

BOUNDARY CONTROL OF TEMPERATURE DISTRIBUTIONS IN A PARALLELEPIPEDON*

H. O. FATTORINI†

Abstract. Let u_0 be the temperature distribution of a homogeneous parallelepipedon at time $t = 0$. We study the problem of regulating the boundary temperature of the parallelepipedon in such a way that the temperature distribution at time $t = T$ coincides with a function fixed in advance. The problem is shown to have a solution under some restrictions on the final temperature distribution.

1. Introduction. We consider the n -dimensional heat flow equation

$$(1.1) \quad \frac{\partial u}{\partial t} = \kappa \Delta u = \kappa \sum_{j=1}^n \frac{\partial^2 u}{\partial x_j^2},$$

$0 \leq x_j \leq X_j$ ($1 \leq j \leq n$), $0 \leq t \leq T$, $\kappa > 0$. This equation governs the evolution of the temperature distribution of a homogenous "body" occupying the n -dimensional parallelepipedon P defined by the inequalities $0 \leq x_j \leq X_j$ ($1 \leq j \leq n$), where X_1, X_2, \dots, X_n are fixed positive numbers. We assume that the temperature $u(x, t) = u(x_1, x_2, \dots, t)$ satisfies the boundary conditions

$$(1.2) \quad u(x, t) = f(x, t) \quad (x \in B = \text{boundary of } P, 0 \leq t \leq T),$$

where f is interpreted as a control or steering function by means of which we try to influence the evolution of the temperature distribution $u(x, t)$ in the whole body in a sense to be defined below.

The *controllability problem* for the system (1.1), (1.2) can be formulated as follows. Given an initial condition

$$(1.3) \quad u(x, 0) = u_0(x)$$

and a final condition

$$(1.4) \quad u(x, T) = u_T(x),$$

can we find a control f such that the solution $u(x, t)$ of (1.1), (1.2), (1.3) also satisfies (1.4)?

To consider this problem we shall need the standard existence theorem for (1.1), (1.2), (1.3), namely, the following.

THEOREM 1.1. *Assume that u_0 is continuous in P , that f is continuous in $B \times [0, T]$ and that*

$$(1.5) \quad u_0(x) = f(x, 0), \quad x \in B.$$

Then there exists a unique solution u of (1.1), (1.2), (1.3). Moreover, u is continuous in $P \times [0, T]$.

* Received by the editors October 31, 1972.

† Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales, Departamento de Matemática, Ciudad Universitaria, Buenos Aires, Argentina. Now at University of California, Department of Mathematics, Los Angeles, California 90024. This research was supported in part by the National Science Foundation under Contract GP-9658.

For a statement and discussion of more general results than Theorem 1.1 see A. Milgram's appendix to [1]; note that $P \times [0, T]$ satisfies the "cone condition" there. A proof for the case $n = 1$ can be found in [11] in a way easily generalizable to the case $n > 1$. Note, finally, that "solution" is understood in the classical sense; u has continuous derivatives up to the order of the ones appearing in (1.1)—in fact, of all orders—in the interior of $P \times [0, T]$. Continuity of u in all of $P \times [0, T]$ guarantees that condition (1.4) can be given meaning.

It is possible to deduce from the results in [3] certain density results for the set of all u_T such that the controllability problem has a solution, keeping u_0 fixed, and also interchanging the roles of u_0 and u_T . We improve here these results by giving precise sufficient conditions on u_0, u_T in order that the controllability problem should have a solution (Theorem 4.1 and Corollary 4.3). In particular, these conditions show that the *null controllability problem* ($u_T = 0$) always has a solution. As can be expected from the smoothing properties of the heat equation, the conditions on u_T for solution of the controllability problem are rather severe.

The same type of problem for parabolic equations in one space variable was treated in [4] by reduction to a moment problem. It was pointed out to the author by D. L. Russell that the present problem can be reduced to a family of moment problems that must be "uniformly" solved in a certain sense. This calls for uniform estimates on the norms of certain biorthogonal sequences, a task that has been carried out in [5] for different classes of sequences (we note that the estimates in [5] are applied there to the solution of the controllability problem in a sphere in n -dimensional space). It should be pointed out that, although the present results are of a much more elementary nature than those of [5], they are not contained in them; on the other hand, we have not made any effort to minimize overlap between the present paper and [5] in order to keep it reasonably self-contained.

A solution of the controllability problem for an arbitrary bounded domain in n -dimensional space has very recently been announced by Russell (private communication). However, the characterization of the final states u_T is less precise than the one given. Besides, the case considered here has two more interesting features of its own. First, it turns out that we do not lose much, in a qualitative sense, if we apply control in only one of the 2^n faces of B , instead of doing it on all of B . Second, one can deduce sufficient conditions on u_T for solution of the controllability problem that do not depend on its Fourier coefficients and are, besides, very simple to check. These two observations should be compared with the results in [5] for a similar situation.

The moment problems mentioned earlier are obtained in § 2. These problems are solved in § 4, after all the necessary estimates on the norms of biorthogonal sequences are carried out in § 3. Section 4 ends with a proof that the results are, in a suitable sense, best possible.

2. Reduction to moment problems. Let A be the operator in $L^2(P)$ defined by

$$(2.1) \quad Au = \kappa \Delta u$$

with domain $D(A)$ consisting of all functions $u \in L^2(P)$ such that Δu (understood in the sense of distributions) belongs to $L^2(P)$ and

$$(2.2) \quad u(x) = 0, \quad x \in B.$$

It is well known (see [3]) that A is self-adjoint and has pure point spectrum; its eigenvalues are $\{-\lambda_\alpha\}$, where

$$(2.3) \quad \lambda_\alpha = \kappa(c_1^2\alpha_1^2 + c_2^2\alpha_2^2 + \cdots + c_n^2\alpha_n^2).$$

(Here $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_n)$ is an arbitrary vector of positive integers and $c_j = \pi/X_j$, $1 \leq j \leq n$.) The (normalized) eigenfunction corresponding to λ_α is

$$(2.4) \quad \varphi_\alpha = 2^{n/2}(X_1 X_2 \cdots X_n)^{-1/2} \sin c_1 \alpha_1 x_1 \cdots \sin c_n \alpha_n x_n.$$

Assume u_0, f satisfy the assumptions in Theorem 1.1 and let $u(x, t)$ be the solution of (1.1), (1.2), (1.3) provided there. Given α and $T > 0$, define

$$w_\alpha(x, t) = e^{\lambda_\alpha(t-T)} \varphi_\alpha(x), \quad 0 \leq t \leq T.$$

Then

$$\frac{\partial w_\alpha}{\partial t} = -\kappa \Delta w_\alpha$$

in $P \times [0, T]$; moreover, w_α vanishes in $B \times [0, T]$. It follows from the divergence theorem that

$$(2.5) \quad \begin{aligned} 0 &= \int_{P \times [0, T]} w_\alpha \left(\frac{\partial u}{\partial t} - \kappa \Delta u \right) dx dt \\ &= \int_P u(x, T) \varphi_\alpha(x) dx - e^{-\lambda_\alpha T} \int_P u_0(x) \varphi_\alpha(x) dx \\ &\quad + \kappa \int_0^T e^{-\lambda_\alpha(T-t)} dt \int_B f(x, t) \frac{\partial}{\partial \nu} \varphi_\alpha(x) d\sigma, \end{aligned}$$

where $\partial/\partial \nu$ and $d\sigma$ indicate, respectively, the outer normal derivative and the element of area in B . We assume from now on that f vanishes in all but one of the $2^n (n-1)$ -dimensional faces that make up the boundary of P (the effect of this as regards the controllability problem will be examined in §4). Obviously, we may assume that the part of B where f does not vanish is

$$B_0 = \{x \in B; x_n = 0\}.$$

In B_0 we have

$$(2.6) \quad \begin{aligned} \frac{\partial}{\partial \nu} \varphi_\alpha(x) &= -2^{n/2} \pi \alpha_n (X_1 \cdots X_{n-1})^{1/2} X_n^{-3/2} \sin c_1 \alpha_1 x_1 \cdots \sin c_{n-1} \alpha_{n-1} x_{n-1} \\ &= -B \alpha_n 2^{(n-1)/2} (X_1 \cdots X_{n-1})^{-1/2} \sin c_1 \alpha_1 x_1 \cdots \sin c_{n-1} \alpha_{n-1} x_{n-1} \\ &= -B \alpha_n \eta_\beta(y), \end{aligned}$$

where we have set $\beta = (\alpha_1, \dots, \alpha_{n-1})$, $y = (x_1, \dots, x_{n-1})$, $B = 2^{1/2} \pi X_n^{-3/2}$. Clearly, $\{\eta_\beta(y)\}$ is a complete orthonormal system in B_0 ; moreover,

$$|\eta_\beta(y)| \leq Q = 2^{(n-1)/2} (X_1 \cdots X_{n-1})^{1/2}, \quad y \in B_0.$$

Assume that the controllability problem of § 1 has a solution f , continuous in $B_0 \times [0, T]$ for some initial state

$$(2.7) \quad u_0(x) = \sum_{\alpha} \mu_{\alpha} \varphi_{\alpha}(x)$$

and some final state

$$(2.8) \quad u_T(x) = \sum_{\alpha} \nu_{\alpha} \varphi_{\alpha}(x).$$

Let, for each β ,

$$(2.9) \quad g_{\beta}(t) = \int_{B_0} g(y, t) \eta_{\beta}(y) dy,$$

where $g(y, t) = f(y, T - t)$. Then we deduce from (2.5) that g_{β} must be a solution of the moment problem

$$(2.10) \quad \int_0^T e^{-\lambda_{\alpha} t} g_{\beta}(t) dt = -(e^{-\lambda_{\alpha} T} \mu_{\alpha} - \nu_{\alpha}) / B \kappa \alpha_n, \quad \alpha_n = 1, 2, \dots$$

These moment problems will be solved in the following way: let β be fixed and denote by $\{\psi_{\alpha}\}$, $\alpha_n = 1, 2, \dots$, a sequence biorthogonal to $\exp(-\lambda_{\alpha} t)$ in $(0, T)$, that is, such that

$$\int_0^T \psi_{\alpha}(t) e^{-\lambda_{\alpha'} t} dt = \delta_{\alpha_n \alpha_n'}.$$

(here δ is the Kronecker delta, $\alpha' = (\alpha_1, \dots, \alpha_{n-1}, \alpha'_n)$). Then, at least formally,

$$(2.11) \quad g_{\beta}(t) = \sum_{\alpha_n=1}^{\infty} [(e^{-\lambda_{\alpha} T} \mu_{\alpha} - \nu_{\alpha}) / B \kappa \alpha_n] \psi_{\alpha}(t)$$

is a solution of (2.10). As for g , it will be given by the expression

$$(2.12) \quad g(y, t) = \sum_{\beta} g_{\beta}(t) \eta_{\beta}(y)$$

provided that (2.11) and (2.12) are shown to be convergent in suitable topologies. This will be justified in § 4 by means of precise estimates for a particular bi-orthogonal sequence.

3. Biorthogonal sequences. Let $\Lambda = \{\lambda_n\}$ be a sequence of real numbers such that $0 < \lambda_0 < \lambda_1 < \dots$,

$$\sum 1/\lambda_n < \infty.$$

The distance $d(n)$ in $L^2(0, \infty)$ from $e^{-\lambda_n t}$ to $E^{(n)}$, the subspace generated by $\{e^{-\lambda_k t}; k \neq n\}$, has been computed by Kaczmarz and Steinhaus [6, Chap. III, § 6]. We have

$$(3.1) \quad d(n) = \frac{1}{(2\lambda_n)^{1/2}} \prod'_{k=0}^{\infty} \left| \frac{\lambda_k - \lambda_n}{\lambda_k + \lambda_n} \right|, \quad n = 0, 1, \dots$$

Here \prod' means that the term corresponding to $k = n$ should be omitted from the infinite product.

If r_n is the unique element of $E^{(n)}$ that lies closest to $e^{-\lambda_n t}$, then it is plain that the sequence

$$(3.2) \quad \psi_n(t) = d(n)^{-2}(e^{-\lambda_n t} - r_n(t)), \quad n = 0, 1, \dots,$$

is biorthogonal to the sequence $e^{-\lambda_n t}$ in $L^2(0, \infty)$. Moreover, it is the biorthogonal sequence *with smallest possible norm*. For, if $\{\tilde{\psi}_n\}$ is another such sequence, then $\tilde{\psi}_n - \psi_n$ belongs to E^\perp (E is the subspace of $L^2(0, \infty)$ generated by *all* the exponentials $\{e^{-\lambda_n t}\}$). Accordingly, $\tilde{\psi}_n = \psi_n + \varphi_n$ ($\varphi_n \in E^\perp$) and, as $\psi_n \in E$,

$$(3.3) \quad \|\psi_n\| \leq \|\tilde{\psi}_n\|$$

(here and afterwards, $\|\cdot\|$ indicates the norm in $L^2(0, \infty)$). The norm of ψ_n can be easily computed from (3.2);

$$\|\psi_n\| = 1/d(n).$$

We shall apply these observations to the sequence (or, rather, family of sequences)

$$(3.4) \quad \lambda_0 = \sigma, \quad \lambda_n = \omega n^2 + \mu, \quad n = 1, 2, \dots$$

Here σ, ω are fixed positive numbers ($\sigma < \omega$) and μ is a parameter varying in the interval $[0, \infty)$. We deduce that, for each $\mu \geq 0$ there exists a sequence $\{\psi_{n,\omega,\sigma,\mu}(t); n \geq 1\}$ in $L^2(0, \infty)$, biorthogonal to $\{e^{-(\omega n^2 + \mu)t}; n \geq 1\}$ such that

$$(3.5) \quad \int_0^\infty e^{-\sigma t} \psi_{n,\omega,\sigma,\mu}(t) dt = 0, \quad n \geq 1, \quad \mu \geq 0.$$

According to (3.1) and (3.2),

$$(3.6) \quad \begin{aligned} \|\psi_{n,\omega,\sigma,\mu}\| &= (2(\omega n^2 + \mu))^{1/2} \frac{\omega n^2 + \mu + \sigma}{\omega n^2 + \mu - \sigma} \prod_{k=1}^\infty \left| \frac{\omega k^2 + \omega n^2 + 2\mu}{\omega k^2 - \omega n^2} \right| \\ &= \frac{\omega n^2(\omega n^2 + \mu + \sigma)}{(2(\omega n^2 + \mu))^{1/2}(\omega n^2 + \mu - \sigma)} \cdot \frac{\prod_{k=1}^\infty (1 + (n^2 + 2\mu/\omega)/k^2)}{\prod_{k=1}^\infty |1 - n^2/k^2|} \\ &= \frac{2^{1/2} \omega^{3/2} (\omega n^2 + \mu + \sigma) n^2}{\pi (\omega n^2 + \mu)^{1/2} (\omega n^2 + 2\mu)^{1/2} (\omega n^2 + \mu - \sigma)} \sinh \pi (n^2 + 2\mu/\omega)^{1/2}. \end{aligned}$$

Let $E(\omega, \sigma, \mu)$ be the subspace of $L^2(0, \infty)$ generated by the exponentials $\{e^{-\lambda_n t}; n \geq 0\}$ where the λ_n are given by (3.4), and let $E_T(\omega, \sigma, \mu)$ be the subspace of $L^2(0, T)$ generated by their restrictions to $(0, T)$. It follows from a result of L. Schwartz [10] that the restriction operator $Q_T(\omega, \sigma, \mu): E(\omega, \sigma, \mu) \rightarrow E_T(\omega, \sigma, \mu)$ has a bounded inverse for each $\mu \geq 0$. A slight modification of his argument (see a more general result in [5]) yields the following lemma,

LEMMA 3.1. *For each $T > 0$,*

$$(3.7) \quad \sup_{\mu \geq 0} \|Q_T(\omega, \sigma, \mu)^{-1}\| = K_{\omega,\sigma}(T) < \infty.$$

Proof. Assume the conclusion of Lemma 3.1 is false for some $T > 0$. Then there exists a sequence $\{p_m\}$ of exponential polynomials

$$\begin{aligned} p_m(t) &= \sum_n a_{mn} e^{-(\omega n^2 + \mu(m))t} + b_m e^{-\sigma t} \\ &= q_m(t) + b_m e^{-\sigma t} \end{aligned}$$

such that

$$(3.8) \quad \|p_m\| = 1$$

while

$$(3.9) \quad \|p_m\|_{(0,T)} \rightarrow 0 \quad \text{as } m \rightarrow \infty$$

(here $\|\cdot\|_{(0,T)}$ indicates the norm in $L^2(0, T)$). If $\psi_{n,\omega,\sigma,\mu}$ is the sequence in (3.6), clearly $a_{mn} = \langle p_m, \psi_{n,\omega,\sigma,\mu} \rangle$ ($\langle \cdot, \cdot \rangle$ is the scalar product in $L^2(0, \infty)$). Hence,

$$(3.10) \quad |a_{mn}| \leq C e^{\pi(n^2 + 2\mu(m)/\omega)^{1/2}}, \quad m, n \geq 1.$$

(Here and in subsequent estimates C indicates a positive constant—not necessarily the same for different inequalities—independent of the parameters subject to variation, in this case n and μ .) Let $\rho > 0$ and z be a complex number such that $\operatorname{Re} z \geq \rho$. Then it is easy to see that (3.10) implies

$$(3.11) \quad |q_m(z)| \leq C e^{-\sigma \operatorname{Re} z}, \quad \operatorname{Re} z \geq \rho, \quad m \geq 1.$$

Accordingly, we can apply Montel's theorem on normal families of holomorphic functions to deduce that, if necessary, passing to a subsequence, $\{q_m\}$ converges uniformly on compacts of $\operatorname{Re} z \geq \rho$ to a holomorphic function q ; in view of (3.11) and of the Lebesgue dominated convergence theorem, $q_m \rightarrow q$ in $L^2(\rho, \infty)$. From this and from boundedness of p_m we deduce boundedness of $b_m e^{-\sigma t}$ in $L^2(\rho, \infty)$; then, if necessary, passing again to a subsequence, we deduce that

$$(3.12) \quad p_m \rightarrow q + b e^{-\sigma t} \quad \text{in } L^2(\rho, \infty).$$

If we combine (3.12) (say, for $\rho = T/2$) with (3.9) we see that p_m converges in $L^2(0, \infty)$ to a function that must vanish almost everywhere in $(0, T)$; since $q + b e^{-\sigma t}$ is holomorphic, it must be identically zero, which is impossible in view of (3.8). This ends the proof of Lemma 3.1.

Now define

$$(3.13) \quad \tilde{\psi}_{n,\omega,\sigma,\mu,T}(t) = e^{-\sigma t} \{ [Q_T(\mu + \sigma)^{-1}]^* \psi_{n,\omega,\sigma,\mu+\sigma} \}(t)$$

for $n \geq 1, \mu \geq 0$. A few simple manipulations show that the sequence $\tilde{\psi}_{n,\omega,\sigma,\mu,T}$ is biorthogonal to $\{e^{-(\omega n^2 + \mu)t}; n \geq 1\}$ in $L^2(0, T)$ for any $\mu \geq 0$ and that

$$(3.14) \quad \int_0^T \tilde{\psi}_{n,\omega,\sigma,\mu,T}(t) dt = 0, \quad n \geq 1, \quad \mu \geq 0.$$

Integrating by parts the biorthogonality relations and taking (3.14) into account we deduce without difficulty that the sequence

$$\psi_{n,\omega,\sigma,\mu,T}(t) = (\omega n^2 + \mu) \int_0^t \tilde{\psi}_{n,\omega,\sigma,\mu,T}(s) ds, \quad n \geq 1,$$

is as well biorthogonal to $\{e^{-(\omega n^2 + \mu)t}; n \geq 1\}$ in $L^2(0, T)$. It follows from (3.13) and from Schwarz's inequality that the $L^1(0, T)$ -norm of $\tilde{\psi}_{n,\omega,\sigma,\mu,T}(t)$ does not exceed

$$(2\sigma)^{-1/2} K_{\omega,\sigma}(T) \|\psi_{n,\omega,\sigma,\mu+\sigma}\|.$$

Then,

$$(3.15) \quad |\psi_{n,\omega,\sigma,\mu,T}(t)| \leq K_{\omega,\sigma}(T)R_{n,\omega,\sigma,\mu}, \quad 0 \leq t \leq T, \quad n \geq 1, \quad \mu \geq 0,$$

where we have set

$$(3.16) \quad R_{n,\omega,\sigma,\mu} = (\omega n^2 + \mu)(2\sigma)^{-1/2} \|\psi_{n,\omega,\sigma,\mu+\sigma}\|.$$

4. Solution of the moment problems. We go back to (2.10) and to its formal solution by means of the series (2.12). The results in §3 will be applied to the following values of the various parameters involved. The index n in §3 will be replaced by α_n ; as for ω, μ ,

$$\omega = \kappa c_n^2, \quad \mu = \kappa(c_1^2 \alpha_1^2 + \cdots + c_{n-1}^2 \alpha_{n-1}^2).$$

Given $T > 0$, the functions $\psi_{n,\omega,\sigma,\mu,T}$ of §3 will be written $\psi_{\alpha,T}$ (thus omitting explicit reference to σ). $K_{\omega,\sigma}(T)$ will be written $K(T)$, $R_{n,\omega,\sigma,\mu}$ will be written R_α , etc. In this notation, inequality (3.16) becomes

$$(4.1) \quad |\psi_{\alpha,T}(t)| \leq K(T)R_\alpha(T), \quad 0 \leq t \leq T, \quad n \geq 1, \quad \mu \geq 0,$$

where, in view of (3.6),

$$(4.2) \quad R_\alpha \leq C|\alpha|^2 \exp \pi(2(X_n/X_1)^2 \alpha_1^2 + \cdots + 2(X_n/X_{n-1})^2 \alpha_{n-1}^2 + \alpha_n^2)^{1/2}.$$

Set $c_\alpha = (e^{-\lambda_\alpha T} \mu_\alpha - \nu_\alpha)/B\kappa\alpha_n$. Given a function f in, say, $L^1(P)$ and some $T \geq 0$ define

$$p_T(f) = (B\kappa)^{-1} \sum_{\alpha} e^{-\lambda_\alpha T} R_\alpha |a_\alpha|/\alpha_n,$$

where the $\{a_\alpha\}$ are the Fourier coefficients of f with respect to the $\{\varphi_\alpha\}$, that is,

$$f \sim \sum a_\alpha \varphi_\alpha.$$

Clearly, $p_T(f) < \infty$ if $T > 0$ for any f .

Assume

$$(4.3) \quad p_0(u_T) < \infty.$$

Then it is clear that the series (2.11) converges absolutely and uniformly in $0 \leq t \leq T$; as for the series (2.12), defining g in terms of the g_β , it is also uniformly and absolutely convergent—this time in $B_0 \times [0, T]$ —and

$$(4.4) \quad |g(y, t)| \leq QK(T)[p_T(u_0) + p_0(u_T)].$$

We note that, as all the ψ_α vanish for $t = 0$ and $t = T$ (an immediate consequence of (3.14)), the same is true of each g_β and thus of g ; accordingly, if the initial state u_0 vanishes in B the compatibility condition (1.5) of Theorem 1.1 holds. We have thus proved the following theorem.

THEOREM 4.1. *The controllability problem of §1 has a solution f continuous in $B \times [0, T]$ (and with support in $B_0 \times [0, T]$) for any initial state u_0 , continuous in P and zero in B and for any final state u_T that satisfies (4.3).*

It should be pointed out that condition (4.3) is far from necessary for solvability of the control problem, even when $n = 1$. In fact, let T be an arbitrary positive number. The solution of

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad 0 \leq x \leq \pi, \quad 0 \leq t \leq T,$$

that satisfies

$$u(x, 0) = 0, \quad 0 \leq x \leq \pi,$$

as well as

$$u(0, t) = 0, \quad u(\pi, t) = \pi t(T - t), \quad 0 \leq t \leq T,$$

is

$$u(x, t) = 2 \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n^3} \left\{ \left(T + \frac{2}{n^2} \right) (e^{-n^2 t} - 1) + 2t \right\} \sin nx + xt(T - t).$$

But it is plain that, if $u_T(x) = u(x, T)$, then

$$p_0(u_T) = \infty.$$

On the other hand, condition (4.3) is the best possible of its type: for details on this, see Remark 2 (especially Lemma 4.5) in the following section.

Condition (4.3) is a consequence of the easier looking inequality

$$(4.5) \quad |a_\alpha| \leq M \exp \{ -(\pi + \varepsilon)(2(X_n/X_1)^2 \alpha_1^2 + \cdots + 2(X_n/X_{n-1})^2 \alpha_{n-1}^2 + \alpha_n^2)^{1/2} \}$$

for some $M, \varepsilon > 0$. An intrinsic characterization of the functions whose Fourier coefficients $\{a_\alpha\}$ satisfy an inequality of the form (4.5) is a consequence of the following result. In its statement and proof, $\gamma = (\gamma_1, \dots, \gamma_n)$ is an n -vector of integers, $\gamma_j = \dots, -1, 0, 1, \dots, 1 \leq j \leq n$.

LEMMA 4.2. *Let $f = f(x_1, x_2, \dots, x_n)$ be $2X_j$ -periodic in $X_j, 1 \leq j \leq n$. Then its Fourier series*

$$(4.6) \quad f \sim \sum_{\gamma} a_{\gamma} \exp i(c_1 \gamma_1 x_1 + \cdots + c_n \gamma_n x_n)$$

satisfies

$$(4.7) \quad |a_{\gamma}| \leq M \exp [-(1 + \varepsilon)(\kappa_1 c_1^2 \gamma_1^2 + \cdots + \kappa_n c_n^2 \gamma_n^2)^{1/2}],$$

$\kappa_1, \dots, \kappa_n > 0$, for some $M, \varepsilon > 0$ if and only if f can be analytically extended to the "ellipsoidal strip" $\Xi = \Xi(\kappa_1, \dots, \kappa_n)$ defined by the inequalities

$$(4.8) \quad -\infty < \operatorname{Re} z_j < \infty,$$

$$(4.9) \quad \sum_{j=1}^n |\operatorname{Im} z_j|^2 / \kappa_j \leq 1.$$

The proof is hardly different from the case of only one variable (see, for instance, [7]) and thus will only be sketched. The direct part of the proof is immediate for we only have to replace (x_1, \dots, x_n) in (4.6) by a complex vector (z_1, \dots, z_n) .

The coefficients of the resulting series can be bounded by

$$M \exp \left[-(1 + \varepsilon) \left(\sum_{j=1}^n \kappa_j c_j^2 \gamma_j^2 \right)^{1/2} + \sum_{j=1}^n c_j \gamma_j |\operatorname{Im} z_j| \right],$$

where, by Schwarz's inequality,

$$\sum_{j=1}^n c_j \gamma_j |\operatorname{Im} z_j| \leq \left(\sum_{j=1}^n |\operatorname{Im} z_j|^2 / \kappa_j \right)^{1/2} \left(\sum_{j=1}^n \kappa_j c_j^2 \gamma_j^2 \right)^{1/2}.$$

The converse part is proved as follows. Observe first that, if f is periodic and analytic in a region defined by (4.8), (4.9) then it will be so in a slightly larger region, say, the one defined by (4.8) and

$$(4.10) \quad \sum_{j=1}^n |\operatorname{Im} z_j|^2 / \kappa_j \leq (1 + \varepsilon)^2, \quad \varepsilon > 0.$$

The γ th Fourier coefficient of f is given by the formula

$$a_\gamma = V^{-1} \int f(x) \exp(-i(c_1 \gamma_1 x_1 + \cdots + c_n \gamma_n x_n)) dx,$$

the integration being carried out in the cube $-X_j \leq x_j \leq X_j$, $1 \leq j \leq n$, with V the volume of the cube. Taking advantage of the periodicity properties of f , the domain of integration can be deformed into the cube defined by the inequalities

$$-X_j \leq \operatorname{Re} z_j \leq X_j$$

and the equalities

$$\operatorname{Im} z_j = -\rho(1 + \varepsilon) \kappa_j c_j \gamma_j = \eta_j,$$

where

$$\rho = \left(\sum_{j=1}^n \kappa_j c_j^2 \gamma_j^2 \right)^{-1/2}.$$

We can then estimate (4.11) by

$$M \exp(-c_1 \gamma_1 \eta_1 - \cdots - c_n \gamma_n \eta_n) = M \exp[-(1 + \varepsilon)(\kappa_1 c_1^2 \gamma_1^2 + \cdots + \kappa_n c_n^2 \gamma_n^2)^{1/2}],$$

which ends the proof of Lemma 4.2.

We can now reformulate (a weaker version of) Theorem 4.1 as follows.

THEOREM 4.3. *The controllability problem has a solution f satisfying the conditions of Theorem 4.1 for any initial state u_0 continuous in P that vanishes in B and for any final state u_T which is odd, $2X_j$ -periodic in x_j , $1 \leq j \leq n$, and can be analytically extended to*

$$\Xi(2X_n^2/\pi, \dots, 2X_n^2/\pi, X_n^2/\pi).$$

It is usually the case in applications of control theory that the control function—in our problem, the boundary temperature f —cannot be arbitrarily large; for instance, f may have to obey the constraint

$$(4.11) \quad |f(y, t)| \leq \theta, \quad y \in B_0, \quad 0 \leq t \leq T,$$

for some $\theta > 0$. Clearly, the null reachability problem ($u_0 = 0$) will have a solution f satisfying (4.11) if

$$(4.12) \quad p_0(u_T) \leq \theta/QK(T).$$

Unfortunately, one needs to compute $K(T)$ explicitly in order to verify (4.12). This kind of information cannot be obtained from Lemma 3.1, whose proof is nonconstructive and sheds no light on the size of $K(T)$, nor on the nature of its dependence on T . Clearly, $K(T)$ exceeds 1 and does not increase as T grows: a slightly more refined analysis shows that $K(T)$ is strictly decreasing and continuous in $T > 0$. From our point of view, a more interesting result is found in the following lemma.

LEMMA 4.4. *Let $\rho > 0$. Then there exists a constant $M > 0$ such that*

$$(4.13) \quad 1 + \frac{1}{2} e^{-2\sigma T} \leq K(T) \leq 1 + M e^{-2\sigma T}, \quad T \geq \rho.$$

Proof. Let $p(t) = e^{-\sigma t} \in E_T(\omega, \sigma, \mu)$. Then

$$K(T) \geq \|p\|_{(0,\infty)}/\|p\|_{(0,T)} = (1 - e^{-2\sigma T})^{-1/2} \geq 1 + \frac{1}{2} e^{-2\sigma T}$$

(here $\|\cdot\|_{(0,T)}$ indicates the L^2 -norm in $(0, T)$ and $\|\cdot\|_{(0,\infty)}$ indicates the L^2 -norm in $(0, \infty)$).

We prove now the right-hand side of (4.13). It was shown in § 3 that there exists a sequence $\{\psi_{n,\omega,\sigma,\mu}; n \geq 0\}$ of functions in $L^2(0, \infty)$ biorthogonal to

$$(4.14) \quad \{e^{-\sigma t}, e^{-(\omega n^2 + \mu)t}; n \geq 1\}$$

and satisfying

$$(4.15.1) \quad \|\psi_{n,\omega,\sigma,\mu}\|_{(0,\infty)} \leq C e^{\pi(n^2 + 2\mu/\omega)^{1/2}}, \quad n \geq 1.$$

The norm of $\psi_{0,\omega,\sigma,\mu}$ was not explicitly calculated in § 3; however, it can be easily obtained from formula (3.1). We have

$$\begin{aligned} \|\psi_{0,\omega,\sigma,\mu}\|_{(0,\infty)} &= (2\sigma)^{1/2} \prod_{k=1}^{\infty} \left(\frac{\omega k^2 + \mu + \sigma}{\omega k^2 + \mu - \sigma} \right) \\ (4.15.2) \quad &= (2\sigma)^{1/2} \frac{\prod_{k=1}^{\infty} (1 + (\mu + \sigma)/(\omega k^2))}{\prod_{k=1}^{\infty} (1 + (\mu - \sigma)/(\omega k^2))} \\ &= (2\sigma)^{1/2} \left(\frac{\mu + \sigma}{\mu - \sigma} \right)^{-1/2} \frac{\sinh \pi[(\mu + \sigma)/\omega]^{1/2}}{\sinh \pi[(\mu - \sigma)/\omega]^{1/2}} \leq C, \quad \mu \geq 0. \end{aligned}$$

If we now set

$$\psi_{n,\omega,\sigma,\mu,\rho} = [Q_\rho(\mu)^{-1}]^* \psi_{n,\omega,\sigma,\mu}, \quad n \geq 1, \quad \mu \geq 0,$$

then $\{\psi_{n,\omega,\sigma,\mu,\rho}; n \geq 0\}$ is a sequence biorthogonal to (4.14) in $L^2(0, \rho)$. By virtue of Lemma 3.1 it satisfies an estimate of the same form as (4.15). Now let

$$p(t) = a_0 e^{-\sigma t} + \sum a_n e^{-(\omega n^2 + \mu)t}$$

be an arbitrary exponential polynomial in $E_T(\omega, \sigma, \mu)$. Since

$$a_n = \langle p, \psi_{n,\omega,\sigma,\mu,\rho} \rangle,$$

we have

$$|a_0| \leq C \|p\|_{(0,T)}, \quad |a_n| \leq C \|p\|_{(0,T)} e^{\pi(n^2 + 2\mu/\omega)^{1/2}}, \quad n \geq 1.$$

Accordingly,

$$\begin{aligned} |p(t)| &\leq C \|p\|_{(0,T)} e^{-\sigma t} \left\{ 1 + \sum_{n=1}^{\infty} \exp [\pi(n^2 + 2\mu/\omega)^{1/2} - (\omega n^2 + \mu - \sigma)t] \right\} \\ &\leq C' \|p\|_{(0,T)} e^{-\sigma t}, \quad t \geq T \geq \rho. \end{aligned}$$

Squaring and integrating,

$$\|p\|_{(T,\infty)}^2 \leq C'' \|p\|_{(0,T)}^2 e^{-2\sigma T}.$$

Then,

$$\begin{aligned} \|p\|_{(0,\infty)} &\leq (1 + C'' e^{-2\sigma T})^{1/2} \|p\|_{(0,T)} \\ &\leq (1 + \tfrac{1}{2} C'' e^{-2\sigma T}) \|p\|_{(0,T)}, \end{aligned}$$

which ends the proof of Lemma 4.4.

Since (4.13) implies that $K(T) \rightarrow 1$ as $T \rightarrow \infty$, we see from (4.12) and the comments preceding it that the null reachability problem has a solution f subject to the constraint (4.11) for sufficiently large T if

$$p_0(u_T) \leq \theta/Q.$$

The null controllability problem ($u_T = 0$) with the constraint (4.11) has a solution for sufficiently large T for any u_0 that satisfies the conditions in Theorem 4.1. This is an immediate consequence of (4.4) and of the easily verifiable fact that $p_T(u_0) \rightarrow 0$ as $T \rightarrow \infty$.

Remark 1. The restriction that u_0 should vanish in B in Theorem 4.1 and subsequent results can be eliminated as follows (we owe this observation to D. L. Russell). Let u_0 be merely continuous in P , and let f be continuous in $B \times [0, T]$, zero in (say) $B \times [T/2, T]$ and such that

$$u_0(x) = f(x, 0), \quad x \in B.$$

Denote by \tilde{u} the solution of (1.1) with

$$\begin{aligned} \tilde{u}(x, 0) &= u_0(x), \quad x \in P, \\ \tilde{u}(x, t) &= \tilde{f}(x, t), \quad x \in B, \quad 0 \leq t \leq T. \end{aligned}$$

It is not difficult to see (due to the fact that \tilde{u} satisfies the homogeneous equation (1.1) for $t \geq T/2$) that $\tilde{u}(\cdot, T)$ satisfies the conditions for a final state in Theorem 4.1. Accordingly, if u_T also satisfies these conditions we can find, making use of Theorem 4.1, a solution of (1.1) with

$$\begin{aligned} u(x, 0) &= 0, \quad x \in P, \\ u(x, T) &= u_T(x) - \tilde{u}(x, T), \quad x \in P. \end{aligned}$$

It is easy to see that

$$u(x, t) + \tilde{u}(x, t)$$

is a solution of the controllability problem with u_0 as initial state and u_T as final state.

Remark 2. It can be easily seen that the constants in inequality (4.5)—and therefore those in Theorem 4.3—are best possible in the following sense.

LEMMA 4.5. Assume the controllability problem has a solution f vanishing outside of B_0 for $u_0 = 0$ and any u_T whose Fourier coefficients satisfy

$$(4.16) \quad |v_\alpha| \leq M \exp \{ -C(\alpha_1, \alpha_2, \dots, \alpha_n) \}$$

for some constant M , where the $C(\alpha_1, \alpha_2, \dots, \alpha_n)$ are positive numbers. Then

$$(4.17) \quad \begin{aligned} & C(\alpha_1, \alpha_2, \dots, \alpha_n) + C_1 \log \alpha_n \\ & \geq \pi(2(X_n/X_1)^2 \alpha_1^2 + \dots + 2(X_n/X_{n-1})^2 \alpha_{n-1}^2 + \alpha_n^2)^{1/2} - C_2. \end{aligned}$$

Proof. Denote by E the space of all multiple sequences $\{v_\alpha\}$ that satisfy (4.16) endowed with the norm

$$\|\{v_\alpha\}\| = \sup_\alpha |v_\alpha| \exp C(\alpha_1, \alpha_2, \dots, \alpha_n).$$

Plainly E is a Banach space. If f is the solution of the controllability problem provided by the assumptions in Lemma 4.5 for a given u_T , then $g(x, t) = f(x, T - t)$ must satisfy (2.9), (2.10) where the v_α are the Fourier coefficients of u_T .

Denote by F the subspace of $L^2(B_0 \times (0, T))$ consisting of all functions $g(x, t)$ such that (2.9), (2.10) hold with

$$\mu_\alpha = v_\alpha = 0 \quad \text{for all } \alpha.$$

Plainly F is closed. Next consider the map

$$P: E \rightarrow L^2(B_0 \times (0, T))/F$$

defined by

$$\{v_\alpha\} \rightarrow g + F,$$

g a solution of (2.9), (2.10) ($g + F$ indicates the equivalence class of g in the quotient space $L^2(B_0 \times (0, T))/F$). It is evident that P is well-defined in all of E . A moment's reflection shows that P is closed; then, as an application of the closed graph theorem we see that P is *bounded*. This implies the existence of a constant $C > 0$ such that, if $\{v_\alpha\} \in E$, there exists a solution $g \in L^2(B_0 \times (0, T))$ of (2.9), (2.10) such that

$$\|g\|_{L^2(B_0 \times (0, T))} \leq C \|\{v_\alpha\}\|$$

(here and afterwards C indicates a constant that may not be the same for different inequalities). Let $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_n)$ be another n -tuple of positive integers. Define

$$v_x^{(\gamma)} = \begin{cases} B\kappa\gamma_n & \text{if } \alpha_1 = \gamma_1, \dots, \alpha_n = \gamma_n, \\ 0 & \text{otherwise.} \end{cases}$$

Plainly, for each fixed γ , $\{v_\alpha^{(\gamma)}\}$ belongs to E and

$$\|\{v_\alpha^{(\gamma)}\}\| = B\kappa\gamma_n \exp C(\gamma_1, \gamma_2, \dots, \gamma_n).$$

Let $g^{(\gamma)}$ be the solution of (2.9), (2.10) corresponding to $\{\nu_\alpha^{(\gamma)}\}$ and let $g_\beta^{(\gamma)}$ be the functions obtained from $g^{(\gamma)}$ through (2.9). Obviously,

$$\|g_\beta^{(\gamma)}\|_{L^2(0,T)} \leq \|g\|_{L^2(B_0 \times (0,T))} \leq C\gamma_n \exp C(\gamma_1, \gamma_2, \dots, \gamma_n).$$

Now, it is not difficult to see that, if $\beta = (\alpha_1, \alpha_2, \dots, \alpha_{n-1}) = (\gamma_1, \gamma_2, \dots, \gamma_{n-1})$ and if we set $\gamma = (\beta, \gamma_n)$, then the sequence $\{g_\beta^{(\gamma)}\}_{\gamma_n}$ is biorthogonal to $\{e^{-\lambda_\alpha t}\}_{\alpha_n}$ in $L^2(0, T)$. If Q_T is the restriction operator of § 3 (we omit parameters for the sake of simplicity) then $\{Q_T^* g_\beta^{(\gamma)}\}_{\gamma_n}$ will provide a sequence biorthogonal to $\{e^{-\lambda_\alpha t}\}_{\alpha_n}$ in $L^2(0, \infty)$ and

$$\|Q_T^* g_\beta^{(\alpha)}\|_{L^2(0, \infty)} \leq C\alpha_n \exp(C\alpha_1, \alpha_2, \dots, \alpha_n).$$

On the other hand, it follows from the considerations at the beginning of § 3 about biorthogonal sequences with smallest possible norm and from (3.6) that we must necessarily have

$$\|Q_T g_\beta^{(\alpha)}\|_{L^2(0, \infty)} \geq C \exp \pi(2(X_n/X_1)^2 \alpha_1^2 + \dots + 2(X_n/X_{n-1})^2 \alpha_{n-1}^2 + \alpha_n^2)^{1/2}$$

from which (4.17) follows.

Remark 3. The results in the present paper can be generalized to bounded, cylindrical domains of the form $D \times [a, b]$, D a bounded domain in $(n-1)$ -dimensional space, the control being applied in, say, $D \times \{a\}$. One loses, however, the useful characterization of final states afforded by Theorem 4.3.

In a qualitative sense, application of control to all of the boundary of P (that is, to all of the faces that make it up) does not change the controllability results.

REFERENCES

- [1] L. BERS, F. JOHN AND M. SCHECHTER, *Partial Differential Equations*, Interscience, New York, 1964.
- [2] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, vol. I, Interscience, New York, 1953.
- [3] H. O. FATTORINI, *Boundary control systems*, this Journal, 6 (1968), pp. 349–385.
- [4] H. O. FATTORINI AND D. L. RUSSELL, *Exact controllability theorems for linear parabolic equations in one space dimension*, Arch. Rational Mech. Anal., 43 (1971), pp. 272–292.
- [5] ———, *Uniform bounds on biorthogonal functions for real exponentials with an application to the control theory of parabolic equations*, Quart. Appl. Math., to appear.
- [6] S. KACZMARZ AND H. STEINHAUS, *Theorie der Orthogonalreihen*, Monografie Matematyczne, tom VI, Warszawa-Lwów, 1935.
- [7] J. P. KAHANE, *Teoría constructiva de funciones*, Cursos y seminarios de matemática, fasc. 5, Universidad de Buenos Aires, Buenos Aires, Argentina, 1959.
- [8] P. D. LAX, *A Phragmén-Lindelöf principle in harmonic analysis and its applications to some questions in the theory of elliptic equations*, Comm. Pure Appl. Math., 10 (1957), pp. 361–389.
- [9] V. J. MIZEL AND T. I. SEIDMAN, *Observation and prediction for the heat equation II*, Rep. 70-35, Department of Mathematics, Carnegie-Mellon University, Pittsburgh, Pa., 1970.
- [10] L. SCHWARTZ, *Etude des sommes d'exponentielles*, 2^{me} édition, Hermann, Paris, 1959.
- [11] W. STERNBERG, *Über die Gleichung der Wärmeleitung*, Math. Ann., 101 (1929), pp. 394–398.

OBSERVABILITY AND RELATED PROBLEMS FOR PARTIAL DIFFERENTIAL EQUATIONS OF PARABOLIC TYPE*

YOSHIYUKI SAKAWA†

Abstract. Questions regarding observability on the basis of the observed measurement data from a finite number of sensors are discussed for the distributed-parameter systems described by linear partial differential equations of parabolic type. Necessary and sufficient conditions for observability are presented. We obtain a sufficient condition for continuous dependence of a state of the system upon the observation. Location of sensors for ensuring observability and continuity are discussed for several examples.

1. Introduction. Determining a state of a distributed-parameter system from observed measurement data is of fundamental importance when we stabilize and optimize the system by the feedback control [1]. It may appear that an infinite number of sensors along the spatial domain would be needed in order to measure spatially distributed physical quantities such as temperature distributions, neutron flux, and so on. However, such a realization of measurement system is impractical. Thus, the following questions will arise:

(i) Is it possible to determine a unique initial condition of a distributed-parameter system on the basis of the observed measurement data from a finite number of sensors? (observability problem).

(ii) Where should the measurement sensors be located and what is the minimum number of sensors in order to ensure observability?

(iii) Does the state at the time $t = T$ which we want to determine depend continuously upon the observed measurement data over the time interval $0 \leq t \leq T$? (continuity problem).

Questions (i) and (iii) correspond to the question of whether the state determination problem is well-posed in the sense of Hadamard. Goodson and Klein [6] and Yu and Seinfeld [12] respectively discussed the observability problem on several examples. Concerning the continuity problem, Mizel and Seidman [9] considered a heat equation in an n -dimensional ball and discussed the continuity of a mapping from the observation along the whole boundary of the ball to the state to be determined. Dolecki [4] considered a single point-sensor for a one-dimensional heat equation and discussed the continuity of the mapping from the observation to the state.

In this paper, we consider a heat equation in high-dimensional space with boundary conditions of general type and discuss both the observability problem by a finite number of sensors and the continuity of a mapping from the observation to the state to be determined. The results are then applied to several systems.

2. Preliminary results. Let D be a bounded domain of an r -dimensional Euclidean space, and let S , the boundary of D , consist of a finite number of $(r - 1)$ -

* Received by the editors May 21, 1973, and in revised form October 4, 1973.

† Department of Control Engineering, Faculty of Engineering Science, Osaka University, Toyonaka, Osaka, Japan.

dimensional hypersurfaces of class $C^{3,1}$. The spatial coordinate vector will be denoted by $x = (x_1, x_2, \dots, x_r) \in D$. Consider a linear parabolic partial differential equation

$$(1) \quad \frac{\partial u(t, x)}{\partial t} = \Delta u(t, x) - q(x)u(t, x),$$

where t is the time and Δ denotes the Laplacian given by

$$\Delta = \frac{\partial^2}{(\partial x_1)^2} + \dots + \frac{\partial^2}{(\partial x_r)^2}.$$

It is assumed that $q(x)$ is Hölder-continuous on the compact domain \bar{D} ($\bar{D} = D \cup S$, the upper bar denotes the closure).

The boundary condition is given by

$$(2) \quad \alpha(\xi)u(t, \xi) + (1 - \alpha(\xi))\frac{\partial u(t, \xi)}{\partial \nu} = 0,$$

where $\xi \in S$, ν is the exterior normal to the surface S at a point $\xi \in S$, and $\alpha(\xi)$ is a function of class C^2 on S satisfying

$$0 \leq \alpha(\xi) \leq 1.$$

The initial condition is given by

$$(3) \quad \lim_{t \rightarrow \infty} u(t, x) = u_0(x) \quad \text{in } L_2(D),$$

where $u_0(x) \in L_2(D)$, and $L_2(D)$ denotes the Hilbert space of all square integrable real-valued functions $u(x)$ with the inner product

$$\langle u_1, u_2 \rangle = \int_D u_1(x)u_2(x) dx.$$

It is shown by Ito [7] that there exists a *fundamental solution* $U(t, x, y)$ ($0 < t$; $x, y \in \bar{D}$) which is of class C^1 in t , of class C^2 in x and y in \bar{D} and that a unique solution to the initial-boundary value problem described by (1), (2) and (3) exists and is given by

$$(4) \quad u(t, x) = \int_D U(t, x, y)u_0(y) dy, \quad 0 < t < \infty, \quad x \in \bar{D}.$$

Furthermore, Ito [7] proved that there exists a sequence $\{\lambda_i, \phi_{ij}; j = 1, \dots, m_i, i = 1, 2, \dots\}$ of eigenvalues and eigenfunctions satisfying the following conditions:

$$(5) \quad (i) \quad C \leq \lambda_1 < \lambda_2 < \dots < \lambda_i < \dots, \quad \lim_{i \rightarrow \infty} \lambda_i = \infty,$$

where $C = \min_{x \in \bar{D}} q(x)$.

¹ In general, C^n denotes the set of all functions having n continuous derivatives.

(ii) $\{\phi_{ij}(x); j = 1, \dots, m_i, i = 1, 2, \dots\}$ is a complete orthonormal system in $L_2(D)$, where the positive integers m_i are finite for any $i < \infty$.

(iii) Each $\phi_{ij}(x)$ satisfies the following equations:

$$(6) \quad \int_D U(t, x, y) \phi_{ij}(y) dy = e^{-\lambda_i t} \phi_{ij}(x),$$

$$(7) \quad \Delta \phi_{ij}(x) - q(x) \phi_{ij}(x) = -\lambda_i \phi_{ij}(x),$$

and the boundary condition (2).

(iv) The fundamental solution is expressed as

$$(8) \quad U(t, x, y) = \sum_{i=1}^{\infty} e^{-\lambda_i t} \sum_{j=1}^{m_i} \phi_{ij}(x) \phi_{ij}(y),$$

where the series in the right side converges uniformly on $[\delta, \infty) \times \bar{D} \times \bar{D}$ for arbitrary $\delta > 0$.

(v) For arbitrary $h \in L_2(D)$ given by

$$h(x) = \sum_{i=1}^{\infty} \sum_{j=1}^{m_i} h_{ij} \phi_{ij}(x),$$

we have

$$(9) \quad (U_t h)(x) = \sum_{i=1}^{\infty} e^{-\lambda_i t} \sum_{j=1}^{m_i} h_{ij} \phi_{ij}(x),$$

where

$$(10) \quad (U_t h)(x) = \int_D U(t, x, y) h(y) dy,$$

and the series converges uniformly on $[\delta, \infty) \times \bar{D}$ for arbitrary $\delta > 0$.

Equations (1) and (2) can be written together as a differential equation in $L_2(D)$:

$$(11) \quad \frac{du(t)}{dt} = Au(t),$$

where the domain $\mathcal{D}(A)$ of the operator A is given by

$$\mathcal{D}(A) = \left\{ u : \Delta u \in L_2(D), \alpha(\xi)u(\xi) + (1 - \alpha(\xi))\frac{\partial u}{\partial \nu} = 0, \xi \in S \right\},$$

and

$$Au = \Delta u - q(x)u, \quad \text{if } u \in \mathcal{D}(A).$$

The positive integers m_i are called the multiplicity of the eigenvalues λ_i . If $\sup \{m_i\} = m < \infty$, we shall say that A has *multiplicity* m ; if $\sup \{m_i\} = \infty$, A is said to have infinite multiplicity. In the following, observability of the system described by (1) and (2) (or (11)) will be considered for the case where A has finite multiplicity.

3. Observability. Let us consider N sensors, whose outputs will be denoted by $y_k(t)$, $k = 1, \dots, N$. We consider two types of measurement such that the outputs of sensors are respectively given by

$$(12) \quad y_k(t) = \int_D w_k(x) u(t, x) dx, \quad k = 1, \dots, N,$$

and

$$(13) \quad y_k(t) = u(t, x^k), \quad k = 1, \dots, N,$$

where $w_k(x)$ are known functions of $L_2(D)$ and represent spatial weighting functions of sensors, and $x^k \in \bar{D}$ represent positions of sensors. The former type of measurement will be called a *type 1 measurement*, the outputs of which are spatial averages of a physical quantity over some effective sensing region. The latter will be called a *type 2 measurement*, which consists of ideal point-sensors.

In the case of the type 2 measurement, the eigenvalues and eigenfunctions are required to satisfy the following assumptions.

Assumption 1.
$$\sum_{i=2}^{\infty} \frac{1}{(\lambda_i - \lambda_1)^2} < \infty.$$

Assumption 2. The normalized eigenfunctions are uniformly bounded, i.e.,

$$|\phi_{ij}(x)| \leq M \quad \text{for all } i, j.$$

By the theory of asymptotic distribution of eigenvalues (see [2, VI, § 4]), it follows that

$$\lim_{n \rightarrow \infty} \frac{n^{2/r}}{\lambda_n} = \text{const.}$$

for an arbitrary r -dimensional spatial domain. Hence, it is clear that Assumption 1 holds if $r \leq 3$.

DEFINITION OF OBSERVABILITY [3]. The system described by (1) and (2) (or (11)) is said to be *observable* if an initial state $u_0(x)$ can be uniquely determined from the observation $Y(t) = (y_1(t), \dots, y_N(t))$, $0 \leq t < \infty$. In particular, the system is said to be *observable in time T* if an initial state can be uniquely determined from the observation $Y(t)$ over the time interval $0 \leq t \leq T$.

Let $L_2(0, T)$ denote the space of functions square integrable over $[0, T]$, and let $L_2^N(0, T)$ denote N products of $L_2(0, T)$. Since the solution $u(t) = U_t u_0$ of (11) (where the operator U_t from $L_2(D)$ into $L_2(D)$ is defined by (10)) is of class C^1 in time t , it follows that

$$Y(t) \in L_2^N(0, T),$$

for both types of measurement. If $u_0(x) \in L_2(D)$ is given, the observation $Y(t) \in L_2^N(0, T)$ is uniquely determined. Hence a linear mapping $P(T)$ from $L_2(D)$ into $L_2^N(0, T)$ is defined by

$$(14) \quad Y = P(T)u_0, \quad u_0 \in L_2(D), \quad Y \in L_2^N(0, T).$$

It is easily seen that the system is observable in time T if and only if

$$(15) \quad P(T)u_0 = 0 \quad \text{implies } u_0 = 0.$$

From (4) and (9) it follows that

$$(16) \quad u(t, x) = \sum_{i=1}^{\infty} e^{-\lambda_i t} \sum_{j=1}^{m_i} u_{ij} \phi_{ij}(x),$$

where $u_{ij} = \langle u_0, \phi_{ij} \rangle$. Therefore,

$$(17) \quad y_k(t) = \sum_{i=1}^{\infty} e^{-\lambda_i t} \sum_{j=1}^{m_i} u_{ij} w_{ij}^k, \quad k = 1, \dots, N,$$

where w_{ij}^k are defined by

$$(18) \quad w_{ij}^k = \langle w_k, \phi_{ij} \rangle \quad \text{for the type 1 measurement,}$$

$$(19) \quad w_{ij}^k = \phi_{ij}(x^k) \quad \text{for the type 2 measurement.}$$

Since $y_k(t)$ given by (17) are analytic in $t \in (0, \infty)$, if $y_k(t) \equiv 0$ over an arbitrary interval $0 < t \leq T$, then it follows that $y_k(t) \equiv 0$ for all $t > 0$. Consequently, if the system is observable, then the system is also observable in any time $T > 0$. In other words, the system is observable in any time $T > 0$ if and only if it is observable.

THEOREM 1. *Suppose that A has finite multiplicity m and that Assumptions 1 and 2 are satisfied for the type 2 measurement. Let us define $N \times m_i$ matrices W_i by*

$$(20) \quad W_i = \begin{bmatrix} w_{i1}^1 & w_{i2}^1 & \cdots & w_{im_i}^1 \\ w_{i1}^2 & w_{i2}^2 & \cdots & w_{im_i}^2 \\ \vdots & \vdots & & \vdots \\ w_{i1}^N & w_{i2}^N & \cdots & w_{im_i}^N \end{bmatrix},$$

where w_{ij}^k are defined by (18) or (19). The system described by (1) and (2) is observable in any finite time if and only if $N \geq m = \max \{m_i\}$ and

$$(21) \quad \text{rank } W_i = m_i \quad \text{for all } i = 1, 2, \dots.$$

Proof. To prove sufficiency, assume that

$$(22) \quad y_k(t) = \sum_{i=1}^{\infty} e^{-\lambda_i t} \sum_{j=1}^{m_i} u_{ij} w_{ij}^k \equiv 0, \quad t > 0, \quad k = 1, \dots, N.$$

For any complex number λ with $\text{Re } \lambda < \lambda_1$, we see from (22) that

$$\int_0^{\infty} \sum_{i=1}^{\infty} e^{-(\lambda_i - \lambda)t} dt \left(\sum_{j=1}^{m_i} u_{ij} w_{ij}^k \right) = 0.$$

If

$$(22') \quad \sum_{i=1}^{\infty} \int_0^{\infty} \left| e^{-(\lambda_i - \lambda)t} \sum_{j=1}^{m_i} u_{ij} w_{ij}^k \right| dt = \sum_{i=1}^{\infty} \frac{1}{\lambda_i - \text{Re } \lambda} \left| \sum_{j=1}^{m_i} u_{ij} w_{ij}^k \right| < \infty,$$

then by the theorem (see [13, Thm. 2.2.4, p. 64]) we see that

$$\begin{aligned}
 (23) \quad 0 &= \sum_{i=1}^{\infty} \int_0^{\infty} e^{-(\lambda_i - \lambda)t} dt \left(\sum_{j=1}^{m_i} u_{ij} w_{ij}^k \right) \\
 &= \sum_{i=1}^{\infty} \frac{1}{\lambda_i - \lambda} \left(\sum_{j=1}^{m_i} u_{ij} w_{ij}^k \right), \quad k = 1, \dots, N.
 \end{aligned}$$

For the type 1 measurement, it is easy to see that (22') holds. In fact, since $u_0(x)$ and $w_k(x)$ belong to $L_2(D)$, by use of the Schwarz inequality,

$$\begin{aligned}
 \sum_{i=1}^{\infty} \frac{1}{\lambda_i - \operatorname{Re} \lambda} \left| \sum_{j=1}^{m_i} u_{ij} w_{ij}^k \right| &\leq \frac{1}{\lambda_1 - \operatorname{Re} \lambda} \sum_{i=1}^{\infty} \sum_{j=1}^{m_i} |u_{ij}| |w_{ij}^k| \\
 &\leq \frac{1}{\lambda_1 - \operatorname{Re} \lambda} \left\{ \sum_{i=1}^{\infty} \sum_{j=1}^{m_i} u_{ij}^2 \right\}^{1/2} \left\{ \sum_{i=1}^{\infty} \sum_{j=1}^{m_i} (w_{ij}^k)^2 \right\}^{1/2} < \infty.
 \end{aligned}$$

For the type 2 measurement, by Assumptions 1 and 2, we obtain

$$\begin{aligned}
 &\sum_{i=1}^{\infty} \frac{1}{\lambda_i - \operatorname{Re} \lambda} \left| \sum_{j=1}^{m_i} u_{ij} \phi_{ij}(x^k) \right| \\
 &\leq \left[\sum_{i=1}^{\infty} (\lambda_i - \operatorname{Re} \lambda)^{-2} \right]^{1/2} \left[\sum_{i=1}^{\infty} \left\{ \sum_{j=1}^{m_i} u_{ij}^2 \right\} \left\{ \sum_{j=1}^{m_i} \phi_{ij}(x^k)^2 \right\} \right]^{1/2} \\
 &\leq \left[(\lambda_1 - \operatorname{Re} \lambda)^{-2} + \sum_{i=2}^{\infty} (\lambda_i - \lambda_1)^{-2} \right]^{1/2} \sqrt{m} M \left[\sum_{i=1}^{\infty} \sum_{j=1}^{m_i} u_{ij}^2 \right]^{1/2} < \infty.
 \end{aligned}$$

Thus, (22') holds in both cases.

By analytic continuation we see that (23) holds for all λ such that $\lambda \neq \lambda_i$, $i = 1, 2, \dots$. Let C_i be a circle in the complex plane of radius ε_i with λ_i as center, where ε_i is such that

$$0 < \varepsilon_i < \min(\lambda_i - \lambda_{i-1}, \lambda_{i+1} - \lambda_i).$$

Because of the relationship (22'), using the Cauchy's formula, we have

$$\begin{aligned}
 (24) \quad \sum_{j=1}^{m_i} u_{ij} w_{ij}^k &= \frac{1}{2\pi\sqrt{-1}} \int_{C_i} \sum_{i=1}^{\infty} \frac{d\lambda}{\lambda - \lambda_i} \left(\sum_{j=1}^{m_i} u_{ij} w_{ij}^k \right) = 0, \\
 &k = 1, \dots, N, \quad i = 1, 2, \dots.
 \end{aligned}$$

If $N \geq m_i$ and $\operatorname{rank} W_i = m_i$, $i = 1, 2, \dots$, then (24) implies $u_{i1} = \dots = u_{im_i} = 0$, $i = 1, 2, \dots$. Since

$$u_0(x) = \sum_{i=1}^{\infty} \sum_{j=1}^{m_i} u_{ij} \phi_{ij}(x),$$

we have $u_0(x) = 0$. This means that the system is observable.

To prove necessity, suppose that $\operatorname{rank} W_i < m_i$ for some i . Then there exists a nonzero m_i -vector $u_i = (u_{i1}, \dots, u_{im_i})$ satisfying (24). This implies that the equation $P(T)u_0 = 0$ has a nonzero solution. Thus the system is not observable.

It is obvious that (21) implies $N \geq m$. Q.E.D.

Goodson and Krein [6] obtained a similar result for the heat equation on a two-dimensional rectangle domain. Theorem 1 is an extension of the Goodson and Krein result to the general case.

COROLLARY 1. *Suppose that A has finite multiplicity m , and that the observation is given by (12) (type 1 measurement). Let a sequence $\{a_i\}$ of numbers be such that*

$$(25) \quad \sum_{i=1}^{\infty} a_i^2 < \infty, \quad a_i \neq 0, \quad i = 1, 2, \dots$$

If $N \geq m$ and the functions $w_k(x)$ are given by

$$(26) \quad w_k(x) = \sum_{i=1}^{\infty} a_i \phi_{ik}(x), \quad k = 1, \dots, N,$$

where $\phi_{ik}(x) = 0$ if $k > m_i$, then the system described by (1) and (2) is observable in any finite time T .

Proof. From the orthogonality of the eigenfunctions we see that

$$w_{ij}^k = \langle w_k, \phi_{ij} \rangle = a_i \delta_{jk},$$

where δ_{jk} is the Dirac δ -function. Therefore it follows that

$$(27) \quad W_i = a_i \begin{bmatrix} I_{m_i \times m_i} \\ 0_{(N-m_i) \times m_i} \end{bmatrix},$$

from which $\text{rank } W_i = m_i, i = 1, 2, \dots$. Q.E.D.

For the type 2 measurement, if we choose the location of sensors $x^k \in \bar{D}$, $k = 1, \dots, N$, so that $w_{ij}^k = \phi_{ij}(x^k)$ satisfy (21), then observability is ensured. When the multiplicity is one, we obtain the following.

COROLLARY 2. *Suppose that the multiplicity of A is 1 and that Assumptions 1 and 2 are satisfied. The system described by (1) and (2) is observable by the observation*

$$(28) \quad y(t) = u(t, x^1), \quad t > 0,$$

from a single sensor at $x^1 \in \bar{D}$ if and only if

$$(29) \quad \phi_i(x^1) \neq 0 \quad \text{for all } i = 1, 2, \dots$$

Remark 1. We have considered the system without any inputs. Now we consider the system described by

$$(30) \quad \begin{aligned} \frac{\partial u(t, x)}{\partial t} &= \Delta u(t, x) - q(x)u(t, x) + f(t, x), \quad x \in D, \\ \alpha(\xi)u(t, \xi) + (1 - \alpha(\xi))\frac{\partial u(t, \xi)}{\partial \nu} &= g(t, \xi), \quad \xi \in S, \end{aligned}$$

where both distributed input $f(t, x)$ and boundary input $g(t, \xi)$ are assumed to be known. Let us write the solution of (30) with an initial condition $u_0(x)$ as $u(t, x; f, g, u_0)$. Since it is known [7] that

$$(31) \quad u(t, x; f, g, u_0) = u(t, x; f, g, 0) + u(t, x; 0, 0, u_0),$$

we see that the same relation holds for the observation, i.e.,

$$(32) \quad Y(t; f, g, u_0) = Y(t; f, g, 0) + Y(t; 0, 0, u_0).$$

Therefore, $Y(t; 0, 0, u_0)$ can be obtained from the observed data $Y(t; f, g, u_0)$ subtracted by the calculated data $Y(t; f, g, 0)$. Thus the problem has been reduced to what we have considered.

Since Corollary 1 gives a general result for the type 1 measurement, we consider the observability problem in the case of type 2 measurement on several examples.

Example 1. We consider a one-dimensional heat equation :

$$(33) \quad \begin{aligned} \frac{\partial u(t, x)}{\partial t} &= \frac{\partial^2 u(t, x)}{\partial x^2}, & 0 < x < 1, \\ u(t, 0) &= u(t, 1) = 0. \end{aligned}$$

It is well known [2] that in this case the multiplicity is 1 and

$$(34) \quad \lambda_n = (n\pi)^2, \quad \phi_n(x) = \sin n\pi x, \quad n = 1, 2, \dots$$

It is clear that Assumptions 1 and 2 are satisfied. If x^1 ($0 < x^1 < 1$) is an arbitrary irrational number, then $\phi_n(x^1) \neq 0$, $n = 1, 2, \dots$. Thus from Corollary 2 the system described by (33) is observable by the observation $y(t) = u(t, x^1)$. For the other type of boundary conditions, the same result is obtained.

Example 2. We next consider a heat equation on a rectangle domain :

$$(35) \quad \begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2}; & 0 < x_1 < 1, \quad 0 < x_2 < \frac{1}{a}, \\ u(t, \xi) &= 0, \quad \xi \in S. \end{aligned}$$

The eigenvalues and eigenfunctions for this problem are given by [2]

$$(36) \quad \begin{aligned} \lambda_{nm} &= \pi^2(n^2 + a^2m^2), & n, m &= 1, 2, \dots, \\ \phi_{nm}(x) &= \sin(n\pi x_1) \sin(m\pi a x_2), & n, m &= 1, 2, \dots, \end{aligned}$$

for which Assumptions 1 and 2 are satisfied. Assume that a^2 is an irrational number. Then, since the relation $n_1^2 + a^2m_1^2 = n^2 + a^2m^2$ implies $n_1 = n$ and $m_1 = m$, the multiplicity is clearly 1. Thus from Corollary 2 the system described by (35) is observable by the observation $y(t) = u(t, x^1)$, where $x^1 = (x_1^1, x_2^1) \in D$ is an arbitrary point such that both x_1^1 ($0 < x_1^1 < 1$) and ax_2^1 ($0 < ax_2^1 < 1$) are irrational.

Remark 2. Since the Lebesgue measure of a set of all irrational numbers on the interval $(0, 1)$ is 1, a randomly chosen point corresponds to an irrational number with probability 1. Thus almost all points on the domains of Examples 1 and 2 are irrational.

Example 3. We consider a heat equation in the polar coordinate form defined on a unit circle :

$$(37) \quad \begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \theta^2}, \\ 0 &\leq r < 1, \quad 0 \leq \theta < 2\pi, \\ \left(\frac{\partial u}{\partial r} \right)_{r=1} &= 0. \end{aligned}$$

The eigenvalues for this problem are given by [2]

$$(38) \quad \lambda_{nm} = \beta_{nm}^2, \quad n = 0, 1, \dots, \quad m = 1, 2, \dots,$$

where β_{nm} are the real roots of the derivative of the Bessel function $J_n(\cdot)$ of n th order, i.e.,

$$(39) \quad J'_n(\beta_{nm}) = 0.$$

The eigenfunctions are given by [2]

$$(40) \quad \begin{aligned} \phi_{0m}(r, \theta) &= J_0(\beta_{0m}r)/\pi J_0^2(\beta_{0m}), \\ \phi_{nm1}(r, \theta) &= J_n(\beta_{nm}r)(\cos n\theta)/c_{nm}, \\ \phi_{nm2}(r, \theta) &= J_n(\beta_{nm}r)(\sin n\theta)/c_{nm}, \\ n, m &= 1, 2, \dots, \end{aligned}$$

where c_{nm} are constants for normalization. From the Bessel's integral (see [11, p. 19]), we see that $|J_n(x)| \leq 1$. Therefore, Assumptions 1 and 2 are satisfied. It is clear that the multiplicity in this case is 2. Therefore at least two sensors are necessary. Assume that two sensors are located on the boundary ($r = 1$), the angular positions of which are θ_1 and θ_2 , respectively.

Applying Theorem 1 to this problem, since

$$\begin{aligned} W_{0m} &= \begin{bmatrix} \phi_{0m}(1, \theta_1) \\ \phi_{0m}(1, \theta_2) \end{bmatrix} = J_0(\beta_{0m})/c_{0m} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \\ W_{nm} &= \begin{bmatrix} \phi_{nm1}(1, \theta_1) & \phi_{nm2}(1, \theta_1) \\ \phi_{nm1}(1, \theta_2) & \phi_{nm2}(1, \theta_2) \end{bmatrix}, \\ |W_{nm}| &= J_n^2(\beta_{nm})(\sin n(\theta_2 - \theta_1))/c_{nm}^2, \end{aligned}$$

we see that the system described by (37) is observable by the observations

$$y_k(t) = u(t, 1, \theta_k), \quad k = 1, 2,$$

from the two sensors if and only if

$$(41) \quad \sin n(\theta_2 - \theta_1) \neq 0, \quad n = 1, 2, \dots$$

The conditions of (41) are satisfied if and only if $(\theta_2 - \theta_1)/\pi$ is irrational.

4. Continuity problem. In this section we consider the following problem: When the system described by (1) and (2) is observable, does the state $u(T, x)$,

$x \in D$, at the time $t = T$ depend continuously on the observation $Y(t)$, $0 \leq t \leq T$? When the system is observable, it is clear from (15) that the inverse mapping $P(T)^{-1}$ of $P(T)$ exists. Therefore

$$(42) \quad u_0 = P(T)^{-1} Y.$$

Using the operator U_t defined by (10), we see that

$$(43) \quad u(T) = U_T u_0 = U_T P(T)^{-1} Y,$$

where $Y \in L_2^N(0, T)$, $u(T) \in L_2(D)$.

Now the problem is to investigate the continuity of a mapping $U_T P(T)^{-1}$ from $L_2^N(0, T)$ to $L_2(D)$, which is equivalent to the existence of a constant γ such that

$$(44) \quad \|u(T)\| \leq \gamma \|Y\|_{L_2^N(0, T)},$$

where $\|\cdot\|$ denotes the norm of $L_2(D)$ and $\|\cdot\|_{L_2^N(0, T)}$ denotes that of $L_2^N(0, T)$.

Since $\{\phi_{ij}(x)\}$ is an ortho-normal system in $L_2(D)$, it follows from (16) that

$$(45) \quad \|u(T, \cdot)\| \leq \sum_{i=1}^{\infty} e^{-\lambda_i T} (u_{i1}^2 + \cdots + u_{im_i}^2)^{1/2}.$$

Suppose that the system is observable by either type 1 measurement or type 2 measurement. Then from Theorem 1, $\text{rank } W_i = m_i$, $i = 1, 2, \dots$. Therefore, defining an m_i -vector $u_i = \text{col}(u_{i1}, \dots, u_{im_i})$, we see that

$$(46) \quad W_i u_i \neq 0, \quad \text{if } u_i \neq 0.$$

Thus, we can define a finite number B_i by

$$(47) \quad B_i = \frac{(u_i, u_i)}{(W_i u_i, W_i u_i)} = \frac{\sum_{j=1}^{m_i} u_{ij}^2}{\sum_{k=1}^N (\sum_{j=1}^{m_i} w_{ijk}^k u_{ij})^2} < \infty,$$

where w_{ij}^k are given by (18) or (19). If $u_i = 0$, we set $B_i = 0$.

To estimate B_i , we define an $m_i \times m_i$ symmetric matrix by

$$(48) \quad V_i = W_i' W_i,$$

where W_i' denotes the transpose of W_i , and we denote the minimum eigenvalue of V_i by μ_i . Then we see that

$$(49) \quad B_i = \frac{(u_i, u_i)}{(u_i, V_i u_i)} \leq \frac{1}{\mu_i}.$$

Let $\Lambda = \{\lambda_{ij}\}$ denote a sequence of numbers, $-\sigma < \lambda_1 < \lambda_2 < \cdots$, σ being a constant, and let $E(\Lambda, T)$ denote a closed subspace of $L_2(0, T)$ spanned by the functions

$$(50) \quad p_i(t) = e^{-\lambda_i t}, \quad 0 \leq t \leq T < \infty, \quad i = 1, 2, \dots.$$

Then it is shown [5] that there exists a biorthogonal sequence of functions $q_j(t) \in E(\Lambda, T) \subset L_2(0, T)$, $j = 1, 2, \dots$, such that

$$(51) \quad (p_i, q_j)_{L_2(0, T)} = \delta_{ij}.$$

Letting

$$(52) \quad a_i^k = u_{i1} w_{i1}^k + \cdots + u_{im_i} w_{im_i}^k,$$

we have from (17)

$$(53) \quad y_k(t) = \sum_{i=1}^{\infty} a_i^k e^{-\lambda_i t} \in E(\Lambda, T).$$

By the Schwarz inequality and (51), it follows that

$$(54) \quad (y_k, q_i)_{L_2(0,T)} = a_i^k \leq \|y_k\|_{L_2(0,T)} \|q_i\|_{L_2(0,T)}.$$

Therefore, we obtain

$$(55) \quad \|Y(t)\|_{L_2^N(0,T)} = \left\{ \sum_{k=1}^N \|y_k(t)\|_{L_2(0,T)}^2 \right\}^{1/2} \geq \frac{1}{\|q_i\|_{L_2(0,T)}} \left\{ \sum_{k=1}^N (a_i^k)^2 \right\}^{1/2}.$$

Using (47) and (49), we see that

$$(56) \quad \|q_i\|_{L_2(0,T)} \|Y\|_{L_2^N(0,T)} \geq \sqrt{\mu_i} (u_{i1}^2 + \cdots + u_{im_i}^2)^{1/2}.$$

From (45) and (56), we obtain

$$(57) \quad \|u(T, \cdot)\| \leq \left[\sum_{i=1}^{\infty} \frac{1}{\sqrt{\mu_i}} \|q_i\|_{L_2(0,T)} e^{-\lambda_i T} \right] \|Y\|_{L_2^N(0,T)}.$$

Thus, if the infinite series in the right side of (57) converges, then the mapping $U_T P(T)^{-1}$ is bounded.

For studying the convergence of the series, an estimation of the norm of q_i is needed. For that purpose, we can make use of the following theorem which was obtained by Fattorini and Russell [5].

THEOREM 2 (Fattorini and Russell). *Let a function F be convex and strictly increasing on $[0, \infty)$, with*

$$(58) \quad F(0) = 0, \quad \int_1^{\infty} \frac{d\lambda}{F(\lambda)} < \infty,$$

and let G be the inverse of F , i.e., $G(F(\lambda)) = \lambda$, $0 \leq \lambda < \infty$. Assume that for some constant $C > 0$,

$$(59) \quad G(\lambda_1 \lambda_2) \leq C G(\lambda_1) G(\lambda_2), \quad 0 \leq \lambda_1, \lambda_2 < \infty.$$

Finally, let a sequence of numbers $\Lambda = \{\lambda_n\}$ be such that

$$(60) \quad G(\lambda_1 + \sigma) \geq \rho,$$

$$(61) \quad G(\lambda_{n+1} + \sigma) - G(\lambda_n + \sigma) \geq \rho, \quad \lambda_n + \sigma > 0, \quad n = 1, 2, \dots,$$

for some nonnegative constant σ and positive constant ρ . Then there exist constants B and K such that

$$(62) \quad \|q_n\|_{L_2(0,T)} \leq B \exp \{K G(\lambda_n + \sigma)\}, \quad n = 1, 2, \dots,$$

where $\{q_n\}$ is the biorthogonal sequence for $\{p_n\}$ ($p_n(t) = e^{-\lambda_n t}$).

Remark 3. It is clear that the function $F(\lambda)$ defined by

$$(63) \quad F(\lambda) = \lambda^\alpha, \quad \alpha > 1,$$

satisfies the conditions of Theorem 2.

From (57) and Theorem 2, we obtain the following theorem.

THEOREM 3. *Suppose that the system is observable and that the sequence of eigenvalues $\{\lambda_n\}$ satisfies the conditions (60) and (61). If*

$$(64) \quad \sum_{n=1}^{\infty} \frac{1}{\sqrt{\mu_n}} \exp[-\lambda_n T + KG(\lambda_n + \sigma)] < \infty,$$

then the mapping $U_T P(T)^{-1}$ from $L_2^N(0, T)$ to $L_2(D)$ is bounded.

In the case of type 1 measurement, where the functions $w_k(x)$, $k = 1, \dots, N$, are given by (26) and $\{a_i\}$ satisfies (25), the matrices V_i are given by

$$V_i = a_i^2 I_{m_i \times m_i}.$$

Therefore,

$$(65) \quad \sqrt{\mu_n} = |a_n|.$$

Thus, we obtain a general result for the type 1 measurement.

