in δx and δu, about $\bar{x}$, $\bar{u}$ is

$$\delta\dot{x} = f_x\,\delta x + f_u\,\delta u + f_{ux}\,\delta u\,\delta x + \tfrac{1}{2}f_{xx}\,\delta x\,\delta x, \qquad \delta x(t_2) \neq 0. \tag{A.2}$$

The solution of (A.2), correct to second order terms is

$$\begin{aligned}\delta x(t) = \phi(t, t_2)\,\delta x(t_0) &+ \int_{t_2}^{t} \phi(t, \tau) f_u(\tau)\,\delta u(\tau)\,d\tau + \tfrac{1}{2}\phi_x(t, t_2)\,\delta x(t)\,\delta x(t_2) \\ &+ \tfrac{1}{2}\int_{t_2}^{t} \phi(t, \tau) f_{ux}(\tau)\,\delta u(\tau)\,\delta x(\tau)\,d\tau \\ &+ \tfrac{1}{2}\int_{t_2}^{t} \phi_x(t, \tau)\,\delta x(t) f_u(\tau)\,\delta u(\tau)\,d\tau,\end{aligned} \tag{A.3}$$

where

$$\dot{\phi}(t, \tau) = f_x\phi(t, \tau), \qquad \phi(\tau, \tau) = I, \tag{A.4}$$

[13] $\psi_{xx}(\bar{x}, t_f)\,\delta x\,\delta x \equiv \sum_{i=1}^{n} \sum_{i=1}^{n} \psi_{x_i x_j}\,\delta x_i\,\delta x_j$; similarly for $f_{ux}\,\delta u\,\delta x$ and $f_{xx}\,\delta x\,\delta x$.

and

$$\dot{\phi}_x(t,\tau) = f_x \phi_x(t,\tau) - \phi_x(t,\tau) f_x + f_{xx}\phi(t,\tau), \qquad \phi_x(\tau,\tau) = 0. \tag{A.5}$$

The quadratic terms on the right-hand side (RHS) of (A.3) contain $\delta x(t)$. For (A.3) to be correct to second order terms, the following expression can be used for $\delta x(t)$ in the RHS of (A.3):

$$\delta x(t) \cong \phi(t, t_2)\,\delta x(t_2) + \int_{t_2}^{t} \phi(t,\tau) f_u(\tau)\,\delta u(\tau)\,d\tau. \tag{A.6}$$

Substituting (A.6) into the RHS of (A.3), we obtain

$$\begin{aligned}\delta x(t) = {} & \phi(t, t_2)\,\delta x(t_2) + \int_{t_2}^{t} \phi(t,\tau) f_u(\tau)\,\delta u(\tau)\,d\tau \\ & + \tfrac{1}{2}\phi_x(t, t_2)\left[\phi(t, t_2)\,\delta x(t_0) + \int_{t_2}^{t} \phi(t,\tau) f_u(\tau)\,\delta u(\tau)\,d\tau\right]\delta x(t_2) \\ & + \tfrac{1}{2}\int_{t_0}^{t} \phi(t,\tau) f_{ux}(\tau)\,\delta u(\tau) \\ & \cdot \left[\int_{t_2}^{\tau} \phi(\tau, t_2)\,\delta x(t_2) + \int_{t_2}^{\tau} \phi(\tau,\tau') f_u(\tau')\,\delta u(\tau')\,d\tau'\right] d\tau \\ & + \tfrac{1}{2}\int_{t_2}^{t} \phi_x(t,\tau)\left[\phi(t, t_2)\,\delta x(t_2) + \int_{t_2}^{t} \phi(t,\tau') f_u(\tau')\,\delta u(\tau')\,d\tau'\right]\delta u(\tau)\,d\tau.\end{aligned} \tag{A.7}$$

The form of $\delta u(\tau)$, based on satisfaction of the terminal constraints to first order is, from (68),

$$\delta u(\tau) = \beta(\tau)\sigma, \tag{A.8}$$

where σ is given by (78) as

$$\sigma = \gamma\,\delta x(t_2) \tag{A.9}$$

so that σ is of first order in $\delta x(t_2)$.

Let us assume now that $\delta u(\tau)$ is of the form

$$\delta u(\tau) = \beta(\tau)[\sigma + \alpha], \tag{A.10}$$

where α is of order $\delta x^2(t_2)$. Then, to second order in $\delta x(t_2)$, $\delta x(t_f)$ is as follows:

$$\begin{aligned}\delta x(t_f) = {} & \phi(t_f, t_2)\,\delta x(t_2) + \int_{t_2}^{t_f} \phi(t_f,\tau) f_u(\tau)\beta(\tau)[\sigma + \alpha]\,d\tau \\ & + \tfrac{1}{2}\phi_x(t_f, t_2)\left[\phi(t_f, t_2)\,\delta x(t_0) + \int_{t_2}^{t_f} \phi(t_f,\tau) f_u(\tau)\beta(\tau)\sigma\,d\tau\right]\delta x(t_2) \\ & + \tfrac{1}{2}\int_{t_0}^{t_f} \phi(t_f,\tau) f_{ux}(\tau)\beta(\tau)\sigma \\ & \cdot \left[\int_{t_0}^{\tau} \phi(\tau, t_2)\,\delta x(t_2) + \int_{t_0}^{\tau} \phi(\tau,\tau') f_u(\tau')\beta(\tau')\sigma\,d\tau'\right] d\tau \\ & + \tfrac{1}{2}\int_{t_2}^{t_f} \phi_x(t_f,\tau)\left[\phi(t_f, t_2)\,\delta x(t_2)\right. \\ & \qquad \left. + \int_{t_2}^{t_f} \phi(t_f,\tau') f_u(\tau')\beta(\tau')\sigma\,d\tau'\right]\beta(\tau)\sigma\,d\tau.\end{aligned} \tag{A.11}$$

If (A.11) is now substituted into (A.1), then the first order terms in σ vanish, because of (73), leaving

$$\text{(A.12)} \qquad \left[\int_{t_2}^{t_f} \psi_x(\bar{x}, t_f) f_u(\tau)\beta(\tau)\, d\tau\right] \alpha + \text{terms of order } \delta x^2(t_2) = 0.$$

The quantity in the square brackets on the LHS of (A.12) is, from (76), just

$$\text{(A.13)} \qquad \psi_x(\bar{x}, t_f) W(t_2, t_b) \psi_x^T(\bar{x}, t_f),$$

which is invertible. So, from (A.12), α can be found and it is of order $\delta x^2(t_2)$. Thus a control variation of form

$$\text{(A.14)} \qquad \delta u(\tau) = \beta(\tau)[\sigma + \alpha],$$

where σ is first order in $\delta x(t_2)$ and α is second order in $\delta x(t_2)$ maintains the terminal equality (4) correct to second order terms. Now if (A.14) is substituted into (85), we find that, because α is second order in $\delta x(t_2)$, and $J_\sigma = 0$, (86) and (87) do not contain α; thus the conclusion (89) is unaffected if we satisfy the terminal constraints to second order rather than to first order. Satisfaction of terminal constraints to higher order yields the same result. This confirms that, in order to include all second order terms in an expansion of the cost functional, it is only necessary to expand the Hamiltonian to second order terms and the dynamic and terminal constraints to first order terms.

Acknowledgments. Stimulating discussions with Dr. J. L. Speyer and Professor Y. C. Ho are gratefully acknowledged.

REFERENCES

[1] L. D. Berkovitz, *Variational methods in problems of control and programming*, J. Math. Anal. Appl., 3 (1961), pp. 145–169.

[2] S. E. Dreyfus, *Variational problems with state variable inequality constraints*, Ibid., 4 (1962), pp. 297–308.

[3] R. V. Gamkrelidze, *Optimal processes with bounded phase coordinates*, Izv. Akad. Nauk SSSR Ser. Mat., 24 (1960), pp. 315–356.

[4] S. S. L. Chang, *Optimal control in bounded state space*, Automatica, 1 (1962), pp. 55–67.

[5] J. McIntyre and B. Paiewonsky, *On optimal control with bounded state variables*, Advances in Control Systems, vol. 5, C. T. Leondes, ed., Academic Press, New York, 1967.

[6] J. L. Speyer, *Optimization and control of nonlinear systems with inflight constraints*, Doctoral thesis, Harvard University, Cambridge, Massachusetts, 1968.

[7] D. H. Jacobson, *Differential dynamic programming methods for solving bang-bang control problems*, IEEE Trans. Automatic Control, AC-13 (1968), pp. 661–675.

[8] C. D. Johnson and J. E. Gibson, *Singular solutions in problems of optimal control*, Ibid., AC-8 (1963), pp. 4–15.

[9] W. M. Wonham and C. D. Johnson, *Optimal bang-bang control with quadratic performance index*, ASME. Trans. J. Basic Engrg., 86 (1964), pp. 107–115.

[10] D. R. Snow, *Singular optimal controls for a class of minimum effort problems*, this Journal, 2 (1964), pp. 203–219.

[11] C. D. Johnson, *Singular solutions in optimal control problems*, Advances in Control Systems, vol. 2, C. T. Leondes, ed., Academic Press, New York, 1965.

[12] M. Athans and M. D. Cannon, *On the fuel optimal singular control of nonlinear second-order systems*, IEEE Trans. Automatic Control, AC-9 (1964), pp. 360–370.

[13] R. W. Bass and R. F. Weber, *On synthesis of optimal bang-bang feedback control systems with quadratic performance criterion*, Proc. 6th Joint Automatic Control Conference, Rensselaer Polytechnic Institute, Troy, New York, 1965, pp. 213–219.

[14] H. HERMES, *Controllability and the singular problem*, this Journal, 2 (1964), pp. 241–260.

[15] H. HERMES AND G. W. HAYNES, *On the nonlinear control problem with control appearing linearly*, this Journal, 1 (1963), pp. 85–107.

[16] H. J. KELLEY, *Singular extremals in Lawden's problem of optimal rocket flight*, J. AIAA, 1 (1963), pp. 1578–1580.

[17] ———, *A transformation approach to singular subarcs in optimal trajectory and control problems*, this Journal, 2 (1964), pp. 234–240.

[18] H. M. ROBBINS, *Optimality of intermediate-thrust arcs of rocket trajectories*, J. AIAA, 3 (1965), pp. 1094–1098.

[19] C. G. PFEIFFER, *Some new results in optimal final value control theory*, J. Franklin Inst., 283 (1967), pp. 404–425.

[20] H. J. KELLEY, *A second variation test for singular extremals*, J. AIAA, 2 (1964), pp. 1380–1382.

[21] H. M. ROBBINS, *A generalized Legendre–Clebsch condition for the singular cases of optimal control*, Rep. 66–825, IBM, Federal Systems Division, Owego, New York, 1966, p. 2043.

[22] K. S. TAIT, *Singular problems in optimal control*, Doctoral thesis, Harvard University, Cambridge, Massachusetts, 1965.

[23] H. J. KELLEY, R. E. KOPP AND H. G. MOYER, *Singular Extremals*, *Topics in Optimization*, G. Leitman, ed., Academic Press, New York, 1967.

[24] D. H. JACOBSON, *New second-order and first-order algorithms for determining optimal control: a differential dynamic programming approach*, J. Opt. Theory Appl., 2 (1968), pp. 411–440.

[25] ———, *Second-order and second-variation methods for determining optimal control: a comparative study using differential dynamic programming*, Internat J. Control, 7 (1968), pp. 175–196.

[26] D. H. JACOBSON AND D. Q. MAYNE, *Differential Dynamic Programming*, American Elsevier, New York, 1969, to appear.

SIAM J. CONTROL
Vol. 7, No. 4, November 1969

SEQUENTIALLY BEST FILTER*

FRANK P. ROMEO, JR.†

Abstract. The noisy-state noisy-observation filtering problem is broached in the language of stochastic differential equations. The state and observation equations are nonlinear and time varying,

$$dx(t) = m(x(t), t)\,dt + \sigma(x(t), t)\,d\xi(t), \tag{1}$$

$$dy(t) = n(x(t), t)\,dt + d\eta(t). \tag{2}$$

By taking $\xi(t)$ and $\eta(t)$ to be independent Brownian motions, (1) and (2) define a two-dimensional diffusion process. The pair $\dot{\xi}$ and $\dot{\eta}$ simulate white Gaussian noise processes which drive the state variable $x(t)$ and corrupt the observation $y(t)$, respectively.

Only recursive estimators will be considered, so only filters whose dynamics can be put into the form

$$dz(t) = g(z(t), t)\,dt + f(z(t), t)\,dy(t) \tag{3}$$

will be admissible. ($z(t)$ is the estimation of $x(t)$, the instantaneous value of the state variable.)

Even though the square loss function is assumed, ambiguity persists with respect to the time t at which $E(z(t) - x(t))^2$ should be minimized. A precise criterion is provided with the definition of the sequentially best filter. A filter is *sequentially best* if every other dynamic scheme which has a smaller error at some instant does so by accruing a larger error beforehand.

A dividend of the above criterion is that the point by point optimization lends itself to an algorithm for the generation of the sequentially best $f(\cdot, \cdot)$ and $g(\cdot, \cdot)$ in (3).

To demonstrate the usefulness of this philosophy, an analytic (linear case) and a computerized (a nonlinear case) example are described.

A heuristic proof of the coincidence of the sequentially best filter and the uniformly best (Kalman–Bucy) filter in the time-invariant case is offered.

1. Statement of the problem. The purpose of this paper is to propose a reasonable recursive estimation scheme for filtering a noise driven state variable from a noisy observation. The behavior of the state variable $x(t)$ is described by the nonlinear time-varying equation:

$$dx(t) = m(x(t), t)\,dt + \sigma(x(t), t)\,d\xi(t), \tag{1.1}$$

where the driving function $\dot{\xi}(t)$ is a sample from a Gaussian white noise process. The observation variable $y(t)$ is related to the state by way of the time-varying function $n(\cdot, t)$ and is corrupted by another additive white noise $\dot{\eta}(t)$:

$$dy(t) = n(x(t), t)\,dt + d\eta(t). \tag{1.2}$$

The restriction to recursive estimates means that only those filters which can be described by a differential equation will be considered. More precisely, if $z(t)$ denotes the estimate of $x(t)$, then

$$dz(t) = g(z(t), t)\,dt + f(z(t), t)\,dy(t) \tag{1.3}$$

or

$$z(t_2) - z(t_1) = \int_{t_1}^{t_2} g(z(\tau), \tau)\,d\tau + \int_{t_1}^{t_2} f(z(\tau), \tau)\,dy(\tau), \tag{1.4}$$

* Received by the editors December 13, 1968, and in final revised form June 6, 1969.

† Ocean Systems Research Department, Bell Telephone Laboratories, Incorporated, Whippany, New Jersey 07981. This work was supported by the National Science Foundation under Grant GK-716 through the Electronics Research Laboratory, University of California, Berkeley.

for some pair $g(\cdot,\cdot)$ and $f(\cdot,\cdot)$, defines the admissible class of filters. Because $\dot{y}(t)$ contains a white noise process, only linear operations on the observation are mathematically justifiable (see [1]). Thus (1.3) embodies considerable generality, the only restrictions being that $z(t)$ is scalar-valued and Markovian. Defining the incremental behavior clearly implies the recursiveness of the estimator, i.e., (1.4) testifies that the new estimate depends only on the old estimate and the received data.

It is expedient to point out here that substitution of (1.2) into (1.3) results in:

$$dz(t) = [g(z(t),t) + f(z(t),t)n(x(t),t)]\,dt + f(z(t),t)\,d\eta(t). \tag{1.5}$$

The transition density for the two-dimensional diffusion process (defined by (1.1) and (1.5)) has an implicit representation provided certain conditions on the coefficients are satisfied. These conditions are mild restrictions on the smoothness of $m(\cdot,\cdot)$, $n(\cdot,\cdot)$ and $\sigma(\cdot,\cdot)$. It is assumed that they are satisfied by the model defined by (1.1) and (1.2). (A thorough discussion of the regularity needed is given in [1].) The joint probability density for the $(x(t), z(t))$ process, defined by

$$\Pr\{x(t) < \alpha \text{ and } z(t) < \beta\} = \int_{-\infty}^{\alpha}\int_{-\infty}^{\beta} P(a,b,t)\,da\,db,$$

can be shown to satisfy:

$$\begin{aligned}\frac{\partial}{\partial t}P(a,b,t) = &-\frac{\partial}{\partial a}(m(a,t)P(a,b,t))\\ &-\frac{\partial}{\partial b}[(f(b,t)n(a,t) + g(b,t))P(a,b,t)]\\ &+\frac{1}{2}\frac{\partial^2}{\partial a^2}(\sigma^2(a,t)P(a,b,t)) + \frac{1}{2}\frac{\partial^2}{\partial b^2}(f^2(b,t)P(a,b,t)).\end{aligned} \tag{1.6}$$

Equation (1.6) comes directly from Kolmogorov's forward equation when the initial distribution on $(x(0), z(0))$ is integrated out.

The usual square loss function is used so all criteria will be based on reducing

$$E(z(t) - x(t))^2. \tag{1.7}$$

The quantity (1.7) is smallest when

$$z(t) = E(x(t) \mid y(\tau), 0 \leqq \tau \leqq t), \tag{1.8}$$

i.e., the conditional expectation (conditioned on the entire observation curve) minimizes the mean-square error. The incompatibility between estimator (1.8) and recursive estimation schemes is well known (see [2], [3], [4] or [5]). The difficulty is essentially that if $z(t)$ is defined by (1.8), then

$$\frac{\partial}{\partial t}z(t) = \mathscr{F}[z(t), E(x^k(t) \mid y(\tau), 0 \leqq \tau \leqq t)], \qquad k = 2, 3, \cdots,$$

for some function $\mathscr{F}$, i.e., the rate of change of the conditional mean is not finite-dimensional in the general (nonlinear) case. If, on the other hand, one imposes

(1.3) on the form of the filter and strives to minimize the error at fixed time T or the average error over $[0, T]$,

$$E(z(T) - x(T))^2 \tag{1.9}$$

or

$$\frac{1}{T}\int_0^T E(z(t) - x(t))^2 \, dt \tag{1.10}$$

respectively, the computational problems are prohibitive. In theory, the technique of minimizing the error at T would be to choose $g(\cdot, \cdot)$ and $f(\cdot, \cdot)$ globally on $(-\infty, \infty) \times [0, T]$ such that the solution to (1.6) minimizes (1.9) or (1.10). Even if implementation of the above were feasible, one would need justification for either putting all the weight on instant T or uniformly on $[0, T]$. Thus the artifice that insures realizability (adoption of (1.3)) raises problems as to a precise definition of the criterion as well as calculus of variation problems of extreme difficulty.

In § 2 a criterion is suggested for which a tractable algorithm for finding $g(\cdot, \cdot)$ and $f(\cdot, \cdot)$ exists. The justification of the criterion lies not in its implementation; rather it will be shown to have intuitive appeal and indeed may be argued to be the ideal in many situations.

2. Sequential criterion and main result. Each filter may be identified with its error-time curve, i.e., for each pair of functions $g(\cdot, \cdot)$ and $f(\cdot, \cdot)$ the resulting solution of (1.6) results in a real-valued function of time $\mathscr{E}(t)$, where

$$\mathscr{E}(t) = E(z(t) - x(t))^2.$$

Necessarily $\mathscr{E}(0)$ will be the same for all filters. A ranking may be placed on the set of filters by way of the error-time curves. The logic is best understood by first considering the discrete analogue of the problem. Suppose one receives information sequentially, i.e., first Δy_1, then Δy_2. To conform with the doctrine of recursive estimation, one must operate on Δy_1 and the initial estimate to get some number N_1, then operate on N_1 and Δy_2 to get the final estimate z_2. (z_2 is the estimate of x_2.)

Obviously the added requirement that N_1 be a good estimate of x_1 may reduce the maximum accuracy of z_2. A pair of discrete error-time curves appears in Fig. 1a. According to the criterion defined below, while filter $2D$ is the better "predictor," the filter with the curve denoted by $1D$ is preferable to $2D$.

DEFINITION. A filter is *sequentially best* if every other filter that has a smaller error at any time t_3 also has a larger error for some interval $(t_1, t_2) \subseteq [0, t_3]$.

The criterion is demonstrated in Fig. 1b. By definition, filter $1C$ is sequentially better than filter $2C$. The sequentially best filter in this sense is the best "tracker," the one that operates on the current estimate and the new data to best update the new estimate without regard to the future. Notice that if a uniformly best estimator exists, i.e., a filter whose output *always* has the smallest error, then it must also be sequentially best. (See Example 1.)

Having made the criterion precise, there remains only the problem of finding this optimal filter. The algorithm is given in the following theorem, the main result.

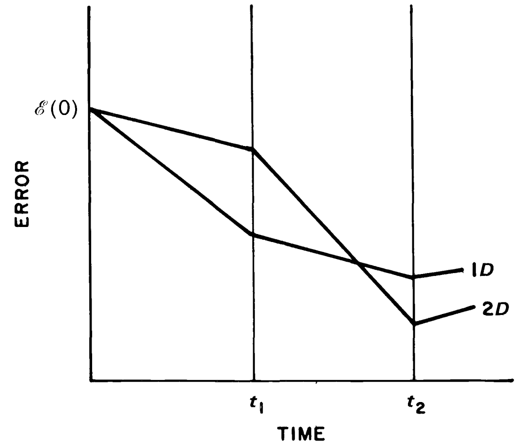


FIG. 1a

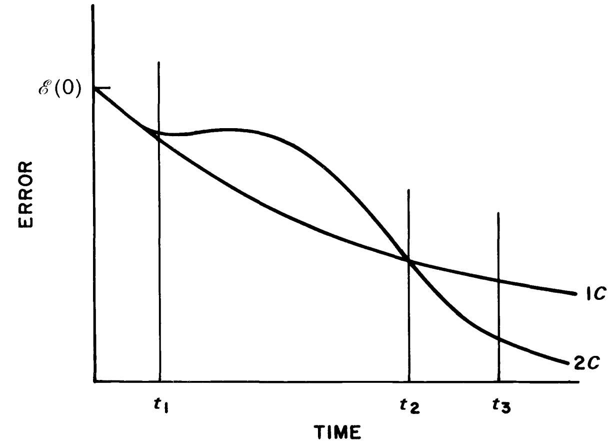


FIG. 1b

THEOREM. *If a sequentially best estimator exists, it must simultaneously satisfy* (1.6) *and*

$$g(b,t) = \bar{m}(b,t) - f(b,t)\bar{n}(b,t), \tag{2.1}$$

$$f(b,t) = \overline{nx}(b,t) - b\bar{n}(b,t), \tag{2.2}$$

where

$$\bar{r}(b,t) = \int_{-\infty}^{\infty} r(a,t)P(a,b,t)\,da\left[\int_{-\infty}^{\infty} P(a,b,t)\,da\right]^{-1}$$

for $r(a,t)$ equal to $m(a,t)$, $n(a,t)$ or $n(a,t)a$.

Thus to find the sequentially best filter one must solve the Kolmogorov forward equation where the coefficients depend on the solution by way of certain conditional expectations.

3. Proof of the main result. All calculations will begin with discrete time versions of the processes involved; thus transition density functions will be needed for the resulting difference equations. The state and observation equations are repeated here:

$$dx(t) = m(x(t), t)\,dt + \sigma(x(t), t)\,d\xi(t), \tag{3.1}$$

$$dy(t) = n(x(t), t)\,dt + d\eta(t), \tag{3.2}$$

where $\xi(t)$ and $\eta(t)$ are independent standard Brownian motions. The reader will recall that the Brownian motion process is the zero-mean Gaussian process with $E(\xi(s) - \xi(t))^2 = E(\eta(s) - \eta(t))^2 = |s - t|$. (Let $\xi(0)$ and $\eta(0)$ be identically zero.) The motive for the incremental equation for $x(\cdot)$ and $y(\cdot)$ instead of $\dot{x}(\cdot)$ and $\dot{y}(\cdot)$ is that the problem may now be treated within the framework of stochastic differential equations as defined by Ito. (Ito's work [6] is definitive, but [7] is sufficiently thorough and more accessible.)

Equations (3.1) and (3.2) are replaced by their corresponding Markov chains.

$$\begin{aligned}
x_{i+1}^n &= x_i^n + m(x_i^n, t_i^n)\,\Delta t^n + \sigma(x_i^n, t_i^n)\,\Delta\xi_i^n,\\
y_{i+1}^n &= y_i^n + n(x_i^n, t_i^n)\,\Delta t^n + \Delta\eta_i^n,\\
z_{i+1}^n &= z_i^n + g(z_i^n, t_i^n)\,\Delta t^n + f(z_i^n, t_i^n)(y_{i+1}^n - y_i^n)\\
&= z_i^n + [f(z_i^n, t_i^n)n(x_i^n, t_i^n) + g(z_i^n, t_i^n)]\Delta t^n + f(z_i^n, t_i^n)\,\Delta\eta_i^n.
\end{aligned}$$

The subscript n denotes partition T_n, where

$$T_n = \{t_0^n = 0, t_1^n, t_2^n, \cdots, t_n^n = T\}$$

and

$$\max_{1 \leqq k \leqq n} |t_k^n - t_{k-1}^n| \to 0 \quad \text{as } n \to \infty.$$

Let

$$x_i^n = x(t_i^n), \quad y_i^n = y(t_i^n), \quad \Delta\xi_i^n = \xi(t_{i+1}^n) - \xi(t_i^n), \quad \text{etc.}$$

Since $\Delta\xi_i^n$ and $\Delta\eta_i^n$ are increments of Brownian motion, the transition densities are:

$$\begin{aligned}
P(x_{i+1}^n \mid x_i^n, z_i^n) &= (2\pi\sigma^2(x_i^n, t_i^n)\,\Delta t^n)^{-1/2}\\
&\quad\cdot \exp\left\{-\frac{(x_{i+1}^n - x_i^n - m(x_i^n, t_i^n)\,\Delta t^n)^2}{2\sigma^2(x_i^n, t_i^n)\,\Delta t^n}\right\},\\
P(z_{i+1}^n \mid x_i^n, z_i^n) &= (2\pi f^2(z_i^n, t_i^n)\,\Delta t^n)^{-1/2}\\
&\quad\cdot \exp\left\{-\frac{(z_{i+1}^n - z_i^n - [f(z_i^n, t_i^n)n(x_i^n, t_i^n) + g(z_i^n, t_i^n)]\,\Delta t^n)^2}{2f^2(z_i^n, t_i^n)\,\Delta t^n}\right\}.
\end{aligned}$$

Lemmas 1, 2 and 3 will serve to prove the form of the discrete sequentially best filter up to terms that vanish faster than Δt, i.e., modulo $O(\Delta t)$. The proofs will be by an inductive construction. It is assumed that $f(\cdot,\cdot)$ and $g(\cdot,\cdot)$ have been specified up to time t_{i-1}^n, so the joint distribution of x_i^n and z_i^n are completely determined. Minimizing the error at the next instant, i.e., minimizing

$$E(x^n_{i+1} - z^n_{i+1})^2,$$

is accomplished by the choice of $f(z^n_i, t^n_i)$ and $g(z^n_i, t^n_i)$. Finally, the theorem will be proved by a limiting argument on Δt^n.

LEMMA 1. *The sequentially best (discrete case) estimator has the property*

$$E(x^n_i \mid z^n_i) = z^n_i. \tag{3.3}$$

Proof. The operation on z^n_i that minimizes

$$E[(x^n_{i+1} - z^n_{i+1})^2 \mid z^n_i] \tag{3.4}$$

will be

$$z^n_{i+1} = E(x^n_{i+1} \mid z_i), \tag{3.5}$$

from which Lemma 1 follows. (The fact that (3.5) implies (3.3) is proved in [9, p. 350].)

Ordinarily results that require a one line proof are not relegated to separate lemmas, but (3.3) has special significance. Its analogue in the continuous case is

$$z(t) = E(x(t) \mid z(t)). \tag{3.6}$$

Relation (3.6) provides the compatibility between the essentials of $z(t)$ being both the estimate and a functional of the observation on which the estimate is based.

LEMMA 2. *The sequentially best (discrete case) estimator has the property*

$$g(z^n_i, t^n_i) = \bar{m}(z^n_i) - f(z^n_i, t^n_i)\bar{n}(z^n_i),$$

where

$$\bar{m}(z^n_i) = \int_{-\infty}^{\infty} m(x^n_i, t^n_i)P(x^n_i \mid z^n_i)\, dx^n_i$$

and

$$\bar{n}(z^n_i) = \int_{-\infty}^{\infty} n(x^n_i, t^n_i)P(x^n_i \mid z^n_i)\, dx^n_i.$$

Proof. The requirement that $E(x^n_{i+1} - z^n_{i+1} \mid z^n_i) = 0$ may be written:

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x^n_{i+1} - z^n_{i+1})P(x^n_{i+1}, z^n_{i+1} \mid z^n_i)\, dx^n_{i+1}\, dz^n_{i+1} \\
&\quad = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x^n_{i+1} - z^n_{i+1})P(x^n_{i+1}, z^n_{i+1} \mid x^n_i, z^n_i)P(x^n_i \mid z^n_i)\, dx^n_{i+1}\, dz^n_{i+1}\, dx^n_i \\
&\quad = \int_{-\infty}^{\infty} [x^n_i + m(x^n_i, t^n_i)\,\Delta t^n - z^n_i - f(z^n_i, t^n_i)n(x^n_i, t^n_i)\,\Delta t^n - g(z^n_i, t^n_i)\,\Delta t^n] \\
&\qquad \cdot P(x^n_i \mid z^n_i)\, dx^n_i \\
&\quad = \bar{m}(z^n_i)\,\Delta t^n - f(z^n_i, t^n_i)\bar{n}(z^n_i)\,\Delta t^n - g(z^n_i, t^n_i)\,\Delta t^n = 0,
\end{aligned}$$

because

$$\int (x_i^n - z_i^n) P(x_i^n \mid z_i^n)\, dx_i^n = 0 \tag{3.7}$$

by Lemma 1 and the fact that the filter has minimized the error at time t_i^n. This proves Lemma 2.

It is interesting to note that if one ignores bias and tries to reduce error by manipulating $g(\cdot,\cdot)$ and $f(\cdot,\cdot)$ independently, the result is that $g(z_i^n, t_i^n)$ comes out to be proportional to $(\Delta t^n)^{-1}$. In the limit, as $\Delta t^n \to 0$, $g(\cdot,\cdot) = \pm\infty$, an unacceptable answer but not a surprising one. What the mathematics is saying is: If the new data causes one to decide the estimate is low, the drift coefficient $g(\cdot,\cdot)$ should be assigned the value that will raise the expectation of the new estimate the fastest. Of course there is no fastest and the derivation leads to a nonsense answer. It is analogous to control problems with a bounded set of controls where the answer is "bang-bang" or one of two extremal points is always optimal.

LEMMA 3. *The sequentially best* (*discrete time*) *estimator has the property*

$$f(z_i^n, t_i^n) = \overline{nx}(z_i^n) - z_i^n \bar{n}(z_i^n) + O(\Delta t^n),$$

where

$$\overline{nx}(z_i^n, t_i^n) = \int_{-\infty}^{\infty} n(x_i^n, t_i^n) x_i^n P(x_i^n \mid z_i^n)\, dx_i^n.$$

Proof. Minimizing $E[(z_{i+1}^n - x_{i+1}^n)^2 \mid z_i^n]$ is equivalent to minimizing

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (-2x_{i+1}^n z_{i+1}^n + (z_{i+1}^n)^2) P(x_{i+1}^n, z_{i+1}^n \mid z_i^n)\, dx_{i+1}\, dz_{i+1}. \tag{3.8}$$

With an expansion of the density function and integration over x_{i+1}^n and z_{i+1}^n, (3.8) becomes:

$$\int_{-\infty}^{\infty} [-2(x_i^n + m(x_i^n, t_i^n)\,\Delta t^n)(z_i^n + f(z_i^n, t_i^n) n(x_i^n, t_i^n)\,\Delta t^n + g(z_i^n, t_i^n)\,\Delta t^n)$$
$$+ (z_i^n + f(z_i^n, t_i^n) n(x_i^n, t_i^n)\,\Delta t^n + g(z_i^n, t_i^n)\,\Delta t^n)^2 + f^2(z_i^n, t_i^n)\,\Delta t^n] P(x_i^n \mid z_i^n)\, dx_i^n.$$

A regrouping of terms yields:

$$\int_{-\infty}^{\infty} (-2x_i^n z_i^n + (z_i^n)^2) P(x_i^n \mid z_i^n)\, dx_i^n$$
$$+ \Delta t^n \int_{-\infty}^{\infty} (f^2(z_i^n, t_i^n) - 2n(x_i^n, t_i^n) x_i^n f(z_i^n, t_i^n)$$
$$+ 2n(x_i^n, t_i^n) z_i^n f(z_i^n, t_i^n)) P(x_i^n \mid z_i^n)\, dx_i^n$$
$$+ \Delta t^n \int_{-\infty}^{\infty} (2z_i^n g(z_i^n, t_i^n) - 2x_i^n g(z_i^n, t_i^n)) P(x_i^n \mid z_i^n)\, dx_i^n + O(\Delta t^n).$$

The first integral is independent of $f(z_i^n, t_i^n)$ and $g(z_i^n, t_i^n)$ and the last integral is zero by Lemma 1. Clearly the second integral is minimized when $f(z_i^n, t_i^n) = \overline{nx}(z_i^n) - \bar{n}(z_i^n) z_i^n$, which proves Lemma 3.

Proof of theorem. The theorem is almost obvious in light of Lemmas 2 and 3. If two estimators agree up to time t (they necessarily agree at $t = 0$) and then differ over an interval $(t, t + \Delta)$, there must be a discrete approximation such that $t_i^n = t$ and $t_{i+1}^n \in (t, t + \Delta)$. Since the algorithm for the discrete case defines $f(\cdot, \cdot)$ and $g(\cdot, \cdot)$ uniquely up to sets of measure zero, the filter that does not satisfy (2.1) and (2.2) (the same as not satisfying Lemmas 2 and 3) must have a larger error and is therefore not sequentially best. This proves the theorem.

To prove that a unique sequentially best estimator exists is to prove that the simultaneous solution of equations (1.6), (2.1) and (2.2) exists and is unique. Any restrictions on the problem will necessarily be expressed in terms of $m(\cdot, \cdot)$ and $n(\cdot, \cdot)$ and $\sigma(\cdot, \cdot)$, the only unspecified quantities. Although suitable conditions on these coefficients to insure the existence of a unique set of solutions $f(\cdot, \cdot)$ and $g(\cdot, \cdot)$ to (1.6), (2.1) and (2.2) have not been discovered at this time, examples of the next section, expecially the linear case, show that the preceding theorem cannot be vacuous. Indeed, it is suspected that it is rather widely applicable. A rigorous proof of existence of a solution would be truly interesting, but will not be pursued here.

4. Examples. If (1.1) and (1.2) are linear, then the model conforms to the case solved by Kalman and Bucy [8]. It is interesting to examine exactly how the algorithm reduces in this instance.

Example 1. First, of course, $\sigma(x(t), t)$, $m(x(t), t)$ and $n(x(t), t)$ must be linear; let $\sigma(x(t), t) = \sigma(t)$, $m(x(t), t) = m(t)x(t)$ and $n(x(t), t) = n(t)x(t)$. By Lemma 1,

$$\begin{aligned} g(z(t), t) &= E(m(t)x(t) \mid z(t)) - f(z(t), t)E(n(t)x(t) \mid z(t)) \\ &= m(t)z(t) - f(z(t), t)n(t)z(t). \end{aligned}$$

Now all the coefficients of (1.6) are specified with the exception of $f(z(t), t)$.

$$\begin{aligned} f(z(t)) &= E[n(t)x(t)x(t) \mid z(t)] - E[n(t)x(t) \mid z(t)]E[x(t) \mid z(t)] \\ &= n(t)E[x^2(t) - (E(x(t) \mid z(t))^2 \mid z(t)], \end{aligned}$$

i.e., $f(z(t), t)$ is proportional to the conditional variance of $x(t)$.

Now consider a Gaussian solution of (1.6). When $P(a, b, t)$ is a bivariate normal distribution, the conditional variance is *not* a function of the conditioning variable, i.e., $f(z(t), t) = f(t)$. Hence all the coefficients of (1.6) are linear and if a solution exists it must be Gaussian.

If the only unknown is $f(t)$, which in the linear case has the property of being the mean-square error at time t, is it still necessary to solve (1.6)? The answer, thanks to Kalman and Bucy, is no. A major contribution of their work is the proof that the evolution of the mean-square error is a Riccati equation. Thus the reduction is complete. The filter, $f(t)$ and $g(z(t), t) = m(t)z(t) - f(t)n(t)z(t) = g(t)z(t)$, does indeed coincide exactly with the Kalman filter.

In lieu of an analytic example of the algorithm for the sequentially best estimator in a nonlinear situation, a numerical solution of the simultaneous equations (1.6), (2.1) and (2.2) was undertaken.

Example 2. In the numerical solution of the algorithm a certain instability is inherent. Equation (1.6) is replaced by a difference equation over quantized time and space; (2.1) and (2.2) are replaced by relations among summations over the space variables. The evolution of the filter coefficients $g(b_i, t_j)$ and $f(b_i, t_j)$ is achieved by using $P(a_1, b_i, t_j)$ to calculate $g(b_i, t_j)$ and $f(b_i, t_j)$ which in turn are used to update the density or find $P(a_i, b_i, t_{j+1})$. Thus errors are cumulative and the error introduced by the truncation of the domain of $P(\cdot, \cdot, \cdot)$ bounds the number of iterations before the density blows up and future calculations are meaningless.

In order to get to an interesting region of operation the linear time-invariant case is assumed to have started at $t = -\infty$; thus at $t = 0$ the state $x(0)$ and estimator $z(0)$ have reached a stationary bivariate normal distribution. At $t = 0$ nonlinearities are introduced through $m(a, \cdot)$ and $n(a, \cdot)$. The resulting growth of error is computed both when the original linear filter is retained and when the filter is modified by the sequential criterion.

The numbers used in the calculations indicate only relative magnitudes, so of course are dimensionless. To facilitate the computations the following were assumed:

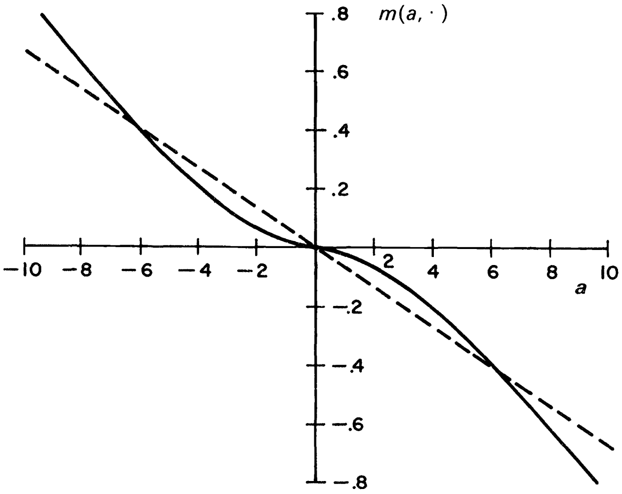
$$m(a,t) = -0.1a, \quad n(a,t) = a \quad \text{and} \quad \sigma(a,t) = 1.$$

By the results of Kalman and Bucy it is straightforward to find the optimal steady state $f(b, \cdot)$ and $g(b, \cdot)$ for this case. Implementation of the above coefficients results in an error of .93, i.e., when the above constants are put into (1.6) the resulting normal density yields $E(x(t) - z(t))^2 = .93$.

Incrementing began at $t = 0$ with $\Delta t = 0.1$ seconds. With everything linear the error was maintained almost constant for seven iterations while the optimal $f(\cdot, t)$ and $g(\cdot, t)$ for $t = 0, .1, .2, \cdots, .7$ were generated and stored on magnetic tape.

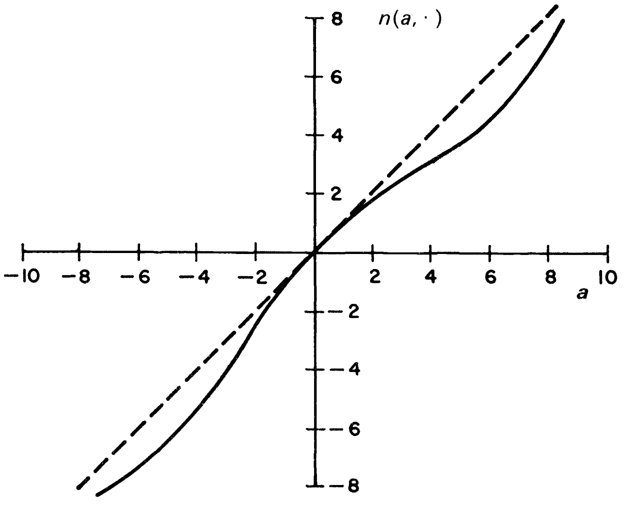
Having carried out the above to test the stability as well as generating the "computerized version" of the linear filter, we imposed nonlinearities on the system. Coefficients $m(b, \cdot)$ and $n(b, \cdot)$ were changed to the solid line graphs shown in Figs. 2 and 3. The resulting $f(b, \cdot)$ and $g(b, \cdot)$ are shown in Figs. 4 and 5. The change of coefficients was accompanied by a growth of mean-square error. The errors are compared in Fig. 6. Lastly, the evolution of the density was recomputed with the nonlinear $m(\cdot, \cdot)$ and $n(\cdot, \cdot)$. This time instead of calculating $f(\cdot, \cdot)$ and $g(\cdot, \cdot)$ each time around the loop, the functions generated in the linear case were read in from the tape. This simulated filtration of the nonlinear system using the best Kalman or linear filters. The resulting growth of error also appears in Fig. 6. The fact that the sequentially best scheme had a smaller error than the linear filter on the nonlinear system indicates the power of the "sequentially best" estimation scheme.

5. Time-invariant case. Consider the time-invariant case where $m(\cdot, \cdot)$, $\sigma(\cdot, \cdot)$ and $n(\cdot, \cdot)$ are functions of one variable and such that $x(\cdot)$ is a stationary process. It is desired to find a time-invariant filter with the smallest possible error.



.8
$m(a,\cdot)$
.6
.4
.2
−10
−8
−6
−4
−2
2
4
6
8
10
a
−.2
−.4
−.6
−.8

FIG. 2



8
$n(a,\cdot)$
6
4
2
−10
−8
−6
−4
−2
2
4
6
8
10
a
−2
−4
−6
−8

FIG. 3

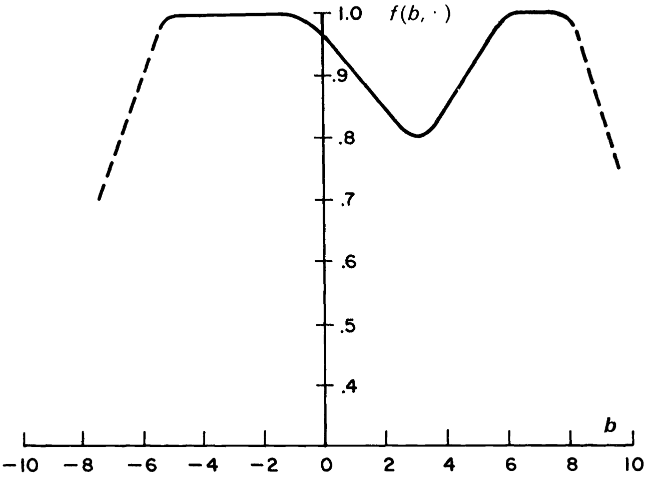

1.0
.9
.8
.7
.6
.5
.4
$f(b, \cdot)$
b
−10
−8
−6
−4
−2
0
2
4
6
8
10

Fig. 4

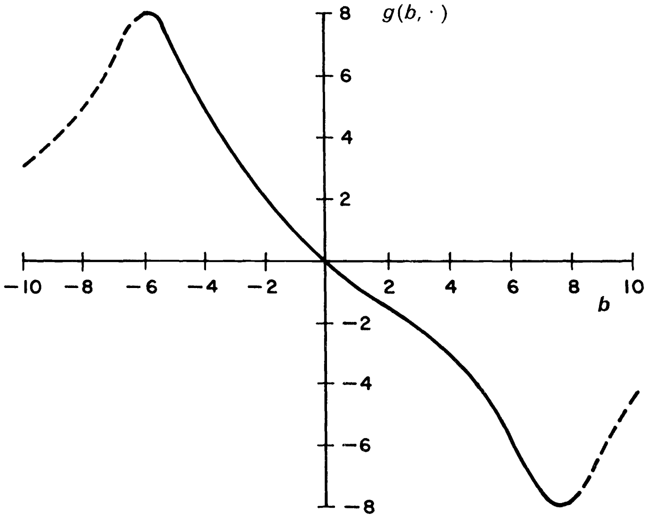

8
6
4
2
−2
−4
−6
−8
$g(b, \cdot)$
b
−10
−8
−6
−4
−2
2
4
6
8
10

Fig. 5

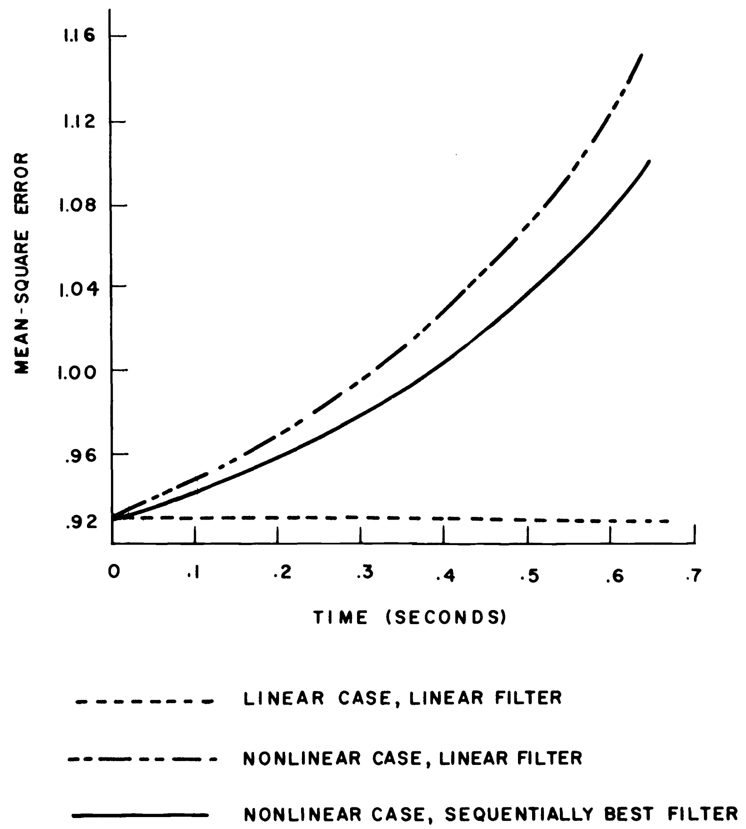


FIG. 6

By suppressing the time dependence, (1.1), (1.3) and (1.6) can be rewritten:

$$
\begin{gathered}
dx = m(x)\,dt + \sigma(x)\,d\xi, \\
dz = g(z)\,dt + f(z)\,dy, \\
-\frac{\partial}{\partial a}[m(a)P(a,b)] - \frac{\partial}{\partial b}[(f(b)n(a) + g(b))P(a,b)] \\
+ \frac{1}{2}\frac{\partial^2}{\partial a^2}[\sigma^2(a)P(a,b)] + \frac{1}{2}\frac{\partial^2}{\partial b^2}[f^2(b)P(a,b)] = 0,
\end{gathered}
\tag{5.1}
$$

where $P(a, b)$ is the joint density for the x and z processes and the left side of (1.6) is identically zero.

Equation (5.1) always has at least one solution because if $g(b)$ and $f(b)$ vanish, the marginal density of the x process satisfies

$$-\frac{\partial}{\partial a}[m(a)P(a,b)] + \frac{1}{2}\frac{\partial^2}{\partial a^2}[\sigma^2(a)P(a,b)] = 0.$$

Now suppose a pair of real-valued functions $g(b)$ and $f(b)$ exist such that the solution of (5.1) defines a two-dimensional process with the property that z is the minimum variance unbiased estimate of x.

Lemma 2 still holds so,

$$g(z) = \bar{m}(z) - f(z)\bar{n}(z), \tag{5.2}$$

where

$$\bar{m}(z) = \int_{-\infty}^{\infty} m(a)P(a \mid z)\,da, \quad \text{etc.}$$

Now if $f(b)$ did *not* satisfy

$$f(z) = \overline{nx}(z) - \bar{n}(z)\bar{x}(z), \tag{5.3}$$

then one could operate on the output of the filter with one satisfying both (5.2) *and* (5.3) to obtain a smaller variance with another time-invariant filter. This argument leads to a contradiction; thus (5.3) must be satisfied.

The implication is that if the state and observation variables have a stationary distribution, then the minimum variance, unbiased, time-invariant filter is given by the simultaneous solution to (5.1), (5.2) and (5.3). Hence the "sequentially best" algorithm yields the uniformly best time-invariant filter.

6. Acknowledgment. The author is extremely indebted to Professor Eugene Wong for his encouragement and counsel throughout the research of the topic and preparation of the manuscript.

REFERENCES

[1] J. L. Doob, *Stochastic Processes*, John Wiley, New York, 1953.

[2] H. J. Kushner, *Dynamical equations for optimal nonlinear filtering*, J. Differential Equations, 2 (1967), pp. 179–190.

[3] R. E. Mortensen, *Optimal control of continuous time stochastic systems*, Rept. ERL-6-1, Electronics Research Laboratory, University of California, Berkeley, 1966.

[4] M. Zakai, *On the optimal filtering of diffusion processes*, E. E. Publ. 80, Israel Institute of Technology, Haifa, Israel, 1967.

[5] F. P. Romeo, *On stochastic differential equations arising in state estimation problems*, Memo. ERL-M246, Electronics Research Laboratory, University of California, Berkeley, 1968.

[6] K. Ito, *Lectures on Stochastic Processes*, Tata Institute for Fundamental Research, Bombay, India, 1961.

[7] A. V. Skorokhod, *Studies in the Theory of Random Processes*, Addison-Wesley, Reading, Massachusetts, 1965.

[8] R. E. Kalman and R. S. Bucy, *New results in linear filtering theory*, J. Basic Engrg. (ASME Trans.), 83D (1961), pp. 95–108.

[9] M. Loève, *Probability Theory*, 3rd ed., Van Nostrand, Princeton, 1963.

SIAM J. CONTROL
Vol. 7, No. 4, November 1969

AN OPTIMAL CONTROL PROBLEM FOR SYSTEMS WITH DIFFERENTIAL-DIFFERENCE EQUATION DYNAMICS*

D. W. ROSS† AND I. FLÜGGE-LOTZ‡

1. Introduction. This paper is concerned with the optimal control of systems in which time delays appear explicitly in the system model. The systems under discussion will be described by constant coefficient, linear differential-difference equations of the type:

$$\dot{x}(t) = Ax(t) + Bx(t-1) + Du(t) \tag{1.1}$$

in which $x(t)$ is an n-dimensional vector, A and B are constant $n \times n$ matrices, $u(t)$ is an r-dimensional vector called the *control* ($r \leqq n$), and D is a constant $n \times r$ matrix. The assumption of a single delay is for convenience; the results for multidelay problems are minor variations of those presented here and are discussed in [1]. Without loss of generality, we assume that the (constant) time delay is normalized to unity.

The differential-difference equation model is applicable in many engineering problems; see [1, Chap. 1], [2], [3], [4, Chap. 8], for examples. Many aspects of the theory of differential-difference equations have been considered by Bellman and Cooke [5].

The time-optimal control of such systems has been discussed by Oguztöreli [6], and more general optimization problems have been considered by Kharatishvili [7, pp. 212–226] and Chyung and Lee [8]. All of these works concerned the derivation of necessary conditions for optimality. In each case, the necessary conditions were presented in the form of a maximum principle.

The problem considered here is more specific than those of references [6]–[8]; it is to investigate conditions to be satisfied by the control $u(t)$, $t \geqq 0$, for the performance criterion

$$J[\varphi, u] = \int_0^\infty [x'(t)Qx(t) + u'(t)Ru(t)]\,dt \tag{1.2}$$

to be minimized.

Here Q and R are positive definite matrices (of conformable dimensions) and φ is the initial condition for (1.1) which is taken to be a collection of n continuous functions (one for each component of x) defined on the interval $[-1, 0]$. Primes will be used throughout this paper to indicate the transpose of vectors or matrices.

This problem has been studied previously by Krasovskii [9], [10] using a sufficient condition for optimality. This approach led to the conclusion that, under certain conditions, a linear control law is optimal for criterion (1.2). How-

* Received by the editors January 9, 1968, and in final revised form January 9, 1969. This research was supported by the Air Force Office of Scientific Research, Office of Aerospace Research, under Contract AF 49(638)1431.

† Information and Control Laboratory, Stanford Research Institute, Menlo Park, California 94024.

‡ Department of Applied Mechanics, Stanford University, Stanford, California 94305.

ever, the resulting analysis did not put in evidence *an explicit characterization of the optimal control* and some key statements remained unproved.

The present paper also addresses the problem using sufficient conditions for optimality, but the emphasis is upon explicit conditions for optimality in the style of Kalman's [12] algebraic Riccati equation characterization of optimal controls for a similar problem without time delay. In this paper we present sufficient conditions for a linear control law to be optimal for criterion (1.2) and we derive a *specific* set of equations to be satisfied by the (optimal) control law parameters. These equations play the same role in our problem as the Riccati equation derived by Kalman [11], [12] does in characterizing the optimal control for criterion (1.2) when $B = 0$ in (1.1). In fact, Kalman's solution [11], [12] is obtained as a special solution of the equations we derive, by simply letting $B = 0$ in our equations.

Sections 2 and 3 of this paper will deal with preliminaries and the discussion of underlying concepts. A statement of the main results (Theorem 4.1) of this investigation will be given in § 4. Section 5 is devoted to the presentation of sufficient conditions for optimality. In § 6 we use this result in deriving our main results (Theorem 4.1). Finally, computational aspects of the problem are treated in § 7.

2. Basic definitions and concepts. Let us first introduce some notation.

$C = C([-1,0], E^n)$ will denote the Banach space of continuous functions with domain $[-1,0]$ and range in E^n. The norm of this space is defined as $\|\varphi\| = \max_{-1 \leqq \theta \leqq 0} |\varphi(\theta)|$, where $|\varphi(\theta)|$ is the usual Euclidean norm of φ in E^n (at time θ). Occasionally $|s|$ will denote the absolute value of the scalar, s; the usage should be clear from context.

For $t \geqq 0$, let x_t denote the element of C defined by $x_t = x(t + \theta)$ for $-1 \leqq \theta \leqq 0$; in other words, x_t denotes the segment of the trajectory of (1.1) on a time interval of length equal to the time delay, prior to time t. If x_{t_1} and $u(t)$ (for $t \geqq t_1$) are specified, then solutions of (1.1) are uniquely determined. Since the quantity, x_t, is the *minimal* set of data which (in conjunction with knowledge of the future control, $u(t)$) allows one to completely determine the evolution of system (1.1) it is called the *state* of the system (1.1) and the corresponding *state space* will be C.

The symbol $x(t)$ will denote the value, in E^n, of x at time t (the reader is strongly urged to note the difference in meanings of the symbols $x(t)$ and x_t and the corresponding norms). A more specific symbol, $x(t; u, \varphi)$, will denote the value, in E^n, of x at time t when the control function is $u(\cdot)$, the initial state is φ, and the initial time is $t_0 = 0$.

2.1. Functional-differential equations. If the control in (1.1) is taken to be an explicit function of the state, i.e., $u(t) = u(x_t)$, then the differential-difference equation becomes an (autonomous) *functional-differential equation*:

$$\dot{x}(t) = f(x(t + \theta)), \quad -1 \leqq \theta \leqq 0, \quad t \geqq t_0 \tag{2.1}$$

(with $x(t_0 + \theta) = \varphi(\theta)$ for $-1 \leqq \theta \leqq 0$).

Solutions to (2.1) exist and are unique if the functional $f(x(\theta))$ satisfies certain conditions. For instance, if $f(x(\theta))$ satisfies a uniform Lipschitz condition, i.e.,

$$|f(x(\theta)) - f(y(\theta))| < L\|x - y\|$$

for some constant, L, then solutions to (2.1) exist and are unique for $t \geqq t_0$ (see [13, pp. 127–128]). Thus solutions to (1.1) will be well-defined if $u(t)$ is a functional, $u(x_t)$, which satisfies a uniform Lipschitz condition.

Let us now assume that the control, $u(t)$, in (1.1) is a functional for which (1.1) becomes an equation of type (2.1) with well-defined solutions. The following stability concepts (see [14] for a more complete discussion) will refer to system (2.1).

DEFINITION 2.1. The null solution ($x \equiv 0$) of (2.1) is said to be *stable* if for any $\varepsilon > 0$, there is a $\delta > 0$ such that if $\|\varphi\| < \delta$, then $\|x_t\| < \varepsilon$ for all $t \geqq 0$.

DEFINITION 2.2. If in addition to being stable we have $\|x_t\| \to 0$ as $t \to \infty$, then $x \equiv 0$ is said to be *asymptotically stable* (and if all solutions approach zero as $t \to \infty$ then $x \equiv 0$ is said to be *globally* asymptotically stable).

3. Admissible controls. Our problem suggests the following question: If the initial state, φ, is an arbitrary function in the state space C, is there a *control function* $u(t)$, for $t \geqq 0$, such that $J[\varphi, u] < \infty$? Of course, the problem posed by (1.1) is completely meaningless if such is not the case for any φ.

As mentioned in § 2, the quantity which describes the state of the system (1.1) and completely determines (for known $u(t)$) its evolution in the future ($t \geqq t_0$) is the segment, $x(t_0 + \theta)$, for $-1 \leqq \theta \leqq 0$, of the system's motion. Thus it is natural to form the control $u(t)$ at each instant of time using the information $x_t = x(t + \theta)$, $-1 \leqq \theta \leqq 0$, on the preceding time interval, $[t - 1, t]$. In fact, it follows from Bellman's *principle of optimality* that an optimal control (if one exists) for our problem is a (time-invariant) functional, $u(t) = u(x_t)$, which maps C into E^r. This observation motivates a more restricted question than asked in the previous paragraph, namely: Is there a *control law*, $u(x_t)$, which yields $J[\varphi, u] < \infty$ for arbitrary φ?

From the design viewpoint, there is another point to consider. If the answer to this last question is affirmative, we would hope that the controlled system is asymptotically stable (hereafter, the term "asymptotically stable" will be used in the global sense).

The preceding paragraphs lead us to the following definition.

DEFINITION 3.1. An *admissible control* for system (1.1) is to satisfy the requirements:

(i) $u(t) = u(x_t)$; in other words, the control at time t is the value, in E^r, of a (time-invariant) functional of the system state, x_t.

(ii) The functional $u(x_t)$ is such that solutions to (1.1) exist and are unique for $t \geqq 0$ and for all initial states, φ. (A *sufficient* condition for this requirement is that $u(x_t)$ be continuous in the topology of space C, and also satisfy a uniform Lipschitz condition.)

(iii) The null solution of (1.1) with control law $u(t) = u(x_t)$ is required to be asymptotically stable.

(iv) The control $u(x_t)$ must yield a finite value for the performance criterion (1.2) for all initial states, φ, from the state space C.

Of course, Definition 3.1 is meaningless unless one can establish conditions for the existence of admissible controls for system (1.1). Necessary and sufficient conditions have recently been established by Osipov [15]; however, those conditions are difficult to apply; in some cases the following simple sufficient condition applies.

PROPOSITION 3.1. *The linear constant system with time delay*

$$\dot{x}(t) = Ax(t) + Bx(t-1) + Du(t) \tag{3.1}$$

has a (linear) admissible control $u(x_t)$ *if*: (i) *the columns of* B *are linear combinations of the columns of* D, (ii) D *has rank* r, *and* (iii) (A, D) *is a completely controllable pair (i.e., the rank of the matrix* $[D|AD|A^2D| \cdots |A^{n-1}D]$ *is* n).

Proof. We first remark that $u(x_t)$ is said to be a *linear control law* if

$$u(\alpha x_t + \beta y_t) = \alpha u(x_t) + \beta u(y_t)$$

for all x_t, y_t in C and all scalars α, β.

Using (*i*), let $\hat{u}(t)$ be defined by $D\hat{u}(t) = Du(t) + Bx(t-1)$; then (3.1) becomes:

$$\dot{x}(t) = Ax(t) + D\hat{u}(t). \tag{3.2}$$

Now, it is well known [12] that if system (3.2) is completely controllable there is an $r \times n$ matrix, K, such that if $\hat{u}(t) = Kx(t)$ then the null solution of (3.2) is asymptotically stable.

Using (*ii*) we have $u(t) = \hat{u}(t) - (D'D)^{-1}D'Bx(t-1)$; consequently the linear law

$$u(t) = Kx(t) - (D'D)^{-1}D'Bx(t-1)$$

causes the null solution of (3.1) to be asymptotically stable. This completes the proof.

Clearly, any nth-order differential-difference equation in a scalar variable and scalar control of the form

$$\frac{d^n y}{dt^n} + \frac{\alpha_{n-1} d^{n-1} y}{dt^{n-1}} + \cdots + \alpha_0 y(t)$$
$$+ \frac{\beta_{n-1} d^{n-1} y(t-1)}{dt^{n-1}} + \cdots + \beta_0 y(t-1) = u(t)$$

meets the conditions of Proposition 3.1, for in that case we may take A, B, D to be

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdot & \cdot & 0 \\ 0 & 0 & 1 & 0 & \cdot & 0 \\ 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & \cdot & & & 1 & 0 \\ 0 & \cdot & & & 0 & 1 \\ -\alpha_0 & -\alpha_1 & \cdot & \cdot & \cdot & -\alpha_{n-1} \end{bmatrix},$$

$$B = \begin{bmatrix} 0 & 0 & 0 & \cdot & \cdot & 0 \\ 0 & \cdot & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & & & \cdot & \cdot \\ \cdot & \cdot & & & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot & 0 \\ -\beta_0 & -\beta_1 & \cdot & \cdot & \cdot & -\beta_{n-1} \end{bmatrix}, \quad D = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \cdot \\ 0 \\ 1 \end{bmatrix}$$

(here $x_i(t) = dy^{i-1}(t)/dt^{i-1}$ for $i = 1, 2, 3, \cdots, n$).

In summary, under either the conditions of Proposition 3.1, or those given in [15], there are admissible controls for our problem.

4. Main results. The following theorem summarizes the main result of this paper; it is a statement of sufficient conditions for a linear control to be optimal for criterion (1.2).

THEOREM 4.1. *A linear control law*

$$u^0(t) = -R^{-1}D'\Pi_0 x(t) - R^{-1}D'\int_{-1}^{0} \Pi_1(\theta)x(t+\theta)\,d\theta, \qquad t \geqq 0,$$

provides the absolute minimum of criterion (1.2) *for the dynamical system* (1.1) *if*:

(a) $u^0(t)$ *is a stable control law* (*since* u^0 *is linear, stability and admissibility are equivalent*);

(b) Π_0 *is a symmetric positive definite matrix which, together with the* $n \times n$ *array*, $\Pi_1(\theta)$, *of functions defined on* $[-1, 0]$, *and an* $n \times n$ *array*, $\Pi_2(\xi, \theta)$, *of functions in two variables having domain* $-1 \leqq \xi \leqq 0$, $-1 \leqq \theta \leqq 0$, *satisfies the relations*:

$$A'\Pi_0 + \Pi_0 A - \Pi_0 DR^{-1}D'\Pi_0 + \Pi_1'(0) + \Pi_1(0) + Q = 0, \tag{4.1a}$$

$$\frac{d\Pi_1(\theta)}{d\theta} = (A' - \Pi_0 DR^{-1}D')\Pi_1(\theta) + \Pi_2(0,\theta), \qquad -1 \leqq \theta \leqq 0, \tag{4.1b}$$

$$\frac{\partial \Pi_2(\xi,\theta)}{\partial \xi} + \frac{\partial \Pi_2(\xi,\theta)}{\partial \theta} = -\Pi_1'(\xi)DR^{-1}D'\Pi_1(\theta), \qquad -1 \leqq \xi \leqq 0, \quad -1 \leqq \theta \leqq 0, \tag{4.1c}$$

$$\Pi_1(-1) = \Pi_0 B, \tag{4.1d}$$

$$\Pi_2(-1,\theta) = B'\Pi_1(\theta), \qquad -1 \leqq \theta \leqq 0. \tag{4.1e}$$

Furthermore, under these conditions, the corresponding representation of (1.2) *in terms of the initial function is*

$$J[\varphi, u^0] = \varphi'(0)\Pi_0\varphi(0) + 2\varphi'(0)\int_{-1}^{0} \Pi_1(\theta)\varphi(\theta)\,d\theta + \int_{-1}^{0}\int_{-1}^{0} \varphi'(\xi)\Pi_2(\xi,\theta)\varphi(\theta)\,d\xi\,d\theta.$$

The proof of Theorem 4.1 will be deferred until § 6 while necessary intermediate material is developed in § 5.