COROLLARY 3. *Suppose that the system is observable and that the sequence of eigenvalues $\{\lambda_n\}$ satisfies the conditions (60) and (61). Let us choose a sequence $\{a_n\}$ so that it satisfies*

$$(66) \quad \sum_{n=1}^{\infty} \frac{1}{|a_n|} \exp[-\lambda_n T + KG(\lambda_n + \sigma)] < \infty$$

and (25), and let us construct the functions w_k , $k = 1, \dots, N$, by (26). Then, the mapping from the type 1 measurement data $Y(t)$, $0 \leq t \leq T$, to the state $u(T, x)$, $x \in D$, is continuous and bounded.

In what follows, we study condition (64) for the type 2 measurement on Example 1.

Example 1. Let $F(\lambda) = \lambda^2$ for this example. Then $G(\lambda) = \sqrt{\lambda}$. Letting $\sigma = 0$, $\rho = \pi$, we see that the sequence of eigenvalues $\{(n\pi)^2\}$ satisfies (60) and (61). Thus we can apply Theorem 3 to this example. It is easily seen that for the type 2 measurement,

$$\sqrt{\mu_n} = |\phi_n(x^1)| = |\sin(n\pi x^1)| > 0$$

by the observability condition.

If α is a real number, we may write $\|\alpha\|$ for the distance between α and the nearest integer. It is easily seen that

$$(67) \quad |\sin(nx^1)| \geq 2\|nx^1\|.$$

A real number α for which there exists a constant c such that

$$(68) \quad \|\alpha\| > c/n \quad \text{for all integers } n > 0$$

is called the number of constant type [8]. Let L be a positive integer which is not a square, and let

$$(69) \quad \alpha = a + b\sqrt{L},$$

where a and b are rational numbers. It is shown [8] that the real number α given by (69) is of constant type.

Now let the irrational number x^1 ($0 < x^1 < 1$) be of constant type. Then, using (67) and (68), we obtain

$$(70) \quad \frac{1}{\sqrt{\mu_n}} \exp[-\lambda_n T + KG(\lambda_n + \sigma)] \leq \frac{1}{2\|nx^1\|} \exp[-\lambda_n T + K\sqrt{\lambda_n}] < \frac{1}{2c} n \exp\left[-T\left(\sqrt{\lambda_n} - \frac{K}{2T}\right)^2 + \frac{K^2}{4T}\right].$$

From (70) it is easily seen that (64) holds in this case.

Dolecki [4] proved for this particular example that for almost all points x^1 ($0 < x^1 < 1$) (64) holds.

Remark 4. Numbers of constant type are generated as roots of quadratic equations.

Remark 5. For Examples 2 and 3, since

$$\lim_{n \rightarrow \infty} \frac{n}{\lambda_n} = \text{const.}$$

(see [2, VI, § 4]), it is not certain that the eigenvalues satisfy the conditions of Theorem 2.

5. Concluding remarks. The observability problem on the basis of the two types of measurement data and the continuity problem of a mapping from the measurement data to the state have been discussed for distributed-parameter systems described by partial differential equations of parabolic type. It has been shown that the multiplicity of the operator, which depends on the shape of domain, boundary conditions, etc., is of importance. When the multiplicity is finite, at least the same number of sensors is necessary for observability.

To calculate the state $u(T, x)$, we have to eliminate $u_0(x)$ from the integral equations

$$(71) \quad \begin{aligned} u(T, x) &= \int_D U(T, x, y) u_0(y) dy, \\ y_k(t) &= \int_D L_k(t, x) u_0(x) dx, \quad k = 1, \dots, N, \end{aligned}$$

where $y_k(t)$ are the observations, and $L_k(t, x)$ are known functions given by

$$L_k(t, x) = \int_D U(t, y, x) w_k(y) dy$$

for the type 1 measurement, and

$$L_k(t, x) = U(t, x^k, x)$$

for the type 2 measurement. If the observability condition is satisfied and the mapping is continuous, then the problem of solving (71) is well-posed. In that case we can solve (71) approximately by applying various numerical methods (e.g., by discretization or by a variational method). Establishing an efficient algorithm for solving (71) is left for future study.

REFERENCES

- [1] M. ATHANS, *Some remarks on the control of distributed parameter systems*, Control of Distributed Parameter Systems, ASME, New York, 1963, pp. 1–12.
- [2] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, vol. 1, Interscience, New York, 1966.
- [3] M. C. DELFOUR AND S. K. MITTER, *Controllability and observability for infinite-dimensional systems*, this Journal, 10 (1972), pp. 329–333.
- [4] S. DOLECKI, *Observation for the one-dimensional heat equation*, Rep. 42, Institute of Mathematics, Polish Academy of Sciences, Warsaw, 1972.
- [5] H. O. FATTORINI AND D. L. RUSSELL, *Uniform bounds on biorthogonal functions for real exponentials with an application to the control theory of parabolic equations*, Quart. Appl. Math., to appear.
- [6] R. E. GOODSON AND R. E. KLEIN, *A definition and some results for distributed system observability*, IEEE Trans. Automatic Control, AC-15 (1970), pp. 165–174.
- [7] S. ITO, *Partial Differential Equations*, Baifukan, Tokyo, 1966. (In Japanese.)
- [8] S. LANG, *Introduction to Diophantine Approximations*, Addison-Wesley, Reading, Mass., 1966.
- [9] V. J. MIZEL AND T. I. SEIDMAN, *Observation and prediction for the heat equation, II*, J. Math. Anal. Appl., 38 (1972), pp. 149–166.
- [10] F. W. J. OLVER, ed., *Bessel Functions, Part III, Zeros and Associated Values*, Cambridge Univ. Press, Cambridge, 1960.
- [11] G. N. WATSON, *Theory of Bessel Functions*, Cambridge Univ. Press, Cambridge, 1966.
- [12] T. K. YU AND J. H. SEINFELD, *Observability of a class of hyperbolic distributed parameter systems*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 495–496.
- [13] E. ASPLUND AND L. BUNGART, *A First Course in Integration*, Holt, Reinhart and Winston, New York, 1966.

GENERALIZED CURVES AND EXTREMAL POINTS*

J. E. RUBIO†

Abstract. A nonparametric variational problem is considered in the setting of the theory of generalized curves. Instead of minimizing a functional dependent on a curve joining two given points, a functional defined on a set of Radon measures is considered; the set of measures is determined by the boundary conditions. It is shown that this functional attains its minimum at an extremal point of the set of measures. Further, an approximation scheme is developed so that the solution of the variational problem can be effected by solving a sequence of finite-dimensional programming problems; it is possible then to construct a sequence of curves such that the functional takes along this sequence values approaching its minimum over the set of measures. It is shown that this minimum is attained at an (extremal) measure which is a generalized curve, that is, the weak* limit of a sequence of curves. This generalized curve is characterized in terms of the minimizing elements of the sequence of discrete programming problems, and conditions on the original problem are obtained so that the generalized curve is actually an ordinary curve.

1. Introduction. This paper presents a method of study of simple nonparametric variational problems based upon the theory of extremal points of certain sets of Radon measures. The problems are considered in the setting provided by the theory of generalized curves of L. C. Young [1]–[4].

For extensions of this theory to optimal control problems, we refer to [3], [5]–[8], [12]. The importance of these developments is due chiefly to the solution of existence problems achieved by the introduction of curve-like elements, known as generalized curves, and control-like elements, relaxed or generalized controls.

Much work has been done on the characterization of the minimizing elements by means of necessary conditions [6]–[13]. After the achievements of the existence theory, one cannot help being somewhat disappointed by these necessary conditions. The development of Euler-type necessary conditions [9]–[11] for the variational problems, or Pontryagin-type conditions [6], [8], [11]–[13] for the optimal control problems, have not been entirely successful as means of characterization of the minimizing elements, since these conditions, just like their classical counterparts, are not constructive. Even worse, the optimal control problem in the generalized setting is singular [14], [15], so that no information is gained on important features of the optimal control by means of a Pontryagin-type approach. Second order conditions, derived to gain further information on these features, do not seem to be very helpful [14], [15].

The objective of the present paper is to present a constructive characterization of the minimizing generalized curve for a nonparametric variational problem. We shall make use of some abstract concepts associated with the theory of extremal points; functional-analytical methods, used to advantage to solve the existence problem, also provide means to construct the minimizing elements. The

* Received by the editors July 18, 1972, and in revised form September 18, 1973.

† Department of Applied Mathematical Studies, School of Mathematics, University of Leeds, Leeds LS2 9JT, England.

end result of our work is a rigorous scheme for discretizing the variational problem; the minimizing generalized curve can be characterized by means of the minimizing elements in a series of discrete mathematical programming problems.

2. Basic considerations. Let x, \dot{x} be vectors in Euclidean n -space R^n with components $x^i, \dot{x}^i, i = 1, \dots, n$; t is a real variable. Consider the continuous function $f_0: \Omega \rightarrow R$, with $\Omega = [t_0, t_f] \times A \times B$, $t_0 < t_f$, $A = \{x: |x^i| \leq a_i, i = 1, \dots, n\}$ for some positive constants a_i , and $B = \{\dot{x}: |\dot{x}| \leq \beta\}$, a closed ball in R^n . Let x_0, x_f be points in A and consider the class \mathcal{F} of absolutely continuous functions $x(t)$, $t_0 \leq t \leq t_f$, such that $x(t) \in A, t \in [t_0, t_f]$, $\dot{x}(t) \in B$ a.e. on $[t_0, t_f]$, $x(t_0) = x_0, x(t_f) = x_f$. We assume that $\beta|t_f - t_0| \geq |x_f - x_0|$, since the class \mathcal{F} is nonempty if and only if this condition is satisfied. Define a functional $J: \mathcal{F} \rightarrow R$,

$$(2.1) \quad J(x(\cdot)) = \int_{t_0}^{t_f} f_0(t, x(t), \dot{x}(t)) dt$$

for $x(\cdot) \in \mathcal{F}$. We consider the problem of minimizing J over \mathcal{F} . Following L. C. Young [3], we recognize the fact that, for fixed $x(\cdot)$, the integral in (2.1) defines a linear, bounded, positive functional mapping $\mathcal{C}(\Omega)$, the space of continuous real-valued functions defined on Ω with the topology of uniform convergence, into the real numbers. There is therefore, by the Riesz representation theorem, a Radon (regular Borel) measure on Ω such that

$$\int_{t_0}^{t_f} f(t, x(t), \dot{x}(t)) dt = \int_{\Omega} f d\mu$$

for all $f \in \mathcal{C}(\Omega)$. This is a positive measure such that, since

$$\left| \int_{t_0}^{t_f} f(t, x(t), \dot{x}(t)) dt \right| \leq (t_f - t_0) \max_{\Omega} |f(t, x, \dot{x})|,$$

then

$$(2.2) \quad \int_{\Omega} d\mu \leq t_f - t_0 \equiv \Delta t.$$

Let $\phi: [t_0, t_f] \times A \rightarrow R$ be any continuously differentiable function. Then, along any curve in \mathcal{F} ,

$$\frac{d}{dt} \phi(t, x(t)) = \dot{x}(t) \text{ grad } \phi(t, x(t)) + \frac{\partial \phi}{\partial t}(t, x(t))$$

a.e. on $[t_0, t_f]$, so that

$$(2.3) \quad \int_{t_0}^{t_f} \left[\dot{x}(t) \text{ grad } \phi(t, x(t)) + \frac{\partial \phi}{\partial t}(t, x(t)) \right] dt = \phi(t_f, x_f) - \phi(t_0, x_0) \equiv \Delta \phi.$$

Further, if an absolutely continuous curve with elements $(t, x, \dot{x}) \in \Omega$ satisfies this relationship for all continuously differentiable functions ϕ , then it joins (t_0, x_0) to (t_f, x_f) , that is, it belongs to the class \mathcal{F} . Indeed, assume that the curve does satisfy (2.3) for all continuously differentiable ϕ , but that it does not go through (t_0, x_0) . Define a function ϕ equal to one at (t_0, x_0) and to zero outside

a neighborhood of (t_0, x_0) chosen so that the points $(t, x(t))$, $t \in [t_0, t_f]$, and the point (t_f, x_f) are not in this neighborhood. Then the left side of (2.3) is zero, while the right side equals -1 , a contradiction. We show in a similar way that the curve goes through (t_f, x_f) .

The basic idea of the theory of generalized curves consists in replacing the original problem as stated above by another problem in which we seek to minimize $\int_{\Omega} f_0 d\mu$ over the set of all positive Radon measures on Ω satisfying (2.2) and

$$(2.4) \quad \int_{\Omega} \left[\dot{x} \operatorname{grad} \phi(t, x) + \frac{\partial \phi}{\partial t}(t, x) \right] d\mu = \Delta \phi$$

for all continuously differentiable $\phi: [t_0, t_f] \times A \rightarrow R$. There are advantages in taking this step; the existence of a minimizing measure is automatically ensured, and a constructive framework can be erected so that it is possible to find a sequence of functions so that the corresponding values of the functional (2.1) approach its infimum on the class \mathcal{F} .

We derive some properties of the set of all positive Radon measures on Ω satisfying (2.2) and (2.4). This set is nonempty since the measures corresponding to the functions in \mathcal{F} are in it. The zero measure is not in this set, since it does not satisfy (2.4) because $t_0 \neq t_f$, and then $\Delta \phi = \phi(t_f, x_f) - \phi(t_0, x_0)$ is not zero for all ϕ . We topologize the linear space of all Radon measures on Ω by the weak* topology.

PROPOSITION 1. *The set of all positive measures on Ω satisfying (2.2) and (2.4) has nonzero extremal points.*

Proof. Let M be the set of positive measures satisfying (2.4), S the closed ball of measures satisfying (2.2), $Q = M \cap S$. Of course, S is compact in the weak* topology. Let μ_k , $k = 1, 2, \dots$, be in M , and let the sequence $\{\mu_k\}$ converge in the weak* topology to a measure μ . Then μ is in M ; since the set of positive Radon measures on Ω is metrizable, we conclude that M is therefore closed; $Q = M \cap S$ is compact. Since M and S are convex, Q is convex and has, by the Krein–Milman theorem, nonzero extremal points.

Consider now the functional $I: Q \rightarrow R$, defined by

$$(2.5) \quad I(\mu) = \int_{\Omega} f_0 d\mu, \quad \mu \in Q.$$

This functional is the restriction to Q of a continuous linear functional $\mu \rightarrow \int_{\Omega} f_0 d\mu$ defined for all Radon measures on Ω . Therefore I attains its minimum over Q at one or some of the extremal points of this set. Putting $\phi(t, x) = t$, $t \in [t_0, t_f]$, $x \in A$, it follows that all $\mu \in M$ satisfy $\int_{\Omega} d\mu = \Delta t$. We have shown the following.

THEOREM 1. *The functional $I: Q \rightarrow R$ defined by (2.5) attains its minimum over Q , at one or some of the extremal points of this set. These extremal points satisfy (together with all measures in M),*

$$(2.6) \quad \int_{\Omega} d\mu = \Delta t.$$

In the following sections of this paper, we construct a sequence of curves at which the functional J in (2.1) attains values arbitrarily close to the minimum of

I over Q . We shall also characterize those extremal points of Q at which I attains its minimum as *generalized curves*, that is, weak* limits of curves.

3. Approximation. Let P be the set of positive Radon measures on Ω satisfying (2.4) and (2.6). Consider a sequence $\{\phi_k, k = 1, 2, \dots\}$ of continuously differentiable functions mapping $[t_0, t_f] \times A \rightarrow R$. Define the sets of Radon measures

$$(3.1) \quad P_v = \left\{ \mu : \mu > 0, \int_{\Omega} d\mu = \Delta t, \int_{\Omega} \left[\dot{x} \operatorname{grad} \phi_k + \frac{\partial \phi_k}{\partial t} \right] d\mu = \Delta \phi_k, \right. \\ \left. k = 1, 2, \dots, v \right\}, \quad v = 1, 2, \dots$$

Then

$$P_1 \supset P_2 \supset \dots \supset P_v \supset \dots \supset P.$$

Further, let $\mu_v \in P_v$ be a measure such that

$$\int_{\Omega} f_0 d\mu_v \leq \int_{\Omega} f_0 d\mu, \quad \mu \in P_v.$$

We can show, by arguments similar to those used in §2, that μ_v exists and is an extremal point of P_v . Then

$$I(\mu_1) \leq I(\mu_2) \leq \dots \leq I(\mu_v) \leq \dots \leq I^*,$$

where I^* is the minimum of I over P . The sequence of real numbers $\{I(\mu_v)\}$ is nondecreasing and bounded above; it converges to a value $\tilde{I} \leq I^*$. We show that, for a special choice of the sequence $\{\phi_k\}$, $\tilde{I} = I^*$.

THEOREM 2. *Let the sequence $\{\phi_k\}$ consist of all monomials in t and the n components of the vector x , in any order. Then $\tilde{I} = I^*$.*

Proof. Let $\bar{P} = \lim P_v = \limsup P_v = \liminf P_v = \bigcap_{v=1}^{\infty} P_v$. Then $\bar{P} \supseteq P$ and

$$\tilde{I} = \lim_{v \rightarrow \infty} I(\mu_v) = \min_{\mu \in \bar{P}} I(\mu).$$

We show that, if the sequence $\{\phi_k\}$ is that of all monomials in t and the components of x , then $P \supseteq \bar{P}$; that is, if

$$(3.2) \quad \int_{\Omega} \left[\dot{x} \operatorname{grad} \phi_k + \frac{\partial \phi_k}{\partial t} \right] d\mu = \Delta \phi_k, \quad k = 1, 2, \dots,$$

then

$$(3.3) \quad \int_{\Omega} \left[\dot{x} \operatorname{grad} \phi + \frac{\partial \phi}{\partial t} \right] d\mu = \Delta \phi$$

for all continuously differentiable functions ϕ mapping $[t_0, t_f] \times A \rightarrow R$. We write $\|\phi\|$ for the sup norm in the space of these functions.

If a positive measure μ satisfying (2.6) satisfies (3.2), then it satisfies (3.3) for all polynomials in t and the components of x , by linearity. First let ϕ be $(n+1)$ -times continuously differentiable. Then the derivative $\phi_{tx^1 \dots x^n}$ exists and is continuous. Let (t, \underline{x}) be an interior point of $[t_0, t_f] \times A$. Then

$$\phi(t, x) = \int_t^t \int_{\underline{x}^1}^{x^1} \dots \int_{\underline{x}^n}^{x^n} \phi_{tx^1 \dots x^n}(\tau, y) d\tau dy + \phi(t, \underline{x}).$$

Since $\phi_{tx^1 \dots x^n}$ is continuous, given $\varepsilon > 0$ a polynomial ψ in t and the components of x exist such that $\|\phi_{tx^1 \dots x^n} - \psi\| \leq \varepsilon$. Define a function $\hat{\psi}$ by

$$\hat{\psi}(t, x) = \int_t^t \int_{x^1}^{x^1} \cdots \int_{x^n}^{x^n} \psi(\tau, y) d\tau dy + \phi(t, x)$$

for all $(t, x) \in [t_0, t_f] \times A$. Then,

$$\|\phi - \hat{\psi}\| \leq \varepsilon \sup \left| \int_t^t \int_{x^1}^{x^1} \cdots \int_{x^n}^{x^n} d\tau dy \right| = \alpha_1 \varepsilon,$$

with α_1 depending only on the set $[t_0, t_f] \times A$. Since

$$\frac{\partial \phi}{\partial t}(t, x) = \int_{x^1}^{x^1} \cdots \int_{x^n}^{x^n} \phi_{tx^1 \dots x^n}(\tau, y) d\tau dy$$

with a similar expression holding for $\partial \hat{\psi}(t, x)/\partial t$,

$$\left\| \frac{\partial \phi}{\partial t} - \frac{\partial \hat{\psi}}{\partial t} \right\| \leq \alpha_2 \varepsilon$$

and

$$\left\| \frac{\partial \phi}{\partial x^i} - \frac{\partial \hat{\psi}}{\partial x^i} \right\| \leq \theta_i \varepsilon, \quad i = 1, 2, \dots, n,$$

with α_2 and $\theta_i, i = 1, 2, \dots, n$, depending only on the set $[t_0, t_f] \times A$. Also, since $\hat{\psi}$ is a polynomial, it satisfies (3.3); then,

$$\begin{aligned} & \left| \int_{\Omega} \left[\dot{x} \operatorname{grad} \phi + \frac{\partial \phi}{\partial t} - \Delta \phi \right] d\mu \right| \\ & \leq \left| \int_{\Omega} \left[\dot{x} \operatorname{grad} (\phi - \hat{\psi}) + \frac{\partial \phi}{\partial t} - \frac{\partial \hat{\psi}}{\partial t} \right] d\mu \right| + |\Delta \phi - \Delta \hat{\psi}| \Delta t \leq \alpha_3 \varepsilon \end{aligned}$$

for some α_3 depending on Ω and for all ε . Thus $\int_{\Omega} [\dot{x} \operatorname{grad} \phi + \partial \phi / \partial t] d\mu = \Delta \phi$; if μ satisfies (3.2) it satisfies (3.3) for all $(n+1)$ -times continuously differentiable functions ϕ . Moreover, this implies that it satisfies (3.3) for all continuously differentiable ϕ ; indeed, given such a ϕ , one can choose $\hat{\phi}$, $(n+1)$ -times continuously differentiable, so that $\dot{x} \operatorname{grad} \phi + \partial \phi / \partial t$ and $\dot{x} \operatorname{grad} \hat{\phi} + \partial \hat{\phi} / \partial t$ are uniformly close (Treves [16, Cor. 1, pp. 157–158]). The rest of the proof is as above.

It follows that, if the functions ϕ_k are the monomials in t and the components of x , then $\bar{P} = P$ and then $\bar{I} = I^*$.

For sufficiently high v , the measure μ_v provides a good approximation to the minimum of the functional I over P . It happens that these measures $\mu_v, v = 1, 2, \dots$, can be characterized quite simply; μ_v is an extremal point of the set P_v , composed of those positive measures which satisfy (2.6) and (3.2) for $k = 1, 2, \dots, v$. A theorem of P. C. Rosenbloom [17, Thm. 38, p. 193] shows that these

extremal points are linear positive combinations of (at most) $v + 1$ atomic measures on Ω . Therefore,

$$(3.4) \quad \int_{\Omega} f_0 d\mu_v = \sum_{i=1}^{v+1} \alpha_i f_0(t_i, x_i, \dot{x}_i),$$

for some constants $\alpha_i \geq 0$, $\sum_{i=1}^{v+1} \alpha_i = \Delta t$, $t_i \in [t_0, t_f]$, $x_i \in A$, $\dot{x}_i \in B$. The constants $\alpha_i, t_i, x_i, \dot{x}_i, i = 1, 2, \dots, v, v + 1$, satisfy the "boundary" conditions:

$$(3.5) \quad \sum_{i=1}^{v+1} \alpha_i \left[\dot{x}_i \operatorname{grad} \phi_k(t_i, x_i) + \frac{\partial \phi_k}{\partial t}(t_i, x_i) \right] = \Delta \phi_k, \quad k = 1, 2, \dots, v.$$

Of course, μ_v is not just any extremal point of P_v but rather one at which the minimum of (3.4) is attained under the constraints (3.5), $\alpha_i \geq 0$, $\sum_{i=1}^{v+1} \alpha_i = \Delta t$. The usefulness of Rosenbloom's theorem consists essentially in the restriction of the class of measures over which we seek the minimum of I , from the whole of P_v to its extremal points. Because of the simple structure of these points, the problem of characterizing and determining the minimizing measure μ_v has become one of nonlinear programming; one seeks the set of quadruples $\{(\alpha_i, t_i, x_i, \dot{x}_i), i = 1, 2, \dots, v + 1\}$, $t_i \in [t_0, t_f]$, $x_i \in A$, $\dot{x}_i \in B$, $\alpha_i \geq 0$, $\sum_{i=1}^{v+1} \alpha_i = \Delta t$, satisfying the v constraints (3.5), so that the expression in the right side of (3.4) is a minimum. Because I achieves its minimum over P_v at one, or some, extremal points, such a set exists, and is not in general unique. For each v there are in general several measures μ_v , and then sets $\{(\alpha_i, t_i, x_i, \dot{x}_i)\}$, at which I attains its minimum.

Let $\{(\alpha_i, t_i, x_i, \dot{x}_i)\}$ be one such set. Some of the numbers α_i may be zero, since the corresponding extremal point may be the linear combination of less than $v + 1$ atomic measures. We shall discard those values of the index i , relabel the rest of the quadruples $(\alpha_i, t_i, x_i, \dot{x}_i)$, and assume that there are left a total of $N(v) \leq v + 1$ different quadruples for which $\alpha_i > 0$. Thus the minimizing measure μ_v gives the functional I the value

$$(3.6) \quad \sum_{i=1}^N \alpha_i f_0(t_i, x_i, \dot{x}_i);$$

we write N for $N(v)$. The quadruples $(\alpha_i, t_i, x_i, \dot{x}_i)$ satisfy:

$$(3.7) \quad \alpha_i > 0, \quad \sum_{i=1}^N \alpha_i = \Delta t, \quad t_i \in [t_0, t_f], \quad x_i \in A, \quad \dot{x}_i \in B, \quad i = 1, 2, \dots, N,$$

$$\sum_{i=1}^N \alpha_i \left[\dot{x}_i \operatorname{grad} \phi_k(t_i, x_i) + \frac{\partial \phi_k}{\partial t}(t_i, x_i) \right] = \Delta \phi_k, \quad k = 1, 2, \dots, v.$$

As we shall see in § 5, it is possible to choose from the sequence $\{\mu_v\}$ a subsequence which converges to a measure, or functional, consisting of a generalized curve; such a sequence provides a characterization of the generalized curve essentially by means of the set of quadruples $\{(\alpha_i, t_i, x_i, \dot{x}_i)\}$ for a sufficiently large value of the index v . However, as we shall show in § 4, it is possible to construct a curve which assigns to the original functional J a value arbitrarily close to (3.6) for any minimizing measure μ_v , regardless of whether it is, or is not, in the converging subsequence.

4. The boundary conditions. We shall derive in this section some properties of those sets of quadruples $\{(\alpha_i, t_i, x_i, \dot{x}_i), i = 1, \dots, N\}$ satisfying the boundary conditions (3.7). It should be clear that these sets depend on the index v ; we wish to study their behavior as this index varies, especially as it tends to infinity. For simplicity, we write $z = (t, x)$, $\dot{z} = (1, \dot{x})$, $\hat{A} = [t_0, t_f] \times A$, $\hat{B} = \{1\} \times B$. The boundary conditions (3.7) are rewritten as

$$(4.1) \quad \alpha_i > 0, \quad \sum_{i=1}^N \alpha_i = \Delta t, \quad \sum_{i=1}^N \alpha_i \dot{z}_i \text{grad } \phi_k(z_i) = \Delta \phi_k, \quad k = 1, \dots, v, \\ i = 1, 2, \dots, N,$$

where $\phi(z) \equiv \phi(t, x)$, $\Delta \phi = \phi(t_f, x_f) - \phi(t_0, x_0) = \phi(z_f) - \phi(z_0)$. We search now for properties of the triples $\{(\alpha_i, z_i, \dot{z}_i), i = 1, \dots, N\}$ satisfying (4.1), $z_i \in \hat{A}$, $\dot{z}_i \in \hat{B}$. Note that these triples satisfy

$$(4.2) \quad \sum_{i=1}^N \alpha_i \dot{z}_i \text{grad } \phi(z_i) = \Delta \phi$$

for all polynomials ϕ which are linear combinations of the first v monomials in the components of z (that is, in t and the components of x). We shall construct some special polynomials, so as to derive from (4.2) some properties of the triples satisfying (4.1). We proceed with this construction in several steps.

(i) We choose an ordering for the first v monomials in the components of z , separating them first into classes, the first class being that of first-degree monomials, the second that of second-degree ones, etc. In each class, we follow a cyclic ordering based on the indices of the components of z ; note that the time t is the first component of this $(n+1)$ -tuple, x^1 the second, x^2 the third, \dots , x^n the $(n+1)$ st.

(ii) The compact set \hat{A} is contained in a ball with center at the origin of radius, say $a/2$. Let \mathcal{A} be a closed ball with center at the origin of radius a . Then the points of \hat{A} , and in particular the points $z_0, z_f, z_i, i = 1, 2, \dots, N$, are at a distance of at least $a/2$ from the boundary of \mathcal{A} . In the development to follow, we shall construct neighborhoods of points of \hat{A} with radii tending to zero as $v \rightarrow \infty$; these neighborhoods are contained in \mathcal{A} for sufficiently high v .

(iii) Consider the odd function $r(\lambda)$, $\lambda \in [-2a, 2a]$, whose graph is shown in Fig. 1. We assume $e(1 + \eta) < 2a$. Consider the problem of approximating this function over the interval $[-2a, 2a]$ by an odd polynomial p of degree M , say, in λ . By Jackson's theorem, a polynomial p exists such that

$$(4.3) \quad \max_{|\lambda| \leq 2a} |p(\lambda) - r(\lambda)| \leq \frac{K_1}{e^2 \eta^2 M},$$

with K_1 depending only on a . Put for $\lambda \in [-2a, 2a]$,

$$p_1(\lambda) = \int_{-2a}^{\lambda} p(\theta) d\theta, \quad r_1(\lambda) = \int_{-2a}^{\lambda} r(\theta) d\theta;$$

then there is a constant K_2 , depending only on a , such that

$$(4.4) \quad \max_{|\lambda| \leq 2a} |p_1(\lambda) - r_1(\lambda)| \leq \frac{K_2}{e^2 \eta^2 M};$$

the even polynomial p_1 is of degree $M + 1$.

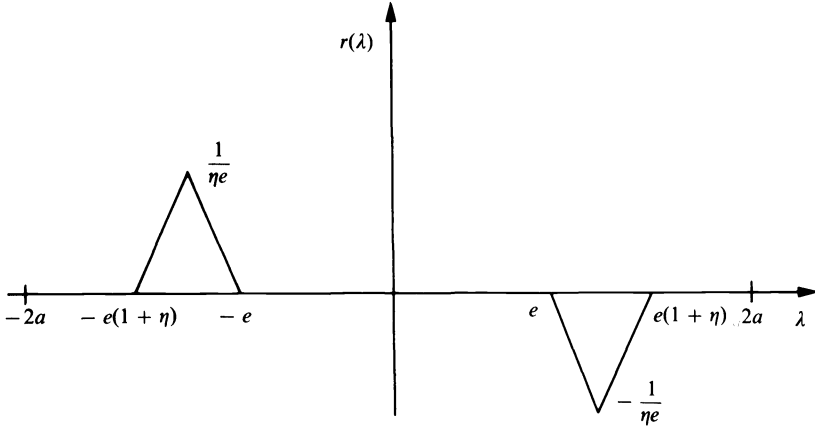


FIG. 1. Graph of the function $r(\lambda)$, $\lambda \in [-2a, 2a]$, used in the construction of some polynomials

(iv) By means of linear combinations of the first v monomials, we wish to construct a polynomial of the type $p_1(|z - \underline{z}|)$, with $\underline{z} \in \hat{A}$. The degree $M + 1$ of the corresponding polynomial $p_1(\lambda)$, $\lambda \in [-2a, 2a]$, is determined by the index v ; let $R = R(v)$ be such that the monomials z^{j2R} , $j = 1, 2, \dots, n + 1$, are included in the first v monomials (but such that not all the monomials z^{j2R+2} are included); then $M + 1 = 2R(v)$. We define a nondecreasing step function $v \rightarrow M(v)$, by $M(v) + 1 = 2R(v)$. Of course, $M(v) \rightarrow \infty$ as $v \rightarrow \infty$. We write M for $M(v)$.

(v) In the definition of the function r , put $e = cM^{-1/4}$, with c any positive constant, $\eta = M^{-1/8}$. Then $e(1 + \eta) \rightarrow 0$ as $v \rightarrow \infty$, and for sufficiently high v , say $v > v_0(c)$, this quantity is less than $2a$, a requirement we put in the definition of the function r . Note that $v_0(c_1) > v_0(c_2)$ for $c_1 > c_2 > 0$. Define

$$\psi_{c,\underline{z}}(z) = p_1(|z - \underline{z}|),$$

$\underline{z} \in \hat{A}$, $z \in \mathcal{A}$; then $|z - \underline{z}| \leq 3a/2$, and (4.4) applies:

$$(4.5) \quad \max_{z \in \mathcal{A}} |\psi_{c,\underline{z}}(z) - r_1(|z - \underline{z}|)| \leq \frac{K_2}{e^2 \eta^2 M} = \frac{K_2}{c^2} M^{-1/4}.$$

Let

$$\begin{aligned} D_1 &= \{z : |z - \underline{z}| \leq e, z \in \mathcal{A}\}, \\ D_2 &= \{z : e \leq |z - \underline{z}| \leq e(1 + \eta), z \in \mathcal{A}\}, \\ D &= D_1 \cup D_2. \end{aligned}$$

Then $D \subset \mathcal{A}$ for $v > v_0(c)$. Since $M^{-1/4} \rightarrow 0$ as $v \rightarrow \infty$, for sufficiently high v the function $\psi_{c,\underline{z}}$ has a value near 1 for $z \in D_1$, and a value near zero for $z \in \mathcal{A} \setminus D$.

Further, a simple computation gives

$$|\text{grad } \psi_{c,\underline{z}}(z)| = p(|z - \underline{z}|);$$

$\psi_{c,\underline{z}}$ is rotation-invariant. Therefore, for $v > v_0(c)$,

$$(4.6) \quad \max_{z \in \mathcal{A}} ||\text{grad } \psi_{c,\underline{z}}(z)| - r(|z - \underline{z}|)| \leq \frac{K_1}{e^2 \eta^2 M} = \frac{K_1}{c^2} M^{-1/4}.$$

Thus, $|\text{grad } \psi_{c,\underline{z}}(z)|$ is near zero for $z \in D_1$ and $z \in \mathcal{A} \setminus D$, and near $1/\eta e = (1/c)M^{3/8}$ for z at a distance $e(1 + \frac{1}{2}\eta)$ from \underline{z} .

We prove now some properties of the triples $\{(\alpha_i, z_i, \dot{z}_i)\}$ by choosing specific values of c and \underline{z} , and then examining the implications of (4.2) for $\phi = \psi_{c,\underline{z}}$.

PROPOSITION 2. *Let $\{(\alpha_i, z_i, \dot{z}_i), i = 1, 2, \dots, N\}$ be a solution of (4.1). There is an integer \hat{v} and a vector in the set $\{z_i, i = 1, 2, \dots, N\}$, to be relabeled z_1 , such that*

$$|z_1 - z_0| \leq 2M^{-1/4}, \quad v > \hat{v}.$$

Proof. Put $\underline{z} = z_0$ and $c = 1$ in the definition of $\psi_{c,\underline{z}}$. Since $z_f \neq z_0$, z_f is not in the set D corresponding to ψ_{1,z_0} for sufficiently high v , say $v > \hat{v}_1$. Put $\phi = \psi_{1,z_0}$ in (4.2). If $v > \max(\hat{v}_1, v_0(1))$, the right side of (4.2) satisfies, by (4.5),

$$(4.7) \quad |\Delta\psi_{1,z_0} - 1| \leq 2K_2M^{-1/4}.$$

Assume there is no \hat{v}_2 such that for all $v > \hat{v}_2$ there is at least one vector z_i , $1 \leq i \leq N$, in D . Then it is possible to find an increasing sequence $\{v^s, s = 1, 2, \dots\}$ such that there are no vectors z_i in D for those values of the index v . For those values of v , since $\dot{z}_i \in \hat{B}$, the left side of (4.2) satisfies, by (4.6),

$$(4.8) \quad \left| \sum_{i=1}^N \alpha_i \dot{z}_i \text{grad } \psi_{1,z_0}(z_i) \right| \leq \sum_{i=1}^N \alpha_i |\dot{z}_i| |\text{grad } \psi_{1,z_0}(z_i)| \\ \leq K_1(\Delta t) \hat{\beta} M^{-1/4},$$

with $\hat{\beta} = (1 + \beta^2)^{1/2}$. For any $v^s > \max(\hat{v}_1, v_0(1))$, the two inequalities (4.7) and (4.8) together with (4.2) imply

$$2K_2M^{-1/4} \geq 1 - K_1(\Delta t) \hat{\beta} M^{-1/4};$$

this is clearly false. We conclude that there is \hat{v}_2 such that for all $v > \hat{v}_2$ there is at least one vector z_i in D . We choose, for each $v > \hat{v}_2$, one such vector, and relabel their associated triples $(\alpha_1, z_1, \dot{z}_1)$; the triples previously labeled in this manner will occupy the places of the ones just relabeled.

The vector z_1 satisfies

$$|z_1 - z_0| \leq e(1 + \eta) = M^{-1/4}(1 + M^{-1/8}) \leq 2M^{-1/4}$$

for $v > \hat{v} = \max(\hat{v}_1, v_0(1), \hat{v}_2, v')$; v' is an index such that $M^{-1/8} < 1$.

This is the first of our structural results on the triples $\{(\alpha_i, z_i, \dot{z}_i)\}$. It can be said the support of μ_v exhibits by means of this result a remarkable curve-like behavior; continuous curves starting at z_0 certainly have points arbitrarily near it. Other results of this kind will permit the construction of an actual curve giving the functional I a value as close as desired to $I(\mu_v)$. We give next one of these results; the (rather lengthy) proof can be found in Appendix A.

PROPOSITION 3. *Let $\{(\alpha_i, z_i, \dot{z}_i), i = 1, 2, \dots, N\}$ be a solution of (4.1). There is an integer \tilde{v} so that it is possible, for any $v > \tilde{v}$, to order these triples such that*

$$|z_{j+1} - z_j| \leq 4M^{-1/4}, \quad v > \tilde{v}, \quad j = 1, 2, \dots, N-1,$$

where the subindices used are those of the new ordering, and z_1 is the vector selected in Proposition 2.

We use this new ordering from now on. We prove a final structural result.

PROPOSITION 4. *Let $\{(\alpha_i, z_i, \dot{z}_i), i = 1, 2, \dots, N\}$ be a solution of (4.1). Then :*

(i) *There is a constant K_3 and an index v^0 such that*

$$|\alpha_i \dot{z}_i| \leq K_3 M^{-1/4}, \quad v > v^0, \quad i = 1, 2, \dots, N;$$

also, $N(v) \rightarrow \infty$ as $v \rightarrow \infty$. (ii) Let $\hat{y}_i = z_1 + \sum_{j=1}^i \alpha_j \dot{z}_j$. Then there exists a constant K_4 and integer v^0 such that

$$|\hat{y}_i - z_i| \leq K_4 M^{-1/4}, \quad v > v^0, \quad i = 1, 2, \dots, N.$$

Proof. (i) Assume first that $z_k \neq z_l, k \neq l, l = 1, \dots, N$, for all v . Assume also that the assertion in (i) is false, that is, that there is an index j and an increasing sequence v^s such that $M^{1/4}|\alpha_j \dot{z}_j| \rightarrow \infty$ along this sequence. Fix v at a value v^r in this sequence. Put $c = \gamma$ in the definition of $\psi_{c,z}$, a constant to be specified below, and

$$z = z_j + \gamma M^{-1/4}(1 + \frac{1}{2}M^{-1/8})e,$$

where $M = M(v^r)$ and e is a unit vector in the direction of $\alpha_j \dot{z}_j$ for $v = v^r$. (Of course, we can assume $\alpha_j \dot{z}_j \neq 0$ at v^r .) Then, $|\text{grad } \psi_{c,z}(z_j)|$ is, for v^r sufficiently large, near $2/e^2 \eta^2 = (2/\gamma^2)M^{3/4}$; further, $\text{grad } \psi_{c,z}(z_j)$ is in the same direction as $z - z_j$, and then in the same direction as e , that is, as $\alpha_j \dot{z}_j$. We have, then,

$$\begin{aligned} |\alpha_j \dot{z}_j \text{ grad } \psi_{c,z}(z_j)| &= |\alpha_j \dot{z}_j| |\text{grad } \psi_{c,z}(z_j)| \\ &\leq |\Delta \psi_{c,z}| + \sum_{i \neq j} |\alpha_i \dot{z}_i| |\text{grad } \psi_{c,z}(z_i)|. \end{aligned}$$

Since no z_i equals $z_j, i \neq j$, no $z_i, i \neq j$, is in the set D corresponding to $\psi_{c,z}$ for sufficiently small γ . Fix γ at such a value; remember v is fixed at a value v^r . The second term in the right side of the inequality above is not higher than $(K_1/\gamma^2)(\Delta t)\hat{\beta}M^{-1/4}$, while $|\Delta \psi_{c,z}|$ is at most near 2 (it can be near 0, or near 1). The left side, however, is near $(2/\gamma^2)|\alpha_j \dot{z}_j|M^{3/4}$, which approaches infinity if v^r is sufficiently large. A contradiction arises, and we conclude that the assertion in (i) is true under the assumption that all z_i 's are different, for all v .

Consider now a solution $\{(\underline{\alpha}_i, z_i, \dot{z}_i), i = 1, 2, \dots, N\}$ of (4.1) for which this assumption is not necessarily true, that is, for which all vectors z_i are not necessarily different for some values of the index v . Consider the equations in $\{(\alpha_i, \dot{z}_i), i = 1, 2, \dots, N\}$:

$$\sum_{i=1}^N \alpha_i \dot{z}_i \text{ grad } \phi_k(z_i + \varepsilon g_i) = \Delta \phi_k, \quad k = 1, 2, \dots, v,$$

where we choose the vectors $g_i, i = 1, 2, \dots, N$, so that for any $\varepsilon > 0$ satisfying $0 < \varepsilon < \varepsilon_0$ for some $\varepsilon_0 > 0$ the vectors $z_i + \varepsilon g_i$ are different; the vectors g_i may of course depend on the index v . The equation above has as a solution for $\varepsilon = 0$ the tuples $\{(\underline{\alpha}_i, \underline{z}_i)\}$; by the implicit function theorem, it has a solution $\{(\alpha_i(\varepsilon), \dot{z}_i(\varepsilon))\}$ for sufficiently small ε . By the argument used in the first part of the proof, since $z_k + \varepsilon g_k \neq z_l + \varepsilon g_l$ for all v ,

$$|\alpha_i(\varepsilon) \dot{z}_i(\varepsilon)| \leq K_3 M^{-1/4}, \quad v > v^0,$$

for all sufficiently small ε , and thus for $\varepsilon = 0$ by continuity. Then $|\alpha_i \dot{z}_i| \leq K_3 M^{-1/4}$, $v > v^0$, $i = 1, 2, \dots, N$.

We show now that $N(v) \rightarrow \infty$ as $v \rightarrow \infty$. Assume that $N(v) \leq N_0 < \infty$ for all v . Then, for $v > v^0$, since

$$\sum_{i=1}^N \alpha_i \dot{z}_i = z_f - z_0,$$

$$|z_f - z_0| \leq N_0 K_3 M^{-1/4},$$

a contradiction since $z_f \neq z_0$.

(ii) By (i) of the proposition, $|\hat{y}_1 - z_1| = |\alpha_1 \dot{z}_1| \leq K_3 M^{-1/4}$, $v > v^0$. Assume there is an index $j \neq 1$ such that $M^{1/4}|\hat{y}_j - z_j| \rightarrow \infty$ as $v \rightarrow \infty$. Then,

$$M^{1/4}|(\hat{y}_j - z_j) - (\hat{y}_{j-1} - z_{j-1})| = M^{1/4}|(\hat{y}_j - \hat{y}_{j-1}) - (z_j - z_{j-1})|$$

$$\leq M^{1/4}|\hat{y}_j - \hat{y}_{j-1}| + M^{1/4}|z_j - z_{j-1}| \leq K_3 + 4,$$

for $v > \max(v^0, \hat{v})$, so that $M^{1/4}|\hat{y}_{j-1} - z_{j-1}| \rightarrow \infty$ as $v \rightarrow \infty$. This, in turn, implies $M^{1/4}|\hat{y}_{j-2} - z_{j-2}| \rightarrow \infty$ as $v \rightarrow \infty$, etc; eventually, $M^{1/4}|\hat{y}_1 - z_1|$ would be shown to tend to infinity as $v \rightarrow \infty$, a contradiction.

We remark, finally, that \hat{y}_N is close to z_f ; indeed, $\hat{y}_N = z_1 + (z_f - z_0)$, so that $\hat{y}_N - z_f = z_1 - z_0$.

We revert to our original notation, write $y_i = x_1 + \sum_{j=1}^i \alpha_j \dot{x}_j$, so that $\hat{y}_i = (t_1 + \sum_{j=1}^i \alpha_j, y_i)$, and put $v_m = \max(\hat{v}, \tilde{v}, v^0, \underline{v}^0)$, $K = \max(4, K_3, K_4)$. We have shown the following,

THEOREM 3. *The quadruples satisfying (3.7) can be ordered in a set $\{(\alpha_i, t_i, x_i, \dot{x}_i), i = 1, 2, \dots, N\}$ such that a constant K and an integer v_m exist such that the following quantities, all functions of the index v , are no higher than $KM^{-1/4}$ for $v > v_m$:*

$$|x_1 - x_0|, \quad |x_i - x_{i-1}|, \quad i = 2, 3, \dots, N(v), \quad |t_1 - t_0|, \quad |t_i - t_{i-1}|,$$

$$i = 2, 3, \dots, N(v),$$

$$|\alpha_i \dot{x}_i|, \quad \alpha_i, \quad |y_i - x_i|, \quad \left| \left(t_1 + \sum_{j=1}^i \alpha_j \right) - t_i \right|, \quad i = 1, 2, \dots, N(v).$$

5. Approximation and convergence. A complete picture of the effects of the constraints (3.7) on the measures μ_v of § 3 was developed in the previous section. We show now that it is possible to construct a sequence of curves $x_v(t)$, $t \in [t_0, t_f]$, such that, for all $f \in \mathcal{C}(\Omega)$,

$$(5.1) \quad \int_{\Omega} f(t, x, \dot{x}) d\mu_v = \int_{t_0}^{t_f} f(t, x_v(t), \dot{x}_v(t)) dt + \xi_v,$$

where $|\xi_v| \rightarrow 0$ as $v \rightarrow \infty$. In particular, this is true for $f = f_0$, the integrand in the functional J in (2.1). Since in this case the left side of (5.1) is $I(\mu_v)$, which tends to I^* as $v \rightarrow \infty$, the curve $x_v(t)$, $t \in [t_0, t_f]$ gives the functional J a value close to I^* for sufficiently high v .

Let the set $\{(\alpha_i, t_i, x_i, \dot{x}_i), i = 1, \dots, N(v)\}$ be as in Theorem 3; we omit a subscript v from the elements of these quadruples for simplicity. Define a curve $x_v(t), t \in [t_0, t_f]$ as follows. Construct first a polygon in R^n , by joining with straight lines the points $x_1 \rightarrow y_1 \rightarrow y_2 \rightarrow \dots \rightarrow y_N$; remember that $y_i = x_1 + \sum_{j=1}^i \alpha_j \dot{x}_j$. Further, assign the value $t = t_0$ to the point x_1 , $t = t_0 + \alpha_1$ to y_1 , $t = t_0 + \alpha_1 + \alpha_2$ to y_2 , \dots , $t = t_0 + \sum_{j=1}^i \alpha_j$ to y_i , \dots , $t = t_f$ to y_N . At last, define the time variation along each segment linearly so that $\dot{x}_v(t) = \dot{x}_1$ on $(t_0, t_0 + \alpha_1)$, \dot{x}_2 on $(t_0 + \alpha_1, t_0 + \alpha_1 + \alpha_2)$, \dots , \dot{x}_i on $(t_0 + \sum_{j=1}^{i-1} \alpha_j, t_0 + \sum_{j=1}^i \alpha_j)$, \dots , \dot{x}_N on $(t_f - \alpha_N, t_f)$.

Consider first a function $f \in \mathcal{C}(\Omega)$ which is the restriction to Ω of a \mathcal{C}^∞ -function defined on an open ball enclosing Ω . Then, for $(t, x, \dot{x}), (\underline{t}, \underline{x}, \dot{x}) \in \Omega$,

$$|f(t, x, \dot{x}) - f(\underline{t}, \underline{x}, \dot{x})| \leq \xi(\dot{x})[|t - \underline{t}| + |x - \underline{x}|] \leq \theta[|t - \underline{t}| + |\underline{x} - x|]$$

for some function $\xi(\dot{x}), \dot{x} \in B$; θ is the maximum of this continuous function over B . Then f is (uniformly in \dot{x}) Lipschitz in x and t . By the mean value theorem,

$$\int_{t_0}^{t_f} f(t, x_v(t), \dot{x}_v(t)) dt = \sum_{i=1}^N \alpha_i f(t_i, \underline{x}_i, \dot{x}_i),$$

where \underline{x}_i is in the segment joining y_i and y_{i-1} , $i = 2, \dots, N$, \underline{x}_1 is in the segment joining y_1 and x_1 , \underline{t}_i is in $(t_0 + \sum_{j=1}^{i-1} \alpha_j, t_0 + \sum_{j=1}^i \alpha_j)$, $i = 2, \dots, N$, and \underline{t}_1 is in $(t_0, t_0 + \alpha_1)$. By Theorem 3,

$$|\underline{x}_i - x_i| < |\underline{x}_i - y_i| + |y_i - x_i| \leq 2KM^{-1/4}, \quad v > v_m, \quad i = 1, 2, \dots, N,$$

and

$$|\underline{t}_i - t_i| \leq \left| \underline{t}_i - \left(t_0 + \sum_{j=1}^i \alpha_j \right) \right| + \left| \left(t_0 + \sum_{j=1}^i \alpha_j \right) - t_i \right| \leq 2KM^{-1/4},$$

$$v > v_m, \quad i = 1, 2, \dots, N.$$

Therefore,

$$\begin{aligned} & \left| \int_{t_0}^{t_f} f(t, x_v(t), \dot{x}_v(t)) dt - \sum_{i=1}^N \alpha_i f(t_i, x_i, \dot{x}_i) \right| \\ & \leq \sum_{i=1}^N \alpha_i |f(\underline{t}_i, \underline{x}_i, \dot{x}_i) - f(t_i, x_i, \dot{x}_i)| \\ & \leq \theta \sum_{i=1}^N \alpha_i \{ |\underline{t}_i - t_i| + |\underline{x}_i - x_i| \} \\ & \leq 2(\Delta t) \theta K M^{-1/4}, \quad v > v_m, \end{aligned}$$

from which (5.1) follows, since $\int_{\Omega} f(t, x, \dot{x}) d\mu_v = \sum_{i=1}^N \alpha_i f(t_i, x_i, \dot{x}_i)$.

Even if f is not necessarily as above, we can show that (5.1) still holds by using the fact that the set of functions in $\mathcal{C}(\Omega)$ which are restrictions to Ω of \mathcal{C}^∞ -functions on an open ball enclosing Ω is dense in $\mathcal{C}(\Omega)$ in the topology of uniform convergence. Let $f \in \mathcal{C}(\Omega)$ and g the restriction of a \mathcal{C}^∞ -function such that

$$\sup_{\Omega} |f(t, x, \dot{x}) - g(t, x, \dot{x})| \leq 1/v.$$

Then,

$$\begin{aligned}
 & \left| \sum_{i=1}^N \alpha_i f(t_i, x_i, \dot{x}_i) - \int_{t_0}^{t_f} f(t, x_v(t), \dot{x}_v(t)) dt \right| \\
 & \leq \sum_{i=1}^N \alpha_i |f(t_i, x_i, \dot{x}_i) - g(t, x_i, \dot{x}_i)| + \int_{t_0}^{t_f} |g(t, x_v(t), \dot{x}_v(t)) \\
 & \quad - f(t, x_v(t), \dot{x}_v(t))| dt + \left| \int_{t_0}^{t_f} g(t, x_v(t), \dot{x}_v(t)) dt - \sum_{i=1}^N \alpha_i g(t_i, x_i, \dot{x}_i) \right| \\
 & \leq (2/v + 2\theta_1 K M^{-1/4}) \Delta t, \quad v > v_m,
 \end{aligned}$$

where θ_1 is the Lipschitz parameter associated with g . Our contention, that (5.1) is valid for all $f \in \mathcal{C}(\Omega)$, is proved.

As mentioned above, any curve $x_v(t)$, $t \in [t_0, t_f]$ gives the functional J a value close to its minimum I^* , for sufficiently high v ; in a sense, any of these curves is a solution to our original problem. It is possible, however, to gain further insight into it by considering the convergence properties of the sequence $\{\mu_v\}$ in the weak* topology. As we remarked in §3, there may be many measures, all necessarily extremal points of the set P_v , at which the minimum of I over P_v is attained; we have chosen arbitrarily one such measure, labeled it μ_v , and formed in this manner the sequence $\{\mu_v\}$. It should be recognized, therefore, that the measures μ_{v_0} and μ_{v_0+1} , for instance, may be quite different, in the sense that in general the two integrals

$$\int_{\Omega} f d\mu_{v_0}, \quad \int_{\Omega} f d\mu_{v_0+1}$$

may be quite different, even if v_0 is very large, for some functions $f \in \mathcal{C}(\Omega)$; it is only when $f = f_0$ that these integrals are known to be quite close for sufficiently large values of v_0 .

It is possible, however, by Alaoglu's theorem, to choose a convergent (in the weak* topology) subsequence from the sequence $\{\mu_v\}$; we shall call this subsequence also $\{\mu_v\}$, and its limit μ_0 . Then, for every $f \in \mathcal{C}(\Omega)$,

$$\lim_{v \rightarrow \infty} \int_{\Omega} f d\mu_v = \int_{\Omega} f d\mu_0.$$

In particular, if $f = f_0$, $I(\mu_0) = I^*$, the minimum of I over P ; μ_0 , necessarily an extremal point of P , has been characterized as being the weak* limit of a subsequence of the measures μ_v . We can characterize it even further; let $\{x_v(t), t \in [t_0, t_f]\}$ be the sequence of curves corresponding to the subsequence $\{\mu_v\}$. Then

$$\lim_{v \rightarrow \infty} \int_{t_0}^{t_f} f(t, x_v(t), \dot{x}_v(t)) dt = \int_{\Omega} f d\mu_0,$$

for all $f \in \mathcal{C}(\Omega)$. Therefore, μ_0 is the weak* limit of a sequence of curves, and is then a generalized curve. We summarize as follows.

THEOREM 4. *The minimum of the functional I over the set of measures P occurs at a generalized curve, that is, a measure which is the weak* limit of a sequence of curves.*

A different proof to the fact that the minimum of I over P occurs at a measure which is the weak* limit of curves is given by L. C. Young in [3, Appendix 2].

In the next section, we seek to characterize the generalized curve μ_0 in terms of the sets of quadruples $\{(\alpha_i, t_i, x_i, \dot{x}_i), i = 1, 2, \dots, N, v = 1, 2, \dots\}$ corresponding to the curves $\{x_v(t), t \in [t_0, t_f]\}$ associated with the subsequence $\{\mu_v\}$. Some further properties of the curves constructed in this section are given in Appendix B.

6. The generalized curve. It is shown in [3] that a generalized curve is characterized by two elements, an absolutely continuous function $x(t), t \in [t_0, t_f]$, and a family of probability measures σ_t defined on B for almost every t in $[t_0, t_f]$ such that $\dot{x}(t) = \int_B \dot{x} d\sigma_t$ a.e. on $[t_0, t_f]$; further, σ_t is completely determined by the values $\int_B f d\sigma_t$ for functions $f(\dot{x}), \dot{x} \in B$ in $\mathcal{C}(B)$, independent of x and t . We seek to characterize both $x(t), \sigma_t, t \in [t_0, t_f]$ in terms of the set of quadruples $\{(\alpha_i, t_i, x_i, \dot{x}_i), i = 1, 2, \dots, N(v), v = 1, 2, \dots\}$ associated with the subsequence $\{\mu_v\}$ tending weakly* to the generalized curve.

Consider the sequence of functions $\{x_v(t), t \in [t_0, t_f]\}$ associated with this subsequence $\{\mu_v\}$; this family is equibounded since $x_v(t) \in A, t \in [t_0, t_f], v = 1, 2, \dots$. It is also equicontinuous; take

$$t_a = \sum_{i=1}^{N_a} \alpha_i + \theta_a \alpha_{N_a+1}, \quad 0 < \theta_a < 1,$$

$$t_b = \sum_{i=1}^{N_b} \alpha_i + \theta_b \alpha_{N_b+1}, \quad 0 < \theta_b < 1,$$

with $N_a \geq N_b$. Then the corresponding values for $x_v(\cdot)$ satisfy

$$\begin{aligned} |x_v(t_a) - x_v(t_b)| &= \left| \sum_{i=1}^{N_a} \alpha_i \dot{x}_i + \theta_a \alpha_{N_a+1} \dot{x}_{N_a+1} - \theta_b \alpha_{N_b+1} \dot{x}_{N_b+1} \right| \\ &\leq \beta \left| \sum_{i=1}^{N_a} \alpha_i + \theta_a \alpha_{N_a+1} - \theta_b \alpha_{N_b+1} \right| = \beta |t_a - t_b|, \end{aligned}$$

from which our contention follows. Then, by Ascoli's theorem, there is a subsequence, to be denoted by $\{x_v(t), t \in [t_0, t_f]\}$ as well, which converges in the topology of uniform convergence to an absolutely continuous function $x(t), t \in [t_0, t_f]$. We denote the subsequence of associated measures by $\{\mu_v\}$.

We shall take a further subsequence of this subsequence, denoted in the same manner. It is shown in [3] that it is possible to extract this subsequence so that

$$\int_0^t f(\dot{x}_v(\tau)) d\tau \xrightarrow{v \rightarrow \infty} \int_B f ds_t,$$

for some measure s_t on B , uniformly in $t \in [t_0, t_f]$ and all $f \in \mathcal{C}(B)$. Let T be any subinterval of $[t_0, t_f]$ of the form $[t, t + \Delta]$ or $[t - \Delta, t]$, $\Delta > 0$, for fixed t . Then

$$\frac{1}{\Delta} \int_T f(\dot{x}_v(\tau)) d\tau \xrightarrow{v \rightarrow \infty} \int_B f ds_{t,\Delta}$$

with $s_{t,\Delta} = (1/\Delta)(s_{t+\Delta} - s_t)$ if $T = [t, t + \Delta]$ and $s_{t,\Delta} = (1/\Delta)(s_t - s_{t-\Delta})$ if $T = [t - \Delta, t]$; this convergence is uniform in t . Note that, by putting $f \equiv 1$, we conclude that $\int_B ds_{t,\Delta} = 1$; the measure $s_{t,\Delta}$ is a probability measure on B . Finally, it is shown in [3] that

$$\int_B f ds_{t,\Delta} \xrightarrow{\Delta \rightarrow 0} \int f d\sigma_t$$

for a probability measure σ_t on B , the convergence being independent of whether the intervals taken are of the form $[t, t + \Delta]$ or $[t - \Delta, t]$, and also independent of the particular value of the variable t .

As mentioned at the beginning of this section, these two objects, the absolutely continuous function $x(t)$, $t \in [t_0, t_f]$, and the probability measure σ_t completely characterize the generalized curve.

Consider this last subsequence, and define a probability measure $\sigma_{t,\Delta}^v$ on B , for all values of t not of the form $t_0 + \sum_{j=1}^i \alpha_j$, for $i = 1, 2, \dots, N(v)$, by

$$(6.1) \quad \int_B f(\dot{x}) d\sigma_{t,\Delta}^v = \frac{1}{\Delta} \int_T f(\dot{x}_v(\tau)) d\tau = \frac{\sum_{i \in \mathcal{N}_{t,\Delta}} \theta_i \alpha_i f(\dot{x}_i)}{\sum_{i \in \mathcal{N}_{t,\Delta}} \theta_i \alpha_i}$$

where T is a subinterval of $[t_0, t_f]$ either of the form $[t, t + \Delta]$ or $[t - \Delta, t]$; for each t we make a fixed, appropriate choice as to the form of T . We examine the situation in detail when T is of the form $[t, t + \Delta]$; a similar analysis can be made in the other case. The set $\mathcal{N}_{t,\Delta} = \{i_t, i_t + 1, \dots, i_t + N_t(\Delta)\}$ is a set of integers dependent on t, Δ, v , so that t is in $(t_0 + \sum_1^{i_t} \alpha_j, t_0 + \sum_1^{i_t+1} \alpha_j)$, and $t + \Delta$ is in $(t_0 + \sum_1^{i_t+N_t(\Delta)-1} \alpha_j, t_0 + \sum_1^{i_t+N_t(\Delta)} \alpha_j)$. Then, $\theta_i < 1$ for $i = i_t, i = i_t + N_t(\Delta)$, $\theta_i = 1$ otherwise. Of course, the quadruples $(\alpha_i, t_i, x_i, \dot{x}_i)$ are those corresponding to $\dot{x}_v(t)$, $t \in [t_0, t_f]$.

We show that the measure $\sigma_{t,\Delta}^v$ is a good approximation to σ_t for large v and small Δ . Indeed, for all t as in (6.1),

$$(6.2) \quad \left| \int_B f(\dot{x}) d\sigma_{t,\Delta}^v - \int_B f(\dot{x}) d\sigma_t \right| \leq \left| \int_B f(\dot{x}) d\sigma_{t,\Delta}^v - \int_B f ds_{t,\Delta} \right| + \left| \int_B f ds_{t,\Delta} - \int_B f(\dot{x}) d\sigma_t \right| < \varepsilon$$

for all $f \in \mathcal{C}(B)$, provided that v is higher than an integer $M(\Delta)$ so as to make the first expression less than $\varepsilon/2$, and $\Delta < \delta(\varepsilon)$, so as to make the second less than $\varepsilon/2$. Thus, given $\varepsilon > 0$, one can choose Δ , and then v , so that (6.2) is satisfied. Of course, $M(\Delta)$ and $\delta(\varepsilon)$ do depend on the particular $f \in \mathcal{C}(B)$. The measure $\sigma_{t,\Delta}^v$ converges in the weak* topology to σ_t as $\Delta \rightarrow 0$, $v \rightarrow \infty$.

Define

$$\beta_j(t) = \frac{\theta_j \alpha_j}{\sum_{i \in \mathcal{N}_{t,\Delta}} \theta_i \alpha_i},$$

with t as in (6.1). Then,

$$(6.3) \quad \sum_{i \in \mathcal{N}_{t,\Delta}} \beta_i(t) \dot{x}_i = \dot{x}_v(t),$$

so that the derivative of the function $x_v(t)$, $t \in [t_0, t_f]$ is a convex average of the values \dot{x}_i associated with a given point on this curve.

Still considering the same subsequence for which (6.3) is valid, we derive a condition as to whether the generalized curve is locally an ordinary curve, that is, whether the measure σ_t is atomic with support $\dot{x}(t)$. From (6.3), it follows that this condition is satisfied if and only if $\dot{x}_v(t) \rightarrow \dot{x}(t)$; indeed, if σ_t is atomic the left side of (6.3) must be close to $\dot{x}(t)$; if $\dot{x}_v(t) \rightarrow \dot{x}(t)$, then the left side of (6.3) is close to $\dot{x}(t)$, for high v , and σ_t is atomic with support $\dot{x}(t)$. Further, the measure $\sigma_{t,\Delta}^v$ must be, for small Δ and high v , nearly atomic, and the values \dot{x}_i , $i \in \mathcal{N}_{t,\Delta}$, should be close; that is, σ_t is atomic if and only if $|\dot{x}_{i+1} - \dot{x}_i| \rightarrow 0$, $i \in \mathcal{N}_{t,\Delta}$, as $v \rightarrow \infty$. The generalized curve is an ordinary curve if these conditions are satisfied a.e. on $[t_0, t_f]$.

Let f_0 be twice continuously differentiable with respect to \dot{x} in the interior of B for all $(t, x) \in [t_0, t_f] \times A$. It is possible under this assumption to put the condition on the atomicity of σ_t in terms of some requirements on f_0 itself and the vectors \dot{x}_i , $i = 1, 2, \dots, N$, for high v ; we obtain in this manner a version of a well-known theorem [18] on the existence of an ordinary curve minimizing the functional (2.1) over the class \mathcal{F} , plus some further insight into the structure of generalized curves.

Let $\{(\alpha_i, t_i, x_i, \dot{x}_i), i = 1, 2, \dots, N(v)\}$ be the set of quadruples minimizing (3.6) under the constraints (3.7). We assume throughout that the vectors \dot{x}_i , $i = 1, 2, \dots, N(v)$, are interior to B . If α_i, t_i, x_i , $i = 1, 2, \dots, N(v)$, are kept fixed, the vectors \dot{x}_i minimize the function

$$\sum_{i=1}^N \alpha_i f_0(t_i, x_i, \dot{x}_i)$$

over the vectors \dot{x}_i in the interior of B satisfying the appropriate constraints in (3.7). Since these constraints are linear, there exist multipliers λ_k , $k = 1, 2, \dots, v$, such that the vectors \dot{x}_i minimize

$$(6.4) \quad \sum_{i=1}^N \alpha_i f_0(t_i, x_i, \dot{x}_i) + \sum_{k=1}^v \lambda_k \left[\sum_{i=1}^N \alpha_i \left[\dot{x}_i \text{grad } \phi_k(t_i, x_i) + \frac{\partial \phi}{\partial t}(t_i, x_i) - \Delta \phi_k \right] \right]$$

so that, for $i = 1, 2, \dots, N(v)$, and $\alpha = \alpha_i$, $t = t_i$, $x = x_i$, the vector \dot{x}_i satisfies

$$(6.5) \quad \alpha \frac{\partial f_0}{\partial \dot{x}}(t, x, \dot{x}) + P(t, x) = 0,$$

with $P(t, x)$ a vector with polynomial entries. Let $(\tilde{t}, \tilde{x}, \tilde{\dot{x}})$ be a solution of (6.5). We consider whether $\tilde{\dot{x}}$ is a continuous function of (\tilde{t}, \tilde{x}) . If the matrix $\partial^2 f_0 / \partial \dot{x}^2$ is definite at $(\tilde{t}, \tilde{x}, \tilde{\dot{x}})$, then (6.4) has a unique continuous solution for (t, x) in a neighborhood of (\tilde{t}, \tilde{x}) ; remember that $\alpha > 0$. Since, however, the vectors \dot{x}_i , $i = 1, 2, \dots, N$, provide a minimum to (6.4) for fixed (t_i, x_i) , this Hessian matrix cannot be negative definite at any triple (t_i, x_i, \dot{x}_i) . We conclude that the solutions of (6.4) with (t, x) in a neighborhood of (t_i, x_i) are unique and continuous in (t, x) if the Hessian matrix is positive definite at (t_i, x_i, \dot{x}_i) , provided of course that the vectors \dot{x}_i are interior to B .

If the Hessian matrix is positive definite at $(t, x(t))$ for all \dot{x} in the interior of B , it is positive definite for all elements (t_i, x_i, \dot{x}_i) , $i \in \mathcal{N}_{t,\Delta}$, for $v > v_0$, say. Since $|x_{i+1} - x_i|$, $|t_{i+1} - t_i|$ tend to zero as $v \rightarrow \infty$, by continuity $|\dot{x}_{i+1} - \dot{x}_i| \rightarrow 0$ as $v \rightarrow \infty$, and σ_t is atomic. Further, those points t at which σ_t is atomic are never isolated. We have shown, in particular, the following.

THEOREM 5. *The generalized curve providing a minimum of the functional I over the set P is an ordinary curve if: (i) the vectors \dot{x}_i , $i = 1, 2, \dots, N(v)$, are interior to B for all v higher than some v_1 ; (ii) the function $f_0(t, x(t), \dot{x})$, $\dot{x} \in B$, is strongly convex in the interior of B for all $t \in [t_0, t_f]$.*

7. Conclusion. A new treatment for variational problems has been presented in this paper. Insight into the problems can be achieved by considering the (simpler) programming problems of § 3, just as was done in deriving Theorem 5 and the material preceding it.

No attempt has as yet been made to develop the potential of the method as a computational technique; considerable effort would be needed to achieve suitable, fast algorithms because of the large dimensionality of the programming problems.

Further applications and extensions of this discretizing scheme will be published elsewhere.

Appendix A. Proof of Proposition 3.

(i) Put $z = z_1$ and $c = 2$ in the definition of $\psi_{c,z}$; then the radius of D_1 is $2M^{-1/4}$, and $z_0 \in D$ by Proposition 2. Let $|z_f - z_0| = l \neq 0$, and let v_1^1 be an index such that $6M^{-1/4}(1 + M^{-1/8}) < l$. Then, for $v > v_1^1$, if $z_0 \in D_1$ then $z_f \notin D$, since the diameter of D is $4M^{-1/4}(1 + M^{-1/8})$. The right side of (4.2) with $\phi = \psi_{2,z_1}$ satisfies, by (4.5),

$$|\Delta\psi_{2,z_1} + 1| \leq 2\frac{K_2}{4}M^{-1/4}$$

for $v > v_0(2)$. Assume there is no v_1^2 such that for all $v > v_1^2$ there is at least one vector z_i , $i = 2, 3, \dots, N$, in D . Then it is possible to find an increasing sequence $\{v^s, s = 1, 2, \dots\}$ such that there are no vectors $z_i \in D$ for those v^s . For these values of v , the left side of (4.2) satisfies

$$\left| \sum_{i=1}^N \alpha_i \dot{z}_i \text{grad } \psi_{2,z_1}(z_i) \right| \leq \frac{1}{4}K_1(\Delta t)\hat{\beta}M^{-1/4},$$

which, together with the previous inequality, implies

$$\frac{1}{2}K_2M^{-1/4} \geq 1 - \frac{1}{4}K_1(\Delta t)\hat{\beta}M^{-1/4}$$

for any $v^s > \max(v_1^1, v_0(2))$, which is false. There is therefore a v_1^2 such that for all $v > v_1^2$ there is at least one vector z_i , $i = 2, 3, \dots, N$, in D . We choose, for each $v > v_1^2$, one such vector, and relabel the associated triples $(\alpha_2, z_2, \dot{z}_2)$; the triples previously labeled in this manner will occupy the places of the ones just relabeled. The vector z_2 satisfies

$$|z_2 - z_1| \leq 2M^{-1/4}(1 + M^{-1/8}) \leq 4M^{-1/4}$$

for $v > v_1 = \max(v_1^1, v_0(2), v_1^2, v)$; v is an index for which $M^{-1/8} < 1$.

(ii) We now put $j = 2$. Let $z = z_2$, $c = 2$. We must consider several cases:

(a) Let $z_0 \in D_1$, $z_1 \notin D_2$. The vector $z_f \notin D$ for $v > v_1^1$. By exactly the same argument as in (i) we show that at least one vector z_i , $i = 3, 4, \dots, N$, must be in D for sufficiently high v . We choose one such vector, relabel its triple with subindex 3, etc; there is v_2 such that $|z_3 - z_2| \leq 4M^{-1/4}$ for $v > v_2$.

(b) Let $z_0 \notin D$, $z_f \notin D$. Put in (4.2) $\phi = \psi_{2,z_2} + \psi_{1,z_0}$, and call $\underline{D}_1, \underline{D}_2, \underline{D}$ the sets associated with ψ_{1,z_0} . For $v > v_1^1$, $z_f \notin \underline{D}$. Also, $z_2 \notin \underline{D}$, since $z_0 \notin D$, and the radius of \underline{D} is one-half of the radius of D . The right side of (4.2) satisfies

$$|\Delta\psi_{2,z_2} + \Delta\psi_{1,z_0} + 1| \leq 2\frac{K_2}{4}M^{-1/4} + 2K_2M^{-1/4}$$

for $v > \max(v_0(2), v_0(1)) = v_0(2)$. Assume $z_1 \notin D_2$ (the case for which $z_1 \in D_2$ will be treated below), and, temporarily, that no z_i , $i = 1, 2, \dots, N$, is in \underline{D}_2 for any v . Assume that no index v_2^1 exists such that for all $v > v_2^1$ there is at least one vector z_i , $i = 3, 4, \dots, N$, in D . Using the same argument as in (i), the left side of (4.2) satisfies along an increasing sequence $\{v^s\}$,

$$\left| \sum_{i=1}^N \alpha_i \dot{z}_i \text{grad}(\psi_{2,z_2} + \psi_{1,z_0})(z_i) \right| \leq (\tfrac{1}{4}K_1 + K_1)(\Delta t)\hat{\beta}M^{-1/4},$$

which implies

$$(A.1) \quad \tfrac{1}{2}K_2M^{-1/4} \geq \tfrac{1}{5} - \tfrac{1}{4}K_1(\Delta t)\hat{\beta}M^{-1/4}$$

for all $v^s > \max(v_1^1, v_0(2))$, which is false. We proceed as in (a), and find z_3 such that $|z_3 - z_2| \leq 4M^{-1/4}$, $v > v_3$.

If there are some z_i 's in \underline{D} , $i = 1, 2, \dots, N$, for some values of v , we replace ψ_{1,z_0} by $\psi_{1/4,z_0}$ and then no z_i 's are in the new \underline{D}_2 . The argument is as before, with (A.1) replaced by

$$(A.2) \quad \tfrac{1}{2}K_2M^{-1/4} \geq \tfrac{1}{65} - \tfrac{1}{4}K_1(\Delta t)\hat{\beta}M^{-1/4}.$$

As above, we find z_3 such that $|z_3 - z_2| \leq 4M^{-1/4}$, $v > v_4$.

(c) Suppose $z_0 \in D_2$, $z_1 \notin D_2$. We can, by taking $c = 1$, $z = z_2$, construct a new neighborhood \hat{D} of z_2 such that $z_0 \notin \hat{D}$ for $v > v'$ (v' is an index for which $M^{-1/8} < 1$). This implies that $z_f \notin \hat{D}$ for $v > \max(v_1^1, v')$; the diameter of \hat{D} is $2M^{-1/4}(1 + M^{-1/8}) < 4M^{-1/4}$ for $v > v'$, and if $z_0 \in D_2$, $|z_0 - z_2| \leq 4M^{-1/4}$, $v > v'$, from which it follows that all points in \hat{D} satisfy $|z - z_0| < 6M^{-1/4}$ for $v > v'$. However, $|z_f - z_0| = l > 6M^{-1/4}$ for $v > v_1^1$, from which our assertion follows.

If $z_1 \notin \hat{D}_2$, this case is like (b), with ψ_{1,z_2} and $\psi_{1/2,z_0}$ instead of ψ_{2,z_2} and ψ_{1,z_0} . The inequality (A.1) is replaced by

$$\tfrac{1}{2}K_2M^{-1/4} \geq \tfrac{1}{18} - \tfrac{1}{4}K_1(\Delta t)\hat{\beta}M^{-1/4}$$

and (A.2) by the inequality corresponding to ψ_{1,z_2} and $\psi_{1/8,z_0}$, that is,

$$\tfrac{1}{2}K_2M^{-1/4} \geq \tfrac{1}{260} - \tfrac{1}{4}K_1(\Delta t)\hat{\beta}M^{-1/4}.$$

If, finally, $z_1 \in \hat{D}_2$, we use $\psi_{1/2, z_2}$, $\psi_{1/4, z_0}$, etc; we obtain two new inequalities. In any case, we can choose z_3 so that $|z_3 - z_2| \leq 4M^{-1/4}$, $v > v_5$, with v_5 higher than or equal to the indices associated with each of the four inequalities.

(d) Let $z_0 \in D_1, z_1 \in D_2$. Construct \hat{D} as in (c). Then $z_1 \notin \hat{D}$ for $v > v$. If $z_0 \in \hat{D}_1, z_f \notin \hat{D}$, for $v > v_1^1$. The proof proceeds now as in (a). We find z_3 such that $|z_3 - z_2| \leq 4M^{-1/4}$, $v > v_6$.

(e) There are two more cases, corresponding to (b) and (c) but with $z_1 \in D_2$, and four more cases, in which z_f takes the place of z_0 . We treat these cases similarly, and obtain $|z_3 - z_2| \leq 4M^{-1/4}$ for indices v_7, \dots, v_{12} respectively. Of course, the vectors labeled z_3 in each of these 12 cases may all be different. Choose for each v the vector z_3 which corresponds to the particular case prevailing at this value of the index; and put $\gamma = \max(v_2, \dots, v_{12})$. Then we have $|z_3 - z_2| \leq 4M^{-1/4}$, $v > \gamma$.

(iii) The rest of the proof proceeds in the same manner. None of the basic inequalities in (ii) depend on the subindex, or on the fact that only two vectors, z_1 and z_2 , have been relabeled previously. If we were considering the situation in which the vectors z_1, \dots, z_j have been relabeled already, the change from (a) to (d), for instance, would be made in the same manner if several vectors from the set $\{z_1, \dots, z_{j-1}\}$ were in D_2 . Then we conclude that the triples can be ordered so that $|z_{j+1} - z_j| \leq 4M^{-1/4}$, $v > \gamma, j = 2, 3, \dots, N$. Finally put $\tilde{v} = \max(v_1, \gamma)$; the proposition follows.