Assuming that one has solved (4.1) and u^0 is stable, we note that the optimal control

$$u^0(t) = -R^{-1}D'\Pi_0 x(t) - R^{-1}D'\int_{-1}^{0} \Pi_1(\theta)x(t+\theta)\,d\theta$$

generalizes the result for ordinary differential equations (see [12]) in a natural way; it is a linear, constant law which includes compensation for hereditary effects through the term

$$-R^{-1}D'\int_{-1}^{0} \Pi_1(\theta)x(t+\theta)\,d\theta.$$

Also note that if $B = 0$ in (1.1) then a solution to equations (4.1) is

$$\Pi_1(\theta) \equiv 0 \quad \text{for } -1 \leqq \theta \leqq 0, \tag{4.2a}$$

$$\Pi_2(\xi, \theta) \equiv 0 \quad \text{for } -1 \leqq \xi \leqq 0, \quad -1 \leqq \theta \leqq 0, \tag{4.2b}$$

$$\Pi_0 \text{ satisfies: } A'\Pi_0 + \Pi_0 A - \Pi_0 D R^{-1} D' \Pi_0 + Q = 0. \tag{4.2c}$$

Noting that (4.2c) is the algebraic Riccati equation derived by Kalman in [11] and [12] we see that the results include those of the linear optimal control problem for ordinary differential equations.

The result of Theorem 4.1 is particularly interesting since a single calculation which yields the quantities Π_0 and $\Pi_1(\theta)$ is all that is needed to optimally control the system for *any* initial conditions. A convenient computation procedure for this determination of Π_0 and $\Pi_1(\theta)$ is presented in § 7.

5. A sufficient condition for optimality. In this section we present a Hamilton–Jacobi type sufficient condition for optimality which will be used in § 6 for the proof of Theorem 4.1.

Some simplifying notation will be used in the theorem below. Let

$$g(x_t, u(t)) = x'(t)Qx(t) + u'(t)Ru(t)$$

(the integrand in (1.2)), and if $V[x_t]$ is a continuous scalar-valued functional of the state, x_t, of system (1.1) at time t, define

$$\dot{V}_u = \lim_{h \to 0^+} (V[x_{t+h}] - V[x_t])/h.$$

(Whenever the limit exists, the subscript indicates that the limit is evaluated along the motion corresponding to control function $u(t)$.)

Now the sufficient condition is stated.

THEOREM 5.1. *If there is an admissible control $u^0(t) = u^0(x_t)$ and a continuous nonnegative scalar functional $V[x_t]$ (which is zero for $x_t = 0$) which together satisfy the conditions*:

$$\dot{V}_{u^0} + g(x_t, u^0(x_t)) = 0, \qquad t \geqq 0, \tag{5.1}$$

$$\dot{V}_{u^0} + g(x_t, u^0(x_t)) \leqq \dot{V}_u + g(x_t, u(t)), \quad t \geqq 0 \tag{5.2}$$

(*for any (measurable) control function $u(t)$*), *then u^0 provides the minimum of the performance criterion* (1.2) *among all admissible controls. Furthermore,*

$$V[\varphi] = J[\varphi, u^0],$$

the optimal value of the performance criterion.

Proof. (The proof is a slightly different version of Theorem 3.1 of [9].)

Integration of (5.1) from $t = 0$ to $t = \infty$ (which is valid due to the asymptotic stability) gives (if φ is the initial state)

$$V[\varphi] = \int_0^\infty g(x_t, u^0(x_t))\, dt = J[\varphi, u^0].$$

But, from (5.2), if $u(x_t)$ is any other admissible control, then

$$\dot{V}_u \geqq -g(x_t, u(x_t)).$$

Thus, integrating both sides of this inequality (again using the asymptotic stability property of admissible controls) yields:

$$V[\varphi] \leqq \int_0^\infty g(x_t, u(x_t))\,dt = J[\varphi, u].$$

Consequently, $J[\varphi, u^0] \leqq J[\varphi, u]$ and so u^0 is an optimal control among the admissible controls. This completes the proof.

Remark. The reader may recognize that this sufficiency condition is similar to that recently used by Boltyanskii [16, Theorem 3] in connection with dynamic programming.

In this paper we shall seek a control $u^0(x_t)$ and a functional $V[x_t]$ which together satisfy (5.1) and (5.2). We shall proceed as follows:

(a) choose a particular *form* of $u^0(x_t)$;

(b) from that choice, express $J[\varphi, u^0(x_t)]$ as an explicit functional of the initial state, i.e., find V such that $V[\varphi] = J[\varphi, u^0(x_t)]$;

(c) use equations (5.1) and (5.2) as constraints on the parameters of the *assumed form* of $u^0(x_t)$.

This approach reduces the solution of the optimal control problem to a solution of equations in which the unknowns are the parameters of the optimal control law. The same state of affairs occurs in [11], where the solution for the optimal control reduces to the solution of the Riccati equation for the parameters of the optimal law.

6. Proof of the main theorem. Since system (1.1) is linear and performance criterion (1.2) is quadratic, one would intuitively feel that in our problem the optimal control is a *linear functional* of the state, x_t; that is, $u^0(x_t) = K(x_t)$ for $t \geqq 0$, where K is a (constant) linear operator mapping the function x_t from space C into E^r. Of course at this point this is only a conjecture, but in view of the comments at the end of § 5, it motivates us to ask the question: "What is the general representation of $J[\varphi, u(x_t)]$ when $u(x_t)$ is a control from the class of *linear* admissible controls?".

One cannot make *arbitrary* choices of $V[x_t]$ and $u^0(x_t)$ in attempting to satisfy conditions (5.1) and (5.2) of Theorem 5.1. More to the point, if $u^0(x_t)$ is considered a candidate for the optimal control, then the corresponding choice of $V[x_t]$ *must* be $J[x_t, u^0]$ which is evaluated from (1.2) for initial state, x_t. For a given *class* of candidates for the optimal control, $u^0(x_t)$, there is a *corresponding class* of functionals, $V[x_t]$.

For the class of *linear* admissible controls one can show [1] that the functional V is a quadratic functional of the form given in the following proposition.

PROPOSITION 6.1 (see [1] for a proof). *If $u_L = u_L(x_t)$, $t \geqq 0$, is a linear admissible control which satisfies a uniform Lipschitz condition, and φ is an arbitrary initial function in space C, then the functional*

$$V[\varphi] = J[\varphi, u_L] = \int_0^\infty [x'(t)Qx(t) + u_L'(x_t)Ru_L(x_t)]\,dt$$

has the quadratic representation:

$$V[\varphi] = \varphi'(0)\Pi_0\varphi(0) + \varphi'(0)\int_{-1}^{0} \Pi_1(\theta)\varphi(\theta)\,d\theta + \left(\int_{-1}^{0} \varphi'(\theta)\Pi_1'(\theta)\,d\theta\right)\varphi(0)$$

$$(6.1) \qquad + \int_{-1}^{0}\int_{-1}^{0} \varphi'(\xi)\Pi_2(\xi,\theta)\varphi(\theta)\,d\xi\,d\theta$$

in which

(i) Π_0 *is a symmetric* $n \times n$ *positive definite matrix*;

(ii) $\Pi_1(\theta)$ *is an* $n \times n$ *array of functions continuous on the interval* $[-1, 0]$;

(iii) $\Pi_2(\xi, \theta)$ *is an* $n \times n$ *array of continuous functions of two arguments defined on the square*: $-1 \leqq \xi \leqq 0$, $-1 \leqq \theta \leqq 0$. *Also* $\Pi_2'(\xi,\theta) = \Pi_2(\theta,\xi)$.

The essence of the proof of Proposition 6.1 is to construct the functional $V[\varphi]$ in a manner similar to the classical Riesz representation theorem of functional analysis, using the linearity (linear, because $u_L(x_t)$ is linear by assumption) of (1.1) (now a functional-differential equation). One first approximates the given initial function, φ, by a finite sum of "step functions" (see the proof given in [1]). This approximation is denoted by $\tilde{\varphi}$. One can express $V[\tilde{\varphi}]$ as the sum of a finite number of terms. This summation is recognized to be a Stieltjes summation. As the approximation, $\tilde{\varphi}$, approaches φ then $V[\tilde{\varphi}] \to V[\varphi]$. In the limit, the expression for $V[\varphi]$ becomes a quadratic functional represented in Stieltjes integral form. Then one can show that the integrals are also defined in the Riemann sense. This Riemann representation involves Π_0, $\Pi_1(\theta)$, and $\Pi_2(\xi, \theta)$.

In the following, we establish conditions for which a *linear* admissible control and its corresponding quadratic function, $V[\varphi]$, of the form (6.1) together satisfy the sufficient conditions for optimality given by (5.1) and (5.2). More to the point, we specify a trio of quantities Π_0, $\Pi_1(\theta)$ and $\Pi_2(\xi,\theta)$ from which both the optimal control and the optimal performance criterion representation can be determined. Explicit equations are given for Π_0, Π_1 and Π_2.

For any $t \geqq 0$, let

$$V[x_t] = x'(t)\Pi_0 x(t) + x'(t)\int_{-1}^{0} \Pi_1(\theta)x(t+\theta)\,d\theta$$

$$(6.2) \qquad + \left(\int_{-1}^{0} x'(t+\theta)\Pi_1'(\theta)\,d\theta\right)x(t)$$

$$+ \int_{-1}^{0}\int_{-1}^{0} x'(t+\xi)\Pi_2(\xi,\theta)x(t+\theta)\,d\xi\,d\theta$$

be a (nonnegative) quadratic functional of the state at time t along motions of the system

$$\dot{x}(t) = Ax(t) + Bx(t-1) + Du(t), \qquad x_0 = \varphi.$$

Now, by letting $u(t)$ be any (measurable) control function ($t \geqq 0$) the calculation (as in (5.1), (5.2)) of the quantity

$$\dot{V}_u + x'(t)Qx(t) + u'(t)Ru(t)$$

when V is the functional of (6.2) yields:

$$\begin{aligned}
&x'(t)[\Pi_0 A + A'\Pi_0 + Q]x(t) + 2x'(t)\Pi_0 Bx(t-1) + 2x'(t)A' \\
&\quad\cdot \int_{-1}^{0} \Pi_1(\theta)x(t+\theta)\,d\theta + 2x'(t-1)B'\int_{-1}^{0} \Pi_1(\theta)x(t+\theta)\,d\theta \\
&\quad + x'(t)\int_{-1}^{0} \Pi_1(\theta)\frac{dx(t+\theta)}{d\theta}\,d\theta + \left(\int_{-1}^{0} \frac{dx'(t+\theta)}{d\theta}\Pi_1'(\theta)\,d\theta\right)x(t) \\
&\quad + \int_{-1}^{0}\int_{-1}^{0} \frac{dx'(t+\xi)}{d\xi}\Pi_2(\xi,\theta)x(t+\theta)\,d\xi\,d\theta \\
&\quad + \int_{-1}^{0}\int_{-1}^{0} x'(t+\xi)\Pi_2(\xi,\theta)\frac{dx(t+\theta)}{d\theta}\,d\xi\,d\theta + 2x'(t)\Pi_0 Du(t) \\
&\quad + 2u'(t)D'\int_{-1}^{0} \Pi_1(\theta)x(t+\theta)\,d\theta + u'(t)Ru(t).
\end{aligned} \tag{6.3}$$

If (6.3) is minimized with respect to u, as is required in (5.2), it is clear that there is a unique u which minimizes (6.3), namely:

$$u^*(t) = -R^{-1}D'\Pi_0 x(t) - R^{-1}D'\int_{-1}^{0} \Pi_1(\theta)x(t+\theta)\,d\theta. \tag{6.4}$$

Moreover, this control is linear and therefore compatible with the assumption of a quadratic $V[x_t]$.

Next we substitute (6.4) into the expression (6.3) and require this to vanish for all $t \geqq 0$, and, in particular, it must vanish for $t = 0$ and any continuously differentiable initial state, φ. After some manipulation, including some integration by parts (assuming $\Pi_1(\theta)$ and $\Pi_2(\xi,\theta)$ have continuous first partials on their respective domains) we obtain:

$$\begin{aligned}
&\varphi'(0)[A'\Pi_0 + \Pi_0 A - \Pi_0 DR^{-1}D'\Pi_0 + \Pi_1'(0) + \Pi_1(0) + Q]\varphi(0) \\
&\quad + 2\varphi'(0)[\Pi_0 B - \Pi_1(-1)]\varphi(-1) \\
&\quad + 2\varphi'(0)\int_{-1}^{0}\left[\frac{-d\Pi_1(\theta)}{\mathrm{d}\theta} + (A' - \Pi_0 DR^{-1}D')\Pi_1(\theta) + \Pi_2(0,\theta)\right]\varphi(\theta)\,d\theta \\
&\quad + 2\varphi'(-1)\int_{-1}^{0} [B'\Pi_1(\theta) - \Pi_2(-1,\theta)]\varphi(\theta)\,d\theta \\
&\quad + \int_{-1}^{0}\int_{-1}^{0} \varphi'(\xi)\left[-\Pi_1'(\xi)DR^{-1}D'\Pi_1(\theta) - \frac{\partial\Pi_2(\xi,\theta)}{\partial\xi} - \frac{\partial\Pi_2(\xi,\theta)}{\partial\theta}\right] \\
&\qquad\qquad \cdot \varphi(\theta)\,d\xi\,d\theta = 0.
\end{aligned} \tag{6.5}$$

Since φ is arbitrary, (6.5) vanishes if and only if Π_0, Π_1 and Π_2 satisfy:

$$A'\Pi_0 + \Pi_0 A - \Pi_0 DR^{-1}D'\Pi_0 + \Pi_1'(0) + \Pi_1(0) + Q = 0, \tag{6.6a}$$

$$\frac{d\Pi_1(\theta)}{d\theta} = (A' - \Pi_0 DR^{-1}D')\Pi_1(\theta) + \Pi_2(0,\theta), \qquad -1 \leqq \theta \leqq 0, \tag{6.6b}$$

$$\text{(6.6c)} \qquad \frac{\partial \Pi_2(\xi,\theta)}{\partial \xi} + \frac{\partial \Pi_2(\xi,\theta)}{\partial \theta} = -\Pi_1'(\xi)DR^{-1}D'\Pi_1(\theta),$$

$$-1 \leqq \xi \leqq 0, \quad -1 \leqq \theta \leqq 0,$$

$$\text{(6.6d)} \qquad \Pi_1(-1) = \Pi_0 B,$$

$$\text{(6.6e)} \qquad \Pi_2(-1,\theta) = B'\Pi_1(\theta), \qquad -1 \leqq \theta \leqq 0.$$

The above conditions are thus necessary in order that (6.3) vanish for $t \geqq 0$ when the control is (6.4). However, these conditions are clearly also *sufficient* for (6.3) to vanish for all $t \geqq 0$.

In summary, conditions (6.6) are necessary and sufficient conditions for the control $u^*(t)$ of (6.4) to minimize expression (6.3) and to make it vanish as required in Theorem 5.1.

There is only one additional condition which $u^*(t)$ must satisfy before Theorem 4.1 is proved. This condition is that $u^*(t)$ must be admissible, in other words, the stability statement (iii) of Definition 3.1 must be satisfied. If the control law $u^*(t)$ of (6.4) formed using a solution for Π_0 and Π_1 obtained from (6.6) is *stable*, then it follows from Theorem 5.1 that $u^*(t)$ is *also optimal*. This is the statement of Theorem 4.1, and so that theorem has been proved.

The results of Theorem 4.1 are difficult to apply *rigorously* to the design of the optimal control law for system (1.1) for two reasons. First, there are no known algorithmic ways in which the stability of control law $u^*(t)$ can be assessed; secondly, only "approximate" methods for solution of the optimality equations (4.1) exist. The "approximate" method discussed in § 7 is characterized by replacing the derivatives in (6.6) by difference quotients. In this manner a control law $\tilde{u}^*(t)$ can be found. The use of difference quotients instead of derivatives necessarily obliges the user of this method to check whether diminishing the step size or mesh size yields convergence of $\tilde{u}^*(t)$ towards a well-defined limit.

In practice, Theorem 4.1 is applied in the following manner: (a) first, one solves (numerically, as described in § 7) equations (6.6) for the parameters Π_0 and Π_1 that determine a control $u^*(t)$ given by (6.4); (b) then one "tests" the stability of the "closed loop" system via simulation for a variety of initial conditions; (c) if only stable behavior is observed, it is probably safe to conclude that $u^*(t)$ is stable, and therefore, by Theorem 4.1, it is optimal. The computational experience of the authors with some practical examples has shown that the solution of the optimality equations (6.6) by the approximate method of § 7 results in a control law that indeed yields stable behavior.

The authors remark that it is possible to derive (6.6a)–(6.6e) from an infinite-dimensional Riccati equation similarly to the development of Falb and Kleinman [17]. One parallels their development using the inner product defined by

$$\langle x_t, y_t \rangle = x'(t)y(t) + \int_{-1}^{0} x'(t+\theta)y(t+\theta)\,d\theta$$

and by defining the linear operator $\Pi(t)$ (see [17]) in terms of Π_0, Π_1 and Π_2 as follows:[1]

$$\Pi(t) * x_t(\theta) = \Pi_0(t)x(t) + 2\int_{-1}^{0} \Pi_1(t,\sigma)x(t+\sigma)\,d\sigma \qquad \text{for } \theta = 0,$$

$$= \int_{-1}^{0} \Pi_2(t;\theta,\sigma)x(t+\sigma)\,d\sigma \qquad \text{for } -1 \leqq \theta < 0.$$

However, this alternate derivation of (6.6) is nonrigorous due to topological difficulties (the space C is not a Hilbert space using the above inner product).

7. Computation of the optimal control. By Theorem 4.1, the solution of the optimality equations (4.1) provides the solution of the given optimal control problem (providing the resulting control law is stable).

The existence of an exact solution to these equations is discussed in the Appendix; here we offer an "approximate" solution in this sense: if one replaces (4.1b), (4.1c) by corresponding finite difference formulas (and requires condition (4.1e) to hold at discrete times in the interval $-1 \leqq \theta \leqq 0$), then the (difference) solution exists and is in a one-to-one correspondence with the solution of a finite-dimensional optimization problem which is uniquely solvable.

Letting m be a positive integer, partition the interval $-1 \leqq \theta \leqq 0$ into equal segments whose endpoints are $\theta = -i/m$, $0 \leqq i \leqq m$. Also, for this same value of m, partition the unit square $-1 \leqq \xi \leqq 0$, $-1 \leqq \theta \leqq 0$ into a grid of smaller squares whose vertices are $(\xi, \theta) = (-i/m, -j/m)$ for $0 \leqq i \leqq m$ and $0 \leqq j \leqq m$. Then replace (4.1) with:

$$A'\Pi_0 + \Pi_0 A - \Pi_0 DR^{-1}D'\Pi_0 + \Pi_1'(0) + \Pi_1(0) + Q = 0, \tag{7.1a}$$

$$\begin{aligned}\frac{\Pi_1(-(i-1)/m) - \Pi_1(-i/m)}{1/m} &= (A' - \Pi_0 DR^{-1}D')\Pi_1(-(i-1)/m)\\ &\quad + \Pi_2(0, -(i-1)/m) \qquad \text{for } 1 \leqq i \leqq m,\end{aligned} \tag{7.1b}$$

$$\begin{aligned}&\frac{\Pi_2(-(i-1)/m, -(j-1)/m) - \Pi_2(-i/m, -(j-1)/m)}{1/m}\\ &+ \frac{\Pi_2(-(i-1)/m, -(j-1)/m) - \Pi_2(-(i-1)/m, -j/m)}{1/m}\\ &= -\Pi_1'(-(i-1)/m)DR^{-1}D'\Pi_1(-(j-1)/m)\\ &\qquad \text{for } 1 \leqq i \leqq m, \quad 1 \leqq j \leqq m,\end{aligned} \tag{7.1c}$$

$$\Pi_1(-1) = \Pi_0 B, \tag{7.1d}$$

$$\Pi_2(-1, -j/m) = B'\Pi_1(-j/m), \qquad 0 \leqq j \leqq m. \tag{7.1e}$$

Let A, B, D, Q, R be the matrices appearing in (1.1) and (1.2), and for this same value of the integer, m, let P be (the) positive semidefinite solution of the algebraic Riccati equation

$$F_m'P + PF_m - PD_mR^{-1}D_m'P + Q_m = 0 \tag{7.2}$$

[1] The symbol $\Pi(t) * x_t(\theta)$ represents the linear operator $\Pi(t)$ operating on the function $x_t(\theta)$.

in which F_m, Q_m and D_m are defined by

$$F_m = \begin{bmatrix} A & 0 & 0 & \cdot & \cdot & B \\ mI & -mI & 0 & 0 & \cdot & 0 \\ 0 & mI & -mI & 0 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & \cdot & \cdot & 0 & mI & -mI \end{bmatrix}, \tag{7.3}$$

$$Q_m = \begin{bmatrix} Q & 0 & 0 & \cdot & \cdot & 0 \\ 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & \cdot & \cdot & \cdot & \cdot & 0 \\ \cdot & & & & & \cdot \\ \cdot & & & & & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot & 0 \end{bmatrix}, \quad D_m = \begin{bmatrix} D \\ 0 \\ \cdot \\ \cdot \\ 0 \end{bmatrix}$$

(I is the $n \times n$ identity matrix, F_m and Q_m are square matrices having $n \times (m + 1)$ rows and columns, D_m is $n \times (m + 1)$ by r).

If one considers the (symmetric) matrix P to consist of $(m + 1)^2$ subblocks, $P_{i,j}$ for $0 \leqq i \leqq m$ and $0 \leqq j \leqq m$, of size $n \times n$:

$$P = \begin{bmatrix} P_{0,0} & P_{0,1} & \cdot & \cdot & P_{0,m} \\ P_{1,0} & P_{1,1} & \cdot & \cdot & P_{1,m} \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ P_{m,0} & \cdot & & \cdot & P_{m,m} \end{bmatrix} \tag{7.4}$$

(we remark that $P_{i,j} = P'_{j,i}$ by the symmetry of P), then (7.2) becomes a set of algebraic equations in the unknowns, $P_{i,j}$. And, if one dissects (7.2) and compares those algebraic relations with the algebraic relations (7.1), the two sets of equations are found to be in direct correspondence (see [1, Chap. 5] for the details). Specifically:

$$\Pi_0 = P_{0,0}, \tag{7.5a}$$

$$\Pi_1(-i/m) = mP_{0,i+1}, \quad 0 \leqq i \leqq m - 1, \quad \Pi_1(-1) = \Pi_0 B = P_{0,0}B, \tag{7.5b}$$

$$\Pi_2(-i/m, -j/m) = m^2 P_{i+1,j+1}, \qquad 0 \leqq i \leqq m - 1 \text{ and } 0 \leqq j \leqq m - 1,$$

$$\Pi_2(-1, -j/m) = B'\Pi_1(-j/m) \qquad \text{for } 0 \leqq j \leqq m, \tag{7.5c}$$

$$\Pi_2(-i/m, -1) = \Pi_1'(-i/m)B \qquad \text{for } 0 \leqq i \leqq m.$$

The equivalence relation (7.5) provides the basis for a computation scheme which has proved successful (see [1, Chap. 7]) for relatively low order systems ($n \leqq 4$). Specifically, we perform a sequence of calculations for $m = 1, 2, \cdots$, and for each m, we solve the associated Riccati equation for the values of $P_{i,j}$; then equations (7.5) are used to determine $\Pi_0, \Pi_1(-i/m)$ for $0 \leqq i \leqq m$, and

$\Pi_2(-i/m, -j/m)$ for $i, j = 0, 1, \cdots, m$. Interpolation between points gives an approximation to the optimal values of $\Pi_0, \Pi_1(\theta)$ and $\Pi_2(\xi, \theta)$. The computation is terminated when no significant improvement in the approximation of Π_0, $\Pi_1(\theta)$, $\Pi_2(\xi, \theta)$ occurs.

The introduction of the matrices F_m, Q_m, D_m can be understood if one consults [18]; there the linear (ordinary) differential equation

$$\dot{y}(t) = F_m y(t) + D_m u(t)$$

is taken as a finite-dimensional approximation to the *dynamics* of the system with delay (1.1). (The first n components of $y(t)$ approximate $x(t)$, the next n components approximate $x(t - 1/m), \cdots$, the last n components approximate $x(t - 1)$ (see [18] or [1, pp. 66–69]).)

This can be given an interesting interpretation in light of the work of this section. The apparent conclusion is that the approximate solution of the equations for optimality (4.1) is equivalent to the solution of an optimal control problem involving a dynamic approximation.

Appendix. In this Appendix we show that the existence of an exact solution to equations (4.1a)–(4.1e) is equivalent to the existence of a solution of a certain hyperbolic quasi-linear partial differential equation.

The proof of Proposition 6.1 of this paper (the proof appears in [1]) shows that if the conditions of Theorem 4.1 are satisfied then Π_0, Π_1 and Π_2 are related through a fourth function, $V(\xi, \theta)$. Namely,

$$\Pi_0 = V(0,0), \tag{A.1a}$$

$$\Pi_1(\theta) = -\frac{\partial V(0,\theta)}{\partial \theta}, \tag{A.1b}$$

$$\Pi_2(\xi, \theta) = \frac{\partial^2 V(\xi, \theta)}{\partial \xi \, \partial \theta} \tag{A.1c}$$

(also $V(\xi, \theta) = V'(\theta, \xi)$).

Using these relations, we can show that the solution of equations (4.1) is equivalent to another problem, in which a second order partial differential equation plays the main role.

Define the function $G(\xi, \theta)$ by

$$G(\xi, \theta) = \frac{\partial V(\xi, \theta)}{\partial \xi}.$$

Then it follows from (A.1) and the fact that $V(\xi, \theta) = V'(\theta, \xi)$ that

$$\Pi_2(\xi, \theta) = \frac{\partial G(\xi, \theta)}{\partial \theta}, \tag{A.2a}$$

$$\Pi_1(\theta) = -G'(\theta, 0), \qquad \Pi_1'(\theta) = -G(\theta, 0). \tag{A.2b}$$

Thus by using equations (4.1c) and (4.1e) in conjunction with (A.2) we have a single second order partial differential equation for $G(\xi, \theta)$:

$$\frac{\partial^2 G(\xi, \theta)}{\partial \xi \, \partial \theta} + \frac{\partial^2 G(\xi, \theta)}{\partial \theta^2} + G(\xi, 0) D R^{-1} D' G'(\theta, 0) = 0 \tag{A.3}$$

with a boundary constraint:

$$\frac{\partial G(-1,\theta)}{\partial\theta} = -B'G'(\theta,0). \tag{A.4}$$

Assuming that a solution, $G(\xi,\theta)$, for (A.3) and (A.4) can be found, we note that the functions $\Pi_1(\theta)$ and $\Pi_2(\xi,\theta)$ can be recovered from $G(\xi,\theta)$ through equations (A.2). Knowing $\Pi_1(\theta)$, we would then know $\Pi_1(0)$ which could be substituted into equation (4.1a) leaving an algebraic Riccati equation in the single unknown, Π_0. Thus the second order partial differential equation (A.3) is potentially the key to the solution of equation (4.1). However, there are two conditions which must be met. First, the choice of boundary conditions for the partial differential equation (A.3) must not violate the constraint (A.4). Secondly, the choice of the boundary conditions must be such that the quantities $\Pi_0, \Pi_1(\theta)$ and $\Pi_2(\xi,\theta)$ recovered from $G(\xi,\theta)$ yield a nonnegative functional (6.1) for any initial function.

Thus the question of existence of a solution to equations (4.1) is equivalent to the question of the existence of a solution of equation (A.3) which satisfies the boundary constraint (A.4) (and has the required nonnegative property for the functional (6.1)). So, the whole question of the existence of solutions to (4.1) reduces to the question of existence of appropriate boundary conditions for $G(\xi,\theta)$.

In the scalar case ($n = 1$ and $G(\xi,\theta)$ is a real-valued function) equations of type (A.3) have been extensively studied. In fact, equation (A.3) is then called a "hyperbolic, quasi-linear partial differential equation" (see [19, Lemma 2, p. 87]). For such equations, the known existence results which appear to be applicable to our problem are contained in the following theorem, which presents the existence conditions for the so-called "Cauchy problem."

THEOREM A.1 (see [19, Theorem 28.5, pp. 125–126]). *Let $Z(\xi)$ be a continuously differentiable function for all ξ and $P(\xi)$ be a continuous function for all ξ.*

Then there exists a function $G(\xi,\theta)$ such that:

(i) *$G(\xi,\theta)$ has continuous second partials*;

(ii) *$G(\xi,\theta)$ is a solution of* (A.3) *in the region* $-1 \leqq \xi \leqq 0$, $-1 \leqq \theta \leqq 0$;

(iii) *for all ξ, $G(\xi,0) = Z(\xi)$ and $\partial G(\xi,\theta)/\partial\theta|_{\theta=0} = P(\xi)$.*

Theorem A.1 guarantees the existence of a solution to (A.3) whenever $G(\xi,\theta)$ and $\partial G(\xi,\theta)/\partial\theta|_{\theta=0}$ are preassigned functions. Consequently, in view of boundary constraint (A.4), the proof of the existence of a solution $G(\xi,\theta)$, $-1 \leqq \xi \leqq 0$, $-1 \leqq \theta \leqq 0$, satisfying (A.3) and (A.4) is complete if one can show that there exists at least one set of preassigned values of $G(\theta,0)$ and $\partial G(\theta,\xi)/\partial\xi|_{\xi=0}$ for which the resulting solution $G(\xi,\theta)$ satisfies: $\partial G(-1,\theta)/\partial\theta = -B'G(\theta,0)$.

Acknowledgments. The authors wish to thank A. J. Tether and Frank J. Rees for many valuable suggestions.

REFERENCES

[1] D. W. ROSS, *Optimal control of systems described by differential-difference equations*, Doctoral thesis, Department of Electrical Engineering, Stanford University, Stanford, California 1967.

[2] R. BELLMAN AND J. M. DANSKIN, *A Survey of the Mathematical Theory of Time Lag, Retarded Control, and Hereditary Processes*, RAND Rep. R-256, Rand Corp., Santa Monica, California, 1954.

[3] N. H. CHOKSY, *Time-lag systems . . . a bibliography*, IRE Trans. Automatic Control, AC-5 (1960), pp. 66–70.

[4] H. S. TSIEN, *Engineering Cybernetics*, McGraw-Hill, New York, 1954.

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

[6] M. N. OGUZTÖRELI, *A time optimal control problem for systems described by differential-difference equations*, this Journal, 1 (1963), pp. 290–310.

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

[8] D. H. CHYUNG AND E. B. LEE, *Linear optimal systems with time delay*, this Journal, 4 (1966), pp. 548–575.

[9] N. N. KRASOVSKII, *On the analytic construction of an optimum control in a system with time lags*, J. Appl. Math. Mech., 26 (1962), pp. 50–67.

[10] ———, *Optimal Processes in Systems with Time Lag*, Proc. 2nd International Federation of Automatic Control Congress (IFAC), Basel, 1963, Butterworths, London, 1964.

[11] R. E. KALMAN, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana (2), 5 (1960), pp. 102–119.

[12] ———, *When is a linear control system optimal?*, Trans. ASME Ser. D. J. Basic Engrg., 86 (1964), pp. 51–60.

[13] N. N. KRASOVSKII, *Stability of Motion*, Stanford University Press, Stanford, California, 1963.

[14] J. K. HALE, *Sufficient conditions for stability and instability of autonomous functional-differential equations*, J. Differential Equations, 1 (1965), pp. 452–484.

[15] YU. S. OSIPOV, *Stabilization of controlled systems with delays*, Differential Equations, 1 (1965), pp. 463–473.

[16] V. G. BOLTYANSKII, *Sufficient conditions for optimality and the justification of the dynamic programming method*, this Journal, 4 (1966), pp. 326–361.

[17] P. L. FALB AND D. L. KLEINMAN, *Remarks on the infinite-dimensional Riccati equation*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 534–536.

[18] M. E. SALUKVADZE, *Concerning the synthesis of an optimal controller in linear delay systems subjected to constantly acting perturbations*, Automat. Remote Control, 23 (1962), pp. 1495–1501.

[19] D. L. BERNSTEIN, *Existence Theorems in Partial Differential Equations*, Princeton University Press, Princeton, 1950.

SIAM J. Control
Vol. 7, No. 4, November 1969

SURVEILLANCE PROBLEMS: POISSON PROCESS UNDER COSTLY SURVEILLANCE[†]

C. BRADLEY RUSSELL[‡]

1. Introduction. This paper is concerned with a one-dimensional Poisson production process $x(t)$ introduced by Savage [13]. Instantaneous errorless inspections can be made at any time at a positive cost L, the costly surveillance criterion. Repair of the process returns it to the origin and takes $m > 0$ units of time at a cost of $K > 0$ per unit of time. The objective function which we seek to maximize is a form of the long run average income.

When the process is in state x and is producing, income is received at a rate $i(x)$. Under the assumption that $i(x)$ is a monotone nonincreasing function satisfying $\lim_{x \to \infty} i(x) < -K$, it is proved (§ 3) that I^*, the maximum (possibly supremum) attainable income rate, can be approximated arbitrarily closely by the income rates obtainable with nonrandomized stationary strategies and (§ 4) that there exists a nonrandomized stationary strategy whose income rate is I^*, i.e., an optimal nonrandomized stationary strategy exists. Many qualitative properties of such optimal strategies are given in Savage [13], and a computer program for obtaining an optimal strategy within this class is given by the author [12]. With the results of this paper, these are placed on a sounder theoretical basis.

The surveillance model considered here is related to a classical model which has appeared in the works of Derman [2], [3], Howard [5], Maitra [8], and Ross [10], [11], among others. It differs from the classical model principally in the fact that observations and actions may be taken at any time rather than at equally spaced time points. This difference necessitates use of a different objective function, and hence precludes direct use of results on the existence of optimal nonrandomized stationary strategies for the classical model [2], [3], [10], [11]. Yet without the existence of optimal stationary strategies, computational methods such as the policy improvement technique of Howard [5], [6] are inapplicable.

The existence of optimal nonrandomized stationary strategies for a model which resembles the surveillance model more closely than does the classical model has been investigated by Chitgopekar [1].

2. Description of model. During production, the process $x(t)$ is a Poisson process with parameter Δ, which we take without loss of generality to be $\Delta = 1$. While production is taking place in state x, income is received at rate $i(x)$. Instantaneous errorless inspections may be made at any time at a cost (paid before the inspection) $L > 0$, and repair, which returns the process to the origin, may be made immediately following an inspection. Such repair will take $m > 0$ units of time at a cost of $K > 0$ per unit of time. Upon completion of repair, no inspection need be made previous to placing the process in production or to repeating the repair.

[†] Received by the editors August 29, 1968, and in revised form February 3, 1969.

[‡] Department of Statistics, Florida State University, Tallahassee, Florida. Now at Department of Mathematics, Clemson University, Clemson, South Carolina 29631. This work was supported by the Office of Naval Research under Contract NONR 988(13).

Let the end of the nth stage of the process be signaled by the $(n + 1)$st occurrence of event $A \cup B$, where A is the taking of an observation and B is the completion of repair. Let X_n, τ_n denote, respectively, the position of the process at the end of the nth stage and the total elapsed time until the end of the nth stage. We will take $X_0 = 0$ and the 0th stage to consist of repair, hence $\tau_0 = m$.

Under a strategy S, let C_n be the income received during the nth stage, the duration of which is T_n. Let $I(N|S)$ be given by

$$I(N|S) = \frac{\sum_{n=0}^{N} EC_n}{\sum_{n=0}^{N} ET_n}.$$

The objective function, which is maximized, is $I(S) = \liminf_{N\to\infty} I(N|S)$, and $I^* \equiv \sup_S I(S)$.

One restrictive assumption will be made throughout the paper. That is, $i(x)$ will be taken to be a nonincreasing function with the property that $\lim_{x\to\infty} i(x) < -K$. It is conjectured that this assumption could be weakened to the condition that $i(x)$ satisfy $\limsup i(x) < -K$, dropping the monotonicity assumption, without changing the results of this paper. The proofs given here, however, will not generalize directly to accommodate this weakened assumption.

Two other assumptions will be made throughout, but without loss of generality. These are (a) that $i(0) = 0$, and (b) the class of strategies $\{S : I(S) > -K\}$ is nonempty. That these assumptions may be made without loss of generality is easily seen. If $i(0) \equiv A \neq 0$, we simply define an analogous problem for which $i'(x) \equiv i(x) - A$, $K' = K + A$, and $(I^*)' = I^* - A$, clearing the first assumption. If $\{S : I(S) > -K\}$ is empty, the strategy of eternal repair is an optimal nonrandomized stationary strategy and the problem is trivial. Thus we need only consider those strategies in the nonempty set $\{S : I(S) > -K\}$.

3. Approximating I^* with nonrandomized stationary strategies. In this section it is proved that if $\mathscr{S}^*$ is the class of nonrandomized stationary strategies, then

$$I^* = \sup_{S \in \mathscr{S}^*} I(S).$$

First, however, some definitions are needed.

Let $\mathscr{A}$ denote the action space. For our present purposes, $\mathscr{A} = \{r\} \cup [0, \infty)$. That is, an action $a \in \mathscr{A}$ consists of either the decision to repair or the decision to continue production for a time $t \in [0, \infty)$ before observing the process again. Later, in Theorem 3.3 and Lemma 4.1, $\mathscr{A}$ will be restricted.

Let $\Delta_n \in \mathscr{A}$ be the action taken at time τ_n.

DEFINITION 3.1. A *decision procedure* given a history

$$H_n = (X_0, \tau_0, \Delta_0, \cdots, X_n, \tau_n)$$

is a probability function (density) on the space $\mathscr{A}$. That is, for each $a \in \mathscr{A}$, $D_a(H_n) = P(\Delta_n = a|H_n)$, where $P(\cdot|H_n)$ is to be interpreted as a probability function or density as the situation dictates.

DEFINITION 3.2. A *strategy* is a sequence $\{\mathscr{D}_n\}$ of decision procedures, where $\mathscr{D}_n = \{D_a(H_n)\}$.

Definition 3.3. A *nonrandomized strategy* is a strategy such that $\mathscr{D}_n$ is degenerate for each n.

Definition 3.4. A *stationary strategy* is a strategy such that for each n, $\mathscr{D}_n = \{D_a(x)\}$, where $X_n = x$. That is, $\mathscr{D}_n$ depends only on the current observed position of the process.

For nonrandomized stationary strategies one need only notice the present position of the process to determine the present action.

Definition 3.5. It was assumed that any strategy S under consideration satisfies $I(S) > -K$. This is equivalent to saying that for each S there exists an integer $N_0(S)$ such that whenever $N \geqq N_0(S)$, $I(N|S) > -K$. The numbers $N_0(S)$ so defined will be extensively used in the proofs to follow.

Theorem 3.1. *To any strategy S there corresponds a strategy S' with the following properties*:

(a) $I(S') \geqq I(S)$, *and*

(b) *Under S', production is continued only when the process is observed in a subset of the finite set $C_0 = \{x : i(x) \geqq -K\}$.*

Proof. Define strategy S' as follows:

Step 1. Let S' be the same as S until a point in $\overline{C_0}$, the complement of C_0, is observed.

Step 2. Then conduct, with no real time, a random experiment simulating the path of the production process under S until a repair is made. Since $I(S) > -K$, a repair will be made with probability one.

Step 3. Under S', immediately make as many observations as were made in the random experiment, then repair.

Immediately upon completion of Steps 2 and 3, the position of the process under S' is the same as the simulated position under S, both being the origin. Up to this point only the histories differ under the two strategies.

Step 4. To continue the definition of S', use the (partially synthetic) history under S as if it were the true history under S' up to this point and proceed as before beginning with Step 1.

Notice that at each stage the histories used in determining the actions under S' have the same distribution as the histories under S. Hence for each stage where repair or production is called for under S', the total expected income and total expected time for that stage equal those for the corresponding stage under S. Furthermore, observations are taken at the same stages under each strategy.

Let $N \geqq N_0(S)$, as in Definition 3.5. In the first N stages of the process under S:

Let A denote the total expected income from production under S from those stages in which production was begun at a point in C_0, minus the costs of observations and repairs.

Let B denote the total expected production time under S from those stages in which production was begun at a point in C_0, plus the time of repairs.

Let T denote the total expected production time under S from those stages in which production was begun at a point in $\overline{C_0}$. Then

$$\frac{A - KT}{B + T} \geqq I(N|S) > -K.$$

Hence $A/B > -K$ and

$$I(N|S') = \frac{A}{B} \geqq \frac{A - KT}{B + T} \geqq I(N|S)$$

for every $N \geqq N_0(S)$. Thus (a) holds, and condition (b) is clear.

Let $I(x, T)$ be defined as the conditional expected income from production given that production is begun in state x and continued for time T.

Lemma 3.2 (Collected properties of $I(x, T)$).

(a) $$I(x, T) = \sum_{n=0}^{\infty} \frac{e^{-T}T^n}{n!} \sum_{y=0}^{n} \frac{i(x + y)T}{n + 1}.$$

(b) *For each $T \geqq 0$, $I(x, T)/T$ is nonincreasing in x.*

(c) *For each $x \geqq 0$, $I(x, T)/T$ is decreasing in T and* $\lim_{T\to\infty} I(x, T)/T < -K$.

(d) *Let $T_x^* = \inf\{T \geqq 0 : I(x, t)/t < -K$ for $t \geqq T\}$ and let T'_x be the radius of convergence of $I(x, T)$ as expressed in* (a). *Then for each x, $T_x^* < \infty$, $T_x^* \leqq T'_x$, and if $I(x, T'_x) = -\infty$, $T_x^* < T'_x$.*

(e) *$I(x, T)$ is continuous on* $[0, T_x^*]$.

Proof. The proof of part (c) will be indicated. The others are straightforward. That $\lim_{T\to\infty} I(x, T)/T < -K$ follows from

$$\lim_{n\to\infty} \frac{\sum_{y=0}^{n} i(x + y)}{n + 1} < -K,$$

which is true since $\lim_{x\to\infty} i(x) < -K$. Showing monotonicity of $I(x, T)/T$ is based on the evident monotonicity of $(\sum_{y=0}^{n} i(x + y))(n + 1)$ and monotone likelihood ratio theory [7, Lemma 2].

Theorem 3.3. *To any strategy S there corresponds a strategy S' such that*

(a) *$I(S') \geqq I(S)$ and*

(b) *$P(t_{(n+1)} \leqq T^*_{x(\tau_n)}$ for each n such that $x(\tau_n) \in C_0 | S') = 1$, where t_n denotes the (random) amount of production time in stage n under S'.*

Proof. Let n_1 be the first stage at which $P(t_n > T^*_{x(\tau_{(n-1)})}|S) > 0$, and let n_i denote the ith such stage. Define S' as follows:

Step 1. Until stage n_1 is reached, S' and S agree.

Step 2. If at the n_1th stage, repair or production and a time $t_{n_1} \leqq T^*_{x(\tau_{(n-1)})}$ before the next observation is called for under S, let S' and S agree.

Step 3. If at the n_1th stage, production and a waiting time $t_{n_1} > T^*_{x(\tau_{(n_1-1)})}$ is called for under S, then for the n_1th stage under S' simply make an observation. Perform a random experiment given t_{n_1} to obtain a synthetic history for the n_1th stage under S. Thus we have a partially synthetic n_1 stage history under S. Let the actions taken under S' from the $(n_1 + 1)$st stage to the n_2th stage be determined by those decisions made under S during that time given the n_1 stage history described above. The actions under S' at stage n_2 and beyond are determined analogously.

Notice that under S', the process is always at least as close to the origin as the hypothesized process under S, and inspections are made at the same stages under each. Hence the rate of income from production under S' is at least as high as it would be under S. Furthermore, (b) is satisfied by the same reasoning.

Let $N \geqq N_0(S)$ be given, and in the first N stages use the following notation:

Let A be the contribution to the expected income under S when repairs or production beginning at a state x and lasting for a time $t \leqq T_x^*$ before observing are called for, minus the costs of observation.

Let A' be the same as above for S'.

Let B be the contribution to the expected time under S (or S') when repairs or production commencing at x and lasting $t \leqq T_x^*$ units are called for.

Let T be the contribution to the total expected time from those stages when production commences at x and lasts for a time $t > T_x^*$ before observation.

Then it is true that

$$I(N|S) = \frac{A + cT}{B + T} \geqq -K,$$

where $c < -K$. But then $A/B > -K$, and $A/B \geqq (A + cT)/(B + T)$. Thus

$$I(N|S') = \frac{A'}{B} \geqq \frac{A}{B} \geqq \frac{A + cT}{B + T} = I(N|S).$$

Thus we have reduced the problem to consideration of a class of strategies which allow production only when the process is observed in a state $x \in C_0$. When such an x is observed and production is called for, the production time t until the next observation satisfies $t \in [0, T_x^*]$.

DEFINITION 3.6 (Lévy metric). Let F and G be distribution functions on the real line. The *Lévy distance* $L(F, G)$ between these functions is given by

$$L(F, G) = \inf\{h : F(x - h) - h \leqq G(x) \leqq F(x + h) + h \text{ for all } x\}.$$

Define $\mathscr{F}_N$ to be the class of discrete distributions whose mass is concentrated at points of the form $k \cdot 2^{-N}$, where k is an integer.

LEMMA 3.4. *For every $\varepsilon > 0$ there exists $N(\varepsilon)$ such that whenever $N \geqq N(\varepsilon)$ and F is a distribution on the real line, there exists a distribution $F_N \in \mathscr{F}_N$ satisfying $L(F_N, F) < \varepsilon$, and $F_N(y) \geqq F(y)$ for each y. Furthermore, if F is concentrated on $[0, T]$, F_N can be taken as concentrated on $[0, T]$.*

Proof. Let $N(\varepsilon)$ be such that $2^{-N(\varepsilon)} < \varepsilon$, and suppose $N \geqq N(\varepsilon)$. Define F_N as follows:

$$F_N(y) = F\left(\frac{k+1}{2^N}\right), \qquad \frac{k}{2^N} \leqq y < \frac{k+1}{2^N}.$$

This lemma, along with the equivalence of Lévy convergence and weak convergence [4, p. 33], yields the following corollary.

COROLLARY 3.5. *For every $\varepsilon > 0$, there exists $N^*(\varepsilon)$ such that if $x \in C_0$, F is a distribution concentrated on $[0, T_x^*]$, and $N \geqq N^*(\varepsilon)$, there exists a distribution $F_N \in \mathscr{F}_N$ concentrated on $[0, T_x^*]$ such that*

$$\left|\int_0^{T_x^*} I(x, t)\, dF_N(t) - \int_0^{T_x^*} I(x, t)\, dF(t)\right| < \varepsilon \tag{3.1}$$

and

$$\left|\int_0^{T_x^*} t\, dF_N(t) - \int_0^{T_x^*} t\, dF(t)\right| < \varepsilon. \tag{3.2}$$

LEMMA 3.6. *Let S be a strategy satisfying $I(S) > -K$. Then for $N \geqq N_0(S)$,*

$$\sum_{n=0}^{N} ET_n > \frac{(N+1)\min(L, mK)}{K}.$$

Proof. Recall that $i(x) \leqq 0$, and that for $N \geqq N_0(S)$,

$$\frac{\sum_{n=0}^{N} EC_n}{\sum_{n=0}^{N} ET_n} > -K.$$

But $\sum_{n=0}^{N} EC_n \leqq -(N+1)\min(L, mK)$, hence

$$\frac{-(N+1)\min(L, mK)}{\sum_{n=0}^{N} ET_n} > -K,$$

or

$$\sum_{n=0}^{N} ET_n > \frac{(N+1)\min(L, mK)}{K}.$$

THEOREM 3.7. *Let S be a strategy. Then for every $\varepsilon > 0$, there is an integer $N(\varepsilon)$ such that whenever $N \geqq N(\varepsilon)$ there is a strategy S_N satisfying*:

(a) $I(S_N) > I(S) - \varepsilon$.

(b) *The production times called for in each stage under S_N are integral multiples of 2^{-N}.*

(c) *Production under S_N is allowed only when the process is observed in a subset of C_0.*

(d) *If production is continued for time t when the process is observed at $x \in C_0$ under S_N, then $0 \leqq t \leqq T_x^*$ a.s.*

Proof. Define S_N as follows:

At each stage where S calls for production for time t, S_N calls for production for time $t' = [2^N t] \cdot 2^{-N}$. At each stage where S calls for repair, S_N does also. More precisely:

S and S_N agree at stage 0, both calling for repair.

At stage 1, the process begins production at the origin, and S calls for a time t_1 until the next observation. For S_N, then, produce for $t_1' = [2^N t_1] \cdot 2^{-N}$ units of time then observe the process at, say, x_1'. Perform a random experiment to determine the position, x_1, of the process under S at time $m + t_1$, given that it was at x_1' at time $m + t_1'$ and was allowed to continue in production until $m + t_1$. Let H_1 be the (partially synthetic) history of the process under S given by the history H_0 of stage 0, the motion of the process from 0 to x_1' at time $m + t_1'$, and by the random experiment from this point to x_1 at time $m + t_1$.

Given H_1, then S dictates an action for stage 2. If repair is called for, repair under S_N. If production time t_2 is called for, produce under S_N for $t_2' = [2^N t_2] \cdot 2^{-N}$ time units, observing the process in a state x_2' at time $m + t_1' + t_2'$. Perform a random experiment to determine the position of the process under S at time $m + t_1 + t_2$ given that at time $m + t_1 + t_2'$ it was in state $x_2' + x_1 - x_1'$ and production was allowed until time $m + t_1 + t_2$. Let H_2 be the (partially synthetic) history of the process under S, given by H_1, the motion of the process from x_1 at time $m + t_1$ to $x'_2 + x_1 - x'_1$ at time $m + t_1 + t'_2$ corresponding to the motion from x'_1 at time $m + t'_1$ to x'_2 at time $m + t'_1 + t'_2$ under S_N, and by the random experiment from this point to x_2 at time $m + t_1 + t_2$.

Proceed to stage 3 and beyond analogously.

Notice that if $F_{x|H}$ is a distribution which determines the waiting time until observation, given that the process was just observed at x with history H under S, then the above method for determining the action under S_N is equivalent to defining a distribution $F_{N,x|H}$ as in the proof of Lemma 3.4. Let $\delta > 0$ be chosen corresponding to $\varepsilon > 0$ as in Lemma A in the Appendix. Corollary 3.5 provides an integer $N^*(\delta) \equiv N(\varepsilon)$ such that for $N \geqq N(\varepsilon)$ and $x \in C_0$, there exists $F_{N,x|H} \in \mathscr{F}_N$ satisfying

$$\left| \int_0^{T_x^*} t \, d F_{N,x|H}(t) - \int_0^{T_x^*} t \, d F_{x|H}(t) \right| < \delta \tag{3.3}$$

and

$$\left| \int_0^{T_x^*} I(x, t) \, d F_{N,x|H}(t) - \int_0^{T_x^*} I(x, t) \, d F_{x|H}(t) \right| < \delta. \tag{3.4}$$

Let A_n and A'_n be the expected incomes in the nth stage under S and S_N, respectively, and let B_n and B'_n be the expected times in the nth stage under S and S_N.

Recalling the definitions of x'_n and x_n made earlier in the proof, it is clear that $x'_n \leqq x_n$ for all n. Hence $I(x'_n, t) \geqq I(x_n, t)$ for each t and n. Thus by (3.3) and (3.4) it is true that for each n,

$$A'_n - A_n > -\delta$$

and

$$B'_n - B_n > -\delta.$$

Hence by Lemma A there is a positive integer $N_0(S)$ such that $M \geqq N_0(S)$ implies that

$$\frac{\sum_{n=0}^{M} A'_n}{\sum_{n=0}^{M} B'_n} > \frac{\sum_{n=0}^{M} A_n}{\sum_{n=0}^{M} B_n} - \varepsilon,$$

or

$$I(S_N) > I(S) - \varepsilon$$

for $N \geqq N(\varepsilon)$. Properties (b), (c) and (d) are clear.

THEOREM 3.8. *Let S be a strategy and let $\varepsilon > 0$ be given. There exists a positive integer $N(\varepsilon)$ such that for $N \geqq N(\varepsilon)$ there exists a nonrandomized stationary strategy S_N^* satisfying*:

(a) $I(S_N^*) > I(S) - \varepsilon$.

(b) *Production under S_N^* is continued only when the process is observed in a subset of C_0.*

(c) *If at state $x \in C_0$, S_N^* calls for production for time $t(x)$, then $0 \leqq t(x) \leqq T_x^*$.*

Proof. Let $\mathscr{S}_N$ be the class of all strategies satisfying (b), (c) and (d) of Theorem 3.7. Consider the set $\overline{C_0}$ to be a single state such that observing the process in $\overline{C_0}$ calls for repair. Then this model with the class of strategies $\mathscr{S}_N$ has finite state and action spaces. Hence by Derman's Theorem 3 (see [2]), (Assumption A is not necessary), there exists a nonrandomized stationary strategy $S_N^* \in \mathscr{S}_N$ satisfying

$$I(S_N^*) = \sup_{S \in \mathscr{S}_N} I(S) \geqq I(S_N) > I(S) - \varepsilon,$$

where S_N is the strategy in $\mathscr{S}_N$ guaranteed by Theorem 3.7.

DEFINITION 3.7. Let $\mathscr{S}^*$ be the class of nonrandomized stationary strategies which allow production only if the process is observed in a state $x \in C_0$, and when $x \in C_0$ and production is allowed, the time $t(x)$ until the next observation satisfies $0 \leqq t(x) \leqq T_x^*$.

COROLLARY 3.9. $I^* = \sup_{S \in \mathscr{S}^*} I(S)$.

4. Existence of an optimal nonrandomized stationary strategy. The fact that $I^* = \sup_{S \in \mathscr{S}^*} I(S)$ suggests the possibility of a limiting procedure to obtain an optimal strategy $S^* \in \mathscr{S}^*$. That possibility is exploited in this section. Briefly, the proof is as follows. There exists a sequence $\{S_n\} \subset \mathscr{S}^*$ such that $\lim_{n \to \infty} I(S_n) = I^*$. A subsequence $\{S_n'\} \subset \{S_n\}$ is obtained which converges to a limiting strategy $S^* \in \mathscr{S}^*$. Finally it is shown that $I(S^*) = I^*$.

For each $S \in \mathscr{S}^*$ let $C(S) \subset C_0$ be the set of states which call for production under S, the *continuation set*, and write (see [2]),

$$I(S) = \frac{\sum_{x \in C(S)} p_s(x)[I(x, t_s(x)) - L] - p_s(R)mK}{\sum_{x \in C(S)} p_s(x)t_s(x) + p_s(R)m}, \tag{4.1}$$

where $p_s(x)$ is the steady state probability under S of being observed in state x, $t_s(x)$ denotes the time called for under S between observing the process in x and observing again, and R equals $\overline{C(S)}$.

An equivalent and equally useful formulation of (4.1) is as follows. A *cycle* is defined as being the (random) sequence of observed states starting from the beginning of production after repair until the recurrence of that event. If $n_s(x)$ is defined as the expected number of times per cycle that the state x is observed under S, it is clear that $n_s(R) = 1$, and, in general

$$n_s(x) = \frac{p_s(x)}{p_s(R)}.$$

Hence (4.1) may be written as

$$I(S) = \frac{\sum_{x \in C(S)} n_s(x)[I(x, t_s(x)) - L] - mK}{\sum_{x \in C(S)} n_s(x)t_s(x) + m}. \tag{4.2}$$

Let $C_0^* = \{x : x \in C(S) \text{ for some } S \in \mathscr{S}^*\}$. Define for each $x \in C_0^*$, ${}_*T_x = \inf\{t_s(x) : S \in \mathscr{S}^* \text{ and } x \in C(S)\}$.

LEMMA 4.1. *For each* $x \in C_0^*$, ${}_*T_x > 0$.

Proof. For any $S \in \mathscr{S}^*$ we can write, by (4.2),

$$-K < \frac{\sum_{x \in C(S)} n_s(x)[I(x, t_s(x)) - L] - mK}{\sum_{x \in C(S)} n_s(x)t_s(x) + m}.$$

Recalling that $i(x) \leqq 0$ for each x, we write

$$-K < \frac{-\sum_{x \in C(S)} n_s(x)L - mK}{\sum_{x \in C(S)} n_s(x)t_s(x) + m}. \tag{4.3}$$

Since the quantity in the denominator of the right-hand side of (4.3) expresses the expected time per cycle under S, it is easy to see that an upper bound for this quantity is

$$G = \max_{x \in C_0} x + \max_{x \in C_0} T_x^* + m,$$

from properties of the Poisson process.

Furthermore, if $N(S)$ denotes the expected number of observations per cycle under S, it is clear that $N(S) = \sum_{x \in C(S)} n_s(x) + 1$. Hence

$$-K < \frac{-N(S) \min(L, mK)}{G}$$

or

$$N(S) < \frac{GK}{\min(L, mK)}$$

for each $S \in \mathscr{S}^*$.

Now for each $S \in \mathscr{S}^*$ and $x \in C(S)$, it is clear that $n_s(x) < N(S)$. Furthermore,

$$n_s(0) = \sum_{m=0}^{\infty} (m + 1) e^{-mt_s(0)}(1 - e^{-t_s(0)}) = \frac{1}{1 - e^{-t_s(0)}}.$$

But then it must be that

$$\frac{1}{1 - e^{-t_s(0)}} < \frac{GK}{\min(L, mK)}. \tag{4.4}$$

Let $T_0 = \inf\{t : t \text{ satisfies (4.4)}\}$. Then $0 < T_0 \leqq {}_*T_0$. Further we can let

$$P_x = \min\left\{\frac{e^{-T_0}T_0^x}{x!}, \frac{e^{-T_0^*}(T_0^*)^x}{x!}\right\}$$

and write for $x \in C(S)$, $x \neq 0$,

$$n_s(x) \geqq P_x \sum_{m=0}^{\infty} (m + 1) e^{-mt_s(x)}(1 - e^{-t_s(x)})$$

and in an analogous way obtain $0 < T_x \leqq {}_*T_x$.

We now define a limiting strategy $S^* \in \mathscr{S}^*$ which Theorem 4.2 will show to be optimal.

By the preceding lemma and § 3 we have obtained a class $\mathscr{S}^*$ of nonrandomized stationary strategies satisfying:

(a) If $S \in \mathscr{S}^*$, the continuation set $C(S)$ is a subset of C_0^*,

(b) $x \in C(S)$ implies $t_s(x) \in [{}_*T_x, T_x^*]$, ${}_*T_x > 0$, and

(c) $I^* = \sup_{S \in \mathscr{S}^*} I(S)$.

By property (c), there exists a sequence $\{S_n\}$ of strategies in $\mathscr{S}^*$ such that $\lim_{n \to \infty} I(S_n) = I^*$. For each point $x \in C_0^*$, S_n determines an action $a_n(x)$ to be taken. Such an action is denoted by a point in the set $\mathscr{A}_x = [{}_*T_x, T_x^*] \cup \{r\}$, where r denotes repair. Define a topology on $\mathscr{A}_x$ as follows: Let the open sets in $\mathscr{A}_x$ be those of form 0 or 0 $\cup$ $\{r\}$, where 0 is open in the usual topology on $[{}_*T_x, T_x^*]$. That $\mathscr{A}_x$ is compact under this topology is easily seen. Due to this fact and the fact that C_0^* is finite, one can obtain a subsequence $\{S_n'\} \subset \{S_n\}$ of strategies such that:

(a) For each $x \in C_0^*$, $\{a_n'(x)\}$, the sequence of actions called for at x by $\{S_n'\}$, converges to a point $a(x) \in [{}_*T_x, T_x^*] \cup \{r\}$, and

(b) $\lim_{n \to \infty} I(S_n') = I^*$.

Let $S^* \in \mathscr{S}^*$ be the strategy calling for the action $a(x)$ at $x \in C_0^*$, then

THEOREM 4.2. $I(S^*) = 1^*$.

Proof. The proof of this theorem follows that of Lemma 4.3.

It is a property of convergence in the topology on $\mathscr{A}_x$ defined above that there exists an integer N such that if $a(x) = r$ and $n \geqq N$, then $a_n'(x) = r$. Hence by considering only the tail of the sequence of strategies, namely $\{S_n'\}_{n=N,\dots,\infty}$, we may speak of a well-defined continuation set C, the same for all strategies in $\{S^*\} \cup \{S_n'\}_{n=N,\dots,\infty}$. The remaining set of states $\bar{C}$, calls for repair, and this set will be designated as R. For brevity of notation write $\pi_n(x) = p_{s_n'}(x)$ and $\pi(x) = p_{s^*}(x)$. Thus, the convergence mentioned in the following lemma is defined.

LEMMA 4.3. *Let* $\pi' = (\pi(0), \pi(1), \cdots, \pi(R))$ *and* $\pi_n' = (\pi_n(0), \pi_n(1), \cdots, \pi_n(R))$ *where the components of these vectors are defined above. Then* $\lim_{n \to \infty} \pi_n = \pi$ *in the sense of elementwise convergence.*

Proof. For each $n \geqq N$, where N is defined above, π_n is the unique solution to

$$\pi_n = Q_n'\pi_n, \qquad U'\pi_n = 1, \tag{4.5}$$

where $Q_n = \{{}_nq_{ij}\}$ is the matrix of transition probabilities under S_n' and $U' = (1, 1, \cdots, 1)$. π is the unique solution to the analogous equations

$$\pi = Q'\pi, \qquad U'\pi = 1. \tag{4.6}$$

The uniqueness of the solutions of these equations is guaranteed by the fact that if $x \in C$ and $n \geqq N$, then $a_n'(x) \in [{}_*T_x, T_x^*]$ and $a(x) \in [{}_*T_x, T_x^*]$. Thus the Markov chains involved are finite and irreducible from which we obtain uniqueness [9, p. 251].

It is easily verified that equivalent formulations of (4.5) and (4.6) are as follows: For $n \geqq N$, π_n is the unique solution to

$$P_n\pi_n = I, \tag{4.5'}$$

where $I' = (0, 0, \cdots, 1)$,

$$P_n = \begin{pmatrix} {}_nq_{00} - 1 & 0 & 0 & \cdots & 0 & 1 \\ {}_nq_{01} & {}_nq_{11} - 1 & 0 & \cdots & 0 & 0 \\ {}_nq_{02} & {}_nq_{12} & {}_nq_{22} - 1 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ {}_nq_{0c} & {}_nq_{1c} & {}_nq_{2c} & \cdots & {}_nq_{cc} - 1 & 0 \\ 1 & 1 & 1 & \cdots & 1 & 1 \end{pmatrix},$$

and c denotes the maximal element of C.

π is the unique solution to an analogous set of equations

$$(4.6') \qquad P\pi = I.$$

Since these solutions are unique, $\det(P_n) \neq 0$ and $\det(P) \neq 0$. Hence by Cramer's rule,

$$\pi_n = \frac{\|P_n\|}{\det(P_n)}$$

and

$$\pi = \frac{\|P\|}{\det(P)},$$

where $\|A\|$ denotes the vector of cofactors of the last column of the matrix A.

Since $P_n \to P$ elementwise, $\|P_n\| \to \|P\|$ and $\det(P_n) \to \det(P)$ by properties of determinants. Since the determinants are nonzero,

$$\frac{\|P_n\|}{\det(P_n)} \to \frac{\|P\|}{\det(P)}.$$

Hence, $\pi_n \to \pi$ as desired.

Proof of Theorem 4.2. By (4.1) and the remarks above, for $n \geqq N$ we can write

$$I(S_n') = \frac{\sum_{x \in C} \pi_n(x)[I(x, a_n'(x)) - L] - \pi_n(R)mK}{\sum_{x \in C} \pi_n(x)a_n'(x) + \pi_n(R)m}$$

and

$$I(S^*) = \frac{\sum_{x \in C} \pi(x)[I(x, a(x)) - L] - \pi(R)mK}{\sum_{x \in C} \pi(x)a(x) + \pi(R)m}.$$

By Lemma 3.2 (e) we know that for $x \in C$, $I(x, a_n'(x))$ converges to $I(x, a(x))$ as $a_n'(x)$ converges to $a(x)$. Lemma 4.3 yields the convergence of $\pi_n(x)$ to $\pi(x)$. Hence $I(S_n')$ converges to $I(S^*)$ giving $I^* = I(S^*)$ by uniqueness of limits.