Appendix B. Some properties of the curves constructed in § 5. It should be noted that some of the curves $x_v(\cdot)$ may take values outside the set A ; this may be the case, for instance, if some of the corresponding vectors x_i are on the boundary of A . Of course, as $v \rightarrow \infty$, the curves $x_v(\cdot)$ will either take values in A , or, if not, the distance from $\{x : x = x_v(t), t \in [t_0, t_f]\}$ to A will tend to zero. If this latter situation is considered unacceptable, the curves $x_v(\cdot)$ may be replaced by the curves $\tilde{x}_v(\cdot)$, defined by

$$\tilde{x}_v(t) = 1 \left(- \frac{4KM^{-1/4}}{a^*} \right) x_v(t),$$

$t \in [t_0, t_f]$, $v > v_m$, where a^* is the smallest side of the polygon A . Then the graph of $\tilde{x}_v(\cdot)$ is in $[t_0, t_f] \times A$; note that by Theorem 3, no point of $x_v(\cdot)$ for $v > v_m$ can be at a distance higher than $2KM^{-1/4}$ from some x_i , and all the x_i 's are in A . The proof of the equality (5.1) for the sequence $\{\tilde{x}_v(\cdot)\}$ follows with minor changes the one given above for $\{x_v(\cdot)\}$.

The curves in the sequence $\{x_v(\cdot)\}$ are not, in general, in \mathcal{F} , because $x_v(t_0) = x_1$, $x_v(t_f) = y_N$; of course, as $v \rightarrow \infty$, $x_1 \rightarrow x_0$ and $y_N \rightarrow x_f$. The definition of the curves $x_v(\cdot)$ can be modified, by prescribing $x_v(t_0) = x_0$, $x_v(t_f) = x_f$, while leaving the rest unchanged; that is, we join $x_0 \rightarrow y_1 \rightarrow \dots \rightarrow y_{N-1} \rightarrow x_f$, assign t_0 to x_0 , $t_0 + \alpha_1$ to y_1 , etc. The proof of (5.1) for the new sequence follows with minor changes the one given above. Theorem 4 can then be modified to state that the functional I attains its minimum over P at a generalized curve which is the weak*—limit of a sequence of curves in \mathcal{F} .

REFERENCES

- [1] L. C. YOUNG, *On approximation by polygons in the calculus of variations*, Proc. Roy. Soc. Ser. A, 141 (1933), pp. 325–341.
- [2] ———, *Generalized curves and the existence of an attained absolute minimum in the calculus of variations*, C.R. Soc. Sci. et Lettres, Varsovie, Cl. III, 30 (1937), pp. 212–234.
- [3] ———, *Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, 1969.
- [4] E. J. MCSHANE, *Generalized curves*, Duke Math. J., 6 (1940), pp. 513–536.
- [5] J. WARGA, *Relaxed variational problems*, J. Math. Anal. Appl., 4 (1962), pp. 111–128.
- [6] ———, *Necessary conditions for minimum in relaxed variational problems*, Ibid., 4 (1962), pp. 129–145.
- [7] T. WAZEWSKI, *Systèmes de commande et equations au contingent*, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys., 9 (1961), pp. 151–155.
- [8] R. V. GAMKRELIDZE, *Optimal sliding regimes*, Dokl. Akad. Nauk SSSR, 143 (1962), pp. 1243–1245.
- [9] L. C. YOUNG, *Necessary conditions in the calculus of variations*, Acta Math., 69 (1938), pp. 239–258.
- [10] E. J. MCSHANE, *Necessary conditions for generalized-curve problems of the calculus of variations*, Duke Math. J., 7 (1940), pp. 1–27.
- [11] ———, *Existence theorems for Bolza problems in the calculus of variations*, Ibid., 7 (1940), pp. 28–61.
- [12] ———, *Relaxed controls and variational problems*, this Journal, 5 (1967), pp. 438–485.
- [13] A. TUROWICZ, *Sur trajectoires et quasitrajectoires des systèmes de commande nonlinéaires*, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys., 10 (1962), pp. 529–531.
- [14] G. GABASOV, *Necessary conditions for optimality of singular control*, Engrg. Cybernetics, 1968, pp. 28–37.
- [15] A. A. BOLONKIN, *Special extrema in optimal control problems*, Ibid., 1969, pp. 170–183.
- [16] F. TREVES, *Topological Vector Spaces, Distributions and Kernels*, Academic Press, New York, 1967.
- [17] P. C. ROSENBLOOM, *Quelques classes de problèmes extremaux*, Bull. Soc. Math. France, 79 (1951), pp. 1–58, 80 (1952), pp. 183–216.
- [18] L. TONELLI, *Fondamenti di calcolo delle variazione*, Zanichelli, Bologna, 1923.

STABILITY AND THE INFINITE-TIME QUADRATIC COST PROBLEM FOR LINEAR HEREDITARY DIFFERENTIAL SYSTEMS*

M. C. DELFOUR,[†] C. McCALLA[‡] AND S. K. MITTER[¶]

Abstract. This paper studies the infinite-time quadratic cost control problem for a general class of linear autonomous hereditary differential systems. It uses an approach which clarifies the system-theoretic relationship between stabilizability, stability and existence of a solution of an associated operator equation of Riccati type. For this purpose the stability problem is studied and an operator equation of the Lyapunov type is derived. In both cases we obtain equations which characterize the kernels of the Lyapunov and the Riccati equations.

1. Introduction. In a previous paper (cf. Delfour–Mitter [8]) we have studied the quadratic cost optimal control problem over a finite time interval for a general class of linear hereditary differential systems. In particular we have characterized the optimal controller as a linear feedback controller acting on the “state” of the system. The feedback operator is determined by the solution of an operational differential equation of Riccati type. The main objective of the present paper is to study the infinite-time quadratic cost problem for a general class of linear autonomous hereditary differential systems. In undertaking this study we insist on an approach which clarifies the system-theoretic relationship between controllability, stabilizability, stability and existence of a solution of an associated operator equation of Riccati type.

For systems described by ordinary differential equations the infinite-time quadratic cost problem is well-studied (cf. R. W. Brockett [1], R. E. Kalman [13], J. C. Willems [22], W. M. Wonham [23]). This problem has been studied for certain classes of infinite-dimensional systems. J. L. Lions [15] has studied this problem for abstract evolution equations of parabolic type and given a complete solution to the problem. Lukes and Russell [16] have studied this problem for abstract evolution equations of the type

$$(1.1) \quad \begin{aligned} \frac{dx(t)}{dt} &= Ax(t) + Bu(t), \\ x(0) &= x_0 \in \mathcal{D}(A), \end{aligned}$$

where A is an unbounded spectral operator (cf. Dunford and Schwartz [11]) and B is also an unbounded operator satisfying certain conditions. Lukes and Russell also allow unbounded operators in the cost function. Using an approach originally

* Received by the editors January 17, 1972, and in final revised form October 18, 1973.

[†] Centre de Recherches Mathématiques, Université de Montréal, Montréal 101, Canada. The research of this author was supported in part by the National Research Council (Canada) under Grant A 8730 at the Centre de Recherches Mathématiques, Université de Montréal.

[‡] Mathematics Department, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

[¶] Electrical Engineering Department, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139. The research of this author was supported in part by the National Science Foundation under Grant GK-25781 and by the Air Force Office of Scientific Research under Grant 72-2273 both at the Electronic Systems Laboratory, M.I.T.; a portion of this work was done while this author was U.K. Science Research Council Senior Visiting Fellow at Imperial College during June–July, 1972.

due to R. E. Kalman [13] they obtain an operational differential equation of Riccati type to characterize the time-varying feedback gain in the finite time case. They also show that under an appropriate stabilizability hypothesis the solution to the infinite-time quadratic cost problem can be obtained in feedback form, where the "feedback gain" is characterized by the solution of an operator equation of quadratic type. The same problem has also been studied by R. Datko [4]. Unfortunately, R. Datko [4] does not characterize the solution as a feedback controller acting on the "state" of the system.

It is felt that the contributions of the present paper are the following:

(i) We present a complete detailed solution to the infinite-time quadratic cost problem for a general class of linear hereditary differential systems. Other than the parabolic case solved by J. L. Lions [15], this appears to be the only other case (so far) where the problem can be solved in a way which is satisfactory from the system-theoretic point of view (that is, no ad hoc mathematical assumptions need to be made).

(ii) The approach we use here is different from that of Lukes and Russell [16] as well as R. Datko [3], [4] and constitutes a synthesis of the work of J. L. Lions [15] and Delfour and Mitter [6], [7], [8].

(iii) The detailed results we obtain exploit the structure of hereditary differential systems in an essential way.

(iv) It gives rigorous derivations of earlier incomplete results of Ross and Flüge-Lotz [19] for a more specialized problem.

The results on the equations for the kernel of the solution of the Lyapunov equation have been announced in 1972 (cf. Delfour [5]).

2. Notation, terminology and preliminary definitions. Let \mathbb{R} be the field of all real numbers and let $a > 0$ be given.

Let X and Y be real Hilbert spaces with norms $|\cdot|_X$, $|\cdot|_Y$ and inner products $(\cdot, \cdot)_X$ and $(\cdot, \cdot)_Y$ respectively.

Let $\mathcal{L}^2(-a, 0; X)$ be the vector space of all m -measurable (m denoting the complete Lebesgue measure on \mathbb{R}) maps $[-a, 0] \rightarrow X$ which are square integrable and let $L^2(-a, 0; X)$ denote the natural Hilbert space associated with $\mathcal{L}^2(-a, 0; X)$ with norm $\|\cdot\|_2$. Consider the space $\mathcal{L}^2(-a, 0; X)$ endowed with the seminorm

$$(2.1) \quad \|f\|_{M^2} = [|f(0)|_X^2 + \|f\|_2^2]^{1/2}.$$

The quotient space of $\mathcal{L}^2(-a, 0; X)$ by the linear subspace of all f such that $\|f\|_{M^2} = 0$ is denoted by $M^2(-a, 0; X)$. $M^2(-a, 0; X)$ endowed with the norm (2.1) and inner product

$$(2.2) \quad (f, g)_{M^2} = (f(0), g(0))_X + \int_{-a}^0 (f(\theta), g(\theta))_X d\theta$$

is a Hilbert space isometrically isomorphic to $X \times L^2(-a, 0; X)$ endowed with the norm

$$(2.3) \quad \|h\| = \left[|h^0|_X^2 + \int_{-a}^0 |h^1(\theta)|_X^2 d\theta \right]^{1/2}$$

and inner product

$$(2.4) \quad (h, k) = (h^0, k^0)_X + \int_{-a}^0 (h^1(\theta), k^1(\theta))_X d\theta.$$

The isomorphism is denoted by κ , where $\kappa(h) = (h(0), h)$. For simplicity we shall often identify h and the pair (h^0, h^1) . For the motivation in introducing M^2 , see Delfour and Mitter [6].

For all $t \in [0, \infty)$, we denote by $W^{1,2}(0, t; X)$ the vector space of all absolutely continuous maps $[0, t] \rightarrow X$ with a distributional derivative Dx in $L^2(0, t; X)$. $W^{1,2}(0, t; X)$ endowed with the norm

$$(2.5) \quad \|x\|_{W^{1,2}} = \left[\int_0^t (|x(s)|_X^2 + |Dx(s)|_X^2) ds \right]^{1/2}$$

is a Hilbert space.

We denote by $L_{\text{loc}}^2(0, \infty; X)$ the Fréchet space of measurable maps $[0, \infty) \rightarrow X$ which are square integrable on every compact subset of $[0, \infty)$. $W_{\text{loc}}^{1,2}(0, \infty; X)$ denotes the Fréchet space of all absolutely continuous maps $[0, \infty) \rightarrow X$ with derivatives in $L_{\text{loc}}^2(0, \infty; X)$, and $C_{\text{loc}}(0, \infty; X)$ denotes the Fréchet space of all continuous maps $[0, \infty) \rightarrow X$.

Let $\mathcal{L}(X, Y)$ denote the real Banach space of all continuous linear maps $\Lambda: X \rightarrow Y$ endowed with the natural norm $\|\Lambda\|$. The adjoint of Λ in $\mathcal{L}(X, Y)$ will be denoted by $\Lambda^* \in \mathcal{L}(Y, X)$. When $X = Y$, we write $\mathcal{L}(X)$ instead of $\mathcal{L}(X, X)$. $\Lambda \in \mathcal{L}(X)$ will be said to be self-adjoint if $\Lambda = \Lambda^*$. A self-adjoint Λ will be said to be positive and written $\Lambda \geq 0$ if $(\Lambda x, x) \geq 0$ for all $x \in X$ and positive definite and written $\Lambda > 0$ if $(\Lambda x, x) > 0, x \neq 0$. The identity in $\mathcal{L}(X)$ is denoted by I_X .

For an operator $\Lambda \in \mathcal{L}(M^2)$ we can exploit the isomorphism between M^2 and $X \times L^2$ to decompose Λ into a matrix of operators

$$(2.6) \quad \begin{pmatrix} \Lambda^{00} & \Lambda^{01} \\ \Lambda^{10} & \Lambda^{11} \end{pmatrix},$$

where $\Lambda^{00} \in \mathcal{L}(X)$, $\Lambda^{01} \in \mathcal{L}(L^2(-a, 0; X), X)$, $\Lambda^{10} \in \mathcal{L}(X, L^2(-a, 0; X))$ and $\Lambda^{11} \in \mathcal{L}(L^2(-a, 0; X))$ are defined in the obvious way.

Let $A: \mathcal{D}(A) \rightarrow X$ be a closed linear operator with dense domain $\mathcal{D}(A)$ in X . The operator A is said to be bounded from below (resp. above) by $\alpha \in \mathbb{R}$ if for all $x \in \mathcal{D}(A)$, $(Ax, x) \geq \alpha \|x\|^2$ (resp. $(Ax, x) \leq \alpha \|x\|^2$).

3. Summary of previous results. Let $N \geq 1$ be an integer, let $a > 0$ and $-a = \theta_N < \dots < \theta_1 < \theta_0 = 0$ be real numbers, let $X = \mathbb{R}^n$ be the Euclidean real Hilbert space of finite dimension n and let U be an arbitrary real Hilbert space.

Consider the autonomous hereditary differential system

$$(L) \quad \begin{aligned} \frac{dx}{dt}(t) &= A_{00}x(t) + \sum_{i=1}^N A_i \begin{cases} x(t + \theta_i), & t + \theta_i \geq 0 \\ h^1(t + \theta_i), & t + \theta_i < 0 \end{cases} \\ &+ \int_{-a}^0 A_{01}(\theta) \begin{cases} x(t + \theta), & t + \theta \geq 0 \\ h^1(t + \theta), & t + \theta < 0 \end{cases} d\theta \\ &+ Bv(t), \quad \text{a.e. in } [0, \infty), \\ x(0) &= h^0, \quad h = (h^0, h^1) \quad \text{in } M^2(-a, 0; X), \end{aligned}$$

where A_{00}, A_i ($i = 1, 2, \dots, N$) are elements of $\mathcal{L}(X)$, $A_{01} \in L^\infty(-a, 0; \mathcal{L}(X))$, $v \in L^2_{\text{loc}}(0, \infty; U)$ and $B \in \mathcal{L}(U, X)$.

It was shown in Delfour and Mitter [8], [10] that the system (L) can be equivalently described by an evolution equation in $M^2(-a, 0; X)$. For this purpose we define the *state at time t* as an element

$$(3.1) \quad \tilde{x}(t; h, v) = (\tilde{x}(t; h, v)^0, \tilde{x}(t; h, v)^1) \in M^2(-a, 0; X)$$

in terms of $h = (h^0, h^1)$ and the solution $x(\cdot; h, v)$ of system (L):

$$(3.2) \quad \tilde{x}(t; h, v)^0 = x(t; h, v), \quad \tilde{x}(t; h, v)^1(\theta) = \begin{cases} x(t + \theta; h, v), & t + \theta \geq 0 \\ h^1(t + \theta) & , \text{ otherwise} \end{cases}$$

We define

$$(3.3) \quad V = \{(h(0), h) | h \in W^{1,2}(-a, 0; X)\}$$

and $\tilde{A}_0: V \rightarrow X$, $\tilde{A}_1: V \rightarrow L^2(-a, 0; X)$ and $\tilde{A}: V \rightarrow M^2(-a, 0; X)$ as follows:

$$(3.4) \quad \tilde{A}_0 h = A_{00}h(0) + \sum_{i=1}^N A_i h(\theta_i) + \int_{-a}^0 A_{01}(\theta)h(\theta) d\theta,$$

$$(3.5) \quad (\tilde{A}_1 h)(\theta) = \frac{dh}{d\theta}(\theta),$$

and

$$(3.6) \quad [\tilde{A}h]^0 = \tilde{A}_0 h, \quad [\tilde{A}h]^1 = \tilde{A}_1 h.$$

Let $v(t) = 0$ in $[0, \infty)$ in (L). We then have (cf. Delfour and Mitter [8], [10]) the following.

THEOREM 3.1. *The map $t \mapsto \tilde{x}(t; h, 0)$ given by (3.1) generates a one-parameter semigroup $\{\tilde{\Phi}(t)\}$ in $\mathcal{L}(M^2)$ satisfying the following properties:*

- (i) *for all h in M^2 , $t \mapsto \tilde{\Phi}(t)h: [0, \infty) \rightarrow M^2$ is continuous;*
- (ii) *$\tilde{\Phi}(0) = I_{M^2}$;*
- (iii) *for $t \geq a$, $\tilde{\Phi}(t)$ is compact (i.e., maps bounded sets into relatively compact sets);*
- (iv) *for all h in V the map $t \mapsto \tilde{\Phi}(t)h: [0, \infty) \rightarrow V$ is continuous;*
- (v) *the operator \tilde{A} defined by (3.4)–(3.6) is the infinitesimal generator of the semigroup $\tilde{\Phi}(t)$.*

Now define the operator $\tilde{B} \in \mathcal{L}(U, M^2(-a, 0; X))$ as

$$(3.7) \quad \tilde{B}u = (Bu, 0).$$

Consider the controlled evolution equation

$$(3.8) \quad \begin{aligned} \frac{d\tilde{x}}{dt}(t) &= \tilde{A}\tilde{x}(t) + \tilde{B}u(t), \\ \tilde{x}(0) &= h. \end{aligned}$$

We then have the following theorem.

THEOREM 3.2.

(i) For all h in V and v in $L^2_{\text{loc}}(0, \infty; U)$, system (\tilde{L}) has a unique solution in

$$(3.8) \quad W_{\text{loc}}(0, \infty; V, M^2) = \{z \in L^2_{\text{loc}}(0, \infty; V) | Dz \in L^2_{\text{loc}}(0, \infty; M^2)\}$$

which coincides with the state $\tilde{x}(\cdot; h, v)$ constructed from h and $x(\cdot; h, v)$.

(ii) The map $(h, v) \mapsto \Lambda(h, u) = \tilde{x}(\cdot; h, v): V \times L^2_{\text{loc}}(0, \infty; U) \rightarrow W_{\text{loc}}(0, \infty; V, M^2)$ is linear and continuous when V is endowed with the $W^{1,2}$ -topology; it can be lifted to a unique continuous linear map $\tilde{\Lambda}: M^2 \times L^2_{\text{loc}}(0, \infty; U) \rightarrow C_{\text{loc}}(0, \infty; M^2)$.

Consider the control system (L) and fix the final time $T \in (0, \infty)$ and the initial time t in $[0, T)$. With a pair (h, v) we associate the cost function

$$(3.9) \quad J_T^t(v, h) = \int_t^T [(x(s; h, v), Qx(s; h, v)) + (v(s), Nv(s))] ds,$$

where $Q \in \mathcal{L}(X)$ self-adjoint, $Q \geq 0$, $N \in \mathcal{L}(U)$ self-adjoint, $(u, Nu) \geq c|u|^2$, $c > 0$.

Consider the optimal control problem of minimizing (3.9) in the interval $[t, T]$. For each h , it can be shown that there exists a unique u in $L^2(t, T; U)$ which minimizes (3.9) over all v in $L^2(t, T; U)$. We can then show that there exists a unique operator $\Pi_T(t) \in \mathcal{L}(M^2)$ which is self-adjoint and positive such that

$$(3.10) \quad (h, \Pi_T(t)h)_{M^2} = \min \{J_T^t(v, h) | v \in L^2(t, T; U)\}.$$

Moreover the optimal control is given by

$$(3.11) \quad u(s) = -N^{-1}\tilde{B}^*\Pi_T(s)\tilde{x}(s; h),$$

where $\tilde{x}(\cdot; h)$ is the solution of

$$(3.12) \quad \begin{aligned} \frac{dy(s)}{ds} &= [\tilde{A} - \tilde{B}N^{-1}\tilde{B}^*\Pi_T(s)]y(s) \quad \text{a.e. in } [t, T], \\ y(t) &= h. \end{aligned}$$

The operator $\Pi_T(s)$ can be shown to satisfy an operator differential equation of Riccati type which (when interpreted appropriately) has a unique solution in $[0, T]$ (cf. Delfour and Mitter [8]).

In the sequel we shall abbreviate $M^2(-a, 0; X)$ by M^2 .

4. Formulation of the infinite-time problem. We now associate with the control system (L) (or equivalently \tilde{L}) the quadratic cost J_∞ which is equal to the quadratic cost (3.9) where $T = \infty$ and $t = 0$. Our objective is to study the problem:

$$(4.1) \quad \text{Minimize } J_\infty(v, h) \quad \text{over all } v \in L^2_{\text{loc}}(0, \infty; U).$$

Our main result may be summarized as follows: Under certain stabilizability hypotheses for each $h \in M^2(-a, 0; X)$, there exists a unique $u \in L^2_{\text{loc}}(0, \infty; U)$ which minimizes $J_\infty(v, h)$ over all $v \in L^2_{\text{loc}}(0, \infty; U)$. Moreover, the minimizing control u can be expressed in "feedback form" in terms of an operator Π for which an operator Riccati equation can be obtained. Under further hypotheses on Q , the resulting closed-loop control is also stable.

The theory is thus as complete as the theory for the corresponding ordinary differential equation case.

5. Solution of the infinite-time problem. The solution to the infinite-time problem proceeds in three parts:

(i) We first have to make sure that the problem is well-posed in the sense that there exists a constant $c > 0$ and for each h a control v_h such that the corresponding cost $J_\infty(v_h, h)$ is bounded by $c\|h\|_{M^2}^2$. This naturally leads to a study of the stability and stabilizability of linear hereditary systems.

(ii) We then study the behavior of $J_T^t(v, h)$ and the feedback operator $\Pi_T(t)$ as $T \rightarrow \infty$. We show in particular that $\Pi_T(t)$ converges to an operator Π .

(iii) Finally we characterize Π and study the stability of the resulting closed-loop system.

5.1. Stability. In this section we shall denote by $x(s; h)$ the solution $x(s; h, 0)$ of (L).

DEFINITION 5.1. The uncontrolled system (L) is said to be L^2 -stable if

$$(5.1) \quad \lim_{t \rightarrow \infty} \int_0^t (x(s; h), x(s; h))_X ds < \infty \quad \forall h \in M^2.$$

By virtue of the choice of M^2 as the space of initial conditions it is easy to show that (5.1) is equivalent to

$$(5.2) \quad \lim_{t \rightarrow \infty} \int_0^t (\tilde{x}(s; h), \tilde{x}(s; h))_{M^2} ds < \infty \quad \forall h \in M^2.$$

DEFINITION 5.2. An operator $R \in \mathcal{L}(M^2)$ is said to be *positive definite on X* if

$$(5.3) \quad (h^0, R^{00}h^0)_X > 0 \quad \forall h^0 \neq 0,$$

where $R^{00} \in \mathcal{L}(X)$ is defined by

$$R^{00}h^0 = [R(h^0, 0)]^0 \quad \forall h^0 \in X.$$

Using the techniques of R. Datko [2] we can state the following equivalent conditions for L^2 -stability.

THEOREM 5.3. Let $R \geq 0$ in $\mathcal{L}(M^2)$ and $Q > 0$ in $\mathcal{L}(X)$ be given. The following statements are equivalent:

- (i) (L) is L^2 -stable.
- (ii) For all h in M^2 ,

$$(5.4) \quad \lim_{t \rightarrow \infty} \int_0^t [(R\tilde{x}(s; h), \tilde{x}(s; h))_{M^2} + (Qx(s; h), x(s; h))_X] ds < \infty.$$

- (iii) There exists a self-adjoint operator $B \geq 0$ in $\mathcal{L}(M^2)$ such that

$$(5.5) \quad (\tilde{A}h, Bk) + (h, B\tilde{A}k) + (h, \tilde{I}k) = 0 \quad \forall h, \forall k \in V,$$

where

$$(5.6) \quad (\tilde{I}h) = (h^0, 0).$$

- (iv) There exists a self-adjoint operator $B \geq 0$ such that

$$(5.7) \quad (\tilde{A}h, Bk) + (h, B\tilde{A}k) + (h, Rk) + (h, \tilde{Q}k) = 0 \quad \forall h, \forall k \in V,$$

where

$$\tilde{Q}h = (Qh^0, 0).$$

(v) There exist $\tilde{\omega} > 0$ and $\tilde{M} \geq 1$ such that

$$(5.8) \quad \|\tilde{x}(t; h)\|_{M^2} \leq \tilde{M} \exp(-\tilde{\omega}t) \|h\|_{M^2} \quad \forall t \geq 0.$$

(vi) There exist $\omega > 0$ and $M > 1$ such that

$$(5.9) \quad \|x(t; h)\|_X \leq M \exp(-\omega t) \|h\|_{M^2} \quad \forall t \geq 0.$$

(vii) There exists $\alpha < 0$ such that the spectrum $\sigma(\tilde{A})$ of \tilde{A} lies entirely in $\{\lambda \in \mathbb{C} \mid \operatorname{Re} \lambda \leq \alpha\}$, where \mathbb{C} is the field of all complex numbers, $\sigma(\tilde{A}) = \{\lambda \in \mathbb{C} \mid \det \Delta(\lambda) = 0\}$ and $\det \Delta(\lambda)$ is the determinant of the matrix

$$(5.10) \quad \Delta(\lambda) = \lambda I - \sum_{i=1}^N A_i \exp(\lambda \theta_i) - \int_{-a}^0 A_{01}(\theta) \exp(\lambda \theta) d\theta.$$

Proof. The equivalence of conditions (i) through (vi) can be easily proved by using the results and techniques of R. Datko [2] and the remark following Definition 5.1. As for condition (vii) it is a straightforward application of the results of J. K. Hale [12] with the space $M^2(-a, 0; X)$ in place of the space $C(-a, 0; X)$. \square

Remark. (i) Equation (5.5) can be rewritten as an equation in $\mathcal{L}(V, V^*)$ (V^* , the topological dual of V):

$$(5.11) \quad \tilde{A}^*B + B\tilde{A} + \tilde{I} = 0.$$

This is the generalization of Lyapunov's equation in the finite-dimensional case. This condition is much sharper than R. Datko's condition (see [2])

$$(5.12) \quad 2(B\tilde{A}x, x) = -|x|^2 \quad \forall x \in V,$$

but obviously equivalent.

(ii) Notice also that a straightforward application of R. Datko's results (see [3]) would have yielded the Lyapunov equation

$$(5.13) \quad \tilde{A}^*B + B\tilde{A} + I = 0,$$

where I is the identity in $\mathcal{L}(M^2)$, or equivalently

$$(5.14) \quad \tilde{A}^*B + B\tilde{A} + Q = 0$$

for some positive self-adjoint Q in $\mathcal{L}(M^2)$ which is *bounded below by some positive nonzero constant*. Conditions (iii) and (iv) are different and make use of the special structure of hereditary systems (cf. remark following Definition 5.1). It is this *subtle difference* that will enable us to solve the infinite-time quadratic cost problem.

In Proposition 5.4 and Theorem 5.5 we further characterize the solutions of equations (5.5) and (5.7).

PROPOSITION 5.4. *Let the hypotheses of Theorem 5.3 be true. If equation (5.5) resp. (5.7) has a positive self-adjoint solution B in $\mathcal{L}(M^2)$, it is unique and for all h and k in M^2 ,*

$$(5.15) \quad (Bh, k)_{M^2} = \int_0^\infty (x(s; h), x(s; k))_X ds$$

$$(5.16) \quad (\text{resp. } (Bh, k)_{M^2} = \int_0^\infty ([R + \tilde{Q}]\tilde{x}(s; h), \tilde{x}(s; k))_{M^2} ds)$$

and B is positive definite on X .

Proof. We prove the proposition only for equation (5.5).

(i) Let B_1 and B_2 be two solutions of (5.5) and let $D = B_2 - B_1$. Then for all h and k in V ,

$$(\tilde{A}h, Dk)_{M^2} + (Dh, \tilde{A}k)_{M^2} = 0.$$

Thus for all $t \geq 0$ and h and k in V ,

$$(5.17) \quad (\tilde{x}(t; h), D\tilde{x}(t; k))_{M^2} = (h, Dk)_{M^2}.$$

Since the system is L^2 -stable, the left-hand side of (5.17) is 0.

(ii) Similarly from equation (5.5) we obtain for all $t \geq 0$, h and k in V ,

$$(\tilde{A}\tilde{x}(t; h), B\tilde{x}(t; k)) + (B\tilde{x}(t; h), \tilde{A}\tilde{x}(t; k)) + (x(t; h), x(t; k)) = 0.$$

This yields

$$(h, Bk) = \int_0^t (x(s; h), x(s; k)) ds + (\tilde{x}(t; h), B\tilde{x}(t; k))$$

and since the system is L^2 -stable, $\tilde{x}(t; h) \rightarrow 0$ and we obtain (5.15) as t goes to infinity.

(iii) Finally for all $h^0 \neq 0$ in X ,

$$(B^{00}h^0, h^0)_X = \int_0^\infty |x(s; (h^0, 0))|_X^2 ds > 0$$

since the map $s \mapsto x(s; (h^0, 0))$ is continuous and $x(0; (h^0, 0)) = h^0$. \square

For linear hereditary differential systems we can exploit the particular structure of the system to further characterize the solution of Lyapunov's equation (5.5).

THEOREM 5.5. *Let $B \geq 0$ in $\mathcal{L}(M^2)$ be the solution of (5.5) in condition (iii) of Theorem 5.3. It is completely characterized by its matrix of operators*

$$(5.18) \quad \begin{bmatrix} B^{00} & B^{01} \\ B^{10} & B^{11} \end{bmatrix}, \quad \begin{array}{l} B^{00} \in \mathcal{L}(X), \quad B^{01} \in \mathcal{L}(L^2(-a, 0; X), X), \\ B^{10} \in \mathcal{L}(X, L^2(-a, 0; X)), \quad B^{11} \in \mathcal{L}(L^2(-a, 0; X)). \end{array}$$

B^{00} is characterized by the equation

$$(5.19) \quad \begin{aligned} B^{00}A_{00} + A_{00}^*B^{00} + B^{10}(0) + B^{10}(0)^* + I &= 0, \\ B^{00} &= (B^{00})^* \geq 0. \end{aligned}$$

B^{10} is characterized in the following way:

$$(5.20) \quad (B^{10}h^0)(\alpha) = B^{10}(\alpha)h^0,$$

where the map

$$(5.21) \quad \alpha \mapsto B^{10}(\alpha): [-a, 0] \rightarrow \mathcal{L}(X)$$

is piecewise absolutely continuous with jumps at $\alpha = \theta_i$ of height $A_i^*B^{00}$, $i = 1, \dots$,

$N - 1$. Moreover the map (5.21) is itself characterized by the differential equation

$$(5.22) \quad \frac{dB^{10}}{d\alpha}(\alpha) = B^{10}(\alpha)A_{00} + A_{01}(\alpha)^*B^{00} + \sum_{i=1}^{N-1} A_i^*B^{00}\delta(\alpha - \theta_i) + B^{11}(\alpha, 0), \quad a.e. \text{ in } [-a, 0],$$

$$B^{10}(-a) = A_N^*B^{00},$$

where $\delta(\alpha - \theta_i)$ is the δ -function at $\alpha = \theta_i$.

B^{01} is obtained from B^{10} :

$$(5.23) \quad B^{01}h^1 = \int_{-a}^0 B^{10}(\alpha)^*h^1(\alpha) d\alpha.$$

B^{11} is characterized in the following way:

$$(5.24) \quad (B^{11}h^1)(\alpha) = \int_{-a}^0 B^{11}(\alpha, \beta)h^1(\beta) d\beta,$$

where the map

$$(5.25) \quad (\alpha, \beta) \mapsto B^{11}(\alpha, \beta): [-a, 0] \times [-a, 0] \rightarrow \mathcal{L}(X)$$

is piecewise absolutely continuous in each variable with jumps of height $A_i^*B^{10}(\beta)^*$ at $\alpha = \theta_i$, $i = 1, \dots, N - 1$ (resp. $B^{10}(\alpha)A_j$ at $\beta = \theta_j$, $j = 1, \dots, N - 1$). Moreover $B^{11}(\alpha, \beta)$ is the solution of

$$(5.26) \quad \left[\frac{\partial}{\partial \alpha} + \frac{\partial}{\partial \beta} \right] B^{11}(\alpha, \beta) = A_{01}(\alpha)^*B^{10}(\beta)^* + B^{10}(\alpha)A_{01}(\beta) + \sum_{i=1}^{N-1} A_i^*B^{10}(\beta)^*\delta(\alpha - \theta_i) + \sum_{j=1}^{N-1} B^{10}(\alpha)A_j\delta(\beta - \theta_j)$$

with boundary conditions

$$(5.27) \quad B^{11}(-a, \beta) = A_N^*B^{10}(\beta)^*, \quad B^{11}(\alpha, -a) = B^{10}(\alpha)A_N,$$

and symmetry property $B^{11}(\alpha, \beta) = B^{11}(\beta, \alpha)^*$.

The solution of the above differential system is

$$(5.28) \quad B^{11}(\alpha, \beta) = \begin{cases} B^{10}(\alpha - \beta - a)A_N, & \alpha \geq \beta \\ A_N^*B^{10}(\beta - \alpha - a)^*, & \alpha < \beta \end{cases} + \sum_{i=1}^{N-1} \begin{cases} A_i^*B^{10}(\beta - \alpha + \theta_i)^*, & -a \leq \beta - \alpha + \theta_i, \theta_i < \alpha \\ 0, & \text{otherwise} \end{cases} + \sum_{j=1}^{N-1} \begin{cases} B^{10}(\alpha - \beta + \theta_j)A_j, & -a \leq \alpha - \beta + \theta_j, \theta_j < \beta \\ 0, & \text{otherwise} \end{cases} \quad (cont.)$$

$$\begin{aligned}
 & + \int_{-a}^{\beta} \left\{ \begin{array}{ll} B^{10}(\alpha - \beta + \xi)A_{01}(\xi), & -a \leq \alpha - \beta + \xi \\ 0 & , \text{ otherwise} \end{array} \right\} d\xi \\
 & + \int_{-a}^{\alpha} \left\{ \begin{array}{ll} A_{01}(\theta)^* B^{10}(\beta - \alpha + \theta)^*, & -a \leq \beta - \alpha + \theta \\ 0 & , \text{ otherwise} \end{array} \right\} d\theta.
 \end{aligned}$$

Proof. See Appendix A.

5.2. Stabilizability. In the control theory of linear ordinary differential equations there is an important result which says that if the system is completely controllable then it is stabilizable, i.e., there exists a constant feedback matrix K such that the resulting closed-loop system matrix can be made to have its eigenvalues strictly in the left half-plane. For hereditary systems we first need a definition of stabilizability.

DEFINITION 5.6. The controlled system (L) (or (\tilde{L})) is said to be *stabilizable* if there exists some operator G in $\mathcal{L}(V, U)$ of the form

$$(5.29) \quad Gh = G_{00}h(0) + \sum_{i=1}^M G_i h(\tau_i) + \int_{-a}^0 G_{01}(\theta)h(\theta) d\theta$$

(for some integer $M \geq 1$, some real numbers $-a = \tau_M < \dots < \tau_1 < \tau_0 = 0$, some G_{00} , G_i ($i = 1, \dots, M$) in $\mathcal{L}(X, U)$ and $G_{01}: [-a, 0] \rightarrow \mathcal{L}(X, U)$ strongly measurable and bounded) such that the resulting *closed-loop system*

$$(5.30) \quad \begin{aligned} \dot{\tilde{x}}(t) &= [\tilde{A} + \tilde{B}G]\tilde{x}(t), \quad \text{a.e. in } [0, \infty), \\ \tilde{x}(0) &= h, \quad h \text{ in } V, \end{aligned}$$

is L^2 -stable.

It is extremely important to notice that for operators of the form (5.29) the map

$$h \mapsto \tilde{x}_G(\cdot; h): M^2 \rightarrow C_{\text{loc}}(0, \infty; M^2)$$

is continuous, where for each h in V , $\tilde{x}_G(\cdot; h)$ denotes the solution of (5.30). This is not true of all operators in $\mathcal{L}(V, U)$. This definition opens the way to the investigation of stabilizability by feedback of a delayed signal (cf. V. M. Popov [18]).

Using the spectral properties of \tilde{A} (cf. J. K. Hale [12]) an analogue of the ordinary differential equation result cited above could be obtained for a linear hereditary differential system. For a study of this question see Y. S. Osipov [17] and H. F. Vandevenne [20], [21].

The importance of the concept of stabilizability and a theorem relating controllability and stabilizability is that it provides us with a verifiable condition for asserting that there exists a constant $c > 0$ and for each h at least one control v such that $J_{\infty}(v, h) \leq c\|h\|^2$. Thus the infinite-time problem is well-posed.

5.3. Asymptotic behavior of $\Pi_T(t)$ as $T \rightarrow \infty$. We know that for the quadratic cost problem over $[0, T]$ the optimal control $u^*(s)$ is given by

$$u^*(s) = -N^{-1}\tilde{B}^*\Pi_T(s)\tilde{x}(s; h), \quad s \in [0, T],$$

and the optimal cost by

$$J_T^0(u^*, h) = (h, \Pi_T(0)h)_{M^2}.$$

We now show the following.

THEOREM 5.7. Assume that (\tilde{L}) is stabilizable. Then:

- (i) For all h in M^2 , $\lim_{t \rightarrow \infty} \Pi_T(t)h = \Pi h$, $t \geq 0$.
- (ii) For all h in M^2 ,

$$(5.31) \quad (\Pi h, h)_{M^2} = \int_0^\infty ([\tilde{Q} + \Pi \tilde{R} \Pi] \tilde{x}(s), \tilde{x}(s)) ds,$$

where $R = BN^{-1}B^*$,

$$(5.32) \quad \tilde{R}h = (Rh^0, 0),$$

and \tilde{x} is the solution of

$$(5.33) \quad \begin{aligned} \frac{dy}{dt}(t) &= (\tilde{A} - \tilde{R}\Pi)y(t), \quad \text{a.e. in } [0, \infty), \\ y(0) &= h, \end{aligned}$$

with initial datum h .

- (iii) For all h in M^2 ,

$$(5.34) \quad (\Pi h, h)_{M^2} = J_\infty(-N^{-1}\tilde{B}^*\Pi\tilde{x}, h).$$

Proof. (i) Consider the optimal control problem on the interval $[s, T]$. By virtue of the stabilizability hypothesis there exists a feedback operator G of the type described in Definition 5.6 such that the operator $\tilde{A} + \tilde{B}G$ is L^2 -stable. Let $\tilde{\Phi}_G$ be the semigroup generated by this operator. For all $T > s \geq 0$,

$$\begin{aligned} (\Pi_T(s)h, h)_{M^2} &= \inf \{J_T^s(v, h) | v \in L^2(s, T, U)\} \\ &\leq \int_s^T [(\tilde{Q}\tilde{\Phi}_G(t-s)h, \tilde{\Phi}_G(t-s)h) + (NG\tilde{\Phi}_G(t-s)h, G\tilde{\Phi}_G(t-s)h)] dt \\ &\leq \int_0^T [(\tilde{Q}\tilde{\Phi}_G(t)h, \tilde{\Phi}_G(t)h) + (NG\tilde{\Phi}_G(t)h, G\tilde{\Phi}_G(t)h)] dt \\ &\leq \|Q\| \int_0^T |z(t)|_X^2 dt + \|N\| \int_0^T |G\tilde{z}(t)|_U dt, \end{aligned}$$

where z is the solution of

$$\begin{aligned} \dot{z}(t) &= (\tilde{A}_0 + BG)\tilde{z}(t), \quad \text{a.e. in } [0, \infty), \\ \tilde{z}(0) &= h, \end{aligned}$$

and \tilde{z} is the state constructed from h and z . But

$$\begin{aligned} \left[\int_0^T |G\tilde{z}(t)|^2 dt \right]^{1/2} &\leq \|G_{00}\| \left[\int_0^T |z(t)|^2 dt \right]^{1/2} \\ &\quad + \sum_{i=1}^M \|G_i\| \left[\int_{\tau_i}^0 |h(\theta)|^2 d\theta + \int_0^T |z(t)|^2 dt \right]^{1/2} \\ &\quad + \|G_{01}\|_\infty a^{1/2} \left[\int_{-a}^0 |h(\theta)|^2 d\theta + \int_0^T |z(t)|^2 dt \right]^{1/2}, \end{aligned}$$

where z is the solution of (5.30). Finally since (5.30) is L^2 -stable there exists a constant $c > 0$ (independent of h , T and s) such that

$$(\Pi_T(s)h, h)_{M^2} \leq c \|h\|_{M^2}^2 \quad \forall h, \forall T \geq s \geq 0.$$

It is now easy to show the following:

(a) $\Pi_{T_2}(s) \geq \Pi_{T_1}(s)$, $T_2 \geq T_1 \geq s$, where \geq denotes the natural partial ordering of positive operators, and

(b) there exists $c > 0$ such that $\|\Pi_T(s)\|_{\mathcal{L}(M^2)} \leq c$ for all $T \geq s$.

Then by a well-known theorem on positive operators (cf. Kantorovich and Akilov [14, p. 189]), for all h in M^2 , $\Pi_T(s)h$ converges to $\Pi(s)h$, for some positive self-adjoint operator $\Pi(s)$ in $\mathcal{L}(M^2)$.

Now for $0 < T_1 - s_1 = T_2 - s_2$, $s_1 \geq s_2 \geq 0$,

$$(h, \Pi_{T_1}(s_1)h) = (h, \Pi_{T_2}(s_2)h)$$

and hence $\Pi_{T_1}(s_1) = \Pi_{T_2}(s_2)$. In particular, for all $s_1 < s_2$ and h in M^2 ,

$$\Pi(s_1)h = \lim_{T_1 \rightarrow \infty} \Pi_{T_1}(s_1)h = \lim_{T_1 \rightarrow \infty} \Pi_{T_1+s_2-s_1}(s_2)h = \Pi(s_2)h$$

and

$$\lim_{T \rightarrow \infty} \Pi_T(s)h = \Pi h \quad \forall s \geq 0.$$

(ii) We now consider the control problem in the interval $[0, \infty)$. Let \tilde{z} denote the solution of (5.30) corresponding to the stabilizing feedback control law G , let \tilde{x} be the solution of (5.33) in $[0, \infty)$ and let \tilde{x}_T be the solution of

$$\begin{aligned} \frac{d\tilde{x}_T}{ds}(s) &= (\tilde{A} - \tilde{R}\Pi_T(s))\tilde{x}_T(s), \quad \text{a.e. in } [0, T], \\ \tilde{x}_T(0) &= h. \end{aligned} \tag{5.35}$$

We first show that for all $t_1 > 0$,

$$\lim_{t_1 < T \rightarrow \infty} \tilde{x}_T(t) \rightarrow \tilde{x}(t) \quad \text{uniformly in } [0, t_1]. \tag{5.36}$$

Fix $t_1 > 0$ and consider T , $T > t_1$. Let

$$y_T(t) = x_T(t) - x(t) \quad \text{in } [0, t_1].$$

Then

$$\begin{aligned} \frac{d\tilde{y}_T}{dt}(t) &= \tilde{A}\tilde{y}_T(t) + \tilde{R}[\Pi\tilde{x}(t) - \Pi_T(t)\tilde{x}_T(t)], \quad \text{a.e. in } [0, t_1], \\ \tilde{y}_T(0) &= 0, \end{aligned}$$

where

$$\tilde{y}_T(s)(\theta) = \begin{cases} y_T(s + \theta), & s + \theta \geq 0, \\ 0 & , \quad \text{otherwise.} \end{cases}$$

As a result there exists $c(t_1) > 0$ such that for all $0 \leq t \leq t_1$,

$$\begin{aligned} |\tilde{y}_T(t)| &\leq c(t_1) \int_0^t |\Pi \tilde{x}(s) - \Pi_T(s) \tilde{x}_T(s)| ds \\ &\leq c(t_1) \int_0^t [|(\Pi - \Pi_T(s)) \tilde{x}(s)| + |\Pi_T(s) \tilde{y}_T(s)|] ds \end{aligned}$$

and we can find $c'(t_1) > 0$ such that

$$\|\tilde{y}_T\|_{C(0,t_1;M^2)} \leq c'(t_1) \int_0^{t_1} |(\Pi - \Pi_T(s)) \tilde{x}(s)| ds.$$

But $\tilde{x} \in L^1(0, t_1; M^2)$. Then $f_T(s) = \Pi_T(s) \tilde{x}(s)$ and $f(s) = \Pi \tilde{x}(s)$ belong to $L^1(0, t_1; M^2)$. Both f_T and f are bounded by the L^1 -function $c|\tilde{x}(s)|$ and for almost all t ,

$$f_T(t) = \Pi_T(t) \tilde{x}(t) \rightarrow f(t) = \Pi \tilde{x}(t) \quad \text{as } T \rightarrow \infty.$$

By the Lebesgue dominated convergence theorem, $f_T \rightarrow f$ in $L^1(0, t_1; M^2)$. This shows that $\tilde{y}_T \rightarrow 0$ and proves (5.36). This also shows that $\tilde{x}_T(t)$ is uniformly bounded in $[0, t_1]$ by a constant independent of T .

We know that for all $T > 0$ (cf. Delfour and Mitter [8])

$$(5.37) \quad (\Pi_T(0)h, h) = \int_0^T ([\tilde{Q} + \Pi_T(s)\tilde{R}\Pi_T(s)] \tilde{x}_T(s), \tilde{x}_T(s)) ds.$$

The left-hand side of (5.37) converges to $(\Pi h, h)$ as T goes to infinity. We now show that the right-hand side of (5.37) converges to

$$\int_0^\infty ([\tilde{Q} + \Pi \tilde{R} \Pi] \tilde{x}(s), \tilde{x}(s)) ds.$$

For this purpose we define

$$\begin{aligned} g_T(t) &= \begin{cases} ([\tilde{Q} + \Pi_T(t)\tilde{R}\Pi_T(t)] \tilde{x}_T(t), \tilde{x}_T(t)), & 0 \leq t \leq T, \\ 0 & \text{otherwise,} \end{cases} \\ g(t) &= ([\tilde{Q} + \Pi \tilde{R} \Pi] \tilde{x}(t), \tilde{x}(t)). \end{aligned}$$

From previous considerations it is now clear that

$$g_T(t) \rightarrow g(t) \quad \text{pointwise in } [0, \infty) \text{ as } T \rightarrow \infty.$$

By Fatou's lemma,

$$\int_0^\infty g(t) dt \leq \lim_{T \rightarrow \infty} \int_0^\infty g_T(t) dt = \lim_{T \rightarrow \infty} (\Pi_T(0)h, h) = (\Pi h, h),$$

and for all $T > 0$,

$$\int_0^T g(t) dt = J_T^0(-N^{-1} \tilde{B}^* \tilde{x}, h) \geq (\Pi_T(0)h, h).$$

(iii) Finally (5.34) has been established at the end of (ii). \square

5.4. Solution to the infinite-time problem.

THEOREM 5.8. Assume that (\tilde{L}) is stabilizable. Then for each h in M^2 , there exists a control function u^* in $L^2_{\text{loc}}(0, \infty; U)$ such that

$$(5.38) \quad J_{\infty}(u^*, h) = \inf \{J_{\infty}(v, h) | v \in L^2_{\text{loc}}(0, \infty; U)\} = (h, \Pi h).$$

Moreover,

$$(5.39) \quad u^*(t) = -N^{-1}\tilde{B}^*\Pi\tilde{x}(s),$$

where \tilde{x} is the solution of

$$\frac{d\tilde{x}}{ds}(s) = (\tilde{A} - \tilde{R}\Pi)\tilde{x}(s), \quad \text{a.e. in } [0, \infty),$$

$$\tilde{x}(0) = h.$$

Proof. The control function u^* defined by (5.38) is clearly an element of $L^2_{\text{loc}}(0, \infty; U)$. Consider any $v \in L^2_{\text{loc}}(0, \infty; U)$. Then for all $T > 0$,

$$(h, \Pi_T(0)h) = \min_{v \in L^2(0, T; U)} J_T^0(v, h) \leq \int_0^T [(Qx(s; v), x(s; v)) + (Nv(s), v(s))] ds,$$

where $x(\cdot; v)$ is the solution of (L) corresponding to h and v . Therefore,

$$(h, \Pi h) \leq \int_0^{\infty} [(Qx(s; v), x(s; v)) + (Nv(s), v(s))] ds,$$

and the result follows from Theorem 5.7 (iii). \square

5.5. Characterization of Π and stability of the closed-loop system.

THEOREM 5.9. Let $Q > 0$. Then:

(i) (\tilde{L}) is stabilizable if and only if there exists a positive self-adjoint operator Π in $\mathcal{L}(M^2)$ which is a solution to the operator equation of Ricatti type

$$(5.40) \quad (\tilde{A}h, \Pi k) + (h, \Pi \tilde{A}k) - (h, \Pi \tilde{R}\Pi k) + (h, \tilde{Q}k) = 0 \quad \forall h, k \text{ in } V.$$

(ii) If a positive self-adjoint solution of (5.40) exists, it is unique and equal to the Π of Theorem 5.7. The operator $\tilde{A} - \tilde{R}\Pi$ is L^2 -stable, the operator $G^* = -N^{-1}\tilde{B}^*\Pi$ defines a stable feedback law and Π is positive definite on X .

Proof. (i) Assume that system (\tilde{L}) is stabilizable. Then equation (5.31) of Theorem 5.7(ii) is true for all h in M^2 . Since $Q > 0$ and $\Pi \tilde{R}\Pi \geq 0$ we can use Theorem 5.3(i) and (ii) to conclude that the operator $\tilde{A} - \tilde{R}\Pi$ is L^2 -stable. Since Q, Π and $\tilde{Q} + \Pi \tilde{R}\Pi$ are positive and self-adjoint, equation (5.31) implies that for all h and k in M^2 ,

$$(\Pi h, k) = \int_0^{\infty} ([\tilde{Q} + \Pi \tilde{R}\Pi]\tilde{x}_h(s), \tilde{x}_k(s)) ds,$$

where \tilde{x}_h (resp. \tilde{x}_k) is the solution of equation (5.33) with initial datum h (resp. k).

Let $\tilde{\Phi}(s)$ be the strongly continuous semigroup generated by $\tilde{A} - \tilde{R}\Pi$, that is, $\tilde{x}(s) = \tilde{\Phi}(s)h$. For all h and k in V ,

$$(\Pi(\tilde{A} - \tilde{R}\Pi)h, k) = \int_0^\infty ((\tilde{Q} + \Pi\tilde{R}\Pi)\tilde{\Phi}(s)(\tilde{A} - \tilde{R}\Pi)h, \tilde{\Phi}(s)k) ds,$$

$$(\Pi h, (\tilde{A} - \tilde{R}\Pi)k) = \int_0^\infty ((\tilde{Q} + \Pi\tilde{R}\Pi)\tilde{\Phi}(s)h, \tilde{\Phi}(s)(\tilde{A} - \tilde{R}\Pi)k) ds,$$

$$\begin{aligned} (\Pi(\tilde{A} - \tilde{R}\Pi)h, k) + (\Pi h, (\tilde{A} - \tilde{R}\Pi)k) &= \int_0^\infty \frac{d}{ds} ((\tilde{Q} + \Pi\tilde{R}\Pi)\tilde{x}_h(s), \tilde{x}_k(s)) ds \\ &= -((\tilde{Q} + \Pi\tilde{R}\Pi)h, k), \end{aligned}$$

since $\tilde{A} - \tilde{R}\Pi$ is L^2 -stable and $\tilde{x}(s) \rightarrow 0$ as $s \rightarrow \infty$. Finally,

$$\begin{aligned} 0 &= \Pi(\tilde{A} - \tilde{R}\Pi) + (\tilde{A} - \tilde{R}\Pi)^*\Pi + \Pi\tilde{R}\Pi + \tilde{Q} \\ &= \Pi\tilde{A} + \tilde{A}^*\Pi - \Pi\tilde{R}\Pi + \tilde{Q}. \end{aligned}$$

Conversely assume that there exists a solution Π to the operator Riccati equation (5.40) which is self-adjoint and positive. Equation (5.40) can be rewritten as

$$(\tilde{A} - \tilde{R}\Pi)^*\Pi + \Pi(\tilde{A} - \tilde{R}\Pi) + \Pi\tilde{R}\Pi + \tilde{Q} = 0.$$

By Theorem 5.3(iv), this means that the system defined by the operator $\tilde{A} - \tilde{R}\Pi$ is L^2 -stable. It is now a simple matter to check that the stabilizing feedback law is $G^* = -N^{-1}B^*\Pi^0$.

(ii) If a positive self-adjoint solution of (5.40) exists, we have shown that system (\tilde{L}) is stabilizable, that Π is a solution of (5.40), that the operator $\tilde{A} - \tilde{R}\Pi$ is L^2 -stable and that G^* is a stable feedback law. By Proposition 5.4 we can also say that Π is positive definite on X . It remains to prove uniqueness. Assume that there exist two solutions $\Pi_1 \geq 0$ and $\Pi_2 \geq 0$ to the Riccati equation (5.40). Let $P = \Pi_1 - \Pi_2$. Then necessarily

$$(\tilde{A}h, Pk) + (Ph, \tilde{A}k) + (h, \Pi_2\tilde{R}\Pi_2k) - (h, \Pi_1\tilde{R}\Pi_1k) = 0$$

or

$$((\tilde{A} - \tilde{R}\Pi_2)h, Pk) + (h, P(\tilde{A} - \tilde{R}\Pi_1)k) = 0.$$

Hence

$$\begin{aligned} \frac{d}{ds}(\tilde{\Phi}_2(s)h, P\tilde{\Phi}_1(s)k) &= ((\tilde{A} - \tilde{R}\Pi_2)\tilde{\Phi}_2(s)h, P\tilde{\Phi}_1(s)k) \\ &\quad + (\tilde{\Phi}_2(s)h, P(\tilde{A} - \tilde{R}\Pi_1)\tilde{\Phi}_1(s)k) = 0, \end{aligned}$$

where $\tilde{\Phi}_2$ (resp. $\tilde{\Phi}_1$) is the semigroup generated by $\tilde{A} - \tilde{R}\Pi_2$ (resp. $\tilde{A} - \tilde{R}\Pi_1$). Then

$$(h, Pk) = (\tilde{\Phi}_2(s)h, P\tilde{\Phi}_1(s)k) \rightarrow 0 \quad \text{as } s \rightarrow \infty,$$

since $\tilde{\Phi}_2$ and $\tilde{\Phi}_1$ are L^2 -stable. Finally $P = 0$ and equation (5.40) has a unique solution which is necessarily equal to the Π of Theorem 5.7.

Remark. Note that the hypothesis $Q > 0$ implies that the pair $(\tilde{A}, Q^{1/2})$ is observable since the map $h^0 \mapsto Q^{1/2}\Phi^0(\cdot)h^0$ is injective (cf. Delfour and Mitter [8, Def. 3.11 and Prop. 3.13]).

6. Detailed characterization of Π . One can exploit the structure of the space M^2 and the fact that Π is a matrix of operators to give a detailed characterization of Π . This is done in the following theorem.

THEOREM 6.1. *Let $\Pi \geq 0$ in $\mathcal{L}(M^2)$ be the solution of (5.40). Then*

$$(6.1) \quad (h, \Pi k) = \int_0^\infty (\tilde{\Phi}(t)h, [\tilde{Q} + \Pi \tilde{R} \Pi] \tilde{\Phi}(t)k) dt.$$

It is completely characterized by its matrix of operators

$$(6.2) \quad \begin{bmatrix} \Pi_{00} & \Pi_{01} \\ \Pi_{10} & \Pi_{11} \end{bmatrix}, \quad \begin{array}{l} \Pi_{00} \in \mathcal{L}(X), \quad \Pi_{01} \in \mathcal{L}(L^2(-a, 0; X), X), \\ \Pi_{10} \in \mathcal{L}(X, L^2(-a, 0; X)), \quad \Pi_{11} \in \mathcal{L}(L^2(-a, 0; X)). \end{array}$$

Π_{00} is characterized by the equation

$$(6.3) \quad \begin{aligned} \Pi_{00}A_{00} + A_{00}^*\Pi_{00} + \Pi_{10}(0) + \Pi_{10}(0)^* + Q - \Pi_{00}R\Pi_{00} &= 0, \\ \Pi_{00}^* &= \Pi_{00} \geq 0. \end{aligned}$$

Π_{10} is characterized in the following way:

$$(6.4) \quad (\Pi_{10}h^0)(\alpha) = \Pi_{10}(\alpha)h^0,$$

where the map

$$(6.5) \quad \alpha \mapsto \Pi_{10}(\alpha): [-a, 0] \rightarrow \mathcal{L}(X)$$

is piecewise absolutely continuous with jumps at $\alpha = \theta_i$ of height $A_i^*\Pi_{00}$, $i = 1, \dots, N-1$. Moreover the map (6.5) is characterized by the differential equation

$$(6.6) \quad \begin{aligned} \frac{d\Pi_{10}}{d\alpha}(\alpha) &= \Pi_{10}(\alpha)[A_{00} - R\Pi_{00}] + \sum_{i=1}^{N-1} A_i^*\Pi_{00}\delta(\alpha - \theta_i) + A_{01}(\alpha)^*\Pi_{00} \\ &+ \Pi_{11}(\alpha, 0), \quad \text{a.e. in } [-a, 0], \\ \Pi_{10}(-a) &= A_N^*\Pi_{00}, \end{aligned}$$

where $\delta(\alpha - \theta_i)$ is the delta function at $\alpha = \theta_i$.

Π_{01} is obtained from Π_{10} :

$$(6.7) \quad \Pi_{01}h^1 = \int_{-a}^0 \Pi_{10}(\alpha)^*h^1(\alpha) d\alpha.$$

Π_{11} is characterized in the following way:

$$(6.8) \quad (\Pi_{11}h^1)(\alpha) = \int_{-a}^0 \Pi_{11}(\alpha, \beta)h^1(\beta) d\beta,$$

where the map

$$(6.9) \quad (\alpha, \beta) \mapsto \Pi_{11}(\alpha, \beta): [-a, 0] \times [-a, 0] \rightarrow \mathcal{L}(X)$$

is piecewise absolutely continuous in each variable with jumps of height $A_i^* \Pi_{10}(\beta)^*$ at $\alpha = \theta_i, i = 1, \dots, N - 1$ (resp. $\Pi_{10}(\alpha) A_j$ at $\beta = \theta_j, j = 1, \dots, N - 1$). Moreover $\Pi_{11}(\alpha, \beta)$ is the solution of

$$(6.10) \quad \left[\frac{\partial}{\partial \alpha} + \frac{\partial}{\partial \beta} \right] \Pi_{11}(\alpha, \beta) = A_{01}(\alpha)^* \Pi_{10}(\beta)^* + \Pi_{10}(\alpha) A_{01}(\beta) \\ + \sum_{i=1}^{N-1} A_i^* \Pi_{10}(\beta)^* \delta(a - \theta_i) + \sum_{j=1}^{N-1} \Pi_{10}(\alpha) A_j \delta(\beta - \theta_j) \\ - \Pi_{10}(\alpha) R \Pi_{10}(\beta)^*$$

with boundary conditions

$$(6.11) \quad \Pi_{11}(-a, \beta) = A_N^* \Pi_{10}(\beta)^*, \quad \Pi_{11}(\alpha, -a) = \Pi_{10}(\alpha) A_N,$$

and symmetry property

$$\Pi_{11}(\alpha, \beta) = \Pi_{11}(\beta, \alpha)^*.$$

The solution of the above differential system is

$$(6.12) \quad \Pi_{11}(\alpha, \beta) = \left\{ \begin{array}{l} \Pi_{10}(\alpha - \beta - a) A_N, \quad \alpha \geq \beta \\ A_N^* \Pi_{10}(\beta - \alpha - a)^*, \quad \alpha < \beta \end{array} \right\} \\ + \sum_{i=1}^{N-1} \left\{ \begin{array}{l} A_i^* \Pi_{10}(\beta - \alpha + \theta_i)^*, \quad -a \leq \beta - \alpha + \theta_i, \theta_i < \alpha \\ 0, \quad \text{otherwise} \end{array} \right\} \\ + \sum_{j=1}^{N-1} \left\{ \begin{array}{l} \Pi_{10}(\alpha - \beta + \theta_j) A_j, \quad -a \leq \alpha - \beta + \theta_j, \theta_j < \beta \\ 0, \quad \text{otherwise} \end{array} \right\} \\ + \int_{-a}^{\alpha} \left\{ \begin{array}{l} A_{01}(\xi)^* \Pi_{10}(\xi - \alpha + \beta)^*, \quad \xi \geq \alpha - \beta - a \\ 0, \quad \text{otherwise} \end{array} \right\} d\xi \\ + \int_{-a}^{\beta} \left\{ \begin{array}{l} \Pi_{10}(\theta - \beta + \alpha) A_{01}(\theta), \quad \theta \geq \beta - \alpha - a \\ 0, \quad \text{otherwise} \end{array} \right\} d\theta \\ - \left\{ \begin{array}{l} \int_{-a}^{\beta} \Pi_{10}(\alpha - \beta + \theta) R \Pi_{10}(\theta)^* d\theta, \quad \alpha \geq \beta \\ \int_{-a}^{\alpha} \Pi_{10}(\xi) R \Pi_{10}(\beta - \alpha + \xi)^* d\xi, \quad \alpha < \beta \end{array} \right\}.$$

Proof. See Appendix B.

Appendix A.

Proof of Theorem 5.5. The reader can find the definitions of Φ^0 , Φ^1 and $\tilde{\Phi}$ in Delfour and Mitter [7], [8] and [10]. We first rewrite equation (5.15) in terms of $\tilde{\Phi}$:

$$(A.1) \quad (h, \Pi k) = \int_0^\infty (\tilde{\Phi}(t) h, [\tilde{Q} + \Pi \tilde{R} \Pi] \tilde{\Phi}(t) k) dt.$$

We shall also use the identity

$$(A.2) \quad [\tilde{\Phi}(t)h]^0 = \Phi^0(t)h^0 + \Phi^1(t)h^1.$$

We first study B^{00} and the kernels $B^{10}(\alpha)$ and $B^{11}(\alpha, \beta)$ of the operators B^{10} and B^{11} . Since we know where the discontinuities can occur we derive differential equations for $B^{10}(\alpha)$ and $B^{11}(\alpha, \beta)$. Finally we solve the equation for $B^{11}(\alpha, \beta)$ and give an explicit expression of $B^{11}(\alpha, \beta)$ in terms of $B^{10}(\cdot)$.

(i) Let $h = (h^0, 0)$ and $k = (k^0, 0)$ in (A.1). Then

$$(A.3) \quad B^{00} = \int_0^\infty \Phi^0(t) * \Phi^0(t) dt.$$

Let $h = (0, h^1)$, $k = (k^0, 0)$ in (A.1). Then

$$(A.4) \quad (h^1, B^{10}k^0) = \int_0^\infty (\Phi^1(t)h^1, \Phi^0(t)k^0) dt.$$

But (cf. Delfour and Mitter [7] and [8])

$$(A.5) \quad \Phi^1(t)h^1 = \int_{-a}^0 \Phi^1(t, \alpha)h^1(\alpha) d\alpha,$$

$$(A.6) \quad (h^1, B^{10}k^0) = \int_{-a}^0 (h^1(\alpha), \int_0^\infty \Phi^1(t, \alpha) * \Phi^0(t)k^0) dt d\alpha,$$

and

$$(A.7) \quad B^{10}(\alpha) = \int_0^\infty \Phi^1(t, \alpha) * \Phi^0(t) dt.$$

We now substitute for $\Phi^1(t, \alpha)$ the expression (cf. Delfour and Mitter [7] and [8])

$$(A.8) \quad \sum_{i=1}^N \left\{ \begin{array}{ll} \Phi^0(t - \alpha + \theta_i)A_i, & t \geq \alpha - \theta_i \geq 0 \\ 0 & , \text{ otherwise} \end{array} \right\} \\ + \int_{\max\{-a, \alpha-t\}}^\alpha \Phi^0(t - \alpha + \theta)A_{01}(\theta) d\theta.$$

Identity (A.7) can now be rewritten in the form

$$(A.9) \quad B^{10}(\alpha) = \sum_{i=1}^N \left\{ \begin{array}{ll} A_i^* \int_{\alpha-\theta_i}^\infty \Phi^0(t - \alpha + \theta_i) * \Phi^0(t) dt, & \theta_i \leq \alpha \\ 0 & , \theta_i > \alpha \end{array} \right\} \\ + \int_0^\infty dt \int_{\max\{-a, \alpha-t\}}^\alpha d\theta A_{01}(\theta) * \Phi^0(t - \alpha + \theta) * \Phi^0(t).$$

Finally we change the order of integration of the last term in (A.9) to obtain

$$(A.10) \quad B^{10}(\alpha) = \sum_{i=1}^N A_i^* \left\{ \begin{array}{ll} \int_{\alpha-\theta_i}^{\infty} \Phi^0(t-\alpha+\theta_i)^* \Phi^0(t) dt, & \theta_i \leq \alpha \\ 0 & , \quad \theta_i > \alpha \end{array} \right\} \\ + \int_{-a}^{\alpha} d\theta A_{01}(\theta)^* \int_{\alpha-\theta}^{\infty} dt \Phi^0(t-\alpha+\theta)^* \Phi^0(t).$$

By inspection it is readily seen that $B^{10}(\alpha)$ has jumps at $\alpha = \theta_i, i = 1, \dots, N-1$, of respective heights $A_i^* B^{00}$. Moreover

$$(A.11) \quad B^{10}(-a) = A_N^* B^{00}.$$

Let $h = (0, h^1)$ and $k = (0, k^1)$ in (A.1). Then

$$(A.12) \quad (h^1, B^{11}k^1) = \int_0^{\infty} (\Phi^1(t)h^1, \Phi^1(t)k^1) dt.$$

In view of (A.5),

$$(A.13) \quad (h^1, B^{11}k^1) = \int_0^{\infty} \left(\int_{-a}^0 \Phi^1(t, \alpha) h^1(\alpha) d\alpha, \int_{-a}^0 \Phi^1(t, \beta) k^1(\beta) d\beta \right) dt$$

and

$$(A.14) \quad B^{11}(\alpha, \beta) = \int_0^{\infty} \Phi^1(t, \alpha)^* \Phi^1(t, \beta) dt.$$

We again use (A.8) to express $B^{11}(\alpha, \beta)$ in terms of Φ^0 :

$$(A.15) \quad B^{11}(\alpha, \beta) = \int_0^{\infty} \left[\sum_{i=1}^N \left\{ \begin{array}{ll} A_i^* \Phi^0(t-\alpha+\theta_i)^*, & t \geq \alpha - \theta_i \geq 0 \\ 0 & , \quad \text{otherwise} \end{array} \right\} + \int_{\max\{-a, \alpha-t\}}^{\alpha} A_{01}(\theta)^* \Phi^0(t-\alpha+\theta)^* d\theta \right] \\ \left[\sum_{j=1}^N \left\{ \begin{array}{ll} \Phi^0(t-\beta+\theta_j) A_j, & t \geq \beta - \theta_j \geq 0 \\ 0 & , \quad \text{otherwise} \end{array} \right\} + \int_{\max\{-a, \beta-t\}}^{\beta} \Phi^0(t-\beta+\xi) A_{01}(\xi) d\xi \right] dt \\ = \sum_{i=1}^N \sum_{j=1}^N \int_0^{\infty} \left\{ \begin{array}{ll} A_i^* \Phi^0(t-\alpha+\theta_i)^* \Phi^0(t-\beta+\theta_j) A_j, & t \geq \alpha - \theta_i \geq 0, t \geq \beta - \theta_j \geq 0 \\ 0 & , \quad \text{otherwise} \end{array} \right\} dt \\ + \sum_{i=1}^N \int_0^{\infty} dt \left\{ \begin{array}{ll} A_i^* \Phi^0(t-\alpha+\theta_i)^*, & t \geq \alpha - \theta_i \geq 0 \\ 0 & , \quad \text{otherwise} \end{array} \right\} \int_{\max\{-a, \beta-t\}}^{\beta} \Phi^0(t-\beta+\xi) A_{01}(\xi) d\xi \\ + \sum_{j=1}^N \int_0^{\infty} dt \left\{ \begin{array}{ll} \int_{\max\{-a, \alpha-t\}}^{\alpha} d\theta A_{01}(\theta)^* \Phi^0(t-\alpha+\theta)^* \Phi^0(t-\beta+\theta_j) A_j, & t \geq \beta - \theta_j \geq 0 \\ 0 & , \quad \text{otherwise} \end{array} \right\} \\ + \int_0^{\infty} dt \int_{\max\{-a, \alpha-t\}}^{\alpha} d\theta \int_{\max\{-a, \beta-t\}}^{\beta} d\xi A_{01}(\theta)^* \Phi^0(t-\alpha+\theta)^* \Phi^0(t-\beta+\xi) A_{01}(\xi) \end{array} \right. \quad (cont.)$$

Given α , term ① has jumps at $\beta = \theta_j, j = 1, \dots, N - 1$, of height

Given β , term ① has jumps at $\alpha = \theta_i, i = 1, \dots, N - 1$, of height

$$(A.17) \quad \sum_{j=1}^N \left\{ \begin{array}{ll} \int_{\beta-\theta_j}^{\infty} dt A_i^* \Phi^0(t) \Phi^0(t-\beta+\theta_j) A_j, & \beta \geq \theta_j \\ 0 & , \quad \beta < \theta_j \end{array} \right\}.$$

Given β , term ② has jumps at $\alpha = \theta_i, i = 1, \dots, N - 1$, of height

$$(A.18) \quad \int_{-a}^{\beta} d\xi \int_{\beta-\xi}^{\infty} dt A_i^* \Phi^0(t) * \Phi^0(t - \beta + \xi) A_{01}(\xi).$$

Given α , term ③ has jumps at $\beta = \theta_j, j = 1, \dots, N - 1$, of height

$$(A.19) \quad \int_{-a}^{\alpha} d\theta \int_{\alpha-\theta}^{\infty} dt A_{01}(\theta) * \Phi^0(t - \alpha + \theta) * \Phi^0(t) A_j.$$

Given α , term ② has no jumps. Given β , term ③ has no jumps. Term ④ has no jumps. Finally, given α the map $\beta \mapsto B^{11}(\alpha, \beta)$ has jumps at $\beta = \theta_j, j = 1, \dots, N - 1$, of height $B^{10}(\alpha) A_j$ and given β the map $\alpha \mapsto B^{11}(\alpha, \beta)$ has jumps at $\alpha = \theta_i, i = 1, \dots, N - 1$, of height $A_i^* B^{10}(\beta)^*$. Moreover,

$$\begin{aligned} B^{11}(-a, \beta) = & \sum_{i=1}^N \left\{ \int_{\beta-\theta_j}^{\infty} A_N^* \Phi^0(t) * \Phi^0(t - \beta + \theta_j) A_j, \quad \beta \geq \theta_j \right\} \\ & 0, \quad \beta < \theta_j \Bigg\} \\ & + \int_0^{\infty} dt \int_{\max\{-a, \beta-t\}}^{\beta} d\xi A_N^* \Phi^0(t) * \Phi^0(t - \beta + \xi) A_{01}(\xi) \end{aligned}$$

and

$$(A.20) \quad B^{11}(-a, \beta) = A_N^* B^{10}(\beta)^*.$$

We now express $B^{11}(\alpha, \beta)$ in terms of $B^{10}(\cdot)$. To do this we consider separately each of the four terms in (A.15).

$$\begin{aligned} \textcircled{1} = & \sum_{j=1}^N \left\{ \sum_{i=1}^N \left\{ \int_{\alpha-\theta_i}^{\infty} dt A_i^* \Phi^0(t - \alpha + \theta_i) * \Phi^0(t - \beta + \theta_j) A_j, \quad \beta - \theta_j - \alpha + \theta_i \leq 0 \right\} \right. \\ & 0, \quad \text{otherwise} \Bigg\} \left. \begin{matrix} \alpha \geq \theta_i \\ \alpha < \theta_i \end{matrix} \right\}, \quad \begin{matrix} \beta \geq \theta_j \\ \beta < \theta_j \end{matrix} \Bigg\} \\ & + \sum_{i=1}^N \left\{ \sum_{j=1}^N \left\{ \int_0^{\infty} dt A_i^* \Phi^0(t - \alpha + \theta_i) * \Phi^0(t - \beta + \theta_j) A_j, \quad \beta - \theta_j - \alpha + \theta_i > 0 \right\} \right. \\ & 0, \quad \beta < \theta_j \Bigg\}, \quad \begin{matrix} \alpha \geq \theta_i \\ \alpha < \theta_i \end{matrix} \Bigg\} \\ = & \sum_{j=1}^N \left\{ \sum_{i=1}^N \left\{ \int_{\alpha-\beta+\theta_j-\theta_i}^{\infty} dt A_i^* \Phi^0(t - \alpha + \beta - \theta_j + \theta_i) * \Phi^0(t) A_j, \quad -\alpha + \beta - \theta_j + \theta_i \leq 0 \right\} \right. \\ & 0, \quad \text{otherwise} \Bigg\}, \quad \begin{matrix} \beta \geq \theta_j \\ \beta < \theta_j \end{matrix} \Bigg\} \\ & + \sum_{i=1}^N \left\{ \sum_{j=1}^N \left\{ \int_{\beta-\alpha+\theta_i-\theta_j}^{\alpha} dt A_i^* \Phi^0(t) * \Phi^0(t - \beta + \alpha - \theta_i + \theta_j), \quad -\beta + \alpha - \theta_i + \theta_j \leq 0 \right\} \right. \\ & 0, \quad \text{otherwise} \Bigg\}, \quad \begin{matrix} \alpha \geq \theta_i \\ \alpha < \theta_i \end{matrix} \Bigg\}. \end{aligned}$$

$$\begin{aligned}
 \textcircled{2} &= \sum_{i=1}^N \left\{ \int_{-a}^{\beta} d\xi \begin{cases} \int_{\beta-\alpha+\theta_i-\xi}^{\infty} dt A_i^* \Phi^0(t) * \Phi^0(t + \alpha - \theta_i - \beta + \xi) A_{01}(\xi), & \beta - \alpha + \theta_i \geq \xi \\ \int_0^{\infty} dt & , \beta - \alpha + \theta_i < \xi \end{cases} \right. \\
 &\quad \left. , \alpha \geq \theta_i \right\} \\
 &= \sum_{i=1}^N A_i^* \left\{ \int_{-a}^{\beta-\alpha+\theta_i} d\xi \int_{\beta-\alpha+\theta_i-\xi}^{\infty} dt \Phi^0(t) * \Phi^0(t + \alpha - \theta_i - \beta + \xi) A_{01}(\xi), \quad -a \leq \beta - \alpha + \theta_i, \alpha \geq \theta_i \right\} \\
 &\quad , \text{ otherwise} \\
 &+ \sum_{i=1}^N A_i^* \left\{ \int_{\beta-\alpha+\theta_i}^{\beta} d\xi \int_0^{\infty} dt \Phi^0(t) * \Phi^0(t + \alpha - \theta_i - \beta + \xi) A_{01}(\xi), \quad -a \leq \beta - \alpha + \theta_i, \alpha \geq \theta_i \right\} \\
 &\quad , \quad -a > \beta - \alpha + \theta_i, \alpha \geq \theta_i \\
 &\quad , \text{ otherwise} \\
 &= \sum_{i=1}^N A_i^* \left\{ \int_{-a}^{\beta-\alpha+\theta_i} d\xi \int_{\beta-\alpha+\theta_i-\xi}^{\infty} dt \Phi^0(t) * \Phi^0(t - \beta + \alpha - \theta_i + \xi) A_{01}(\xi), \quad -a \leq \beta - \alpha + \theta_i, \alpha \geq \theta_i \right\} \\
 &\quad , \text{ otherwise} \\
 &+ \int_{-a}^{\beta} d\xi \sum_{i=1}^N \left\{ A_i^* \int_{\alpha-\beta+\xi-\theta_i}^{\infty} dt \Phi^0(t - \alpha + \beta - \xi + \theta_i) * \Phi^0(t), \quad \xi \geq \beta - \alpha + \theta_i, \alpha \geq \theta_i \right\} \\
 &\quad , \text{ otherwise} \quad A_{01}(\xi).
 \end{aligned}$$

Notice that we can drop $\alpha \geq \theta_i$ in the last term since

$$\beta \geq \xi \quad \text{and} \quad \xi \geq \beta - \alpha + \theta_i \Rightarrow \alpha \geq \theta_i.$$

By symmetry

$$\begin{aligned}
 \textcircled{3} &= \sum_{j=1}^N \left\{ \int_{-a}^{\alpha-\beta+\theta_j} d\theta A_{01}(\theta) * \int_{\alpha-\beta+\theta_j-\theta}^{\infty} dt \Phi^0(t - \alpha + \beta - \theta_j + \theta) * \Phi^0(t) A_j, \quad -a \leq \alpha - \beta + \theta_j, \beta \geq \theta_j \right\} \\
 &\quad , \text{ otherwise} \\
 &+ \int_{-a}^{\alpha} d\theta A_{01}(\theta) * \sum_{j=1}^N \left\{ \int_{\beta-\alpha+\theta-\theta_j}^{\infty} dt \Phi^0(t) * \Phi^0(t - \beta + \alpha - \theta) A_j, \quad \theta \geq \alpha - \beta + \theta_j \right\} \\
 &\quad , \text{ otherwise}
 \end{aligned}$$

Finally,

$$\begin{aligned}
 \textcircled{4} &= \int_{-a}^{\alpha} d\theta \int_{-a}^{\beta} d\xi \left\{ \int_{\alpha-\theta}^{\infty} dt A_{01}(\theta) * \Phi^0(t - \alpha + \theta) * \Phi^0(t - \beta + \xi) A_{01}(\xi), \quad \alpha - \theta \geq \beta - \xi \right\} \\
 &\quad \left\{ \int_{\beta-\xi}^{\infty} dt A_{01}(\theta) * \Phi^0(t - \alpha + \theta) * \Phi^0(t - \beta + \xi) A_{01}(\xi), \quad \alpha - \theta < \beta - \xi \right\} \\
 &= \int_{-a}^{\beta} d\xi \left\{ \int_{-a}^{\alpha-\beta+\xi} d\theta \int_{\alpha-\beta+\xi-\theta}^{\infty} dt A_{01}(\theta) * \Phi^0(t - \alpha + \beta - \xi + \theta) * \Phi^0(t) A_{01}(\xi), \quad \alpha - \beta + \xi \geq -a \right\} \\
 &\quad , \text{ otherwise}
 \end{aligned}$$

(cont.)

$$+ \int_{-a}^{\alpha} d\theta \left\{ \int_{-a}^{\beta-\alpha+\theta} d\theta \int_{\beta-\alpha+\theta-\xi}^{\infty} dt A_{01}(\theta) * \Phi^0(t) * \Phi^0(t - \beta + \alpha - \theta + \xi) A_{01}(\xi), \quad \beta - \alpha + \theta \geq -a \right\} \\ , \quad \text{otherwise}$$

(ii) We now derive equations (5.19), (5.22) and (5.26). Our starting point is the Lyapunov equation

$$(A.21) \quad 0 = (\tilde{A}h, Bk) + (Bh, \tilde{A}k) + (\tilde{I}h, k) \quad \forall h, k \text{ in } V$$

or

$$(A.22) \quad \left(A_{00}h(0) + \sum_{i=1}^N A_i h(\theta_i) + \int_{-a}^0 A_{01}(\alpha) h(\alpha) d\alpha, B^{00}k(0) + \int_{-a}^0 B^{01}(\theta) k(\theta) d\theta \right) \\ + \int_{-a}^0 \left(\frac{dh}{d\alpha}(\alpha), B^{10}(\alpha) k(0) + \int_{-a}^0 B^{11}(\alpha, \theta) k(\theta) d\theta \right) d\alpha \\ + \left(B^{00}h(0) + \int_{-a}^0 B^{01}(\alpha) h(\alpha) d\alpha, A_{00}k(0) + \sum_{i=1}^N A_i k(\theta_i) \right. \\ \left. + \int_{-a}^0 A_{01}(\theta) k(\theta) d\theta \right) \\ + \int_{-a}^0 \left(B^{10}(\theta) h(0) + \int_{-a}^0 B^{11}(\theta, \alpha) h(\alpha) d\alpha, \frac{dk}{d\theta}(\theta) \right) d\theta + (h(0), k(0)) = 0.$$

Let

$$h_n(\theta) = \begin{cases} h^0 \left(1 + n \frac{\theta}{a} \right), & -\frac{a}{n} \leq \theta \leq 0, \\ 0, & \text{otherwise,} \end{cases}$$

where n is chosen in such a way that $n > a\theta_1^{-1}$. Then

$$h_n(0) \rightarrow h^0 \quad \text{and} \quad h_n \rightarrow 0 \text{ in } L^2(-a, 0; X).$$

Let k_i be chosen in $W^{1,2}(-a, 0; X)$ in such a way that

$$\text{supp } k_i \subset (\theta_i, \theta_{i-1}) \cup (\theta_1, 0].$$

Let $h = h_n$ and $k = k_i$ in (A.22):

$$(A.23) \quad 0 = \left(A_{00}h^0 + \int_{\theta_1}^0 A_{01}(\theta) h_n(\theta) d\theta, B^{00}k_i(0) + \left\{ \int_{\theta_i}^{\theta_{i-1}} + \int_{\theta_1}^0 \right\} B^{01}(\theta) k_i(\theta) d\theta \right) \\ + \int_{\theta_1}^0 \left(\frac{dh_n}{d\alpha}(\alpha), B^{10}(\alpha) k_i(0) + \left\{ \int_{\theta_i}^{\theta_{i-1}} + \int_{\theta_1}^0 \right\} B^{11}(\alpha, \theta) k_i(\theta) d\theta \right) d\alpha \\ + \left(B^{00}h^0 + \int_{\theta_1}^0 B^{01}(\alpha) h_n(\alpha) d\alpha, A_{00}k_i(0) + \left\{ \int_{\theta_i}^{\theta_{i-1}} + \int_{\theta_1}^0 \right\} A_{01}(\theta) k_i(\theta) d\theta \right) \\ + \left\{ \int_{\theta_i}^{\theta_{i-1}} + \int_{\theta_1}^0 \right\} \left(B^{10}(\theta) h^0 + \int_{\theta_1}^0 B^{11}(\theta, \alpha) h_n(\alpha) d\alpha, \frac{dk_i}{d\theta}(\theta) \right) d\theta + (h^0, k_i(0)).$$

Since $\alpha \mapsto \Pi_{01}(\alpha)$, $\alpha \mapsto \Pi_{11}(\alpha, \theta)$ and $\theta \mapsto \Pi_{11}(\alpha, \theta)$ are absolutely continuous in (θ_i, θ_{i-1}) and $(\theta_1, 0)$ we can now integrate by parts.