Thus we have proved the existence of an optimal nonrandomized stationary strategy. Many properties of such strategies are to be found in [13]. The author has obtained [12] a computer program for obtaining optimal strategies for these problems when $i(x)$ is a polynomial function.

Appendix.

LEMMA A. *Let $\{A_n\}$, $\{A'_n\}$, $\{B_n\}$, $\{B'_n\}$ be sequences of real numbers and K and c be positive real numbers such that for $N \geqq N_0 \geqq 0$,*

$$\sum_{n=0}^{N} B_n > c(N+1) \quad \textit{and} \quad \frac{\sum_{n=0}^{N} A_n}{\sum_{n=0}^{N} B_n} > -K.$$

Then for every $\varepsilon > 0$ there exists a $\delta > 0$ such that whenever $A'_n - A_n > -\delta$ and $B'_n - B_n > -\delta$ for each n, it is true that $N \geqq N_0$ yields

$$\frac{\sum_{n=0}^{N} A'_n}{\sum_{n=0}^{N} B'_n} - \frac{\sum_{n=0}^{N} A_n}{\sum_{n=0}^{N} B_n} > -\varepsilon.$$

Proof. Let $\delta = \min\left(\frac{c}{2}, \frac{c\varepsilon}{2(K+1)}\right)$. First we notice that for $N \geqq N_0$,

$$\sum_{n=0}^{N} B'_n = \sum_{n=0}^{N} B_n + \sum_{n=0}^{N} (B'_n - B_n) > c(N+1) - \delta(N+1) \geqq \frac{c}{2}(N+1).$$

And since $\sum_{n=0}^{N} B'_n$ and $\sum_{n=0}^{N} B_n$ are positive for $N \geqq N_0$, the conclusion of the lemma is equivalent to

$$\sum_{n=0}^{N} A'_n \sum_{n=0}^{N} B_n - \sum_{n=0}^{N} A_n \sum_{n=0}^{N} B'_n > -\varepsilon \sum_{n=0}^{N} B_n \sum_{n=0}^{N} B'_n$$

which we now show.

$$\begin{aligned}\sum_{n=0}^{N} A'_n \sum_{n=0}^{N} B_n - \sum_{n=0}^{N} A_n \sum_{n=0}^{N} B'_n &= \sum_{n=0}^{N} (A'_n - A_n) \sum_{n=0}^{N} B_n + \sum_{n=0}^{N} A_n \sum_{n=0}^{N} (B_n - B'_n) \\ &> -(N+1)\delta \sum_{n=0}^{N} B_n - K \sum_{n=0}^{N} B_n(N+1)\delta \\ &> -\frac{2\delta}{c} \sum_{n=0}^{N} B_n \sum_{n=0}^{N} B'_n - \frac{2K\delta}{c} \sum_{n=0}^{N} B_n \sum_{n=0}^{N} B'_n \\ &= -\frac{2(K+1)}{c}\delta \sum_{n=0}^{N} B_n \sum_{n=0}^{N} B'_n \\ &\geqq -\varepsilon \sum_{n=0}^{N} B_n \sum_{n=0}^{N} B'_n.\end{aligned}$$

REFERENCES

[1] S. S. CHITGOPEKAR, *Continuous time Markovian sequential control processes*, F.S.U. Statistics Rep. M127, The Florida State University, Tallahassee, 1967.

[2] C. DERMAN, *On sequential decisions and Markov chains*, Management Sci., 9 (1962), pp. 16–24.

[3] ———, *Denumerable state Markovian decision processes—average cost criterion*, Ann. Math. Statist., 37 (1966), pp. 1545–1554.

[4] B. V. GNEDENKO AND A. N. KOLMOGOROV, *Limit Distributions for Sums of Independent Random Variables*, revised ed., Addison-Wesley, Reading, Massachusetts, 1968.

[5] R. HOWARD, *Dynamic Programming and Markov Processes*, Technology Press, John Wiley, New York, 1960.

[6] ———, *Semi-Markovian decision processes*, Bull. Inst. Internat. Statist., 40 (1963), pp. 625–652.

[7] S. KARLIN AND H. RUBIN, *The theory of decision procedures for distributions with monotone likelihood ratio*, Ann. Math. Statist., 27 (1956), pp. 272–299.

[8] A. MAITRA, *Dynamic programming for countable state systems*, Sankhyā Ser. A., 27 (1965), pp. 241–248.

[9] E. PARZ ͗N, *Stochastic Processes*, Holden-Day, San Francisco, 1962.

[10] S. M. ROSS, *Non-discounted denumerable Markovian decision models*, Ann. Math. Statist., 39 (1968), pp. 412–423.

[11] ———, *Arbitrary state Markovian decision processes*, Ann. Math. Statist., 39 (1968), pp. 2118–2122.

[12] C. B. RUSSELL, *Surveillance problems: a program for obtaining optimal costly surveillance strategies for a Poisson production process*, F.S.U. Statistics Rep. M125, The Florida State University, Tallahassee, 1967.

[13] I. R. SAVAGE, *Surveillance problems*, Naval Res. Logist. Quart., 12 (1962), pp. 187–209.

SIAM J. CONTROL
Vol. 7, No. 4, November 1969

NECESSARY CONDITIONS FOR OPTIMAL STRATEGIES IN A CLASS OF NONCOOPERATIVE N-PERSON DIFFERENTIAL GAMES*

I. G. SARMA, R. K. RAGADE AND U. R. PRASAD†

1. Introduction. Examples of N-person differential games appear to have been first constructed by Petrosyan [1]–[4]. The study of N-person games in general was initiated by von Neumann and Morgenstern [5] and Nash [6]. It has been applied to various team game situations, squadron warfare, and so forth. An exhaustive list of papers is found in [7]–[10].

In this paper we wish to extend some concepts of noncooperative N-person games to a dynamic situation. Further results about noncooperative and cooperative N-person differential games with their several ramifications shall be reported elsewhere. A rigorous theory to the study of a class of differential games has been advanced by Berkovitz [11]–[13]. This paper draws inspiration from his work and is an attempt to extend his results in [12] to the N-person noncooperative situation. The main results are the necessary conditions for equilibrium strategies of a class of N-person differential games and the properties of the value function vector. We assume perfect information to all the players implying that they know the state of the game during the course of the play.

The terminology and notation in this paper shall be to a large extent that in [12]. We shall suppose that the reader is familiar with its contents and reference is made to it at a number of places in this paper. Thus the terms region, closed region and decomposition are understood as in [12]. However, superscripts on control and adjoint variables, and payoff and value functions indicate the player to which they belong. The collection of corresponding quantities of all players is represented by bold letters.

2. Formulation of the game. In an N-person differential game the N-players have N payoff functions, in general all different. The state of the game satisfies a set of differential equations, and each player, knowing the state at every instant of time and knowing how the game proceeds, selects his optimal strategy from a permissible set, to minimize his payoff function. We obtain necessary conditions for a Nash equilibrium point under certain assumptions, by considering the optimality of each player's strategy against the optimal play of the rest of the $N - 1$ players. The terms strategies, Nash equilibrium point and optimality in this sense will be defined presently.

Let t denote time, x an n-dimensional vector and u^l, $l = 1, 2, \cdots, N$, an r^l-dimensional vector. We shall be concerned with a bounded region $\mathscr{S}$ of $(t, x, \mathbf{u}) = (t, x, u^1, \cdots, u^N)$ space and a region $\mathscr{G}$ of (t, x) space contained in the projection of $\mathscr{S}$ into (t, x) space. We consider two real vector valued functions

$$f(t, x, \mathbf{u}) = (f^1(t, x, \mathbf{u}), \cdots, f^N(t, x, \mathbf{u}))$$

and

$$G(t, x, \mathbf{u}) = (G^1(t, x, \mathbf{u}), \cdots, G^n(t, x, \mathbf{u}))$$

* Received by the editors June 6, 1968, and in revised form January 28, 1969.

† Department of Electrical Engineering, Indian Institute of Technology, Kanpur, U.P., India. The work of the second author was supported by the National Institute of Sciences of India, New Delhi.

of class $C^{(1)}$ on $\mathscr{S}$ with ranges contained in Euclidean spaces of dimensions N and n respectively.

We also consider functions $\Omega^l(t,x)$ from $\mathscr{G}$ into subsets of E^{r^l}, $l = 1, 2, \cdots, N$, such that

$$(t, x, \Omega^1(t,x), \cdots, \Omega^N(t,x)) = (t, x, \mathbf{\Omega}(t,x)) \in \mathscr{S} \qquad \text{for all } (t,x) \in \mathscr{G}.$$

We shall define the terminal surface $\mathscr{T}$ of the game as a connected manifold which separates $\mathscr{G}$ and which can be written as

$$\mathscr{T} = \bigcup_{i=1}^{\alpha} \mathscr{T}_i,$$

where each $\mathscr{T}_i$ is an n-dimensional $C^{(1)}$ manifold given parametrically by the equations

$$t = T_i(\sigma), \qquad x = X_i(\sigma),$$

where $\sigma = (\sigma^1, \cdots, \sigma^n)$ ranges over a cube in E^n. Let $g = (g^1, \cdots, g^N)$ be a real vector-valued function defined and of class $C^{(1)}$ in a neighborhood of $\mathscr{T}$. Then g^l is the terminal payoff to the lth player.

The game takes place in a region $\mathscr{R}$ whose closure $\bar{\mathscr{R}}$ is contained in $\mathscr{G}$ such that $\mathscr{T}$ forms a part of the boundary of $\mathscr{R}$. Let $\mathscr{U}^l$ denote the nonvoid class of functions that are piecewise $C^{(1)}$ in x on $\bar{\mathscr{R}}$ with their range in E^{r^l} space and satisfy the condition $U^l(t,x) \in \Omega^l(t,x)$ for all $(t,x) \in \bar{\mathscr{R}}$ for every $l = 1, 2, \cdots, N$.

Let $U^l \in \mathscr{U}^l$, $l = 1, \cdots, N$, and consider the differential equation

$$\dot{x} = G(t, x, \mathbf{U}(t,x)) \tag{2.1}$$

with the initial condition $x(\tau) = \xi$.

Solutions to this may bifurcate or coalesce at (τ, ξ), points of discontinuity of any one U^l or a number of them. If (τ, ξ) be a point such that every U^l is continuous in a neighborhood of (τ, ξ), then there is a unique solution in a neighborhood of (τ, ξ).

We say that $U^l \in \mathscr{U}^l$, $l = 1, \cdots, N$, or $\mathbf{U} \in \mathbf{\mathscr{U}}$ is a playable N-tuple if, for every (τ, ξ) in $\mathscr{R}$, every solution of (2.1) stays in $\mathscr{R}$ and reaches $\mathscr{T}$ in finite time. For each playable N-tuple we can define for the game a real, possibly multivalued, vector function

$$\mathbf{P}(t_o, x_o, \mathbf{U}) = (P^1(t_o, x_o, \mathbf{U}), \cdots, P^N(t_o, x_o, \mathbf{U})) \tag{2.2}$$

in $\mathscr{R}$ with range in E^N space, where

$$P^l(t_o, x_o, \mathbf{U}) = g^l(t_f, x_f) + \int_{t_0}^{t_f} f^l(t, x(t), \mathbf{U}(t, x(t)))\, dt, \tag{2.3}$$

where $x(t)$ is the solution to (2.1) called path and (t_f, x_f) is its point of intersection with $\mathscr{T}$. P^l is the payoff to the lth player.

We assume a maximal nonvoid subclass called the pure strategy set $\mathscr{U}^l_1 \subseteq \mathscr{U}^l$ for each player such that $U^l \in \mathscr{U}^l_1$ implies that the N-tuple $\mathbf{U}$ is playable. The objective of each player is to minimize his payoff function by choosing a pure strategy.

DEFINITION. *Nash equilibrium point*: Let $\mathbf{U}^*$ be a playable N-tuple for which the payoff $\mathbf{P}(t, x, \mathbf{U}^*)$ is single valued in $\mathscr{R}$. Then $\mathbf{U}^*$ is said to be a Nash equilibrium point strategy for the N-person differential game relative to the strategies $\mathscr{U}_1^l$, $l = 1, \cdots, N$, if

$$P^l(t, x, (\mathbf{U}^*; U^l)) \geqq P^l(t, x, \mathbf{U}^*) \tag{2.4}$$

for all $l = 1, 2, \cdots, N$, where

$$(U^{1*}, U^{2*}, \cdots, U^{(l-1)*}, U^l, U^{(l+1)*}, \cdots, U^{N*}) \triangleq (\mathbf{U}^*; U^l). \tag{2.5}$$

The strategies $\mathbf{U}^*$ are said to be optimal in the sense of Nash and $\mathbf{P}(t, x, \mathbf{U}^*)$ denoted by $\mathbf{W}(t, x)$ is called the value function of the game.

The question of under what conditions the preceding assumptions are valid for a game and the analysis of other classes of games will be answered in future.

The paths resulting from $\mathbf{U}^*$ are called optimal paths and denoted by $\phi^*(t, \tau, \xi)$ with a subscript sometimes to distinguish the possible different paths.

For the class of games studied in the paper we assume that (i) the game has an equilibrium point $\mathbf{u}^*$ and (ii) the decomposition associated with $\mathbf{u}^*$ is regular; (iii) if (τ, ξ) is a point of $\mathscr{R}_{ij}^-$, $i = 1, \cdots, x$, $j = 1, \cdots, j_i$, then there is a unique optimal path $\phi^*(t; \tau, \xi)$ in $\mathscr{R}_i$ for $\tau < t < t_{i,j_i}$, where t_{i,j_i} is the final time, and, further the path is never tangent to a manifold $\mathscr{M}_{i,k}$, $k = j, \cdots, j_i$ or to a manifold $\mathscr{N}_{i_1,\cdots,i_p}$. Further properties of $\phi^*(t; \tau, \xi)$ are given in the form of three lemmas in [12].

We make an extensive use of these properties in further discussion.

3. The value function. In this section we derive some properties satisfied by the value function $\mathbf{W}$ for the game. The main objective is to establish a partial differential equation for each of the components of $\mathbf{W}$ corresponding to the different players. To do this, we consider the game as viewed by player l, when all he knows is that perhaps the other players have chosen optimal strategies.

Let (τ, ξ) be a point of $\mathscr{R}_{ij}$, $1 \leqq j \leqq j_i$. Then for the lth player,

$$W^l(\tau, \xi) = g^l(t_{ij_i}, x_{ij_i}) + \int_\tau^{t_{ij}} \bar{f}^l \, dt + \sum_{k=j}^{j_i - 1} \int_{t_{ik}}^{t_{i,k+1}} \bar{f}^l \, dt, \tag{3.1}$$

where (t_{ij_i}, x_{ij_i}) is the end point of the optimal path and the super bar notation is used whenever the arguments are as in the following equation.

$$\bar{f}^l = f^l(t, \phi^*(t; \tau, \xi), \mathbf{U}^*(t, \phi^*(t; \tau, \xi))). \tag{3.2}$$

The bar notation is used in § 4 also in this sense. Thus from the properties of the function f^l and Lemma 3 in [12], it follows that W_ξ^l, W_τ^l exist and are continuous on $\mathscr{R}_{ij}$ with the interpretation that for points on $\mathscr{M}_{i,j-1}$ and $\mathscr{M}_{i,j}$ they represent the unique one-sided limits.

By arguments similar to those of [12] we obtain

$$\begin{aligned} W_\tau^l(\tau, \xi) &\leqq f^l(\tau, \xi, (\mathbf{U}^*(\tau, \xi); U^l(\tau, \xi))) \\ &\quad + W_\xi^l(\tau, \xi)G(\tau, \xi), (\mathbf{U}^*(\tau, \xi); U^l(\tau, \xi)), \end{aligned} \tag{3.3}$$

with the equality holding for

$$U^l(\tau, \xi) = U^{l*}(\tau, \xi),$$

that is

$$W_\tau^l(\tau, \xi) = f^l(\tau, \xi, \mathbf{U}^*(\tau, \xi)) + W_\xi^l(\tau, \xi)G(\tau, \xi, \mathbf{U}^*(\tau, \xi)). \tag{3.4}$$

The relations (3.3) and (3.4) hold for each player $l = 1, 2, \cdots, N$ by similar arguments. The implication of the inequality (3.3) will be discussed more fully in § 4.

By defining $\mathbf{u} = \mathbf{U}(\tau, \xi)$ and

$$H^l(t, x, \mathbf{u}, \lambda^l) = f^l(t, x, \mathbf{u}) + \lambda^l G(t, x, \mathbf{u}), \tag{3.5}$$

(3.4) can be written as

$$H^l(t, x, \mathbf{U}^*, W^l_x(t, x)) + W^l_t(t, x) = 0 \tag{3.6}$$

which is a Hamilton–Jacobi equation.

The preceding results, that is, (3.3), (3.4) and (3.6), which hold for (τ, ξ) in the interior of $\mathscr{R}_{ij}$, also apply by continuity to the appropriate one-sided limits of W^l_τ and W^l_ξ if (τ, ξ) is a point of a manifold $\mathscr{M}_{ij}$. By similar arguments as in [12] it can be shown that W^l_τ and W^l_ξ are continuous across $\mathscr{M}_{ij}$ if all the other players excepting l use strategies continuous across $\mathscr{M}_{ij}$.

We put down all these results in the form of the following theorem.

THEOREM 1. *The value function* $\mathbf{W}$ *consisting of* N *components corresponding to the different players for the game are continuous on* $\mathscr{R}$. *On each* $\mathscr{R}_{ij}$ *the functions* $\mathbf{W}_t$ *and* $\mathbf{W}_x$ *exist, are continuous and have continuous extentions to* $\mathscr{R}_{ij}$. *If* $\mathscr{M}_{ij}$ *is a manifold of discontinuity of only one player, say l-th, then* W^l_t *and* W^l_x *are continuous at points of* $\mathscr{M}_{ij}$. *The function* $\mathbf{W}$ *satisfies* (3.3) *and* (3.4), *respectively, at all points of* $\mathscr{R} \cup \mathscr{T}$, *provided we interpret* $\mathbf{W}_t$, $\mathbf{W}_x$ *and* $\mathbf{U}^*$ *as the appropriate limits at points of* $\mathscr{T}$, $\mathscr{M}_{ij}$ *and* $\mathscr{N}_{i_1, \cdots, i_k}$ *manifolds. At a manifold* $\mathscr{M}_{ij}$, *where only* U^{l*} *is discontinuous,* (3.3) *and* (3.4) *hold for* U^{l*}_{ij} *and* $U^{l*}_{i,j+1}$. *Finally* W *satisfies the Hamilton–Jacobi equation* (3.6).

4. The adjoint variables or Lagrange multipliers. In this section we shall introduce a set of Lagrange multipliers for each player, say λ^l for the lth player and study their relationship to W^l_t and W^l_x for the same player.

Let (τ, ξ) be a point of $\mathscr{R}_{ij}$ and on the interval $t_{i,j-1}(\tau, \xi) \leqq t \leqq t_{ij_i}(\tau, \xi)$. Consider the differential equation

$$\dot{\lambda}^l = -\left(\bar{H}^l_x + \sum_{m=1}^{N} \bar{H}^l_{u^m} \bar{U}^{m*}_x\right), \tag{4.1}$$

where H^l and the bar notation are defined in § 3, with the initial condition $\lambda(t_{ij_i}) = \lambda_{j_i}$.

We shall introduce some notations:

$$\begin{aligned}
(p^-_{ik}) &= (t_{ik}, x_{ik}, \mathbf{U}^*_{ik}(t_{ik}, x_{ik})),\\
(p^+_{ik}) &= (t_{ik}, x_{ik}, \mathbf{U}^*_{ik+1}(t_{ik}, x_{ik})),\\
(\pi^{l^-}_{i,k}) &= t_{ik}, x_{ik}, \mathbf{U}^*_{ik}(t_{ik}, x_{ik}), \lambda^{l^-}),\\
(\pi^{l^+}_{i,k}) &= (t_{ik}, x_{ik}, \mathbf{U}^*_{ik+1}(t_{ik}, x_{ik}), \lambda^{l^+}),
\end{aligned} \tag{4.2}$$

which will be used in the sequel.

Consider the following system of linear equations in the components of λ_{j_i}:

$$g^l_\sigma + f^l(p_{ij_i})\frac{\partial T_{ij_i}}{\partial \sigma} + \lambda^l_{j_i}\left(G(p_{ij_i})\frac{\partial T_{ij_i}}{\partial \sigma} - \frac{\partial X_{ij_i}}{\partial \sigma}\right) = 0, \tag{4.3}$$

where the value of σ corresponds to $t_{ij_i}(\tau, \xi)$ and $x_{ij_i}(\tau, \xi)$ (which is the end point of the optimal path from (τ, ξ)). By Lemma 3 in [12] (4.3) defines $\lambda^l_{j_i}$ uniquely as a continuous function of (τ, ξ) on $\mathscr{R}_{ij}$. Equation (4.3) can also be written in terms of H^l as follows:

$$g^l_\sigma + H^l(\pi^l_{ij_i})\frac{\partial T_{ij_i}}{\partial \sigma} - \lambda^l_{j_i}\frac{\partial X_{ij_i}}{\partial \sigma} = 0. \tag{4.4}$$

Post-multiplying (4.4) by $\partial\sigma/\partial\xi$ and using results in Lemma 3 in [12] we obtain

$$g^l_\sigma \frac{\partial \sigma}{\partial \xi} + H^l(\pi^l_{ij_i})\frac{\partial t_{ij_i}}{\partial \xi} - \lambda^l_{j_i}\frac{\partial x_{ij_i}}{\partial \xi} = 0. \tag{4.5}$$

Since $\lambda^l_{j_i}$ and t_{ij_i} are continuous functions of (τ, ξ) on $\mathscr{R}_{ij}$ it follows that the solution to (4.1) with the initial condition $\lambda^l(t_{ij_i}) = \lambda^l_{j_i}$ can be written as

$$\lambda^l = \lambda^l(t; t_{ij_i}, \lambda_{j_i}) = \lambda^l(t; \tau, \xi).$$

Now with the notation in (4.2) we can define the corner conditions in any of the following three equivalent forms:

$$\begin{aligned} -\lambda^{l-}_k\left[G(p^-_{ik})\frac{\partial T_{ik}}{\partial \sigma} - \frac{\partial X_{ik}}{\partial \sigma}\right] & \\ = [f^l(p^-_{ik}) - f^l(p^+_{ik})]\frac{\partial T_{ik}}{\partial \sigma} &- \lambda^{l+}_k\left[G(p^+_{ik})\frac{\partial T_{ik}}{\partial \sigma} - \frac{\partial X_{ik}}{\partial \sigma}\right], \\ & k = j, \cdots, j_{i-1}, \end{aligned}$$

$$H^l(\pi^{l-}_{ik})\frac{\partial T_{ik}}{\partial \sigma} - \lambda^{l-}_k\frac{\partial X_{ik}}{\partial \sigma} = H^l(\pi^{l+}_{ik})\frac{\partial T_{ik}}{\partial \sigma} - \lambda^{l+}_k\frac{\partial X_{ik}}{\partial \sigma}, \tag{4.6}$$

or

$$H^l(\pi^{l-}_{ik})\frac{\partial t_{ik}}{\partial \xi} - \lambda^{l-}_k\frac{\partial x_{ik}}{\partial \xi} = H^l(\pi^{l+}_{ik})\frac{\partial t_{ik}}{\partial \xi} - \lambda^{l+}_k\frac{\partial x_{ik}}{\partial \xi}.$$

Since the left-hand limits of λ^l can be uniquely determined in terms of the right-hand limits at the $\mathscr{M}_{ik}$ manifolds, the function $\lambda^l(t; \tau, \xi)$ is defined for (τ, ξ) in $\mathscr{R}_{ij}$ and all $t_{i,k-1} < t < t_{ik}$, $k = j, \cdots, j_i$, and is continuous in its arguments, possesses unique one-sided limits at the manifolds of discontinuity $\mathscr{M}_{ik}$ and satisfies the transversality and corner conditions (4.5) and (4.6).

Also from the properties of ϕ^* and λ functions it follows that $\lambda(t; \tau, \xi) = \lambda(t; \hat{\tau}, \hat{\xi})$ for $\tau < \hat{\tau} < t$ and $\hat{\xi} = \phi^*(\hat{\tau}; \tau, \xi)$.

Now consider the function W^l on $\mathscr{R}_{ij}$.

$$W^l_\xi(\tau,\xi) = g^l_\sigma\frac{\partial^\sigma}{\partial\xi} + f^l(p_{ij_i})\frac{\partial t_{ij_i}}{\partial\xi} + \sum_{k=1}^{j_i-1}[f^l(p_{ik-0}) - f^l(p_{ik+0})]\frac{\partial t_{ik}}{\partial\xi} \tag{4.7}$$
$$+ \left(\int_\tau^{t_{ij}} + \sum_{k=j}^{j_i-1}\int_{t_{ik}}^{t_{i,k+1}}\right)\left(\bar{f}^l_x + \sum_{m=1}^{N}\bar{f}^l_{u^m}\bar{U}^{mx}_x\right).$$

By arguments of [12] applied to $\mathbf{W}$ and λ^l we have

$$W^l_\xi(\tau,\xi) = \lambda^l(\tau;\tau,\xi) \tag{4.8}$$

and

$$W^l_x(t,x) = \lambda^l(t;t,x) = \lambda^l(t;\tau,\xi), \tag{4.9}$$

where $x = \phi(t,\tau,\xi)$.

The properties of λ^l are thus given as those satisfied by W^l_x in Theorem 1.

The preceding analysis can be carried out to each player $l = 1,2,\cdots,N$.

Now, we shall discuss the implications of the inequalities (3.3). Let (t,x) be a point on the optimal path from (τ,ξ), thus $x = \phi^*(t,\tau,\xi)$. It then follows from (3.3) and (4.9) that

$$\begin{aligned} H^l(t,x,(\mathbf{u}^*;u^l),\lambda^l) &= f^l(t,x,(\mathbf{u}^*;u^l)) + \lambda^l G(t,x,(\mathbf{u}^*;u^l)) \\ &= f^l(t,x,(\mathbf{u}^*;u^l)) + W^l_x G(t,x,(\mathbf{u}^*;u^l)) \\ &\geqq f^l(t,x,\mathbf{u}^*) + W^l_x G(t,x,\mathbf{u}^*) \\ &= H^l(t,x,\mathbf{u}^*,\lambda^l) \\ &= -W^l_t(t,x), \end{aligned} \tag{4.10}$$

where $x = \phi(t;\tau,\xi)$, $\lambda^l = \lambda^l(t;\tau,\xi)$ and

$$u^l \in E[u^l | u^l = U^l(t,x), u^l \in \mathscr{U}^l_1].$$

At points $t = t_{ik}$ (4.10) holds for one-sided limits.

Thus (4.10) shows that at every point (t,x) the game $\Gamma(t,x)$ with payoffs defined by $H^l(t,x,\mathbf{u},\lambda^l)$ has a pure strategy equilibrium point $\mathbf{u}^*$. The value of this game $\Gamma(t,x)$ is

$$[H^l_u(t,x,\mathbf{u}^*,\lambda^l),\cdots,H^N(t,x,\mathbf{u}^*,\lambda^N)] = -\mathbf{W}_\tau(t,x). \tag{4.11}$$

This is an important result as the optimal strategy for the original game is necessarily the optimal strategy for the extensive game in the Hamiltonians along the optimal path of the original game.

We summarize the principal results in the following theorem.

THEOREM 2. *Let $\phi^*(t;\tau,\xi)$ be the optimal path from a point (τ,ξ) in $\mathscr{R}_{ij}$. Then there exist functions $\lambda^l(t;\tau,\xi)$, $l = 1,2,\cdots,N$, defined for $\tau < t < t_{ij}$ (τ,ξ) and $t \neq t_{ik}$, $k = j, j+1,\cdots,j_i - 1$, such that the following hold:*

(i) *Each λ^l is piecewise continuous on its domain of definition, and at points t_{ik} possesses unique one-sided limits λ^{l-}_k and λ^{l+}_k.*

(ii) *The functions* λ^l, $l = 1, \cdots, N$ *and* ϕ^* *satisfy the following system of differential equations*:

$$\dot{x} = H^l_{\lambda^l}(t, x, \mathbf{U}^*(t, x), \lambda^l) \qquad (4.12)$$

$$\dot{\lambda}^l = -\frac{\partial H^l}{\partial x}(t, x, \mathbf{U}^*(t, x), \lambda^l),$$

where

$$\frac{\partial H^l}{\partial x} = H^l_x + \sum_{m=1}^{N} H^l_{u^m} U^{m*}_x \qquad (4.13)$$

and at $t = t_{ik}$, $k = j, \cdots, j_i - 1$ *the equations hold for the one-sided limits. These equations are the characteristic equations of the Hamilton–Jacobi equations.*

(iii) *If* $\mathcal{M}_{ik}$, $j \leqq k \leqq j_i - 1$, *is a manifold of discontinuity of only one of the functions say* U^{l*}, *then* λ^l *is continuous at* $t = t_{ik}$, *i.e.,* $\lambda^{l+}_k = \lambda^{l-}_k$. *Otherwise* (4.5) *holds.*

(iv) *At* $t = t_{ij_i}$ *the transversality relation* (4.5) *holds.*

(v) *If* $x = \phi^*(t; \tau, \xi)$, $t \geqq T$ *then*

$$W^l_x(t, x) = \lambda^l(t; \tau, \xi). \qquad (4.14)$$

(vi) *For all* $\tau \leqq t < t_{ij_i}$ *and* $t \neq t_{ik}$, $k = j, \cdots, j_i - 1$,

$$H^l(t, \phi^*(t), (u^*(t); u^l(t)), \lambda^l(t)) \geqq H^l(t, \phi^*(t), \mathbf{u}^*(t), \lambda^l(t)), \qquad (4.15)$$

where $\mathbf{u}^*(t) = \mathbf{U}^*(t, \phi^*(t))$ *and* $u^l(t) = U^l(t, \phi^*(t))$ *for some* $U^l \in \mathcal{U}^l_1$.

We shall give two corollaries for the case when the sets $\Omega^l(t, x)$ for each player are defined by systems of inequalities, i.e., let $K^l(t, x, u^l)$ be a function with domain in $(t, x, \mathcal{U}^l)$ space and range in E^{p^l}. A vector u^l is in $\Omega^l(t, x)$ if and only if $k^l(t, x, u^l) \geqq 0$. Further let K^l satisfy the constraint condition that if $p^l > r^l$, then at each point $(t, x, \mathbf{u}^l)$ at most r^l components of K^l can vanish and the matrix $[\partial K^{li}/\partial u^{lj}]$ formed from these components of K^l that vanish at (t, x, u^l) has maximum rank at (t, x, u^l), this being true for all $l = 1, 2, \cdots, N$.

COROLLARY 1. *Let* $\Omega^l(t, x)$ *be given by a system of inequalities* $K^l(t, x, u^l) \geqq 0$ *for* $l = 1, 2, \cdots, N$, *where* K^l *satisfy the constraint condition. Then there exist functions* $\mu^l(t; \tau, \xi)$ *defined for* $\tau \leqq t \leqq t_{ij_i}$ *and* $t \neq t_{ik}$ *such that the following hold. At all points* $(t, \phi^*(t), \mathbf{u}^*(t), \lambda^l(t))$, *where* $\mathbf{u}^*(t) = \mathbf{U}^*(t, \phi^*(t))$,

$$H^l_{u^l} + \mu^l K^l_{u^l} = 0, \qquad (4.16)$$

$$\mu^{l_i} K^{l_i} = 0 \qquad \text{for each } l_i,$$

$$\mu^l \leqq 0.$$

COROLLARY 2. *Let* (τ, ξ) *be a point of* $\mathcal{R}_{ij}$. *Suppose that on each interval* $t_{i,k-1} \leqq t \leqq t_{ik}$, $k = j, \cdots, j_i$ *the components* K^{l_i} *of* K^l *do not change, for all* $l = 1, 2, \cdots, N$. *Then* μ^l *are continuous functions of* $(t; \tau, \xi)$ *for* (τ, ξ) *in* $\mathcal{R}_{ij}$ *and* $t_{i,k-1} \leqq t < t_{ik}$ *and have one-sided limits at the end points* $t_{i,k-1}$ *and* t_{ik}. *At the points* t_{ik}, (4.16) *hold for the one-sided limits.*

The second equation in (4.12) can be replaced by the equation

$$\dot{\lambda}^l = -H_x^l - \sum_{\substack{m=1\\ m\neq l}}^{N} H_{u^m}^l U_x^m - \mu^l K_x^l. \tag{4.17}$$

The proofs of these corollaries follow exactly on the lines of [12] and will not be repeated here.

5. Conclusions. In the course of the paper we have established the necessary conditions for the optimal strategies for a class of N-person differential games. Some of the assumptions made such as n-dimensionality of the terminal surface and nontangency requirements of the optimal paths [13, p. 342] are rather restrictive. These and other problems like the existence of equilibrium points and particularly of strategies which induce regular decomposition on the (t, x) space of interest exist and remain to be investigated.

REFERENCES

[1] L. A. Petrosyan, *On a family of differential games of survival in the space R^n*, Dokl. Akad. Nauk. SSR, 161 (1965), pp. 52–54.

[2] ———, *Differential games of survival with many participants*, Ibid., 161 (1965), pp. 285–287.

[3] ———, *On the reduction of the solution of a game of pursuit and survival to the solution of the Cauchy problem for a first order partial differential equation*, Dokl. Akad. Nauk. Armajn. SSR, 40 (1965), pp. 193–196.

[4] ———, *A game of pursuit in the half plane*, Ibid., 40 (1965), pp. 265–269.

[5] J. Von Neumann and O. Morgenstern, *Theory of Games and Economic Behaviour*, Princeton University Press, Princeton, 1947.

[6] J. Nash, *Non-cooperative games*, Ann. of Math., 54 (1951), pp. 286–295.

[7] H. W. Kuhn and A. W. Tucker, eds., *Contributions to the Theory of Games*, Princeton University Press, Princeton, 1950 and 1953.

[8] M. Dresher, A. W. Tucker and P. Wolfe, eds., *Contributions to the Theory of Games*, vol. III, Annals of Math. Studies No. 39, Princeton University Press, Princeton, 1957.

[9] A. W. Tucker and R. D. Luce, eds., *Contributions to the Theory of Games*, vol. IV, Annals of Math. Studies No. 40, Princeton University Press, Princeton, 1959.

[10] M. Dresher, L. S. Shapley and A. W. Tucker, eds., *Advances in Game Theory*, Annals of Math. Studies No. 52, Princeton University Press, Princeton, 1964.

[11] L. D. Berkovitz, *A variational approach to differential games*, Advances in Game Theory, Annals of Math. Studies No. 52, M. Dresher, L. S. Shapley and A. W. Tucker, eds., Princeton University Press, Princeton, 1964.

[12] ———, *Necessary conditions for optimal strategies in a class of differential games and control problem*, this Journal, 5 (1967), pp. 1–24.

[13] ———, *A survey of differential games*, Mathematical Theory of Control, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, New York, 1967.

SIAM J. Control
Vol. 7, No. 4, November 1969

STABILITY, INSTABILITY, INVERTIBILITY AND CAUSALITY*

JAN C. WILLEMS†

1. Introduction. A number of interesting stability criteria for feedback systems have recently appeared in the control theory literature. The procedures used in proving these criteria can roughly be divided into three classes; the first based on Popov-like methods, the second using Lyapunov theory with Lyapunov functions derived from spectral factorizations or Riccati-type algebraic matrix equations, and the third treating the stability problem from a functional analysis point of view.

Each of these methods has relative merits, e.g., the Lyapunov methods seem to be the only ones which allow us to obtain an estimate of the domain of attraction in the case of nonglobal stability. However, the method based on functional analysis appears to be the more satisfactory one, in view of the essential simplicity, of the intuitive nature of the results (loop gain less than one, passivity conditions), and of the fact that it unifies the various criteria (as, e.g., the circle criterion and the Popov criterion). It therefore deserves more investigation and exposition than it has thus far been given.

A peculiarity of this method, as presently employed, is that most of the analysis and estimates have to be made on extended spaces which, although derived from normed spaces, are themselves not normed. This entails in general rather cumbersome mathematical manipulations. One however suspects the introduction of extended spaces to be merely a tool which enables one to make a satisfactory definition of stability, and that the stability properties of the system only depend on how the operators in the forward and the feedback loop operate on elements of the nonextended space. Another aspect which has not been explored as yet is the use of these functional analysis methods to generate instability criteria. The present paper investigates these facets of the functional analysis methods as applied in stability analysis. More specifically, the stability and instability properties of the feedback system are expressed in terms of the properties of the inverse of the closed loop operator on the nonextended space, and procedures for generating instability criteria are described.

When considering the stability properties of a system defined through an input-output relation, one is generally asked to determine conditions under which an operator, F, qualifies as a bounded transformation on appropriate normed spaces. This then yields a bound on the norm of the output, $y = Fu$, in terms of the norm of the input, u. The question of stability of a *feedback* system, however, is a more intricate matter both at a conceptual level and at the point of deriving specific conditions. The reason for this difficulty is that the equations governing a feedback system give the output, y, in terms of the input, u, only through an *implicit* equation of the type $u = Fy$.

* Received by the editors November 21, 1968, and in revised form April 16, 1969.

† Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139. This work was supported in part by the National Aeronautics and Space Administration under Grant NGR-22-009(124) with the M.I.T. Electronic Systems Laboratory and by the U.S. Army Research Office, Durham.

Sandberg [1], [2] and Zames [3], [4] introduced the idea of extended spaces in order to give a definition of stability which appears, at least for causal systems, to be entirely satisfactory. The output, y, is then allowed to belong to a larger space (i.e., the extended space) than the input (which is for stability purposes assumed to belong to the nonextended space). The fact that the output should actually belong to the nonextended space is then taken as the basic requirement for stability. This a priori enlargement of the solution space has been the key to the very successful application of functional analysis to the stability of feedback systems, and has therewith provided a class of rather elegant applications of classical analysis techniques to modern control engineering problems.

The feedback equation $u = Fy$ is in this paper considered as an equation relating a priori inputs u in the extended space to outputs y in the extended space. In order to get the ideas of the paper through, assume, for the purpose of this discussion, that the time interval of definition is $(-\infty, +\infty)$, that F is a causal linear operator from

$$L_{2e}(-\infty, +\infty) \triangleq \left\{x(t) \middle| \int_{-\infty}^{T} |x(t)|^2 \, dt < \infty \text{ for all finite } T\right\}$$

into itself, and that u and y are real-valued functions in $L_{2e}(-\infty, +\infty)$. The results, which are shown schematically in Fig. 1, can then roughly be summarized as follows: Consider in the class of causal linear operators from $L_{2e}(-\infty, +\infty)$ into itself the subset of those for which the zero solution can be continued in a causal way in $L_{2e}(-\infty, +\infty)$; i.e., for any $u \in L_{2e}(-\infty, +\infty)$ with $u(t) = 0$ for $t \leqq T$, there exists a unique $y \in L_{2e}(-\infty, +\infty)$ with $y(t) = 0$ for $t \leqq T$, such that $u = Fy$, and this y is related to u in a nonanticipatory sense. The result then states that F defines a stable feedback system *if and only if* (i) F is invertible on $L_2(-\infty, +\infty)$ and (ii) this inverse is causal on $L_2(-\infty, +\infty)$. Considering the algebra of linear operators from $L_2(-\infty, +\infty)$ into itself and considering the causal operators as a subalgebra, we then state the result that F defines a stable feedback system if and only if F is a regular (i.e., invertible) element of this subalgebra.

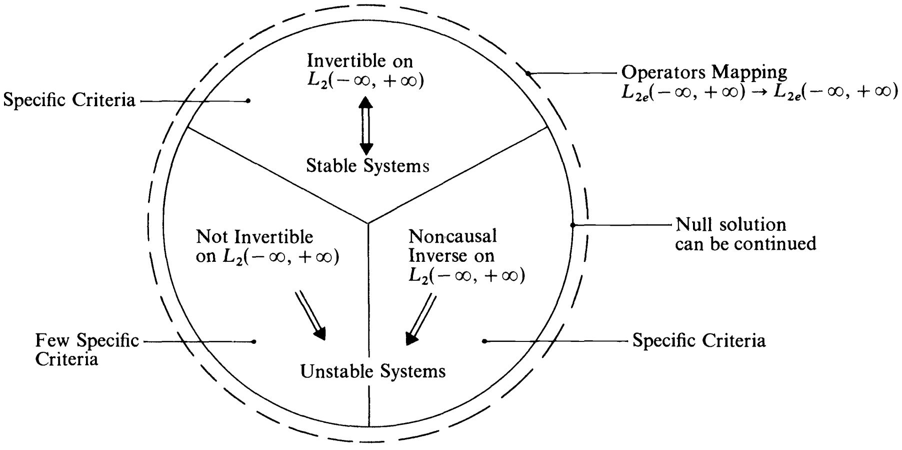

Invertible on
$L_2(-\infty, +\infty)$
Stable Systems
Operators Mapping
$L_{2e}(-\infty, +\infty) \to L_{2e}(-\infty, +\infty)$
Specific Criteria
Not Invertible on $L_2(-\infty, +\infty)$
Noncausal Inverse on $L_2(-\infty, +\infty)$
Null solution can be continued
Few Specific Criteria
Specific Criteria
Unstable Systems

FIG. 1. *Summary of results when the time interval of definition is* $(-\infty, +\infty)$

Not only has this algebraic characterization a great deal of mathematical appeal, but it also identifies two clearly distinct methods for generating instability criteria: the first based on showing that the inverse on $L_2(-\infty, +\infty)$ does not exist, and the second based on showing that the inverse exists but is not causal. The former method remains largely unexplored and is even in the present paper only exploited in a very restricted setting. The latter method appears to be the setting in which instability criteria will most easily be generated. In fact, the frequency-domain instability criteria which have thus far appeared in the literature can actually be put into this mathematical framework.[1] Note finally the interesting analogy between this equivalence of noncausality and instability and the fact that one can choose to interpret, depending on the region of convergence, a bilateral Laplace transform with a pole in the right half-plane either as the transform of an impulse response which does not vanish for negative time or as the transform of an impulse response whose integral diverges.

The paper also treats the case when the time interval of definition is $[T_0, \infty)$ and the results are then best summarized as follows (see Fig. 2): Consider the class of causal linear operators from $L_{2e}(T_0, \infty)$ into itself which have a causal inverse on $L_{2e}(T_0, \infty)$, i.e., any input $u \in L_{2e}(T_0, \infty)$ yields a unique output $y \in L_{2e}(T_0, \infty)$ with $Fy = u$, and this y is related to u in a nonanticipatory sense. The result then states that F defines a stable feedback system if and only if F is invertible on $L_2(T_0, \infty)$. This inverse will then necessarily be causal since it is the restriction of the inverse on $L_{2e}(T_0, \infty)$ to $L_2(T_0, \infty)$. The paper also gives a procedure to establish the noninvertibility of a class of operators on $L_2(T_0, \infty)$. This is done by suitably extending F to an operator, F', on $L_2(-\infty, +\infty)$ and showing that F' has a noncausal inverse on $L_2(-\infty, +\infty)$. It is then shown that this implies that F is not invertible on $L_2(T_0, \infty)$.

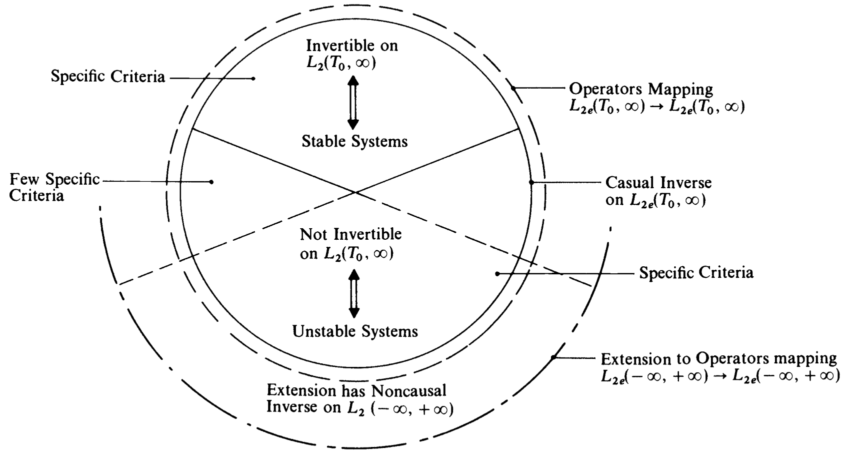

Invertible on $L_2(T_0, \infty)$
Stable Systems
Specific Criteria
Operators Mapping $L_{2e}(T_0, \infty) \to L_{2e}(T_0, \infty)$
Few Specific Criteria
Casual Inverse on $L_{2e}(T_0, \infty)$
Not Invertible on $L_2(T_0, \infty)$
Unstable Systems
Specific Criteria
Extension to Operators mapping $L_{2e}(-\infty, +\infty) \to L_{2e}(-\infty, +\infty)$
Extension has Noncausal Inverse on $L_2(-\infty, +\infty)$

FIG. 2. *Summary of results when the time interval of definition is* (T_0, ∞)

The analysis behind the ensuing demonstrations of noncausality of certain inverses are lengthy and somewhat subtle.

[1] At least when the system is open loop stable. The open loop unstable case, which is not void of interest, has not been considered here.

The general results are finally applied to a particular class of feedback systems and lead to a generalization of the Nyquist criterion and the circle criterion.

The paper is divided as follows:

1. Introduction
2. Mathematical preliminaries
3. Problem formulation
4. Interrelations between invertibility, stability and continuity
5. Existence and uniqueness of solutions
6. Stability and continuity
7. Instability
8. An example
9. Concluding remarks

The material in §§ 2 and 3 are standard in papers treating stability from the point of view adopted here. The new results are contained mainly in §§ 4, 7 and 8 and will be the ones of primary interest to the readers already familiar with the definitions of § 3. Sections 5 and 6 have been included mainly for the sake of completeness, in order to make the paper a reasonably self-contained treatment of stability and instability. In any case, §§ 5, 6 and 7 are independent and the instability part of the paper with its relation to causality can be grasped by reading only §§ 3, 4 and 7.

2. Mathematical preliminaries. A certain familiarity with the notions of metric spaces, linear spaces, inner product spaces, norms and inner products is assumed. The norm of an element x of a normed linear space V will be denoted by $\|x\|_V$, and the subscript will be deleted whenever no confusion can occur. The same holds for inner products.

A normed linear space is called a *Banach space* if it is complete in the metric induced by its norm. (A metric space is called *complete* if every Cauchy sequence converges.)

An *operator*, or *transformation*, F, from a space X into a space Y is a law which associates with every element $x \in X$ an element $Fx \in Y$. F is said to be *invertible* if it is one-to-one and onto. I denotes the identity operator from a space into itself. An operator L from a linear space X into a linear space Y, defined over the same field as X, is said to be *linear* if $L(\alpha x_1 + \beta x_2) = \alpha L x_1 + \beta L x_2$ for all $x_1, x_2 \in X$ and scalars α, β.

Let F be an operator from a normed linear space X_1 into a normed linear space X_2. F is said to be *bounded* on X_1 if

$$\sup_{\substack{x \in X_1 \\ x \neq 0}} \frac{\|Fx\|_{X_2}}{\|x\|_{X_1}} < \infty .$$

This least upper bound will be denoted by $\|F\|$. F is said to be *Lipschitz continuous* on X_1 if

$$\sup_{\substack{x, y \in X_1 \\ x \neq y}} \frac{\|Fx - Fy\|_{X_2}}{\|x - y\|_{X_1}} < \infty .$$

This least upper bound will be denoted by $\|\tilde{F}\|$. Boundedness, continuity at some point and Lipschitz continuity are equivalent for linear operators.

Let V denote a finite-dimensional normed vector space and let (T_1, T_2), $-\infty \leqq T_1 \leqq T_2 \leqq \infty$, be an interval in the real numbers. $L_p(T_1, T_2)$, $1 \leqq p < \infty$, denotes the linear space of all V-valued measurable functions for which the pth power of the pointwise norm has a finite Lebesgue integral. $L_\infty(T_1, T_2)$ denotes all V-valued measurable functions whose pointwise norm is almost everywhere smaller than some number. The norm[2] on $L_p(T_1, T_2)$ is defined as

$$\left(\int_{T_1}^{T_2} \|x(t)\|_V^p \, dt\right)^{1/p}$$

for $1 \leqq p < \infty$, and as the infimum of the set $\{M| \, \|x(t)\|_V \leqq M \text{ a.e. in } (T_1, T_2)\}$ for $L_\infty(T_1, T_2)$. L_p-spaces are complete in the metric induced by their norm, and are thus Banach spaces.

The natural setting for studying bounded linear operators from a Banach space into itself is a Banach algebra. This point of view will be an advantageous one for the purposes of the paper, and the appropriate notions are therefore introduced below.

A *Banach algebra* is a normed linear space $\mathscr{A}$ over the real or complex field which is complete in the topology induced by its norm, and a map (*multiplication*) from $\mathscr{A} \times \mathscr{A}$ into $\mathscr{A}$. This multiplication is associative and is distributive with respect to addition, i.e., $x(yz) = (xy)z$, $(x + y)z = xz + yz$, $x(y + z) = xy + xz$ for all x, y and $z \in \mathscr{A}$. It is related to scalar multiplication by $\alpha(xy) = x(\alpha y) = (\alpha x)y$, and to the norm on $\mathscr{A}$ by $\|xy\| \leqq \|x\| \, \|y\|$ for all $x, y \in \mathscr{A}$ and scalars α. A Banach algebra is said to have a *unit* if there exists an element $e \in \mathscr{A}$ such that $xe = ex = x$ for all $x \in \mathscr{A}$. An element x of a Banach algebra with a unit is said to be *invertible* if there exists an element $x^{-1} \in \mathscr{A}$ such that $xx^{-1} = x^{-1}x = e$. It is easily seen that there exists at most one unit and one inverse. If x and y are invertible, so is xy, and $(xy)^{-1} = y^{-1}x^{-1}$.

A subset $\mathscr{A}^+ \subset \mathscr{A}$ is said to be a *subalgebra* of a Banach algebra if $\mathscr{A}^+$ is itself a Banach algebra under the operations induced by $\mathscr{A}$.

An essential fact which will be needed is the completeness and the algebraic structure of the bounded linear transformations on a Banach space. More precisely, let $\mathscr{B}$ be a Banach space and let $\mathscr{L}(\mathscr{B}, \mathscr{B})$ denote the linear space of all bounded linear transformations from $\mathscr{B}$ into itself. Let multiplication on $\mathscr{L}(\mathscr{B}, \mathscr{B})$ be defined as composition of maps, and let the norm on $\mathscr{L}(\mathscr{B}, \mathscr{B})$ be defined as $\|L\| = \sup_{x \in \mathscr{B}, \|x\| = 1} \|Lx\|$ for $L \in \mathscr{L}(\mathscr{B}, \mathscr{B})$. Then $\mathscr{L}(\mathscr{B}, \mathscr{B})$ forms a Banach algebra with a unit. The only property in this statement which is not immediate is the completeness of $\mathscr{L}(\mathscr{B}, \mathscr{B})$. A proof of this well-known fact can be found, e.g., in [9, p. 61]. For more details on the preliminaries, see, e.g., [9] or [10].

3. Formulation of the problem. In this section the functional equations describing a feedback system are introduced and the definition of a solution is given. Other important concepts which are formally defined are those of stable, unstable, and continuous feedback loops. These definitions involve the idea of extended spaces as introduced in this context by Sandberg [1], [2] and Zames [3],

[2] Under the assumption that two functions are considered equal if they are equal almost everywhere. Hence whenever L_p-spaces are involved, equality of elements should be interpreted as equality of the functions almost everywhere.

[4]. Finally a number of important assumptions on the operators in the forward and feedback loop and on the solution space are made.

Let R denote the real line and let $S \subset R$ be given. The set S will be referred to as the *time interval of definition* and is typically $(-\infty, +\infty)$, $[T_0, \infty)$, the integers or the integers larger than some given real number.

For ɛny linear space V, let $Y(V)$ denote the linear space of all maps from S into V, i.e., $Y(V) = \{x | x : S \to V\}$. Every element of $Y(V)$ is thus a V-valued function defined on S. Let P_T, $T \in S$, denote the projection operator from $Y(V)$ into itself defined by

$$P_T y(t) \triangleq \begin{cases} y(t) & \text{for } t \leqq T, \quad t \in S, \\ 0 & \text{for } t \geqq T, \quad t \in S \end{cases}$$

for $y \in Y(V)$. $P_T y$ will be called the *T-truncation* of y. Let $Z \subset Y(V)$. The *extended space*, Z_e, is defined as all elements of $Y(V)$ for which all T-truncations belong to Z as T ranges over S, i.e., $Z_e \triangleq \{y \in Y(V) | P_T y \in Z \text{ for all } T \in S\}$.

3.1. The feedback equation. Let $W_1 \subset Y(V_1)$ and $W_2 \subset Y(V_2)$ be given normed linear spaces, and let W_{1e}, W_{2e} denote their extensions. Consider now the feedback system shown in Fig. 3. The inputs and the signals in the loop are assumed to be V_1- or V_2-valued and (as functions defined on S) to belong to W_{1e} or W_{2e}. Let G_1 and G_2 denote operators respectively from W_{1e} into W_{2e} and from W_{2e} into W_{1e}, and let $u_1 \in W_{1e}$ and $u_2 \in W_{2e}$ be given. It is convenient to introduce the product space $W = W_1 \times W_2$, with $\|x\|_W = \|(x_1, x_2)\|_W \triangleq \|x_1\|_{W_1} + \|x_2\|_{W_2}$. G then denotes the operator from W_e into itself defined by $Ge = (G_2 e_2, -G_1 e_1)$ for $e = (e_1, e_2) \in W_e$. The element $u = (u_1, u_2) \in W_e$ will be referred to as the *input*.

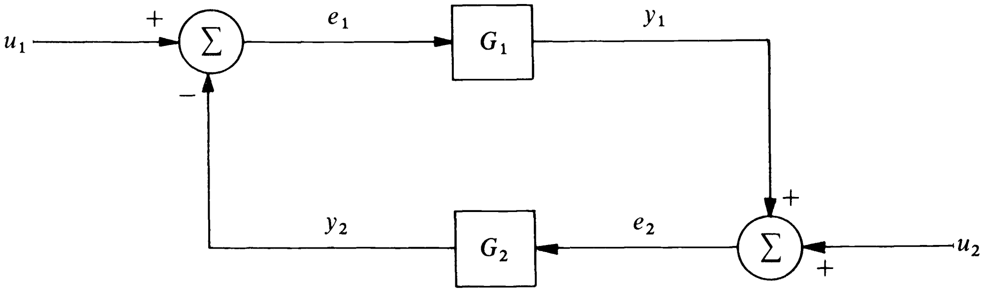


FIG. 3. *The feedback system*

DEFINITION 3.1. An element $e \in W_e$ is said to be a *solution* of the equations describing the feedback loop if $(I + G)e = u$, i.e., $e(t) + (Ge)(t) = u(t)$ for all $t \in S$.

Remark 3.1. If the input-output relation G_1 or G_2 is obtained through a dynamical system description, then the initial conditions can very often be modeled either as equivalent inputs, or as part of the description of G. The first method is certainly to be preferred for the purposes of the paper.

Remark 3.2. The signs in the definition of G were chosen such that $(I + G)e = u$, in accordance with the equations describing the usual unit negative gain feedback system.

3.2. Stability, instability and continuity. The next notions which will be introduced are those of stability and instability. This will be done in terms of convergence in the norm of W.

DEFINITION 3.2. The feedback system under consideration is said to be *stable* if for any given $u \in W$,

(i) any solution $e \in W_e$ actually belongs to W itself,

(ii) there exists a real number K, independent of u, such that $\|e\|_W \leqq K\|u\|_W$.

The feedback system under consideration is said to be *unstable* if it is not stable. It is said to be *continuous* if it is stable and if any input sequence $\{u_n\}$ with $\lim_{n\to\infty} u_n = u$ yields $\lim_{n\to\infty} e_n = e$, where e_n and e are solutions corresponding to u_n and u.

The above definition of stability is rather strong because of condition (ii) and yields a correspondingly weak type of instability: the system could be unstable even if $e \in W$ for any $u \in W$. It will turn out however that for linear systems stability, continuity and stability without condition (ii) are equivalent.

In examples one usually calls a feedback system which is stable in the above sense more explicitly *W-stable.*

3.3. Assumptions. The following restrictive assumptions will be made throughout the paper. As will be indicated in the concluding remarks, some of these assumptions can be relaxed at various points.

A1. The space W satisfies the following conditions:

(i) It is closed under the projection P_T for all $T \in S$, i.e., if $x \in W$ then $P_T x \in W$ for all $T \in S$.

(ii) $$\sup_{T\in S} \|P_T x\|_W = \begin{cases} \|x\|_W & \text{if } x \in W, \\ \infty & \text{if } x \in W_e - W. \end{cases}$$

(iii) The elements of W can be approximated by functions which vanish up to some time $T \in S$. More precisely, let W_T denote the subspace of W defined by $W_T = \{x \in W | P_T x = 0,\ T \in S \text{ given}\}$, and let W_r denote the subspace of W defined by $W_r = \{x \in W | P_T x = 0 \text{ for some } T \in S\}$, i.e., $W_r = \bigcup_{T\in S} W_T$. It is assumed that W_r is *dense* in W. Thus for any $x \in W$, $\inf_{T\in S} \|P_T x\|_W = 0$, i.e., the closure of $W_r = W$.

(iv) W is a Banach space. Moreover, the subspace of W defined by $\{x \in W | x(t) = 0 \text{ for } t \in S, t \notin [T_1, T_2]\}$ is a Banach space for any $T_1, T_2 \in S$, $T_1 \leqq T_2$. In other words, it is assumed that the set of all elements of W which are zero outside any interval $[T_1, T_2]$ is closed in W.

A2. The operator G satisfies the following conditions:

(i) G is a *causal* operator on W_e, i.e., $P_T G$ and P_T commute on W_e and thus $P_T G x = P_T G P_T x$ for all $x \in W_e$. This assumption is equivalent to requiring G_1 and G_2 to be causal on W_{1e} and W_{2e} respectively.

(ii) G maps W into itself, $G(0) = 0$, and G is *continuous* on W.

(iii) G_1 is *bounded* on W_1 and G_2 is *Lipschitz continuous* on W_2. Note that this assumption together with (ii) implies that G is bounded on W.

The causality restriction is of course a natural one if the feedback system describes a physically realizable feedback controller. Many other interesting functional equations can successfully be modeled into a feedback configuration which indeed satisfies the causality condition. This additional structure on the operators appearing in the functional equations and the corresponding particular role played by the "past" as compared to the "future," leads for instance to the possibility of establishing the invertibility of certain operators on extended spaces, although these spaces are in general *not* normed.

The notions on stability, continuity, and instability are, at least from a mathematical point of view, well-defined even when G is not causal. It should be remarked however that in these definitions the "future" again plays a particular role and that therefore these definitions would appear ill conceived when G is itself not causal.

The restriction that W_r be dense in E is, when S is doubly infinite, somewhat more severe than one might like. It is satisfied for l_p- and L_p-spaces with $1 \leqq p < \infty$, but unfortunately excludes for instance $L_\infty(-\infty, +\infty)$.

3.4. Objectives and historical remarks. The remainder of this paper discusses some of the fundamental properties of the feedback system under consideration. In particular, attention is focussed on the relationship between the invertibility on W of the operator $I + G$, and stability, instability and continuity. In analyzing this relationship, the question of existence and uniqueness of solutions is encountered and is therefore briefly treated.

Many previous authors have treated similar problems. More specifically, the questions of existence and uniqueness have been considered in essentially the same setting by Zames in [5], [6]. The question of stability has received a great deal of attention in the last decade. Particularly the work of Sandberg and Zames (see, e.g., [2], [4]) for general functional equations, and of Brockett and Willems [11] for differential equations deserves mention. More recently, a number of instability criteria have appeared in the literature. The results here in fact extend[3] and reinterpret in a functional setting those obtained using Lyapunov theory by Brockett and Lee [12] for feedback systems described by ordinary differential equations.

Since the mathematical approach to the problem used here is inspired by the work of Sandberg and particularly of Zames, there is some unavoidable overlap of results.

4. Interrelations between invertibility, stability and continuity. In this section some general stability and continuity theorems for feedback systems are derived. They expose the interrelationship between invertibility of the closed-loop operator (on the nonextended space) and stability and continuity. The resulting Theorems 4.1 and 4.2 are believed to be of great interest and constitute the main results of the paper. This approach makes the large mathematical literature on invertibility of operators directly applicable to the problem of stability and continuity of feedback systems. Vice versa, it appears important to realize that every stability theorem actually yields an inverse function theorem.

Recall that the equations describing the feedback system could be written as $(I + G)e = u$. Theorem 4.1 presupposes (of course in addition to the assumptions of § 3) that the operator $I + G$ has a causal inverse on W_e. Theorem 3.2

[3] Under the assumptions of § 3.3. Note that particularly the boundedness assumption of G is not needed in [12].

assumes that the operator $I + G$ has a causal inverse on W_{Te} for all $T \in S$, where $W_{Te} \triangleq \{x \in W_e | P_T x = 0, T \in S \text{ given}\}$. As will be discussed in § 5, this invertibility condition on W_{Te} is weak. The invertibility condition on W_e however is weak only when the set S is bounded from below but is rather severe—about equivalent to continuity of the feedback system—when S is *not* bounded from below.

THEOREM 4.1. *Assume that $I + G$ has a causal inverse on W_e. Then the feedback system under consideration is stable if and only if $I + G$ has a causal bounded inverse on W.*

Proof. (i) The condition is sufficient. $I + G$ is thus assumed to have a causal bounded inverse on W. Let $u \in W$ be given, and let $e \in W_e$ be a solution. Then $P_T(I + G)P_T e = P_T u$, and $P_T(I + G)^{-1}P_T u = P_T(I + G)^{-1}P_T(I + G)P_T e = P_T e$. Thus $\|P_T e\|_W \leqq \|(I + G)^{-1}\| \, \|u\|_W$, for any $T \in S$, which yields stability.

(ii) The condition is necessary; the feedback system is thus assumed to be stable. Since $I + G$ has a causal inverse on W_e it is one-to-one and onto on W_e. By stability, $(I + G)e = u$, $u \in W$, and $e \in W_e$ imply that $e \in W$. Hence $I + G$ is one-to-one and onto on W. Causality follows since the inverse of $I + G$ on W is the restriction to W of the inverse of $I + G$ on W_e which is assumed to be causal. Boundedness is then a direct consequence of the second condition in the definition of stability.

THEOREM 4.2. *Assume that, for any $T \in S$, $I + G$ has a causal inverse on $W_{Te} \triangleq \{x \in W_e | P_T x = 0, T \in S \text{ given}\}$. Then the feedback system under consideration is continuous if and only if $I + G$ has a continuous causal bounded inverse on W.*

Proof. (i) The condition is sufficient; $I + G$ is thus assumed to have a continuous, causal, bounded inverse on W. Stability follows from Theorem 4.1, and continuity of the feedback system thus follows immediately from continuity of the inverse of $I + G$ on W.