Equation (A.23) now reduces to

$$\begin{aligned}
 0 = & \left(A_{00}h^0 + \int_{\theta_1}^0 A_{01}(\theta)h_n(\theta) d\theta, B^{00}k_i(0) + \left\{ \int_{\theta_i}^{\theta_{i-1}} + \int_{\theta_1}^0 \right\} B^{01}(\theta)k_i(\theta) d\theta \right) \\
 & + \left(h_n(0), B^{10}(0)k_i(0) + \left\{ \int_{\theta_i}^{\theta_{i-1}} + \int_{\theta_1}^0 \right\} B^{11}(0, \theta)k_i(\theta) d\theta \right) \\
 & - \int_{\theta_1}^0 \left(h_n(\alpha), B^{10}(\alpha)k_i(0) + \left\{ \int_{\theta_i}^{\theta_{i-1}} + \int_{\theta_1}^0 \right\} B^{11}(\alpha, \theta)k_i(\theta) d\theta \right) \\
 (A.24) \quad & + \left(B^{00}h^0 + \int_{\theta_1}^0 B^{01}(\alpha)h_n(\alpha) d\alpha, A_{00}k_i(0) + \left\{ \int_{\theta_i}^{\theta_{i-1}} + \int_{\theta_1}^0 \right\} A_{01}(\theta)k_i(\theta) d\theta \right) \\
 & + \left(B^{10}(0)h^0 + \int_{\theta_1}^0 B^{11}(0, \alpha)h_n(\alpha) d\alpha, k_i(0) \right) - \left\{ \int_{\theta_i}^{\theta_{i-1}} + \int_{\theta_1}^0 \right\} \left(\frac{dB^{01}}{d\theta}(\theta)h^0 \right. \\
 & \left. + \int_{\theta_1}^0 \frac{d\Pi_{11}}{d\theta}(\theta, \alpha)h_n(\alpha) d\alpha, k_i(\theta) \right) d\theta + (h^0, k_i(0)).
 \end{aligned}$$

Notice that

$$\int_{-a}^0 |A_{01}(\theta)h_n(\theta)| d\theta \leq \left[\int_{-a}^0 |A_{01}(\theta)|^2 d\theta \right]^{1/2} \left[\int_{-a}^0 |h_n(\theta)|^2 d\theta \right]^{1/2}$$

and

$$\lim_{n \rightarrow \infty} \|h_n\|_{L^2(-a, 0; X)} = 0$$

imply that

$$\int_{-a}^0 |A_{01}(\theta)h_n(\theta)| d\theta \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Similarly given any f in $L^2(-a, 0; X)$,

$$\left| \int_{-a}^0 (h_n(\theta), f(\theta)) d\theta \right| \leq \|h_n\|_{L^2} \|f\|_{L^2}$$

and

$$\lim_{n \rightarrow \infty} \int_{-a}^0 (h_n(\theta), f(\theta)) d\theta \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

As a result equation (A.24) yields

$$\begin{aligned}
 0 = & \left(A_{00}h^0, B^{00}k_i(0) + \left\{ \int_{\theta_i}^{\theta_{i-1}} + \int_{\theta_1}^0 \right\} B^{01}(\theta)k_i(\theta) d\theta \right) \\
 & + \left(h^0, B^{10}(0)k_i(0) + \left\{ \int_{\theta_i}^{\theta_{i-1}} + \int_{\theta_1}^0 \right\} B^{11}(0, \theta)k_i(\theta) d\theta \right) \\
 (A.25) \quad & + \left(B^{00}h^0, A_{00}k_i(0) + \left\{ \int_{\theta_i}^{\theta_{i-1}} + \int_{\theta_1}^0 \right\} A_{01}(\theta)k_i(\theta) d\theta \right) \\
 & + (B^{10}(0)h^0, k_i(0)) - \left\{ \int_{\theta_i}^{\theta_{i-1}} + \int_{\theta_1}^0 \right\} \left(\frac{dB^{01}}{d\theta}(\theta)h^0, k_i(\theta) \right) d\theta + (h^0, k_i(0)).
 \end{aligned}$$

To obtain equation (5.19) we use

$$k_i(\theta) = k_m(\theta) = \begin{cases} k^0 \left(1 + m \frac{\theta}{a} \right), & -\frac{a}{m} \leq \theta \leq 0 \\ 0, & \text{otherwise} \end{cases},$$

where m is chosen in such a way that $m > a\theta_1^{-1}$. When we take the limit of equation (A.25) as m goes to infinity we obtain

$$([B^{00}A_{00} + B^{10}(0)^* + A_{00}^*B^{00} + B^{10}(0) + I]h^0, k^0) = 0$$

for all h^0 and k^0 in X .

To obtain equation (5.22) in the open interval (θ_i, θ_{i-1}) we choose k_i such that

$$\text{supp } k_i \subset (\theta_i, \theta_{i-1}).$$

Then equation (A.25) yields

$$0 = \int_{\theta_i}^{\theta_{i-1}} \left[\left(B^{01}(\theta)^*A_{00} + B^{11}(0, \theta)^* + A_{01}(\theta)^*B^{00} - \left(\frac{dB^{01}}{d\theta}(\theta) \right)^* \right) h^0, k_i(\theta) \right] d\theta.$$

By density of the set of absolutely continuous maps with support in (θ_i, θ_{i-1}) in $L^2(\theta_i, \theta_{i-1}; X)$ and the properties

$$(A.26) \quad B^{10}(\theta) = B^{01}(\theta)^*, \quad B^{11}(\alpha, \theta)^* = B^{11}(\theta, \alpha),$$

the above equation yields for all h^0 in X

$$\left[-\frac{dB^{10}}{d\theta}(\theta) + B^{10}(\theta)A_{00} + A_{01}(\theta)^*B^{00} + B^{11}(\theta, 0) \right] h^0 = 0,$$

almost everywhere in (θ_i, θ_{i-1}) .

To obtain (5.26) in the region

$$\{(\alpha, \theta) \in [-a, 0] \times [-a, 0] | \alpha \in (\theta_i, \theta_{i-1}), \theta \in (\theta_j, \theta_{j-1})\}$$

we choose

$$\begin{aligned}
 h &= h_i, \quad \text{supp } h_i \subset (\theta_i, \theta_{i-1}), \\
 k &= k_j, \quad \text{supp } k_j \subset (\theta_j, \theta_{j-1})
 \end{aligned}$$

and substitute in (A.22) which reduces to the following expression :

$$\begin{aligned}
 (A.27) \quad 0 = & \left(\int_{\theta_i}^{\theta_{i-1}} A_{01}(\alpha) h_i(\alpha) d\alpha, \int_{\theta_j}^{\theta_{j-1}} B^{01}(\theta) k_j(\theta) d\theta \right) \\
 & + \int_{\theta_i}^{\theta_{i-1}} \left(\frac{dh_i}{d\alpha}(\alpha), \int_{\theta_j}^{\theta_{j-1}} B^{11}(\alpha, \theta) k_j(\theta) d\theta \right) d\alpha \\
 & + \left(\int_{\theta_i}^{\theta_{i-1}} B^{01}(\alpha) h_i(\alpha) d\alpha, \int_{\theta_j}^{\theta_{j-1}} A_{01}(\theta) k_j(\theta) d\theta \right) \\
 & + \int_{\theta_j}^{\theta_{j-1}} \left(\int_{\theta_i}^{\theta_{i-1}} B^{11}(\theta, \alpha) h_i(\alpha) d\alpha, \frac{dk_j}{d\theta}(\theta) \right) d\theta.
 \end{aligned}$$

The two terms with a derivative can be integrated by parts :

$$\begin{aligned}
 & \int_{\theta_i}^{\theta_{i-1}} \left(\frac{dh_i}{d\alpha}(\alpha), \int_{\theta_j}^{\theta_{j-1}} B^{11}(\alpha, \theta) k_j(\theta) d\theta \right) d\alpha \\
 & = - \int_{\theta_i}^{\theta_{i-1}} \left(h_i(\alpha), \int_{\theta_j}^{\theta_{j-1}} \frac{\partial}{\partial \alpha} B^{11}(\alpha, \theta) k_j(\theta) d\theta \right) d\alpha
 \end{aligned}$$

and

$$\begin{aligned}
 & \int_{\theta_j}^{\theta_{j-1}} \left(\int_{\theta_i}^{\theta_{i-1}} B^{11}(\theta, \alpha) h_i(\alpha) d\alpha, \frac{dk_j}{d\theta}(\theta) \right) d\theta \\
 & = - \int_{\theta_j}^{\theta_{j-1}} \int_{\theta_i}^{\theta_{i-1}} \left(\frac{\partial}{\partial \theta} B^{11}(\theta, \alpha) h_i(\alpha) d\alpha, k_j(\theta) \right) d\theta.
 \end{aligned}$$

Finally equation (A.27) takes the form

$$\begin{aligned}
 & \int_{\theta_i}^{\theta_{i-1}} d\alpha \int_{\theta_j}^{\theta_{j-1}} d\theta \left(\left[B^{01}(\theta)^* A_{01}(\alpha) - \left(\frac{\partial B^{11}}{\partial \alpha}(\alpha, \theta) \right)^* + A_{01}(\theta)^* B^{01}(\alpha) \right. \right. \\
 & \quad \left. \left. - \frac{\partial B^{11}}{\partial \theta}(\theta, \alpha) \right] h_i(\alpha), k_j(\theta) \right).
 \end{aligned}$$

By using relations (A.26) and the density argument we obtain

$$\frac{\partial B^{11}}{\partial \theta}(\alpha, \theta) + \frac{\partial B^{11}}{\partial \alpha}(\alpha, \theta) = A_{01}(\alpha)^* B^{10}(\theta)^* + B^{10}(\alpha) A_{01}(\theta)$$

for almost all (α, θ) in $(\theta_i, \theta_{i-1}) \times (\theta_j, \theta_{j-1})$.

(iii) We now solve equation (5.26) with boundary conditions (5.27). We let $\eta = \alpha - \beta$ and consider two cases. First let $a \geq \eta \geq 0$; then

$$-a \leq \beta \leq 0 \Rightarrow \eta - a \leq \alpha \leq 0.$$

If we change the variable β to $\eta = \alpha - \beta$, equation (5.26) becomes

$$\begin{aligned}
 \frac{d}{d\alpha} B^{11}(\alpha, \alpha - \eta) = & A_{01}(\alpha)^* B^{10}(\alpha - \eta)^* + B^{10}(\alpha) A_{01}(\alpha - \eta) \\
 & + \sum_{i=1}^{N-1} A_i^* B^{10}(\alpha - \eta)^* \delta(\alpha - \theta_i) + \sum_{j=1}^{N-1} B^{10}(\alpha) A_j \delta(\alpha - \eta - \theta_j).
 \end{aligned}$$

This last equation can be integrated from $\eta - a$ to α :

$$\begin{aligned}
 B^{11}(\alpha, \alpha - \eta) &= B^{11}(\eta - a, -a) \\
 &+ \int_{\eta-a}^{\alpha} A_{01}(\xi)^* B^{10}(\xi - \eta)^* d\xi + \int_{\eta-a}^{\alpha} B^{10}(\xi) A_{01}(\xi - \eta) d\xi \\
 &+ \sum_{i=1}^{N-1} \left\{ A_i^* B^{10}(\theta_i - \eta)^*, \quad \eta - a \leq \theta_i < \alpha \right\} \\
 &\quad \left\{ 0, \quad \text{otherwise} \right\} \\
 &+ \sum_{j=1}^{N-1} \left\{ B^{10}(\eta + \theta_j) A_j, \quad \eta - a \leq \eta + \theta_j < \alpha \right\} \\
 &\quad \left\{ 0, \quad \text{otherwise} \right\}.
 \end{aligned}$$

Finally for $\alpha \geq \beta$,

$$\begin{aligned}
 B^{11}(\alpha, \beta) &= B^{10}(\alpha - \beta - a) A_N + \sum_{j=1}^{N-1} \left\{ B^{10}(\alpha - \beta + \theta_j) A_j, \quad \theta_j < \beta \right\} \\
 &\quad \left\{ 0, \quad \text{otherwise} \right\} \\
 &+ \sum_{i=1}^{N-1} \left\{ A_i^* B^{10}(\beta - \alpha + \theta_i)^*, \quad -a \leq \beta - \alpha + \theta_i, \theta_i < \alpha \right\} \\
 &\quad \left\{ 0, \quad \text{otherwise} \right\} \\
 &+ \int_{-a}^{\alpha} \left\{ A_{01}(\xi)^* B^{10}(\beta - \alpha + \xi)^*, \quad \xi \geq \alpha - \beta - a \right\} d\xi \\
 &\quad \left\{ 0, \quad \text{otherwise} \right\} \\
 &+ \int_{-a}^{\beta} B^{10}(\alpha - \beta + \theta) A_{01}(\theta) d\theta,
 \end{aligned}$$

$$\begin{aligned}
 B^{11}(\alpha, \beta) &= B^{10}(\alpha - \beta - a) A_N \\
 &+ \sum_{j=1}^{N-1} \left\{ B^{10}(\alpha - \beta + \theta_j) A_j, \quad \theta_j < \beta, -a \leq \alpha - \beta + \theta_j \right\} \\
 &\quad \left\{ 0, \quad \text{otherwise} \right\} \\
 &+ \sum_{i=1}^{N-1} \left\{ A_i^* B^{10}(\beta - \alpha + \theta_i)^*, \quad -a \leq \beta - \alpha + \theta_i, \theta_i < \alpha \right\} \\
 &\quad \left\{ 0, \quad \text{otherwise} \right\} \\
 &+ \int_{-a}^{\alpha} \left\{ A_{01}(\xi)^* B^{10}(\beta - \alpha + \xi)^*, \quad \xi \geq \alpha - \beta - a \right\} d\xi \\
 &\quad \left\{ 0, \quad \text{otherwise} \right\} \\
 &+ \int_{-a}^{\beta} \left\{ B^{10}(\alpha - \beta + \theta) A_{01}(\theta), \quad \theta \geq \beta - \alpha - a \right\} d\theta.
 \end{aligned}$$

Notice that in the above expression for $B^{11}(\alpha, \beta)$ all terms but the first are symmetrical. Hence for $\alpha \leq \beta$ we shall obtain the same expression with the exception of the first term which will be equal to

$$A_N^* B^{10}(\beta - \alpha - a)^*.$$

But

$$\lim_{\alpha \leq \beta, \beta \rightarrow \alpha} B^{10}(\alpha - \beta - a) A_N = B^{10}(-a) A_N = A_N^* B^{00} A_N$$

and

$$\lim_{\beta \leq \alpha, \beta \rightarrow \alpha} A_N^* B^{10}(\beta - \alpha - a)^* = A_N^* B^{10}(-a)^* = A_N^* B^{00} A_N$$

imply that this first term is continuous at (α, α) , $-a \leq \alpha < \theta_{N-1}$. This makes it possible to write the first term as follows:

$$\begin{aligned} B^{10}(\alpha - \beta - a)A_N, & \quad \alpha \geq \beta, \\ A_N^* B^{10}(\beta - \alpha - a)^*, & \quad \alpha < \beta. \end{aligned}$$

This yields identity (5.28).

Appendix B.

Proof of Theorem 6.1. The reader can find the definitions of Φ^0 , Φ^1 and $\tilde{\Phi}$ in Delfour and Mitter [7], [8], [10].

We first study Π_{00} and the kernels $\Pi_{10}(\alpha)$ and $\Pi_{11}(\alpha, \beta)$ of the operators Π_{10} and Π_{11} . Since we know where the discontinuities can occur we derive differential equations for $\Pi_{10}(\alpha)$ and $\Pi_{11}(\alpha, \beta)$. Finally we solve the equation for $\Pi_{11}(\alpha, \beta)$ and give an explicit expression for $\Pi_{11}(\alpha, \beta)$ in terms of $\Pi_{10}(\cdot)$. We shall use the following results (cf. Delfour and Mitter [7], [8], [10]):

$$(B.1) \quad [\tilde{\Phi}(t)h]^0 = \Phi^0(t)h^0 + \Phi^1(t)h^1,$$

$$(B.2) \quad \begin{aligned} [\Pi\tilde{\Phi}(t)h]^0 &= \Pi_{00}\Phi^0(t)h^0 + \int_{-a}^0 \Pi_{01}(\alpha) \begin{cases} \Phi^0(t+\alpha)h^0, & t+\alpha \geq 0 \\ 0, & \text{otherwise} \end{cases} d\alpha \\ &+ \Pi_{00}\Phi^1(t)h^1 + \int_{-a}^0 \Pi_{01}(\alpha) \begin{cases} \Phi^1(t+\alpha)h^1, & t+\alpha \geq 0 \\ h^1(t+\alpha), & \text{otherwise} \end{cases} d\alpha. \end{aligned}$$

(i) Let $h = (h^0, 0)$ and $k = (k^0, 0)$ in (6.1). Then

$$(B.3) \quad \begin{aligned} (\Pi_{00}h^0, k^0) &= \int_0^\infty \left\{ (Q\Phi^0(t)h^0, \Phi^0(t)k^0) \right. \\ &+ \left(R \left[\Pi_{00}\Phi^0(t)h^0 + \int_{-\min(t,a)}^0 \Pi_{01}(\alpha)\Phi^0(t+\alpha)h^0 d\alpha \right], \right. \\ &\quad \left. \Pi_{00}\Phi^0(t)k^0 + \int_{-\min(t,a)}^0 \Pi_{01}(\beta)\Phi^0(t+\beta)k^0 d\beta \right\} dt. \end{aligned}$$

Notice that

$$\int_0^\infty dt \int_{-\min(a,t)}^0 d\alpha = \int_{-a}^0 d\alpha \int_{-\alpha}^\infty dt$$

and

$$\int_0^\infty dt \int_{-\min(a,t)}^0 d\beta \int_{-\min(a,t)}^0 d\alpha = \int_{-a}^0 d\beta \int_{-a}^0 d\alpha \begin{cases} \int_{-\alpha}^\infty dt, & \alpha \leq \beta \\ \int_{-\beta}^\infty dt, & \alpha > \beta \end{cases}.$$

Hence

$$\begin{aligned}
 \Pi_{00} &= \int_0^\infty \Phi^0(t)^* [Q + \Pi_{00} R \Pi_{00}] \Phi^0(t) dt \\
 &+ \int_{-a}^0 d\beta \int_{-\beta}^\infty dt \Phi^0(t + \beta)^* \Pi_{01}(\beta)^* R \Pi_{00} \Phi^0(t) \\
 \text{(B.4)} \quad &+ \int_{-a}^0 d\alpha \int_{-\alpha}^\infty dt \Phi^0(t)^* \Pi_{00} R \Pi_{01}(\alpha) \Phi^0(t + \alpha) \\
 &+ \int_{-a}^0 d\alpha \int_{-a}^0 d\beta \left\{ \int_{-\alpha}^\infty dt \Phi^0(t + \beta)^* \Pi_{01}(\beta)^* R \Pi_{01}(\alpha) \Phi^0(t + \alpha), \quad \alpha \leq \beta \right\} \\
 &\left\{ \int_{-\beta}^\infty dt \Phi^0(t + \beta)^* \Pi_{01}(\beta)^* R \Pi_{01}(\alpha) \Phi^0(t + \alpha), \quad \alpha > \beta \right\}.
 \end{aligned}$$

Let $h = (0, h^1)$, $k = (k^0, 0)$ in (6.1). Then

$$\begin{aligned}
 \Pi_{01} h^1 &= \int_0^\infty \Phi^0(t)^* Q \Phi^1(t) h^1 dt \\
 &+ \int_0^\infty \left[\Phi^0(t)^* \Pi_{00}^* + \int_{-\min(t,a)}^0 \Phi^0(t + \theta)^* \Pi_{01}(\theta)^* d\theta \right] \\
 &\cdot R \left[\Pi_{00} \Phi^1(t) h^1 + \int_{-a}^0 \Pi_{01}(\alpha) \left\{ \begin{array}{l} \Phi^1(t + \alpha) h^1, \quad t + \alpha \geq 0 \\ h^1(t + \alpha), \quad \text{otherwise} \end{array} \right\} d\alpha \right] dt \\
 \text{(B.5)} \quad &= \int_0^\infty \Phi^0(t)^* Q \Phi^1(t) h^1 dt \\
 &+ \int_0^\infty \left[\Phi^0(t)^* \Pi_{00}^* + \int_{-\min(t,a)}^0 \Phi^0(t + \theta)^* \Pi_{01}(\theta)^* d\theta \right] \\
 &\cdot R \left[\Pi_{00} \Phi^1(t) h^1 + \int_{-\min(t,a)}^0 \Pi_{01}(\alpha) \Phi^1(t + \alpha) h^1 d\alpha \right] dt \\
 &+ \int_0^\infty \left[\Phi^0(t)^* \Pi_{00}^* + \int_{-\min(t,a)}^0 \Phi^0(t + \theta)^* \Pi_{01}(\theta)^* d\theta \right] \\
 &\cdot R \int_{-a}^{-\min(t,a)} \Pi_{01}(\alpha) h^1(t + \alpha) d\alpha dt.
 \end{aligned}$$

In view of (B.3) and (B.4) and the fact that

$$\Phi^1(t + \alpha)h^1 = \int_{-a}^0 \Phi^1(t + \alpha, \xi)h^1(\xi) d\xi,$$

$$\begin{aligned} \Pi_{01}(\xi) &= \int_0^\infty dt \Phi^0(t)^*[Q + \Pi_{00}R\Pi_{00}]\Phi^1(t, \xi) \\ &+ \int_{-a}^0 d\theta \int_{-\theta}^\infty dt \Phi^0(t + \theta)^*\Pi_{01}(\theta)^*R\Pi_{00}\Phi^1(t, \xi) \\ &+ \int_{-a}^0 d\alpha \int_{-\alpha}^\infty dt \Phi^0(t)^*\Pi_{00}^*R\Pi_{01}(\alpha)\Phi^1(t + \alpha, \xi) \\ (B.6) \quad &+ \int_{-a}^0 d\alpha \int_{-a}^0 d\theta \left\{ \begin{aligned} &\int_{-\alpha}^\infty dt \Phi^0(t + \theta)^*\Pi_{01}(\theta)^*R\Pi_{01}(\alpha)\Phi^1(t + \alpha, \xi), \quad \alpha \leq \theta \\ &\int_{-\theta}^\infty dt \Phi^0(t + \theta)^*\Pi_{01}(\theta)^*R\Pi_{01}(\alpha)\Phi^1(t + \alpha, \xi), \quad \alpha > \theta \end{aligned} \right\} \\ &+ \int_{-a}^\xi d\alpha \left[\Phi^0(\xi - \alpha)^*\Pi_{00}^* + \int_{\alpha - \xi}^0 \Phi^0(\xi - \alpha + \theta)^*\Pi_{01}(\theta)^* d\theta \right] R\Pi_{01}(\alpha), \end{aligned}$$

where the last term is obtained from the last term in (B.5) after changes in the order of integration and changes of variable:

$$\begin{aligned} \int_0^\infty dt \int_{-a}^{-\min(a,t)} d\alpha &= \int_{-a}^0 d\alpha \int_0^{-\alpha} dt = \int_{-a}^0 d\alpha \int_\alpha^0 d\xi \quad \text{with } \xi = t + \alpha, \\ \int_{-a}^0 d\alpha \int_\alpha^0 d\xi \int_{-\min(a,\xi-\alpha)}^0 &= \int_{-a}^0 d\xi \int_{-a}^\xi d\alpha \int_{-\min(a,\xi-\alpha)}^0 d\theta = \int_{-a}^0 d\xi \int_{-a}^\xi d\alpha \int_{\alpha-\xi}^0 d\theta, \end{aligned}$$

since $0 \leq \xi - \alpha \leq a + \xi \leq a$.

We can now use the identity (cf. Delfour and Mitter [7], [8])

$$\begin{aligned} \Phi^1(t, \xi) &= \sum_{i=1}^N \left\{ \begin{aligned} &\Phi^0(t - \xi + \theta_i)A_i, \quad t \geq \xi - \theta_i \geq 0 \\ &0, \quad \text{otherwise} \end{aligned} \right\} \\ (B.7) \quad &+ \int_{\max\{-a, \xi-t\}}^\xi \Phi^0(t - \xi + \theta)[A_{01}(\theta) - R\Pi_{01}(\theta)] d\theta \end{aligned}$$

to eliminate Φ^1 of the expression (B.6) for $\Pi_{01}(\xi)$. The term $\Phi^1(t, \xi)$ will always be integrated with respect to t and the only discontinuities that can occur are at $\xi = \theta_i$ where $\Phi^1(t, \xi)$ has a jump of height $\Phi^0(t)A_i$, $i = 1, \dots, N - 1$. This will

produce a jump in $\Pi_{01}(\xi)$ of height

$$\begin{aligned}
 & \int_0^\infty dt \Phi^0(t)^* [Q + \Pi_{00} R \Pi_{00}] \Phi^0(t) A_i \\
 & + \int_{-a}^0 d\theta \int_{-\theta}^\infty dt \Phi^0(t + \theta)^* \Pi_{01}(\theta)^* R \Pi_{00} \Phi^0(t) A_i \\
 & + \int_{-a}^0 d\alpha \int_{-a}^\infty dt \Phi^0(t)^* \Pi_{00}^* R \Pi_{01}(\alpha) \Phi^0(t + \alpha) A_i \\
 (B.8) \quad & + \int_{-a}^0 d\alpha \int_{-a}^0 d\theta \left\{ \int_{-\alpha}^\infty dt \Phi^0(t + \theta)^* \Pi_{01}(\theta)^* R \Pi_{01}(\alpha) \Phi^0(t + \alpha) A_i, \quad \alpha \leq \theta \right. \\
 & \left. \int_{-\theta}^\infty dt \Phi^0(t + \theta)^* \Pi_{01}(\theta)^* R \Pi_{01}(\alpha) \Phi^0(t + \alpha) A_i, \quad \alpha > \theta \right\} \\
 & = \Pi_{00} A_i
 \end{aligned}$$

at the points $\xi = \theta_i, i = 1, \dots, N - 1$. Moreover as $\xi \rightarrow -a$,

$$\lim_{\xi \rightarrow -a} \Phi^1(t, \xi) = \begin{cases} \Phi^0(t) A_N, & t \geq 0, \\ 0 & , \text{ otherwise,} \end{cases}$$

and if we let

$$(B.9) \quad \Pi_{01}(-a) = \lim_{\xi \rightarrow -a} \Pi_{01}(\xi)$$

we obtain in a similar way

$$(B.10) \quad \Pi_{01}(-a) = \Pi_{00} A_N.$$

Let $h = (0, h^1)$ and $k = (0, k^1)$ in (6.1). Then

$$\begin{aligned}
 (\Pi_{11} h^1, k^1) &= \int_0^\infty (Q \Phi^1(t) h^1, \Phi^1(t) k^1) dt \\
 &+ \int_0^\infty dt \left[R \left[\Pi_{00} \Phi^1(t) h^1 + \int_{-a}^0 \Pi_{01}(\alpha) \left\{ \begin{array}{l} \Phi^1(t + \alpha) h^1, \quad t + \alpha \geq 0 \\ h^1(t + \alpha), \quad \text{otherwise} \end{array} \right\} d\alpha \right] \right. \\
 (B.11) \quad & \left. \Pi_{00} \Phi^1(t) k^1 + \int_{-a}^0 \Pi_{01}(\beta) \left\{ \begin{array}{l} \Phi^1(t + \beta) k^1, \quad t + \beta \geq 0 \\ k^1(t + \beta), \quad \text{otherwise} \end{array} \right\} d\beta \right].
 \end{aligned}$$

The right-hand side of (B.11) can be rewritten in the following form :

$$\begin{aligned}
 & \int_0^\infty ([Q + \Pi_{00} R \Pi_{00}] \Phi^1(t) h^1, \Phi^1(t) k^1) dt \\
 & + \int_0^\infty dt \int_{-a}^0 d\alpha \left(R \Pi_{01}(\alpha) \left\{ \begin{array}{l} \Phi^1(t + \alpha) h^1, \quad t + \alpha \geq 0 \\ h^1(t + \alpha), \quad \text{otherwise} \end{array} \right\}, \Pi_{00} \Phi^1(t) k^1 \right) \\
 & + \int_0^\infty dt \int_{-a}^0 d\beta \left(R \Pi_{00} \Phi^1(t) h^1, \Pi_{01}(\beta) \left\{ \begin{array}{l} \Phi^1(t + \beta) k^1, \quad t + \beta \geq 0 \\ k^1(t + \beta), \quad \text{otherwise} \end{array} \right\} \right) \quad (cont.)
 \end{aligned}$$

$$+ \int_0^\infty dt \int_{-a}^0 d\alpha \int_{-a}^0 d\beta \left(R\Pi_{01}(\alpha) \begin{cases} \Phi^1(t + \alpha)h^1, & t + \alpha \geq 0 \\ h^1(t + \alpha), & \text{otherwise} \end{cases}, \right. \\ \left. \Pi_{01}(\beta) \begin{cases} \Phi^1(t + \beta)k^1, & t + \beta \geq 0 \\ h^1(t + \beta), & \text{otherwise} \end{cases} \right)$$

and

$$(\Pi_{11}h^1, k^1)$$

$$= \int_0^\infty dt ([Q + \Pi_{00}R\Pi_{00}]\Phi^1(t)h^1, \Phi^1(t)k^1) \\ + \int_0^\infty dt \int_{-\min(a,t)}^0 d\alpha (R\Pi_{01}(\alpha)\Phi^1(t + \alpha)h^1, \Pi_{00}\Phi^1(t)k^1) \\ + \int_0^\infty dt \int_{-\min(a,t)}^0 d\beta (R\Pi_{00}\Phi^1(t)h^1, \Pi_{01}(\beta)\Phi^1(t + \beta)k^1) \\ + \int_0^\infty dt \int_{-\min(a,t)}^0 d\alpha \int_{-\min(a,t)}^0 d\beta (R\Pi_{01}(\alpha)\Phi^1(t + \alpha)h^1, \Pi_{01}(\beta)\Phi^1(t + \beta)k^1)$$

$$(B.12) \quad + \int_0^\infty dt \int_{-a}^{-\min(a,t)} d\alpha (R\Pi_{01}(\alpha)h^1(t + \alpha), \Pi_{00}\Phi^1(t)k^1) \\ + \int_0^\infty dt \int_{-a}^{-\min(a,t)} d\beta (\Pi_{00}\Phi^1(t)h^1, \Pi_{01}(\beta)k^1(t + \beta)) \\ + \int_0^\infty dt \int_{-a}^{-\min(a,t)} d\alpha \int_{-\min(a,t)}^0 d\beta (R\Pi_{01}(\alpha)h^1(t + \alpha), \Pi_{01}(\beta)\Phi^1(t + \beta)k^1) \\ + \int_0^\infty dt \int_{-\min(a,t)}^0 d\alpha \int_{-a}^{-\min(a,t)} d\beta (R\Pi_{01}(\alpha)\Phi^1(t + \alpha)h^1, \Pi_{01}(\beta)k^1(t + \beta)) \\ + \int_0^\infty dt \int_{-a}^{-\min(a,t)} d\alpha \int_{-a}^{-\min(a,t)} d\beta (R\Pi_{01}(\alpha)h^1(t + \alpha), \Pi_{01}(\beta)k^1(t + \beta)).$$

We number ①, ②, \dots , ⑤ the last five terms in the right-hand side of (B.12).

Since

$$\int_0^\infty dt \int_{-a}^{-\min(a,t)} d\alpha = \int_{-a}^0 d\alpha \int_0^{-\alpha} dt = \int_{-a}^0 d\alpha \int_\alpha^0 d\xi = \int_{-a}^0 d\xi \int_{-a}^\xi d\alpha$$

with the change of variable t to $\xi = t + \alpha$,

$$(B.13) \quad \textcircled{1} = \int_{-a}^0 d\xi \int_{-a}^\xi d\alpha (R\Pi_{01}(\alpha)h^1(\xi), \Pi_{00}\Phi^1(\xi - \alpha)k^1).$$

Similarly,

$$(B.14) \quad \textcircled{2} = \int_{-a}^0 d\theta \int_{-a}^\theta d\beta (R\Pi_{00}\Phi^1(\theta - \beta)h^1, \Pi_{01}(\beta)k^1(\theta)).$$

Also,

$$\int_0^\infty dt \int_{-a}^{-\min(a,t)} d\alpha \int_{-\min(a,t)}^0 d\beta = \int_0^a dt \int_{-a}^{-t} d\alpha \int_{-t}^0 d\beta = \int_{-a}^0 d\alpha \int_0^{-\alpha} dt \int_{-t}^0 d\beta$$

and the change of variable t to $\xi = t + \alpha$ yields

$$\begin{aligned}
 \textcircled{3} &= \int_{-a}^0 d\alpha \int_{\alpha}^0 d\xi \int_{\alpha-\xi}^0 d\beta (R\Pi_{01}(\alpha)h^1(\xi), \Pi_{01}(\beta)\Phi^1(\xi - \alpha + \beta)k^1) \\
 \text{(B.15)} \quad &= \int_{-a}^0 d\xi \int_{-a}^{\xi} d\alpha \int_{\alpha-\xi}^0 d\beta (R\Pi_{01}(\alpha)h^1(\xi), \Pi_{01}(\beta)\Phi^1(\xi - \alpha + \beta)k^1).
 \end{aligned}$$

Similarly,

$$\text{(B.16)} \quad \textcircled{4} = \int_{-a}^0 d\theta \int_{-a}^{\theta} d\beta \int_{\beta-\theta}^0 d\alpha (R\Pi_{01}(\alpha)\Phi^1(\theta - \beta + \alpha)h^1, \Pi_{01}(\beta)k^1(\theta)).$$

Finally,

$$\int_0^{\infty} dt \int_{-a}^{-\min(a,t)} d\alpha \int_{-a}^{-\min(a,t)} d\beta = \int_{-a}^0 d\alpha \int_{\alpha}^0 d\xi \int_{-a}^{\alpha-\xi} d\beta$$

with the change of the variable t to $\xi = t + \alpha$ and

$$\textcircled{5} = \int_{-a}^0 d\alpha \int_{\alpha}^0 d\xi \int_{-a}^{\alpha-\xi} d\beta (R\Pi_{01}(\alpha)h^1(\xi), \Pi_{01}(\beta)k^1(\xi - \alpha + \beta)).$$

We change the variable β to $\theta = \xi - \alpha + \beta$,

$$\textcircled{5} = \int_{-a}^0 d\alpha \int_{\alpha}^0 d\xi \int_{\xi-\alpha-a}^0 d\theta (R\Pi_{01}(\alpha)h^1(\xi), \Pi_{01}(\alpha - \xi + \theta)k^1(\theta)),$$

and change the order of integration. But

$$\begin{aligned}
 \int_{-a}^0 d\alpha \int_{\alpha}^0 d\xi \int_{\xi-\alpha-a}^0 d\theta &= \int_{-a}^0 d\xi \int_{-a}^{\xi} d\alpha \int_{\xi-\alpha-a}^0 d\theta \\
 &= \int_{-a}^0 d\xi \int_{-a}^0 d\theta \left\{ \int_{-a}^{\xi} d\alpha, \quad \xi \leq \theta \right. \\
 &\quad \left. \int_{\xi-\theta-a}^{\xi} d\alpha, \quad \xi > \theta \right\},
 \end{aligned}$$

and by changing once more the variable α to $\alpha - \xi$ we finally obtain

$$\text{(B.17)} \quad \textcircled{5} = \int_{-a}^0 d\xi \int_{-a}^0 d\theta \left\{ \int_{-a-\xi}^0 d\alpha (R\Pi_{01}(\alpha + \xi)h^1(\xi), \Pi_{01}(\alpha + \theta)k^1(\theta)), \quad \xi \leq \theta \right. \\
 \left. \int_{-a-\theta}^0 d\alpha (R\Pi_{01}(\alpha + \xi)h^1(\xi), \Pi_{01}(\alpha + \theta)k^1(\theta)), \quad \xi > \theta \right\}.$$

By analogy with equations (B.3) and (B.4) and with the help of equations (B.13)

to (B.17), identity (B.12) yields

$$\begin{aligned}
 & \Pi_{11}(\xi, \theta) \\
 &= \int_0^\infty \Phi^1(t, \xi)^* [Q + \Pi_{00} R \Pi_{00}] \Phi^1(t, \theta) dt \\
 &+ \int_{-a}^0 d\alpha \int_{-\alpha}^\infty dt \Phi^1(t + \alpha, \xi)^* \Pi_{01}(\alpha)^* R \Pi_{00} \Phi^1(t, \theta) \\
 &+ \int_{-a}^0 d\beta \int_{-\beta}^\infty dt \Phi^1(t, \xi)^* \Pi_{00} R \Pi_{01}(\beta) \Phi^1(t + \beta, \theta) \\
 &+ \int_{-a}^0 d\alpha \int_{-a}^0 d\beta \left\{ \begin{aligned} & \int_{-\alpha}^\infty dt \Phi^1(t + \alpha, \xi)^* \Pi_{01}(\alpha)^* R \Pi_{01}(\beta) \Phi^1(t + \beta, \theta), \quad \alpha \leq \beta \\ & \int_{-\beta}^\infty dt \Phi^1(t + \alpha, \xi)^* \Pi_{01}(\alpha)^* R \Pi_{01}(\beta) \Phi^1(t + \beta, \theta), \quad \alpha > \beta \end{aligned} \right\} \\
 &+ \int_{-a}^\xi d\alpha \Pi_{01}(\alpha)^* R \Pi_{00} \Phi^1(\xi - \alpha, \theta) + \int_{-a}^\theta d\beta \Phi^1(\theta - \beta, \xi)^* \Pi_{00} R \Pi_{01}(\beta) \\
 &+ \int_{-a}^\xi d\alpha \int_{\alpha - \xi}^0 d\beta \Pi_{01}(\alpha)^* R \Pi_{01}(\beta) \Phi^1(\xi - \alpha + \beta, \theta) \\
 &+ \int_{-a}^\theta d\beta \int_{\beta - \theta}^0 d\alpha \Phi^1(\theta - \beta + \alpha, \xi)^* \Pi_{01}(\alpha)^* R \Pi_{01}(\beta) \\
 &+ \left\{ \begin{aligned} & \int_{-a - \xi}^0 d\alpha \Pi_{01}(\alpha + \xi)^* R \Pi_{01}(\alpha + \theta), \quad \xi \leq \theta \\ & \int_{-a - \theta}^0 d\alpha \Pi_{01}(\alpha + \xi)^* R \Pi_{01}(\alpha + \theta), \quad \xi < \theta \end{aligned} \right\}.
 \end{aligned}
 \tag{B.18}$$

In the light of identity (B.7), $\Phi^1(t, \xi)$ has discontinuities of height $\Phi^0(t)A_i$ at $\xi = \theta_i$, $i = 1, \dots, N - 1$, and

$$\lim_{\xi \rightarrow -a} \Phi^1(t, \xi) = \Phi^0(t)A_N.
 \tag{B.19}$$

Fix θ and consider the map $\xi \mapsto \Pi_{11}(\xi, \theta)$. Since everywhere $\Phi^1(t, \xi)$ is integrated with respect to t , discontinuities can only occur at $\xi = \theta_i$, $i = 1, \dots, N - 1$. At $\xi = \theta_i$, $\Pi_{11}(\xi, \theta)$ has a jump of height

$$\begin{aligned}
 & \Pi_{11}(\xi, \theta) \\
 &= \int_0^\infty A_i^* \Phi^0(t)^* [Q + \Pi_{00} R \Pi_{00}] \Phi^1(t, \theta) dt \\
 &+ \int_{-a}^0 d\alpha \int_{-\alpha}^\infty dt A_i^* \Phi^0(t + \alpha)^* \Pi_{01}(\alpha)^* R \Pi_{00} \Phi^1(t, \theta) \\
 &+ \int_{-a}^0 d\beta \int_{-\beta}^\infty dt A_i^* \Phi^0(t)^* \Pi_{00} R \Pi_{01}(\beta) \Phi^1(t + \beta, \theta)
 \end{aligned}
 \tag{B.20}$$

(cont.)

$$\begin{aligned}
& + \int_{-a}^0 d\alpha \int_{-a}^0 d\beta \left\{ \int_{-\alpha}^{\infty} dt A_i^* \Phi^0(t + \alpha)^* \Pi_{01}(\alpha)^* R \Pi_{01}(\beta) \Phi^1(t + \beta, \theta), \alpha \leq \beta \right\} \\
& \left\{ \int_{-\beta}^{\infty} dt A_i^* \Phi^0(t + \alpha)^* \Pi_{01}(\alpha)^* R \Pi_{01}(\beta) \Phi^1(t + \beta, \theta), \alpha > \beta \right\} \\
& + \int_{-a}^{\theta} d\beta \int_{\beta-\theta}^0 d\alpha A_i^* \Phi^0(\theta - \beta + \alpha)^* \Pi_{01}(\alpha)^* R \Pi_{01}(\beta) \\
& + \int_{-a}^{\theta} d\beta A_i^* \Phi^0(\theta - \beta)^* \Pi_{00} R \Pi_{01}(\beta) = A_i^* \Pi_{01}(\theta) = A_i^* \Pi_{10}(\theta)^*.
\end{aligned}$$

By symmetry for each ξ the map $\theta \mapsto \Pi_{11}(\xi, \theta)$ has jump discontinuities of height $\Pi_{10}(\xi)A_j$ at $\theta = \theta_j, j = 1, \dots, N-1$. As for the boundary conditions we fix θ and evaluate

$$\lim_{\xi \rightarrow -a} \Pi_{11}(\xi, \theta) = \Pi_{11}(-a, \theta)$$

using (B.19). This yields

$$(B.21) \quad \Pi_{11}(-a, \theta) = A_N^* \Pi_{01}(\theta) = A_N^* \Pi_{10}(\theta)^*,$$

and by symmetry

$$(B.22) \quad \Pi_{11}(\xi, -a) = \Pi_{11}(-a, \xi)^* = \Pi_{01}(\xi)A_N.$$

(ii) Now that we know where the jumps are we can derive equations (6.3), (6.6) and (6.10). Our starting point is the Riccati equation

$$(B.23) \quad 0 = (\tilde{A}h, \Pi k) + (\Pi k, \tilde{A}k) - (\tilde{R}\Pi h, \Pi k) + (\tilde{Q}h, k),$$

or in expanded form

$$\begin{aligned}
& \left(A_{00}h(0) + \sum_{i=1}^N A_i h(\theta_i) + \int_{-a}^0 A_{01}(\alpha)h(\alpha) d\alpha, \Pi_{00}k(0) + \int_{-a}^0 \Pi_{01}(\theta)k(\theta) d\theta \right) \\
& + \int_{-a}^0 \left(\frac{dh}{d\alpha}(\alpha), \Pi_{10}(\alpha)k(0) + \int_{-a}^0 \Pi_{11}(\alpha, \theta)k(\theta) d\theta \right) d\alpha \\
& + \left(\Pi_{00}h(0) + \int_{-a}^0 \Pi_{01}(\alpha)h(\alpha) d\alpha, A_{00}k(0) + \sum_{i=1}^N A_i k(\theta_i) + \int_{-a}^0 A_{01}(\theta)k(\theta) d\theta \right) \\
& + \int_{-a}^0 \left(\Pi_{10}(\theta)h(0) + \int_{-a}^0 \Pi_{11}(\theta, \alpha)h(\alpha) d\alpha, \frac{dk}{d\theta}(\theta) \right) d\theta \\
& - \left(\Pi_{00}h(0) + \int_{-a}^0 \Pi_{01}(\alpha)h(\alpha) d\alpha, R \left[\Pi_{00}k(0) + \int_{-a}^0 \Pi_{01}(\theta)k(\theta) d\theta \right] \right) \\
& + (Qh(0), k(0)).
\end{aligned}$$

Let

$$(B.25) \quad h_n(\theta) = \begin{cases} h^0 \left(1 + n \frac{\theta}{a} \right), & -\frac{a}{n} \leq \theta < 0 \\ 0, & \text{otherwise} \end{cases},$$

where n is chosen in such a way that $n > a\theta_1^{-1}$. Then

$$h_n(0) \rightarrow h^0 \quad \text{and} \quad h_n \rightarrow 0 \text{ in } L^2(-a, 0; X).$$

Let k_i be chosen in $W^{1,2}(-a, 0; X)$ in such a way that

$$\text{supp } k_i \subset (\theta_i, \theta_{i-1}) \cup (\theta_1, 0].$$

Let $h = h_n$ and $k = k_i$ in (B.24):

$$\begin{aligned} & \left(A_{00}h^0 + \int_{\theta_1}^0 A_{01}(\theta)h_n(\theta) d\theta, \Pi_{00}k_i(0) + \left\{ \int_{\theta_i}^{\theta_{i-1}} + \int_{\theta_1}^0 \right\} \Pi_{01}(\theta)k_i(\theta) d\theta \right) \\ & + \int_{\theta_1}^0 \left(\frac{dh_n}{d\alpha}(\alpha), \Pi_{01}(\alpha)k_i(0) + \left\{ \int_{\theta_i}^{\theta_{i-1}} + \int_{\theta_1}^0 \right\} \Pi_{11}(\alpha, \theta)k_i(\theta) d\theta \right) d\alpha \\ & + \left(\Pi_{00}h^0 + \int_{\theta_1}^0 \Pi_{01}(\alpha)h_n(\alpha) d\alpha, A_{00}k_i(0) + \left\{ \int_{\theta_i}^{\theta_{i-1}} + \int_{\theta_1}^0 \right\} A_{01}(\theta)k_i(\theta) d\theta \right) \\ (B.26) \quad & + \left\{ \int_{\theta_i}^{\theta_{i-1}} + \int_{\theta_1}^0 \right\} \left(\Pi_{10}(\theta)h^0 + \int_{\theta_1}^0 \Pi_{11}(\theta, \alpha)h_n(\alpha) d\alpha, \frac{dk_i}{d\theta}(\theta) \right) d\theta - \left(\Pi_{00}h^0 \right. \\ & + \int_{\theta_1}^0 \Pi_{01}(\alpha)h_n(\alpha) d\alpha, R \left[\Pi_{00}k_i(0) + \left\{ \int_{\theta_i}^{\theta_{i-1}} + \int_{\theta_1}^0 \right\} \Pi_{01}(\theta)k_i(\theta) d\theta \right] \right) \\ & + (Qh^0, k_i(0)) = 0. \end{aligned}$$

Since $\alpha \mapsto \Pi_{01}(\alpha)$, $\alpha \mapsto \Pi_{11}(\alpha, \theta)$ and $\theta \mapsto \Pi_{11}(\alpha, \theta)$ are absolutely continuous in (θ_i, θ_{i-1}) and $(\theta_1, 0)$ we can now integrate by parts.

Equation (B.26) now reduces to

$$\begin{aligned} & \left(A_{00}h^0 + \int_{\theta_1}^0 A_{01}(\theta)h_n(\theta) d\theta, \Pi_{00}k_i(0) + \left\{ \int_{\theta_i}^{\theta_{i-1}} + \int_{\theta_1}^0 \right\} \Pi_{01}(\theta)k_i(\theta) d\theta \right) \\ & + \left(h^0, \Pi_{10}(0)k_i(0) + \left\{ \int_{\theta_i}^{\theta_{i-1}} + \int_{\theta_1}^0 \right\} \Pi_{11}(0, \theta)k_i(\theta) d\theta \right) \\ & - \int_{\theta_1}^0 \left(h_n(\alpha), \Pi_{10}(\alpha)k_i(0) + \left\{ \int_{\theta_i}^{\theta_{i-1}} + \int_{\theta_1}^0 \right\} \Pi_{11}(\alpha, \theta)k_i(\theta) d\theta \right) \\ & + \left(\Pi_{00}h^0 + \int_{\theta_1}^0 \Pi_{01}(\alpha)h_n(\alpha) d\alpha, A_{00}k_i(0) + \left\{ \int_{\theta_i}^{\theta_{i-1}} + \int_{\theta_1}^0 \right\} A_{01}(\theta)k_i(\theta) d\theta \right) \\ (B.27) \quad & + \left(\Pi_{10}(0)h^0 + \int_{\theta_1}^0 \Pi_{11}(\theta, \alpha)h_n(\alpha) d\alpha, k_i(\theta) \right) \\ & - \left\{ \int_{\theta_i}^{\theta_{i-1}} + \int_{\theta_1}^0 \right\} \left(\frac{d\Pi_{01}}{d\theta}(\theta)h^0 + \int_{\theta_1}^0 \frac{\partial \Pi_{11}}{\partial \theta}(\theta, \alpha)h_n(\alpha) d\alpha, k_i(\theta) \right) d\theta - \left(\Pi_{00}h^0 \right. \\ & \quad \left. (cont.) \right) \end{aligned}$$

$$\begin{aligned}
& + \int_{\theta_i}^0 \Pi_{01}(\alpha) h_n(\alpha) d\alpha, \quad R \left[\Pi_{00} k_i(0) + \left\{ \int_{\theta_i}^{\theta_{i-1}} + \int_{\theta_1}^0 \right\} \Pi_{01}(\theta) k_i(\theta) d\theta \right] \\
& + (Qh^0, k_i(0)) = 0.
\end{aligned}$$

Notice that

$$\int_{-a}^0 |A_{01}(\theta) h_n(\theta)| d\theta \leq \left[\int_{-a}^0 |A_{01}(\theta)|^2 d\theta \right]^{1/2} \left[\int_{-a}^0 |h_n(\theta)|^2 d\theta \right]^{1/2}$$

and

$$\lim_{n \rightarrow \infty} \|h_n\|_{L^2(-a, 0; X)} = 0$$

imply that

$$\int_{-a}^0 |A_{01}(\theta) h_n(\theta)| d\theta \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Similarly given an f in $L^2(-a, 0; X)$,

$$\left| \int_{-a}^0 (h_n(\theta), f(\theta)) d\theta \right| \leq \|h_n\|_{L^2} \|f\|_{L^2}$$

and

$$\lim_{n \rightarrow \infty} \int_{-a}^0 (h_n(\theta), f(\theta)) d\theta \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

As a result equation (B.27) yields

$$\begin{aligned}
& \left(A_{00} h^0, \Pi_{00} k_i(0) + \left\{ \int_{\theta_i}^{\theta_{i-1}} + \int_{\theta_1}^0 \right\} \Pi_{01}(\theta) k_i(\theta) d\theta \right) \\
& + \left(h^0, \Pi_{10}(0) k_i(0) + \left\{ \int_{\theta_i}^{\theta_{i-1}} + \int_{\theta_1}^0 \right\} \Pi_{11}(0, \theta) k_i(\theta) d\theta \right) \\
& + \left(\Pi_{00} h^0, A_{00} k_i(0) + \left\{ \int_{\theta_i}^{\theta_{i-1}} + \int_{\theta_1}^0 \right\} A_{01}(\theta) k_i(\theta) d\theta \right) \\
& + \left(\Pi_{10}(0) h^0, k_i(0) \right) - \left\{ \int_{\theta_i}^{\theta_{i-1}} + \int_{\theta_1}^0 \right\} \left(\frac{d\Pi_{01}}{d\theta}(\theta) h^0, k_i(\theta) \right) d\theta \\
& - \left(\Pi_{00} h^0, R \left[\Pi_{00} k_i(0) + \left\{ \int_{\theta_i}^{\theta_{i-1}} + \int_{\theta_1}^0 \right\} \Pi_{01}(\theta) k_i(\theta) d\theta \right] \right) \\
& + (Qh^0, k_i(0)) = 0.
\end{aligned}
\tag{B.28}$$

To obtain equation (6.3) we use

$$k_i(\theta) = k_m(\theta) = \begin{cases} k^0 \left(1 + m \frac{\theta}{a} \right), & -\frac{a}{m} \leq \theta \leq 0 \\ 0, & \text{otherwise} \end{cases}$$

where m is chosen in such a way that $m > a\theta_1^{-1}$. When we take the limit of equation (B.28) as m goes to infinity we obtain

$$(B.29) \quad ([\Pi_{00}A_{00} + \Pi_{10}(0)^* + A_{00}^*\Pi_{00} + \Pi_{10}(0) - \Pi_{00}R\Pi_{00} + Q]h^0, k^0) = 0$$

for all h^0 and k^0 in X .

To obtain equation (6.6) in the open interval (θ_i, θ_{i-1}) we choose k_i such that

$$\text{supp } k_i \subset (\theta_i, \theta_{i-1}).$$

The equation (B.28) yields

$$0 = \int_{\theta_i}^{\theta_{i-1}} \left(\left[\Pi_{01}(\theta)^* A_{00} + \Pi_{11}(0, \theta)^* + A_{01}(\theta)^* \Pi_{00} \right. \right. \\ \left. \left. - \left(\frac{d\Pi_{01}}{d\theta}(\theta) \right)^* - \Pi_{01}(\theta)^* R \Pi_{00} \right] h^0, k_j(\theta) \right) d\theta.$$

By density of the set of absolutely continuous maps with support in (θ_i, θ_{i-1}) in $L^2(\theta_i, \theta_{i-1}; X)$ and the properties

$$(B.30) \quad \Pi_{10}(\theta) = \Pi_{01}(\theta)^*, \quad \Pi_{11}(\alpha, \theta)^* = \Pi_{11}(\theta, \alpha),$$

the above equation yields for h^0 in X ,

$$\left[-\frac{d\Pi_{10}}{d\theta}(\theta) + \Pi_{10}(\theta)[A_{00} - R\Pi_{00}] + A_{01}(\theta)^*\Pi_{00} + \Pi_{11}(\theta, 0) \right] h^0 = 0,$$

a.e. in (θ_i, θ_{i-1}) .

To obtain (6.10) in the region

$$\{(\alpha, \theta) \in [-a, 0] \times [-a, 0] | \alpha \in (\theta_i, \theta_{i-1}), \theta \in (\theta_j, \theta_{j-1})\},$$

we choose

$$h = h_i, \quad \text{supp } h_i \subset (\theta_i, \theta_{i-1}),$$

$$k = k_j, \quad \text{supp } k_j \subset (\theta_j, \theta_{j-1})$$

and substitute in (B.24) which reduces to the following expression:

$$(B.31) \quad \left(\int_{\theta_i}^{\theta_{i-1}} A_{01}(\alpha) h_i(\alpha) d\alpha, \int_{\theta_j}^{\theta_{j-1}} \Pi_{01}(\theta) k_j(\theta) d\theta \right) \\ + \int_{\theta_i}^{\theta_{i-1}} \left(\frac{dh_i}{d\alpha}(\alpha), \int_{\theta_j}^{\theta_{j-1}} \Pi_{11}(\alpha, \theta) k_j(\theta) d\theta \right) d\alpha \\ + \left(\int_{\theta_i}^{\theta_{i-1}} \Pi_{01}(\alpha) h_i(\alpha) d\alpha, \int_{\theta_j}^{\theta_{j-1}} A_{01}(\theta) k_j(\theta) d\theta \right) \\ + \int_{\theta_j}^{\theta_{j-1}} \left(\int_{\theta_i}^{\theta_{i-1}} \Pi_{11}(\theta, \alpha) h_i(\alpha) d\alpha, \frac{dk_j}{d\theta}(\theta) \right) d\theta \\ - \left(\int_{\theta_i}^{\theta_{i-1}} \Pi_{01}(\alpha) h_i(\alpha) d\alpha, R \int_{\theta_j}^{\theta_{j-1}} \Pi_{01}(\theta) k_j(\theta) d\theta \right).$$

The two terms with a derivative can be integrated by parts:

$$\begin{aligned} & \int_{\theta_i}^{\theta_{i-1}} \left(\frac{dh_i}{d\alpha}(\alpha), \int_{\theta_j}^{\theta_{j-1}} \Pi_{11}(\alpha, \theta) k_j(\theta) d\theta \right) d\alpha \\ &= - \int_{\theta_i}^{\theta_{i-1}} \left(h_i(\alpha), \int_{\theta_j}^{\theta_{j-1}} \frac{\partial \Pi_{11}}{\partial \alpha}(\alpha, \theta) k_j(\theta) d\theta \right) d\alpha \end{aligned}$$

and

$$\begin{aligned} & \int_{\theta_j}^{\theta_{j-1}} \left(\int_{\theta_i}^{\theta_{i-1}} \Pi_{11}(\theta, \alpha) h_i(\alpha) d\alpha, \frac{dk_j}{d\theta}(\theta) \right) d\theta \\ &= - \int_{\theta_j}^{\theta_{j-1}} \int_{\theta_i}^{\theta_{i-1}} \left(\frac{\partial \Pi_{11}}{\partial \theta}(\theta, \alpha) h_i(\alpha) d\alpha, k_j(\theta) \right) d\theta. \end{aligned}$$

Finally equation (B.31) takes the form

$$\begin{aligned} & \int_{\theta_i}^{\theta_{i-1}} d\alpha \int_{\theta_j}^{\theta_{j-1}} d\theta \left[\left[\Pi_{01}(\theta)^* A_{01}(\alpha) - \left(\frac{\partial \Pi_{11}}{\partial \alpha}(\alpha, \theta) \right)^* + A_{01}(\theta)^* \Pi_{01}(\alpha) \right. \right. \\ & \quad \left. \left. - \frac{\partial \Pi_{11}}{\partial \theta}(\theta, \alpha) - \Pi_{01}(\theta)^* R \Pi_{01}(\alpha) \right] h_i(\alpha), k_j(\theta) \right]. \end{aligned}$$

By using relations (B.30) and the density argument we obtain

$$\frac{\partial \Pi_{11}}{\partial \theta}(\theta, \alpha) + \frac{\partial \Pi_{11}}{\partial \alpha}(\theta, \alpha) = \Pi_{10}(\theta) A_{01}(\alpha) + A_{01}(\theta)^* \Pi_{10}(\alpha)^* - \Pi_{10}(\theta) R \Pi_{10}(\alpha)^*$$

for almost all (α, θ) in $(\theta_i, \theta_{i-1}) \times (\theta_j, \theta_{j-1})$.

(iii) We now solve equation (6.10) with boundary conditions (6.11). We let $\eta = \alpha - \beta$ and consider two cases. First let $a \geq \eta \geq 0$; then

$$-a \leq \beta \leq 0 \Rightarrow \eta - a \leq \alpha \leq 0.$$

If we change the variable β to $\eta = \alpha - \beta$, equation (6.10) becomes

$$\begin{aligned} \frac{d}{d\alpha} \Pi_{11}(\alpha, \alpha - \eta) &= A_{01}(\alpha)^* \Pi_{10}(\alpha - \eta)^* + \Pi_{10}(\alpha) A_{01}(\alpha - \eta) - \Pi_{10}(\alpha) R \Pi_{10}(\alpha - \eta) \\ &+ \sum_{i=1}^{N-1} A_i^* \Pi_{10}(\alpha - \eta)^* \delta(\alpha - \theta_i) + \sum_{j=1}^{N-1} \Pi_{10}(\alpha) A_j \delta(\alpha - \eta - \theta_j). \end{aligned}$$

This last equation can be integrated from $\eta - a$ to α :

$$\begin{aligned} \Pi_{11}(\alpha, \alpha - \eta) &= \Pi_{11}(\eta - a, -a) + \int_{\eta-a}^{\alpha} A_{01}(\xi)^* \Pi_{10}(\xi - \eta)^* d\xi \\ &+ \int_{\eta-a}^{\alpha} \Pi_{10}(\xi) [A_{01}(\xi - \eta) - R \Pi_{10}(\xi - \eta)^*] d\xi \\ &+ \sum_{i=1}^{N-1} \begin{cases} A_i^* \Pi_{10}(\theta_i - \eta)^*, & \eta - a \leq \theta_i < \alpha \\ 0, & \text{otherwise} \end{cases} \\ &+ \sum_{j=1}^{N-1} \begin{cases} \Pi_{10}(\eta + \theta_j) A_j, & \eta - a \leq \eta + \theta_j < \alpha \\ 0, & \text{otherwise} \end{cases}. \end{aligned}$$

Finally for $\alpha \geq \beta$,

$$\begin{aligned} \Pi_{11}(\alpha, \beta) &= \Pi_{10}(\alpha - \beta - a)A_N + \sum_{j=1}^{N-1} \left\{ \begin{array}{ll} \Pi_{10}(\alpha - \beta + \theta_j)A_j, & \theta_j < \beta \\ 0, & \text{otherwise} \end{array} \right\} \\ &+ \sum_{i=1}^{N-1} \left\{ \begin{array}{ll} A_i^* \Pi_{10}(\beta - \alpha + \theta_i)^*, & -a \leq \beta - \alpha + \theta_i, \theta_i < \alpha \\ 0, & \text{otherwise} \end{array} \right\} \\ &+ \int_{\alpha - \beta - a}^{\alpha} A_{01}(\xi)^* \Pi_{10}(\xi - \alpha + \beta)^* d\xi \\ &+ \int_{-a}^{\beta} \Pi_{10}(\alpha - \beta + \theta)[A_{01}(\theta) - R\Pi_{10}(\theta)^*] d\theta \end{aligned}$$

and

$$\begin{aligned} \Pi_{11}(\alpha, \beta) &= \Pi_{10}(\alpha - \beta - a)A_N \\ &+ \sum_{j=1}^{N-1} \left\{ \begin{array}{ll} \Pi_{10}(\alpha - \beta + \theta_j)A_j, & -a \leq \alpha - \beta + \theta_j, \theta_j < \beta \\ 0, & \text{otherwise} \end{array} \right\} \\ &+ \sum_{i=1}^{N-1} \left\{ \begin{array}{ll} A_i^* \Pi_{10}(\beta - \alpha + \theta_i)^*, & -a \leq \beta - \alpha + \theta_i, \theta_i < \alpha \\ 0, & \text{otherwise} \end{array} \right\} \\ &+ \int_{-a}^{\alpha} \left\{ \begin{array}{ll} A_{01}(\xi)^* \Pi_{10}(\xi - \alpha + \beta)^*, & \xi \geq \alpha - \beta - a \\ 0, & \text{otherwise} \end{array} \right\} d\xi \\ &+ \int_{-a}^{\beta} \left\{ \begin{array}{ll} \Pi_{10}(\theta - \beta + \alpha)A_{01}(\theta), & \theta \geq \beta - \alpha - a \\ 0, & \text{otherwise} \end{array} \right\} d\theta \\ &- \left\{ \begin{array}{ll} \int_{-a}^{\beta} \Pi_{10}(\alpha - \beta + \theta)R\Pi_{10}(\theta)^* d\theta, & \alpha \geq \beta \\ \int_{-a}^{\alpha} \Pi_{10}(\xi)R\Pi_{10}(\beta - \alpha + \xi)^* d\xi, & \alpha < \beta \end{array} \right\}. \end{aligned}$$

Notice that in the above expression for $B^{11}(\alpha, \beta)$ all terms but the first are symmetrical. Hence for $\alpha \leq \beta$ we shall obtain the same expression with the exception of the first term which will be equal to

$$A_N^* \Pi_{10}(\beta - \alpha - a)^*.$$

But

$$\lim_{\alpha \leq \beta, \beta \rightarrow \alpha} \Pi_{10}(\alpha - \beta - a)A_N = \Pi_{10}(-a)A_N = A_N^* \Pi_{00}A_N$$

and

$$\lim_{\beta \leq \alpha, \beta \rightarrow \alpha} A_N^* \Pi_{10}(\beta - \alpha - a)^* = A_N^* \Pi_{10}(-a)^* = A_N^* \Pi_{00}A_N$$

imply that this first term is continuous at (α, α) , $-a \leq \alpha < \theta_{N-1}$. This makes it

possible to write the first term as follows :

$$\begin{aligned}\Pi_{10}(\alpha - \beta - a)A_N, & \quad \alpha \geq \beta, \\ A_N^* \Pi_{10}(\beta - \alpha - a)^*, & \quad \alpha < \beta.\end{aligned}$$

This yields identity (6.12).

REFERENCES

- [1] R. W. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.
- [2] R. DATKO, *Extending a theorem of A. M. Lyapunov to Hilbert spaces*, J. Math. Anal. Appl., 32 (1970), pp. 610–616.
- [3] ———, *A linear control problem in an abstract Hilbert space*, J. Differential Equations, 9 (1971), pp. 346–359.
- [4] ———, *Unconstrained control problems with quadratic cost*, this Journal, 11 (1973), pp. 32–52.
- [5] M. C. DELFOUR, *Theory of differential delay systems in the space M^2 ; Stability and Lyapunov equation*, Proc. Symposium on Differential Delay and Functional Equations, Control Theory Centre Rep. 12, University of Warwick, England, 1972.
- [6] M. C. DELFOUR AND S. K. MITTER, *Hereditary differential systems with constant delays, I—General case*, J. Differential Equations, 12 (1972), pp. 213–235.
- [7] ———, *Hereditary differential systems with constant delays, II—A class of affine systems and the adjoint problem*, Rep. CRM-293, Centre de Recherches Mathématiques, Université de Montréal, Montréal, Canada.
- [8] ———, *Controllability, observability and optimal feedback control of hereditary differential systems*, this Journal, 10 (1972), pp. 298–328.
- [9] ———, *Controllability and observability for infinite dimensional systems*, this Journal, 10 (1972), pp. 329–333.
- [10] ———, *State theory of linear hereditary differential systems*, in preparation.
- [11] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Part III*, Wiley-Interscience, New York, 1971.
- [12] J. K. HALE, *Functional Differential Equations*, Springer-Verlag, New York, 1971.
- [13] R. E. KALMAN, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana, 5 (1960), pp. 102–119.
- [14] L. V. KANTOROVICH AND G. P. AKILOV, *Functional Analysis in Normed Spaces*, Pergamon Press, London, 1964.
- [15] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.
- [16] D. L. LUKES AND D. L. RUSSELL, *The quadratic criterion for distributed systems*, SIAM J. Control, 7 (1969), pp. 101–121.
- [17] YU. S. OSIPOV, *Stabilizability of controlled systems with delays*, Differential Equations, 1 (1965), pp. 463–475 (Differencial'nye Uravnenija, 1 (1965), pp. 605–618).
- [18] V. M. POPOV, *Delay-feedback, time-optimal, linear time-invariant control systems, in ordinary differential equations*, 1971 NRL-MRC Conference, L. Weiss, ed., Academic Press, New York, 1972, pp. 545–552.
- [19] D. W. ROSS AND I. FLÜGGE-LOTZ, *An optimal control problem for systems with differential-difference equation dynamics*, this Journal, 7 (1969), pp. 609–623.
- [20] H. F. VANDEVENNE, *Qualitative properties of a class of infinite dimensional systems*, Doctoral dissertation, Electrical Engineering Department, M.I.T., Cambridge, Mass., 1972.
- [21] ———, *Controllability and stabilizability properties of delay systems*, Proc. 1972 IEEE Decision and Control Conference, New Orleans, La., 1972.
- [22] J. C. WILLEMS, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automatic Control, AC-16 (1972), pp. 621–634.
- [23] W. M. WONHAM, *On a matrix Riccati equation of stochastic control*, this Journal, 6 (1968), pp. 681–697.

INFINITE-DIMENSIONAL FILTERING*

RUTH F. CURTAIN†

Abstract. This paper presents a generalization of the standard Kalman–Bucy linear filtering problem to infinite dimensions. The infinite-dimensional linear stochastic dynamical system is represented as a stochastic evolution equation

$$du(t, \omega) = \mathcal{A}(t)u(t, \omega) dt + \mathcal{B}(t) dw(t, \omega),$$

where $\mathcal{A}(t)$ is an unbounded operator, $\mathcal{B}(t)$ is a bounded operator, $w(t, \omega)$ is a Hilbert space-valued Wiener process and $u(t, \omega)$ is then a Hilbert space-valued stochastic process. The observation process is represented by

$$dz(t, \omega) = \mathcal{C}(t)u(t, \omega) dt + \mathcal{F}(t) dv(t, \omega),$$

where $\mathcal{C}(t)$ and $\mathcal{F}(t)$ are bounded operators and $v(t)$ is a finite-dimensional Wiener process. Using a combination of evolution equation techniques and abstract probability theory, the existence of an optimal filter for $u(t, \omega)$ based on the observation $z(t, \omega)$, $0 \leq s \leq t$, is established. As in the finite-dimensional theory, the filter may be obtained recursively by solving an infinite-dimensional Riccati equation. Similar results have been obtained by A. Bensoussan, *Filtrage optimal des systèmes Lineaires*, 1971, where $\mathcal{A}(t)$ are specifically partial differential operators satisfying slightly stronger conditions. However, he also presents a theory for the case where $v(t)$ may be infinite-dimensional.

Introduction. Here we give a generalization of the standard Kalman–Bucy linear filtering theory to infinite dimensions. The types of systems covered are very general, as the semigroup or abstract evolution equation approach allows us to consider a wide class of unbounded operators. As the theory draws heavily on results on abstract evolution equations, infinite-dimensional probability theory, stochastic differential equations in a Hilbert space and the infinite-dimensional Riccati equation, we give a summary of these in § 1 before proving the main results in § 2. As there has been much recent work published on infinite-dimensional filtering, there is a brief discussion of these in § 3.