(ii) The condition is necessary; the feedback system is thus assumed to be continuous. Moreover $I + G$ has, by assumption, a causal inverse on

$$W_{re} \triangleq \{x \in W_e | P_T x = 0 \text{ for } \textit{some } T \in S\}.$$

By stability and continuity $I + G$ is invertible on $W_r \triangleq \{x \in W | P_T x = 0$ for some $T \in S\}$, and this inverse is continuous, causal and bounded. Let $(I + G)_r^{-1}$ denote this inverse (defined on W_r) and let F denote its continuous extension to W. This is possible since W_r is, by assumption, dense in W. Thus

$$Fx \triangleq \lim_{n \to \infty} (I + G)_r^{-1} x_n$$

with x_n a sequence in W_r with $\lim_{n \to \infty} x_n = x$. It remains to be shown that F is indeed the inverse of $I + G$ on W and that it is continuous, causal and bounded. That F is indeed continuous, causal and bounded is rather immediate and will not be shown explicitly. For the invertibility, recall that G was assumed to be continuous, and consider, with x_n a sequence as above,

$$(I + G)Fx = (I + G) \lim_{n \to \infty} (I + G)_r^{-1} x_n = \lim_{n \to \infty} x_n = x,$$

$$F(I + G)x = F \lim_{n \to \infty} (I + G)x_n = \lim_{n \to \infty} F(I + G)x_n = \lim_{n \to \infty} x_n = x.$$

Remark 4.1. Requiring invertibility of $I + G$ on W_{re} as in Theorem 4.2 is a lot weaker than requiring it on W_e as in Theorem 4.1 and constitutes an important facet of Theorem 4.2. Note also that Theorem 4.2 states that continuity implies causal invertibility of $I + G$ on W_e where this was only assumed on W_{re}. Note finally that, by the assumption that $I + G$ has a causal inverse on W_e, the inverse on W will necessarily be causal whenever it exists. This was stated in Theorem 4.1 merely for emphasis.

Remark 4.2. If the second condition in the definition of stability (i.e., the existence of a K) is not taken as an a priori requirement, then Theorem 1.1 should be modified so as not to require boundedness but only existence. Recall however that, by a theorem of Banach [13, p. 47], if $I + G$ is an invertible bounded linear transformation on W, then so is $(I + G)^{-1}$, and thus for linear systems condition (i) in the definition of stability implies, by Theorem 4.1, condition (ii).

Remark 4.3. Theorem 4.1 shows (under the invertibility assumption on W_e) the equivalence of the present definition of stability and the *alternate* definition in which the feedback system under consideration would be called stable if for any given $u \in W_e$ a solution $e \in W_e$ satisfies $\|P_T e\|_W \leqq K\|P_T u\|$ for all $T \in S$ and some constant K. This definition is also entirely reasonable and in fact allows a larger and somewhat more realistic class of testing inputs which yields a bounded response. The fact that stability in the second sense implies stability in the first one follows immediately. The converse is a consequence of Theorem 4.1 and the following lemma which will also be used in § 6. First however we have a definition.

DEFINITION 4.1. Let F be an operator from W_e into itself. F is said to be *bounded* on W_e if

$$\sup_{\substack{x \in W_e \\ T \in S \\ P_T x \neq 0}} \frac{\|P_T F x\|_W}{\|P_T x\|_W} < \infty .$$

This least upper bound will be denoted by $\|F\|_e$.

LEMMA 4.1. *Let F be a causal operator from W_e into itself. If F is bounded on W_e, then F maps W into itself, is bounded on W, and $\|F\| = \|F\|_e$. Conversely, if F maps W into itself and is bounded on W, then F is bounded on W_e, and $\|F\|_e = \|F\|$.*

Proof. The proof of this lemma which merely involves the properties of extended spaces is left to the reader.

Remark 4.4. A similar redefinition of uniform continuity which would require that given any $\varepsilon > 0$ there exists a $\delta > 0$ such that $u', u'' \in W_e$ and $\|P_T(u' - u'')\|_W \leqq \delta$ implies that any corresponding solutions $e', e'' \in W_e$ satisfy $\|P_T(e' - e'')\|_W \leqq \varepsilon$ for any $T \in S$ is, by Theorem 4.2, equivalent to the one used here.

5. Existence and uniqueness of solutions. The question of existence, uniqueness, and causal dependence of solutions on inputs is by no means of solely academic interest. First of all, the operators G_1 and G_2 in the forward and the feedback loop of the system need not be pathological in nature in order to induce, of course in a strict mathematical sense, signals in the loop which depend on future values of the inputs. Furthermore if the conditions for existence, uniqueness, and causal dependence of solutions are not satisfied then it can usually be concluded that the mathematical model of the feedback system does not represent a reasonable model of a physical system. In fact, under such circumstances the consideration of certain additional factors in the description of the feedback system (as, e.g., the introduction of infinitesimal time delays or filtering effects in the loops) will very often modify some of the important properties of the system as, for instance, its stability. Finally, as indicated in § 4 there exists a close relation between the questions of existence, uniqueness, and causality on one hand, and stability and instability on the other hand.

5.1. An example. As an example illustrating the above comments, consider the feedback loop shown in Fig. 4. All signals are real-valued, the forward loop consists of a unit gain minus a simple time delay, and the feedback loop consists of a unit gain.

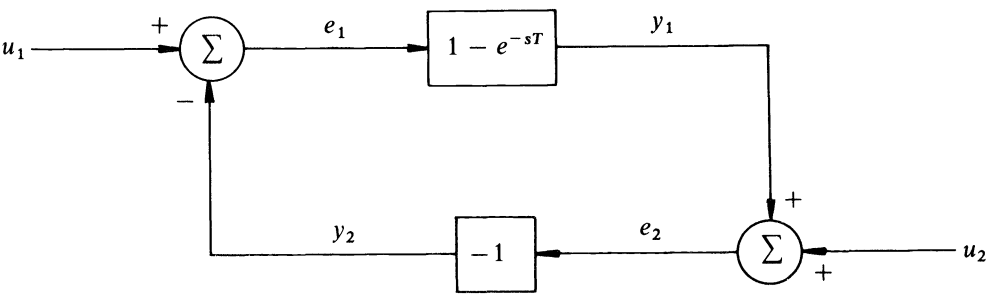


FIG. 4. *A predictor*

More precisely, let $S = R$, and let $V_1 = V_2 = R$. The operators G_1 and G_2 are defined by

$$G_1x(t) = x(t) - x(t - T), \qquad T > 0,$$

and

$$G_2x(t) = -x(t).$$

G_1 and G_2 are clearly causal. A simple calculation shows that a unique solution exists for any u_1, u_2, and

$$e_1(t) = u_1(t + T) + u_2(t + T),$$

$$e_2(t) = -u_1(t) + u_1(t + T) + u_2(t + T).$$

The closed-loop system thus acts as a predictor and certainly does not represent, even approximately, a physical realizable system.

The system is clearly stable and continuous. In particular, this is true with $W_1 = W_2 = L_p$, $1 \leqq p \leqq \infty$. It can easily be verified however that the introduction of an unavoidable infinitesimal time delay in the feedback loop, i.e.,

$$G_2x(t) = -x(t - \varepsilon), \qquad \varepsilon > 0,$$

will in fact lead to an unstable system.

5.2. Theorems on existence and uniqueness. The existence theorems which follow are based on the contraction mapping principle. They rely on the completeness properties of W expressed in assumption A1 (iv) of § 3. The proof, which will be deleted, is completely analogous to the usual proofs in showing existence and uniqueness of initial value problems in ordinary differential equations: one proceeds to show that a unique solution exists over a small initial interval and that this process can be continued. For details and additional related results, see Zames [5], [6]. The theorems will be stated with $S = [T_0, \infty)$ or $(-\infty, +\infty)$. The extension to cover, e.g., discrete systems, is straightforward.

THEOREM 5.1. *Let $S = [T_0, \infty)$. Then $I + G$ has a causal inverse on W_e if there exists a $\Delta T > 0$ such that for all $T \geqq T_0$, and all $x = (x_1, x_2)$ and $y = (y_1, y_2) \in W$ with $P_T x = P_T y$, either of the following conditions is satisfied*:

$$\text{(i)}\qquad \|P_{\Delta T}(G_1 x_1 - G_1 y_1)\|_{W_1} \leqq \rho_1 \|P_{\Delta T}(x_1 - y_1)\|_{W_1},$$

$$\|P_{\Delta T}(G_2 x_2 - G_2 y_2)\|_{W_2} \leqq \rho_2 \|P_{\Delta T}(x_2 - y_2)\|_{W_2}$$

with $\rho_1 \rho_2 < 1$ and where $P_{\Delta T} \triangleq P_{T+\Delta T} - P_T$.

(ii) *$W_1 = W_2$ and there exists a scalar c such that $I + cG_2$ has a causal inverse in W_{1e} and $G'_1 = G_1 - cI$ and $G'_2 = G_2(I + cG_2)^{-1}$ satisfy conditions* (i).

THEOREM 5.2. *Let $S = (-\infty, +\infty)$. Then $I + G$ has a causal inverse on $W_{T_0 e} \triangleq \{x \in W_e | P_{T_0} x = 0, T_0 \in S \text{ given}\}$ if the condition in Theorem* 5.1 *is satisfied.*

Theorem 5.1 thus gives conditions under which a unique solution to the feedback equations exists while Theorem 5.2 gives conditions under which solutions can be continued. Note that condition (i) will be satisfied whenever there is an infinitesimal delay present in the loop. Condition (ii) was introduced for the example in § 8 and to indicate what conditions for realizability could be required when it is decided to neglect certain time delays or filtering effects.

6. Stability and continuity.

General stability criteria involving gains, conicity and positivity have been given by Zames [4] and Sandberg [2] in a setting which is slightly less general than the one considered here. However, the extension does not represent any difficulties. This section starts by stating a general stability theorem and derives some specific results similar to the ones obtained in the above papers.

6.1. A general stability condition.

THEOREM 6.1. *The feedback system under consideration is stable if the inequality $\|P_T(I + G_2 G_1)e_1\|_{W_1} \geqq \varepsilon \|P_T e_1\|_{W_1}$ holds for all $e_1 \in W_1$, all $T \in S$, and some constant $\varepsilon > 0$. This condition is also necessary for stability if $I + G$ has a causal inverse on W_e.*

Proof. (i) The condition is sufficient; it is thus assumed that the inequality in the statement of Theorem 6.1 is satisfied. Let M denote $\|\tilde{G}_2\|$. Let $u = (u_1, u_2) \in W$ be given, and let $e = (e_1, e_2) \in W_e$ be a solution. Thus

$$e_1 + G_2 G_1 e_1 = u_1 - (G_2(G_1 e_1 + u_2) - G_2 G_1 e_1)$$

and consequently

$$P_T(I + G_2 G_1)e_1 = P_T u_1 - P_T(G_2(P_T G_1 e_1 + P_T u_2) - G_2 P_T G_1 e_1)$$

for all $T \in S$. Hence from the conditions of the theorem and the triangle inequality it follows that

$$\varepsilon \|P_T e_1\|_{W_1} \leqq \|u_1\|_{W_1} + M \|u_2\|_{W_2}$$

which, since $\varepsilon > 0$, yields

$$\|P_T e_1\|_{W_1} \leqq \varepsilon^{-1}\|u_1\|_{W_1} + \varepsilon^{-1}M\|u_2\|_{W_2}.$$

This shows that indeed $e_1 \in W_1$ and that

$$\|e_1\|_{W_1} \leqq \varepsilon^{-1}\|u_1\|_{W_1} + \varepsilon^{-1}M\|u_2\|_{W_2}.$$

Since $e_2 = G_1 e_1 + u_1$ and since G_1 is bounded on W_1 it follows that

$$\|P_T e_2\|_{W_2} \leqq \|G_1\|\ \|e_1\|_{W_1} + \|u_1\|_{W_1}$$

which yields $e_2 \in W_2$ and the boundedness condition required for stability.

(ii) The condition is necessary if $I + G$ has a causal inverse on W_e. This part of the theorem follows by letting $u_2 = 0$ and considering the equation

$$(I + G_2 G_1)e_1 = u_1.$$

$I + G_2 G_1$ has, by Theorem 4.1 and stability, a causal bounded inverse on W_1. This then yields, for all $T \in S$,

$$\begin{aligned}\|P_T e_1\|_{W_1} &= \|P_T(I + G_2G_1)^{-1}P_T(I + G_2G_1)e_1\|_{W_1}\\ &\leqq \|(I + G_2G_1)^{-1}\|\ \|P_T(I + G_2G_1)e_1\|_{W_1}\end{aligned}$$

and the conclusion with $\varepsilon = \|(I + G_2G_1)^{-1}\|^{-1}$. This completes the proof of Theorem 6.1.

Some simple conditions for the inequality condition in the statement of the theorem are given below. Notice that the sufficiency condition in Theorem 4.1 is, formally at least, completely disjoint from the questions of existence and uniqueness. Finally, if the second condition in the definition of stability is relaxed, then the sufficiency part of the theorem remains. The necessity part however is then false (unless, e.g., linearity of G is assumed).

COROLLARY 6.1. *The feedback system under consideration is stable if*

$$\|G_2G_1\| < 1.$$

Proof. Since by Lemma 4.1, $\|P_T G_2 G_1 e_1\|_{W_1} \leqq \|G_2G_1\|\ \|P_T e_1\|_{W_1}$, the corollary follows from Theorem 6.1 with $\varepsilon = 1 - \|G_2G_1\|$. Indeed,

$$\|P_T e_1 + P_T G_2 G_1 e_1\|_{W_1} \geqq \|P_T e_1\|_{W_1} - \|P_T G_2 G_1 e_1\|_{W_1} \geqq (1 - \|G_2G_1\|)\|P_T e_1\|_{W_1}.$$

COROLLARY 6.2. *Let $W_1 = W_2$. Then the feedback system under consideration is stable if there exists a scalar c such that $\|G_1 - cI\| < r$ and if for all $e_1 \in W_1$ and $T \in S$, $\|P_T(I + cG_2)e_1\|_{W_1} \geqq r\|P_T G_2 e_1\|_{W_1}$.*

Proof. Let $c \neq 0$. Since $c(I + G_2G_1) = -(G_1 - cI) + (I + cG_2)G_1$, it follows from the triangle inequality and the assumptions of the corollary that

$$\begin{aligned}\|P_T(I + G_2G_1)e_1\|_{W_1} &\geqq |c^{-1}|(\|P_T(I + cG_2)G_1e_1\|_{W_1} - \|P_T(G_1 - cI)e_1\|_{W_1})\\ &\geqq |c^{-1}|(r\|P_T G_2 G_1 e_1\|_{W_1} - \|G_1 - cI\|\ \|P_T e_1\|_{W_1})\end{aligned}$$

which yields the condition of Theorem 6.1 whenever

$$\|P_T D_2 G_1 e_1\|_{W_1} \geqq \frac{1}{2}\left(1 + \frac{\|G_1 - cI\|}{r}\right)\|P_T e_1\|_{W_1}.$$

For $c = 0$ and whenever

$$\|P_T G_2 G_1 e_1\|_{W_1} \leqq \frac{1}{2}\left(1 + \frac{\|G_1 - cI\|}{r}\right)\|P_T e_1\|_{W_1},$$

the proof of Corollary 6.1 can be used unaltered. Thus in all cases the condition of Theorem 6.1 is satisfied which yields stability as claimed.

Remark 6.1. If $I + cG_2$ has a bounded causal inverse on W_1, then the second condition of the above corollary is equivalent to requiring $\|G_2(I + cG_2)^{-1}\| \leqq r^{-1}$.

DEFINITION 6.1. An operator, F, from an inner product space, X, into itself is said to be *positive*[4] on X if $\operatorname{Re}\langle x, Fx\rangle_X \geqq 0$ for all $x \in X$.

Remark 6.2. If W_1 is an inner product space (and only then), then the last condition of Corollary 6.2 can be stated somewhat more simply by requiring G_2 to satisfy[5] the following conditions:

(i) $$\left\|G_2 - \frac{c}{r^2 - |c|^2}I\right\| \leqq \frac{r}{r^2 - |c|^2} \qquad \text{if } r > |c|,$$

(ii) $$\left\|P_T\left(G_2 - \frac{c}{|c|_2 - r^2}I\right)e_1\right\|_{W_1} \geqq \frac{r}{|c|^2 - r^2}\|P_T e_1\|_{W_1}$$

for all $e_1 \in W_1$ and $T \in S$, if $r < |c|$.

(iii) $G_2 + 1/r$ is positive on W_1, if $r = |c| \neq 0$.

The fact that no truncations appear in (i) is due to Lemma 4.1.

COROLLARY 6.3. *Let $W_1 = W_2$ be an inner product space. Then the feedback system under consideration is stable if G_1 and $G_2 - \varepsilon I$ are positive on W_1 for some $\varepsilon > 0$.*

Proof. The proof of this limiting case (with $r, c \to \infty$) of Corollary 6.2 is left to the reader.

Remark 6.3. Note that Corollary 6.3 requires the verification of positivity conditions on the nonextended space only.

Remark 6.4. As can immediately be deduced from the statement of Theorem 6.1, it is of course possible to restate the conditions of the previous corollaries in terms of two causal operators G_2' and G_1' which are a factorization of G_2G_1 (i.e., $G_2'G_1' = G_2G_1$ on W_1).

6.2. A general continuity condition. The following continuity theorem follows immediately using the same methods as were used in the stability analysis.

THEOREM 6.2. *The feedback system under consideration is continuous if the inequality*

$$\|P_T(I + G_2G_1)e_1' - P_T(I + G_2G_1)e_1''\|_{W_1} \geqq \varepsilon\|P_T(e_1' - e_1'')\|_{W_1}$$

holds for all $e_1', e_1'' \in W_1$, all $T \in S$ and some $\varepsilon > 0$.

Corollaries 6.1, 6.2 and 6.3 have obvious counterparts as continuity conditions. The condition of Theorem 6.2 is necessary for Lipschitz continuity of the feedback systems if $I + G$ has a causal inverse on W_{re}.

7. Instability. Theorem 4.1 gives a procedure through which it is possible to generate instability criteria. It suffices therefore to consider the invertibility of

[4] The terms *dissipative*, *passive* and *accretive* have been used in the same context.

[5] These conicity conditions, which are very useful in practice since the spaces involved are more often than not Hilbert spaces (e.f., L_2 or ℓ_2), are due to Zames [4].

$I + G$ on W. If this inverse does not exist, or exists but is unbounded or noncausal, then instability results. However, in order for Theorem 4.1 to be applicable, it is necessary for $I + G$ to have a causal inverse on W_e. Thus whenever the inverse of $I + G$ on W exists, it will unavoidably be causal and hence for instability it needs to be established that a certain inverse does not exist or exists but is unbounded. Note that for linear operators the inverse will necessarily be bounded whenever it exists and that then instability is thus equivalent to noninvertibility of $I + G$.

It is in general not an easy matter to show that an inverse does not exist, and Theorem 4.1 does not seem at first sight very useful for instability purposes. One procedure which accomplishes this however will be described at the end of this section.

Theorem 4.2 on the other hand gives a procedure through which it is possible to prove that the system is not continuous. One needs therefore to show that either $I + G$ is not invertible on W or that it is invertible but that the inverse is not bounded, not continuous, or not causal. This last condition will indeed be very useful and a large class of operators which have a noncausal inverse on W will be constructed. Note that Theorem 4.2 only presupposed continuability of the null solution, a condition which in general causes no difficulties. Since for linear systems continuity and stability are equivalent, Theorem 4.2 will thus be used to prove a rather general instability theorem for linear systems.

For the remainder of this section attention will be restricted to linear systems.[6] It might not be clear at this point how one would go about showing that the inverse of a particular operator, given that it exists, is noncausal. Some thought however reveals that this will, in at least some cases, be possible, namely, when G is linear and time invariant. The remainder of this section explains a procedure by which this can be achieved under more general circumstances. This method is to modify the system such that the inverse of the modified closed-loop operator exists but is actually noncausal, and to "follow" the causality of this inverse as this modification is removed. The modified system will be chosen to be simpler (e.g., time invariant) so that much more can be said about its inverse. This procedure will be explained in more detail later on.

7.1. Two lemmas on Banach algebras. As pointed out by Zames and Falb [7], the natural setting for linear operators in a Banach space is in a Banach algebra, with the causal operators characterized by a subalgebra. The relevant notions on Banach algebras and subalgebras have been introduced in § 2. The following relationship between the spectrum of an element of a subalgebra when considered as an element of a subalgebra or of the basic algebra is essential in the proof of the instability theorems which follow. It is believed to be of some independent interest.

LEMMA 7.1. *Let x and y be elements of a subalgebra $\mathscr{A}^+$ of a Banach algebra $\mathscr{A}$ with unit e. Let $e \in \mathscr{A}^+$. Assume that $x + \alpha y$ is invertible in the algebra for all α in some connected set C of the complex plane and that the collection of elements $(x + \alpha y)^{-1} \in \mathscr{A}$ is bounded on C (i.e., there exists a constant K such that*

[6] It should be noticed that the resulting theorems do have implications about nonlinear systems as well. Indeed, instability of the equations linearized around the null solution implies instability of the original nonlinear system.

$\|(x + \alpha y)^{-1}\| \leqq K$ for all $\alpha \in C$). Then $(x + \alpha y)^{-1} \in \mathscr{A}^+$ (i.e., $x + \alpha y$ is invertible in the subalgebra) for all $\alpha \in C$ if and only if $(x + \alpha_0 y)^{-1} \in \mathscr{A}^+$ for some $\alpha_0 \in C$.

Proof. (i) The condition is necessary: this part is obvious.

(ii) The condition is sufficient; it is thus assumed that $(x + \alpha_0 y)^{-1} \in \mathscr{A}^+$ and that $\alpha_0 \in C$. The conclusion clearly holds if $\|y\| = 0$. Assume therefore that $\|y\| \neq 0$ and notice that $K \neq 0$. It will first be shown that if $(x + \alpha' y)^{-1} \in \mathscr{A}^+$, then $(x + \alpha y)^{-1} \in \mathscr{A}^+$ for all complex numbers $\alpha \in N \triangleq \{\alpha | |\alpha - \alpha'| \leqq \|y\|^{-1}K^{-1}\}$. Write therefore $x + \alpha y$ as $(x + \alpha' y)(e + (\alpha - \alpha')(x + \alpha' y)^{-1} y)$. Since $e - z$ is invertible whenever $\|z\| < 1$, with $(e - z)^{-1} = \sum_{k=0}^{\infty} z^k$, and since thus $(e - z)^{-1} \in \mathscr{A}^+$ if $z \in \mathscr{A}^+$, the claim follows by the obvious estimate

$$\|(\alpha - \alpha')(x + \alpha' y)^{-1} y\| \leqq |\alpha - \alpha'| \, \|(x + \alpha' y)^{-1}\| \, \|y\|.$$

Let P be the set in the complex plane defined by $P \triangleq \{\alpha | (x + \alpha y)^{-1}$ exists and belongs to $\mathscr{A}^+\}$, and let P^c denote its complement. The lemma claims that $P^c \cap C$ is empty. Assume therefore that $P^c \cap C$ is not empty. Then

$$d(P \cap C, P^c \cap C) = \inf_{\substack{\alpha' \in P \cap C \\ \alpha'' \in P^c \cap C}} |\alpha' - \alpha''| \geqq \|y\|^{-1}K^{-1}.$$

Let

$$N_1 \triangleq \bigcup_{\alpha \in P \cap C} \{\alpha' | |\alpha' - \alpha| < \|y\|^{-1}K^{-1}/3\}$$

and let N_2 be similarly defined with $\alpha \in P^c \cap C$. The sets N_1 and N_2 are open, disjoint, have a nonempty intersection with C (with N_1 since $\alpha_0 \in P \cap C$ and with N_2 since $P^c \cap C$ is assumed to be nonempty) and their union contains C. Hence C is not connected. This contradiction ends the proof of the lemma.

Since W is assumed to be a Banach space, $\mathscr{L}(W, W)$ forms a Banach algebra. Let $\mathscr{L}^+(W, W)$ be the set of all elements of $\mathscr{L}(W, W)$ which are in addition causal. The next lemma exposes the algebraic properties of $\mathscr{L}^+(W, W)$.

LEMMA 7.2. *$\mathscr{L}^+(W, W)$ forms a subalgebra of $\mathscr{L}(W, W)$ and contains the unit.*

Proof. $\mathscr{L}^+(W, W)$ is clearly closed under addition, multiplication of elements, multiplication by scalars, and contains the unit. It remains to be proven that $\mathscr{L}^+(W, W)$ is complete. Let $\{L_n^+\}$ be a convergent (in $\mathscr{L}(W, W)$) sequence of elements in $\mathscr{L}^+(W, W)$ and let L be its limit. Thus $\lim_{n\to\infty} \|(L - L_n^+)x\| = 0$ for all $x \in W$. Assume that L is not causal; then there exists an element in $x \in W$ and a $T \in S$ such that $P_T x = 0$ but $P_T L x \neq 0$. But since, for all n,

$$\|(L - L_n^+)x\| \geqq \|P_T(L - L_n^+)x\| = \|P_T L x\|,$$

a contradiction follows.

7.2. An inverse function theorem. To obtain instability theorems, it will be established that the operator $I + G$ is invertible and that its inverse is noncausal. The previous section gives the structure which enables one to prove noncausality. This section gives the inverse function theorem which suffices for the purposes of this paper.

LEMMA 7.3. *Let $W_1 = W_2$, and let $G \in \mathscr{L}(W, W)$. Then $I + G$ is invertible on W if there exists a scalar c such that $I + cG_2$ is invertible on W_1 and*

$$\|(I + cG_2)^{-1}G_2(G_1 - cI)\| < 1.$$

Moreover, $(I + G)^{-1} \in \mathscr{L}(W, W)$.

Proof. Since

$$I + G_2G_1 = I + cG_2 + G_2(G_1 - cI) = (I + cG_2)(I + (I + cG_2)^{-1}G_2(G_1 - cI)),$$

it follows that $I + G_2G_1$ is indeed invertible on W_1. It is easily verified that the operator $F \in \mathscr{L}(W, W)$ defined by

$$Fx = F(x_1, x_2) = ((I + G_2G_1)^{-1}(x_1 - G_2x_2), x_2 + G_1(I + G_2G_1)^{-1}(x_1 - G_2x_2))$$

is indeed the inverse of $I + G$ on W and belongs to $\mathscr{L}(W, W)$ and to $\mathscr{L}^+(W, W)$ if and only if $(I + G_2G_1)^{-1}$ is causal.

7.3. An instability theorem. The following general instability theorem is the main result of this section, and is, in a sense, the converse of the stability conditions involving conicity obtained by Zames [4].

THEOREM 7.1. *Let $W_1 = W_2$ and let $G \in \mathscr{L}(W, W)$. Assume that, for any $T \in S$, $I + G$ has a causal inverse on $W_{Te} \triangleq \{x \in W_e | P_Tx = 0,\ T \in S \text{ given}\}$. Then the feedback system under consideration is unstable if there exists a scalar c such that $I + cG_2$ has a noncausal inverse on W_1 and $\|(I + cG_2)^{-1}G_2(G_1 - cI)\| < 1$.*

Proof. It follows from Lemma 7.3 that $I + G$ is invertible on W, and hence by Theorem 4.2 it suffices to show that this inverse is not causal on W. This inverse $(I + G)^{-1}$ is causal on W if and only if $(I + G_2G_1)^{-1}$ is causal on W_1. However, since $I + G_2G_1 = I + cG_2 + G_2(G_1 - cI)$ it suffices to prove, by Lemmas 7.1 and 7.2, that $I + cG_2 + rG_2(G_1 - cI)$ is invertible for all $|r| \leqq 1$ and that the norm of this collection of inverses is bounded. Since however

$$I + cG_2 + rG_2(G_1 - cI) = (I + cG_2)(I + r(I + cG_2)^{-1}G_2(G_1 - cI)),$$

it follows immediately from the inequality $\|r(I + cG_2)^{-1}G_2(G_1 - cI)\| < |r|$ that the inverse exists for all $|r| \leqq 1$ and that in fact

$$\|(I + cG_2 + rG_2(G_1 - cI))^{-1}\| \leqq \|(I + cG_2)^{-1}\|(1 - |r|\,\|(I + cG_2)^{-1}G_2(G_1 - cI)\|)^{-1}.$$

This then yields

$$\|(I + cG_2)^{-1}\|(1 - \|(I + cG_2)^{-1}G_2(G_1 - cI)\|)^{-1}$$

as the desired bound.

Remark 7.1. There is no problem in identifying an input $u \in W$ which yields a solution $e \in W_e$, with $e \notin W$, i.e., $e \in W_e - W$. In fact, let $u \in W$ and $T \in S$ be such that $P_T(I + G)^{-1}u \neq 0$ with $P_Tu = 0$. Such a T and a u exist since $(I + G)^{-1}$ is not causal. Since $u \in W_{Te}$ and $I + G$ has a causal inverse on W_{Te} there is an $e'' \in W_e$ such that $(I + G)e'' = u$. This continuation, e'', is *not* in W since $e' = (I - G)^{-1}u$ is, by invertibility, the *unique* element of W which satisfies $(I + G)e' = u$, and since $e' \neq e''$ (in fact, $P_Te'' = 0$ but $P_Te' \neq 0$ by assumption). Thus $e'' \in W_e - W$, as claimed.

Remark 7.2. The above remark also illustrates that if the system is unstable because the inverse $(I + G)^{-1}$ exists on W but is noncausal, then there will in general *not* be a unique solution $e \in W_e$ for certain inputs $u \in W_e$. Indeed, certain inputs $u \in W$ will give a solution $e'' \in W_e - W$ and a solution $e' \in W$. Naturally only e'' will be a solution in the "dynamical system" sense.

7.4. Instability theorems when S has a lower bound. The case when the time interval of definition, S, has a lower bound is somewhat more realistic from a practical viewpoint than the case when S has no lower bound. Theorem 7.1 gives a procedure which allows one to prove instability for a large class of feedback systems and which is based on showing that the inverse of $I + G$ exists on W, but is not causal. This procedure is not promising whenever S actually has a lower bound since very weak conditions on the operator G will then ensure that $I + G$ has a causal inverse on W_e and thus it is impossible for $I + G$ to have a noncausal inverse on W. Hence, in order to obtain instability theorems when S has a lower bound, it will be necessary to design a procedure through which it is possible to show that a particular operator is not invertible. The remainder of this section describes one such procedure. It is based on an extension of S, W and G and on the following lemma which is believed to be of some interest in its own right.

LEMMA 7.4. *Assume that, for any* $T \in S$, $I + G$ *has a causal inverse on* $W_{Te} = \{x \in W_e | P_T x = 0,\ T \in S \text{ given}\}$. *Let* $G \in \mathscr{L}(W, W)$, *and let* $T_1, T_2 \in S$ *be arbitrary. Then* $I + G$ *has a causal inverse on* $W_{T_1} \triangleq \{x \in W | P_{T_1} x = 0,\ T_1 \in S \text{ given}\}$ *if and only if it has a causal inverse on* W_{T_2}.

Proof. Assume that $T_1 \leqq T_2$. Then the inverse on W_{T_1} clearly qualifies as the inverse on W_{T_2}. The converse however is not as immediate. Assume thus that $I + G$ has a causal inverse on W_{T_2}, and let $x \in W_{T_1}$ be given. Let $e \in W_{T_1 e}$ be the unique solution (in $W_{T_1 e}$) of $(I + G)e = x$, and let $x_2 = (I - P_{T_2})x$. Let $e_2 \in W_{T_2}$ be the unique solution (in W_{T_2}) of $(I + G)e_2 = (I - P_{T_2})(x - (I + G)P_{T_2} e)$. Then $P_{T_2} e_1 + e_2 \in W_{T_1}$ and, as a simple calculation shows, $(I + G)(P_{T_2} e + e_2) = x$. Thus $e = P_{T_2} e + e_2 \in W_{T_1}$ and $I + G$ has a causal inverse on W_{T_1} as claimed. This completes the proof of the lemma.

The idea behind the remainder of the procedure is to extend the interval of definition, say $[T_0, \infty)$, to $(-\infty, +\infty)$, and extend the space W accordingly to W'. The operator G, originally defined on W, is then extended to G', defined on W'. The pairs W, W' and G, G' will of course have to satisfy certain compatibility conditions. If G' is properly chosen, within its restrictions, then it can very well happen that $I + G'$ has a noncausal inverse on W'. This fact and the properties of G and G' will then lead to a demonstration of the noninvertibility of G.

To make this procedure more precise, let $S_1 \subset R$ be such that $S_1 \cap S$ is empty and $\sup S_1 \leqq \inf S$. Let $S' = S \cup S_1$. Let W'_1 and W'_2 be Banach spaces of respectively V_1- and V_2-valued functions on S', and let $W' = W'_1 \times W'_2$. W and W' are related as follows: for any $x' \in W'$, consider the function defined by $x(t) = x'(t)$ for $t \in S$. It is then assumed that $x \in W$. Conversely, it is assumed that if $x \in S$ then the function $x'(t)$ with $x'(t) = x(t)$ for $t \in S$, and $x'(t) = 0$ for $t \in S' - S$, is an element of W'. The next step is to define an appropriate extension of G. Let G'_1 and G'_2 be operators from W'_1 into W'_2 and from W'_2 into W'_1 respectively and let G' map W' into itself according to $G'e = G'_2 e_2, -G'_1 e_1$ for $e = (e_1, e_2) \in W'$. The operators G and G' are related as follows: for any $x' \in W'$ with support on S, let x be the element of W such that $x'(t) = x(t)$ for $t \in S$. x indeed belongs to W by the assumption on the relation between W and W'. It is then assumed that $(G'x')(t) = (Gx)(t)$ for all $t \in S$. Typically, neither W' nor G' is uniquely defined by the above assumptions. There is more often than not a natural choice for W'_1

and W_2', e.g., $L_p(-\infty, +\infty)$ if W_1 and $W_2 = L_p(T_0, \infty)$. In general many possible choices for G' remain.