1. Preliminaries. Let $(\Omega, \mathcal{S}, \mu)$ be our basic probability space which is assumed complete. Let $T = [0, T]$ be a real, finite interval and \mathcal{H}, \mathcal{K} real Hilbert spaces. Then an \mathcal{H} -valued random variable is a function $u(\cdot): \Omega \rightarrow \mathcal{H}$ which is measurable with respect to the μ -measure. If $u(\cdot)$ is also integrable, then we define the expectation $E\{u(\cdot)\} = \int_{\Omega} u(\omega) d\mu$. An \mathcal{H} -valued stochastic process is a function $u(\cdot, \cdot): T \times \Omega \rightarrow \mathcal{H}$ which is measurable in the pair (t, ω) , using Lebesgue measure on T . We also recall the definition of an \mathcal{H} -valued Wiener process from [6] or [7]:

(1.1) $w(t)$ is a *Wiener process* on \mathcal{H} if it is an \mathcal{H} -valued stochastic process with the following properties:

- (i) $E\{w(t) - w(s)\} = 0 \quad \forall s, t \in T$;
- (ii) $w(t)$ is continuous in t on T with probability one (w.p.1);
- (iii) $E\{(w(t) - w(s)) \circ (w(t) - w(s))\} = (t - s)\mathcal{W} \quad \forall s < t \in T$;

* Received by the editors June 25, 1973.

† Control Theory Centre, University of Warwick, Coventry, England. The Control Theory Centre is supported by the Leverhulme Trust and by the Science Research Council under Grant B/SR/9186.

where $\mathcal{W} \in \mathcal{L}(\mathcal{H})$ and is a positive, nuclear operator with eigenvalues $\{\lambda_i\}$ and orthonormal eigenvectors $\{e_i\}$ and is called the *covariance operator of $w(\cdot)$* .

$(u \circ v \in \mathcal{L}(\mathcal{H}))$ is defined $\forall u, v \in \mathcal{H}$ by $u \circ v(h) = u\langle v, h \rangle \forall h \in \mathcal{H}$.

(iv) $E\{\|w(t) - w(s)\|^2\} < \infty \forall s, t \in T$ and $\langle w(t_2) - w(t_1), e_i \rangle, \langle w(t_4) - w(t_3), e_i \rangle$ are independent real random variables for $t_1 < t_2 \leq t_3 < t_4$ and all eigenvectors e_i of \mathcal{W} .

Then it is shown that $w(t)$ has the unique representation

$$(1.2) \quad w(t) = \sum_{i=0}^{\infty} \beta_i(t, \omega) e_i \quad (t, \omega)\text{-almost everywhere,}$$

where $\beta_i(t, \omega)$ are mutually orthogonal real Wiener processes and $\{e_i\}$ is the orthonormal basis of \mathcal{H} generated by the eigenvectors of \mathcal{W} . We remark that this definition may be generalized by replacing assumption (iii) by

$$(iii)' \quad E\{(w(t) - w(s)) \circ (w(t) - w(s))\} = \int_s^t \mathcal{W}(\tau) d\tau,$$

where $\mathcal{W}(\cdot) \in L_{\infty}(T; \mathcal{L}(\mathcal{H}))$ and is nuclear, positive for almost all t and $\int_T \text{tr } \mathcal{W}(\tau) d\tau < \infty$. However, for stochastic dynamical system models, there is no loss of generality in using (iii) (see [5] for a discussion of this point).

(1.3) Generalize the \circ operation between 2 different Hilbert spaces as follows:

$$u \circ v(x) = u\langle v, x \rangle$$

for fixed $u \in \mathcal{H}, v \in \mathcal{K}$ and $\forall x \in \mathcal{K}$.

Then $u \circ v \in \mathcal{L}(\mathcal{K}, \mathcal{H})$. We also need some properties of the \mathcal{H} -valued Itô stochastic integral,

$$\int_T \mathcal{B}(t) dw(t, \omega),$$

where $\mathcal{B}(t) \in \mathcal{L}(\mathcal{H}, \mathcal{K})$ and $\int_T \|\mathcal{B}(t)\|^2 dt < \infty$. (For a complete study of $\int_T \mathcal{B}(t) dw(t)$, where $\mathcal{B}(t)$ is a stochastic process with values in $\mathcal{L}(\mathcal{H}, \mathcal{K})$, see [6]).

(1.4) *Properties of the Itô integral $\int_T \mathcal{B}(t) dw(t)$ for $\mathcal{B}(t)$ nonrandom* (see [5], [6]).

(i) $E\{\|\int_T \mathcal{B}(t) dw(t)\|^2\} \leq \text{tr } \mathcal{W} \int_T \|\mathcal{B}(t)\|^2 dt$;

(ii) $\int_T \mathcal{B}(t) dw(t) = \sum_{i=0}^{\infty} \int_T \mathcal{B}(t) e_i d\beta_i(t)$,

where $\{\beta_i(t)\}, \{e_i\}$ are as in (1.2);

(iii) if $w(t)$ and $v(t)$ are independent Wiener processes on \mathcal{H}, \mathcal{K} respectively, then

$$E\left\{\int_{t_1}^{t_2} \mathcal{B}_1(t) dw(t) \circ \int_{s_1}^{s_2} \mathcal{B}_2(t) dv(t)\right\} = 0$$

for nonrandom $\mathcal{B}_1(\cdot) \in L_2(T; \mathcal{L}(\mathcal{H}, \mathcal{H}_1))$ and $\mathcal{B}_2(\cdot) \in L_2(T; \mathcal{L}(\mathcal{K}, \mathcal{K}_1))$, respectively, and $(t_1, t_2), (s_1, s_2)$ are any intervals contained in T ;

(iv) $E\{\int_0^{s_1} \mathcal{B}_1(t) dw(t) \circ \int_0^{s_2} \mathcal{B}_2(t) dw(t)\} = \int_0^{\min(s_1, s_2)} \mathcal{B}_1(t) \mathcal{W} \mathcal{B}_2^*(t) dt$
for $\mathcal{B}_1(\cdot) \in L_2(T; \mathcal{L}(\mathcal{H}, \mathcal{H}_1))$ and $\mathcal{B}_2(\cdot) \in L_2(T; \mathcal{L}(\mathcal{H}, \mathcal{K}_2))$.

(1.5) *The Itô differential.* If $z_0 \in \mathcal{H}$, $a(\cdot) \in L_1(T; \mathcal{H})$, $\mathcal{B}(\cdot) \in L_2(T; \mathcal{L}(\mathcal{H}, \mathcal{H}))$ and $w(t)$ is an \mathcal{H} -valued Wiener process, then

$$z(t) = z_0 + \int_0^t a(s) ds + \int_0^t \mathcal{B}(s) dw(s)$$

is a well-defined \mathcal{H} -valued stochastic process with $E\{\|z(t)\|^2\} < \infty$, and we write it in differential notation:

$$dz(t) = a(t) dt + \mathcal{B}(t) dw(t),$$

$$z(0) = z_0.$$

(1.6) *Linear evolution equations.* Consider the linear evolution equation on \mathcal{H}

$$(1.7) \quad \begin{aligned} \dot{u}(t) &= \mathcal{A}(t)u(t) + f(t), \\ u(0) &= u_0, \end{aligned}$$

where $u_0 \in \mathcal{H}$, $f(t): T \rightarrow \mathcal{H}$, and $\mathcal{A}(t)$ is a closed linear operator on \mathcal{H} which generates an evolution operator $\mathcal{U}(t, s)$ on \mathcal{H} with the properties:

- (1.8) (i) $\mathcal{U}(t, s): T \times T \rightarrow \mathcal{L}(\mathcal{H}, \mathcal{H})$ and is strongly continuous in s and t for $0 \leq s < t \leq T$;
(ii) $\mathcal{U}(t, s) = \mathcal{U}(t, r)\mathcal{U}(r, s)$ if $0 \leq s \leq r \leq t \leq T$,
 $\mathcal{U}(t, t) = \mathcal{I}$;
(iii) $\mathcal{U}(t, s)$ is strongly continuously differentiable in t for $t > s$ and $\partial \mathcal{U}(t, s)/\partial t = \mathcal{A}(t)\mathcal{U}(t, s)$, where $\|\mathcal{A}(t)\mathcal{U}(t, s)\| \leq c_1/|t - s|$ for $0 \leq s < t \leq T$;
(iv) $\mathcal{U}(t, s) = \exp(-(t - s)\mathcal{A}(t)) + \mathcal{M}(t, s)$ for $0 \leq s < t \leq T$, where $\|\mathcal{A}(t)\mathcal{M}(t, s)\| \leq c_2/|t - s|^\theta$ for $0 \leq s < t \leq T$ and $\theta < 1$.

Then if $f(t)$ is Hölder continuous in t on T , (1.6) has the unique solution

$$u(t) = \mathcal{U}(t, 0)u_0 + \int_0^t \mathcal{U}(t, s)f(s) ds.$$

If $f(t)$ is merely Bochner integrable on T , then this is still well-defined and may be called a *weak* or *mild* solution, i.e., $u(t)$ is continuous on $(s, T]$ and

$$\int_s^T \langle u(t), g'(t) - \mathcal{A}^*(t)g(t) \rangle dt + \langle u(s), g(s) \rangle = 0,$$

where $g(t) \in \mathcal{D}(\mathcal{A}^*(t))$, and $g(t)$, $g'(t)$ and $\mathcal{A}^*(t)g(t)$ are continuous in $(s, T]$ and $g(T) = 0$ (see [10]). For technical conditions for $\mathcal{A}(t)$ to generate such an evolution operator see [10] and [11]. They include a large class of operators $\mathcal{A}(t)$ which generate analytic semigroups for each t and also $\mathcal{A}(t) + \mathcal{A}_1(t)$, where $\mathcal{A}_1(t)$ is any bounded operator which is uniformly bounded in norm in t on T .

(1.9) *The stochastic analogue* (see [2], [7]). Consider the linear stochastic evolution equation on \mathcal{H}

$$(1.10) \quad \begin{aligned} du(t, \omega) &= \mathcal{A}(t)u(t, \omega) dt + \mathcal{B}(t) dw(t, \omega), \\ u(0) &= u_0, \end{aligned}$$

where $\mathcal{A}(t)$ generates an evolution operator $\mathcal{U}(t, s)$ satisfying properties (1.8),

$u_0 \in \mathcal{H}$, $w(t, \omega)$ is a \mathcal{H} -valued Wiener process and $\mathcal{B}(t) \in \mathcal{L}(\mathcal{H}, \mathcal{H})$ and is uniformly bounded in norm in t on T . Then (1.10) has the unique weak or mild solution

$$u(t, \omega) = \mathcal{U}(t, 0)u_0 + \int_0^t \mathcal{U}(t, s)\mathcal{B}(s)dw(s, \omega)$$

which is an \mathcal{H} -valued stochastic process with $E\{\|u(t)\|^2\} < \infty$ uniformly on T . (This is actually a special case of a more general existence theorem for (1.10) in [2] where $\mathcal{B}(t)$ is an $\mathcal{L}(\mathcal{H}, \mathcal{H})$ -valued stochastic process and strong solutions are also considered).

(1.10) *The infinite-dimensional Riccati equation* (see [8]). Consider the following inner product versions of the infinite-dimensional Riccati equation:

$$(1.11) \left\langle \left[\frac{d\mathcal{P}(t)}{dt} + \mathcal{P}(t)\mathcal{A}(t) + \mathcal{A}^*(t)\mathcal{P}(t) + \mathcal{Q}(t) - \mathcal{P}(t)\mathcal{B}(t)\mathcal{R}^{-1}(t)\mathcal{B}^*(t)\mathcal{P}(t) \right] y, x \right\rangle = 0,$$

$$\mathcal{P}(T) = \mathcal{G} \quad \text{for arbitrary } x, y \in \bigcap_{t \in T} \mathcal{D}(\mathcal{A}(t)),$$

where $\mathcal{Q}(t)$, $\mathcal{B}(t)$, $\mathcal{R}(t)$, $\mathcal{R}(t)^{-1} \in L_\infty(T; \mathcal{L}(\mathcal{H}))$; \mathcal{G} , $\mathcal{Q}(t)$ are self-adjoint, positive semidefinite; $\mathcal{R}(t)$, $\mathcal{R}(t)^{-1}$ are self-adjoint and positive definite, and $\mathcal{A}(t)$ generates the evolution operator $\mathcal{U}(t, s)$ with properties (1.8). Then (1.11) has the unique solution $\mathcal{P}(t) \in \mathcal{L}(\mathcal{H})$ which is strongly continuous in t on T .

(1.12) *Remarks on the time invariant $\mathcal{A}(t) \equiv \mathcal{A}$ case.* When $\mathcal{A}(t) \equiv \mathcal{A}$, similar results to those in (1.6), (1.9) and (1.10) hold under the much weaker condition that \mathcal{A} should be the infinitesimal generator of a semigroup $\mathcal{T}(t)$. However, $\mathcal{T}(t - s)$ does not have all the properties of $\mathcal{U}(t, s)$ in (1.8) and so the proofs are technically a little tricky, and often one needs to impose other conditions on the other coefficients. For example, in (1.7) u_0 must be in $\mathcal{D}(\mathcal{A})$, and in (1.11) $\mathcal{Q}(t)$, $\mathcal{B}(t)$, $\mathcal{R}(t)$, $\mathcal{R}(t)^{-1}$ must be strongly continuously differentiable in t on T . We note that if \mathcal{A} generates a contraction semigroup or, more generally, an analytic semigroup, then $\mathcal{T}(t - s)$ does satisfy properties (1.8).

2. The filtering problem. Let \mathcal{H} , \mathcal{K} be real Hilbert spaces and $(\Omega, \mathcal{S}, \mu)$ an underlying probability space. Consider the following infinite-dimensional linear system:

$$(2.1) \quad \begin{aligned} du(t, \omega) &= \mathcal{A}(t)u(t, \omega) dt + \mathcal{B}(t) dw(t, \omega), \\ u(0) &= u_0, \end{aligned} \quad t \in T = [0, T],$$

$$(2.2) \quad \begin{aligned} dz(t, \omega) &= \mathcal{C}(t)u(t, \omega) dt + \mathcal{F}(t) dv(t, \omega), \\ z(0) &= 0, \end{aligned} \quad t \in T = [0, T],$$

where $\mathcal{A}(t)$ is a linear closed operator on \mathcal{H} which generates an evolution operator $\mathcal{U}(t, s)$ with properties (1.8), $\mathcal{B}(\cdot) \in L_\infty(T; \mathcal{L}(\mathcal{H}))$, $\mathcal{C}(\cdot) \in L_\infty(T; \mathcal{L}(\mathcal{H}, \mathcal{K}))$, $\mathcal{F}(\cdot)$, $\mathcal{F}(\cdot)^{-1} \in L_\infty(T; \mathcal{L}(\mathcal{K}))$, u_0 is an \mathcal{H} -valued random variable independent of

$w(t)$ and $v(t)$ and has zero expectation and covariance operator \mathcal{P}_0 . $w(t)$ and $v(t)$ are independent Wiener processes on \mathcal{H} and \mathcal{K} with covariance operators \mathcal{W} and \mathcal{V} , respectively. We assume that the observation space, \mathcal{K} , is finite-dimensional. Then from the results (1.9), (2.1) has the unique mild solution

$$u(t) = \mathcal{U}(t, 0)u_0 + \int_0^t \mathcal{U}(t, s)\mathcal{B}(s)dw(s)$$

and from (1.5), we see that $z(t)$ is a well-defined stochastic process with differential given by (2.2).

The filtering problem we propose is to find the best estimate of the state $u(t)$ based on observations $z(s)$, $0 \leq s \leq t$, which has the form

$$\hat{u}(t) = \int_0^t \mathcal{K}(t, s)dz(s),$$

where $\mathcal{K}(t, \cdot) \in L_2(T; \mathcal{L}(\mathcal{K}, \mathcal{H}))$ for almost all t and which minimizes $E\{\langle h, u(t) - \hat{u}(t) \rangle^2\} \forall h \in \mathcal{H}$.

THEOREM 2.1. $\hat{u}(t) = \int_0^t \mathcal{K}(t, s)dz(s)$ is a solution of the filtering problem if and only if $E\{\tilde{u}(t) \circ z(\sigma) - z(\tau)\} = 0$ for all σ, τ such that $0 \leq \tau \leq \sigma \leq T$, where $\tilde{u}(t) = u(t) - \hat{u}(t)$.

Proof. Let h be fixed and define the Hilbert space $X(h): X(h) = \{\langle u, h \rangle, \text{ where } u \text{ is an } \mathcal{H}\text{-valued random variable with } E\{\|u\|^2\} < \infty\}$. The inner product is $[\langle u, g \rangle, \langle v, h \rangle] = E\{\langle u, h \rangle \langle v, h \rangle\}$.

(We note that $E\{u \circ v\} = 0$ iff $[\langle u, h \rangle, \langle v, h \rangle] = 0 \forall h \in \mathcal{H}$). We also define, for fixed t , the subspace

$$X_t(h) = \left\{ \langle y_t, h \rangle, \text{ where } y(t) = \int_0^t \mathcal{B}(t, s)dz(s) \text{ and } \mathcal{B}(t, \cdot) \in L_2(T; \mathcal{L}(\mathcal{K}, \mathcal{H})) \text{ for almost all } t \right\}.$$

Now $\langle \hat{u}(t), h \rangle \in X_t(h)$, and we wish to minimize $\tilde{u}(t) = u(t) - \hat{u}(t)$ in the $X(h)$ -norm for all h . By the orthogonal projection lemma, this is equivalent to requiring $\langle \tilde{u}(t), h \rangle \perp X_t(h)$ in $X(h)$ for all $h \in \mathcal{H}$, i.e., $E\{\langle h, \tilde{u}(t) \rangle \langle h, y(t) \rangle\} = 0 \forall \langle h, y(t) \rangle \in X_t(h) \forall h \in \mathcal{H}$, i.e., iff $\langle h, E\{\tilde{u}(t) \circ y(t)\}h \rangle = 0$ by definition of \circ . So we need only establish that $E\{\tilde{u}(t) \circ y(t)\} = 0$ iff $E\{\tilde{u}(t) \circ z(\sigma) - z(\tau)\} = 0$ for $0 \leq \tau \leq \sigma \leq T$ (where we use the generalized definition (1.3) of \circ). Suppose first that $E\{\tilde{u}(t) \circ z(\sigma) - z(\tau)\} = 0$. Consider $y(t) = \int_0^t \mathcal{B}(t, s)dz(s)$, where $\mathcal{B}(t, s)$ is a step function in s . Then (with the obvious incremental notation)

$$\begin{aligned} E\{\tilde{u}(t) \circ y(t)\} &= \sum_j E\{\tilde{u}(t) \circ \mathcal{B}_j \Delta z_j\} \\ &= \sum_j E\{\tilde{u}(t) \circ \Delta z_j\} \mathcal{B}_j^* \quad (\text{since } \mathcal{B} \text{ is nonrandom}) \\ &= 0 \quad (\text{by assumption}). \end{aligned}$$

For general $\mathcal{B}(t, s)$, we can approximate by a sequence $\{\mathcal{B}_n(t, s)\}$ of step functions such that $\int_0^t \|\mathcal{B}(t, s) - \mathcal{B}_n(t, s)\|^2 ds \rightarrow 0$ as $n \rightarrow \infty$. Then

$$\begin{aligned}
& E \left\{ \left\| \int_0^t [\mathcal{B}(t, s) - \mathcal{B}_n(t, s)] dz(s) \right\|^2 \right\} \\
& \leq 2E \left\{ \left\| \int_0^t (\mathcal{B}(t, s) - \mathcal{B}_n(t, s)) \mathcal{C}(s) u(s) ds \right\|^2 \right\} + 2E \left\{ \left\| \int_0^t (\mathcal{B} - \mathcal{B}_n) \mathcal{F} dv \right\|^2 \right\} \\
& \hspace{25em} \text{(by (2.2))} \\
& \leq 2 \int_0^t \|\mathcal{B} - \mathcal{B}_n\|^2 ds \sup_{s \in T} \{\|\mathcal{C}(s)\|^2\} \cdot \int_0^t E\{\|u(s)\|^2\} ds \\
& \quad + 2 \operatorname{tr} \mathcal{V} \sup_{s \in T} \{\|\mathcal{F}(s)\|^2\} \cdot \int_0^t \|\mathcal{B} - \mathcal{B}_n\|^2 ds \\
& \hspace{25em} \text{(by Schwarz' inequality and property (1.4(i)))} \\
& \leq \text{const.} \int_0^t \|\mathcal{B} - \mathcal{B}_n\|^2 ds \quad \text{(since } \mathcal{C}(\cdot), \mathcal{F}(\cdot) \text{ are uniformly bounded} \\
& \hspace{25em} \text{in norm and (1.9))} \\
& \rightarrow 0 \quad \text{as } n \rightarrow \infty.
\end{aligned}$$

But

$$\begin{aligned}
\|E\{\tilde{u}(t) \circ y(t)\}\| & \leq E\{\|\tilde{u}(t) \circ y(t)\|\} \\
& = E\{\|\tilde{u}(t)\| \cdot \|y(t)\|\} \quad \text{(by definition of } \circ) \\
& \leq (E\{\|\tilde{u}(t)\|^2\} E\{\|y(t)\|^2\})^{1/2} \quad \text{(by Schwarz' inequality).}
\end{aligned}$$

So by approximating $y(t)$ by $y_n(t) = \int_0^t \mathcal{B}_n(t, s) dz(s)$, we see that $E\{\tilde{u}(t) \circ y(t)\} = 0$ for all $y(\cdot)$ such that $\langle y(t), h \rangle \in X_t(h)$. Clearly the argument is independent of h .

Suppose, conversely, that $E\{\tilde{u}(t) \circ z(\sigma) - z(\tau)\} \neq 0$ for some σ, τ . Then define

$$\mathcal{B}(t, s) = \begin{cases} E\{\tilde{u}(t) \circ z(\sigma) - z(\tau)\} & \text{for } \tau \leq s \leq \sigma, \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\begin{aligned}
\int_0^t \|\mathcal{B}(t, s)\|^2 ds & \leq \int_\tau^\sigma E\{\|\tilde{u}(t)\|^2\} ds \cdot \int_\tau^\sigma E\{\|z(\sigma) - z(\tau)\|^2\} ds \\
& \hspace{25em} \text{(by the usual inequality arguments)} \\
& < \infty.
\end{aligned}$$

So $y(t) = \int_0^t \mathcal{B}(t, s) ds$ is such that $\langle y(t), h \rangle \in X_t(h)$. Now

$$\begin{aligned}
\langle h, E\{\tilde{u}(t) \circ y(t)\} h \rangle & = \langle h, E\{\tilde{u}(t) \circ z(\sigma) - z(\tau)\} E\{\tilde{u}(t) \circ z(\sigma) - z(\tau)\}^* h \rangle \\
& = \|E\{\tilde{u}(t) \circ z(\sigma) - z(\tau)\}^* h\|^2 \\
& \neq 0 \quad \text{for some } h.
\end{aligned}$$

So $E\{\tilde{u}(t) \circ y(t)\} \neq 0$.

We now need the following technical results.

LEMMA 2.1. Let $\Lambda(t, s) = E\{u(t) \circ u(s)\}$, where $u(t)$ is the mild solution of (2.1). Then (a)

$$(2.3) \quad \Lambda(t, s) = \mathcal{U}(t, 0)\mathcal{P}_0\mathcal{U}^*(s, 0) + \int_0^{\min(t,s)} \mathcal{U}(t, \tau)\mathcal{B}(\tau)\mathcal{W}\mathcal{B}^*(\tau)\mathcal{U}^*(s, \tau) d\tau;$$

(b) $\Lambda(t, s)$ is strongly continuously differentiable in t for $t > s$ with $\partial\Lambda(t, s)/\partial t = \mathcal{A}(t)\Lambda(t, s)$;

(c) under the extra assumptions

(i) $\mathcal{B}(s)e_i \in \mathcal{D}(\mathcal{A}(s)) \forall s \in T$,

(ii) $\int_T \|\mathcal{A}(t)\mathcal{B}(s)e_i\| ds < \infty$

uniformly in t on T for all eigenvectors e_i of \mathcal{W} , then $\Lambda(t, t)$ is differentiable in t in the following inner product sense:

$$\left\langle \left[\frac{\partial\Lambda(t, t)}{\partial t} - \mathcal{A}(t)\Lambda(t, t) + \Lambda(t, t)\mathcal{A}^*(t) + \mathcal{B}(t)\mathcal{W}\mathcal{B}^*(t) \right] y, x \right\rangle = 0 \quad \forall x, y \in \mathcal{H}.$$

Proof. (a) This follows by substituting $u(t) = \mathcal{U}(t, 0)u_0 + \int_0^t \mathcal{U}(t, s)\mathcal{B}(s)dw(s)$ and applying the result (1.4(iv)).

(b) This follows using property (1.8(iii)) of the evolution operator $\mathcal{U}(t, s)$. To justify differentiation under the integral sign, we note that $\|\mathcal{A}(t)\mathcal{U}(t, s)\| \leq c_1/|t - s|$, and all other operators are uniformly bounded in norm on T . Since $\mathcal{A}(t)$ is closed, we can put it outside the integral.

(c) Again one can formally verify the result by using the differentiation property (1.8(iii)) of $\mathcal{U}(t, s)$. The problem is to justify differentiation under the integral sign, which is delicate. Essentially, we need to show that

$$\langle \mathcal{A}(t)\mathcal{U}(t, s)\mathcal{B}(s)\mathcal{W}\mathcal{B}^*(s)\mathcal{U}^*(t, s)y, x \rangle$$

is bounded by an integrable function of s uniformly in t .

Using the decomposition (1.8(iv)), we see that

$$\mathcal{A}(t)\mathcal{U}(t, s)\mathcal{B}(s)e_i = \mathcal{A}(t)\mathcal{U}(t, s)\mathcal{B}(s)e_i + \exp(-(t-s)\mathcal{A}(t))\mathcal{A}(t)\mathcal{B}(s)e_i.$$

Therefore,

$$\|\mathcal{A}(t)\mathcal{U}(t, s)\mathcal{B}(s)e_i\| \leq \frac{\text{const.}}{|t-s|^\theta} + \text{const.} \|\mathcal{A}(t)\mathcal{B}(s)e_i\|, \quad \text{where } \theta < 1.$$

So under assumptions (i) and (ii), $\mathcal{A}(t)\mathcal{U}(t, s)\mathcal{B}(s)e_i$ is bounded in norm by an integrable function of s uniformly in t . Also $\mathcal{W} : \mathcal{H} \rightarrow \text{span of its eigenvectors}$, and all other operators are uniformly bounded in norm. So we can differentiate under the integral sign. The inner product is necessary when we consider the analogous term

$$\langle \mathcal{U}(t, s)\mathcal{B}(s)\mathcal{W}\mathcal{B}^*(s)\mathcal{U}^*(t, s)\mathcal{A}^*(t)y, x \rangle.$$

LEMMA 2.2. Let $\mathcal{K}(t, \cdot) \in L_2(T; \mathcal{L}(\mathcal{K}, \mathcal{H}))$ and $z(\cdot)$ be as in (2.1) and $u(\cdot)$ be the mild solution of (2.1); then

$$E \left\{ \int_0^t \mathcal{K}(t, s) dz(s) \circ u(s) \right\} = \int_0^t \mathcal{K}(t, s)\mathcal{C}(s)\Lambda(s, s) ds.$$

Proof.

$$\begin{aligned}
& E \left\{ \int_0^t \mathcal{K}(t, s) dz(s) \circ u(\sigma) \right\} \\
&= E \left\{ \left[\int_0^t \mathcal{K}(t, s) \mathcal{C}(s) \mathcal{U}(s, 0) u_0 ds + \int_0^t \mathcal{K}(t, s) \mathcal{F}(s) dv(s) \right. \right. \\
&\quad \left. \left. + \int_0^t \mathcal{K}(t, s) \mathcal{C}(s) \int_0^s \mathcal{U}(s, \tau) \mathcal{B}(\tau) dw(\tau) ds \right] \circ \mathcal{U}(\sigma, 0) u_0 + \int_0^\sigma \mathcal{U}(\sigma, s) \mathcal{B}(s) dw(s) \right\} \\
&\quad \text{(by definition of } u(\cdot) \text{ and } z(\cdot)) \\
&= \int_0^t \mathcal{K}(t, s) \mathcal{C}(s) \mathcal{U}(s, 0) ds \mathcal{P}_0 \mathcal{U}^*(\sigma, 0) \\
&\quad + \int_0^t \mathcal{K}(t, s) \mathcal{C}(s) E \left\{ \int_0^s \mathcal{U}(s, \tau) \mathcal{B}(\tau) dw(\tau) \circ \int_0^\sigma \mathcal{U}(\sigma, s) \mathcal{B}(s) dw(s) \right\} ds \\
&\quad \text{(since } u_0, v(\cdot) \text{ and } w(\cdot) \text{ are mutually independent)} \\
&= \int_0^t \mathcal{K}(t, s) \mathcal{C}(s) \mathcal{U}(s, 0) \mathcal{P}_0 \mathcal{U}^*(\sigma, 0) ds \\
&\quad + \int_0^t \mathcal{K}(t, s) \mathcal{C}(s) \int_0^{\min(s, \sigma)} \mathcal{U}(s, \tau) \mathcal{B}(\tau) \mathcal{W} \mathcal{B}^*(\tau) \mathcal{U}^*(\sigma, \tau) d\tau ds \\
&\quad \text{(by 1.4(iv))} \\
&= \int_0^t \mathcal{K}(t, s) \mathcal{C}(s) \Lambda(s, \sigma) ds \quad \text{(by definition of } \Lambda(\cdot, \cdot) \text{.)} \quad \text{Q.E.D.}
\end{aligned}$$

Consider the operator-valued integral equation

$$(2.4) \quad \int_0^t \mathcal{K}(t, s) \mathcal{C}(s) \Lambda(s, \sigma) \mathcal{C}^*(\sigma) ds + \mathcal{K}(t, \sigma) \mathcal{F}(\sigma) \mathcal{V} \mathcal{F}^*(\sigma) = \Lambda(t, \sigma) \mathcal{C}^*(\sigma).$$

This is also a key element in A. Bensoussan's development of the filtering theory, and the following existence and uniqueness theorem is essentially his.

THEOREM 2.2. *If $\Lambda(\cdot, \cdot)$ is given by (2.3) and $\mathcal{C}(\cdot)$, $\mathcal{F}(\cdot)$, \mathcal{V} satisfy the assumptions in our problem statement, then (2.4) has a unique solution $\mathcal{K}(t, \cdot) \in L_\infty([0, t]; \mathcal{L}(\mathcal{H}, \mathcal{H}))$ for each fixed $t \in T$.*

Proof. (see [1] for details). Let t be fixed and define the operators \mathcal{Q}_1 , \mathcal{Q}_2 by

$$\begin{aligned}
\mathcal{Q}_1 &= \mathcal{C}(\tau) \Lambda(\tau, t), \\
\mathcal{Q}_2 f(\sigma) &= \mathcal{F}(\sigma) \mathcal{V} \mathcal{F}^*(\sigma) f(\sigma) + \int_0^t \mathcal{C}(\sigma) \Lambda(\sigma, s) \mathcal{C}^*(s) f(s) ds.
\end{aligned}$$

Then

$$\begin{aligned}
\mathcal{Q}_1 &: \mathcal{H} \rightarrow L_\infty([0, t]; \mathcal{H}), \\
\mathcal{Q}_2 &: L_2([0, T]; \mathcal{H}) \rightarrow L_2([0, t]; \mathcal{H}),
\end{aligned}$$

and

$$\begin{aligned} \langle\langle \mathcal{Q}_2 f, f \rangle\rangle &= \int_0^t \langle \mathcal{F}(\sigma) \mathcal{V} \mathcal{F}^*(\sigma) f(\sigma), f(\sigma) \rangle d\sigma \\ &\quad + \int_0^t \left\langle \int_0^s \mathcal{C}(\sigma) \Lambda(\sigma, s) \mathcal{C}^*(s) f(\sigma) ds, f(\sigma) \right\rangle d\sigma \\ &> 0 \quad \forall f(\cdot) \in L_2([0, t], \cdot, \mathcal{H}) \end{aligned}$$

from the form of $\Lambda(\cdot, \cdot)$ in (2.3) and since $\mathcal{F}(\sigma) \mathcal{V} \mathcal{F}(\sigma)^*$ is invertible. Hence \mathcal{Q}_2^{-1} exists. Define $\mathcal{K}(t, \sigma) = (\mathcal{Q}_2^{-1} \mathcal{Q}_1)^*$, and let $k(\sigma) = \mathcal{K}(t, \sigma)^* h$ for some $h \in \mathcal{H}$. Then $\mathcal{Q}_2 k(\sigma) = \mathcal{Q}_1 h$ or

$$(2.5) \quad \mathcal{F}(\sigma) \mathcal{V} \mathcal{F}(\sigma)^* k(\sigma) + \int_0^t \mathcal{C}(\sigma) \Lambda(\sigma, s) \mathcal{C}^*(s) k(s) ds = \mathcal{C}(\sigma) \Lambda(\sigma, t) h$$

and $k(\cdot) \in L_2([0, t], \mathcal{H})$. But $k(\sigma) = \mathcal{K}(t, \sigma)^* h$, and we see that $\mathcal{K}(t, \cdot) \in L_\infty([0, t], \mathcal{L}(\mathcal{H}, \mathcal{H}))$. $\mathcal{K}(t, \cdot)$ satisfies the integral equation (2.4); we see this by substituting $k(\sigma) = \mathcal{K}(t, \sigma)^* h$ in (2.5) and taking adjoints. So $\mathcal{K}(t, \sigma) = (\mathcal{Q}_2^{-1} \mathcal{Q}_1)^*$ is a solution of (2.4). The uniqueness is similarly proved using the linearity of (2.4).

Remarks. Here we need for the first time the conditions $\mathcal{F}(\cdot)^{-1} \in L_\infty(T; \mathcal{L}(\mathcal{H}))$ and \mathcal{V}^{-1} exists. As \mathcal{V} is the covariance operator of the Wiener process $v(\cdot)$, it is nuclear, and its invertibility means that \mathcal{H} must be finite-dimensional.

THEOREM 2.3. *There is a solution $\hat{u}(t) = \int_0^t \mathcal{K}(t, s) dz(s)$ to the filtering problem iff the integral equation (2.4) has a solution.*

Proof. Suppose first that there is a solution $\tilde{u}(t) = \int_0^t \mathcal{K}(t, s) dz(s)$ to the filtering problem. Let t be fixed and define, for $0 \leq \sigma < t$,

$$y(\sigma) = \int_0^\sigma \mathcal{C}(s) u(s) ds = z(\sigma) - z(0) - \int_0^\sigma \mathcal{F}(s) dv(s).$$

Now

$$\begin{aligned} \frac{d}{d\sigma} E\{\tilde{u}(t) \circ y(\sigma)\} &= E\{\tilde{u}(t) \circ \mathcal{C}(\sigma) u(\sigma)\} \quad (\text{since } \sigma < t) \\ (2.6) \quad &= E\{u(t) \circ u(\sigma)\} \mathcal{C}^*(\sigma) - E\left\{\int_0^t \mathcal{K}(t, s) dz(s) \circ u(\sigma)\right\} \mathcal{C}^*(\sigma) \\ &= \Lambda(t, \sigma) \mathcal{C}^*(\sigma) - \int_0^t \mathcal{K}(t, s) \mathcal{C}(s) \Lambda(s, \sigma) \mathcal{C}^*(\sigma) ds \end{aligned}$$

(by Lemmas 2.1 and 2.2).

But

$$\begin{aligned}
 E\{\tilde{u}(t) \circ y(\sigma)\} &= E\left\{\tilde{u}(t) \circ - \int_0^\sigma \mathcal{F}(s) dv(s)\right\} \quad (\text{applying Theorem (2.1)}) \\
 &= E\left\{\int_0^t \mathcal{K}(t, s) \mathcal{F}(s) dv(s) \circ \int_0^\sigma \mathcal{F}(s) dv(s)\right\} \\
 &\quad (\text{expanding } u(\cdot) \text{ and } \hat{u}(\cdot), \text{ since } v(\cdot) \text{ is independent of } u_0 \text{ and } \\
 &\quad w(\cdot) \text{ and using property (1.4(iii))}). \\
 &= \int_0^\sigma \mathcal{K}(t, s) \mathcal{F}(s) \mathcal{V} \mathcal{F}(s)^* ds \quad (\text{by (1.4(iv))}).
 \end{aligned}$$

Therefore

$$(2.7) \quad \frac{d}{d\sigma} E\{\tilde{u}(t) \circ y(\sigma)\} = \mathcal{K}(t, \sigma) \mathcal{F}(\sigma) \mathcal{V} \mathcal{F}(\sigma)^*.$$

Equating (2.6) and (2.7) gives us the integral equation (2.4). Suppose now that $\mathcal{K}(t, s)$ is the unique solution of (2.4). It remains to prove that

$$\hat{u}(t) = \int_0^t \mathcal{K}(t, s) dz(s)$$

satisfies Theorem 2.1, i.e., $E\{\tilde{u}(t) \circ z(\sigma) - z(\tau)\} = 0$ for all $0 \leq \tau \leq \sigma \in T$. From the linearity, we may let $\tau = 0$. Now

$$\begin{aligned}
 E\{\tilde{u}(t) \circ z(\sigma) - z(0)\} &= E\{\tilde{u}(t) \circ y(\sigma)\} + E\left\{\tilde{u}(t) \circ \int_0^\sigma \mathcal{F}(s) dv(s)\right\} \\
 &\quad (\text{from the definition of } y(\sigma)) \\
 &= \int_0^\sigma \left[\Lambda(t, \alpha) \mathcal{C}^*(\alpha) - \int_0^t \mathcal{K}(t, s) \mathcal{C}(s) \Lambda(s, \alpha) \mathcal{C}^*(\alpha) ds \right] d\alpha \\
 &\quad - \int_0^\sigma \mathcal{K}(t, s) \mathcal{F}(s) \mathcal{V} \mathcal{F}(s)^* ds \quad (\text{from (2.6) and (1.4(iv))}) \\
 &= \int_0^\sigma \mathcal{K}(t, \alpha) \mathcal{F}(\alpha) \mathcal{V} \mathcal{F}(\alpha)^* d\alpha - \int_0^\sigma \mathcal{K}(t, s) \mathcal{F}(s) \mathcal{V} \mathcal{F}(s)^* ds \\
 &= 0 \quad (\text{since } \mathcal{K}(t, s) \text{ is a solution of (2.4)}).
 \end{aligned}$$

COROLLARY. *Under the assumptions of our problem, there exists a unique solution to the filtering problem $\hat{u}(t) = \int_0^t \mathcal{K}(t, s) dz(s)$, where $\mathcal{K}(t, s)$ is the unique solution of (2.4).*

The following theorem gives necessary conditions for an optimal filter under rather technical assumptions on $\mathcal{K}(t, s)$. It should be regarded as a preliminary to the later theorem on sufficient conditions for an optimal filter.

THEOREM 2.4. *If the solution $\mathcal{K}(t, \cdot)$ of (2.4) satisfies the additional regularity properties*

(i) $\mathcal{K}(t, s)$ is strongly differentiable in t on T for $t > s$ and

$$\left\| \frac{\partial \mathcal{K}}{\partial t}(t, s) \mathcal{C}(s) \Lambda(s, \sigma) \mathcal{C}^*(\sigma) x \right\| \leq \beta(s, \sigma),$$

where $\beta(\cdot, \sigma) \in L_2(T)$ and is independent of t ,

(ii) $\mathcal{A}(t) \mathcal{K}(t, s) \mathcal{C}(s) \Lambda(s, \sigma) \mathcal{C}^*(\sigma) x$ is integrable in t on T for all $x \in \mathcal{H}$; then

$$(2.8) \quad \left[\mathcal{A}(t) \mathcal{K}(t, s) - \mathcal{K}(t, t) \mathcal{C}(t) \mathcal{K}(t, s) - \frac{\partial \mathcal{K}}{\partial t}(t, s) \right] x = 0$$

for all $x \in \mathcal{D}(\mathcal{A}(t) \mathcal{K}(t, s))$.

Proof. We differentiate (2.4) (in the strong sense) for $t > \sigma$.

$$\begin{aligned} & \frac{d}{dt} [\Lambda(t, \sigma) \mathcal{C}^*(\sigma) - \mathcal{K}(t, \sigma) \mathcal{F}(\sigma) \mathcal{V} \mathcal{F}(\sigma)^*] x \\ &= \mathcal{A}(t) \Lambda(t, \sigma) \mathcal{C}^*(\sigma) x - \frac{\partial \mathcal{K}}{\partial t}(t, \sigma) \mathcal{F}(\sigma) \mathcal{V} \mathcal{F}(\sigma)^* x \\ & \quad \text{(by Lemma (2.1(a)) and assumption (i) on } \mathcal{K}) \\ &= \mathcal{A}(t) \left[\int_0^t \mathcal{K}(t, s) \mathcal{C}(s) \Lambda(s, \sigma) \mathcal{C}^*(\sigma) ds + \mathcal{K}(t, \sigma) \mathcal{F}(\sigma) \mathcal{V} \mathcal{F}^*(\sigma) \right] x \\ & \quad - \frac{\partial \mathcal{K}}{\partial t}(t, \sigma) \mathcal{F}(\sigma) \mathcal{V} \mathcal{F}^*(\sigma) x \quad \text{(by (2.4))} \\ &= \mathcal{A}(t) \mathcal{K}(t, \sigma) \mathcal{F}(\sigma) \mathcal{V} \mathcal{F}(\sigma)^* x - \frac{\partial \mathcal{K}}{\partial t}(t, \sigma) \mathcal{F}(\sigma) \mathcal{V} \mathcal{F}^*(\sigma) x \\ & \quad + \int_0^t \mathcal{A}(t) \mathcal{K}(t, s) \mathcal{C}(s) \Lambda(s, \sigma) \mathcal{C}^*(\sigma) ds x \\ & \quad \text{(since } \mathcal{A}(t) \text{ is a closed operator and by assumption (ii)).} \end{aligned}$$

But

$$\begin{aligned} & \frac{d}{dt} \left[\int_0^t \mathcal{K}(t, s) \mathcal{C}(s) \Lambda(s, \sigma) \mathcal{C}^*(\sigma) ds x \right] \\ &= \int_0^t \frac{\partial \mathcal{K}}{\partial t}(t, s) \mathcal{C}(s) \Lambda(s, \sigma) \mathcal{C}^*(\sigma) ds x + \mathcal{K}(t, t) \mathcal{C}(t) \Lambda(t, \sigma) \mathcal{C}^*(\sigma) x \\ & \quad \text{(by assumption (i))} \\ &= \int_0^t \frac{\partial \mathcal{K}}{\partial t}(t, s) \mathcal{C}(s) \Lambda(s, \sigma) \mathcal{C}^*(\sigma) ds x \\ & \quad + \mathcal{K}(t, t) \mathcal{C}(t) \left(\int_0^t \mathcal{K}(t, s) \mathcal{C}(s) \Lambda(s, \sigma) \mathcal{C}^*(\sigma) ds x \right. \\ & \quad \left. + \mathcal{K}(t, \sigma) \mathcal{F}(\sigma) \mathcal{V} \mathcal{F}^*(\sigma) x \right) \quad \text{(by (2.4)).} \end{aligned}$$

Therefore

$$\begin{aligned} \int_0^t \left[\mathcal{A}(t)\mathcal{K}(t, s) - \frac{\partial \mathcal{K}}{\partial t}(t, s) - \mathcal{K}(t, t)\mathcal{C}(t)\mathcal{K}(t, s) \right] \mathcal{C}(s)\Lambda(s, \sigma)\mathcal{C}^*(\sigma) ds x \\ + \left[\mathcal{A}(t)\mathcal{K}(t, \sigma) - \frac{\partial \mathcal{K}}{\partial t}(t, \sigma) - \mathcal{K}(t, t)\mathcal{C}(t)\mathcal{K}(t, \sigma) \right] \mathcal{F}(\sigma)\mathcal{V}\mathcal{F}^*(\sigma)x = 0 \\ \forall x \in \mathcal{H}. \end{aligned}$$

Letting

$$\Delta(t, \sigma) = \mathcal{A}(t)\mathcal{K}(t, \sigma) - \frac{\partial \mathcal{K}}{\partial t}(t, \sigma) - \mathcal{K}(t, t)\mathcal{C}(t)\mathcal{K}(t, \sigma),$$

we have that

$$\int_0^t \Delta(t, s)\mathcal{C}(s)\Lambda(s, \sigma)\mathcal{C}^*(\sigma) ds x + \Delta(t, \sigma)\mathcal{F}(\sigma)\mathcal{V}\mathcal{F}^*(\sigma)x = 0 \quad \forall x \in \mathcal{H}.$$

Therefore

$$\int_0^t \langle \Delta(t, s)\mathcal{C}(s)\Lambda(s, \sigma)\mathcal{C}^*(\sigma)x, y \rangle ds + \langle \Delta(t, \sigma)\mathcal{F}(\sigma)\mathcal{V}\mathcal{F}^*(\sigma)x, y \rangle = 0 \quad \forall x, y \in \mathcal{H}.$$

Thus

$$\int_0^t \langle \mathcal{C}(s)\Lambda(s, \sigma)\mathcal{C}^*(\sigma)x, \Delta^*(t, s)y \rangle ds + \langle \mathcal{F}(\sigma)\mathcal{V}\mathcal{F}^*(\sigma)x, \Delta^*(t, \sigma)y \rangle = 0$$

if $y \in \mathcal{D}(\Delta^*(t, s))$, i.e., $\langle \mathcal{Q}_2 x, \Delta^*(t, \sigma)y \rangle = 0$, where \mathcal{Q}_2 is as defined in the proof of Theorem 2.2. But \mathcal{Q}_2 is strictly positive, and hence $\Delta^*(t, \sigma) = 0$ on its domain.

LEMMA 2.3. *Under the assumption of our problem, the following inner product version of the infinite-dimensional Riccati equation has a unique solution $\mathcal{P}(t) \in \mathcal{L}(\mathcal{H})$ which is strongly continuous in t on T :*

$$\begin{aligned} (2.9) \quad & \left\langle \left[\frac{d\mathcal{P}(t)}{dt} - \mathcal{A}(t)\mathcal{P}(t) - \mathcal{P}(t)\mathcal{A}^*(t) - \mathcal{B}(t)\mathcal{W}\mathcal{B}^*(t) \right. \right. \\ & \left. \left. + \mathcal{P}(t)\mathcal{C}^*(t)(\mathcal{F}(t)\mathcal{V}\mathcal{F}^*(t))^{-1}\mathcal{C}(t)\mathcal{P}(t) \right] x, y \right\rangle = 0, \\ & \mathcal{P}(0) = \mathcal{P}_0 \quad \text{for arbitrary } x, y \in \bigcap_{t \in T} \mathcal{D}(\mathcal{A}^*(t)). \end{aligned}$$

Proof. Let $t' = T - t$ and $\mathcal{A}_1^*(t') = \mathcal{A}^*(T - t')$; $\mathcal{P}_1(t') = \mathcal{P}(T - t')$; $\mathcal{B}_1(t') = \mathcal{B}(T - t)$, etc. Then (2.9) may be rewritten

$$\begin{aligned} & \left\langle \left[-\frac{d\mathcal{P}_1}{dt'} - \mathcal{A}_1^*(t')\mathcal{P}_1(t') - \mathcal{P}_1(t')\mathcal{A}_1(t') - \mathcal{B}_1(t')\mathcal{W}\mathcal{B}_1^*(t') \right. \right. \\ & \left. \left. + \mathcal{P}_1(t')\mathcal{C}_1^*(t')(\mathcal{F}_1(t')\mathcal{V}\mathcal{F}_1^*(t'))^{-1}\mathcal{C}_1(t')\mathcal{P}_1(t') \right] x, y \right\rangle = 0, \\ & \mathcal{P}_1(T) = \mathcal{P}_0, \end{aligned}$$

which is now the form of (1.11). Then all bounded operators are L_∞ on T and satisfy the adjointness and positivity conditions of result (1.10). Moreover, $\mathcal{A}_1(t') = \mathcal{A}^*(T - t')$ is a closed, linear operator which generates an evolution operator with properties (1.8) (see [10] for details), so we appeal to the result of (1.10).

THEOREM 2.5. *If $\mathcal{P}(t)$ is the unique solution of (2.9), let*

$$\mathcal{K}(t) = \mathcal{P}(t)\mathcal{C}^*(t)(\mathcal{F}(t)\mathcal{V}\mathcal{F}^*(t))^{-1}$$

and $\mathcal{Y}(t, s)$ the evolution operator generated by $\mathcal{A}(t) - \mathcal{K}(t)\mathcal{C}(t)$. Then

$$\hat{u}(t) = \int_0^t \mathcal{Y}(t, s)\mathcal{K}(s) dz(s)$$

is the solution to the filtering problem and is the weak solution of the stochastic evolution equation

$$(2.10) \quad \begin{aligned} d\hat{u}(t) &= (\mathcal{A}(t) - \mathcal{K}(t)\mathcal{C}(t))\hat{u}(t) dt + \mathcal{K}(t)\mathcal{C}(t)u(t) dt + \mathcal{K}(t)\mathcal{F}(t) dv(t), \\ \hat{u}(0) &= 0. \end{aligned}$$

Proof. (a) $\mathcal{K}(t, s) = \mathcal{Y}(t, s)\mathcal{K}(s) = \mathcal{Y}(t, s)\mathcal{P}(s)\mathcal{C}^*(s)(\mathcal{F}(s)\mathcal{V}\mathcal{F}^*(s))^{-1}$ satisfies assumptions (i) and (ii) of Theorem 2.4, since $\mathcal{Y}(t, s)$ satisfies properties (1.8). In particular, by property (1.8(iii)), $\mathcal{K}(t, s)$ satisfies the necessary condition (2.8) of Theorem 2.4, i.e., $\mathcal{K}(t, s)$ satisfies the differentiated version of the integral equation (2.4), and so we have

$$\begin{aligned} \int_0^t \mathcal{Y}(t, s)\mathcal{K}(s)\mathcal{C}(s)\Lambda(s, \sigma)\mathcal{C}^*(\sigma) ds - \Lambda(t, \sigma)\mathcal{C}^*(\sigma) \\ + \mathcal{Y}(t, \sigma)\mathcal{K}(\sigma)(\mathcal{F}(\sigma)\mathcal{V}\mathcal{F}^*(\sigma))^{-1} = \mathcal{G}(\sigma), \end{aligned}$$

where \mathcal{G} is some operator-valued function of σ , independent of t . We show that $\mathcal{G}(\sigma) = 0$ and so $\mathcal{Y}(t, s)\mathcal{K}(s) = \mathcal{K}(t, s)$ satisfies (2.4). It is sufficient to let $t = \sigma$ and to show that

$$\mathcal{K}(\sigma)\mathcal{F}(\sigma)\mathcal{V}\mathcal{F}^*(\sigma) = \Lambda(\sigma, \sigma) - \int_0^\sigma \mathcal{Y}(\sigma, s)\mathcal{K}(s)\mathcal{C}(s)\Lambda(s, \sigma) ds$$

i.e.,

$$(2.11) \quad \mathcal{P}(\sigma) = \Lambda(\sigma, \sigma) - \int_0^\sigma \mathcal{Y}(\sigma, s)\mathcal{P}(s)\mathcal{C}^*(s)(\mathcal{F}(s)\mathcal{V}\mathcal{F}^*(s))^{-1}\mathcal{C}(s)\Lambda(s, \sigma) ds.$$

But this is just an integrated version of the Riccati equation (2.9), as we now verify. Let $\mathcal{N}(\sigma)$ denote the left side. In order to differentiate $\Lambda(\sigma, \sigma)$ we must differentiate $\langle \mathcal{N}(\sigma)x, y \rangle$ and assume extra assumptions (i) and (ii) of Lemma 2.1(c).

Then

$$\begin{aligned}
\frac{d}{d\sigma} \langle \mathcal{N}(\sigma)x, y \rangle &= \langle [\mathcal{A}(\sigma)\Lambda(\sigma, \sigma) + \Lambda(\sigma, \sigma)\mathcal{A}^*(\sigma) + \mathcal{B}(\sigma)\mathcal{W}\mathcal{B}^*(\sigma)]x, y \rangle \\
&\quad - \left\langle \left[\mathcal{P}(\sigma)\mathcal{C}^*(\sigma)(\mathcal{F}(\sigma)\mathcal{V}\mathcal{F}^*(\sigma))^{-1}\Lambda(\sigma, \sigma) \right. \right. \\
&\quad + \int_0^\sigma (\mathcal{A}(\sigma) - \mathcal{K}(\sigma)\mathcal{C}(s))\mathcal{Y}(\sigma, s)\mathcal{P}(s)\mathcal{C}^*(s) \\
&\quad \cdot (\mathcal{F}(s)\mathcal{V}\mathcal{F}^*(s))^{-1}\mathcal{C}(s)\Lambda(s, \sigma) ds \\
&\quad \left. + \int_0^\sigma \mathcal{Y}(\sigma, s)\mathcal{P}(s)\mathcal{C}^*(s)(\mathcal{F}(s)\mathcal{V}\mathcal{F}^*(s))^{-1}\mathcal{C}(s)\Lambda(s, \sigma)\mathcal{A}^*(\sigma) ds \right] x, y \rangle \\
&= \langle [\mathcal{A}(\sigma)\mathcal{N}(\sigma) + \mathcal{N}(\sigma)\mathcal{A}^*(\sigma) - \mathcal{K}(\sigma)\mathcal{C}(\sigma)\mathcal{N}(\sigma) \\
&\quad + \mathcal{B}(\sigma)\mathcal{W}\mathcal{B}^*(\sigma)]x, y \rangle,
\end{aligned}$$

substituting for $\mathcal{N}(\sigma)$ and also $\mathcal{N}(0) = \mathcal{P}_0$. So $\mathcal{N}(\sigma)$ satisfies (2.9) and by the uniqueness of the solution, $\mathcal{P}(\sigma) \equiv \mathcal{N}(\sigma)$. Since (2.11) is well-defined without the assumptions needed to differentiate $\Lambda(\sigma, \sigma)$, we have verified (2.11) as required.

(b) (2.10) satisfies all the conditions of the theorem as stated in (1.9), and so $\hat{u}(t)$ is a weak solution of (2.10).

3. General remarks and conclusions. Using an abstract evolution approach, we have proved the existence and uniqueness of an optimal filter for a very general class of stochastic linear infinite-dimensional dynamical systems. Furthermore, we have shown that it may be obtained recursively from an infinite-dimensional Riccati equation as in the finite-dimensional case. Our results may be generalized in the following ways:

1. An easy generalization of our filtering problem is to consider systems like $du(t) = \mathcal{A}(t)u(t) dt + f(t) dt + \mathcal{B}(t) dw(t)$ on $[T_1, T_2] = T$, where $f(\cdot) \in L_2(T; \mathcal{H})$. All arguments go through for the filter

$$\hat{u}(t) = \int_{T_1}^t \mathcal{K}(t, s) dz(s) + E\{u(t)\} - E\left\{ \int_{T_1}^t \mathcal{K}(t, s) dz(s) \right\},$$

where $\mathcal{K}(t, s)$ is the same as in the homogeneous ($f \equiv 0$) case.

2. *Perturbation result.* Our conclusions include systems of the form $du(t) = \mathcal{A}(t)u(t) + \mathcal{A}_1(t)u(t) dt + \mathcal{B}(t) dw(t)$, where $\mathcal{A}(t)$ is as before, but $\mathcal{A}_1(t)$ satisfies

- (i) $\mathcal{A}_1(t)$ is a closed linear operator on \mathcal{H} with $\mathcal{D}(\mathcal{A}_1(t)) \supset \mathcal{D}(\mathcal{A}(t))$;
- (ii) $\mathcal{A}_1(t)\|(\gamma\mathcal{I} + \mathcal{A}(t))^{-1}\| \leq c/(|\gamma|^{1-\theta})$ for $\gamma \notin \text{spectrum of } \mathcal{A}(t)$;
- (iii) $\mathcal{A}_1(t)\mathcal{A}(t)^{-1}$ is Hölder continuous;

since then $\mathcal{A}(t) + \mathcal{A}_1(t)$ generates an evolution operator $\mathcal{U}(t, s)$ with properties (1.8) (see [11]). For example, this includes $\mathcal{A}_1(\cdot) \in L_\infty(T; \mathcal{L}(\mathcal{H}))$.

3. We have assumed that $\mathcal{A}(t)$ is time-dependent and generates an evolution operator $\mathcal{U}(t, s)$ with properties (1.8), which include the time-invariant version of the filtering problem :

$$du(t) = \mathcal{A}u(t) dt + \mathcal{B} dw(t),$$

$$u(0) = u_0;$$

$$dz(t) = \mathcal{C}u(t) dt + \mathcal{F} dv(t),$$

$$z(0) = 0;$$

where \mathcal{A} is the infinitesimal generator of a strongly continuous semigroup and \mathcal{B} , \mathcal{C} , \mathcal{F} and \mathcal{F}^{-1} are bounded operators. (2.9) has a unique solution $\mathcal{P}(t)$ under these assumptions (see [8]) and generates a semigroup $\mathcal{Y}(t, s)$ (see [13]). So $\hat{u}(t) = \int_0^t \mathcal{Y}(t, s) \mathcal{P}(s) \mathcal{C}(s) (\mathcal{F}(s) \mathcal{V} \mathcal{F}^*(s))^{-1} dz(s)$ is well-defined, but $\mathcal{Y}(t, s)$ has weaker properties than before, and our proofs would need modification.

Finally, as there has been much recent work published on the infinite-dimensional filtering problem, we discuss their relationship with this paper. Much of the work has been of a formal nature or otherwise applied to a very specific type of system, for example, [9] and [12]. (For a full discussion of these, see [3].) The major contribution in this field is by A. Bensoussan in his recent book [1], and where our assumptions overlap, the results are of course identical, although his approach is quite different. The results of this paper are applicable to a very wide class of operators $\mathcal{A}(t)$ in an infinite-dimensional setting and could, for example, be applied to differential delay systems and parabolic equations of first and higher orders in t (with suitable extensions, see [4]). Whereas A. Bensoussan's results are specifically for a wide, but less general class, parabolic partial differential equations of first and second orders in t . However, A. Bensoussan's general theory gives a wider concept of estimation, which gives a filtering theory even for the case when \mathcal{X} is infinite-dimensional. He also proves that the filter is the best global filter in an appropriate sense, so the two theories are complementary.

REFERENCES

- [1] A. BENSOUSSAN, *Filtrage Optimal des Systèmes Lineaires*, Dunod, Paris, 1971 (1967).
- [2] R. F. CURTAIN, *Stochastic differential equations in a Hilbert space*, Ph.D. dissertation, Brown University, Providence, R.I., 1969.
- [3] ———, *Infinite Dimensional Filtering: A Survey Introduction*, Report no. 7301, Lund Inst. Techn., Sweden, 1973.
- [4] ———, *Stochastic parabolic equations of higher order in t* , J. Math. Anal. Appl., to appear.
- [5] ———, *On the Itô stochastic integral in a Hilbert space*, Report no. 11, Control Theory Centre, University of Warwick, Coventry, England, 1972.
- [6] R. F. CURTAIN AND P. L. FALB, *Itô's lemma in infinite dimensions*, J. Math. Anal. Appl., 31 (1970), pp. 434–448.
- [7] ———, *Stochastic differential equations in Hilbert space*, J. Differential Equations, 10 (1971), pp. 412–430.
- [8] R. F. CURTAIN AND A. J. PRITCHARD, *The infinite dimensional Riccati equation*, Report no. 11, Control Theory Centre, University of Warwick, Coventry, England, 1972.
- [9] P. L. FALB, *Infinite Dimensional Filtering*, Information and Control, 11 (1967), pp. 102–137.
- [10] T. KATO, *Abstract evolution equation of parabolic type in Banach and Hilbert spaces*, Nagoya Math. J., 19 (1961), pp. 93–125.

- [11] T. KATO AND H. TANABE, *On the abstract evolution equation*, Osaka J. Math., 14 (1962), pp.107–133.
- [12] H. J. KUSHNER, *Filtering for Linear Distributed Parameter Systems*, this Journal, 8 (1970), pp. 346–359.
- [13] R. S. PHILLIPS, *Perturbation theory for semigroups of linear operators*, Trans. Amer. Math. Soc., 74 (1954), pp. 199–221.

A NOTE ON GENERALIZED PURSUIT-EVASION GAMES*

ROBERT J. ELLIOTT AND AVNER FRIEDMAN†

Abstract. Two player zero sum differential games are an extension of optimal control problems. When the cost or payoff is the integral of some function h up to the first time the trajectory enters a "terminal set" the differential game is one of survival. If $h \equiv 1$, the payoff is just the time elapsed up to the "capture time" and the game is one of pursuit and evasion. If $h \geq 0$, the game is called a generalized pursuit-evasion game, and in previous papers it has been shown that when the Isaacs condition is satisfied the upper and lower "extended" values of such a differential game are equal—that is, the game "has extended value". In the present note this result is proved under the weaker condition that $\max_y \min_z h(t, x, y, z) \geq 0$.

1. Introduction. The present note is a sequel to the papers [2] and [4] on generalized pursuit-evasion games; the notation and terminology of [4] will be followed. A generalized pursuit-evasion game consists of a dynamical system

$$(1.1) \quad \frac{dx}{dt} = f(t, x, y, z), \quad x \in R^m, \quad t_0 \leq t \leq T_0,$$

an initial condition

$$(1.2) \quad x(t_0) = x_0,$$

a terminal set $F \supset [T_0, \infty) \times R^m$ and a payoff

$$(1.3) \quad P(y, z) = \int_{t_0}^{\tilde{t}} h(t, x, y, z) dt,$$

where \tilde{t} is the first time the trajectory $(t, x(t))$ enters F and

$$(1.4) \quad h(t, x, y, z) \geq 0.$$

It is proved in [2] and [4] that such a game has an extended value; in this note we prove the existence of extended value when (1.4) is replaced by the weaker condition

$$(1.5) \quad \max_{y \in Y} \min_{z \in Z} h(t, x, y, z) \geq 0.$$

This result was recently proved by Kalton [5] by an entirely different method; the proof given here is much simpler. It is based on an extension of Lemma 2.2 of [4] to the case where (1.5) holds. Once this is done, the results of [4] can be applied immediately to establish the existence of an extended value.

* Received by the editors April 16, 1973.

† Department of Mathematics, Northwestern University, Evanston, Illinois 60201.

2. The main theorem. As in [4], we shall assume:

(A₁) $f(t, x, y, z)$ is continuous on $[t_0, T_0] \times R^m \times Y \times Z$,

$$x \cdot f(t, x, y, z) \leq k(t)(1 + |x|^2),$$

$$\int_{t_0}^{T_0} k(t) dt < \infty.$$

(A₂) For any $R > 0$, if $|x| < R$ and $|\bar{x}| < R$, then

$$|f(t, x, y, z) - f(t, \bar{x}, y, z)| \leq k_R(t)|x - \bar{x}|,$$

$$\int_{t_0}^{T_0} k_R(t) dt < \infty.$$

(A₃) $h(t, x, y, z)$ is continuous on $[t_0, T_0] \times R^m \times Y \times Z$.

(A₄) The Isaacs condition holds:

$$\min_{z \in Z} \max_{y \in Y} \{f(t, x, y, z) \cdot p + h(t, x, y, z)\} = \max_{y \in Y} \min_{z \in Z} \{f(t, x, y, z) \cdot p + h(t, x, y, z)\}$$

for any $p \in R^m$, $x \in R^m$ and $t \in [t_0, T_0]$.

(A₅) For any $t \in [t_0, T_0]$, $x \in R^m$ the set

$$(f(t, y, Y, Z), h(t, x, Y, Z)) = \{(f(t, x, y, z), h(t, x, y, z)); y \in Y, z \in Z\}$$

is convex.

THEOREM. *If (A₁)–(A₅) and (1.5) hold, then the differential game has extended value.*

We initially suppose that for all $t \in [t_0, T_0]$ and $x \in R^m$,

$$(2.1) \quad \max_{y \in Y} \min_{z \in Z} h(t, x, y, z) \geq h_0 > 0;$$

the case where $\max_{y \in Y} \min_{z \in Z} h(t, x, y, z) \geq 0$ can then be treated as in [2], [4] and [5] by considering the approximating cases obtained by replacing h by

$$h + 1/n \quad \text{for } n = 1, 2, 3, \dots$$

For any pair of control functions $y(s), z(s)$, the dynamical equations determine a trajectory $x(s) \in R^m$. The new variable x_{m+1} is introduced and

$$x_{m+1}(t) = \int_{t_0}^t h(s, x(s), y(s), z(s)) ds.$$

$x'(t) = (x(t), x_{m+1}(t))$ denotes this augmented trajectory. For any such trajectory x' and any $E > 0$ write $t_E(x')$ for the first time that $x_{m+1}(t) \geq E$. The following payoff is then introduced:

$$Q_E(x') = \inf \{\rho(t, x(t)); t_0 \leq t \leq t_E(x')\},$$

where $\rho(t, x)$ is the distance from (t, x) to F .

Let n be any positive integer and write $\delta = (T_0 - t_0)/2^n$; $t_j = t_0 + j\delta$ ($1 \leq j \leq 2^n$), $\Lambda = T_0 - t_0$, $\eta = \delta h_0$. For $E > 0$, write

$$\chi_E(r) = \begin{cases} 1 & \text{if } r \leq E, \\ 1 + \Lambda(r - E)/\eta^2 & \text{if } E < r \leq E + \eta, \\ 1 + \Lambda/\eta & \text{if } E + \eta < r. \end{cases}$$

For any trajectory $x'(t) \in R^{m+1}$ define a payoff

$$\begin{aligned} M_{E,n}(x') &= M_{E,n}(y, z) = M_{E,n}(x'(t_1), \dots, x'(t_{2^n})) \\ &= \min_{0 \leq j \leq 2^n} \{(\rho(t_j, x(t_j)) + \eta)\chi_E(x_{m+1}(t_j))\}. \end{aligned}$$

Note that $M_{E,n+1}(x') \leq M_{E,n}(x')$ so

$$V^+(M_{E,n+1}) \leq V^+(M_{E,n}) \quad \text{and} \quad V^-(M_{E,n+1}) \leq V^-(M_{E,n}).$$

Consequently, $\lim_{n \rightarrow \infty} V^+(M_{E,n})$ and $\lim_{n \rightarrow \infty} V^-(M_{E,n})$ exist.

Because of the weight factor χ_E , $M_{E,n}$ selects approximately the smallest distance on the whole trajectory to F , where the x_{m+1} coordinate is less than E . However, Q_E only looks for the smallest distance until $x_{m+1}(t) \geq E$, and in the present game h may be negative and $x_{m+1}(t)$ may decrease. Therefore, for any trajectory x' ,

$$M_{E,n}(x') \leq Q_E(x') + (h_0 + B)/n,$$

where B is the uniform bound on $|f(t, x, y, z)|$.

With the usual notation for upper and lower values, it is therefore the case that

$$(2.2) \quad \lim_{n \rightarrow \infty} V^+(M_{E,n}) \leq V^+(Q_E)$$

and

$$(2.3) \quad \lim_{n \rightarrow \infty} V^-(M_{E,n}) \leq V^-(Q_E).$$

LEMMA.

$$\lim_{n \rightarrow \infty} V^+(M_{E,n}) = V^+(Q_E).$$

Proof. We suppose that

$$(2.4) \quad L = \lim_{n \rightarrow \infty} V^+(M_{E,n}) < V^+(Q_E)$$

and derive a contradiction; in view of (2.2) the proof of the lemma is then complete. From (2.4), it follows that there is $c > 0$ such that

$$L \leq V^+(Q_E) - c.$$

Choose N_0 such that for $n \geq N_0$,

$$V^+(M_{E,n}) \leq L + c/4.$$

Using the notation of [3], for $c/4$ there is a δ_0 such that if $\delta < \delta_0$,

$$V^\delta(M_{E,N_0}) \leq V^+(M_{E,N_0}) + c/4.$$

There is, therefore, a lower δ strategy (with $\delta \leq \delta_0$) Δ_δ^* for the player z such that for any upper strategy Γ^δ for y :

$$M_{E,N_0}(\Delta_\delta^*, \Gamma^\delta) \leq V^\delta(M_{E,N_0}) + c/4.$$

With this lower strategy Δ_δ^* and any upper strategy Γ^δ , consider the payoff $Q_E(\Delta_\delta^*, \Gamma^\delta)$. If $x'(t)$ is the trajectory generated by $\Delta_\delta^*, \Gamma^\delta$, write, as above, $t_E(x')$ for the first time that $x_{m+1}(t) \geq E$. Because of (2.1), strategy Γ^δ can be modified from $t_E(x')$ onwards to a strategy Γ_E^δ which ensures that $x_{m+1}(t) \geq E$ for $t \geq t_E(x')$.

Then

$$Q_E(\Delta_\delta^*, \Gamma^\delta) = Q_E(\Delta_\delta^*, \Gamma_E^\delta)$$

but also

$$\begin{aligned} Q_E(\Delta_\delta^*, \Gamma_E^\delta) &\leq M_{E,N_0}(\Delta_\delta^*, \Gamma_E^\delta) \\ &\leq V^\delta(M_{E,N_0}) + c/4 \\ &\leq V^+(M_{E,N_0}) + c/2 \\ &\leq L + 3c/4 \leq V^+(Q_E) - c/4. \end{aligned}$$

Therefore,

$$\sup_{\Gamma^\delta} Q_E(\Delta_\delta^*, \Gamma^\delta) \leq V^+(Q_E) - c/4,$$

so

$$\begin{aligned} V^+(Q_E) &= \inf_{\Delta_\delta} \sup_{\Gamma^\delta} Q_E(\Delta_\delta, \Gamma^\delta) \\ &\leq V^+(Q_E) - c/4 \end{aligned}$$

which is a contradiction. This completes the proof of the lemma.

Proof of theorem. By [1],

$$V^+(M_{E,n}) = V^-(M_{E,n}).$$

Combining this with (2.3) and the lemma we get

$$V^+(Q_E) \leq V^-(Q_E).$$

Hence

$$V^+(Q_E) = V^-(Q_E).$$

This value $V(Q_E)$ is a nonincreasing function of E and $V(Q_E) = 0$ for sufficiently large E . We may suppose $\rho(t_0, x_0) > 0$ so by (2.1), $V(Q_E) > 0$ for sufficiently small E .

Thus

$$E^* = \inf \{E : V(Q_E) = 0\}$$

is positive and we prove as in [4] that

$$V_e^+ = E^*, \quad V_e^- = E^*.$$

The proof that $V_e^+ \leq E^*$ is exactly as in [4]; we now prove $V_e^- \geq E^*$.

For any $\varepsilon > 0$ there is an $E > E^* - \varepsilon$ for which $V(Q_E) > 0$. By [3, § 2.5], there is a generalized saddle point $(\Delta^{**}, \Gamma^{**})$ for this game, so that

$$(2.5) \quad Q_E^0(\Delta, \Gamma_\delta^{**}) \geq V(Q_E) > 0$$

for all strategies Δ . Suppose $\Gamma^{**} = \{\Gamma_\delta^{**}\}$. $\delta = 1/n$. For small enough δ the components Γ_δ^{**} of Γ^{**} can be modified into $\tilde{\Gamma}_\delta^{**}$ so that for $t > t_E$ the new y control is such that for all z ,

$$h(t, x(t), y(t), z(t)) > 0.$$

Then if $x'(t)$ is an outcome of $(\Delta, \tilde{\Gamma}^{**})$, where $\tilde{\Gamma}_\delta^{**} = \{\tilde{\Gamma}_\delta^{**}\}$, $x_{m+1}(t)$ is monotone increasing in t for $t > t_E(x')$. The inequality (2.5) implies

$$Q_E^0(\Delta, \tilde{\Gamma}^{**}) \geq V(Q_E) > 0.$$

This implies

$$P_0[\Delta, \tilde{\Gamma}^{**}] \geq E > E^* - \varepsilon.$$

As ε is arbitrary,

$$V_e^- \geq E^*.$$

Therefore

$$V_e^+ = V_e^- = E^*.$$

Remark. With the definition of extended value as in [2] an extended value also exists.

REFERENCES

- [1] R. J. ELLIOT AND N. J. KALTON, *The existence of value in differential games*, Mem. Amer. Math. Soc., No. 126, Providence, R.I., 1972.
- [2] ———, *The existence of value in differential games of pursuit and evasion*, J. Differential Equations, 12 (1972), pp. 504–523.
- [3] A. FRIEDMAN, *Differential Games*, Pure and Applied Mathematics, vol. 25, Wiley-Interscience, New York, 1971.
- [4] ———, *Existence of extended value for differential games of generalized pursuit-evasion*, J. Differential Equations, 13 (1973), pp. 172–181.
- [5] N. J. KALTON, *Differential games with optimal stopping*, to appear.

THE LINEAR QUADRATIC COST PROBLEM WITH LINEAR STATE CONSTRAINTS AND THE NONSYMMETRIC RICCATI EQUATION*

D. D. THOMPSON† AND R. A. VOLZ‡

Abstract. Application of Neustadt's necessary conditions for the optimal control of a linear system with quadratic cost under a closed, hyperplane-bounded half-space state constraint is considered here. Significantly it is proved that the jump-discontinuities generally required in these conditions are not possible except at the initial and final times. This permits simplification of these conditions to a form similar to that obtained for the unconstrained problem of the same type. A nonsymmetric matrix Riccati equation is obtained to solve the resulting split-boundary conditions reducing them to a set-valued fixed-point condition on the boundary times. An iterative computational scheme is proposed to solve this fixed-point problem and is applied to two examples.

Additionally the existence of the solution to a nonsymmetric Riccati equation of the type obtained here is treated in detail and bounds on this solution are obtained for a large number of general cases.

1. Introduction. Almost since the inception of Pontryagin's necessary conditions for the solution of optimal control problems involving control and terminal constraints, similar conditions have been available for the state inequality constrained problem as well [1], [2], [3]. However, the conditions for the state constrained problem are not as well known as those pertaining only to control constrained problems, due primarily to the fact that their application is generally exceedingly difficult. Instead, approximate techniques have become popular for the solution of these problems.

Most popular among the approximate techniques have been the penalty function procedures. While there are a number of different schemes employing this method (see Kelley [4], Lasdon [5] and Russell [6]) all have the basic feature of replacing the state constraints by an additional term in the cost function which becomes very large either for constraint violation or as the constraint is approached. In addition, other approximate methods such as the cutting plane algorithms (see Kapur and Van Slyke [7]) have been suggested.

For certain special cases (when the state dimension is the same as the "order of the constraint") Speyer [8] has shown that the state constrained optimization problem can be separated into two unconstrained problems. This method, however, is only applicable for a single boundary interval and requires explicit solutions to certain implicit relations.

In this paper we return to consideration of the use of necessary conditions. Based upon a more convenient form of the necessary conditions developed by Neustadt [9] and sufficient conditions of Gilbert and Funk [10], the solution to the nonsingular, linear problem with quadratic cost and linear, recoverable (first order) state constraint is expressed, as in the solution to the unconstrained problem, in terms of the solution to a Riccati equation. In this case, however, the Riccati equation is nonsymmetric. With the aid of this Riccati equation, a novel

* Received by the editors January 18, 1972, and in revised form April 23, 1973.

† Research and Development Department, Atlantic Richfield Co., Dallas, Texas 75221.

‡ Department of Electrical and Computer Engineering, University of Michigan, Ann Arbor, Michigan 48104.

fixed-point problem is introduced and shown to be equivalent to the constrained quadratic problem.

The paper is essentially in two parts. In the first the linear quadratic cost problem with linear recoverable state constraint is considered, the Riccati form for the problem developed, and the fixed-point theorem proved. This latter form suggests several possibilities for new computational algorithms. An example of a rather simple iterative procedure is proposed and successfully applied to two examples, one of which illustrates multiple boundary intervals.

The lack of symmetry in the Riccati system poses new problems, however. None of the usual theorems on the existence of solution to the system apply. Indeed, examples can be found for finite escape. The second portion of the paper deals with this problem and presents several examples and theorems on the existence of solutions to the nonsymmetric system.

2. Problem definition. The problem to be considered is a linear quadratic cost problem with linear constraints (QCP) and may be stated as follows:

PROBLEM (QCP). Find a measurable, essentially bounded control function $u(t) \in R^m$ and an absolutely continuous state function $x(t) \in R^n$ defined on the fixed interval $\bar{I} = [t_0, t_1]$ which minimize the cost functional

$$(1) \quad J(x, u) = \frac{1}{2} x'(t_1) Q x(t_1) + \frac{1}{2} \int_{t_0}^{t_1} [u'(t) W(t) u(t) + x'(t) R(t) x(t)] dt$$

subject to the constraints

$$(2) \quad \dot{x}(t) = A(t)x(t) + B(t)u(t) + w(t) \quad \text{a.e. on } \bar{I},$$

$$(3) \quad x(t_0) = x^0,$$

$$(4) \quad x(t_1) = \text{"free"},$$

$$(5) \quad s'(t)x(t) \leq \sigma(t) \quad \text{for all } t \in \bar{I},$$

where $A(\cdot)$ and $B(\cdot)$ are C^1 -matrix functions on \bar{I} of dimensions $n \times n$ and $n \times m$ respectively; $w(\cdot): \bar{I} \rightarrow R^n$ and is C^1 ; $W(\cdot): \bar{I} \rightarrow R^{m \times m}$ and $R(\cdot): \bar{I} \rightarrow R^{n \times n}$ are C^1 , symmetric matrix functions on \bar{I} , $W(\cdot)$ positive definite and $R(\cdot)$ positive semidefinite; Q is a constant $n \times n$ positive semidefinite matrix; $s(\cdot): \bar{I} \rightarrow R^n$ and $\sigma(\cdot): \bar{I} \rightarrow R^1$ and are C^2 .

Equation (5) is the constraint of most interest and will be referred to as the *state constraint*. It constrains the state to a closed, smoothly moving half-space of R^n at each $t \in \bar{I}$. This half-space is bounded by a moving hyperplane having an outer normal $s(t)$ and a displacement from the origin of a distance $\sigma(t)/(s'(t)s(t))^{1/2}$ (assuming that $s(t) \neq 0$). With the omission of this state constraint the QCP is seen to be the classical problem considered by Kalman [11] and others.

DEFINITION 1. Any state-control pair (x, u) satisfying all the conditions of QCP except, possibly, for minimizing the functional $J(\cdot, \cdot)$ is an *admissible arc*. If $J(\cdot, \cdot)$ is also minimized, (x, u) is an *optimal arc*.

DEFINITION 2. The time $t \in \bar{I}$ is a *boundary time* if $s'(t)x(t) = 0$.

DEFINITION 3. The time $t \in \bar{I}$ is an *interior time* if $s'(t)x(t) < 0$.

DEFINITION 4. The boundary time $t \in (t_0, t_1]$ is an *entry time* if each interval of the form $(\tau, t) \subset I$ contains an interior time.

DEFINITION 5. The boundary time $t \in [t_0, t_1)$ is an *exit time* if every interval of the form $(t, \tau) \subset I$ contains an interior time.

DEFINITION 6. The interval $[t_a, t_b] \subset \bar{I}$ is a *boundary interval* if it consists entirely of boundary times and if t_a is either an entry time or t_0 , and t_b is either an exit time or t_1 .

DEFINITION 7. The QCP is recoverable if $B'(t)s(t) \neq 0$ for all $t \in \bar{I}$.

The recoverability condition is equivalent for the QCP to Denham and Bryson's "first order constraint" condition [12]. Essentially it insures that the state can always be controlled in the $s(t)$ direction.

3. Necessary and sufficient conditions. Necessary and sufficient conditions for this problem can be found by application of a minor extension of the theorems of Neustadt [9] and Gilbert and Funk [10]. The results are summarized in the following. The proof is straightforward and omitted. It can be found in detail in [13].

THEOREM 1. If (x, u) is an optimal arc of the QCP, then there exist measurable functions $\lambda(\cdot): \bar{I} \rightarrow R^1$ and $\eta(\cdot): \bar{I} \rightarrow R^n$ such that (i) $\lambda(t)$ is nonincreasing for all $t \in \bar{I}$ and constant if $s'(t)x(t) < \sigma(t)$; (ii) $\lambda(t)$ is right continuous for all $t \in I$; (iii) $\eta(\cdot)$ is absolutely continuous on \bar{I} ; and the conditions

$$(6) \quad \lambda(t_1) = 0,$$

$$(7) \quad \dot{\eta}(t) = -A'(t)[\eta(t) - \lambda(t)s(t)] + \lambda(t)\dot{s}(t) + \mu R(t)x(t) \quad \text{a.e. on } \bar{I},$$

$$(8) \quad \eta(t_1) = -\mu Qx(t_1) + \lambda(t_1)s(t_1)$$

and either

$$(9a) \quad u(t) = W^{-1}(t)B'(t)[\eta(t) - \lambda(t)s(t)] \quad \text{a.e. on } \bar{I}$$

and

$$\mu = 1$$

or

$$B'(t)[\eta(t) - \lambda(t)s(t)] = 0 \quad \text{a.e. on } \bar{I},$$

$$(9b) \quad \lambda(t_0) \neq 0,$$

$$\eta(t) - \lambda(t)s(t) \neq 0 \quad \text{a.e. on } \bar{I},$$

and

$$\mu = 0$$

are satisfied. Furthermore, any admissible arc (x, u) of the QCP which satisfies conditions (1)–(3) and equations (6)–(9a) for some $\eta(\cdot)$ and $\lambda(\cdot)$ is optimal.

The major difficulty arising from the application of these conditions is that, in general, $\lambda(t)$ may have unknown jump discontinuities at the unknown entry and exit times. However, as an important consequence of the form of the QCP and the continuity assumptions, it will be shown here that $\lambda(\cdot)$ must be continuous on (t_0, t_1) .

DEFINITION 8. An optimal arc (x, u) which does satisfy Theorem 1 with (9a) being valid is said to be a *nonsingular arc*.

For convenience we introduce two additional definitions.

DEFINITION 9. $\bar{B}(t) = B(t)W^{-1}(t)B'(t)$.

DEFINITION 10.

$$\lambda_b(t) = \frac{\dot{s}'(t)x(t) - \dot{\sigma}(t) + s'(t)[A(t)x(t) + \bar{B}(t)\eta(t) + w(t)]}{s'(t)\bar{B}(t)s(t)}.$$

From the condition on $s(t)$, $\sigma(t)$, $A(t)$, $B(t)$, $W(t)$, $w(t)$ and $x(t)$, it is evident that $\lambda_b(t)$ is continuous. In the following three lemmas, λ is presumed to satisfy the conditions of Theorem 1.

LEMMA 1. For a nonsingular optimal arc of the recoverable QCP and for almost all $t \in \bar{I}$,

$$\frac{d}{dt}(s'(t)x(t) - \sigma(t)) = s'(t)\bar{B}(t)s(t)(\lambda_b(t) - \lambda(t)).$$

Proof. Direct computation verifies the result. Q.E.D.

LEMMA 2. If $\hat{t} \in I = (t_0, t_1)$ is a boundary time for a nonsingular, optimal arc of the recoverable QCP, then $\lambda(\hat{t}) = \lambda_b(\hat{t})$.

Proof. Let \hat{t} be a boundary time in I . Suppose that $\lambda(\hat{t}) < \lambda_b(\hat{t})$. Then, since $\lambda(\cdot)$ is nonincreasing there exists an interval $(\hat{t}, \tilde{t}) \subset I$ such that $\lambda(t) < \lambda_b(t)$ for all $t \in (\hat{t}, \tilde{t})$. Thus, by Lemma 1,

$$\frac{d}{dt}[s'(t)x(t) - \sigma(t)] > 0 \quad \text{a.e. on } (\hat{t}, \tilde{t}).$$

But \hat{t} is a boundary time so

$$s'(t)x(t) - \sigma(t) > 0 \quad \text{for all } t \in (\hat{t}, \tilde{t}),$$

which is a violation of the constraint. Hence, $\lambda(\hat{t}) \geq \lambda_b(\hat{t})$.

Next suppose that $\lambda(\hat{t}) > \lambda_b(\hat{t})$. Then there exists an interval $(\bar{t}, \hat{t}) \in I$ on which $\lambda(t) > \lambda_b(t)$. Again by Lemma 1,

$$\frac{d}{dt}[s'(t)x(t) - \sigma(t)] < 0 \quad \text{a.e. on } (\bar{t}, \hat{t}).$$

Since \hat{t} is a boundary time this implies constraint violation on (\bar{t}, \hat{t}) which is a contradiction. Hence, $\lambda(\hat{t}) = \lambda_b(\hat{t})$ as required.

Before proving the continuity of $\lambda(\cdot)$, observe that $\lambda_b(\cdot)$ is C^1 and that by direct (but tedious) computation,

$$\dot{\lambda}_b(t) = y'(t)x(t) + z'(t)\eta(t) + \beta(t),$$

where

$$(10) \quad y'(t) = \frac{1}{s'(t)\bar{B}(t)s(t)} \{ \ddot{s}'(t) + 2\dot{s}'(t)A(t) + s'(t)[\dot{A}(t) + A^2(t) + \bar{B}(t)R(t)] \} \\ - \left(\frac{1}{s'(t)\bar{B}(t)s(t)} \right)^2 s'(t)[2\bar{B}(t)\dot{s}(t) + \dot{\bar{B}}(t)s(t)][\dot{s}'(t) + s'(t)A(t)],$$

$$(11) \quad \begin{aligned} z'(t) = & \frac{1}{s'(t)\bar{B}(t)s(t)} \{ 2\dot{s}'(t)\bar{B}(t) + s'(t)[A(t)\bar{B}(t) - \bar{B}(t)A'(t) + \dot{\bar{B}}(t)] \} \\ & - \left(\frac{1}{s'(t)\bar{B}(t)s(t)} \right)^2 s'(t)[2\bar{B}(t)\dot{s}(t) + \dot{\bar{B}}(t)s(t)]s'(t)\bar{B}(t), \end{aligned}$$

$$(12) \quad \begin{aligned} \beta(t) = & \frac{1}{s'(t)\bar{B}(t)s(t)} \{ 2\dot{s}'(t)w(t) - \dot{\sigma}(t) + s'(t)[A(t)w(t) + \dot{w}(t)] \} \\ & - \left(\frac{1}{s'(t)\bar{B}(t)s(t)} \right)^2 s'(t)[2\bar{B}(t)\dot{s}(t) + \dot{\bar{B}}(t)s(t)][s'(t)w(t) - \dot{\sigma}(t)]. \quad \text{Q.E.D} \end{aligned}$$

LEMMA 3. For a nonsingular, optimal arc of the recoverable QCP, define for all $t \in \bar{I}$,

$$(13) \quad \lambda^*(t) = \begin{cases} \lim_{\tau \rightarrow t_0^+} \lambda(\tau) & \text{if } t = t_0, \\ \lambda(t) & \text{if } t \in (t_0, t_1), \\ \lim_{\tau \rightarrow t_1^-} \lambda(\tau) & \text{if } t = t_1, \end{cases}$$

where $\lambda(\cdot)$ satisfies the conditions of Theorem 1. Then $\lambda^*(\cdot)$ is absolutely continuous on \bar{I} and satisfies the differential equation

$$(14) \quad \lambda^*(t) = \begin{cases} \dot{\lambda}_b(t) & \text{if } t \text{ is a boundary time,} \\ 0 & \text{otherwise,} \end{cases}$$

for almost all $t \in \bar{I}$.

Proof.

Part A. The ordinary continuity of $\lambda(\cdot)$ on I will be proved.

Case 1. Let $\hat{t} \in I$ be an interior time. Then there is an open interval containing \hat{t} on which $\lambda(t)$ is constant. Hence $\lambda(\cdot)$ is, trivially, continuous at \hat{t} .

Case 2. Let $\hat{t} \in I$ be a boundary time. Suppose that $\lambda(\cdot)$ is discontinuous at \hat{t} . Then, since $\lambda(\cdot)$ is nonincreasing and right-continuous, Lemma 2 implies that there is an $\varepsilon > 0$ such that

$$\lim_{\tau \rightarrow \hat{t}^-} \lambda(\tau) - \varepsilon > \lim_{\tau \rightarrow \hat{t}^+} \lambda(\tau) = \lambda(\hat{t}) = \lambda_b(\hat{t}).$$

But since $\lambda_b(\cdot)$ is continuous, there is an interval $(\bar{t}, \hat{t}) \subset I$ on which

$$\lambda_b(t) - \lambda_b(\hat{t}) \leq |\lambda_b(t) - \lambda_b(\hat{t})| < \varepsilon.$$

Thus, since $\lambda(\cdot)$ is nonincreasing,

$$\lambda_b(t) - (\lambda(t) - \varepsilon) \leq \lambda_b(t) - \left[\lim_{\tau \rightarrow \hat{t}^-} \lambda(\tau) - \varepsilon \right] < \lambda_b(t) - \lambda(\hat{t}) < \varepsilon$$

for all $t \in (\bar{t}, \hat{t})$. Hence, $\lambda(t) > \lambda_b(t)$ on (\bar{t}, \hat{t}) and so Lemma 1 implies that

$$\frac{d}{dt}[s'(t)x(t) - \sigma(t)] < 0 \quad \text{a.e. on } (\bar{t}, \hat{t}).$$

But since \hat{t} is a boundary time, this implies state violation on (\bar{t}, \hat{t}) which is a contradiction. Hence, we conclude, as required, that $\lambda(\cdot)$ is continuous on I .

Part B. The absolute continuity of $\lambda^*(\cdot)$ on \bar{I} will be proved. Define

$$\Lambda = \{t_0\} \cup \{t_1\} \cup \{\tau \in \bar{I} : s'(\tau)x(\tau) < \sigma(\tau)\},$$

$$\tilde{\Lambda} = \bar{I} \setminus \Lambda.$$

Since $\lambda_b(\cdot)$ is C^1 , and thus, a.c., there exists for each $\varepsilon > 0$ a $\delta(\varepsilon) > 0$ such that for all finite collections of pairwise disjoint subintervals of \bar{I} , $\{[a_i, b_i]\}_{i=1}^N$, satisfying

$$\sum_{i=1}^N |b_i - a_i| < \delta(\varepsilon),$$

we have the bound

$$\sum_{i=1}^N |\lambda_b(b_i) - \lambda(a_i)| < \varepsilon.$$

Define the following subintervals $[\hat{a}_i, \hat{b}_i] \subset [a_i, b_i]$:

If $[a_i, b_i] \not\subset \Lambda$, then

$$\hat{a}_i = \inf \{\tau \in [a_i, b_i] : \tau \in \tilde{\Lambda}\},$$

$$\hat{b}_i = \sup \{\tau \in [a_i, b_i] : \tau \in \tilde{\Lambda}\}.$$

Clearly \hat{a}_i and \hat{b}_i exist since there must be at least one point of $[a_i, b_i]$ in $\tilde{\Lambda}$. Now since intervals (a_i, \hat{a}_i) and (b_i, \hat{b}_i) are in Λ , Theorem 1 requires that $\lambda(\cdot)$ be constant on these intervals, so continuity of $\lambda^*(\cdot)$ implies that $\lambda^*(a_i) = \lambda^*(\hat{a}_i)$ and $\lambda^*(b_i) = \lambda^*(\hat{b}_i)$. Furthermore, since the constraint function $s'(t)x(t) - \sigma(t)$ is continuous, $\hat{a}_i \in \tilde{\Lambda}$ or $\hat{a}_i = t_0$. If $\hat{a}_i \in \tilde{\Lambda}$, then Lemma 2 requires that $\lambda^*(a_i) = \lambda_b(\hat{a}_i)$. If, however, $\hat{a}_i = t_0$, then there are boundary times arbitrarily close to \hat{a}_i and so, again, by continuity of $\lambda^*(\cdot)$ and $\lambda_b(\cdot)$ and Lemma 2, $\lambda^*(a_i) = \lambda_b(\hat{a}_i)$. Similarly, $\lambda^*(b_i) = \lambda_b(\hat{b}_i)$.

If $[a_i, b_i] \subset \Lambda$, then define $\hat{a}_i = a_i, \hat{b}_i = b_i$. Since $\lambda(\cdot)$ must be constant on Λ we have, trivially,

$$|\lambda^*(b_i) - \lambda^*(a_i)| = |\lambda_b(\hat{b}_i) - \lambda_b(\hat{a}_i)| = 0.$$

Hence, for all intervals,

$$\sum_{i=1}^N |\hat{b}_i - \hat{a}_i| \leq \sum_{i=1}^N |b_i - a_i| < \delta(\varepsilon)$$

and

$$\sum_{i=1}^N |\lambda^*(b_i) - \lambda^*(a_i)| = \sum_{i=1}^N |\lambda_b(\hat{b}_i) - \lambda_b(\hat{a}_i)| < \varepsilon$$

which verifies that $\lambda^*(\cdot)$ is a.c. on \bar{I} .

Part C. Finally, we verify the differential equation given for $\lambda^*(\cdot)$. If $t \in I$ is an interior time, then by continuity of the constraint function, there is an open interval of interior times containing t . Thus, trivially, since $\lambda(\cdot)$ must be constant for interior times, then $\dot{\lambda}(t) = 0$.

Next suppose $t_b \in I$ is a boundary time for which $\lambda(\cdot)$ is differentiable. (Since $\lambda(\cdot)$ is a.c., t_b is almost any boundary time.) Define

$$\bar{\lambda}(\tau) = \lambda_b(\tau) - \lambda^*(\tau).$$

By Lemma 2, $\bar{\lambda}(t_b) = 0$. First suppose $\dot{\lambda}(t_b) > 0$. Then there is a $\delta(t_b) > 0$, where

$$\dot{\lambda}(t_b) - \frac{\bar{\lambda}(t_b + \tau)}{\tau} \leq \left| \dot{\lambda}(t_b) - \frac{\bar{\lambda}(t_b + \tau)}{\tau} \right| < \frac{1}{2}\dot{\lambda}(t_b) \quad \text{if } 0 < \tau < \delta(t_b),$$

which implies that

$$\bar{\lambda}(t_b + \tau) > \frac{1}{2}\dot{\lambda}(t_b)\tau > 0 \quad \text{for all } \tau \in (0, \delta(t_b)).$$

Similarly, if $\dot{\lambda}(t_b) < 0$, there is a $\delta(t_b) > 0$, where

$$\frac{\bar{\lambda}(t_b + \tau)}{\tau} - \dot{\lambda}(t_b) \leq \left| \frac{\bar{\lambda}(t_b + \tau)}{\tau} - \dot{\lambda}(t_b) \right| < -\frac{1}{2}\dot{\lambda}(t_b) \quad \text{if } 0 < \tau < \delta(t_b),$$

which implies that

$$\bar{\lambda}(t_b - \tau) < \frac{1}{2}\dot{\lambda}(t_b)\tau < 0 \quad \text{if } 0 < \tau < \delta(t_b).$$

By Lemma 2, in either case for each such a boundary time t_b , either the interval $(t_b, t_b + \delta(t_b))$ or $(t_b - \delta(t_b), t_b)$ contains no boundary times and can, thus, supply a unique rational number. Hence, the set of all boundary times t_b for which $\dot{\lambda}^*(t_b) \neq \dot{\lambda}_b(t_b)$ has measure zero. This verifies the differential equation for $\lambda^*(\cdot)$ and, so, completes the proof. Q.E.D

Note that Lemma 3 does *not* rule out jumps in $\lambda(\cdot)$ at the end times t_0 and t_1 . In fact, discontinuities at these points can occur if the trajectory leaves the boundary at t_0 or enters it at t_1 . This is a result of the fact that at the endpoints the boundary may be hit in such a manner that the constraint would be violated were the system to be operated outside of \bar{I} .

Proof of the absolute continuity of $\lambda^*(\cdot)$ and its description by a differential equation now permit a considerable simplification of the optimality conditions of Theorem 1. In the next theorem the function $\lambda(\cdot)$ is eliminated by absorption into a new costate vector ($p(\cdot)$). The only vestige of $\lambda(\cdot)$ yet appearing in these new optimality conditions will be a parameter (α) which serves to permit a jump in $\lambda(\cdot)$ at t_1 (a possibility not ruled out by Lemma 3).

THEOREM 2. *An admissible arc (x, u) of the recoverable QCP is a nonsingular, optimal arc if and only if there exist a constant $\alpha \in R^1$ and functions $v(\cdot): \bar{I} \rightarrow R^1$ and $p(\cdot): \bar{I} \rightarrow R^n$, where $p(\cdot)$ is absolutely continuous on \bar{I} , which satisfy the conditions*

$$(15) \quad v(t) = \begin{cases} 1 & \text{if } s'(t)x(t) = \sigma(t), \\ 0 & \text{otherwise,} \end{cases}$$

$$(16) \quad \alpha \begin{cases} = 0 & \text{if } s'(t_1)x(t_1) < \sigma(t_1), \\ \leq 0 & \text{otherwise,} \end{cases}$$

$$(17) \quad \begin{aligned} \dot{p}(t) &= -A'(t)p(t) + R(t)x(t) \\ &\quad - v(t)s(t)[y'(t)x(t) + z'(t)p(t) + \beta(t)] \quad \text{a.e. on } I, \end{aligned}$$

$$(18) \quad p(t_1) = \alpha_1 s(t_1) - Qx(t_1),$$

$$(19) \quad v(t)[y'(t)x(t) + z'(t)p(t) + \beta(t)] \leq 0 \quad \text{for all } t \in I$$

and

$$(20) \quad u(t) = W^{-1}(t)B'(t)p(t) \quad \text{a.e. on } I,$$

where $y(\cdot)$, $z(\cdot)$ and $\beta(\cdot)$ are defined by equations (10)–(12).

Proof. Suppose that (x, u) is a nonsingular, optimal arc for the recoverable QCP. Then the conditions of Theorem 1 apply. Define λ^* by equation (13) and define

$$(21) \quad p(t) = \eta(t) - \lambda^*(t)s(t) \quad \text{for all } t \in \bar{I}.$$

Since it can be shown by direct computation that $z'(t)s(t) = 0$, Lemma 3 implies that

$$(22) \quad \dot{\lambda}^*(t) = v(t)[y'(t)x(t) + z'(t)p(t) + \beta(t)],$$

where $v(t)$ is defined as in (15).

By setting $\alpha = (\lambda(t_1) - \lambda^*(t_1))$ and using equations (21) and (22) in Theorem 1, equations (14) to (20) can be verified by direct substitution.

Conversely, suppose that (x, u) is an admissible arc of the recoverable QCP for which there exist parameters α , $p(\cdot)$ and $v(\cdot)$, with $p(\cdot)$ absolutely continuous, which satisfy equations (15)–(20). Define $\tilde{\lambda}(\cdot)$ to be an absolutely continuous scalar-value function on \bar{I} satisfying

$$\dot{\tilde{\lambda}}(t) = v(t)[y'(t)x(t) + z'(t)p(t) + \beta(t)] \quad \text{a.e. on } \bar{I},$$

with $\tilde{\lambda}(t_1) = -\alpha$. Then set

$$(23) \quad \lambda(t) = \begin{cases} \tilde{\lambda}(t) & \text{if } t \in [t_0, t_1), \\ \tilde{\lambda}(t_1) + \alpha & \text{if } t = t_1, \end{cases}$$

and

$$\eta(t) = p(t) + \tilde{\lambda}(t)s(t) \quad \text{for all } t \in \bar{I}.$$

Direct computation will now verify that the sufficient conditions of Theorem 1 are satisfied by $\lambda(\cdot)$ and $\eta(\cdot)$ and hence that (x, u) is an optimal arc. Q.E.D.

4. Riccati equation and fixed-point theorem. The linearity of the state-costate system and the absence of $\lambda(\cdot)$ in the optimality conditions of Theorem 2 suggests that, as in the unconstrained QCP, a matrix Riccati equation might be applicable. Indeed, for any particular choice of $v(\cdot)$ a Riccati equation can be obtained by straightforward application of the principles used in the unconstrained case. Based on equations (2), (17), (18) and (20), however, the solution to the resultant Riccati equation may not exist. An example of this is given in [13]. It is, therefore, convenient to modify equations (2) and (17) slightly before forming the Riccati equation. Since for an optimal arc equation (15) implies that $v(t)[s'(t)x(t) - \sigma(t)] = 0$ for all $t \in \bar{I}$, then the terms $v(t)z(t)[s'(t)x(t) - \sigma(t)]$ and $v(t)y(t)[s'(t)x(t) - \sigma(t)]$ may be added to the right-hand sides of equations (2) and (17), respectively. Rearranging the terms yields

$$(24) \quad \dot{x}(t) = \hat{A}(t)x(t) + \bar{B}(t)p(t) + w(t) - v(t)\sigma(t)z(t)$$

and

$$(25) \quad \dot{p}(t) = \hat{R}(t)x(t) - \hat{A}'(t)p(t) - v(t)[\beta(t)s(t) + \sigma(t)y(t)],$$

where

$$(26) \quad \hat{A}(t) = A(t) + v(t)z(t)s'(t)$$

and

$$(27) \quad \hat{R}(t) = R(t) + v(t)[y(t)s'(t) - s(t)y'(t)].$$

One can now use the standard procedure exactly as in the unconstrained case to express $p(t)$ in terms of $x(t)$. The result has the form

$$(28) \quad p(t) = -K(t)x(t) + \alpha s^*(t) + r(t).$$

Direct substitution of (28) into equations (24) and (25) and the use of the boundary conditions of (18) yields differential equations and boundary conditions for the Riccati matrix $K(t)$ and the vectors $s^*(t)$, $r(t)$.

$$(29) \quad \dot{K}(t) = K(t)\bar{B}(t)K(t) - K(t)\hat{A}(t) - \hat{A}'(t)K(t) - \hat{R}(t),$$

$$(30) \quad \dot{s}^*(t) = [K(t)\bar{B}(t) - \hat{A}'(t)]s^*(t),$$

$$(31) \quad \begin{aligned} \dot{r}(t) = & [K(t)\bar{B}(t) - \hat{A}'(t)]r(t) + K(t)[w(t) - v(t)\sigma(t)z(t)] \\ & - v(t)[\beta(t)s(t) + \sigma(t)y(t)]. \end{aligned}$$

$$(32) \quad K(t_1) = Q,$$

$$(33) \quad s^*(t_1) = s(t_1),$$

$$(34) \quad r(t_1) = 0.$$

For the unconstrained problem ($s(t) \equiv 0$, $\sigma(t) \equiv 0$) the above formulation gives the optimal control in feedback form via equation (20). In the constrained case, however, two problems arise. First, does a solution to (29) exist? Since the matrix $\hat{R}(t)$ in (27) has a skew-symmetric component, the matrix $K(t)$ will *not* be symmetric as in the unconstrained case and none of the usual existence proofs apply. Unfortunately, this existence question has not been completely answered. However, for a number of important cases existence is proved in § 7. The second difficulty is that the Riccati matrix $K(t)$ depends upon the unknown function $v(\cdot)$ and the resultant solution must also satisfy equations (5) and (15).

By examining this from a different point of view, one is led to a type of fixed-point theorem which can form the basis for an iterative computational algorithm to generate (an approximation to) the optimal control. Let V be a measurable subset of \bar{I} and define

$$(35) \quad v(t) = \begin{cases} 1 & \text{if } t \in V, \\ 0 & \text{if } t \notin V. \end{cases}$$

Now, given a set $V \subset \bar{I}$, one can compute a trajectory (dependent on V) by using the control given in (20) and the solution to the Riccati and related equations (29)–(34) in (28). Defining

$$(36) \quad \Gamma(V) = \{\tau \in \bar{I} : s'(\tau)x(\tau) \geq \sigma(\tau)\},$$

the sort of fixed-point result expected is that if the set V yields a nonsingular optimal arc (x, u) , then $\Gamma(V) = V$. Indeed, such a result is obtained.

LEMMA 4. *Let $v(t)$ be defined as in (35). For the recoverable QCP let $x(\cdot)$ and $p(\cdot)$ satisfy equations (2), (17) and (20). Then the state constraint and (15) are satisfied for all $t \in \bar{I}$ if and only if the state constraint holds at t_0 and t_1 and*

$$(36a) \quad V = \Gamma(V),$$

where $\Gamma(\cdot)$ is as in (36).

Proof. Assume for some measurable $V \subset \bar{I}$ that $v(t)$ is defined by (35) and that (15) and the state constraint are satisfied. The state constraint implies that

$$\Gamma(V) = \{t \in \bar{I} : s'(t)x(t) = \sigma(t)\}$$

and so, trivially, (15) implies that (36a) is satisfied.

Proving the converse, now, assume that the constraint is satisfied at t_0 and t_1 and that (36a) holds. Suppose there exists a $\tau \in \bar{I}$ such that $s'(\tau)x(\tau) > \sigma(\tau)$. Since the constraint holds at t_0 and t_1 , $\tau \in (t_0, t_1)$. Define

$$\tau_a = \inf \{\hat{t} \in \bar{I} : [\hat{t}, \tau] \subset \Gamma(V)\}$$

and

$$\tau_b = \sup \{\hat{t} \in \bar{I} : [\tau, \hat{t}] \subset \Gamma(V)\}.$$

(Clearly τ_a and τ_b exist since $\tau \in \Gamma(V)$.) Then $s'(\tau_a)x(\tau_a) = \sigma(\tau_a)$ follows from continuity if $t_0 < \tau_a$ and by the assumption that the constraint is satisfied at t_0 if $\tau_a = t_0$. Similarly $s'(\tau_b)x(\tau_b) = \sigma(\tau_b)$. Now since $[\tau_a, \tau_b] \subset \Gamma(V)$, (35) and (36a) imply that

$$v(t) = 1 \quad \text{for all } t \in [\tau_a, \tau_b].$$

Thus, after tedious but routine calculation using equations (2), (10), (11), (12), (17) and (20), we have

$$\frac{d}{dt} \left\{ \left[\frac{d}{dt} (s'(t)x(t) - \sigma(t)) \right] / (s'(t)\bar{B}(t)s(t)) \right\} = 0 \quad \text{for all } t \in [\tau_a, \tau_b].$$

Since for the recoverable QCP, $s'(t)\bar{B}(t)s(t)$ is strictly positive, we thus conclude that

$$\frac{d}{dt} [s'(t)x(t) - \sigma(t)]$$

is either identically zero or of strictly constant sign on $[\tau_a, \tau_b]$. In fact, since $s'(\tau)x(\tau) - \sigma(\tau) > 0$ and $s'(\tau_a)x(\tau_a) - \sigma(\tau_a) = 0$ we conclude that

$$\frac{d}{dt} [s'(t)x(t) - \sigma(t)] > 0 \quad \text{for all } t \in [\tau_a, \tau_b].$$

But this implies that

$$s'(\tau_b)x(\tau_b) - \sigma(\tau_b) \geq s'(\tau)x(\tau) - \sigma(\tau) > 0,$$

which contradicts the previous conclusion that $s'(\tau_b)x(\tau_b) - \sigma(\tau_b) = 0$. Hence, constraint violation cannot occur as assumed. From this we conclude that

$$\Gamma(V) = \{t \in \bar{I} : s'(t)x(t) = \sigma(t)\}$$

and so, that (15) holds, and the proof is complete. Q.E.D

The solution is thus seen to depend on determination of the set V and the parameter α . The necessary and sufficient conditions can be rewritten in a form more amenable for the determination of α by separating $x(t)$ and $p(t)$ each into two parts. When combined with Lemma 4, an interesting fixed-point theorem results.

THEOREM 3. *The pair (x, u) is an admissible, nonsingular, optimal arc for the recoverable QCP if*

$$(37) \quad x(t) = x^1(t) + \alpha x^2(t),$$

$$(38) \quad u(t) = W^{-1}(t)B'(t)p(t),$$

$$(39) \quad V = \Gamma(V) = \Gamma^*(V) \cap V,$$

where

$$(40) \quad p(t) = p^1(t) + \alpha p^2(t),$$

and the mappings $\Gamma(\cdot), \Gamma^*(\cdot) : \bar{I} \rightarrow \bar{I}$ are defined by

$$(41) \quad \Gamma(V) = \{\tau \in \bar{I} : s'(\tau)x(\tau) \geq \sigma(\tau)\}$$

and

$$(42) \quad \Gamma^*(V) = \{t_0\} \cup \{t_1\} \cup \{\tau \in (t_0, t_1) : y'(\tau)x(\tau) + z'(\tau)p(\tau) + \beta(\tau) \leq 0\}.$$

The vector functions $x^i(\cdot)$ and $p^i(\cdot)$ are defined to be absolutely continuous and to satisfy

$$(43) \quad \dot{x}^i(t) = A(t)x^i(t) + \bar{B}(t)p^i(t) + (2 - i)w(t), \quad i = 1, 2,$$

and

$$(44) \quad \begin{aligned} \dot{p}^i(t) = & -A'(t)p^i(t) + R(t)x^i(t) \\ & - v(t)s(t)[y'(t)x^i(t) + z'(t)p^i(t) + (2 - i)\beta(t)], \quad i = 1, 2, \end{aligned}$$

with $v(t)$ given by

$$(45) \quad v(t) = \begin{cases} 1 & \text{if } t \in V, \\ 0 & \text{otherwise,} \end{cases}$$

and the terminal conditions by

$$(46) \quad x^1(t_0) = x^0,$$

$$(47) \quad p^1(t_0) = -K(t_0)x^0 + r(t_0),$$

$$(48) \quad x^2(t_0) = 0$$

and

$$(49) \quad p^2(t_0) = s^*(t_0),$$

where $K(t_0)$, $r(t_0)$ and $s^*(t_0)$ are determined from the unique solution to the Riccati system defined by equations (29)–(34).

If $s'(t_1)x^1(t_1) > \sigma(t_1)$, it is sufficient for optimality that

$$(50) \quad \alpha = \bar{\alpha} \neq 0,$$

otherwise, sufficiency is guaranteed for either

$$(51) \quad \alpha = 0 \quad \text{or} \quad \alpha = \bar{\alpha},$$

where

$$(52) \quad \bar{\alpha} = \begin{cases} \text{if } [\sigma(t_1) - s'(t_1)x^1(t_1)][s'(t_1)x^2(t_1)] \geq 0, \\ \frac{\sigma(t_1) - s'(t_1)x^1(t_1)}{s'(t_1)x^2(t_1)} \quad \text{otherwise.} \end{cases}$$

Proof. Suppose that the pair (x, u) satisfies conditions (37)–(52). If $s'(t_1)x^1(t_1) > \sigma(t_1)$, then equations (50) and (52) provide the required nonzero $\alpha = \bar{\alpha}$ which by superposition forces $x(t_1)$ to the boundary. If, however, $s'(t_1)x^1(t_1) \leq \sigma(t_1)$, both $\alpha = 0$ and $\alpha = \bar{\alpha}$ satisfy the state constraint at t_1 . Note also from (52) that $\bar{\alpha} \leq 0$, so in all cases $\alpha \leq 0$ with $\alpha = 0$ unless $x(t_1)$ is on the boundary as required.

Thus, in either event, the constraint is satisfied at t_1 (and presumably at t_0 by x^0). Lemma 4 and the condition $V = \Gamma(V)$ from (39) thus imply that the constraint is satisfied for all $t \in \bar{I}$ and that $v(t)$, in fact, satisfies (15). Furthermore, α defined by equations (50)–(52) satisfies (16). The costate equation (17) holds by superposition from (40) and (44). From equations (37), (40), (46)–(49), superposition implies (as required) that

$$x(t_0) = x^0$$

and

$$p(t_0) = -K(t_0)x^0 + r(t_0) + \alpha s^*(t_0).$$

Thus the Riccati system assures that (18) is satisfied. The condition $V = \Gamma^*(V) \cap V$ from (39) implies that (19) holds. Finally, equations (38) and (20) are identical. Thus, the arc (x, u) has been shown to be admissible and to satisfy the conditions of Theorem 2. Hence, (x, u) is a nonsingular, optimal arc. Q.E.D

The only difficulty in showing the necessity of the conditions of Theorem 3 lies in verifying the existence of the solution to the Riccati system of equations (29)–(34). This question is still not entirely resolved, though Theorem 10 gives several alternative conditions sufficient for the existence of a unique solution to the Riccati system. Thus we state the following theorem.

THEOREM 3A. *Let (x, u) be an admissible nonsingular optimal arc for the recoverable QCP and for the measurable function $v(t)$ satisfying the conditions of Theorem 3 and let the Riccati system of equations (29)–(34) have a unique solution (for example, see Theorem 10). Then the conditions of Theorem 3 are necessary.*

Proof. Since (x, u) is an admissible, nonsingular, optimal arc then Theorem 2 requires that equations (15)–(20) hold for some scalar constant α , scalar function $v(\cdot)$ and absolutely continuous vector $p(\cdot)$. We shall show that, in fact, the same α , $v(\cdot)$ and $p(\cdot)$ which satisfy equations (15)–(20) also satisfy the conditions of Theorem 3. For the $v(t)$ which satisfies (15), define the set V by (45). Then, since (x, u) is admissible, Lemma 4 implies that $\Gamma(V) = V$, where $\Gamma(V)$ is defined by (41). Also, with V defined as above, equation (19) implies that $V = \Gamma^*(V) \cap V$, where $\Gamma^*(V)$ is defined by (43). Equation (38) follows from equations (20) and (16), respectively. It thus remains to verify that equations (37) and (40)–(50) hold for $x(\cdot)$, α , $v(\cdot)$ and $p(\cdot)$. By superposition the unique $x(\cdot)$ and $p(\cdot)$ which satisfy equations (37), (40), (43), (44) and (46)–(49) must satisfy the state equation

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

with

$$u(t) = W^{-1}(t)B'(t)p(t)$$

and the costate equation (17), with the boundary conditions

$$\begin{aligned} (*) \quad x(t_0) &= x^0, \\ p(t_0) &= -K(t_0)x^0 + s^*(t_0) + r(t_0). \end{aligned}$$

Thus, by the uniqueness of solutions for linear systems, we need only verify that the final condition of (18) implies (*) and that α of Theorem 2 satisfies equations (50)–(52) to complete the proof. Since there is a unique solution $K(\cdot)$ on $[t_0, t_1]$ to the Riccati system of equations (29) and (32), equation (18) implies that $p(t_0)$ is defined as in (*). Verifying α we note that if $s'(t_1)x^1(t_1) > \sigma(t_1)$, then by superposition, (16) and the state constraint require that (5) be satisfied with $\bar{\alpha}$ as defined by (52). On the other hand, if $s'(t_1)x^1(t_1) \leq \sigma(t_1)$, then (16) and the state constraint require (by superposition) that either $\alpha = 0$ or $\alpha = \bar{\alpha}$ (which might be zero). Q.E.D.

Observe that the sufficient (and often necessary) conditions of Theorem 3 encompass both admissibility and optimality. Actually, these conditions can be viewed as consisting only of equations (39) and (50) by regarding the remaining equations of Theorem 3 as mere definitions. Aside from the ambiguity in the definition for α in (51), the mappings $\Gamma(V)$ and $\Gamma^*(V)$ are uniquely defined and can be computed readily for any closed set $V \subset \bar{I}$. Thus, adopting either choice in (51) to define α , the QCP can be viewed as a set-valued, fixed-point problem as defined by (39). If a set V can be found satisfying (39) and also the inequality (50), the QCP is solved. If not, the other definition for α must be used and the fixed-point problem solved again.

As is probably apparent, the parameter α serves to prevent boundary violation at $t = t_1$. The possibility that two different definitions for α might have to be tried to solve the QCP is not at all unreasonable when one compares the QCP with the much simpler, terminal-constrained QCP problem in which the state constraint is only imposed at $t = t_1$. The conditions of Theorem 3, with the exception of (39), apply to the terminal-constrained QCP if $V = \emptyset$. In that case one solves first the unconstrained problem ($\alpha = 0$) and if this arc does not violate

the constraint the problem is solved. If, however, the constraint is violated, $\alpha = \bar{\alpha} \neq 0$ is computed by superposition to force the trajectory to the boundary at $t = t_1$, and the trajectory is recomputed to solve the problem. Thus, it is reasonable that imposition of the constraint throughout the trajectory still results in a certain ambiguity for the parameter α .

5. Computational algorithm. As an illustration of the validity of Theorem 3, a set-valued, fixed-point, iterative algorithm is developed and applied to two examples. The proposed algorithm is certainly not the last word on the solution of this problem, but it is hoped that it may stimulate further research into the application of Theorem 3.

Theorem 3 suggests that for an iterative computational procedure one could guess a set $V_0 \subset \bar{I}$ and compute $\Gamma(V_0)$ and $\Gamma^*(V_0)$. If V_0 corresponds to a solution, then $\Gamma(V_0) = \Gamma^*(V_0) \cap V_0 = V_0$. If not, V_0 must be modified in some way to more nearly satisfy this condition. An intuitively appealing notion is that in some sense the points in $V_0 \setminus (\Gamma(V_0) \cap \Gamma^*(V_0))$ may represent "extra" points that do not belong in V_0 , while points in $(\Gamma(V_0) \cap \Gamma^*(V_0)) \setminus V_0$ represent points which perhaps should be in V_0 . Algorithms can then be developed by selectively adding or deleting points from V_0 with the objective of satisfying (39). We present one such algorithm. To formalize this idea define the mappings $D_i(\cdot): \bar{I} \rightarrow \bar{I}$ as follows:

$$(53) \quad D_0(V) = V \cap \Gamma(V) \cap \Gamma^*(V);$$

$$(54) \quad D_1(V) = V \cap \Gamma(V) \cap \tilde{\Gamma}^*(V);^1$$

$$(55) \quad D_2(V) = V \cap \tilde{\Gamma}(V) \cap \Gamma^*(V);$$

$$(56) \quad D_3(V) = V \cap \tilde{\Gamma}(V) \cap \tilde{\Gamma}^*(V);$$

$$(57) \quad D_4(V) = \tilde{V} \cap \Gamma(V) \cap \Gamma^*(V);$$

$$(58) \quad D_5(V) = \tilde{V} \cap \Gamma(V) \cap \tilde{\Gamma}^*(V).$$

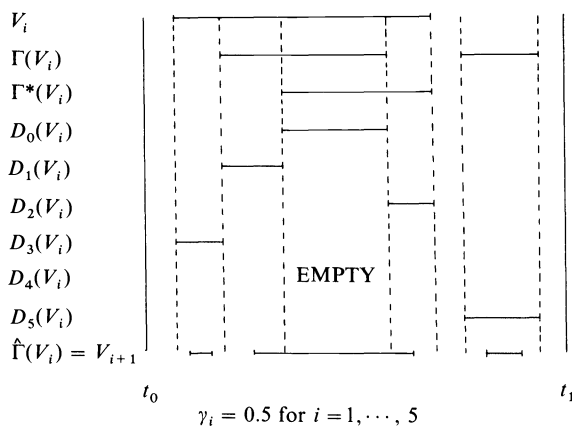
The set $D_0(V)$ forms the core which is to be kept in forming a new V . In addition, subsets of each of the $D_i(V)$, $i = 1, \dots, 5$, will be added to this. For illustration, suppose the sets V , $\Gamma(V)$ and $\Gamma^*(V)$ are as shown in Fig. 1. Then the sets $D_i(V)$ are as illustrated.

It is evident that the sets $D_i(V)$ need not be connected. Thus let

$$(59) \quad D_i(V) = \bigcup_{j=1}^{k_i} D_i^j(V), \quad i = 0, \dots, 5,$$

where the sets $D_i^j(V)$ are pairwise disjoint nontouching subintervals of \bar{I} . Let constants γ_i be given ($i = 1, \dots, 5$) with $\gamma_i \in [0, 1]$ for $i = 1, \dots, 3$ and $\gamma_i \in (0, 1]$ for $i = 4, 5$. The subsets of the $D_i(V)$ to be used will be found by taking a fractional length γ_i of each of the $D_i^j(V)$. For a reasonable procedure, the geometry of the situation must be considered. Thus,

¹ The notation " \tilde{V} " will be used here to denote the complement of the set V with respect to the interval \bar{I} .

FIG. 1. Generation of $\hat{\Gamma}(V_i)$ from V_i , $\Gamma(V_i)$ and $\Gamma^*(V_i)$

- I. Suppose that $D_i^j(V)$ is *not* adjacent (either on the left or the right) to an interval of $D_0(V)$. Then define $\hat{D}_i^j(V)$ to be a *centered* subinterval of $D_i^j(V)$ having a fractional length γ_i of that of $D_i^j(V)$.
- II. Suppose that $D_i^j(V)$ is adjacent on the right (left) but *not* on the left (right) to an interval of $D_0(V)$. Then define $\hat{D}_i^j(V)$ to be a subinterval of $D_i^j(V)$ terminating (originating) at the extreme right (left) of the interval $D_i^j(V)$ and having fractional length γ_i of that of $D_i^j(V)$.
- III. Suppose that $D_i^j(V)$ is adjacent *both* on the left and the right to intervals of $D_0(V)$. Then define $\hat{D}_i^j(V)$ to consist of two subintervals of $D_i^j(V)$ each having fractional length $\gamma_i/2$ of that of $D_i^j(V)$. One of the subintervals originates at the extreme left and the other terminates at the extreme right of the interval $D_i^j(V)$.

Now define

$$(60) \quad \hat{\Gamma}(V) = D_0(V) \cup \hat{D}_1(V) \cup \hat{D}_2(V) \cup \hat{D}_3(V) \cup \hat{D}_4(V) \cup \hat{D}_5(V).$$

It can be shown that for $D_i(V) \neq \emptyset$, if the $\hat{D}_i(V)$ are proper subsets of $D_i(V)$ for $i = 1, \dots, 3$, and nonempty subsets for $i = 4, 5$, and if $\hat{D}_i(V) = \emptyset$ if $D_i(V) = \emptyset$ for $i = 1, \dots, 5$, then the fixed-point condition of Theorem 3 is satisfied if and only if $V = \hat{\Gamma}(V)$. Since the above conditions are met by \hat{D}_i as defined by I–III provided $\gamma_i \in (0, 1)$ for $i = 1, \dots, 5$, we shall consider a computational algorithm based on iterates of the form

$$V_{i+1} = \hat{\Gamma}(V_i).$$

Fig. 1 illustrates the generation of the new iterate V_{i+1} for $\gamma_i = 1/2$, $i = 1, \dots, 5$.

In developing the computational algorithm, it will be convenient to treat the fractional constants γ_i as step sizes, much as one might do with an ordinary gradient procedure. Further, one can think of the sets $\hat{D}_1(V_i), \dots, \hat{D}_3(V_i)$ as retaining part of the old V_i and $\hat{D}_4(V_i)$ and $\hat{D}_5(V_i)$ as introducing new points into V_{i+1} . Thus, set

$$(61) \quad \gamma_i = \begin{cases} \gamma \bar{\gamma}_i, & i = 1, \dots, 3, \\ (1 - \gamma) \bar{\gamma}_i, & i = 4, 5, \end{cases}$$

where the $\bar{\gamma}_i \in (0, 1)$. $\gamma \in [0, 1]$ may thus be treated as a step size parameter. For $\gamma = 0$, $V_{i+1} = V_i$. Many measures of the satisfaction of the desired conditions could be used. Here we use

$$(62) \quad e(V) = \sup_{\tau \in V \cup \bar{\Gamma}(V)} |s'(\tau)x(\tau) - \sigma(\tau)|.$$

The algorithm may now be stated.

1. Choose an initial V_0 (the empty set is a convenient choice), the constants $\bar{\gamma}_i$, $i = 1, \dots, 5$, and a stopping parameter ε . Set $i = 0$.
2. Solve the Riccati equation and obtain $K(\cdot)$, $s^*(\cdot)$ and $r(\cdot)$.
3. Obtain α and solve the state and costate equation.
4. Determine $e(V_i)$. If $e(V_i) \leq \varepsilon$, stop. If not, determine $\gamma \in [0, 1]$ to minimize $e(V_i)$ (one dimensional search).
5. Set $V_{i+1} = \hat{\Gamma}(V_i)$, and go to 2.

In practice, γ is held at 1 for the first few iterations until the vicinity of the solution is reached. This eliminates a few solutions of the Riccati equation. Also, in the one dimensional search, three trial values of γ are used to fit a quadratic function. The predicted minimum is compared with the other trial values and the best one chosen. No attempt at a further refinement of γ is made.

6. Application of the algorithm. Two examples will be presented to illustrate the behavior of the algorithm. The computational results to be given were obtained from an IBM 360/67 digital computer using single precision arithmetic. It should also be remarked that the coding was done exclusively in the FORTRAN IV G Level compiler language. The state-costate and Riccati differential equations were approximately solved using a fourth order Runge-Kutta integration scheme with uniform step size.

Example 1. Consider the second order QCP with scalar control and fixed initial and free final state defined on the interval $\bar{I} = [0, 1]$ by

$$(63) \quad J(x, u) = \frac{1}{2}(x_2(1))^2 + \frac{1}{2} \int_0^1 (x_1(t))^2 + (u(t))^2 dt,$$

$$\dot{x} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u, \quad x(0) = \begin{bmatrix} -1 \\ 0 \end{bmatrix},$$

$$x_2(t) \leq 0.05 \quad \text{for all } t \in [0, 1].$$

ε was chosen to be 0.01 and $\bar{\gamma}_i = 0.5$, $i = 1, \dots, 5$. All integration was performed using a step size of 0.01. The procedure converged to within the prescribed limits after 7 iterations. A further improvement is not possible without reducing the integration step size. Fig. 2 illustrates the progress of the algorithm. $x_2(t) - 0.05$ is shown for each iteration. The optimal cost for the constrained problem was 0.4864.

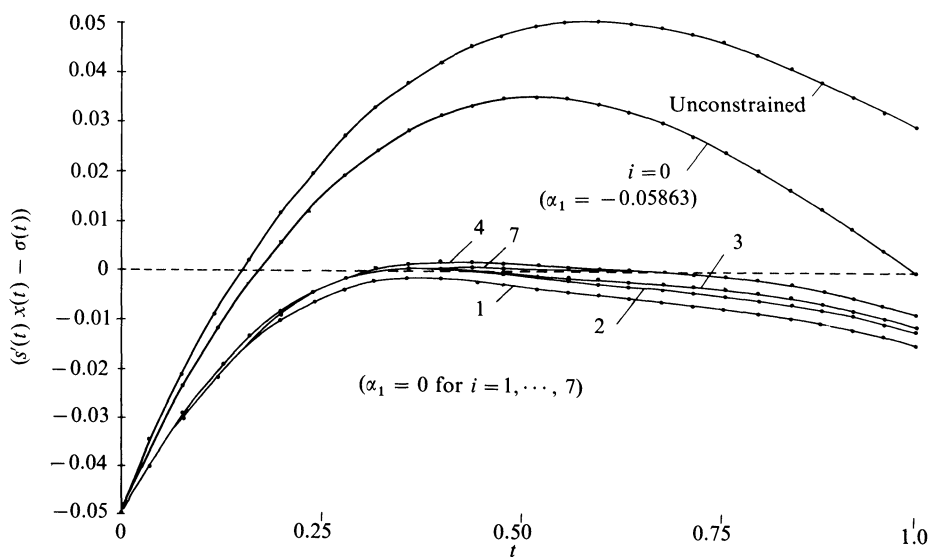


FIG. 2. Iterative improvement in the state constraint for Example 1

Example 2. To illustrate the use of the algorithm for multiple boundary intervals, a problem artificially constructed to have two boundary intervals $[-2, -1]$ and $[1, 2]$ is considered.

$$\dot{x} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} u; \quad x(-3) = \begin{bmatrix} -5869.2 \\ 2488.8 \\ -701.64 \\ 72.0 \end{bmatrix},$$

$$x_4 \leq 103.0.$$

The parameters in the cost function of (1) are

$$R = 0, \quad Q = 1$$

and

$$W = 1,$$

and an integration interval $[-3, 3]$ was used with an integration step size of 0.015. Fig. 3 shows the progression of $x_4(t) - 103.0$ for the first few iterations of the algorithm. The optimal cost found was 2471.7. It can be seen that the algorithm successfully found the two boundary intervals.

7. Existence and uniqueness of the Riccati solution. Existence, uniqueness and stability of solutions to the classical, symmetric Riccati equation have been considered in great detail by Kalman [11], Potter [14], Kleinman [15] and others. In addition, various modifications to the classical problem have been investigated. Kleinman and Falb [16] have extended the existence proofs for the classical problem to a Hilbert space setting. Wonham [17] has considered the addition of

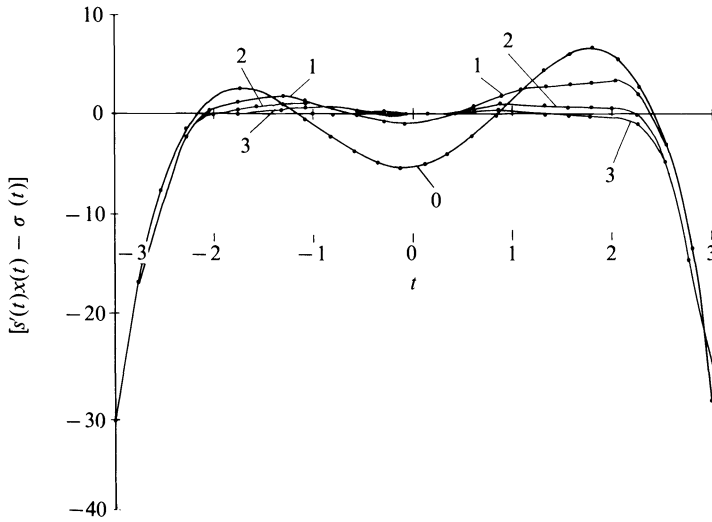


FIG. 3. Iterative improvements in the state constraint for Example 2

a term of the form $\Pi(t, K(t))$ to the expression for \dot{K} of the standard Riccati equation. However, in both cases, symmetry (self-adjointness in Hilbert space) has been maintained. In view of the results obtained in the above, however, we shall be interested here in relaxing the symmetry requirements.

While it is only required that skew symmetries be allowed in the driving function $\hat{R}(t)$ of (27), we shall consider a somewhat more general problem. In particular, let us define the nonsymmetric Riccati system (NRS) as

$$(64) \quad \dot{\tilde{K}}(t) = \tilde{K}(t)\tilde{B}(t)\tilde{K}(t) - \tilde{A}'(t)\tilde{K}(t) - \tilde{K}(t)\tilde{A}(t) - \tilde{R}(t) \quad \text{a.e. on } \bar{I},$$

with the terminal condition

$$(65) \quad \tilde{K}(t_1) = \tilde{Q}.$$

The following assumptions are made for the $n \times n$ dimensional matrix parameters of the NRS:

(NRS-1) $\tilde{B}(t)$ is integrable and positive semidefinite on $[t_0, t_1]$;

(NRS-2) $\tilde{A}(t)$ is integrable on $[t_0, t_1]$;

(NRS-3) $\tilde{R}(t)$ is integrable and positive semidefinite on $[t_0, t_1]$;

(NRS-4) \tilde{Q} is a positive semidefinite constant.

(Note that contrary to common practice the term "positive semidefinite" as used here does not imply symmetry.)

For convenience we shall consider what appears to be a special case of the NRS. Let the $n \times n$ matrix $T(t)$ be required to satisfy

$$(66) \quad \dot{T}(t) = T(t)E(t)T(t) - F(t) \quad \text{a.e. on } \bar{I}$$

and

$$(67) \quad T(t_1) = \tilde{Q},$$

where

$$(68) \quad E(t) = \Phi(t_1, t)\tilde{B}(t)\Phi'(t_1, t),$$

$$(69) \quad F(t) = \Phi'(t, t_1)\tilde{R}(t)\Phi(t, t_1),$$

$$(70) \quad \dot{\Phi}(t, \tau) = \tilde{A}(t)\Phi(t, \tau) \quad \text{for almost all } t \in \bar{I} \text{ and for all } \tau \in \bar{I}$$

and

$$(71) \quad \Phi(\tau, \tau) = (\text{identity matrix}) \quad \text{for all } t, \tau \in \bar{I}.$$

Notice that $E(\cdot)$ and $F(\cdot)$ possess the same properties required of $\tilde{B}(\cdot)$ and $\tilde{R}(\cdot)$ by conditions (NRS-1) and (NRS-3), respectively.

LEMMA 5. *If $\tilde{K}(t)$ is a solution for the NRS on some interval $(t_a, t_1] \subset \bar{I}$, then*

$$(72) \quad T(t) = \Phi'(t, t_1)\tilde{K}(t)\Phi(t, t_1)$$

satisfies (67) and is a solution to the system of (66) on $(t_a, t_1]$. Conversely, if $T(t)$ satisfies equations (66) and (67) on an interval $(t_a, t_1]$, then

$$(73) \quad \tilde{K}(t) = \Phi'(t_1, t)T(t)\Phi(t_1, t)$$

is a solution for the NRS on $(t_a, t_1]$.

Proof. Direct computation verifies the results. Q.E.D.

Thus, equations (66)–(71) simply represent another form for the NRS. We shall refer to both forms as the NRS. The appropriate system will be made clear by whether $T(\cdot)$ or $\tilde{K}(\cdot)$ is mentioned.

For the symmetric Riccati equation the positive semidefiniteness of the solution is well known. (See Kleinman [15].) However, this result is generally obtained by making use of the symmetry. We offer here a proof of positive semidefiniteness which applies to the NRS.

THEOREM 4. *If $\tilde{K}(t)(T(t))$ is a solution for the NRS on an interval $(t_a, t_1] \subset \bar{I}$, then $\tilde{K}(t)(T(t))$ is positive semidefinite for all $t \in (t_a, t_1]$. If, in addition, \tilde{Q} is positive definite, then $\tilde{K}(t)(T(t))$ is positive definite for all $t \in (t_a, t_1]$.*

Proof. Assume that $T(t)$ is a solution for the NRS on $(t_a, t_1] \subset \bar{I}$. Define the transition matrix $\Phi_{ET}(\cdot, \cdot)$ by

$$\frac{d}{dt}\Phi_{ET}(t, \tau) = -E(t)T(t)\Phi_{ET}(t, \tau)$$

with

$$\Phi_{ET}(\tau, \tau) = (\text{identity matrix}) \quad \text{for all } \tau \in (t_a, t_1].$$

On any closed subinterval $[\hat{t}_a, t_1] \subset (t_a, t_1]$, the continuous matrix function $T(t)$ is bounded and by (NRS-1) and (68), $E(t)$ is integrable. Thus, the product $(-E(t)T(t))$ is integrable on $[\hat{t}_a, t_1]$. Hence, both $\Phi_{ET}(t, \tau)$ and

$$\Phi_{ET}^{-1}(t, \tau) = \Phi_{ET}(\tau, t)$$

exist on $[\hat{t}_a, t_1]$ and hence, on $(t_a, t_1]$. By direct computation,

$$\frac{d}{d\tau} [\Phi'_{ET}(\tau, t) T(\tau) \Phi_{ET}(\tau, t)] = -\Phi'_{ET}(\tau, t) [T'(\tau) E'(\tau) T(\tau) + F(\tau)] \Phi_{ET}(\tau, t).$$

Hence, upon integration from t to t_1 , we obtain

$$\begin{aligned} T(t) &= \Phi'_{ET}(t_1, t) \tilde{Q} \Phi_{ET}(t_1, t) \\ &\quad + \int_t^{t_1} \Phi'_{ET}(\tau, t) [T'(\tau) E'(\tau) T(\tau) + F(\tau)] \Phi_{ET}(\tau, t) d\tau \quad \text{for all } t \in (t_a, t_1]. \end{aligned}$$

With the above representation for $T(t)$, the conclusions of the theorem follow, trivially, from the definiteness conditions on \tilde{Q} , $E(t)$ and $F(t)$. (Observe that symmetry is of no consequence for these parameters.) Positive (semi) definiteness of $T(t)$ implies positive (semi) definiteness of $\tilde{K}(t)$ by (73). Thus, the conclusions apply to both $T(t)$ and $\tilde{K}(t)$. Q.E.D.

As a prelude to the existence and uniqueness theorem we now verify that the NRS satisfies a *local* Lipschitz condition.

LEMMA 6. *For any $n \times n$ matrices T_1 and T_2 satisfying $\|T_1\|, \|T_2\| \leq \mu$, then,*

$$\|[T_1 E(t) T_1 - F(t)] - [T_2 E(t) T_2 - F(t)]\| \leq 4\mu \|E(t)\| \|T_1 - T_2\|.$$

Proof. Let $\|T_1\|, \|T_2\| \leq \mu$. Then

$$\begin{aligned} &\|T_1 E(t) T_1 - T_2 E(t) T_2\| \\ &= \|(T_1 + T_2) E(t) (T_1 - T_2) + (T_1 - T_2) E(t) T_1 - T_1 E(t) (T_1 - T_2)\| \\ &\leq \|E(t)\| [3\|T_1\| + \|T_2\|] \|T_1 - T_2\| \leq 4\mu \|E(t)\| \|T_1 - T_2\|. \quad \text{Q.E.D.} \end{aligned}$$

We begin the discussion of existence by presenting the well-known result for the case where $E(t)$, $F(t)$ and \tilde{Q} are symmetric for all $t \in [t_0, t_1]$. (See Kalman [11].) The bound obtained for this case will serve as an interesting comparison with results to be presented for the more general problem.

THEOREM 5. *If $E(t)$, $F(t)$ and \tilde{Q} are symmetric for all $t \in [t_0, t_1]$, then there is a unique solution $T(t)$ to the NRS defined on $[t_0, t_1]$. Furthermore, this solution is symmetric and satisfies the bound*

$$(74) \quad \|T(t)\| \leq \|\tilde{Q}\| + \int_t^{t_1} \|F(\tau)\| d\tau \quad \text{for all } t \in [t_0, t_1]:$$

Proof. The local Lipschitz property of Lemma 6 guarantees existence and uniqueness on some interval $(\hat{t}_a, t_1]$. (See [18].) Define $t_a = \inf \{\hat{t}_a \in [t_0, t_1] : \text{solution to NRS exists on } (\hat{t}_a, t_1]\}$. Clearly $T'(t)$ is a solution to the NRS wherever $T(t)$ is by the symmetry of the parameters $E(t)$, $F(t)$ and \tilde{Q} . Thus, by the uniqueness of the solution, $T(t)$ is symmetric. From Theorem 10.5.5 of Dieudonné [18], if we can show that $T(t)$ is bounded on $(t_a, t_1]$ and if $t_0 < t_a$, then there exists a time $t_b \in [t_0, t_a)$ such that the solution exists on $(t_b, t_1]$. This would contradict the definition of t_a . Hence, it suffices to obtain the above bound to verify that $t_a = t_0$ and complete the proof.

Since $T(t)$ is both symmetric and positive semidefinite (by Theorem 4) on $(t_a, t_1]$, then $\|T(t)\| = \sup_{\|x\|=1} x'T(t)x$. For any $x \in R^n$ the solution must satisfy

$$x'T(t)x = x'\tilde{Q}x + \int_t^{t_1} -x'T(\tau)E(\tau)T(\tau)x + x'F(\tau)x d\tau \leq x'\tilde{Q}x + \int_t^{t_1} x'F(\tau)x d\tau$$

for all $t \in (t_a, t_1]$ since $T(t)$ is symmetric by the above argument and $E(t)$ is positive semidefinite by condition (NRS-1). Hence,

$$\|T(t)\| \leq \|\tilde{Q}\| + \int_t^{t_1} \|F(\tau)\| d\tau \quad \text{for all } t \in (t_a, t_1],$$

since $T(t)$, $F(t)$ and \tilde{Q} are positive semidefinite and symmetric. Q.E.D.

The interesting point to note concerning Theorem 5 is that the bound obtained for the $\|T(t)\|$ is independent of the parameter $E(t)$. In a sense the nonlinear term $T(t)E(t)T(t)$ of (66) actually contributes a stabilizing influence on the solution in the symmetric case as a consequence of the positive semidefiniteness assumptions (NRS-1, 3 and 4). That this is *not* the case for the more general NRS will be illustrated by the following example.

Example 3. Let the parameters for the second order NRS be given by

$$E(t) \equiv \begin{bmatrix} 0 & 0 \\ 0 & \mu^2 \end{bmatrix},$$

$$F(t) \equiv 0$$

and

$$\tilde{Q} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix},$$

where μ is some fixed scalar. For this simple example the skew symmetry is only introduced through the terminal parameter \tilde{Q} . (A similar example can be generated by interchanging the values of \tilde{Q} and $F(t)$.) Let us see what this skew symmetry does to the solution (if it exists). Define the components of $T(t)$ as

$$T(t) = \begin{bmatrix} T_{11}(t) & T_{12}(t) \\ T_{21}(t) & T_{22}(t) \end{bmatrix}.$$

Then from (66) we have

$$\dot{T}_{11}(t) = \mu^2 T_{12}(t)T_{21}(t),$$

$$\dot{T}_{12}(t) = \mu^2 T_{12}(t)T_{22}(t),$$

$$\dot{T}_{21}(t) = \mu^2 T_{22}(t)T_{21}(t)$$

and

$$\dot{T}_{22}(t) = \mu^2 T_{22}(t).$$

The terminal conditions are given by (67) as

$$\tilde{Q} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

Hence, we see that

$$T_{22}(t) \equiv 0$$

and, thus, that

$$T_{12}(t) \equiv 1 = -T_{21}(t)$$

which imply that

$$T_{11}(t) = \mu^2(t_1 - t).$$

This solution, thus, indicates that a bound cannot be obtained for the $\|T(t)\|$ which is independent of $E(t)$, (i.e., of μ) in general, when skew symmetries are injected into the system.

Example 3 raises serious doubts as to the ability to prove the existence of the solution for the NRS by any procedure similar to that of Theorem 5. Clearly, the nonlinear term $T(t)E(t)T(t)$ of (66) can contribute to the growth of the solution in the nonsymmetric case. Thus, the question of the existence of a solution for the general NRS is considerably subtler than for the symmetric case.

Up to this point we have framed the NRS in a way which allows for skew symmetries in all of the parameters. However, we should note from the outset that it is impossible to obtain a general existence result for this problem. In particular, if skew symmetries are permitted in the parameter $E(t)(\tilde{B}(t))$, then finite escape can occur as is evidenced by the next example.

Example 4. For the second order NRS define the parameters

$$E(t) \equiv \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix},$$

$$F(t) \equiv 0$$

and

$$\tilde{Q} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

Define

$$T(t) = \begin{bmatrix} T_{11}(t) & T_{12}(t) \\ T_{21}(t) & T_{22}(t) \end{bmatrix}.$$

Then from equations (66) and (67) we have

$$\dot{T}_{11}(t) = T_{11}(t)[T_{21}(t) - T_{12}(t)],$$

$$\dot{T}_{12}(t) = T_{11}(t)T_{22}(t) - T_{12}^2(t),$$

$$\dot{T}_{21}(t) = T_{21}^2(t) - T_{22}(t)T_{11}(t),$$

$$\dot{T}_{22}(t) = T_{22}(t)[T_{21}(t) - T_{12}(t)],$$

$$T_{11}(t_1) = T_{22}(t_1) = 0$$

and

$$T_{12}(t_1) = -T_{21}(t_1) = 1.$$

Clearly, $T_{11}(t) \equiv T_{22}(t) \equiv 0$. Consider only the component $T_{12}(t)$. It satisfies

$$\dot{T}_{12}(t) = -T_{12}^2(t).$$

Thus,

$$T_{12}(t) = \frac{1}{(t - t_1 + 1)}$$

and finite escape occurs at $t = t_1 - 1$.

Example 4 does not clearly incriminate the skew symmetric $E(t)$ as being solely responsible for escape since a skew symmetric \tilde{Q} was also employed for convenience. However, the example does demonstrate that, in general, existence cannot be guaranteed for the NRS. In addition, when coupled with the existence theorems to be presented for the case when $E(t)$ is symmetric, Example 4 will verify that the symmetry requirement on $E(t)$ cannot be dropped. It should be emphasized, of course, that for the QCP, the corresponding $E(t)$ is symmetric since $\bar{B}(t)$ is symmetric, and so we suffer no loss of generality for our intended application of the Riccati equation. (See (29).)

From the proof to Theorem 5 we see that an essential element in our existence proofs must be the establishment of a bound on $T(t)$ on any interval $(t_a, t_1] \subset (t_0, t_1]$ on which it is defined. We state next a continuity lemma which will help in this.

LEMMA 7. *Let $M(t)$ be an $n \times n$ differentiable matrix defined on the interval $(t_a, t_1]$ satisfying $M(t_1) = M_1$ and*

$$(75) \quad \|M(t)X\| \frac{d}{dt} \|M(t)X\| \geq -\|M(t)X\| \|N(t)X\| \quad \text{a.e. on } (\hat{t}_a, t_1]$$

for all $X \in E^n$, where $N(t)$ is $n \times n$ integrable matrix for $t \in (\hat{t}_a, t_1]$. Then

$$(76) \quad \|M(t)\| \leq \|M_1\| + \int_t^{t_1} \|N(\tau)\| d\tau \quad \text{for all } t \in (\hat{t}_a, t_1].$$

Proof. First suppose that $M(t)X$ has no roots on some interval $[\hat{t}, t_1] \subset (t_a, t_1]$. Then both sides of (75) may be divided by $\|M(t)X\|$ in this interval. Then, integrating from t to t_1 we obtain

$$\|M(t)X\| \leq \|M_1X\| + \int_t^{t_1} \|N(\tau)X\| d\tau.$$

On the other hand, if on the interval $[\hat{t}, t_1] \subset (t_a, t_1]$ the function $\|M(t)X\|$ has at least one root, define

$$t^* = \inf \{t \in [\hat{t}, t_1] : \|M(t)X\| = 0\}.$$

Then, by continuity $\|M(t)X\| = 0$. If $t^* \neq \hat{t}_1$, $\|M(t)X\| \neq 0$ for all $t \in (\hat{t}, t^*)$. Thus,

$$\|M(\hat{t})X\| \leq \int_{\hat{t}}^{t^*} \|N(\tau)X\| d\tau.$$

If $t = \hat{t}$, then $\|M(\hat{t})X\| = 0$. Combining the above results, we have, in any event,

$$\|M(t)X\| \leq \|M_1X\| + \int_t^{t_1} \|N(\tau)X\| d\tau \quad \text{for all } t \in (t_a, t_1].$$

Since this holds for all $X \in E^n$, we have

$$\|M(t)\| \leq \|M_1\| + \int_t^{t_1} \|N(\tau)\| d\tau \quad \text{for all } t \in (t_a, t_1]. \quad \text{Q.E.D.}$$

THEOREM 6. *For the second order NRS with $E(t)$ symmetric for all $t \in [t_0, t_1]$, a unique solution $T(t)$ exists for all $t \in [t_0, t_1]$. This solution satisfies the bound*

$$(77) \quad \|T_s(t)\| \leq \|\tilde{Q}_s\| + \int_t^{t_1} \|F_s(\tau)\| d\tau + \left[\|\tilde{Q}_a\| + \int_t^{t_1} \|F_a(\tau)\| d\tau \right]^2 \int_t^{t_1} \|E(\tau)\| d\tau$$

and

$$(78) \quad \|T_a(t)\| \leq \|\tilde{Q}_a\| + \int_t^{t_1} \|F_a(\tau)\| d\tau \quad \text{for all } t \in [t_0, t_1],$$

where the subscript “s” denotes “symmetric part” and the subscript “a” denotes “skew symmetric part”.

Proof. Define

$$\mathcal{S} \begin{cases} \dot{T}_1(t) = T_1(t)E(t)T_1(t) + T_2(t)E(t)T_2(t) - F_s(t) & \text{a.e. on } \tilde{I}, \\ T_1(t_1) = \tilde{Q}_s, \\ \dot{T}_2(t) = T_1(t)E(t)T_2(t) + T_2(t)E(t)T_1(t) - F_a(t) & \text{a.e. on } \tilde{I}, \\ T_2(t_1) = \tilde{Q}_a, \end{cases}$$

where $T_1(t)$ and $T_2(t)$ are $n \times n$ matrixes. We first verify that the above system satisfies a local Lipschitz condition in a way similar to Lemma 6. Let us consider the pair $[T_1, T_2] \in R^{n^2} \times R^{n^2}$, and define

$$\|[T_1, T_2]\| = \max(\|T_1\|, \|T_2\|),$$

where $\|T_i\|$ is the usual “sup norm”. (Clearly the above norm is a valid norm in $R^{n^2} \times R^{n^2}$.) Let $[T_1^a, T_2^a], [T_1^b, T_2^b] \in R^{n^2} \times R^{n^2}$ with

$$\|[T_1^a, T_2^a]\|, \|[T_1^b, T_2^b]\| \leq \mu.$$

Then

$$\begin{aligned} & \|(T_1^a E(t) T_1^a + T_2^a E(t) T_2^a) - (T_1^b E(t) T_1^b + T_2^b E(t) T_2^b)\| \\ (*) \quad &= \left\| (T_1^a + T_1^b)E(t)(T_1^a - T_1^b) + (T_1^a - T_1^b)E(t)T_1^a - T_1^a E(t)(T_1^a - T_1^b) \right. \\ & \quad \left. + (T_2^a + T_2^b)E(t)(T_2^a - T_2^b) + (T_2^a - T_2^b)E(t)T_2^a - T_2^a E(t)(T_2^a - T_2^b) \right\| \\ & \leq 4\mu \|E(t)\| (\|T_1^a - T_1^b\| + \|T_2^a - T_2^b\|) \\ & \leq 8\mu \|E(t)\| \|[T_1^a - T_1^b, T_2^a - T_2^b]\|. \end{aligned}$$

Next consider

$$\begin{aligned}
 & \| (T_1^a E(t) T_2^a + T_2^a E(t) T_1^a) - (T_1^b E(t) T_2^b + T_2^b E(t) T_1^b) \| \\
 (**) \quad &= \left\| \begin{pmatrix} (T_1^a - T_1^b) E(t) T_2^a + T_1^b E(t) (T_2^a - T_2^b) \\ + (T_2^a - T_2^b) E(t) T_1^a + T_2^b E(t) (T_1^a - T_1^b) \end{pmatrix} \right\| \\
 &\leq 2\mu \|E(t)\| (\|T_1^a - T_1^b\| + \|T_2^a - T_2^b\|) \leq 4\mu \|E(t)\| \|(T_1^a - T_1^b), (T_2^a - T_2^b)\|.
 \end{aligned}$$

Taking the max of (*) and (**) we obtain the desired local Lipschitz property in $R^{n^2} \times R^{n^2}$:

$$\begin{aligned}
 & \left\| \begin{pmatrix} (T_1^a E(t) T_1^a + T_2^a E(t) T_2^a - F_s(t)) \\ - (T_1^b E(t) T_1^b + T_2^b E(t) T_2^b - F_s(t)) \end{pmatrix}, \begin{pmatrix} (T_1^a E(t) T_2^a + T_2^a E(t) T_1^a - F_a(t)) \\ - (T_1^b E(t) T_2^b + T_2^b E(t) T_1^b - F_a(t)) \end{pmatrix} \right\| \\
 &\leq 8\mu \|E(t)\| \|(T_1^a - T_1^b), (T_2^a - T_2^b)\|.
 \end{aligned}$$

Thus, a unique solution $[T_1(t), T_2(t)]$ exists on some interval $(\hat{t}_a, t_1] \subset [t_0, t_1]$ to the system \mathcal{S} . (See Dieudonné [18].) Define

$$t_a = \inf \{ \hat{t}_a \in [t_0, t_1] : T_1(t), T_2(t) \text{ exist on } (\hat{t}_a, t_1] \}.$$

Observe that $[T_1'(t), -T_2'(t)]$ also is a solution to \mathcal{S} on $(t_a, t_1]$. Also note that

$$T(t) = T_1(t) + T_2(t)$$

is a solution to the NRS on $(t_a, t_1]$. Hence, by the uniqueness of solutions for the NRS and \mathcal{S} we have

$$T_s(t) = T_1(t)$$

and

$$T_a(t) = T_2(t),$$

where the subscripts “s” and “a” are as defined above. Thus, as before, it suffices to obtain bounds on $T_1(t)$ and $T_2(t)$ for $t \in (t_a, t_1]$ corresponding to the above bounds on $T_s(t)$ and $T_a(t)$, respectively, to complete the proof.