The following additional assumptions are made on W' and G':

(i) W' satisfies the conditions A1 (i)–(iv);

(ii) G' satisfies the conditions A2 (i)–(iii);

(iii) $G' \in \mathscr{L}(W', W')$ whenever $G \in \mathscr{L}(W, W)$.

If W' and G' satisfy all the above hypothesis then they will be called *backward extensions* of W and G.

Examples of such extensions of W and G will be given in § 8. It is in general rather easy to come up with a suitable choice for W' and G'. The above procedure leads to the following instability theorem.

THEOREM 7.2. *Let $W_1 = W_2$, and let $G \in \mathscr{L}(W, W)$. Let W' and G' be backward extensions of W and G. Assume that for any $T \in S'$, $I + G'$ has a causal inverse on $W'_{Te} = \{x \in W'_e | P_T x = 0,\ T \in S' \text{ given}\}$. Then the feedback system under consideration is unstable if there exists a scalar c such that $I + cG'_2$ has a noncausal inverse on W'_1 and $\|(I + cG'_2)^{-1}G'_2(G'_1 - cI)\| < 1$.*

Proof. As in Theorem 7.1, it follows from Lemmas 7.1 and 7.2 that $I + G'$ has a *noncausal* inverse on W'. Hence $I + G'$ is not invertible on some $W_{T'}$, $T' \in S'$. Thus by Lemma 7.4, $I + G'$ is not invertible on W'_T for *any* $T \in S$. Thus $I + G'$ is either not one-to-one on W'_T or not onto W'_T. The claim is that this implies that also $I + G$ is either not one-to-one or not onto. Assume first that $I + G'$ is not one-to-one on W'_T. Then there exist $x'_1, x'_2 \in W'_T$, $x'_1 \neq x'_2$, with $(I + G')x'_1 = (I + G')x'_2$. Consider now the elements $x_1, x_2 \in W_T$ for which $x_1(t) = x'_1(t)$ and $x_2(t) = x'_2(t)$ for $t \in S$. By the assumptions on W' and G', x_1, x_2 exist and $(I + G)x_1 = (I + G)x_2$. Since x'_1 and x'_2 have support on S and $x'_1 \neq x'_2$, $x_1 \neq x_2$. Thus $I + G$ is not one-to-one on W_T. Assume next that $I + G'$ is not onto W'_T. Since for all $x \in W_T$ there exists an $x' \in W'_T$ (i.e., $x'(t) = x(t)$ for $t \in S$) such that $(I + G')x' = (I + G)x$ on S it follows that also $I + G$ cannot be onto. Thus $I + G$ is not invertible on W_T, which yields instability as claimed.

8. An example.

As an illustration of how the results obtained in the previous sections translate in a more concrete situation, consider the feedback loop shown in Fig. 5. In terms of the notation used in the previous sections, let $S = [T_0, \infty)$ or $(-\infty, +\infty)$, $V_1 = V_2 = R$, i.e., u_1 and u_2 are real-valued, and let G_1 and G_2 be defined by

$$G_1x(t) \triangleq \sum_{n=0}^{\infty} g_n x(t - t_n) + \int_S g(t - \tau)x(\tau)\, d\tau ,$$

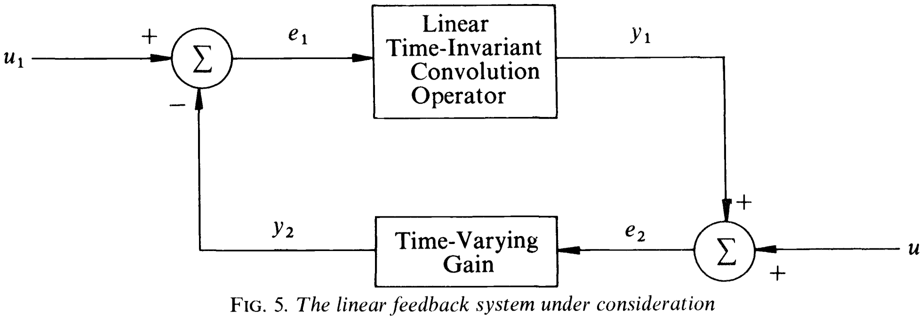


FIG. 5. *The linear feedback system under consideration*

where $t_0 = 0$, $t_n > 0$ for $n \geqq 1$, $g(t) = 0$ for $t < 0$, $\{g_n\} \in l_1$ and $g \in L_1(0, \infty)$, and

$$G_2x(t) \triangleq k(t)x(t)$$

with $k \in L_\infty(S)$. Let $\underline{k} = \text{ess inf}_{t \in S}\, k(t)$, and let $\bar{k} = \text{ess sup}_{t \in S}\, k(t)$. Thus $\underline{k} \leqq k(t) \leqq \bar{k}$ for almost all $t \in S$.

Assumption. It will be assumed that $g_0 = 0$ or that $g_0^{-1} \notin [\underline{k}, \bar{k}]$.

Let $W \triangleq L_2(S) \times L_2(S)$. It can then be verified using some elementary properties of L_2-spaces that the assumptions A1 (i)–(iv) of § 3 are then indeed satisfied.

Standard calculations based on Minkowski's inequality show that G_1 and G_2 define causal bounded linear operators from $L_{2e}(S)$ into itself. This then shows that G_1 and G_2 satisfy assumptions A2 (i)–(iii) listed in § 3.

The assumption on g_0 suffices to ensure that the conditions of Theorem 5.1 in the case $S = [T_0, \infty)$, and of Theorem 5.2 in the case $S = (-\infty, +\infty)$, are satisfied. (The example of § 5 indicates that some restriction would have to be made.) To verify this, let, in Theorem 5.1 or 5.2, $c = g_0$, and let ΔT be such that

$$\sum_{0 < t_n \leqq \Delta T} |g_n| + \int_0^{\Delta T} |g(\tau)|\, d\tau < \|k\|^{-1}_{L_\infty(S)} \cdot \|1 + g_0 k\|^{-1}_{L_\infty(S)}.$$

Such a $\Delta T > 0$ indeed exists since, by assumption, $\{g_n\} \in l_1$ and $g \in L_1(0, \infty)$. Hence the operator $I + G$, specialized to the example, has a causal inverse on W_e when $S = [T_0, \infty)$ and on $W_{Te} \triangleq \{x \in W_e | P_T x = 0,\ T \in S \text{ given}\}$ for any $T \in R$, when $S = (-\infty, +\infty)$.

8.1. Stability and instability in the time-invariant case. The stability results obtained here constitute a generalization of the well-known Nyquist criterion. There is some overlap between these results and similar generalizations of the Nyquist criterion which have recently been derived by Desoer [14], [15].

DEFINITION 8.1. Let LA denote the algebra consisting of elements determined by a real-valued $L_1(-\infty, +\infty)$-function, g, a real-valued l_1-sequence, $\{g_n\}$, $n = 0, 1, \cdots$, and a sequence of real numbers $\{t_n\}$, $n = 0, 1, \cdots$. Addition of $x_1 = (g_1, \{g_n, t_n\}_1)$ and $x_2 = (g_2, \{g_n, t_n\}_2)$ is defined as $(g_1 + g_2, \{g_n, t_n\}_3)$ where the sequence $\{g_n, t_n\}_3$ consists of exactly all pairs $\{g_n, t_n\}_1$ and $\{g_n, t_n\}_2$. Multiplication by scalars is defined as

$$\alpha x = \alpha(g, \{g_n, t_n\}) = (\alpha g, \{\alpha g_n, t_k\}),$$

and multiplication of elements is defined by

$$x_1 x_2 \triangleq \left(\int_{-\infty}^{+\infty} g_1(t - \tau) g_2(\tau)\, d\tau + \sum_n (g_n)_1 g_2(t - (t_n)_1) + \sum_n (g_n)_2 g_1(t - (t_n)_2), \right.$$
$$\left. \{g_n, t_n\}_3 \right) \quad \text{where} \quad \{t_n\}_3 = \{t_n\}_1 \oplus \{t_n\}_2$$

(i.e., all elements of the form $t_n = t_{n_1} + t_{n_2}$ where t_{n_1} and t_{n_2} range over $\{t_n\}_1$ and $\{t_n\}_2$ respectively), and the element g_n corresponding to t_n in $\{g_n, t_n\}_3$ is given by $g_{n_1} \cdot g_{n_2}$, with $t_n = t_{n_1} + t_{n_2}$. Let the norm on LA be defined by $\|g\|_{L_1} + \|\{g_n\}\|_{l_1}$. It can be shown that LA as defined above is a real commutative Banach algebra with the unit $e = (0, \{g_n, t_n\})$ with $g_0 = 1$, $t_0 = 0$, and $g_k = 0$ otherwise.

Consider now the subset of LA, $(LA)^+$, which consist of all elements of LA which satisfy $g(t) = 0$ for $t < 0$ and $t_n \geqq 0$ for all n. It is easily verified that $(LA)^+$ is a subalgebra and that it contains the unit. The details of the proofs of the above claims can be found in [13, pp. 141–157].

The *Laplace transform* of an element of LA is defined as the function of the complex variable s

$$G(s) \triangleq \sum_{n=0}^{\infty} g_n e^{-st_n} + \int_{-\infty}^{\infty} g(t)\, e^{-st}\, dt.$$

$G(s)$ is well-defined for $\operatorname{Re} s = 0$ for elements of LA. It is well-defined and analytic in $\operatorname{Re} s > 0$ for elements of $(LA)^+$.

The following lemma plays an essential role in the results which follow.

LEMMA 8.1. *Let $(g, \{g_n, t_n\}) \in (LA)^+$. Then it is invertible in $(LA)^+$ if and only if* $\inf_{\operatorname{Re} s \geqq 0} |G(s)| > 0$, *and in LA if and only if* $\inf_{\operatorname{Re} s = 0} |G(s)| > 0$.

Proof. The first statement is proven in [13, p. 150], and the second in [13, p. 155]. See also [16, p. 71].

Application of the above lemma leads to the following condition for stability of time-invariant systems.

LEMMA 8.2. *Assume that in the feedback system under consideration $k(t)$* $=$ const. $= K$. *Let $G_1(s)$ be the Laplace transform of $(g, \{g_n, t_n\})$. Let $W = L_2(S) \times L_2(S)$ be the space with respect to which stability is defined, and assume that $1 + Kg_0 \neq 0$. Then the feedback system under consideration is stable if and only if* $\inf_{\operatorname{Re} s \geqq 0} |1 + KG_1(s)| > 0$.

Proof. Since by assumption $|1 + Kg_0| > 0$ it follows from the existence results that Theorem 4.1 is applicable in the case $S = [T_0, \infty)$ and that Theorem 4.2 is applicable in the case $S = (-\infty, +\infty)$. There are three mutually exclusive possibilities:

(i) $$\inf_{\operatorname{Re} s \geqq 0} |1 + KG_1(s)| > 0,$$

(ii) $$\inf_{\operatorname{Re} s = 0} |1 + KG_1(s)| = 0,$$

(iii) $$\inf_{\operatorname{Re} s > 0} |1 + KG_1(s)| = 0$$

and

$$\inf_{\operatorname{Re} s = 0} |1 + KG_1(s)| > 0.$$

It needs to be shown that (i) yields stability and that (ii) and (iii) yield instability. In the first case, it follows from Lemma 8.1 that $I + G$ has a bounded causal inverse on W which then by Theorems 4.1 or 4.2 yields stability. Assume next that (ii) is satisfied. $I + G$ has a bounded inverse on W_2 if and only if $I + KG_1$ has a bounded inverse on $L_2(S)$. $I + KG_1$ multiplies the limit-in-the-mean transform of the element on which it operates by $1 + KG_1(j\omega)$; thus the only candidate for the inverse is the operator which divides the limit-in-the-mean transform of the element on which it operates by $1 + KG(j\omega)$. Thus for this inverse to be bounded, $(1 + KG_1(j\omega))^{-1}$ ought to exist for almost all $\omega \in R$ and belong to L_∞. Since $G_1(j\omega)$ is continuous and, by assumption,

$$\inf_{\omega \in R} |1 + KG_1(j\omega)| = 0,$$

$I + G$ has thus no bounded inverse on W, which by Theorems 4.1 and 4.2 yields instability. The third case requires a different proof depending whether $S = (-\infty, +\infty)$ or $S = [T_0, \infty)$. Consider first the case $S = (-\infty, +\infty)$. It then follows from Lemma 8.1 that $I + G$ has a bounded inverse on W but that this inverse is not causal, which then by Theorem 4.2 yields instability. If $S = [T_0, \infty)$, then the claim is that $I + G$ has no causal bounded inverse on W. Consider the obvious backward extension of W and G by letting $W' = L_2(-\infty, +\infty) \times L_2(-\infty, +\infty)$ and defining G' on W' by

$$G'_1x(t) = \sum_{n=0}^{\infty} g_n x(t - t_n) + \int_{-\infty}^{+\infty} g(t - \tau)x(\tau)\, d\tau$$

and $G'_2x(t) = Kx(t)$ with $x \in L_2(-\infty, +\infty)$. Since by the previous part of the proof, G' has a noncausal inverse on W', it then follows, by the arguments used in the proof of Theorem 7.2, that G is not invertible on W, which yields instability as claimed.

Remark 8.1. The above lemma is well known although the usual proofs assume the equivalence of stability and the absence of singularities of $(1 + KG_1(s))^{-1}$ in $\operatorname{Re} s \geqq 0$, and lack therefore a certain amount of justification. The papers by Desoer [14], [15] are an exception to this, but only treat stability. Notice that since the system is linear, instability actually implies that there exists a $u \in W(S)$ such that $e \in W_e(S) - W(S)$.

Remark 8.2. If stability is defined with $W = L_p(S) \times L_p(S)$, $1 \leqq p \leqq \infty$, then it can be shown that the condition of the lemma are still sufficient for stability. If the condition fails because of (iii) then instability results. The proofs of these claims follow the proof of Lemma 8.2 when $S = [T_0, \infty)$, or when $S = (-\infty, +\infty)$ and $1 \leqq p \leqq \infty$. The case $p = \infty$ and $S = (-\infty, +\infty)$ cannot be treated by the above methods since then assumption A1 (iii) is not satisfied. It can however be treated directly. If the condition fails because of (ii), the situation appears to be more complex. It can still be shown that then instability results at least when $p = 1$ or ∞, since $I + KG_1$ is invertible in $\mathscr{L}^+(L_p(S), L_p(S))$, $p = 1, \infty$, *if and only if* it is invertible in LA^+.

Remark 8.3. It is possible to verify, at least in some cases, the condition $\inf_{\operatorname{Re} s \geqq 0} |1 + KG_1(s)| > 0$ by establishing (i) $\inf_{\omega \in R} |1 + KG_1(j\omega)| > 0$, and (ii) checking whether $KG_1(j\omega)$ encircles the $-1 + 0.j$ point A proof of this for the case $g_n = 0$ for all $n \geqq 1$ can be found for instance in [4]. It has not been possible as yet to generalize this condition to the case under consideration, mainly because it appears to be no easy matter to give a suitable generalization of the no encirclement condition. One important particular case is stated below, namely when the delays are equally spaced, i.e., when $t_n = nT$ for some $T > 0$.

DEFINITION 8.2. The *argument*, $\theta(\omega)$, of $1 + KG_1(j\omega)$, with $1 + KG_1(j\omega) \neq 0$, is defined as the *continuous* function with $\theta(0) = 0$ or $\theta(0) = \pi$ such that for all $\omega \in R$,

$$1 + KG_1(j\omega) = |1 + KG_1(j\omega)|\, e^{j\theta(\omega)}.$$

LEMMA 8.3. *The condition*

$$\inf_{\operatorname{Re} s \geqq 0} |1 + KG_1(s)| > 0$$

is equivalent to the conditions

(i) $$\inf_{\operatorname{Re} s = 0} |1 + kG_1(s)| > 0$$

and

(ii) $\lim_{N\to\infty} \theta(N2\pi T^{-1})$ *exists and is zero.*

Proof. (i) Conditions (i) and (ii) imply that $\inf_{\operatorname{Re} s \geqq 0} |1 + KG_1(s)| > 0$. Let

$$A(s) \triangleq \sum_{n=0}^{\infty} g_n e^{-snT} \quad \text{and} \quad L(s) \triangleq \int_0^{\infty} g(t)\, e^{-st}\, dt.$$

Since $G_1(j\omega) = A(j\omega) + L(j\omega)$, it is the sum of a periodic function, $A(j\omega)$, and a bounded function, $L(j\omega)$, which, by the Riemann–Lebesgue lemma, approaches zero as $|\omega| \to \infty$. Since $\inf_{\operatorname{Re} s = 0} |1 + KG(s)| > 0$ by assumption (i), it thus follows that $\inf_{\omega \in R} |1 + KA(j\omega)| > 0$. Since $\lim_{N\to\infty} \theta(N2\pi T^{-1})$ exists by assumption (ii) it follows that the argument $\Phi(\omega)$ of $1 + KA(j\omega)$ satisfies $\Phi(2\pi T^{-1}) = \Phi(0)$. Thus by the principle of the argument there are no zeros of the function

$$R(z) = \sum_{n=0}^{\infty} g_n z^n$$

inside the unit circle since $R(z)$ is analytic inside the unit circle and since the increase in its argument as z moves around the unit circle equals zero. Thus the function $I + KA(s)$ has no zeros in $\operatorname{Re} s \geqq 0$.

Consider now the contour in the complex plane shown in Fig. 6. The increase of the argument of $1 + KG_1(s)$ as s moves around this contour is zero for N and σ sufficiently large. Indeed, along C_1 it is zero by the assumption

$$\lim_{N\to\infty} \theta(N2\pi T^{-1}) = 0,$$

and along C_2, C_3, C_4 it is zero since $G_1(s)$ is arbitrarily close to $A(s)$ along that part of the contour. Hence $1 + KG_1(s)$ has, by the principle of the argument, no zeros in any finite part of the half-plane $\operatorname{Re} s \geqq 0$. It is bounded away from zero in $\operatorname{Re} s \geqq 0$ since it arbitrarily closely approximates $|1 + KG_1(s)|$ for large values of

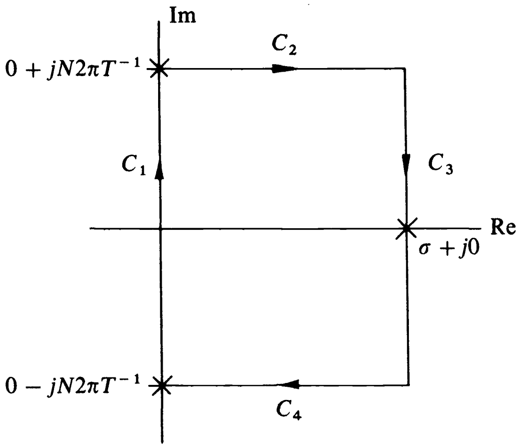


FIG. 6. *A contour in the complex plane*

$|s|$ in $\operatorname{Re} s \geqq 0$. Thus $|1 + KG_1(s)|$ is indeed bounded away from zero in $\operatorname{Re} s \geqq 0$, as claimed.

(ii) The above argument is easily reversed to yield the converse. The details are left to the reader.

Example. Consider the case where the forward loop is a pure delay, i.e.,

$$G_1 x(t) = x(t - T), \qquad T > 0 \text{ given.}$$

Then $1 + KG(j\omega) = 1 + K\,e^{-j\omega T}$, and

$$\theta(N2\pi T^{-1}) = \begin{cases} 0 & \text{when } |K| < 1, \\ -2\pi N & \text{when } |K| > 1, \end{cases}$$

$$\inf_{\omega} |1 + KG(j\omega)| = 0 \qquad \text{when } |K| = 1.$$

Thus the system is stable if and only if $|K| < 1$. The instability conclusion can in fact easily be verified directly by considering the input

$$u(t) = \begin{cases} 1, & 0 \leqq t \leqq T/2, \\ 0, & \text{otherwise}. \end{cases}$$

8.2. Stability and instability in the time-varying case. All the elements are now available to state a converse of the well-known circle criterion, when applied to linear time-varying systems. The results which follow are similar to those obtained by Brockett and Lee [12]. Let $0 < \underline{k} \leqq \bar{k}$, and let C be the disc centered on the real axis of the complex plane which passes through the points $-\underline{k}^{-1} + 0.j$ and $-\bar{k}^{-1} + 0.j$; let P denote the set in the complex plane determined by

$$P \triangleq \{G_1(s) | \operatorname{Re} s \geqq 0\};$$

and let Z denote the set in the complex plane determined by $Z \triangleq \{G_1(s) | \operatorname{Re} s = 0\}$. Let $d(C, P) \triangleq \inf_{x \in C, y \in P} |x - y|$, and let $d(C, Z)$ be similarly defined.

THEOREM 8.1. *Let $G_1(s)$ denote the Laplace transform of $(g, \{g_k, t_k\})$. Let $W = L_2(S) \times L_2(S)$ be the space with respect to which stability is defined and assume that $g_0 = 0$ or $g_0^{-1} \notin [\underline{k}, \bar{k}]$. Assume that $d(C, Z) > 0$. Then the feedback system under consideration is stable if $d(C, P) > 0$; it is unstable if $d(C, P) = 0$.*

Proof. Consider first the case $S = (-\infty, +\infty)$. It then suffices to prove, by the results in § 8.1 and Theorems 4.2 and 7.1, that $\|(I + cG_1)^{-1}G_1(G_2 - cI)\| < 1$ with $c = (\underline{k} + \bar{k})/2$. If $\underline{k} = \bar{k}$, then the theorem specializes to Lemma 7.1. Assume thus that $\underline{k} \neq \bar{k}$. The operator $(I + cG_1)^{-1}G_1$ corresponds to multiplication of the limit-in-the-mean transform of the element on which it operates by $G'(j\omega) = (1 + cG_1(j\omega))^{-1}G_1(j\omega)$ and thus $\|(I + cG_1)^{-1}G_1\| = \|G'(j\omega)\|_{L_\infty}$. By assumption, $\|G'(j\omega)\|_{L_\infty} < 2(\bar{k} + \underline{k})^{-1}$. The operator $G_2 - cI$ corresponds to multiplication (in the time domain) by $k(t) - c$ and $\|G_2 - cI\| = \|k(t) - c\|_{L_\infty}$, which equals $(\underline{k} + \bar{k})/2$. Thus

$$\|(I + cG_1)^{-1}G_1(G_2 - cI)\| \leqq \|(I + cG_1)^{-1}G_1\| \, \|G_2 - cI\| < 1,$$

and the conclusion follows.

For the case $S = [T_0, \infty)$, consider the backward extension of W and G by letting $W' = L_2(-\infty, +\infty) \times L_2(-\infty, +\infty)$ and defining G' on W' by

$$G'_1x(t) = \sum_{n=0}^{\infty} g_nx(t - t_n) + \int_{-\infty}^{+\infty} g(t - \tau)x(\tau)\,d\tau$$

and $G'_2x(t) = k'(t)x(t)$ with

$$k'(t) = \begin{cases} k(t) & \text{for } t \geqq T_0, \\ \frac{1}{2}(\underline{k} + \bar{k}) & \text{for } t < T_0, \end{cases}$$

and $x(t) \in L_2(-\infty, +\infty)$. It is a simple matter to verify that W' and G' indeed qualify as backward extensions of W and G. The remainder of the proof is then based on Theorems 4.1 and 7.2 by identical estimates as used above for the case $S = (-\infty, +\infty)$.

Remark 8.4. It is again possible to replace the condition $d(C, P) > 0$ or $d(C, P) = 0$ by an encirclement condition, at least in the case when all the delays are equally spaced, i.e., $t_k = kT$, $T > 0$. In fact, $d(C, P) > 0$ if and only if $d(C, Z) > 0$ and $\lim_{N\to\infty} \theta(N2\pi T^{-1})$ exists and is zero where $\theta(\omega)$ is the argument of $1 + \alpha G(j\omega)$ and α is an arbitrary element of C.

Remark 8.5. Let $Z_\sigma = \{G_1(s)|\mathrm{Re}\, s = \sigma\}$. It can be shown that it suffices for instability that $d(C, Z_\sigma) > 0$ for some $\sigma \geqq 0$ and that $d(C, P) = 0$. This in fact leads to an improved instability criterion in situations as the one illustrated in Fig. 7.

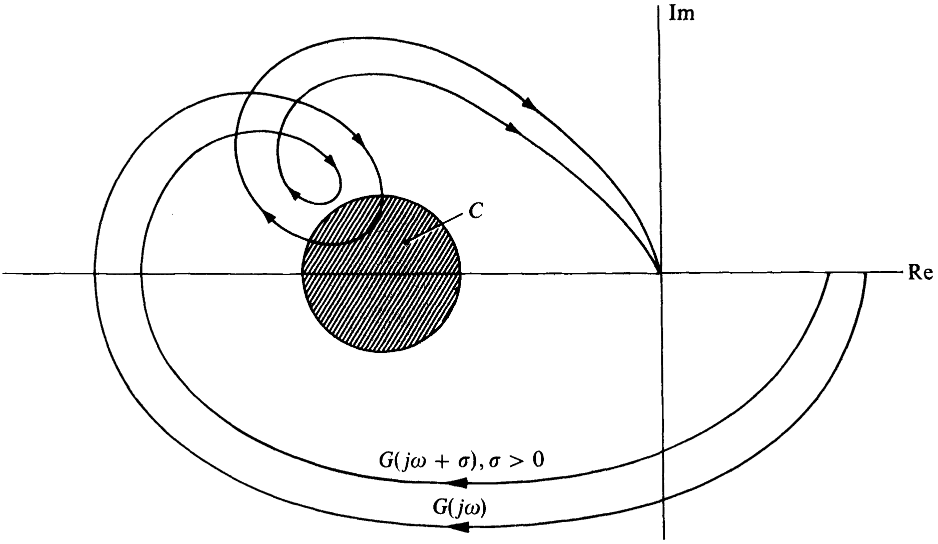


FIG. 7. *Illustration of Remark* 8.5

9. Concluding remarks.

Remark 9.1. The main results of this paper consist of Theorems 4.1 and 4.2. Although rather straightforward in their proofs, they spell out precisely what a priori conditions on invertibility in the extended space one needs in order to draw the conclusions. Note that whether the question is continuity or just stability leads to an essential difference in these a priori conditions. On a semi-infinite interval of definition, Theorems 4.1 and 4.2 are in fact completely parallel. On the doubly-infinite interval of definition, however, Theorem 4.1 will for all practical purposes be useless as it stands since the a priori existence assumption of the inverse on the extended space becomes then very stringent. It could be questioned whether or not, also in Theorem 4.1, continuability in the extended space is sufficient as an a priori assumption. The necessity part of Theorem 4.1 still stands, but it is not clear whether the conditions are then still sufficient for stability. The answer to the above question is in the negative, and simple counterexamples consisting of a delay in the forward loop and an instantaneous nonlinearity in the feedback loop can be constructed.

Other contributions of the paper are believed to be the framework and estimations involved in obtaining the instability Theorems 7.1 and 7.2.

Remark 9.2. Not all of the assumptions enumerated in § 3.3 are strictly needed in various theorems. The following meaningful relaxations can be made:[7] Assumption A1 (iii) can be relaxed in Theorems 4.1, 5.1, 5.2, 6.1, 7.2, and in the sufficiency part of Theorem 6.2. Assumption A1 (iv) can be relaxed in the sufficiency part of Theorems 6.1 and 6.2. Assumption A2 (iii) is not needed in Theorems 4.1, 4.2, 5.1, 5.2, and in the necessity part of Theorems 6.1 and 6.2. Assumption A2 (ii) does not enter in Theorem 5.1, only $G(0) = 0$ enters in Theorem 5.2, and only that G maps W into itself enters in Theorems 4.1 and 6.1.

Remark 9.3. All of the theorems in the paper remain valid if the roles of G_1 and G_2 are reversed (this was implicitly done in § 8), or if G_1 and G_2 are replaced by MG_1N and $N^{-1}G_2M^{-1}$ with M, N causal operators with causal inverses M^{-1} and N^{-1}. Some of the more interesting stability criteria (e.g., [8]) can thus be given an instability converse.

Remark 9.4. It is interesting to note that stability theorems yield inverse function theorems. More specifically, assume causal invertibility of $I + G_2G_1$ on W_e (a weak assumption for causal systems), and assume that $\|G_2G_1\| < 1$. Then G_2G_1 is invertible on W. Observe that the contraction mapping principle would require $\|\widetilde{G_2G_1}\| < 1$. It is precisely the causality of the operators involved which allows this relaxation.

Remark 9.5. Possible extensions of the results in this paper include, for instance, the treatment of open-loop unstable systems which violate the assumption that G maps W into itself. Note that an extension of Theorems 7.1 and 7.2 to nonlinear systems does not appear as important as it seems at first sight. More specifically, if the resulting theorems imply the instability of the equations linearized around the null solution then some direct estimates will lead to the instability conclusion more directly.

Acknowledgments. The author is indebted to Professor R. A. Skoog, Professor R. W. Brockett, and the referees, particularly the associate editor-in-charge,

[7] This does not include relaxations which are essentially only formal in the sense that then the assumptions of the theorem will rarely be satisfied. For instance, assumption A1 (iv) is not strictly needed in Theorem 4.1 but will most likely enter in the verification of the existence assumption which was presupposed in that theorem.

for making some essential suggestions. He also wishes to thank Mr. J. Davis for clarifying some points.

REFERENCES

[1] I. W. Sandberg, *On the L_2-boundedness of solutions of nonlinear functional equations*, Bell System Tech. J., 43 (1964), pp. 1581–1599.

[2] ———, *Some results on the theory of physical systems governed by nonlinear functional equations*, Ibid., 44 (1965), pp. 871–898.

[3] G. Zames, *On the stability of nonlinear, time-varying feedback systems*, Proc. 1964 National Electronics Conference, vol. 20, Chicago, Illinois, 1964, pp. 725–730.

[4] ———, *On the input-output stability of time-varying nonlinear feedback systems. Part I: Conditions derived using concepts of loop gain, conicity, and positivity. Part II: Conditions involving circles in the frequency plane and sector nonlinearities*, IEEE Trans. Automatic Control, 11 (1966), pp. 228–239, 465–476.

[5] ———, *Functional analysis applied to nonlinear feedback systems*, IEEE Trans. Circuit Theory, 10 (1936), pp. 392–404.

[6] ———, *Realizability conditions for nonlinear feedback systems*, Ibid., 11 (1964), pp. 186–194.

[7] G. Zames and P. L. Falb, *Stability conditions for systems with monotone and slope-restricted nonlinearities*, this Journal, 6 (1968), pp. 89–108.

[8] M. Freedman and G. Zames, *Logarithmic variation criteria for the stability of systems with time-varying gains*, this Journal, 6 (1968), pp. 487–507.

[9] N. Dunford and J. T. Schwartz, *Linear Operators*, Interscience, New York, 1967.

[10] F. Riesz and B. Sz-Nagy, *Functional Analysis*, Frederick Ungar, New York, 1955.

[11] R. W. Brockett and J. L. Willems, *Frequency domain stability criteria. Parts I and II*, IEEE Trans. Automatic Control, 10 (1965), pp. 255–261, 401–413.

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

[13] E. Hille and R. S. Phillips, *Functional Analysis and Semi-Groups*, vol. 31, 2nd ed., Colloquium Publications, American Mathematical Society, Providence, 1957.

[14] C. A. Desoer, *A general formulation of the Nyquist criterion*, IEEE Trans. Circuit Theory, 12 (1965), pp. 230–234.

[15] C. A. Desoer and M. Y. Wu, *Stability of linear time-invariant systems*, Ibid., 15 (1968), pp. 245–250.

[16] L. H. Loomis, *An Introduction to Abstract Harmonic Analysis*, Van Nostrand, New York, 1953.