For any $X \in E^n$, compute (note that $x' \dot{T}_2'(t) T_2(t) x = x' T_2'(t) \dot{T}_2(t) x$)

$$\begin{aligned}
 \|T_2(t)x\| \frac{d}{dt} \|T_2(t)x\| &= x' T_2'(t) \dot{T}_2(t) x \\
 &= x' G(t)x - x' T_2'(t) F_a(t)x \quad \text{a.e. on } (\hat{t}_a, t_1],
 \end{aligned}$$

where

$$G(t) = T_2'(t)[T_1(t)E(t)T_2(t) + T_2(t)E(t)T_1(t)].$$

It is claimed that for the *second order* case $G(t)$ is positive semidefinite. To see this note that since $T_1(t) = T_s(t)$, Theorem 4 implies that $T_1(t)$ is positive semidefinite and symmetric, and write the most general 2nd order positive semidefinite, symmetric matrix for $T_1(t)$ as

$$T_1(t) = \begin{bmatrix} \alpha & \beta \\ \beta & \gamma \end{bmatrix} \begin{bmatrix} \alpha & \beta \\ \beta & \gamma \end{bmatrix} = \begin{bmatrix} (\alpha^2 + \beta^2) & \beta(\alpha + \gamma) \\ \beta(\alpha + \gamma) & (\beta^2 + \gamma^2) \end{bmatrix}$$

for any scalars α, β and γ . Similarly, we write the most general form for $E(t)$ (which is symmetric and positive semidefinite by assumption)

$$E(t) = \begin{bmatrix} (\lambda^2 + \rho^2) & \rho(\lambda + \omega) \\ \rho(\lambda + \omega) & (\rho^2 + \omega^2) \end{bmatrix}$$

for any scalars λ, ρ and ω . The most general skew symmetric matrix $T_2(t)$ can be represented as

$$T_2(t) = \begin{bmatrix} 0 & \sigma \\ -\sigma & 0 \end{bmatrix}$$

for any scalar σ . After tedious but routine computation we find that

$$G(t) = \begin{bmatrix} \sigma^2 q & 0 \\ 0 & \sigma^2 q \end{bmatrix},$$

where

$$q = (\beta\rho + \gamma\omega)^2 + (\alpha\rho + \beta\omega)^2 + (\alpha\lambda + \rho\beta)^2 + (\beta\lambda + \rho\gamma)^2.$$

Hence, $G(t)$ is positive semidefinite (and symmetric) for each $t \in (t_a, t_1]$. Thus,

$$\|T_2(t)x\| \frac{d}{dt} \|T_2(t)x\| \geq -x'T'_2(t)F_a(t)x \geq -\|T_2(t)x\| \|F_a(t)x\| \quad \text{a.e. on } (t_a, t_1].$$

Then, by Lemma 7,

$$\|T_2(t)\| \leq \|\tilde{Q}_a\| + \int_t^{t_1} \|F_a(\tau)\| d\tau \quad \text{for all } t \in (t_a, t_1].$$

Now, since $T_1(t)$ is positive semidefinite and symmetric,

$$\|T_1(t)\| = \sup_{\|x\|=1} x'T_1(t)x$$

and

$$\begin{aligned} x'\dot{T}_1(t)x &= x'T'_1(t)E(t)T_1(t)x - x'T'_2(t)E(t)T_2(t)x - x'F_s(t)x \\ &\geq -[\|T_2(t)\|^2 \|E(t)\| + \|F_s(t)\|] \|x\|^2 \end{aligned}$$

since $E(t)$ is positive semidefinite. Thus, integrating and taking the sup for $\|x\| = 1$ we have

$$\|T_1(t)\| \leq \|\tilde{Q}_s\| + \int_t^{t_1} [\|T_2(\tau)\|^2 \|E(\tau)\| + \|F_s(\tau)\|] d\tau.$$

Substituting the bound on $\|T_2(t)\|$ we have

$$\begin{aligned} \|T_1(t)\| &\leq \|\tilde{Q}_s\| + \int_t^{t_1} \|F_s(\tau)\| d\tau + \int_t^{t_1} \left[\|\tilde{Q}_a\| + \int_\tau^{t_1} \|F_a(\hat{\tau})\| d\hat{\tau} \right]^2 \|E(\tau)\| d\tau \\ &\leq \|\tilde{Q}_s\| + \int_t^{t_1} \|F_s(\tau)\| d\tau + \left[\|\tilde{Q}_a\| + \int_t^{t_1} \|F_a(\tau)\| d\tau \right]^2 \int_t^{t_1} \|E(\tau)\| d\tau, \end{aligned}$$

which completes the proof. Q.E.D.

Unfortunately, the above argument does not generalize even to the third order case. To see this we have only to let (for some $t \in [t_0, t_1]$)

$$T_s(t) = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 2 \end{bmatrix},$$

$$E(t) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad T_a(t) = \begin{bmatrix} 0 & 4 & 0 \\ -4 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix}.$$

Then

$$G(t) = T'_a(t)[T_s(t)E(t)T_a(t) + T_a(t)E(t)T_s(t)]$$

$$= \begin{bmatrix} 12 & 0 & 8 \\ 0 & 10 & 0 \\ -3 & 0 & -2 \end{bmatrix},$$

which is clearly not positive semidefinite. It should be emphasized, however, that this example does *not* imply that existence of the solution for higher order systems fails, in general, but simply that the above proof breaks down for $n \geq 3$.

The next theorem and three corollaries verify the existence of the solution $T(t)$ under various assumptions on the differential properties of $E(t)$.

THEOREM 7. *Let $E(t) = \rho(t)\bar{E}(t)$, where $\rho(t) \geq 0$ is an integrable scalar function and $\bar{E}(t)$ is symmetric, positive semidefinite and C^1 with $\dot{\bar{E}}(t)$ positive semidefinite for all $t \in [t_0, t_1]$. Then there exists a unique solution $T(t)$ to the NRS for all $t \in [t_0, t_1]$. Furthermore, this solution satisfies the bound*

$$(79) \quad \|T(t)\| \leq \left(\|\bar{Q}\| + \int_t^{t_1} \|F(\tau)\| d\tau \right) + \left(\|\bar{Q}\| + \int_t^{t_1} \|F(\tau)\| d\tau \right)^2 \int_t^{t_1} \rho(\tau) \|E(t_1)\| d\tau \quad \text{for all } t \in [t_0, t_1].$$

Proof. The local Lipschitz property of Lemma 6 establishes the existence of a unique solution $T(t)$ on some interval $(\hat{t}_a, t_1] \subset [t_0, t_1]$. Define

$$t_a = \inf \{ \hat{t}_a \in [t_0, t_1] : T(t) \text{ exists on } (\hat{t}_a, t_1] \}.$$

Then, as before, it suffices to obtain the above bound on $T(t)$ for $t \in (t_a, t_1]$ to complete the proof.

Since $\bar{E}(t)$ is positive semidefinite and symmetric on $[t_0, t_1]$, there is a symmetric, positive semidefinite matrix $\bar{E}_{1/2}(t)$ such that

$$\bar{E}(t) = \bar{E}_{1/2}(t)\bar{E}_{1/2}(t) \quad \text{for all } t \in [t_0, t_1].$$

For any $x \in R^n$, compute

$$\begin{aligned} \|\bar{E}_{1/2}(t)T(t)x\| \frac{d}{dt} \|\bar{E}_{1/2}(t)T(t)x\| &= \frac{1}{2} \frac{d}{dt} [\|\bar{E}_{1/2}(t)T(t)x\|^2] \\ &= x'T'(t)\bar{E}(t)\dot{T}(t)x + \frac{1}{2}x'T'(t)\dot{\bar{E}}(t)T(t)x \\ &= \rho(t)x'T'(t)\bar{E}(t)T(t)\bar{E}(t)T(t)x \\ &\quad - x'T'(t)\bar{E}(t)F(t)x + \frac{1}{2}x'T'(t)\dot{\bar{E}}(t)T(t)x \\ &\quad \text{for almost all } t \in (t_a, t_1]. \end{aligned}$$

But since $\bar{E}(t)$ is symmetric and $\dot{\bar{E}}(t)$ is positive semidefinite by assumption and $T(t)$ is positive semidefinite from Theorem 4, we obtain the inequality

$$\begin{aligned} \|\bar{E}_{1/2}(t)T(t)x\| \frac{d}{dt} \|\bar{E}_{1/2}(t)T(t)x\| &\geq -x'T'(t)\bar{E}(t)F(t)x \\ &\geq -\|\bar{E}_{1/2}(t)T(t)x\| \|\bar{E}_{1/2}(t)F(t)x\| \quad \text{a.e. on } (t_a, t_1]. \end{aligned}$$

Then, by Lemma 7 we have

$$\|\bar{E}_{1/2}(t)T(t)\| \leq \|\bar{E}_{1/2}(t_1)\| \|\tilde{Q}\| + \int_t^{t_1} \|\bar{E}_{1/2}(\tau)\| \|F(\tau)\| d\tau \quad \text{for all } t \in (t_a, t_1].$$

(Note that the above integral exists since $F(\cdot)$ is integrable and $\bar{E}_{1/2}(\cdot)$ is bounded since $\bar{E}(\cdot)$ is C^1 .) Furthermore, since $\bar{E}(t)$ is positive semidefinite and symmetric its norm is given by

$$\|\bar{E}(t)\| = \sup_{\|x\|=1} x'\bar{E}(t)x.$$

But for any $x \in R^n$,

$$x'\dot{\bar{E}}(t)x \geq 0 \quad \text{for all } t \in [t_0, t_1].$$

Hence,

$$\|\bar{E}(t)\| \leq \|\bar{E}(t_1)\| \quad \text{for all } t \in [t_0, t_1].$$

However,

$$\|\bar{E}_{1/2}(t)\| = \|\bar{E}(t)\|^{1/2}.$$

Thus,

$$(*) \quad \|\bar{E}_{1/2}(t)T(t)\| \leq \|\bar{E}(t_1)\|^{1/2} \left[\|\tilde{Q}\| + \int_t^{t_1} \|F(\tau)\| d\tau \right] \quad \text{for all } t \in (t_a, t_1].$$

In a similar fashion we obtain

$$(**) \quad \|\bar{E}_{1/2}(t)T'(t)\| \leq \|\bar{E}(t_1)\|^{1/2} \left[\|\tilde{Q}\| + \int_t^{t_1} \|F(\tau)\| d\tau \right] \quad \text{for all } t \in (t_a, t_1].$$

These bounds would suffice to guarantee existence if it were not for the possibility that $\bar{E}(t)$ (and, thus, also $\bar{E}_{1/2}(t)$) might be singular. To circumvent this difficulty

we compute for any $x \in R^n$,

$$\begin{aligned} \|T(t)x\| \frac{d}{dt} \|T(t)x\| &= x' T'(t) \dot{T}(t)x \\ &= \rho(t)x' T'(t) T(t) \bar{E}(t) T(t)x - x' T'(t) F(t)x \\ &\geq -\|T(t)x\| [\rho(t) \|\bar{E}_{1/2}(t) T'(t)\| \|\bar{E}_{1/2}(t) T(t)\| + \|F(t)\|] \|x\|. \end{aligned}$$

Again, by Lemma 7,

$$\|T(t)\| \leq \|\tilde{Q}\| + \int_t^{t_1} \|F(\tau)\| d\tau + \int_t^{t_1} \rho(\tau) \|\bar{E}_{1/2}(\tau) T'(\tau)\| \|\bar{E}_{1/2}(\tau) T(\tau)\| d\tau.$$

Applying the bounds (*) and (**) we have

$$\begin{aligned} \|T(t)\| &\leq \|\tilde{Q}\| + \int_t^{t_1} \|F(\tau)\| d\tau \\ &\quad + \int_t^{t_1} \rho(\tau) \|\bar{E}(t_1)\| \left[\|\tilde{Q}\| + \int_t^{t_1} \|F(\tau)\| d\tau \right]^2 d\tau \\ &\leq \left(\|\tilde{Q}\| + \int_t^{t_1} \|F(\tau)\| d\tau \right) \\ &\quad \cdot \left[1 + \left(\|\tilde{Q}\| + \int_t^{t_1} \|F(\tau)\| d\tau \right) \int_t^{t_1} \rho(\tau) \|\bar{E}(t_1)\| d\tau \right] \quad \text{for all } t \in (t_a, t_1]. \end{aligned}$$

Thus, the desired bound has been established and the proof is complete. Q.E.D.

As an obvious but important consequence of Theorem 7 we present the following conclusion.

COROLLARY 1. *If $E(t)$ is symmetric and piecewise constant with only a finite number of discontinuities, then on $[t_0, t_1]$ a unique solution exists to the NRS.*

Proof. For each of the intervals (τ_i, τ_{i+1}) on which $E(t)$ is constant, choose $\rho(t) \equiv 1$ and $\bar{E}(t) \equiv E(t)$. Then if the solution exists at τ_{i+1} , Theorem 7 asserts that it exists on $[\tau_i, \tau_{i+2}]$. But the solution exists at t_1 ($T(t_1) = \tilde{Q}$). Thus, existence can be extended by induction over any finite number of adjacent intervals of constant E working down from T_1 . Since only a finite number of such steps are required to reach t_0 , existence is guaranteed for the entire interval $[t_0, t_1]$. Q.E.D.

The curious thing about Corollary 1 is that *no* restrictions were necessary concerning the total variation of $E(t)$ (other than that it be finite) to guarantee existence. One might think that this result would be sufficient to obtain general existence for the case where $E(t)$ is C^1 by making a piecewise constant approximation to $E(t)$. However, the situation is deceptive. For any piecewise constant $\bar{E}(t)$ and any $\varepsilon > 0$, there is a $\delta(\varepsilon) > 0$ such that the corresponding solution $\bar{T}(t)$ will be modified by no more than some ε provided $\bar{E}(t)$ is modified by no more than $\delta(\varepsilon)$ on $[t_0, t_1]$. (See Coddington and Levinson [19].) However, by taking more steps and obtaining a closer approximation to the desired $E(t)$, the number $\delta(\varepsilon)$ may decrease so that no matter how close $E(t)$ is to $\bar{E}(t)$ it is not within $\delta(\varepsilon)$. This problem stems, of course, from the fact that we are choosing as the nominal trajectory the approximating trajectory rather than the desired trajectory. But

this is required since the “ ε, δ ” approximation only holds if the nominal trajectory exists on $[t_0, t_1]$. Hence, such an argument leads nowhere.

Let us next consider a somewhat less obvious consequence of Theorem 7.

COROLLARY 2. *If $E(t)$ can be represented as*

$$(80) \quad E(t) = \hat{\rho}(t)\hat{E}(t),$$

where $\hat{\rho}(t) \geq 0$ is an integrable scalar function and $\hat{E}(t)$ is symmetric, positive definite and C^1 , then the conditions of Theorem 7 are satisfied with

$$(81) \quad \rho(t) = \hat{\rho}(t) \exp [\mu(t_1 - t)]$$

and

$$(82) \quad \bar{E}(t) = \exp [-\mu(t_1 - t)]\hat{E}(t),$$

where

$$(83) \quad \mu = \left\{ \frac{\sup_{\substack{\|x\|=1 \\ t \in [t_0, t_1]}} |x' \dot{\hat{E}}(t)x|}{\inf_{\substack{\|x\|=1 \\ t \in [t_0, t_1]}} x' \hat{E}(t)x} \right\}.$$

Proof. First, note that $\rho(t)\bar{E}(t) = \hat{\rho}(t)\hat{E}(t)$. Thus, it suffices to show that $\hat{\rho}(t)$ and $\hat{E}(t)$ satisfy the condition of Theorem 7. Since $|x' \dot{\hat{E}}(t)x|$ is continuous in both x and t ($\hat{E}(\cdot)$ is C^1), then the $\sup |x' \dot{\hat{E}}(t)x|$ exists on the compact set of $R^n \times R^n$ given by $\|x\| = 1$ and $t \in [t_0, t_1]$. Similarly, the $\inf \| \hat{E}(t)x \|$ taken over this same set is realized for some x . Furthermore, since $\hat{E}(t)$ is positive definite then

$$x' \hat{E}(t)x > 0 \quad \text{if } \|x\| = 1 \text{ and } t \in [t_0, t_1].$$

Hence the $\inf \| \hat{E}(t)x \|$ taken on the above compact set *cannot* be zero. Thus, we conclude that $0 \leq \mu < \infty$ is well-defined. It suffices to show that $\bar{E}(t)$ is positive semidefinite where $\bar{E}(t)$ is as above.

For any $x \in R^n$ and $t \in [t_0, t_1]$, compute

$$x' \dot{\bar{E}}(t)x = x' [\mu \hat{E}(t) + \dot{\hat{E}}(t)]x \exp [-\mu(t_1 - t)] \geq 0$$

by the definition of μ . Thus, the representation

$$E(t) = \rho(t)\bar{E}(t)$$

is valid and $\bar{E}(t)$ is positive semidefinite, symmetric and C^1 with $\dot{\bar{E}}(t)$ also positive semidefinite. Hence, Theorem 7 applies to this case. Q.E.D.

As a final existence result which depends upon conditions on the differential properties of $E(\cdot)$ we present the following theorem.

THEOREM 8. *Let $E(t)$ be symmetric, C^1 , and of constant rank $r < n$ for all $t \in [t_0, t_1]$. Let $P(t)$ be a C^1 , nonsingular matrix function on $[t_0, t_1]$ which is a Dolezal transformation for the matrix $E(t)$. In other words,*

$$P'(t)E(t)P(t) = \begin{bmatrix} E_r(t) & 0 \\ 0 & 0 \end{bmatrix} \quad \text{for all } t \in [t_0, t_1],$$

where $E_r(t)$ is an $r \times r$, full rank matrix for all $t \in [t_0, t_1]$. Define

$$D(t) = P'(t) \frac{d}{dt} [P'^{-1}(t)]$$

and partition $D(t)$ as

$$D(t) = \left[\begin{array}{c|c} D_{11}(t) & D_{12}(t) \\ \hline D_{21}(t) & D_{22}(t) \end{array} \right],$$

where $D_{11}(t)$ is an $r \times r$ matrix. If the $(n-r) \times r$ matrix $D_{21}(t) = 0$ for all $t \in [t_0, t_1]$, then the NRS has a unique solution for all $t \in [t_0, t_1]$.

Proof. From Corollary 4 of Weiss and Falb [20], since $E(t)$ is symmetric and C^1 , then a matrix $P(t)$ exists as defined above. (Note that it is well known that a $P(t)$ exists for each $t \in [t_0, t_1]$ which transforms $E(t)$ as above but we need to turn to Dolezal's theorem to verify that, in fact, the matrix function $P(t)$ can be chosen to have the same smoothness properties as $E(t)$, in this case C^1 , so that it can be differentiated.)

Consider the system (*):

$$\dot{T}^*(t) = T^*(t) \left[\begin{array}{c|c} E_r(t) & 0 \\ \hline 0 & 0 \end{array} \right] T^*(t) + D'(t)T^*(t) + T^*(t)D(t) - F^*(t),$$

where

$$(*) \quad F^*(t) = P^{-1}(t)F(t)P'^{-1}(t)$$

and where $D(t)$ and $E_r(t)$ are as above. Establish the terminal condition

$$T^*(t_1) = P^{-1}(t_1)\tilde{Q}P'^{-1}(t_1) = Q^*.$$

Note that $F^*(t)$ and Q^* are positive semidefinite since $F(t)$ and \tilde{Q} are for the NRS. If we can show that a unique solution $T^*(t)$ exists to this system, then

$$T(t) = P(t)T^*(t)P'(t)$$

is a solution to the NRS. (Furthermore, by the local Lipschitz property of Lemma 6, the solution to the NRS is unique if it exists.) Thus, it suffices to consider the system (*). Partition $T^*(t)$, Q^* and $F^*(t)$ in the same way as $D(t)$. Then

$$\begin{aligned} \dot{T}_{11}^*(t) &= T_{11}^*(t)E_r(t)T_{11}^*(t) + T_{11}^*(t)D_{11}(t) + D_{11}'(t)T_{11}^*(t) \\ &\quad + T_{12}^*(t)D_{21}(t) + D_{21}'(t)T_{21}^*(t) - F_{11}^*(t), \end{aligned}$$

$$T_{11}^*(t_1) = Q_{11}^*,$$

$$\begin{aligned} \dot{T}_{12}^*(t) &= T_{11}^*(t)E_r(t)T_{12}^*(t) + T_{11}^*(t)D_{12}(t) + T_{12}^*(t)D_{22}(t) \\ &\quad + D_{11}'(t)T_{21}^*(t) + D_{21}'(t)T_{22}^*(t) - F_{12}^*(t), \end{aligned}$$

$$T_{12}^*(t_1) = Q_{12}^*,$$

$$\begin{aligned} \dot{T}_{21}^*(t) &= T_{21}^*(t)E_r(t)T_{11}^*(t) + T_{21}^*(t)D_{11}(t) + T_{22}^*(t)D_{21}(t) \\ &\quad + D_{12}'(t)T_{11}^*(t) + D_{22}'(t)T_{21}^*(t) - F_{21}^*(t), \end{aligned}$$

$$T_{21}^*(t_1) = Q_{21}^*$$

and

$$\begin{aligned}\dot{T}_{22}^*(t) &= T_{21}^*(t)E_r(t)T_{12}^*(t) + T_{21}^*(t)D_{12}(t) + T_{22}^*(t)D_{22}(t) \\ &\quad + D'_{12}(t)T_{12}^*(t) + D'_{22}(t)T_{22}^*(t) - F_{22}^*(t), \\ T_{22}^*(t_1) &= Q_{22}^*.\end{aligned}$$

Note that with the assumption that

$$D_{21}(t) \equiv 0,$$

however, we have

$$\begin{aligned}\dot{T}_{11}^*(t) &= T_{11}^*(t)E_r(t)T_{11}^*(t) + T_{11}^*(t)D_{11}(t) + D'_{11}(t)T_{11}^* - F_{11}^*, \\ T_{11}^*(t_1) &= Q_{11}^*.\end{aligned}$$

This is precisely of the standard form for an r -dimensional NRS (see (64)) and so, by Lemma 5 and Corollary 2, a unique solution $T_{11}^*(t)$ exists on $[t_0, t_1]$. In addition, since $D_{21}(t) \equiv 0$, note that the T_{12}^* and T_{21}^* systems are simply linear in themselves (dependent on T_{11}^* though) and so their unique solution exists. Finally, the system for T_{22}^* is linear in itself (but dependent on $T_{12}^*(t)$ and $T_{21}^*(t)$) and so a unique solution $T_{22}^*(t)$ exists. Thus, under the assumption that $D_{21}(t) \equiv 0$, we have proved that a unique solution $T^*(t)$ exists for (*) and, thus, that a unique solution $T(t)$ exists for the NRS. Q.E.D.

It should be remarked that the condition that $E(t)$ be of constant rank is relatively mild when applied to the QCP. From equations (29), (64) and (68) and Definition 9, we identify for the QCP

$$E(t) = \Phi(t_1, t)B(t)W^{-1}(t)B'(t)\Phi'(t_1, t).$$

Since $W^{-1}(t)$ is positive *definite* and symmetric and the transition matrix $\Phi(t, \tau)$ is always nonsingular, we see that $E(t)$ is of constant rank if and only if $B(t)B'(t)$ is of constant rank. Furthermore, this constant rank requirement could always be relaxed to piecewise constancy. Thus, nearly all problems of interest would satisfy this condition. However, the condition that $D_{21}(t) \equiv 0$ is not easily traced back to corresponding conditions on the parameters of the QCP and seriously weakens Theorem 8.

We depart, now, from cases which place restrictions on the differential properties of $E(\cdot)$ and, instead, consider the case where the skew symmetric parts of both $E(t)$ and $F(t)$ are zero. In this case the only channel through which skew symmetries can enter the solution is through the terminal condition $T(t_1) = \tilde{Q}$.

THEOREM 9. *If both $E(t)$ and $F(t)$ are symmetric, then there is a unique solution $T(t)$ to the NRS. If, in addition, $F(t) = 0$ for all $t \in [t_0, t_1]$, then the solution satisfies the bound*

$$(84) \quad \|T(t)\| \leq \|\tilde{Q}\| \left[1 + \|\tilde{Q}\| \int_t^{t_1} \|E(\tau)\| d\tau \right] \quad \text{for all } t \in [t_0, t_1].$$

Proof. Proving the last part first, suppose that $E(t)$ is symmetric and $F(t) = 0$ for all $t \in [t_0, t_1]$. As before, let $(t_a, t_1]$ be the interval of existence of the

unique solution $T(t)$ for the NRS. Then we need only verify that the above bound holds on $(t_a, t_1]$ to complete the proof for the case that $F(t) \equiv 0$.

As for Theorem 4, define $\Phi_{ET}(t, \tau)$ by

$$\frac{d}{dt}\Phi_{ET}(t, \tau) = -E(t)T(t)\Phi_{ET}(t, \tau),$$

$$\Phi_{ET}(\tau, \tau) = (\text{the identity matrix}) \quad \text{for all } t, \tau \in (t_a, t_1].$$

As argued previously, $\Phi_{ET}(t, \tau)$ and $\Phi_{ET}^{-1}(t, \tau) = \Phi_{ET}(\tau, t)$ exist for all $t, \tau \in (t_a, t_1]$. Since $F(t) \equiv 0$ we compute

$$\frac{d}{dt}[T(t)\Phi_{ET}(t, t_1)] = 0 \quad \text{for all } t \in (t_a, t_1].$$

Thus,

$$T(t)\Phi_{ET}(t, t_1) = \tilde{Q} \quad \text{on } (t_a, t_1]$$

and so

$$(*) \quad T(t) = \tilde{Q}\Phi_{ET}^{-1}(t, t_1).$$

Also we have

$$\dot{\Phi}_{ET}(t, t_1) = -E(t)T(t)\Phi_{ET}(t, t_1) = -E(t)\tilde{Q}.$$

Thus

$$(**) \quad \Phi_{ET}(t, t_1) = (\text{the identity matrix}) + \int_t^{t_1} E(\tau) d\tau \tilde{Q}.$$

Let $\hat{t} \in (t_a, t_1)$ be arbitrary. If we replace, throughout, the matrix function $E(t)$ by the *constant* matrix

$$\frac{1}{(t_1 - \hat{t})} \int_{\hat{t}}^{t_1} E(\tau) d\tau,$$

then by (**) the transition matrix $\Phi_{ET}(\hat{t}, \tau)$ remains unchanged. Thus, by (*) the solution $T(t)$ *evaluated at* $t = \hat{t}$ is also *unchanged*. Applying Theorem 7, we thus have the bound

$$\begin{aligned} \|T(\hat{t})\| &\leq \|\tilde{Q}\| \left[1 + \|\tilde{Q}\| \left\| \int_{\hat{t}}^{t_1} E(\tau) d\tau \right\| \right] \\ &\leq \|\tilde{Q}\| \left[1 + \|\tilde{Q}\| \int_{\hat{t}}^{t_1} \|E(\tau)\| d\tau \right]. \end{aligned}$$

But this bound holds for any $\hat{t} \in (t_a, t_1]$, and so existence and the required bound have been established for the case $F(t) \equiv 0$.

Next, consider the case where both $E(t)$ and $F(t)$ are symmetric for all $t \in [t_0, t_1]$. Define the absolutely continuous matrix $T_1(t)$ by the system

$$\dot{T}_1(t) = T_1(t)E(t)T_1(t) - F(t) \quad \text{a.e. on } [t_0, t_1]$$

with

$$T_1(t_1) = \frac{1}{2}(\tilde{Q} + \tilde{Q}').$$

By Theorem 5, a unique *symmetric* solution $T_1(t)$ exists for the above system on $[t_0, t_1]$. Define the transition matrix $\Phi_{ET_1}(t, t_1)$ by

$$\dot{\Phi}_{ET_1}(t, \tau) = -E(t)T_1(t)\Phi_{ET_1}(t, \tau)$$

and

$$\Phi_{ET_1}(\tau, \tau) = (\text{the identity matrix}).$$

Also define the positive semidefinite, symmetric and integrable matrix function

$$E^*(t) = \Phi_{ET_1}(t_1, t)E(t)\Phi_{ET_1}'(t_1, t).$$

Then as just proved, the system

$$\dot{T}_2(t) = T_2(t)E^*(t)T_2(t),$$

with

$$T_2(t_1) = \frac{1}{2}(\tilde{Q} - \tilde{Q}'),$$

has a unique solution on $[t_0, t_1]$. But, by direct calculation (recall that $T_1(t)$ is symmetric),

$$T(t) = T_1(t) + \Phi_{ET_1}'(t_1, t)T_2(t)\Phi_{ET_1}(t_1, t)$$

is a solution to the NRS and must be the unique solution (by the local Lipschitz property of Lemma 5). Hence, there is a unique solution to the NRS if both $E(t)$ and $F(t)$ are symmetric. Q.E.D.

Unfortunately, each of the various existence proofs presented in this section suffers some type of limitation. Hence, it has not yet been possible to conclusively demonstrate that finite escape cannot occur for the general NRS under the condition that $E(t)$ be symmetric. Curiously, though, there does *not* appear to be a central difficulty common to each of the proofs given. For example, the possibility of a nonsymmetric driving function $F(t)$ presents no problem in the proof of Theorem 7, but the time-varying behavior of $E(t)$ must be restricted there. On the other hand, the time variations of $E(t)$ present no difficulty, but the presence of a nonsymmetric driving function $F(t)$ cannot be accommodated in the proof of Theorem 9. For this reason, the authors feel that at the very least, existence can be proved for more general cases than given here and very possibly for the general NRS with $E(t)$ symmetric. However, based upon the cases for which existence has thus far been established, we summarize in the next theorem the corresponding conditions on the parameters of the recoverable QCP for which the solution to the system of equations (29) and (32) can be guaranteed to exist.

THEOREM 10. *A unique solution $K(t)$ exists on $[t_0, t_1]$ to the Riccati system of equations (29) and (32) for the recoverable QCP if one or more of the following conditions is met:*

- (i) *The dimension of the state $n = 1$.*
- (ii) *The dimension of the state $n = 2$.*

(iii) The vector function $y(\cdot)$ given by (10) can be written as

$$y(t) = \rho(t)s(t) \quad \text{for all } t \in [t_0, t_1]$$

for some scalar function $\rho(\cdot)$.

(iv) $B(t)$ is nonsingular for all $t \in [t_0, t_1]$.

(v) The matrix function

$$\dot{\bar{B}}(t) - A(t)\bar{B}(t) - \bar{B}(t)A'(t)$$

is positive semidefinite for all $t \in [t_0, t_1]$.

Proof. Condition (i) is trivial since there are no skew symmetries if $n = 1$.

Hence, Theorem 5 applies.

Condition (ii) follows directly from Theorem 6.

Condition (ii) implies that

$$\hat{R}(t) = R(t) \quad \text{for all } t \in [t_0, t_1].$$

Thus, since $R(t)$ and Q_1 are symmetric, Theorem 5 applies.

Condition (iv) follows from Corollary 2.

Condition (v) implies that (see (68) and (70))

$$\dot{E}(t) = \Phi(t_1, t)(\dot{\bar{B}}(t) - A(t)\bar{B}(t) - \bar{B}(t)A'(t))\Phi'(t_1, t)$$

is positive semidefinite and so Theorem 7 applies. Q.E.D.

8. Conclusion. We have considered a special class of state constrained optimal control problems, the linear state quadratic cost problem with a linear state constraint. By use of the results of Neustadt and Gilbert and Funk, a number of important properties of the solution to this problem are obtained including the important conclusion that the costate jumps normally associated with a state constrained problem can occur only at the initial and final times for the QCP. Using a nonsymmetric Riccati equation a simple set-valued fixed-point problem has been obtained, the solution of which defines an admissible optimal arc for the QCP. This representation suggests a simple computational scheme which is applied to two examples for their direct solution. The fixed-point representation obtained for this class of problems thus eliminates the need to resort to approximate methods such as penalty functions for a solution.

The results obtained for the QCP are dependent upon the existence of a solution to a nonsymmetric Riccati system. A completely satisfactory general solution to this problem has, unfortunately, not yet been found. What has been done is to develop existence theorems for a number of special cases. In addition to providing existence for a number of important cases, these, together with the examples given, illustrate the difficulties to be encountered with the more general cases.

In summary, a new viewpoint has been given to the linear quadratic cost, linear state constrained optimal control problem, and the significance of the nonsymmetric Riccati shown. It is hoped that this will stimulate further work in these areas.

REFERENCES

- [1] L. S. PONTRYAGIN, R. V. BOLTYANSKI, R. V. GAMKRELIDZE AND E. F. MISCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.
- [2] R. V. GAMKRELIDZE, *Optimal control processes with restricted phase coordinates*, Izv. Akad. Nauk SSSR Ser. Mat., 24 (1960), pp. 315–356.
- [3] L. D. BERKOVITZ, *On control problems with bounded state variables*, J. Math. Anal. Appl., 5 (1962), pp. 488–498.
- [4] H. J. KELLEY, *Method of gradients*, Optimization Techniques, Academic Press, New York, 1962, Chap. 6.
- [5] L. S. LASDON, A. D. WARREN AND R. K. RICE, *An interior penalty method for inequality constrained optimal control problems*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 388–395.
- [6] D. L. RUSSELL, *Penalty functions and bounded phase coordinate control*, this Journal, 2 (1965), pp. 409–422.
- [7] K. C. KAPUR AND R. M. VAN SLYKE, *Cutting plane algorithms and state space constrained linear optimal control problems*, J. Comput. System Sci., submitted.
- [8] J. L. SPEYER, R. K. MEHRA AND A. E. BRYSON, JR., *The separate computation of arcs for optimal flight paths with state variable inequality constraints*, Tech. Rep. 526, Div. of Eng. and Appl. Physics, Harvard Univ., Cambridge, Mass., 1967.
- [9] L. W. NEUSTADT, *A general theory of extrema's*, J. Comput. System Sci., 4 (1969), pp. 57–92.
- [10] E. G. GILBERT AND J. E. FUNK, *Some sufficient conditions for optimality in control problems*, this Journal, 8 (1970), pp. 498–504.
- [11] R. E. KALMAN, *Contributions to the theory of optimal control*, Bul. Soc. Mat. Mex., 5 (1960), pp. 102–119.
- [12] W. F. DENHAM AND A. E. BRYSON, JR., *Optimal programming with inequality constraints II: Solution by steepest ascent*, AIAA J., 2 (1964), pp. 25–34.
- [13] DAVID D. THOMPSON, *Optimal control of a linear system with quadratic cost under a linear half-space state constraint*, Ph.D. thesis, Univ. of Michigan, Ann Arbor, 1970.
- [14] J. E. POTTER, *A matrix equation arising in statistical filter theory*, NASA Contractor Rep. CR-270, 1965.
- [15] D. L. KLEINMAN, *On the linear regulator problem and the matrix Riccati equation*, MIT Elect. Systems Lab., Rep. ESL-R-271, Cambridge, Mass., 1966.
- [16] D. L. KLEINMAN AND P. L. FALB, *Remarks on the infinite dimensional Riccati equation*, IEEE Trans. Automatic Control, AC 11 (1966), pp. 534–536.
- [17] W. M. WONHAM, *On a matrix Riccati equation stochastic control*, this Journal, 6 (1968), pp. 681–697.
- [18] J. DIEUDONNÉ, *Foundations of Modern Analysis*, Academic Press, New York, 1960.
- [19] E. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.
- [20] L. WEISS AND P. L. FALB, *Dolezal's theorem, linear algebra with continuously parameterized elements and time varying systems*, Math. Systems Theory, 3 (1969), pp. 67–75.

NONCONVEX OPTIMIZATION PROBLEMS DEPENDING ON A PARAMETER*

J. BARANGER† AND R. TEMAM‡

Abstract. This paper deals with infinite-dimensional nonconvex optimization problems (or nonconvex control problems), which usually admit no solutions. The problem is perturbed by adding to the cost functional certain expressions depending on a parameter. The main result is that for "almost" all values of the parameter the optimization problem possesses at least one solution. After deriving the general result with the use of convex analysis and the properties of normed spaces, we present an example of a control problem for a system governed by a partial differential equation with boundary conditions of the Dirichlet type.

Introduction. The difficulties of nonconvex optimization are well known; very often a nonconvex continuous bounded function does not attain its minimum; or it can happen that this minimum point exists, but that minimizing sequences do not necessarily converge to this minimum. This makes the study of nonconvex optimization problems very difficult when compactness tools are not available.

Nevertheless some interesting results have been obtained for families of nonconvex optimization problems depending on a parameter: in a simplified form, they assert that in some cases, for almost all values of a parameter (i.e., for a dense set of values of the parameter), the optimization problem does possess a solution. The first result of this type concerning farthest points seems to be that of Edelstein [8]: let X be a uniformly convex Banach space, and S a (not necessarily convex) closed bounded set in X ; then the points of S which are the farthest from some point y are solutions of the maximization problem ($\|\cdot\|$ = norm in X)

$$(A) \quad \sup_{x \in S} \|y - x\|.$$

Edelstein proved in [8] that a solution of (A) exists for all the y 's of a dense subset of X . This result has been extended in several ways (briefly described below) by Asplund [1], [3], Baranger [5], [6], Bidaut [7], Edelstein [8], and Zizler [18].

Our goal here is to establish a similar result for a family of optimization problems of the type:

$$(B) \quad \sup_x \{\omega(\|x - y\|) - f(x)\}.$$

The hypotheses on X , f (a function from X into $\mathbb{R} \cup \{+\infty\}$) and ω (a function from \mathbb{R}_+ into \mathbb{R}_+) are precisely given in the text. The main result is given in Theorem 1; an application and further remarks are given in § 2.

* Received by the editors August 24, 1973, and in revised form October 10, 1973.

† Département de Mathématiques, Université de Lyon I, Lyon, France.

‡ Département de Mathématiques, Université de Paris XI, Orsay, France.

1. The main result.

1.1. Definitions and notations.

We recall here a few definitions and notations.
(a) A Banach space X is called a strongly differentiable space (SDS space) if the following property holds (see Asplund [2], [3]):

any convex continuous function f from X into $\mathbb{R} \cup \{+\infty\}$ is Fréchet differentiable in a dense G_δ subset of its domain of continuity, $\text{dom } f$, where $\text{dom } f = \{x \in X | f(x) < \infty\}$.

(b) A strictly convex or rotund normed space (R space) is characterized by the property that $\|x + y\| = 2$ and $\|x\| = \|y\| = 1$ imply $x = y$.

A locally uniformly rotund normed space (LUR space) is a space satisfying the following condition (see Lovaglia [12]):

If $\|x_n + x\| \rightarrow 2$ as $n \rightarrow \infty$, and $\|x_n\| = \|x\| = 1$ then $\|x_n - x\| \rightarrow 0$.

(c) If φ is a function from $\mathbb{R}_+ = [0, \infty)$ into $\overline{\mathbb{R}}_+ = [0, \infty]$, one can define its conjugate function φ^* by setting

$$\varphi^*(t) = \sup_{s \geq 0} \{t \cdot s - \varphi(s)\} \quad \text{for all } t \geq 0.$$

Let Γ denote the class of convex and lower semicontinuous (l.s.c.) functions from \mathbb{R}_+ into $\overline{\mathbb{R}}_+$ such that $\varphi(0) = 0$. Then if φ belongs to Γ , φ^* belongs to Γ too and $\varphi^{**} = \varphi$. Put

$$\Gamma_U = \{\varphi \in \Gamma, \varphi(t) > 0 \text{ for } t > 0\},$$

$$\Gamma_L = \{\varphi \in \Gamma, \lim_{t \rightarrow 0} \varphi(t)/t = 0\}.$$

According to a result of Asplund [3], φ belongs to Γ_L if and only if φ^* belongs to Γ_U .

1.2. Statement of the result. Let X be a reflexive Banach space satisfying the following property:

(H) If a sequence x_n converges weakly to x and $\|x_n\|$ converges to $\|x\|$, then $\|x_n - x\| \rightarrow 0$.

Some examples of (H)-spaces are briefly given in § 2.

We consider a l.s.c. function f from X into $\mathbb{R} \cup \{+\infty\}$, which is not identically equal to $+\infty$ and a scalar function ω from \mathbb{R}_+ into \mathbb{R} which is convex, continuous, and strictly increasing. We are interested in the family of maximization problems:

$$(1.1) \quad \sup_{x \in X} \{\omega(\|x - y\|) - f(x)\}.$$

The supremum in (1.1) is denoted $f_\omega(y)$, and we assume that

$$(1.2) \quad f_\omega(y) < +\infty \quad \text{for each } y \in X,$$

and

$$(1.3) \quad \text{every maximizing sequence of problem (1.1) is bounded.}^1$$

¹ This is true as usual, if, for each fixed y :

$$-\omega(\|x - y\|) + f(x) \rightarrow +\infty \quad \text{as } \|x\| \rightarrow \infty, \quad x \in \text{dom } f.$$

Property (1.3) guarantees the existence of weakly convergent maximizing sequences. In the nonconvex case, by lack of a weak semicontinuity property of the functional, the limit of a weakly maximizing sequence is not necessarily a solution of the maximization problem. For this reason the usual arguments in the calculus of variations (the direct method in particular) do not allow us to show the existence of solutions of a problem like (1.1).

Using completely different methods, we will prove the following result (our main result).

THEOREM 1. *Under the preceding assumptions, in particular the condition (H) on X , (1.2) and (1.3), there exists a dense G_δ subset of X such that for each y in this set, problem (1.1) possesses a solution:*

$$(1.4) \quad \text{There exists } \bar{x} \in X \text{ such that } f_\omega(y) = \omega(\|\bar{x} - y\|) - f(\bar{x}).$$

This theorem is proved in § 1.3. It implies the following.

COROLLARY 1. *Let X be a reflexive Banach space satisfying the (H)-property and let S be a closed bounded subset of X . We denote by J an u.s.c. function from S into \mathbb{R} bounded from above, and by ω a convex continuous strictly increasing function from \mathbb{R}_+ into \mathbb{R} . Then there exists a dense G_δ subset of X , say G , such that for each $y \in G$, there exists an $\bar{x} \in S$ with*

$$(1.5) \quad J(\bar{x}) + \omega(\|\bar{x} - y\|) = \sup_{x \in S} \{J(x) + \omega(\|x - y\|)\}.$$

Proof. We set $f(x) = -J(x)$, $x \in S$, and $f(x) = +\infty$, otherwise, and we apply Theorem 1; (1.2) and (1.3) are clearly satisfied (see preceding footnote 1).

1.3. Proof of Theorem 1.

(i) With (1.2) and since $\text{dom } f \neq \emptyset$, the function $f_\omega: y \rightarrow f_\omega(y)$ is defined from X into \mathbb{R} ; this function is convex l.s.c. as an upper bound of such functions. According to a result of Rockafellar [15] and since X is barrelled, the function f_ω is continuous on X . This implies that f_ω is everywhere subdifferentiable.²

On the other hand, by a result of Trojanski [17], any reflexive Banach space is an SD space; since f_ω is continuous, there exists a G_δ dense subset of X , say G , on which f_ω is Fréchet differentiable. We will prove that (1.4) holds for each y in this G , at least.

Asplund gives in [2], [3] a characterization of Fréchet differentiability of convex functions: a convex function $g: X \rightarrow \mathbb{R}$ is Fréchet differentiable at the point ζ with differential $\eta \in X'$ if and only if there exists a function φ in Γ_L such that

$$g(y) - g(\zeta) - \langle \eta, y - \zeta \rangle \leq \varphi(\|y - \zeta\|) \quad \text{for all } y \in X.$$

Thus in particular, for each $\zeta \in G$, f_ω is Fréchet differentiable at ζ with differential η , and there exist $\eta \in X'$ and $\varphi \in \Gamma_L$ such that

$$(1.6) \quad 0 \leq f_\omega(y) - f_\omega(\zeta) - \langle \eta, y - \zeta \rangle \leq \varphi(\|y - \zeta\|) \quad \text{for all } y \in X.$$

Since the function $y \rightarrow \omega(\|y - x\|)$ is convex and continuous, it is subdifferentiable everywhere and in particular at the point ζ . Let t be some element of this

² For results related to convex analysis, the reader is referred to Moreau [13], Rockafellar [15]; see also Laurent [10] and Ekeland and Temam [9].

subdifferential :

$$(1.7) \quad 0 \leq \omega(\|x - y\|) - \omega(\|x - \zeta\|) - \langle t, y - \zeta \rangle.$$

It follows from (1.6) and (1.7) that, for each $x \in X$,

$$\begin{aligned} \omega(\|x - \zeta\|) + \langle t - \eta, y - \zeta \rangle - f(x) - f_\omega(\zeta) \\ \leq f_\omega(y) - f_\omega(\zeta) - \langle \eta, y - \zeta \rangle \leq \varphi(\|y - \zeta\|) \end{aligned}$$

and hence

$$\omega(\|x - \zeta\|) - f_\omega(\zeta) + \langle t - \eta, y - \zeta \rangle - \varphi(\|y - \zeta\|) \leq f(x) \quad \text{for all } x, y \in X.$$

Taking the supremum with respect to y of the left-hand side, we obtain

$$(1.8) \quad \omega(\|x - \zeta\|) - f_\omega(\zeta) + \varphi^*(\|t - \eta\|_*) \leq f(x) \quad \text{for all } x \in X,$$

where $\varphi^* \in \Gamma_U$ (see § 1.1), and (1.9) is thus established.

(ii) Let x_n be a maximizing sequence of problem (1.1) where $y = \zeta$:

$$\omega(\|x_n - \zeta\|) - f(x_n) \rightarrow f_\omega(\zeta) \quad \text{as } n \rightarrow \infty.$$

For each n , let $t_n \in \partial\theta_{x_n}(\zeta)$. Relation (1.8) gives

$$(1.9) \quad \varphi^*(\|t_n - \eta\|_*) \leq f(x_n) + f_\omega(\zeta) - \omega(\|x_n - \zeta\|) \rightarrow 0.$$

Since $\varphi^*(\|t_n - \eta\|) \geq 0$ we infer from (1.9) that $\varphi(\|t_n - \eta\|_*) \rightarrow 0$ as $n \rightarrow \infty$, and since $\varphi^* \in \Gamma_U$, this implies

$$(1.10) \quad \|t_n - \eta\|_* \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Besides that, (1.3) implies that x_n is a bounded sequence and X being a reflexive space, one can extract from x_n a subsequence (still denoted x_n) such that

$$(1.11) \quad x_n \rightarrow \bar{x} \quad \text{in } X \text{ weakly as } n \rightarrow \infty.$$

We will finally prove the following two properties:

$$(1.12) \quad \|x_n - \bar{x}\| \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

$$(1.13) \quad \bar{x} \text{ is a solution of problem (1.1) with } y = \zeta.$$

To establish (1.12) we write, using (1.7),

$$(1.14) \quad \omega(\|\bar{x} - \zeta\|) \geq \omega(\|x_n - \zeta\|) + \langle t_n, x_n - \bar{x} \rangle.$$

According to (1.10), (1.11), $\langle t_n, x_n - \bar{x} \rangle \rightarrow 0$, as $n \rightarrow \infty$. Passing to the limit and taking into account the monotonicity of ω , we find

$$\begin{aligned} \omega(\|\bar{x} - \zeta\|) &\geq \overline{\lim_{n \rightarrow \infty}} \omega(\|x_n - \zeta\|) \\ &\geq \omega(\lim_{n \rightarrow \infty} \|x_n - \zeta\|) \\ &\geq \omega(\lim_{n \rightarrow \infty} \|x_n - \zeta\|) \\ &\geq \omega(\|\bar{x} - \zeta\|). \end{aligned}$$

Therefore,

$$\|x_n - \zeta\| \rightarrow \|\bar{x} - \zeta\|$$

and, because of the (H)-property,

$$\|x_n - \bar{x}\| \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

It remains to show that \bar{x} is a solution of (1.1) for $y = \zeta$. But

$$\begin{aligned} f_\omega(\zeta) &= \lim_{n \rightarrow \infty} \{\omega(\|x_n - \zeta\|) - f(x_n)\} \\ &\leq \omega(\|\bar{x} - \zeta\|) - \lim_{n \rightarrow \infty} f(x_n) \quad (\text{because of (1.14)}) \\ &\leq \omega(\|\bar{x} - \zeta\|) - f(\bar{x}) \quad (\text{by the l.s.c. of } f), \end{aligned}$$

and finally,

$$f_\omega(\zeta) = \omega(\|\bar{x} - \zeta\|) - f(\bar{x}).$$

The proof of Theorem 1 is completed.

2. Complements and applications.

2.1. Miscellaneous remarks.

(i) Theorem 1 and Corollary 1 extend some results of Edelstein [8] (where $J \equiv 0$ and X is uniformly convex), Asplund [1] (where $J \equiv 0$ and X is a LUR reflexive Banach space), Baranger [5], [6] (J as in Corollary 1, X is a uniformly convex Banach space and $\omega(\|x - y\|) = \|x - y\|$), Bidaut [7] (J and X as before, $\omega(\|x - y\|) = \|x - y\|^s$, $1 \leq s < \infty$), and Zizler [18] ($J \equiv 0$, $\omega(\|x - y\|) = \|x - y\|$, X reflexive Fréchet differentiable and satisfying the (H)-condition).

The relations between the different hypotheses on J are clear. Concerning the hypotheses on the space, we recall that a uniformly convex space is reflexive and LUR; it is also easy to see that a reflexive LUR Banach space satisfies the (H)-property: Indeed, let $x_n \rightarrow x$ be a weakly convergent sequence in a reflexive LUR Banach space, and let us assume that $\|x_n\| \rightarrow \|x\|$. If $x = 0$, x_n converges strongly to 0; if $x \neq 0$, set $x'_n = x_n/\|x_n\|$, $x' = x/\|x\|$. Then $\|x'_n\| = \|x'\| = 1$. There exists f in X^* such that $f(x) = \|f\| = 1$, and

$$f(x'_n) + f(x') \leq \|x'_n + x'\| \leq 2.$$

Thus $\|x'_n + x'\| \rightarrow 2$, and the LUR property implies $\|x'_n - x'\| \rightarrow 0$: finally $\|x_n - x\| \rightarrow 0$.

Conversely a reflexive Banach space satisfying the (H)-condition could even be nonrotund (take \mathbb{R}^2 equipped with the norm $\|x\| = |x_1| + |x_2|$).

(ii) Some of the previous results can be extended to more general families of nonconvex problems of type

$$(2.1) \quad \sup_{x \in X} g(x, y),$$

where X is some set, and y a parameter belonging to an SD Banach space.

We assume that the supremum in (2.1) is finite for each y , and that the function

$$g_x : y \rightarrow g(x, y)$$

is convex and l.s.c. for each x . Denoting by $f(y)$ the supremum in (2.1), we see that the function $y \rightarrow f(y)$ is convex continuous and thus Fréchet differentiable on a dense G_δ subset of Y . Let ζ belong to this set, $\eta = f'(\zeta)$; if x_n is a maximizing sequence of problem (2.1) with $y = \zeta$, and $t_n \in \partial g_{x_n}(\zeta)$, then

$$(2.2) \quad t_n \rightarrow \eta \quad \text{as } n \rightarrow \infty.$$

2.2. An application to optimal control. We now describe briefly an application of this type of result to nonconvex optimal control problems; a more detailed description of the following example, and many other examples are given in Baranger [6]; see also Bidaut [7].

Let Ω be an open bounded set in \mathbb{R}^n . For each given function $u \in L^2(\Omega)$, with

$$(2.3) \quad 0 < \alpha \leq u(x) \leq \beta \quad \text{a.e.,}$$

there exists a unique function $y = y(u)$ in $H^1(\Omega)$ (space of functions in $L^2(\Omega)$ with derivatives in $L^2(\Omega)$) such that

$$(2.4) \quad \sum_{i=1}^n \frac{\partial}{\partial x_i} \left(u \frac{\partial y}{\partial x_i} \right) = f \quad (f \in L^2(\Omega), \text{ given}),$$

$$(2.5) \quad y = 0 \quad \text{on } \partial\Omega.$$

We are interested in the following control problem: for a given y_0 in $H^1(\Omega)$, find a measurable function u satisfying (2.3) which minimizes

$$(2.6) \quad \|y(u) - y_d\|_{H^1(\Omega)}.$$

This problem has no solution in general; see for instance Lions [11], Murat [14].

Let ω be as in Theorem 1. For each $\varepsilon > 0$ and $v \in L^2(\Omega)$, we consider the following perturbation of problem (2.6):

$$(2.7) \quad \text{minimize } (\|y(u) - y_d\|_{H^1(\Omega)} - \varepsilon\omega(\|u - v\|_{L^2(\Omega)})).$$

Theorem 1 asserts that for all $\varepsilon > 0$ and for all the v 's of a dense G_δ subset of $L^2(\Omega)$, the minimum in problem (2.7) is attained by a function satisfying (2.3).

Acknowledgments. The authors wish to thank R. T. Rockafellar for a simplification of the proof he suggested to us.

REFERENCES

- [1] E. ASPLUND, *Farthest points in reflexive locally uniformly rotund Banach spaces*, Israel J. Math., 4 (1966), pp. 213–216.
- [2] ———, *Fréchet differentiability of convex functions*, Acta Math., 121 (1968), pp. 31–47.
- [3] ———, *Topics in the theory of convex functions*, Theory and Applications of Monotone Operators, Aldo Ghizetti, ed., Edizioni "Oderisi". (Proceedings of NATO, Venice, June 1968.)
- [4] ———, *Positivity of duality mappings*, Bull. Amer. Math. Soc., 73 (1967), pp. 200–203.
- [5] J. BARANGER, *Norm perturbation of supremum problems*, 5th IFIP Conference on optimization techniques, Rome, June 1973. Lectures Notes in Computer Science, Springer-Verlag, New York, 1973.
- [6] ———, *Quelques résultats en optimisation non convexe*, Thesis, University of Grenoble, France, 1973, and J. Math. Pures Appl., to appear.

- [7] M. F. BIDAUT, *Théorèmes d'existence et d'existence en général d'un contrôle optimal pour des systèmes régis par des équations aux dérivées partielles non linéaires*, Thesis, University of Paris, 1973.
- [8] M. EDELSTEIN, *Farthest points of sets in uniformly convex Banach spaces*, Israel J. Math., 4 (1966), no. 3, pp. 171–176.
- [9] I. EKELAND AND R. TEMAM, *Analyse convexe et problèmes variationnels*, Dunod, Paris, 1974.
- [10] P. J. LAURENT, *Approximation et optimisation*, Hermann, Paris, 1972.
- [11] J. L. LIONS, *Contrôle optimal de système gouvernés par des équations aux dérivées partielles*, Dunod, Paris, 1969; English transl., Springer-Verlag, New York, 1972.
- [12] A. R. LOVAGLIA, *Locally uniformly convex Banach spaces*, Trans. Amer. Math. Soc., 78 (1955), pp. 225–238.
- [13] J. J. MOREAU, *Fonctionnelles convexes*, Séminaire sur les Equations aux Dérivées Partielles, Collège de France, Paris, 1966–67.
- [14] F. MURAT, *Un contre exemple pour le problème du contrôle dans les coefficients*, C.R. Acad. Sci. Ser. A Paris, 273 (1971), pp. 708–711.
- [15] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, 1970.
- [16] V. L. SMULIAN, *Sur la dérivabilité de la norme dans l'espace de Banach*, Dokl. Akad. Nauk SSSR, 27 (1940), pp. 643–648.
- [17] S. L. TROYANSKI, *On locally uniformly convex and differentiable norms in certain non separable Banach spaces*, Studia Math., 37 (1971), pp. 173–180.
- [18] V. ZIZLER, *On some extremal problems in Banach spaces*, Math. Scand., 32 (1973).

HILBERT NETWORKS. II: SOME QUALITATIVE PROPERTIES*

VACLAV DOLEZAL AND ARMEN ZEMANIAN†

Abstract. The concept of a Hilbert network, introduced in another paper, extends network theory to finite or infinite networks whose elements are described by operators on a Hilbert space. The present work investigates a variety of qualitative properties possessed by such networks. In particular, some operators associated with the entire network, such as the driving-point impedances as well as the admittance operator which relates the branch-voltage vector to the branch-current vector, are shown to be—under suitable conditions—either positive, monotonic, or convex. Also, generalized versions of Jeans' least power theorem and the Shannon–Hagelbarger theorem are proved. Finally, a bound on the power dissipation is determined.

1. Introduction. The present paper develops certain qualitative properties of the Hilbert networks introduced in [1]. Actually, some analogous results can also be established for the alternative approach to infinite networks of operators developed in the series of papers [2]–[5]. The ideas and even the manipulations needed for these results are quite similar in both approaches. However, the formalisms and notations are very different. Therefore, to avoid repetition, we shall present our development only in the formalism of [1].

As our first result, we give necessary and sufficient conditions for the power dissipated in the network to be positive or incrementally positive. This amounts to conditions for the positivity or monotonicity of the admittance operator of the network. Next, we will show that, under certain conditions, the admittance operator of a linear network is a convex nonincreasing function of the impedance operator of the network.

As a further result, we will discuss a generalization of Jeans' least power theorem [6, p. 322]. Using this and a theorem on the existence of a driving-point impedance, we will show that the driving point impedance of a linear network satisfying a certain positivity condition is a concave nondecreasing function of the impedance operator. This result generalizes the classical Shannon–Hagelbarger theorem [7], [8] valid for finite resistive networks. Finally, a bound is established on the power dissipation in the network.

2. Results. In this sequel, we will consistently use the terminology and notation introduced in [1].

Let φ be a fixed function from the complex plane into the real line having the property that $\varphi(\bar{\xi}) = \varphi(\xi)$ for every ξ ; furthermore, let \mathcal{H} be a fixed Hilbert space and let M be a (not necessarily linear) operator from \mathcal{H} into itself. We will write

- (a) $M \in \mathfrak{P}_\varphi(\mathcal{H})$ if $\varphi(\langle Mx, x \rangle) \geq 0$ for all $x \in \mathcal{H}$,
- (b) $M \in \mathfrak{M}_\varphi(\mathcal{H})$ if $\varphi(\langle Mx_1 - Mx_2, x_1 - x_2 \rangle) \geq 0$ for all $x_1, x_2 \in \mathcal{H}$.

For example, we can put $\varphi(\xi) = \operatorname{Re} \xi$, or $\varphi(\xi) = \alpha \operatorname{Re} \xi + \beta |\operatorname{Im} \xi|$, α, β real, etc.

* Received by the editors June 14, 1973, and in revised form October 9, 1973.

† Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, Stony Brook, N.Y. 11790. This research was supported by the National Science Foundation under Grant PO33568-XOO.

THEOREM 1. Let $\hat{\mathcal{N}} = (\hat{\mathcal{Z}}, G)$ be a regular (not necessarily linear) Hilbert network, and let A be the admittance operator of $\hat{\mathcal{N}}$; then

- (i) $A \in \mathfrak{P}_\varphi(H^{c_2}) \Leftrightarrow \hat{\mathcal{Z}} \in \mathfrak{P}_\varphi(N_{\hat{a}}) \Leftrightarrow W \in \mathfrak{P}_\varphi(H^{c_0})$,
(ii) $A \in \mathfrak{M}_\varphi(H^{c_2}) \Leftrightarrow \hat{\mathcal{Z}} \in \mathfrak{M}_\varphi(N_{\hat{a}}) \Leftrightarrow W \in \mathfrak{M}_\varphi(H^{c_0})$, where $W = \hat{X}^* \hat{\mathcal{Z}} \hat{X} : H^{c_0} \rightarrow H^{c_0}$.

Proof. Choose arbitrarily $e \in H^{c_2}$ and let $i = Ae$. By definition of a solution i of $\hat{\mathcal{N}}$ corresponding to e , we have $i \in N_{\hat{a}}$ and $\hat{\mathcal{Z}}i - e \in N_{\hat{a}}^\perp$; hence $\langle i, \hat{\mathcal{Z}}i - e \rangle_{c_2} = 0$, and consequently

$$(1) \quad \langle Ae, e \rangle_{c_2} = \overline{\langle \hat{\mathcal{Z}}i, i \rangle_{c_2}}.$$

Now, if $\hat{\mathcal{Z}} \in \mathfrak{P}_\varphi(N_{\hat{a}})$, then (1) shows that $A \in \mathfrak{P}_\varphi(H^{c_2})$.

Conversely, let $A \in \mathfrak{P}_\varphi(H^{c_2})$ and arbitrarily choose $i \in N_{\hat{a}}$. Putting $e = \hat{\mathcal{Z}}i \in H^{c_2}$, we have $\hat{\mathcal{Z}}i - e = 0 \in N_{\hat{a}}^\perp$, and (1) holds. Hence, $\hat{\mathcal{Z}} \in \mathfrak{P}_\varphi(H^{c_2})$.

Moreover, since by [1, Lemma 2.2], $N_{\hat{a}} = \hat{X}H^{c_0}$ and \hat{X} is one-to-one, then for each $i \in N_{\hat{a}}$ there exists a unique $x \in H^{c_0}$ with $i = \hat{X}x$ and vice versa. Thus we have

$$(2) \quad \langle \hat{\mathcal{Z}}i, i \rangle_{c_2} = \langle \hat{\mathcal{Z}}\hat{X}x, \hat{X}x \rangle_{c_2} = \langle \hat{X}^* \hat{\mathcal{Z}} \hat{X}x, x \rangle_{c_0} = \langle Wx, x \rangle_{c_0}.$$

This relation completes the proof of (i).

Next, choose $e_j \in H^{c_2}$, $j = 1, 2$, and let $i_j = Ae_j$; then

$$(3) \quad \begin{aligned} \langle Ae_1 - Ae_2, e_1 - e_2 \rangle_{c_2} &= \langle i_1 - i_2, e_1 - e_2 \rangle_{c_2} \\ &= \langle i_1, e_1 \rangle_{c_2} - \langle i_2, e_1 \rangle_{c_2} - \langle i_1, e_2 \rangle_{c_2} + \langle i_2, e_2 \rangle_{c_2}. \end{aligned}$$

However, since $\langle c, \hat{\mathcal{Z}}i_j \rangle_{c_2} = \langle c, e_j \rangle_{c_2}$ for every $c \in N_{\hat{a}}$, and $i_j \in N_{\hat{a}}$, we obtain the relations

$$\begin{aligned} \langle i_1, e_1 \rangle_{c_2} &= \langle i_1, \hat{\mathcal{Z}}i_1 \rangle_{c_2}, & \langle i_2, e_1 \rangle_{c_2} &= \langle i_2, \hat{\mathcal{Z}}i_1 \rangle_{c_2}, \\ \langle i_1, e_2 \rangle_{c_2} &= \langle i_1, \hat{\mathcal{Z}}i_2 \rangle_{c_2}, & \langle i_2, e_2 \rangle_{c_2} &= \langle i_2, \hat{\mathcal{Z}}i_2 \rangle_{c_2}. \end{aligned}$$

Introducing this into (3), we readily get

$$(4) \quad \langle Ae_1 - Ae_2, e_1 - e_2 \rangle_{c_2} = \overline{\langle \hat{\mathcal{Z}}i_1 - \hat{\mathcal{Z}}i_2, i_1 - i_2 \rangle_{c_2}}.$$

If $\hat{\mathcal{Z}} \in \mathfrak{M}_\varphi(N_{\hat{a}})$, (4) implies that $A \in \mathfrak{M}_\varphi(H^{c_2})$. Conversely, if $A \in \mathfrak{M}_\varphi(H^{c_2})$, then putting $e_j = \hat{\mathcal{Z}}i_j$ for chosen $i_j \in N_{\hat{a}}$, $j = 1, 2$, we conclude from (4) as before that $\hat{\mathcal{Z}} \in \mathfrak{M}_\varphi(N_{\hat{a}})$.

Finally, due to the one-to-one correspondence between $N_{\hat{a}}$ and H^{c_0} mentioned above, for any chosen $i_j \in N_{\hat{a}}$, $j = 1, 2$, there exists a unique $x_j \in H^{c_0}$ such that $i_j = \hat{X}x_j$. Hence

$$\begin{aligned} \langle \hat{\mathcal{Z}}i_1 - \hat{\mathcal{Z}}i_2, i_1 - i_2 \rangle_{c_2} &= \langle \hat{\mathcal{Z}}\hat{X}x_1 - \hat{\mathcal{Z}}\hat{X}x_2, \hat{X}(x_1 - x_2) \rangle_{c_2} \\ &= \langle Wx_1 - Wx_2, x_1 - x_2 \rangle_{c_0}, \end{aligned}$$

which concludes the proof.

COROLLARY 1. If A is the admittance operator of a regular network, then

- (i*) $\hat{\mathcal{Z}} \in \mathfrak{P}_\varphi(H^{c_2}) \Rightarrow A \in \mathfrak{P}_\varphi(H^{c_2})$,
(ii*) $\hat{\mathcal{Z}} \in \mathfrak{M}_\varphi(H^{c_2}) \Rightarrow A \in \mathfrak{M}_\varphi(H^{c_2})$.

(The proof follows immediately from Theorem 1 and the fact that $N_{\hat{a}} \subset H^{c_2}$.)

Obviously, Theorem 1 gives necessary and sufficient conditions for the power and incremental power dissipated in the network to be nonnegative. The following theorem gives bounds for the incremental power in a particular case of network.

THEOREM 2. Let $\hat{\mathcal{N}} = (\hat{Z}, G)$ be a Hilbert network, and let $W = \hat{X}^* \hat{Z} \hat{X} : H^{c_0} \rightarrow H^{c_0}$. Assume that there exist constants $\gamma, \lambda > 0$ such that

$$(i) \operatorname{Re} \langle Wx_1 - Wx_2, x_1 - x_2 \rangle_{c_0} \geq \gamma \|x_1 - x_2\|_{c_0}^2,$$

$$(ii) \|Wx_1 - Wx_2\|_{c_0} \leq \lambda \|x_1 - x_2\|_{c_0}$$

for all $x_1, x_2 \in H^{c_0}$. Then $\hat{\mathcal{N}}$ is regular and its admittance operator A satisfies the inequality

$$(5) \quad \gamma \lambda^{-2} \|X^*(x_1 - x_2)\|_{c_0}^2 \leq \operatorname{Re} \langle Ax_1 - Ax_2, x_1 - x_2 \rangle_{c_2} \leq \gamma^{-1} \|X^*(x_1 - x_2)\|_{c_0}^2$$

for all $x_1, x_2 \in H^{c_2}$.

Proof. Regularity of $\hat{\mathcal{N}}$ follows from [2, Thm. 2.4]. Also, the proof of Theorem 2.4 shows that (i), (ii) imply that

$$(6) \quad \operatorname{Re} \langle \hat{Z}y_1 - \hat{Z}y_2, y_1 - y_2 \rangle_{c_2} \geq \gamma \|y_1 - y_2\|_{c_2}^2,$$

$$(7) \quad \|P(\hat{Z}y_1 - \hat{Z}y_2)\|_{c_2} \leq \lambda \|y_1 - y_2\|_{c_2}$$

for all $y_1, y_2 \in N_{\hat{a}}$, where $P = \hat{X} \hat{X}^*$ is the orthogonal projection from H^{c_2} onto $N_{\hat{a}}$. However, since $Py = y$ for each $y \in N_{\hat{a}}$, (6) and (7) yield

$$(8) \quad \begin{aligned} \operatorname{Re} \langle \hat{Z}y_1 - \hat{Z}y_2, P(y_1 - y_2) \rangle_{c_2} &= \operatorname{Re} \langle P\hat{Z}y_1 - P\hat{Z}y_2, y_1 - y_2 \rangle_{c_2} \\ &\geq \gamma \|y_1 - y_2\|_{c_2}^2 \geq \gamma \lambda^{-2} \|P\hat{Z}y_1 - P\hat{Z}y_2\|_{c_2}^2. \end{aligned}$$

Choose now $x_j \in H^{c_2}$, $j = 1, 2$, and put $y_j = Ax_j = (P\hat{Z})^{-1}Px_j$ (see [1, Thm. 1.1]). Then, by (8) and the fact $Ax_j \in N_{\hat{a}}$,

$$\begin{aligned} &\operatorname{Re} \langle Px_1 - Px_2, Ax_1 - Ax_2 \rangle_{c_2} \\ &= \operatorname{Re} \langle x_1 - x_2, P(Ax_1 - Ax_2) \rangle_{c_2} \\ &= \operatorname{Re} \langle Ax_1 - Ax_2, x_1 - x_2 \rangle_{c_2} \geq \gamma \lambda^{-2} \|Px_1 - Px_2\|_{c_2}^2 \\ &= \gamma \lambda^{-2} \|\hat{X} \hat{X}^*(x_1 - x_2)\|_{c_2}^2 = \gamma \lambda^{-2} \|\hat{X}^*(x_1 - x_2)\|_{c_0}^2. \end{aligned}$$

(See [1, Lemma 2.2].) This proves the first part of (5).

Furthermore, from (6) we infer by Schwarz's inequality that

$$(9) \quad \gamma \|y_1 - y_2\|_{c_2} \leq \|P\hat{Z}y_1 - P\hat{Z}y_2\|_{c_2}$$

for all $y_1, y_2 \in N_{\hat{a}}$. Also by (6) and (9),

$$(10) \quad \begin{aligned} &\operatorname{Re} \langle \hat{Z}y_1 - \hat{Z}y_2, y_1 - y_2 \rangle_{c_2} \\ &\leq |\langle P\hat{Z}y_1 - P\hat{Z}y_2, y_1 - y_2 \rangle_{c_2}| \\ &\leq \|P\hat{Z}y_1 - P\hat{Z}y_2\|_{c_2} \cdot \|y_1 - y_2\|_{c_2} \leq \gamma^{-1} \|P\hat{Z}y_1 - P\hat{Z}y_2\|_{c_2}^2 \end{aligned}$$

for all $y_1, y_2 \in N_{\hat{a}}$. Using the above substitution $y_j = Ax_j$, we obtain from (10),

$$\operatorname{Re} \langle Ax_1 - Ax_2, x_1 - x_2 \rangle_{c_2} \leq \gamma^{-1} \|Px_1 - Px_2\|_{c_2}^2 = \gamma^{-1} \|\hat{X}^*(x_1 - x_2)\|_{c_0}^2,$$

hence the proof.

Remark 1. From (10) we see that we have proved somewhat more than (5), i.e.,

$$|\langle Ax_1 - Ax_2, x_1 - x_2 \rangle_{c_2}| \leq \gamma^{-1} \|\hat{X}^*(x_1 - x_2)\|_{c_0}^2.$$

Let \mathcal{H} be a Hilbert space, and let $M \in [\mathcal{H}, \mathcal{H}]$ be a self adjoint operator; M will be called *positive* if $\langle Mx, x \rangle$ is real and nonnegative for each $x \in \mathcal{H}$. In this case we will write $M \geq 0$. If $M_1 - M_2 \geq 0$, we will write $M_1 \geq M_2$ or $M_2 \leq M_1$.

Next, let $\mathcal{N} = (\hat{Z}, G)$ be a Hilbert network; if \mathcal{N} is regular, we will denote its admittance operator by $A(\hat{Z})$.

THEOREM 3. *Let G be a fixed oriented graph and let $\hat{Z}_j \in [H^{c_2}, H^{c_2}]$, $j = 1, 2$, be self-adjoint. For each $\lambda \in [0, 1]$, let $\mathcal{N}_\lambda = (\lambda\hat{Z}_1 + (1 - \lambda)\hat{Z}_2, G)$, and let $W_j = \hat{X}^* \hat{Z}_j \hat{X} : H^{c_0} \rightarrow H^{c_0}$, $j = 1, 2$.*

Assume that

$$(i) \quad W_1 W_2 = W_2 W_1,$$

(ii) there exists a $\gamma > 0$ such that $W_j - \gamma I \geq 0$ for $j = 1, 2$, where I is the identity operator on H^{c_0} .

Then each network \mathcal{N}_λ is regular, and

$$(11) \quad A(\lambda\hat{Z}_1 + (1 - \lambda)\hat{Z}_2) \leq \lambda A(\hat{Z}_1) + (1 - \lambda)A(\hat{Z}_2).$$

Moreover, if $\hat{Z}_1 \leq \hat{Z}_2$, then

$$(12) \quad A(\hat{Z}_1) \geq A(\hat{Z}_2).$$

Proof. From [1, Thm. 2.2] and (ii), it follows immediately that for every $\lambda \in [0, 1]$, \mathcal{N}_λ is regular. Moreover, [1, Lemma 1.1] and (ii) show that operators W_1^{-1} , W_2^{-1} , $(\lambda W_1 + (1 - \lambda)W_2)^{-1}$ exist, are in $[H^{c_0}, H^{c_0}]$, and clearly are positive and self-adjoint. Also note the fact that due to (i), any two operators from the collection $\{W_1, W_2, W_1^{-1}, W_2^{-1}, \lambda W_1 + (1 - \lambda)W_2, (\lambda W_1 + (1 - \lambda)W_2)^{-1}\}$ commute. Since $Q^*Q \geq 0$ for any operator Q and W_1, W_2 are self-adjoint, we have $(W_1 - W_2)^2 \geq 0$ and $(W_1 - W_2)^2$ is self-adjoint; hence, by commutativity,

$$(13) \quad W_1^2 + W_2^2 - 2W_1 W_2 \geq 0.$$

Moreover, since the product of two commuting positive self-adjoint operators is again positive and self-adjoint (see [9, p. 149]), (13) and the fact that $W_2^{-1} W_1^{-1} \geq 0$ yield

$$(14) \quad W_2^{-1} W_1 + W_1^{-1} W_2 - 2I \geq 0.$$

However, $\lambda(1 - \lambda) \geq 0$ for any $\lambda \in [0, 1]$; hence (14) implies that

$$(15) \quad \begin{aligned} & -2\lambda(1 - \lambda)I + \lambda(1 - \lambda)W_1^{-1}W_2 + \lambda(1 - \lambda)W_1W_2^{-1} \\ & = -I + \lambda[\lambda W_1 + (1 - \lambda)W_2]W_1^{-1} + (1 - \lambda)\lambda W_1 + (1 - \lambda)W_2]W_2^{-1} \geq 0. \end{aligned}$$

Thus, since $(\lambda W_1 + (1 - \lambda)W_2)^{-1} \geq 0$, we get from (15),

$$(16) \quad -(\lambda W_1 + (1 - \lambda)W_2)^{-1} + \lambda W_1^{-1} + (1 - \lambda)W_2^{-1} \geq 0.$$

On the other hand, using the definition of W_j , (16) can be written as

$$(17) \quad -[\hat{X}^*(\lambda\hat{Z}_1 + (1 - \lambda)\hat{Z}_2)\hat{X}]^{-1} + \lambda(\hat{X}^*\hat{Z}_1\hat{X})^{-1} + (1 - \lambda)(\hat{X}^*\hat{Z}_2\hat{X})^{-1} \geq 0.$$

However, by [1, Thm. 2.2], the admittance operator $A(\hat{Z})$ of any regular network $\mathcal{N} = (\hat{Z}, G)$ is given by $A(\hat{Z}) = \hat{X}(\hat{X}^*\hat{Z}\hat{X})^{-1}\hat{X}^*$; also, $Q \geq 0$ implies that $T^*QT \geq 0$ for any operator T . Hence, we infer from (17) that

$$-A(\lambda\hat{Z}_1 + (1 - \lambda)\hat{Z}_2) + \lambda A(\hat{Z}_1) + (1 - \lambda)A(\hat{Z}_2) \geq 0,$$

which proves (11).

Next, assume that $\hat{Z}_1 \leq \hat{Z}_2$; then $\hat{X}^* \hat{Z}_2 \hat{X} - \hat{X}^* \hat{Z}_1 \hat{X} = W_2 - W_1 \geq 0$. Hence, by commutativity and positivity of W_1^{-1} , W_2^{-1} , $W_1^{-1}(W_2 - W_1)W_2^{-1} = W_1^{-1} - W_2^{-1} \geq 0$. Thus, by the same argument as before, $\hat{X}W_1^{-1}\hat{X}^* - \hat{X}W_2^{-1}\hat{X}^* = A(Z_1) - A(Z_2) \geq 0$, which proves (12).

Remark 2. The assumption $\hat{Z}_1 \leq \hat{Z}_2$ in Theorem 3 may clearly be replaced by the weaker condition $W_1 \leq W_2$.

Let us now prepare the way for stating the least power theorem.

Let $\hat{\mathcal{N}} = (\hat{Z}, G)$ be a Hilbert network and \hat{a} be its structural operator; if $e \in H^{c_2}$, define

$$(18) \quad F_e = \{x : x \in N_{\hat{a}}, \operatorname{Re} \langle e, x \rangle_{c_2} = 0\}.$$

It is clear that F_e is a real linear subspace of $N_{\hat{a}}$, i.e., $\alpha x + \beta y \in F_e$ whenever $x, y \in F_e$ and α, β are real; also note that $\{e\}^\perp \cap N_{\hat{a}} \subset F_e$.

Then we have the following theorem.

THEOREM 4. *Let $\hat{\mathcal{N}} = (\hat{Z}, G)$ be a linear Hilbert network (i.e., \hat{Z} is linear but not necessarily bounded) such that*

- (i) $\operatorname{Re} \langle \hat{Z}x, y \rangle_{c_2} = \operatorname{Re} \langle x, \hat{Z}y \rangle_{c_2}$ for all $x, y \in H^{c_2}$,
- (ii) $\operatorname{Re} \langle \hat{Z}x, x \rangle_{c_2} > 0$ for all $x \in H^{c_2}$, $x \neq 0$.

Let $e \in H^{c_2}$ be such that $F_e \neq \{0\}$, and assume that $i \in H^{c_2}$ is a (not necessarily unique) solution of $\hat{\mathcal{N}}$ corresponding to e . Let $\mathcal{P} : i + F_e \rightarrow \mathbb{R}^1$ be defined by

$$(19) \quad \mathcal{P}(x) = \operatorname{Re}(\langle \hat{Z}x, x \rangle_{c_2} - \langle \hat{Z}i, i \rangle_{c_2}).$$

Then $\mathcal{P}(x) > 0$ for every $x \in i + F_e$, $x \neq i$.

Proof. We have $i \in N_{\hat{a}}$ and $\langle c, e \rangle_{c_2} = \langle c, \hat{Z}i \rangle_{c_2}$ for all $c \in N_{\hat{a}}$; thus, for any $i' \in F_e$, $\operatorname{Re} \langle i', e \rangle_{c_2} = 0 = \operatorname{Re} \langle i', \hat{Z}i \rangle_{c_2}$. Hence, by (i),

$$(20) \quad \operatorname{Re} \langle \hat{Z}i, i' \rangle_{c_2} = 0, \quad \operatorname{Re} \langle \hat{Z}i', i \rangle_{c_2} = 0.$$

Now, let $x \in i + F_e$, $x \neq i$; then $x = i + i'$, $i' \in F_e$, $i' \neq 0$, and we have by (20) and (ii),

$$\begin{aligned} \mathcal{P}(x) &= \operatorname{Re}(\langle \hat{Z}(i + i'), i + i' \rangle_{c_2} - \langle \hat{Z}i, i \rangle_{c_2}) \\ &= \operatorname{Re} \langle \hat{Z}i, i \rangle_{c_2} + \operatorname{Re} \langle \hat{Z}i', i \rangle_{c_2} + \operatorname{Re} \langle \hat{Z}i, i' \rangle_{c_2} + \operatorname{Re} \langle \hat{Z}i', i' \rangle_{c_2} - \operatorname{Re} \langle \hat{Z}i, i \rangle_{c_2} \\ &= \operatorname{Re} \langle \hat{Z}i', i' \rangle_{c_2} > 0, \end{aligned}$$

hence the proof.

Observe that assumptions (i), (ii) of Theorem 4 are satisfied if \hat{Z} is a self-adjoint strictly positive operator, i.e., $\langle \hat{Z}x, x \rangle$ is real and positive for all $x \in H^{c_2}$, $x \neq 0$.

For our next purpose it will be convenient to state Theorem 4 in a slightly weaker form. To this end, let us introduce the following concepts.

Let $s = \{n_i\}$ be a (finite or infinite) increasing sequence of positive integers. A c -vector $x = [x_k]$ will be said to have the s -pattern if $x_k = 0$ for all $k \notin s$; similarly, a c -vector $y = [y_k]$ has the s -copattern if $y_k = 0$ for all $k \in s$.

If N is a linear subspace of H^c , then N^{cs} will signify the linear subspace of all c -vectors in N which have the s -copattern.

COROLLARY 2. Let $\hat{\mathcal{N}} = (\hat{\mathcal{Z}}, G)$ be a linear Hilbert network, and let $\hat{\mathcal{Z}}$ be self-adjoint and strictly positive. Furthermore, let $e \in H^{c_2}$ have an s -pattern, let $N_a^{cs} \neq \{0\}$, and assume that $i \in H^{c_2}$ is a solution of $\hat{\mathcal{N}}$ corresponding to e . Then the function

$$(21) \quad \mathcal{P}(x) = \langle \hat{\mathcal{Z}}x, x \rangle_{c_2} - \langle \hat{\mathcal{Z}}i, i \rangle_{c_2}$$

is positive on $i + N_a^{cs}$ except for $x = i$.

For the proof it is sufficient to realize that $N_a^{cs} \subset F_e$.

Let us now discuss the existence of the driving-point impedance. To this end, let us introduce the following terminology.

Let G be an oriented graph with the set of branches $\{b_j\}$; a branch b_k will be called *regular* if the matrix $X = [\xi^i]$ contains at least one column $\xi^{i*} = [\xi_j^{i*}]$ such that $\xi_k^{i*} \neq 0$. Using the terminology of the graph theory, it is clear that a branch b_k is regular exactly if it is contained in at least one loop of G .

THEOREM 5. Let $\hat{\mathcal{N}} = (\hat{\mathcal{Z}}, G)$ be a linear Hilbert network such that $\hat{\mathcal{Z}} \in [H^{c_2}, H^{c_2}]$ and

$$(22) \quad |\langle Wx, x \rangle_{c_0}| \geq \gamma \|x\|_{c_0}^2$$

for some $\gamma > 0$ and all $x \in H^{c_0}$, where $W = \hat{X}^* \hat{\mathcal{Z}} \hat{X} : H^{c_0} \rightarrow H^{c_0}$. Furthermore, assume that the branch b_1 of G is regular. Then there exists an operator $R \in [H, H]$ with the following property: for any $j \in H$ there exists a unique $i = [i_k] \in H^{c_2}$, with $i_1 = j$, such that i is the solution of $\hat{\mathcal{N}}$ corresponding to $e = [Rj, 0, 0, \dots]^T \in H^{c_2}$.

Proof. First of all, from [1, Thm. 2.2] it follows that $\hat{\mathcal{N}}$ is regular and its admittance operator A is given by $A = \hat{X} Y \hat{X}^*$, where $Y = W^{-1}$. Let x_1^T, x_2^T, \dots be the rows of the matrix X (thus, each x_k is a c_0 -vector); due to our assumption on regularity of the branch b_1 , it follows that $x_1 \neq 0$.

Next, choose $f \in H$, put $e = [f, 0, 0, \dots]^T \in H^{c_2}$, and denote by $i = [i_k]$ the solution of $\hat{\mathcal{N}}$ corresponding to e . Using [1, Lemma 2.2], it follows that

$$[i_k] = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \end{bmatrix} \cdot Y \cdot [\bar{x}_1, \bar{x}_2, \dots] \cdot \begin{bmatrix} f \\ 0 \\ 0 \\ \vdots \end{bmatrix} = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \end{bmatrix} \cdot Y \cdot (\bar{x}_1 f).$$

Consequently, $i_1 = x_1^T \cdot Y \cdot (\bar{x}_1 f)$.

Now define the operator $B : H \rightarrow H$ by

$$(23) \quad Bv = x_1^T \cdot Y \cdot (\bar{x}_1 v).$$

We can easily verify that B is bounded. Indeed, from (22) it follows by [1, Lemma 1.1] that $W^{-1} = Y \in [H^{c_0}, H^{c_0}]$ and $\|Y\| \leq \gamma^{-1}$. Also, (22) implies readily that

$$(24) \quad |\langle Yu, u \rangle_{c_0}| \geq \gamma \|W\|^{-2} \|u\|_{c_0}^2$$

for each $u \in H^{c_0}$. Now, put $\eta = [1, 0, 0, \dots]^T \in l^2$; then $\bar{x}_1 = \bar{X}^T \cdot \eta = \tilde{X}^* \eta$, and consequently, by [1, Lemma 2.1], $\|\bar{x}_1\| = \|x_1\| \leq \|\tilde{X}^*\| \cdot \|\eta\| = 1$. Hence

$$\|Bv\|_1 \leq \|x_1^T\| \cdot \|Y\| \cdot \|\bar{x}_1 v\| \leq \gamma^{-1} \|v\|_1,$$

and the boundedness of B is proved.

On the other hand, we have according to (24),

$$(25) \quad \begin{aligned} |\langle Bv, v \rangle_1| &= |\langle v, x_1^T \cdot Y(\bar{x}_1 f) \rangle_1| = |\langle \bar{x}_1 v, Y(\bar{x}_1 v) \rangle_{c_0}| \\ &\geq \gamma \|W\|^{-2} \|\bar{x}_1 v\|_{c_0}^2 = \lambda \|v\|_1^2, \end{aligned}$$

where $\lambda = \gamma \|W\|^{-2} \|x_1\|^2 > 0$ since $x_1 \neq 0$.

Now invoking [1, Lemma 1.1] and (25), we obtain that B possesses a bounded inverse. Thus, letting $R = B^{-1}$, then for any $j \in H$ we can construct $e = [Rj, 0, 0, \dots]^T \in H^{c_2}$, and the corresponding solution $i = [i_k]$ of $\hat{\mathcal{N}}$ will have the property that $i_1 = BRj = j$, hence the proof.

In agreement with the terminology of classical network theory, the operator R mentioned in Theorem 5 will be called the *driving-point impedance of the branch* b_1 . (Obviously the fact that we considered b_1 instead of branch b_k is no loss of generality.)

Let us now discuss a generalization of the Shannon–Hagelbarger theorem [7], [8]. To this purpose, let us introduce the following notation.

Let \mathcal{G} be the set of all self-adjoint operators in $[H^{c_2}, H^{c_2}]$ having the property that for each $\hat{Z} \in \mathcal{G}$, $\hat{Z} - \gamma_z I \geq 0$ for some $\gamma_z > 0$.

Clearly, \mathcal{G} is a convex subset of $[H^{c_2}, H^{c_2}]$.

Furthermore, let G be an oriented graph with the set of branches $\{b_k\}$ having cardinality $c_2 \leq \aleph_0$, and assume that the branch b_1 is regular. Finally, we will denote by $R(\hat{Z})$ the driving-point impedance of b_1 (provided it exists) whenever the network (\hat{Z}, G) is under consideration. Then we have the following.

THEOREM 6. *Let G be an oriented graph with branch b_1 being regular, and let $R(\hat{Z})$ have the meaning defined above. Then, for each $\hat{Z} \in \mathcal{G}$, the network $\hat{\mathcal{N}} = (\hat{Z}, G)$ is regular, the driving point impedance $R(\hat{Z})$ exists, is a positive operator and is a concave nondecreasing function on \mathcal{G} , i.e.,*

$$(26) \quad R(\lambda \hat{Z}_1 + (1 - \lambda) \hat{Z}_2) \geq \lambda R(\hat{Z}_1) + (1 - \lambda) R(\hat{Z}_2)$$

for all $\lambda \in [0, 1]$ and any pair $\hat{Z}_1, \hat{Z}_2 \in \mathcal{G}$, and $R(\hat{Z}_1) \leq R(\hat{Z}_2)$ whenever $\hat{Z}_1 \leq \hat{Z}_2$, $\hat{Z}_1, \hat{Z}_2 \in \mathcal{G}$.

Proof. Recalling the definition of \mathcal{G} , we see that, for any $\hat{Z} \in \mathcal{G}$ and $x \in H^{c_0}$, $\langle Wx, x \rangle_{c_0} = \langle \hat{X}^* \hat{Z} \hat{X} x, x \rangle_{c_0} = \langle \hat{Z} \hat{X} x, \hat{X} x \rangle_{c_2} \geq \gamma_z \|\hat{X} x\|_{c_2}^2 = \gamma_z \|x\|_{c_0}^2$; hence by [1, Thm. 2.2], the network (\hat{Z}, G) is regular. Also, this inequality shows that assumptions of Theorem 5 are met, and consequently, $R(\hat{Z})$ exists and is in $[H, H]$.

Next, let $s = \{1\}$ and, as above, let N_a^{cs} be the space of all vectors $x = [x_k] \in N_a$ with $x_1 = 0$. Arbitrarily choose $\hat{Z}_1, \hat{Z}_2 \in \mathcal{G}$, $\lambda \in [0, 1]$, $j \in H$, and denote $e_\lambda = [R(\lambda \hat{Z}_1 + (1 - \lambda) \hat{Z}_2)j, 0, 0, 0, \dots]^T \in H^{c_2}$ and $e_m = [R(\hat{Z}_m)j, 0, 0, 0, \dots]^T \in H^{c_2}$ for $m = 1, 2$. Let $i^\lambda = [i_k^\lambda] \in H^{c_2}$ be the solution of $\hat{\mathcal{N}}_\lambda = (\lambda \hat{Z}_1 + (1 - \lambda) \hat{Z}_2, G)$ corresponding to e_λ and $i^m = [i_k^m] \in H^{c_2}$ the solution of $\hat{\mathcal{N}}_m = (\hat{Z}_m, G)$ corresponding to e_m , $m = 1, 2$. Thus, we have by the definition of the driving-point impedance, $i_1^\lambda = i_1^1 = i_1^2 = j$. Moreover,

$$(27) \quad \begin{aligned} \langle i^\lambda, e_\lambda \rangle_{c_2} &= \langle j, R(\lambda \hat{Z}_1 + (1 - \lambda) \hat{Z}_2)j \rangle_1 = \langle i^\lambda, (\lambda \hat{Z}_1 + (1 - \lambda) \hat{Z}_2) i^\lambda \rangle_{c_2} \\ &= \lambda \langle i^\lambda, \hat{Z}_1 i^\lambda \rangle_{c_2} + (1 - \lambda) \langle i^\lambda, \hat{Z}_2 i^\lambda \rangle_{c_2}. \end{aligned}$$

Also,

$$(28) \quad \langle i^m, e_m \rangle_{c_2} = \langle j, R(\hat{Z}_m)j \rangle_1 = \langle i^m, \hat{Z}_m i^m \rangle_{c_2}, \quad m = 1, 2.$$

Relation (28) shows that $R(\hat{Z})$ is a positive operator for every $\hat{Z} \in \mathcal{G}$. On the other hand, $i^\lambda - i^1 \in N_{\hat{a}}^{cs}$, i.e., $i^\lambda \in i^1 + N_{\hat{a}}^{cs}$; hence, by Corollary 2 (its assumptions are clearly met), $\langle i^\lambda, \hat{Z}_1 i^\lambda \rangle_{c_2} \geq \langle i^1, \hat{Z}_1 i^1 \rangle_{c_2}$. By the same argument, $i^\lambda \in i^2 + N_{\hat{a}}^{cs}$, and consequently, $\langle i^\lambda, \hat{Z}_2 i^\lambda \rangle_{c_2} \geq \langle i^2, \hat{Z}_2 i^2 \rangle_{c_2}$. Thus, using these inequalities and (27), (28), we obtain

$$\langle j, R(\lambda \hat{Z}_1 + (1 - \lambda) \hat{Z}_2) j \rangle_1 \geq \langle j, \lambda R(\hat{Z}_1) j \rangle_1 + \langle j, (1 - \lambda) R(\hat{Z}_2) j \rangle_1,$$

which proves (26).

Finally, assume that $\hat{Z}_1 \leq \hat{Z}_2$, $\hat{Z}_1, \hat{Z}_2 \in \mathcal{G}$. Choosing $j \in H$ and using the same notation as above, we have (28). Since $i^2 \in i^1 + N_{\hat{a}}^{cs}$, it follows that $\langle i^2, \hat{Z}_1 i^2 \rangle_{c_2} \geq \langle i^1, \hat{Z}_1 i^1 \rangle_{c_2}$, and consequently, $\langle i^2, \hat{Z}_2 i^2 \rangle_{c_2} \geq \langle i^1, \hat{Z}_1 i^1 \rangle_{c_2}$. Hence, by (28), $\langle j, R(\hat{Z}_2) j \rangle_1 \geq \langle j, R(\hat{Z}_1) j \rangle_1$ which concludes the proof.

As our final result, we establish a bound on the power dissipated in a regular linear Hilbert network (\hat{Z}, G) when it is driven by an emf vector $e \in H^{c_2}$. As we shall see, such a bound is $\langle \hat{Z}^{-1} e, e \rangle$ when \hat{Z} is invertible. Furthermore, a result of this nature can be obtained under a weaker condition than the invertibility of \hat{Z} .

As before, let $s = \{n_i\}$ be a finite or infinite increasing sequence of positive integers. H^s and H^{cs} will denote those subspaces of H^{c_2} consisting of all c_2 -vectors $[x_k] \in H^{c_2}$ having the s -pattern and the s -copattern, respectively. Thus, $H^{c_2} = H^s \oplus H^{cs}$. Also, \hat{Z}_s and \hat{Z}_{cs} will denote the restrictions of $\hat{Z} \in [H^{c_2}; H^{c_2}]$ to H^s and H^{cs} respectively.

THEOREM 7. *Let $\hat{\mathcal{N}} = (\hat{Z}, G)$ be a regular linear Hilbert network, and let \hat{Z} be positive. Let $e \in H^{c_2}$ have an s -pattern, and assume that H^s reduces \hat{Z} . Also, assume that \hat{Z}_s is invertible on H^s . If $i \in H^{c_2}$ is a solution of $\hat{\mathcal{N}}$ corresponding to e , then*

$$(29) \quad 0 \leq \langle i, \hat{Z} i \rangle_{c_2} \leq \langle \hat{Z}_s^{-1} e, e \rangle_{c_2}.$$

Proof. Since \hat{Z} is positive, so, too, are \hat{Z}_s and \hat{Z}_{cs} . Moreover, \hat{Z}_s possesses a positive square root $\hat{Z}_s^{1/2}$. Also, since $\hat{Z}_s^{1/2} \hat{Z}_s^{1/2} = \hat{Z}_s$ and \hat{Z}_s is invertible on H^s , $\hat{Z}_s^{-1/2}$ exists as a positive mapping of H^s onto H^s .

Now, $i = i_s + i_{cs}$, where $i_s \in H^s$ and $i_{cs} \in H^{cs}$.

Thus,

$$\langle i, e \rangle_{c_2} = \langle i_s, e \rangle_{c_2} = \langle \hat{Z}_s^{-1/2} \hat{Z}_s^{1/2} i_s, e \rangle_{c_2} = \langle \hat{Z}_s^{1/2} i_s, \hat{Z}_s^{-1/2} e \rangle_{c_2}.$$

By Schwarz's inequality,

$$(30) \quad |\langle i, e \rangle_{c_2}|^2 \leq \|\hat{Z}_s^{1/2} i_s\|_{c_2}^2 \|\hat{Z}_s^{-1/2} e\|_{c_2}^2 = \langle i_s, \hat{Z}_s i_s \rangle_{c_2} \langle \hat{Z}_s^{-1/2} e, e \rangle_{c_2}.$$

Since H^s reduces \hat{Z} , we have that

$$\langle i, \hat{Z} i \rangle_{c_2} = \langle i_s, \hat{Z}_s i_s \rangle_{c_2} + \langle i_{cs}, \hat{Z}_{cs} i_{cs} \rangle_{c_2}.$$

Since both terms on the right-hand side are nonnegative,

$$\langle i_s, \hat{Z}_s i_s \rangle_{c_2} \leq \langle i, \hat{Z} i \rangle_{c_2}.$$

Therefore, (30) yields

$$(31) \quad |\langle i, e \rangle_{c_2}|^2 \leq \langle i, \hat{Z} i \rangle_{c_2} \langle \hat{Z}_s^{-1} e, e \rangle_{c_2}.$$

Finally, $\langle i, e \rangle_{c_2} = \langle i, \hat{Z} i \rangle_{c_2} \geq 0$ because i is a solution corresponding to e . Therefore, (31) yields (29).

It is worth pointing out that the right-hand side of (29), which is a bound on the power dissipation $\langle i, \hat{Z}i \rangle_{c_2}$ in $\hat{\mathcal{N}}$, is independent of the topology and the element values in that part of $\hat{\mathcal{N}}$ corresponding to the s -copattern. Also, a special case of Theorem 7 arises when every branch containing an emf source $e_j \in H$ also contains a positive invertible resistor $r_j \in [H, H]$ in series with e_j . In this case, (29) can be reduced to

$$\langle i, \hat{Z}i \rangle_{c_2} \leq \sum_j \langle r_j^{-1} e_j, e_j \rangle_1,$$

where the summation is over those branches for which $e_j \neq 0$.

REFERENCES

- [1] V. DOLEZAL, *Hilbert networks. I*, this Journal, 12 (1974), pp. 755–778.
- [2] H. FLANDERS, *Infinite networks: I—Resistive networks*, IEEE Trans. Circuit Theory, 18 (1971), pp. 326–331.
- [3] A. H. ZEMANIAN, *Passive operator networks*, IEEE Trans. Circuits and Systems, CAS-21 (1974), pp. 184–193.
- [4] ———, *Infinite networks of positive operators*, Internat. J. Circuit Theory and Appl., 2 (1974), pp. 69–78.
- [5] ———, *Countably infinite networks that need not be locally finite*, IEEE Trans. Circuits and Systems, CAS-21 (1974), pp. 274–277.
- [6] J. JEANS, *Electricity and Magnetism*, 5th ed., Cambridge Univ. Press, London, 1927.
- [7] C. E. SHANNON AND D. W. HAGELBARGER, *Concavity of resistance functions*, J. Appl. Phys., 27 (1956), pp. 42–43.
- [8] H. M. MELVIN, *On concavity of resistance functions*, Ibid, 27 (1956), pp. 658–659.
- [9] S. K. BERBERIAN, *Introduction to Hilbert Space*, Oxford Univ. Press, New York, 1961.

MARKOVIAN REPRESENTATION OF STOCHASTIC PROCESSES BY CANONICAL VARIABLES*

HIROTUGU AKAIKE†

Abstract. The structure of the information interface between the future and the past of a discrete-time stochastic process is analyzed by using the concepts of canonical correlation analysis. Two extreme Markovian representations are obtained with states defined by the sets of canonical variables which represent the past information projected on the future and the future information projected on the past, respectively. The result completely clarifies the probabilistic structure of the Faurre algorithm of realization of stochastic systems. By an extension of the basic result the Ho-Kalman algorithm of realization of general systems is also given a stochastic interpretation.

1. Introduction. Given a sequence of covariance matrices $\{R_l, l = 0, 1, \dots\}$ of a zero mean d -dimensional stationary stochastic process $\{y_n; n = \dots - 1, 0, 1, \dots\}$, where $R_l = Ey_{n+l}y_n'$ and $'$ denotes the transpose, the problem to be considered here is to find a Markovian representation of y_n which is given by

$$(1.1) \quad \begin{aligned} x_{n+1} &= Fx_n + w_n, \\ y_n &= Hx_n, \end{aligned}$$

where the dimensions of x_n, w_n, F and H are $e \times 1, e \times 1, e \times e$ and $d \times e$, respectively, and w_n is an e -dimensional zero mean white noise, i.e., $Ew_{n+l}w_n' = 0$ for $l \neq 0$ and Q for $l = 0$ and $Ew_nx_{n-l}' = 0$ for $l = 0, 1, 2, \dots$, where 0 denotes a zero matrix. It is assumed that the process is Gaussian but the results have a natural interpretation in the sense of mean squares without this assumption. Ho and Kalman [1] gave an algorithm which realizes the minimal factorization $R_l = HF^lG$ ($l = 0, 1, 2, \dots$) which is with the minimum possible value of e . The original algorithm was developed generally for the realization of a linear constant system specified by F, G , and H which provides the given sequence of matrices, now R_l ($l = 0, 1, 2, \dots$), as the sequence of its impulse response matrices. Assuming the minimal factorization of R_l realized by the Ho-Kalman algorithm, Faurre [2], [3] proceeded further to give an algorithm for the factorization $G = PH'$ with P $e \times e$ -positive definite and $Q = P - FPF'$ nonnegative definite. In the paper by Ho and Kalman the minimal factorization was attained by reducing the dimension of a nonminimal factorization by a procedure which was considered to be difficult to motivate [1]. Faurre's result is based on the analysis of some quadratic functional of the process, but apparently the motivation for the use of this functional is not explicitly given.

In the present paper it is shown that if the covariance function R_l of a discrete-time stationary stochastic process y_n admits a finite-dimensional factorization $R_l = HF^lG$, it has two specific Markovian representations. The state x_n of one of these representations is defined as a set of mutually orthogonal random variables which contains the full information of the past of the process to be expressed by

* Received by the editors January 2, 1973, and in revised form July 16, 1973.

† The Institute of Statistical Mathematics, Minami-Azabu, Minato-ku, Tokyo 106, Japan. This research was supported in part by the National Science Foundation under Grants GK-31472 and GP-31074 at the University of Hawaii.

the present and future, and the state of the other representation contains the full information of the future of the process to be expressed by the present and past. This is proved by using the concept of canonical variables which is well developed in the field of multivariate statistical analysis.

When two sets of random variables are given, the set of canonical variables of one of the two sets is defined as the set of mutually orthogonal, or uncorrelated, random variables which form a basis of the space of linear combinations of the random variables of the original set and are ordered successively to give the highest possible correlation coefficients with one of the variables in the space of linear combinations of the random variables of the other set. Between the two sets of canonical variables, only the pair of canonical variables of one and the same order are correlated and the correlation coefficient is called the canonical correlation coefficient. If one of these two sets of random variables is composed of the present and past values of a stochastic process y_n and another of the present and future, the corresponding two sets of canonical variables with positive canonical correlations will form a minimal information interface between the past and the future of the process. It is shown that this idea can be made rigorous and these two sets can be used as the state variables x_n of two extreme Markovian representations of the process y_n .

These representations completely clarify the probabilistic background and the underlying motivation of both the Ho-Kalman and the Faurre algorithm as applied to the realization of stochastic systems. Also, with the aid of the canonical correlation coefficients, it provides a rational basis for the decision of fitting a lower dimensional approximation. The basic idea can further be extended to realize a minimum dimensional stochastic system which explains the covariance between two stationarily correlated stationary processes. This extension enables a probabilistic interpretation of the Ho-Kalman algorithm as applied to the general non-stochastic system realization problems. By specifying the state x_n of the Markovian representation of y_n to be physically realizable, or obtainable from y_n, y_{n-1}, \dots , this extension and the original procedure provide Kailath's two innovation process-type representations of y_n . Also, the ideas can be extended to cover the non-stationary case.

The purpose of the present paper is the conceptual clarification of the subject of realization of stochastic and nonstochastic systems by using the Markovian representation based on the notation of information interface between the future and the past of the stochastic processes under consideration. No attention has been given to the algorithmic aspect of the problem, for which there are important contributions by Rissanen [4], [5].

2. Canonical variables as an information interface. Consider two vectors of zero mean Gaussian random variables $u = (u_1, u_2, \dots, u_l)'$ and $v = (v_1, v_2, \dots, v_m)'$, where $'$ denotes transpose. If they share any statistical information in common this must be reflected in that their simultaneous distribution must be different from the distribution for which u and v are independent. From the definition of the characteristic function it is clear that the simultaneous distribution of u and v is determined by the distribution of every possible linear combination of u_i 's and v_j 's. Thus the analysis of the dependence between u and v is the analysis of the depend-

ence between the two spaces of random variables which are defined as the sets of all possible linear combinations of the u_i 's and the v_j 's, respectively. These spaces will be denoted by $R(u)$ and $R(v)$.

The theory of canonical correlations and variables, introduced by Hotelling [6] and now a fundamental concept in multivariate statistical analysis [7], says that these spaces $R(u)$ and $R(v)$ have orthonormal bases $\bar{u} = (\bar{u}_1, \bar{u}_2, \dots, \bar{u}_r)'$ and $\bar{v} = (\bar{v}_1, \bar{v}_2, \dots, \bar{v}_s)'$ such that

$$\begin{aligned} E\bar{u}_i &= E\bar{v}_j = 0, & i = 1, 2, \dots, r, \quad j = 1, 2, \dots, s, \\ E\bar{u}_i\bar{u}_j &= \delta_{ij}, & i, j = 1, 2, \dots, r, \\ E\bar{v}_i\bar{v}_j &= \delta_{ij}, & i, j = 1, 2, \dots, s, \\ E\bar{u}_i\bar{v}_j &= r_{ij}, & i = 1, 2, \dots, r, \quad j = 1, 2, \dots, s, \end{aligned}$$

where $\delta_{ij} = 1 (i = j) = 0 (i \neq j)$, $r_{ij} = 0 (i \neq j)$, $1 \geq r_{ii} \geq 0$, and E denotes expectation. It is assumed that r_{ii} are arranged in descending order of magnitude so that $r_{ii} \geq r_{i+1, i+1}$. r and s are equal to the ranks of the covariance matrices of u and v , respectively, and accordingly $r \leq l$ and $s \leq m$. The variables \bar{u}_i and \bar{v}_i are the canonical variables and r_{ii} is the canonical correlation coefficient of the pair of canonical variables \bar{u}_i and \bar{v}_i ($i = 1, 2, \dots, \min(r, s)$). If only the first k components of \bar{u} and \bar{v} have nonzero canonical correlation coefficients, the information shared by u and v , or the cause of the dependence between u and v , is completely contained within the components of the vectors $U = (\bar{u}_1, \bar{u}_2, \dots, \bar{u}_k)'$ and $V = (\bar{v}_1, \bar{v}_2, \dots, \bar{v}_k)'$. U and V or the spaces $R(U)$ and $R(V)$, respectively spanned by the components of U and V , may be considered to be the information interfaces of $R(u)$ and $R(v)$.

It is clear from the structure of U and V that $R(U)$ and $R(V)$ are respectively the projections of $R(v)$ on $R(u)$ and of $R(u)$ on $R(v)$. Since only the random variables with finite second order moments are considered, the projections can be understood either in the sense of conditional expectation or simply in the sense of mean square [8]. As to the positively correlated pairs of canonical variables, the canonical correlation analysis of any two sets of random variables which respectively span the spaces $R(U)$ and $R(V)$ will give an essentially identical result to that of the original u and v . The above observations of the structure of $R(U)$ and $R(V)$ are most useful in extending the concepts of canonical variables and correlation coefficients to the analysis of dependence of stochastic processes. It should be mentioned here that with a proper definition of the amount of information the same idea of canonical correlation analysis has been used by Gelfand and Yaglom [9] for the calculation of the amount of information about a random function contained in another random function.

3. Information interface and the Markovian representation. Denote by *past* the history $y_0, y_{-1}, y_{-2}, \dots$ and by *future* y_0, y_1, y_2, \dots . Consider the space $R(y)$ which is the closure with respect to the mean square norm of the space spanned by finite linear combinations of the components of y_n ($n = 0, \pm 1, \pm 2, \dots$). Denote by $R(\text{past})$ and $R(\text{future})$ the subspaces of $R(y)$ spanned by the components of the elements of past and future, respectively. The projection of $R(\text{past})$ on $R(\text{future})$ will be denoted by $R(\text{past}|\text{future})$ and that of $R(\text{future})$ on $R(\text{past})$ by $R(\text{future}|\text{past})$. These projections are respectively spanned by $y_0|\text{future}, y_{-1}|\text{future}, \dots$ and by

$y_0|\text{past}, y_1|\text{past}, \dots$, where $y_m|\text{future}$ and $y_m|\text{past}$ represent the projections of y_m on $R(\text{future})$ and $R(\text{past})$ and are specified by the relations

$$(3.1) \quad Ey_m y'_l = E(y_m|\text{future})y'_l, \quad l = 0, 1, 2, \dots,$$

and

$$(3.2) \quad Ey_m y'_{-l} = E(y_m|\text{past})y'_{-l}, \quad l = 0, 1, 2, \dots.$$

When the covariance function $R_l = Ey_{n+l}y'_n$ can be factored into the form $R_l = HF^lG$ ($l = 0, 1, 2, \dots$) with F of finite dimension, there exist a finite integer r and constants a_1, a_2, \dots, a_r , which can be determined from the minimal polynomial of F , [1], such that for all $j \geq 0$,

$$(3.3) \quad R_{r+j} = \sum_{i=1}^r a_i R_{r+j-i}.$$

This relation implies the equations

$$(3.4) \quad Ey_l \left(y_{-r-m} - \sum_{i=1}^r a_i y_{-r-m+i} \right)' = 0, \quad l, m = 0, 1, 2, \dots,$$

and

$$(3.5) \quad E \left(y_{r+m} - \sum_{i=1}^r a_i y_{r+m-i} \right) y'_{-l} = 0, \quad l, m = 0, 1, 2, \dots.$$

With the aid of (3.1) and (3.2), (3.4) and (3.5) show that

$$y_{-r-m}|\text{future} = \sum_{i=1}^r a_i y_{-r-m+i}|\text{future}, \quad m = 0, 1, 2, \dots,$$

and

$$y_{r+m}|\text{past} = \sum_{i=1}^r a_i y_{r+m-i}|\text{past}, \quad m = 0, 1, 2, \dots,$$

where equalities denote the equivalence with respect to the mean square norm. This result shows that the spaces $R(\text{past}|\text{future})$ and $R(\text{future}|\text{past})$ are finite-dimensional and are spanned by the components of the vectors $(y'_0|\text{future}, y'_{-1}|\text{future}, \dots, y'_{-r+1}|\text{future})'$ and $(y'_0|\text{past}, y'_1|\text{past}, \dots, y'_{r-1}|\text{past})'$, respectively.

By identifying u and v of § 2 with the above two vectors, the observation made at the end of the section shows that the canonical correlation analysis applied to these two sets of random variables provides $U = (\bar{u}_1, \bar{u}_2, \dots, \bar{u}_k)'$ and $V = (\bar{v}_1, \bar{v}_2, \dots, \bar{v}_k)'$, of which components form orthonormal bases of the spaces $R(\text{past}|\text{future})$ and $R(\text{future}|\text{past})$, respectively. These two spaces form the information interface between $R(\text{future})$ and $R(\text{past})$. Either U or V can be used to define a minimal Markovian representation of y_n . To see this, a study of the dynamics of the system is necessary. This forms a point of departure from the classical static multivariate analysis to the dynamic time series analysis.

The dynamics of the system under observation is reflected in the effect of translation of time on U or V . To analyze this, we adopt the convention to denote by S_n the vector or the set of vectors obtained by replacing y_i by y_{i+n} in the definition

of S , which is a vector or a set of vectors composed of elements of $R(y)$, especially $S_0 = S$. Since y_n is stationary, U_n and V_n ($n = 0, \pm 1, \pm 2, \dots$) are stationary and are stationarily correlated. Since U_n is a basis of $R(\text{past}_n | \text{future}_n)$ and U_{n+1} is a vector composed of elements of $R(\text{future}_n)$, there is a unique representation

$$(3.6) \quad U_{n+1} = FU_n + W_n,$$

where F is a $k \times k$ matrix of regression coefficients of U_{n+1} on U_n and W_n is a k -dimensional random vector which is independent of or uncorrelated with U_n . Since W_n is an element of $R(\text{future}_n)$ and is independent of U_n which spans $R(\text{past}_n | \text{future}_n)$, it is independent of the elements of $R(\text{past}_n)$. Since $R(\text{past}_n)$ is contained in $R(\text{past}_{n+m})$ ($m = 0, 1, 2, \dots$), W_{n+m} is independent of $R(\text{past}_n)$ and, since W_{n+m} is a vector composed of elements of $R(\text{future}_n)$, it is independent of U_n . From this it can be seen that W_n is a k -dimensional white noise. Since $y_n = y_n | \text{future}_n$, y_n is an element of $R(\text{past}_n | \text{future}_n)$ and thus there is a $d \times k$ matrix H such that

$$(3.7) \quad y_n = HU_n.$$

By identifying $x_n = U_n$, $w_n = W_n$, $e = k$ and $Q = I - FF'$, (3.6) and (3.7) give the desired Markovian representation (1.1) of y_n .

If we consider the projection of (3.6) on the space $R(\text{past}_{n+1})$, we get

$$(3.8) \quad \Lambda V_{n+1} = F\Lambda V_n + w_n,$$

where Λ is a $k \times k$ diagonal matrix of positive canonical correlation coefficients between U_n and V_n and w_n is the projection of $FU_n + W_n$ on the space spanned by the components of the innovation $y_{n+1} - y_{n+1} | \text{past}_n$. Since y_n is an element of $R(\text{past}_n)$, (3.7) is transformed into

$$(3.9) \quad y_n = H\Lambda V_n.$$

By identifying $x_n = \Lambda V_n$, $e = k$ and $Q = \Lambda^2 - F\Lambda^2 F'$, we get a second Markovian representation of y_n . Since in any Markovian representation of y_n its state x_n must contain the full information to be transmitted from past_n into future_n , it is clear that the dimension of x_n cannot be less than that of U_n and V_n . Thus the above two representations are minimal in the sense that k is the smallest possible value of e or the dimension of x_n of (1.1).

In the next section it is shown that the above two representations give the maximum and the minimum of the covariance matrix P of the state x_n of the Markovian representation of y_n with the transition matrix F and the observation matrix H . By Faurre's notation [2], [3] $P^* = I$ and $P_* = \Lambda^2$.

It should be mentioned here that if the diagonal elements of Λ are all different and the process is ergodic, so that the sample covariances converge to the true covariances with probability one as the length of available observation is indefinitely increased, and is full rank, in the sense that the covariance matrix of its arbitrary finite portion is always of full rank, the realization procedure of the Markovian representation described in this section can be applied to a properly chosen positive definite consistent estimate of covariance function to provide consistent estimates of the related quantities. If Λ has some equal elements on its diagonal, there remains an obvious indeterminacy. A well developed numerical procedure for the canonical correlation analysis is already available [10]. The

problem of statistical identification of Markovian representation will be discussed elsewhere.

4. Extreme properties of the representations. Faurre [3] has shown that when a minimal factorization of R_l is given in the form $R_l = HF^lG$ ($l = 0, 1, 2, \dots$), G can be factored into $G = PH'$ with P positive definite and $Q = P - FPF'$ non-negative definite. He has shown that the set of such P 's has a maximum P^* and a minimum P_* , in the sense that $P^* - P$ and $P - P_*$ are nonnegative definite for all P in this set. Now it can be shown that when there is a Markovian representation of y_n with the transition matrix F and the observation matrix H and with the state x_n which is a vector composed of elements of $R(\text{past}_n)$, then $Ex_n x'_n = P_*$. If x_n is a vector of elements of $R(\text{future}_n)$, then $Ex_n x'_n = P^*$. The two Markovian representations obtained in the preceding section form a pair of these two extreme representations. Before going into the discussion of these properties it should be noticed that if $R_l = HF^lG$ ($l = 0, 1, 2, \dots$) is a minimal factorization of R_l in the sense that F has the minimal possible dimension, then the system defined by $\{F, G, H\}$ is completely observable and completely controllable [1]. From the observability of the pair (F, H) , $\text{rank } [H' \ F'H' \ \dots \ F'^{r-1}H'] = k$, where r is as defined in § 3. Thus, given F and H , G in the minimal factorization $R_l = HF^lG$ ($l = 0, 1, 2, \dots$) is uniquely determined by the relation

$$G'[H' \ F'H' \ \dots \ F'^{r-1}H'] = [R'_0 \ R'_1 \ \dots \ R'_1 \ \dots \ R'_{r-1}].$$

Assume that y_n has a minimal Markovian representation

$$(4.1) \quad \begin{aligned} x_{n+1} &= Fx_n + w_n, \\ y_n &= Hx_n. \end{aligned}$$

The state covariance matrix $Ex_n x'_n$ is denoted by P . For an arbitrary representation (4.1) it holds that

$$(4.2) \quad Ex_n y'_{n-l} = F^l P H', \quad l = 0, 1, 2, \dots$$

As was noticed above, $G = PH'$ is uniquely determined when F and H are given. With the aid of this result (4.2) shows that the projection of x_n on $R(\text{past}_n)$ is identical for every representation (4.1). By following the derivation of (3.7) and (3.9) from (3.6) and (3.7), it can be shown that this projection defines the state of a Markovian representation of y_n with F and H of (4.1). Since the variance matrix reduces by the operation of projection, this projection has the minimum covariance matrix P_* within the set of the states of the Markovian representations of y_n with F and H of (4.1). Especially when x_n is a vector composed of elements of $R(\text{past}_n)$, it holds that $Ex_n x'_n = P_*$.

If P is the covariance matrix of the state of a minimal Markovian representation of y_n with the transition matrix F and the observation matrix H , then P^{-1} is the covariance matrix of the state of another Markovian representation of y_n which runs in the reversed direction of time and with the transition matrix F' and the observation matrix $G' (= HP)$. This representation is called the dual representation of the original Markovian representation. To prove the existence of the

dual representation, consider x_n of (4.1) and define z_n by

$$z_n = P^{-1}x_{n-1} - F'P^{-1}x_n,$$

where $P = Ex_nx'_n$. Since (4.1) is a minimal representation, P is nonsingular and $P^{-1}x_n$ is well-defined. From (4.1) we have

$$\begin{aligned} Ez_nx'_{n+l} &= P^{-1}Ex_{n-1}x'_{n+l} - F'P^{-1}Ex_nx'_{n+l} \\ &= 0, \end{aligned} \quad l = 0, 1, 2, \dots$$

From this relation it can be seen that the desired dual representation is given by

$$\begin{aligned} (4.3) \quad P^{-1}x_{n-1} &= F'P^{-1}x_n + z_n, \\ y_n &= HPP^{-1}x_n. \end{aligned}$$

It should be remembered that $G' (= HP)$ is a constant matrix determined by F and H . The state of this representation is $P^{-1}x_n$ and its covariance matrix is P^{-1} .

From the reasoning which was applied to show the minimality of $Ex_nx'_n$ when x_n is a vector composed of elements of $R(\text{past}_n)$, it can be seen that $P^{-1}x_n$ has the minimum covariance matrix within the set of the states of the dual representations defined with F' and $G' (= HP)$ of (4.3), when x_n is composed of elements of $R(\text{future}_n)$. For this case, since the covariance matrix of the state of the dual representation is equal to the inverse of the original state covariance matrix, it obviously holds that $Ex_nx'_n = P^*$, the maximum within the set of the state covariance matrices of Markovian representations of y_n with F and H of (4.1).

To show the identity of the present definitions of P^* and P_* to those of Faure, it is only necessary to show the identity of both of the definitions of P^* . Faure [3] starts with a minimal factorization $R_l = HF'G$ obtained by the Ho-Kalman algorithm. The $k \times k$ matrix P^* was defined by

$$(4.4) \quad x'P^*x = \inf_{C(x)} \sum_{l=0}^{\infty} \sum_{m=0}^{\infty} u'_l R_{l-m} u_m,$$

where u_l is a d -dimensional vector and the infimum is taken over the set $C(x)$ of $u = (u'_0, u'_1, \dots)'$ which is defined by

$$(4.5) \quad C(x) = \left\{ u \left| \sum_{l=0}^{\infty} F'^l H' u_l = x \right. \right\}.$$

From the observability of the pair (F, H) , $\text{rank} [H' \ F'H' \ \dots \ F'^{r-1}H'] = k$ and $C(x)$ is nonvoid for arbitrary x . Assume that y_n has a k -dimensional Markovian representation (4.1) with x_n composed of elements of $R(\text{future}_n)$. From (4.1), $Ex_ny'_{n+l} = PF'^lH'$, where $P = Ex_nx'_n$. By replacing F'^lH' of (4.5) by $P^{-1}Ex_ny'_{n+l}$ and putting $z = \sum_{l=0}^{\infty} y'_{n+l}u_l$, one can rewrite (4.4) in the form

$$(4.6) \quad x'P^*x = \inf_{D(x)} Ez^2,$$

where z is a one-dimensional random variable and the infimum is taken over the set $D(x)$ which is defined by

$$(4.7) \quad D(x) = \{z | z \in R(\text{future}_n) \text{ and } P^{-1}Ex_nz' = x\}.$$

(4.6) and (4.7) show that a z which has x as its vector of regression coefficients on x_n and with a minimum mean square value is to be found in the space of $R(\text{future}_n)$. Since x_n is assumed to be in $R(\text{future}_n)$, the desired solution is simply given by $x'x_n$ and from (4.6), $P^* = P$. This result shows the identity of Faurre's definition of P^* to the definition of the present paper. It was shown by Faurre that P^* as defined by (4.4) gives the maximum of P which appears in the factorization of $G = PH'$ with nonnegative definite $Q = P - FPF'$.

The results obtained above completely clarify the statistical or probabilistic structure of Faurre's results. It is now clear that for any Markovian representation (4.1) of y_n , the difference $P - P_*$ is the covariance of the portion of x_n which cannot be "explained" by past_n and vanishes when premultiplied by H . Thus only V_n or its nonsingular transformation can give a minimal Markovian representation, the state of which is physically realizable in the sense that it can be defined as a vector composed of elements of $R(\text{past}_n)$. This point is important when the Markovian representation is used for the realization of a predictor. The physically realizable Markovian representation can be obtained by first applying the Ho-Kalman algorithm to realize a minimal factorization $R_l = HF^lG$ ($l = 0, 1, 2, \dots$) and then applying the Faurre algorithm to realize the factorization $G = P_*H'$ with the corresponding variance matrix Q_* of the white noise determined by $Q_* = P_* - FP_*F'$. For an arbitrary real number c , $P_c = c^2P^* + (1 - c^2)P_*$ realizes the factorization $G = P_cH'$. P_c is the covariance matrix of $x_{cn} = c(x_n^* - x_{*n}) + x_{*n}$, where x_n^* and x_{*n} are the states of the extreme Markovian representations with $Ex_n^*x_n^{*'} = P^*$ and $Ex_{*n}x_{*n}' = P_*$, respectively. Although P_c realizes the factorization $G = P_cH'$, it is not clear whether x_{cn} is a state of a Markovian representation when c is not equal to either 0 or 1.

5. Analysis of the Ho-Kalman algorithm. In the middle of § 3, it was noticed that the canonical correlation analysis of the two sets of random variables $u = (y_0|\text{future}, y_{-1}'|\text{future}, \dots, y_{-r+1}'|\text{future})'$ and $v = (y_0|\text{past}, y_1'|\text{past}, \dots, y_{r-1}'|\text{past})'$ will give U and V as the sets of normalized canonical variables with positive canonical correlation coefficients. From this there must be $dr \times dr$ matrices T_u and T_v such that the first k components of $T_u u$ and $T_v v$ are respectively U and V and the rest are uncorrelated with any other components of $T_u u$ and $T_v v$. When the ranks of Euu' and Evv' are r and s , respectively, only the first r and s components of $T_u u$ and $T_v v$ are with unit variance and others are identically kept equal to zero.

We have

$$(5.1) \quad T_u Euv' T_v' = \begin{bmatrix} EUV' & 0 \\ 0 & 0 \end{bmatrix},$$

where 0 stands for a zero matrix of appropriate dimension. Since U spans the space $R(\text{past}|\text{future})$, y_j ($j = 0, 1, \dots, r-1$) has a (not necessarily unique) representation

$$y_j = \sum_{l=0}^{r-1} A_{jl} y_{-l}|\text{future} + r_j,$$

where r_j is an element of $R(\text{future})$ and $Er_j p' = 0$ for any element p of $R(\text{past}|\text{future})$.

Analogously, y_{-i} ($i = 0, 1, \dots, r-1$) has a representation

$$y_{-i} = \sum_{k=0}^{r-1} B_{ik} y_k | \text{past} + r_{-i},$$

where r_{-i} is an element of $R(\text{past})$ and $Er_{-i}t' = 0$ for any element t of $R(\text{future}|\text{past})$. A_{jl} and B_{ik} are $d \times d$ matrices. From the definition of r_j we have $Er_j y'_{-i} = Er_j(y_{-i}|\text{future})' = 0$. Also

$$E \left(\sum_{l=0}^{r-1} A_{jl} y_{-l} | \text{future} \right) r'_{-i} = E \left\{ \sum_{l=0}^{r-1} A_{jl} (y_{-l} | \text{future}) | \text{past} \right\} r'_{-i} = 0.$$

From these relations we get

$$Ey_j y'_{-i} = \sum_{l=0}^{r-1} \sum_{k=0}^{r-1} A_{jl} E(y_{-l} | \text{future}) (y_k | \text{past})' B'_{ik}.$$

This result shows that there are $rd \times rd$ matrices A and B , composed of the blocks A_{jl} and B_{ik} , such that

$$(5.2) \quad \begin{aligned} AEuv'B' &= E \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{r-1} \end{bmatrix} \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{-r+1} \end{bmatrix}' \\ &= \begin{bmatrix} R_0 & R_1 & \cdots & R_{r-1} \\ R_1 & R_2 & \cdots & R_r \\ \vdots & \vdots & \ddots & \vdots \\ R_{r-1} & R_r & \cdots & R_{2r-2} \end{bmatrix}. \end{aligned}$$

This is a generalized Hankel matrix which forms the starting point of the Ho-Kalman algorithm [1]. By the Ho-Kalman algorithm, the factorization of R_l is accomplished by first finding two nonsingular $rd \times rd$ matrices L and R which realize the transformation

$$(5.3) \quad LS_r R = \begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix},$$

where S_r denotes the generalized Hankel matrix defined by (5.2) and I_k is a $k \times k$ identity matrix. Since $EUUV' (= \Lambda)$ is a positive diagonal matrix, the difference between (5.1) and (5.3) is only a nonsingular k -dimensional linear transformation. While it was stated by Ho and Kalman [1] that the procedure to get through (5.3) a minimal factorization of R_l in the form $HF'G$ was difficult to motivate, the present realization of the Markovian representation based on the concept of information interface between future_n and past_n gives an illumination into the basic structure underlying the formula (5.3).

Though the difference between (5.1) and (5.3) is apparently only a linear transformation, the difference becomes significant when the problem of approxi-

mation of the system by a lower dimensional model is considered. In this case, to determine the best fitting model with dimension e which is less than k , a criterion of fit must be defined explicitly. When the realization of the Markovian representation through U_n or V_n is considered, any reasonable measure of the amount of information contained in each component of U_n or V_n will be dependent on the value of its canonical correlation coefficient, the higher the value the more the amount of information. Thus the best e -dimensional approximation will be obtained by keeping only the first e components of U_n or V_n , which correspond to the largest e canonical correlation coefficients, as the state of the e -dimensional model. The definition of the best fit will be made explicit by using the Kullback–Leibler mean amount of information for discrimination of two probability distributions [11], [12]. The same idea will be able to treat the case where the process has no finite-dimensional Markovian representation, which will always be the case when a real stochastic process is considered. The Kullback–Leibler definition of the amount of information has a natural connection with the log-likelihood-ratio statistic used for the discrimination of two distributions based on a finite number of observations and there is a closely related statistic used for statistical model identification [13]. This suggests that a reasonable statistical identification procedure may be developed along the line of the present approach for the case where only a finite length record of observations is available instead of the theoretical values of the covariance function.

6. Some extensions. Consider the case where there are two zero mean stationary stochastic processes y_n and z_n which are stationarily correlated with covariance function $R_l = Ey_{n+l}z'_n$ which is factorable into the form $R_l = HF^lG$. For this case define past as the past history of z_n , i.e., z_0, z_{-1}, \dots , and future as the future history of y_n , i.e., y_0, y_1, \dots . By proceeding entirely analogously as in the case of § 3 where $z_n = y_n$, one arrives at the representation

$$(6.1) \quad \begin{aligned} U_{n+1} &= FU_n + W_n, \\ y_n &= HU_n + N_n, \end{aligned}$$

where U_n and F are defined analogously as in § 3, W_n is independent of U_n, U_{n-1}, \dots and is a white noise, W_n and N_n are both independent of z_n, z_{n-1}, \dots and N_n is independent of U_n . This representation gives a factorization of R_l in the form $R_l = Ey_{n+l}z'_n = HF^lG$ with $G = EU_nz'_n$. Every R_l which admits a factorization in the form HF^lG with a stability matrix F can be considered as the covariance function $Ey_{n+l}z'_n$ of two stationary time series y_n and z_n , where z_n is a white noise of appropriate dimension with a covariance matrix equal to the identity matrix, and y_n is given by the relation

$$(6.2) \quad \begin{aligned} Y_{n+1} &= FY_n + Gz_n, \\ y_n &= HY_n + n_n, \end{aligned}$$

where n_n is an arbitrary stationary process of the same dimension as y_n , independent of the process z_n and with a finite covariance matrix. For this case even without the explicit specification of the covariance matrices of y_n , the desired factorization can be carried out through (5.3). This result gives a probabilistic explanation of the

structure of the Ho–Kalman algorithm as applied to the general, not necessarily stochastic, system realization problems.

(6.1) also shows that by assuming $z_n = y_{n-m}$ with some fixed positive m , we can get various representations of y_n which might be called Markovian representations of y_n with additive noise terms. By projecting U_n on $R(\text{past}_n)$, one gets a representation of the form (6.1) with U_n replaced by its projection. The representation thus obtained with $z_n = y_{n-1}$ corresponds to Kailath's innovation representation (IR1) and the original representation obtained with $z_n = y_n$ corresponds to the innovation representation (IR2), of y_n , discussed by Gevers [14]. Since the state variable of the representation corresponding to the choice $z_n = y_{n-1}$ can be obtained as a projection of the state variable obtained by putting $z_n = y_n$, it is clear that the former has a smaller variance matrix than the latter. By increasing the value of m further the variance matrix of the corresponding state variable becomes smaller, suggesting the decrease of information contained within the state variable.

It should also be mentioned here that if only the finiteness of the dimension of U_n or V_n is assumed, the basic results of § 3 are not dependent on the assumption of stationarity. Thus under the finite dimensionality assumption the present approach can also give Markovian representations of nonstationary processes.

Acknowledgments. The author would like to express his thanks to Professor R. E. Kalman, University of Florida, for providing the chance to meet the problem and to Professor W. Gersch, University of Hawaii, for providing the latest information on the subject with continuous stimulus and encouragement. Thanks are also due to Professor R. H. Jones, University of Hawaii, for encouragement during the preparation of the present paper. The author is deeply indebted to Professor J. Rissanen, Linköping University, for the helpful comments which led to an improvement of the original version of the present paper.

REFERENCES

- [1] B. L. HO AND R. E. KALMAN, *Effective construction of linear state-variable models from input/output functions*, Regelungstechnik, 14 (1966), pp. 545–548.
- [2] P. FAURRE AND J. P. MARMORAT, *Un algorithme de realisation stochastique*, C. R. Acad. Sci. Paris, Serie A, 268 (1969), pp. 978–981.
- [3] P. FAURRE, *Identification par minimisation d'une representation Markovienne de processus aleatoire*, Symposium on Optimization, Lecture Notes in Mathematics 132, Springer, Berlin, 1970, pp. 83–107.
- [4] J. RISSANEN, *Recursive identification of linear systems*, this Journal, 9 (1971), pp. 420–430.
- [5] J. RISSANEN AND T. KAILATH, *Partial realization of random systems*, Automatica, 8 (1972), pp. 389–396.
- [6] H. HOTELLING, *Relations between two sets of variables*, Biometrika, 28 (1936), pp. 321–377.
- [7] T. W. ANDERSON, *Introduction to Multivariate Statistical Analysis*, John Wiley, New York, 1958.
- [8] J. NEVEU, *Mathematical Foundations of the Calculus of Probability*, Holden-Day, San Francisco, 1965.
- [9] I. M. GELFAND AND A. M. YAGLOM, *Calculation of the amount of information about a random function contained in another such function*, Amer. Math. Soc. Transl. (2), 12 (1959), pp. 199–246.
- [10] G. H. GOLUB, *Matrix decompositions and statistical calculations*, Statistical Computation, R. C. Milton and J. A. Nelder, eds., Academic Press, New York, 1969, pp. 365–397.
- [11] S. KULLBACK AND R. A. LEIBLER, *On information and sufficiency*, Ann. Math. Statist., 22 (1951), pp. 79–86.

- [12] S. KULLBACK, *Information Theory and Statistics*, Dover, New York, 1968.
- [13] H. AKAIKE, *Use of an information theoretic quantity for statistical model identification*, Proc. 5th Hawaiian International Conference on System Sciences, 1972, pp. 249–250.
- [14] M. R. GEVERS, *Structural properties of realizations of discrete time Markovian processes*, Tech. Rep. 7050-19, Information Systems Laboratory, Stanford University, Stanford, Calif., 1972.

BOUNDARY VALUE CONTROL OF THE WAVE EQUATION IN A SPHERICAL REGION*

KEITH D. GRAHAM† AND DAVID L. RUSSELL‡

Abstract. In this paper we study controllability problems for the wave equation

$$\frac{\partial^2 w}{\partial t^2} - \sum_{j=1}^n \frac{\partial^2 w}{(\partial x^j)^2}, \quad t \geq 0, \quad x \in \Omega,$$

where Ω is a spherical region in R^n . The control force f enters in the boundary condition

$$\frac{\partial w}{\partial \nu} = f,$$

assumed to hold for $t \geq 0$, $x \in \Gamma = \partial\Omega$. Our main result is that all "finite energy" initial states can be steered to the zero state in time τ , using a control $f \in L^2(\Gamma \otimes [0, \tau])$, provided $\tau > 2$.

Beginning with standard existence results for solutions of the wave equation, we show the control problem to be equivalent to a collection of trigonometric moment problems. These are solved using the theory of nonharmonic Fourier series together with certain results concerning the separation of eigenvalues of the Helmholtz operator in $L^2(\Omega)$ with the Neumann boundary condition. Essential use is also made of the theory of interpolation spaces and a priori estimates for solutions of elliptic boundary value problems.

1. Introduction and summary of results. In this paper we shall be concerned with the wave equation

$$(1.1) \quad \frac{\partial^2 w}{\partial t^2} = \Delta w = \sum_{i=1}^n \frac{\partial^2 w}{(\partial x^i)^2}, \quad x \in \Omega, \quad t \geq 0,$$

where Ω is the interior of the sphere in R^n . With $\|\cdot\|_e$ denoting the Euclidean norm,

$$(1.2) \quad \Omega = \{x \in R^n \mid \|x\|_e < 1\}$$

and $\partial\Omega \equiv \Gamma$ is the set

$$(1.3) \quad \Gamma = \{x \in R^n \mid \|x\|_e = 1\}.$$

In most physical processes, control is applied at the boundary of the spatial region in which the process evolves. As the equation (1.1) is the archetype for wave processes, which are of importance in the study of structural vibration, counter-current heat exchangers, tubular catalytic reactions, etc., the control of solutions of this equation by means of boundary forcing functions is of considerable interest.

If a force $f(x, t)$ is applied to the above system, and if $f(x, t)$ is defined for $x \in \Gamma = \partial\Omega$ and for $t \geq 0$, the effect of such a force is described mathematically by the requirement

$$(1.4) \quad \frac{\partial w}{\partial \nu}(x, t) = f(x, t), \quad x \in \Gamma, \quad t \geq 0,$$

* Received by the editors May 24, 1973. This research was supported in part by the Air Force Office of Scientific Research/AFSC, United States Air Force, under Contract F44620-72-C-0044.

† Honeywell, Inc., Minneapolis, Minnesota 55413.

‡ Departments of Mathematics and Computer Sciences, University of Wisconsin, Madison, Wisconsin 53706 and Honeywell, Inc., Minneapolis, Minnesota 55413.

where ν is the unit outward normal vector to Γ at the point $x \in \Gamma$. In the present paper the admissible control functions f are those which satisfy

$$(1.5) \quad f \in L^2(\Gamma \otimes [0, T])$$

for some $T > 0$.

Our restriction of Z to the very particular set (1.2) is motivated by our wish to obtain controllability results of the strongest sort. Weaker controllability results for more general domains Ω are already available. In [21], [22] the set of controllable states is shown to be dense in $H^1(\Omega) \otimes L^2(\Omega)$ without any particular geometric properties being imposed on Ω . In [24], again with general geometry for Ω , the set of controllable states is shown to include $H^2(\Omega) \otimes H^1(\Omega)$. Very strong results are known for the case $x \in R^1$ (see, e.g., [9], [19], [20], [23]). In the present paper, as we indicate in much greater detail below, we show that the set of controllable states for the system (1.1), (1.4) includes $H^1(\Omega) \otimes L^2(\Omega)$ in the special case where Ω is the domain described by (1.2).

Let the Sobolev spaces $H^r(\Omega)$, $H^s(\Gamma \otimes [0, T])$ of real orders r, s be defined as in [14], [15]. We pose for (1.1) the initial conditions

$$(1.6) \quad w(x, 0) \equiv w_0(x), \quad \frac{\partial w}{\partial t}(x, 0) \equiv v_0(x), \quad x \in \Omega,$$

where the initial displacement w_0 and initial velocity v_0 satisfy

$$(1.7) \quad w_0 \in H^1(\Omega), \quad v_0 \in H^0(\Omega) \equiv L^2(\Omega).$$

In the remainder of this work the term “control problem” refers to the following.

CONTROL PROBLEM. *Let $T > 0$ be prescribed. Given an initial state (w_0, v_0) as specified in (1.7), find a control f satisfying (1.5) so that the solution $w(x, t)$ of (1.1), (1.4), (1.6) also satisfies*

$$(1.8) \quad w(x, T) \equiv \frac{\partial w}{\partial t}(x, T) \equiv 0.$$

Remark. Since the wave equation is time reversible, if the above control problem is solvable for all initial conditions (1.6), (1.7)— T remaining fixed—then we can also solve a more general control problem wherein

$$(1.9) \quad w(x, T) \equiv w_1(x) \in H^1(\Omega), \quad \frac{\partial w}{\partial t}(x, T) \equiv v_1(x) \in H^0(\Omega),$$

by letting $g(x, t) \in L^2(\Gamma \otimes [0, T])$ be a control such that the solution $\tilde{w}(x, t)$ of (1.1) satisfying these terminal conditions and

$$\frac{\partial \tilde{w}}{\partial \nu}(x, t) \equiv g(x, t), \quad x \in \Omega, \quad 0 \leq t \leq T,$$

satisfies

$$\tilde{w}(x, 0) \equiv \frac{\partial \tilde{w}}{\partial t}(x, 0) \equiv 0.$$

The general control problem (1.1), (1.6), (1.9) is then solved by adding the above solutions w , \tilde{w} , whose sum satisfies

$$\frac{\partial(w + \tilde{w})}{\partial v}(x, t) = f(x, t) + g(x, t) \in L^2(\Gamma \otimes [0, T]).$$

The main result proved in this paper is summarized in the following theorem.

THEOREM 1.1. *If $T > 2$, the control problem is solvable for $f \in L^2(\Gamma \otimes [0, T])$; indeed, there is a constant $K > 0$ such that*

$$(1.10) \quad \|f\|_{L^2(\Gamma \otimes [0, T])}^2 \leq K(\|w_0\|_{H^1(\Omega)}^2 + \|v_0\|_{H^0(\Omega)}^2).$$

If $T < 2$, the control problem is not solvable in general.

However, even if $T > 2$, it is not possible to guarantee that the solution $w(x, t)$ will have the property that $\|w(\cdot, t)\|_{H^1(\Omega)}$ and $\|(\partial w / \partial t)(\cdot, t)\|_{H^0(\Omega)}$ are finite for all $t \in [0, T]$. This is a consequence of the rather weak regularity properties of solutions of (1.1), (1.4), (1.5) when the only restriction on f is (1.5). The best result proved in [15] is to the effect that for $\varepsilon > 0$,

$$\int_0^T \|w(\cdot, t)\|_{H^{1/2-\varepsilon}(\Omega)}^2 dt < \infty$$

and

$$\int_0^T \left\| \frac{\partial w}{\partial t}(\cdot, t) \right\|_{H^{-1/2-\varepsilon}(\Omega)}^2 dt < \infty$$

under these assumptions. This is reinforced by a further result of the present paper, which demonstrates by a constructive procedure that if $w_0 \equiv v_0 \equiv 0$, in (1.7), there is a control $f \in H^0(\Gamma \otimes [0, T])$ such that the terminal state $w(x, T)$, $(\partial w / \partial t)(x, T)$ does not lie in $H^1(\Omega) \oplus H^0(\Omega)$. Because of the time reversibility of our system, this is a direct consequence of the following theorem.

THEOREM 1.2. *There exist initial states w_0, v_0 which can be steered to the terminal state $w(x, T) \equiv (\partial w / \partial t)(x, T) \equiv 0$ with a control $f \in L^2(\Gamma \otimes [0, T])$, $T > 2$, and for which*

$$(w_0, v_0) \notin H^1(\Omega) \oplus H^0(\Omega).$$

In addition to the implications noted before its statement, Theorem 1.2 also shows that, strong as it is (compared with the results in [19], [21], [24] for more general domains Ω), Theorem 1.1 is not a "best possible" result. What we have obtained in Theorem 1.1 are sufficient, not necessary, conditions for controllability. In the way of necessary conditions, not too much is available. A starting point is provided by the following theorem.

THEOREM 1.3. *In the case $n = 2$, $T > 2$, a necessary condition for solution of the control problem with a control function $f \in L^2(\Gamma \otimes [0, T])$ is that the initial state w_0, v_0 satisfy*

$$w_0 \in H^{2/3}(\Omega), \quad v_0 \in H^{-1/3}(\Omega).$$

2. Solutions of $w_{tt} - \Delta w = 0$. In this section we will describe some basic existence, uniqueness and regularity theorems relevant to the partial differential equation and boundary conditions introduced in § 1. This material is not new but is required for effective presentation of our controllability results. The theorems

stated here are proved in [14], to which reference we have also referred the reader for definitions and basic properties of the Sobolev spaces H^r , where r is a non-negative real number. References [3] and [13] are likewise helpful here.

The trace theorem [14] shows that if $w \in H^2(\Omega)$, then $\partial w / \partial \nu$ is defined on Γ as an element of the interpolation space $H^{1/2}(\Gamma)$ and

$$\left\| \frac{\partial w}{\partial \nu} \right\|_{H^{1/2}(\Gamma)} \leq K \|w\|_{H^2(\Omega)}$$

for some positive constant K . It follows that

$$V = \left\{ w \in H^2(\Omega) \left| \frac{\partial w}{\partial \nu} = 0 \text{ on } \Gamma \right. \right\}$$

is a closed subspace of $H^2(\Omega)$ and hence a Hilbert space in its own right with the inner product induced from $H^2(\Omega)$. It can be shown that V is dense in $H^1(\Omega)$.

The dual, V' of V with respect to $H^1(\Omega)$, is defined as follows. We let $w \in H^1(\Omega)$ and define a continuous linear functional on $H^1(\Omega)$ by

$$(2.1) \quad \ell_w(u) = (u, w)_{H^1(\Omega)}, \quad u \in H^1(\Omega).$$

Then

$$|\ell_w(u)| \leq \|u\|_{H^1(\Omega)} \|w\|_{H^1(\Omega)}.$$

For $u \in V \subseteq H^2(\Omega)$ we have

$$\|u\|_V = \|u\|_{H^2(\Omega)} \geq \|u\|_{H^1(\Omega)}$$

and we obtain

$$|\ell_w(u)| \leq \|u\|_V \|w\|_{H^1(\Omega)},$$

showing that ℓ_w is also a continuous linear functional on V . Since V is a Hilbert space, there is a unique $v(w) \in V$ such that

$$(2.2) \quad \ell_w(u) = (u, v(w))_V, \quad u \in V.$$

We define

$$(2.3) \quad \|w\|_{V'} = \|v(w)\|_V,$$

and define V' itself to be the completion of $H^1(\Omega)$ with respect to this norm.

For the purpose of this paper it is important that we be able to identify the space V' as $L^2(\Omega)$. We do this as follows. For each $w \in H^1(\Omega)$,

$$\|w\|_{V'} = \|v(w)\|_V = \sup_{\substack{u \in V \\ u \neq 0}} \frac{|(u, v(w))_V|}{\|u\|_V}.$$

But, comparing (2.1) and (2.2),

$$\begin{aligned} (u, v(w))_V &= (u, w)_{H^1(\Omega)} \\ &= \int_{\Omega} \left[u(x)w(x) + \sum_{i=1}^n \frac{\partial u}{\partial x^i} \frac{\partial w}{\partial x^i} \right] dx. \end{aligned}$$

Then integrating by parts and using the fact that $u \in V$,

$$(u, v(w))_V = \int_{\Omega} \left(u - \sum_{i=1}^n \frac{\partial^2 u}{\partial x_i^2} \right) w(x) dx.$$

Hence,

$$(2.4) \quad \|w\|_{V'} = \sup_{\substack{u \in V \\ u \neq 0}} \frac{|\int_{\Omega} (u(x) - \sum_{i=1}^n \partial^2 u / \partial x_i^2) w(x) dx|}{\|u\|_V}.$$

Using the Schwarz inequality,

$$\begin{aligned} \|w\|_{V'} &\leq \sup_{\substack{u \in V \\ u \neq 0}} \frac{[\int_{\Omega} (u(x) - \sum_{i=1}^n \partial^2 u / \partial x_i^2)^2 dx]^{1/2} \|w\|_{L^2(\Omega)}}{\|u\|_V} \\ &\leq \sup_{\substack{u \in V \\ u \neq 0}} \frac{K \|u\|_V \|w\|_{L^2(\Omega)}}{\|u\|_V} = K \|w\|_{L^2(\Omega)} \end{aligned}$$

for some fixed positive number K . On the other hand it is known [14, Chap. 2, § 6] that the equation

$$u(x) - \sum_{i=1}^n \frac{\partial^2 u}{\partial x_i^2} = w(x)$$

has a unique solution $\hat{u} \in V$ with

$$\|\hat{u}\|_V \leq \hat{K} \|w\|_{L^2(\Omega)}$$

for some fixed positive \hat{K} . Equation (2.4) with $u = \hat{u}$ gives

$$\begin{aligned} \left| \int_{\Omega} \left(\hat{u}(x) - \sum_{i=1}^n \frac{\partial^2 \hat{u}}{\partial x_i^2} \right) w(x) dx \right| &= \|w\|_{L^2(\Omega)}^2 \leq \|\hat{u}\|_V \|w\|_{V'} \\ &\leq \hat{K} \|w\|_{L^2(\Omega)} \|w\|_{V'}, \end{aligned}$$

so that

$$\|w\|_{L^2(\Omega)} \leq \hat{K} \|w\|_{V'}.$$

Thus we see that for $w \in H^1(\Omega)$, the $L^2(\Omega)$ and V' norms of w are equivalent. Since V' is the completion of $H^1(\Omega)$ with respect to the V' norm it is then also the completion of $H^1(\Omega)$ with respect to the $L^2(\Omega)$ norm, which is, of course, $L^2(\Omega)$. Thus, topologically, V' and $L^2(\Omega)$ are the same space and the norms are equivalent, i.e.,

$$(2.5) \quad \frac{1}{K} \|w\|_{V'} \leq \|w\|_{L^2(\Omega)} \leq \hat{K} \|w\|_{V'}.$$

In [13] and [14] we find the proof of the following theorem.

THEOREM 2.1. *Let $T_1 > 0$ and suppose that $\varphi_0 \in V$, $\varphi_1 \in H^1(\Omega)$, $h \in L^2([0, T_1]; H^1(\Omega))$. Then there is a unique function $\varphi(x, t)$ such that*

$$(2.6) \quad (i) \ \varphi(\cdot, t) \text{ is continuous from } [0, T_1] \text{ into } V \text{ and } \varphi(\cdot, 0) = \varphi_0;$$

- (2.7) (ii) $(\partial\varphi/\partial t)(\cdot, t)$ is continuous from $[0, T_1]$ into $H^1(\Omega)$ and $(\partial\varphi/\partial t)(\cdot, 0) = \varphi_1$;
 (iii) $\partial^2\varphi/\partial t^2$ and $\sum_{i=1}^n \partial^2\varphi/(\partial x^i)^2$ lie in $L^2(0, T_1; V')$ and satisfy, in $L^2(0, T_1; V')$,

$$(2.8) \quad \frac{\partial^2\varphi}{\partial t^2} - \sum_{i=1}^n \frac{\partial^2\varphi}{(\partial x^i)^2} = h \in L^2(0, T_1; H^1(\Omega)) \quad (\subseteq L^2([0, T_1]; V')).$$

Remark. Property (i) implies, in particular, that $\varphi(\cdot, t) \in V$, $t \in [0, T_1]$, and thus that φ satisfies

$$(2.9) \quad \frac{\partial\varphi}{\partial\nu}(x, t) = 0, \quad x \in \Gamma.$$

Using this result, one can define, in a rather weak sense, solutions of (1.1), (1.4) for functions f satisfying (1.5). This is done in [13, pp. 319 ff.] using Green's formula. For our case, the result becomes the following.

THEOREM 2.2. *Let w_0, v_0 satisfy (1.7) and let f satisfy (1.5) with T replaced by T_1 . Then there exists a unique function $w \in L^2(\Omega \otimes [0, T_1])$ such that for all solutions φ of (2.8), as described in Theorem 2.1, for which*

$$(2.10) \quad \varphi(x, T_1) \equiv \frac{\partial\varphi}{\partial t}(x, T_1) \equiv 0,$$

we have

$$(2.11) \quad \int_{\Omega \otimes [0, T_1]} w \left(\frac{\partial^2\varphi}{\partial t^2} - \sum_{i=1}^n \frac{\partial^2\varphi}{(\partial x^i)^2} \right) dx dt = (\text{cf. (2.8)}) \int_{\Omega \otimes [0, T_1]} w(x, t) h(x, t) dx dt$$

$$= \int_{\Omega} \left[v_0(x) \varphi(x, 0) - w_0(x) \frac{\partial\varphi}{\partial t}(x, 0) \right] dx$$

$$+ \int_{\Gamma \otimes [0, T_1]} f(x, t) \varphi(x, t) ds dt.$$

Using Green's formula one sees that smooth solutions of (1.1) satisfy (2.11). Here we use (2.11) as the defining property for certain weak solutions of (1.1).

Remarks. Since the wave equation is invariant under time reversal, the replacement of the initial conditions stated in Theorem 2.1 by the terminal conditions (2.10) causes no problem.

Our identification of V' with $L^2(\Omega)$ combines with (iii) of Theorem 2.1 to show that the left-hand side of (2.11) is defined for $w \in L^2(\Omega \otimes [0, T_1])$.

Weak as it is, Theorem 2.2 is all that we need to pose the controllability problem in an exact manner and to characterize its solutions.

All we know about the solution $w(x, t)$ of (1.1), (1.4), as interpreted in Theorem 2.2, is that $w \in L^2(\Omega \otimes [0, T_1])$. Thus $w(\cdot, T)$, $(\partial w/\partial t)(\cdot, T)$ are not very well defined for a fixed T , as required in our statement of the control problem. We therefore replace the condition (1.8) by the following. We let $T_1 > T$ and extend f (cf. (1.4), (1.5)) from $[0, T]$ to $[0, T_1]$ by setting

$$(2.12) \quad f(x, t) \equiv 0, \quad x \in \Gamma, \quad t \in (T, T_1].$$

We let $w(x, t)$ be the solution obtained from Theorem 2.2. Then (1.8) is replaced by

$$(2.13) \quad w(x, t) \equiv 0, \quad (x, t) \in \Omega \otimes [T, T_1].$$

It is clear that (1.8) and (2.13) are equivalent in the case of classical solutions of (1.1), (1.4).

3. The operator Δ . Consider the Laplacian operator

$$\Delta w = \sum_{i=1}^n \frac{\partial^2 w}{(\partial x^i)^2}$$

defined in the domain V introduced in § 2. So defined, Δ is an unbounded, self-adjoint operator in $L^2(\Omega)$ with domain V . It is known [3] that its spectrum consists entirely of eigenvalues, $-\lambda$, which, together with an associated eigenfunction U , satisfy

$$(3.1) \quad \Delta U + \lambda U = 0 \quad \text{in } \Omega, \quad \frac{\partial U}{\partial \nu} = 0 \quad \text{on } \Gamma = \partial\Omega.$$

The real numbers λ are known to be nonnegative. They may be identified in our case, where Ω is the unit ball in R^n , by the introduction of "hyperspherical coordinates" r, θ, ϕ , ([6, p. 233]), where r is the radial coordinate, $\theta = \{\theta_j | 0 \leq \theta_j \leq \pi, j = 1, 2, \dots, p \equiv n - 2\}$ are $n - 2$ coordinates of "longitude" and $0 \leq \phi \leq 2\pi$ is the coordinate of "latitude". When $n = 2$, no θ coordinates are required. In terms of these coordinates,

$$(3.2) \quad \Delta = r^{1-n} \frac{\partial}{\partial r} \left(r^{n-1} \frac{\partial}{\partial r} \right) + r^{-2} L_{\theta, \phi},$$

where $L_{\theta, \phi}$ is a second order self-adjoint partial differential operator involving only the angular variables [6]. The boundary condition $\partial U / \partial \nu \equiv 0$ now becomes

$$(3.3) \quad \frac{\partial U}{\partial r}(1, \theta, \phi) \equiv 0.$$

Writing U in separated form,

$$(3.4) \quad U(r, \theta, \phi) = R(r)Y(\theta, \phi),$$

and substituting (3.4) and (3.2) in (3.1) we find that, for some constant C ,

$$L_{\theta, \phi} Y(\theta, \phi) = -C Y(\theta, \phi).$$

This equation has nontrivial solutions for

$$C = C_k = k(k + p), \quad k = 0, 1, 2, \dots,$$

where

$$p = n - 2.$$

It is known [6, p. 237] that there are exactly

$$h = h(k, p) = (2k + p) \frac{(k + p - 1)!}{p!k!}$$

linearly independent solutions for each value of k , which we denote by $Y_{km}(\theta, \phi)$, $m = 1, 2, \dots, h(k, p)$. These are the spherical harmonics of degree k . Since $L_{\theta, \phi}$ is self-adjoint, we may assume these solutions orthonormal in $L^2(\Gamma)$. For $k = 0$, there is exactly one spherical harmonic

$$Y_0(\theta, \phi) = \left[\frac{\Gamma(1 + p/2)}{2\pi^{1+p/2}} \right]^{1/2} = \left[\frac{\Gamma(n/2)}{2\pi^{n/2}} \right]^{1/2},$$

with Γ here denoting Euler's gamma function.

With $Y(\theta, \phi) = Y_{k,m}(\theta, \phi)$, we find that (3.4) solves (3.1), (3.3) if and only if $R(r) \equiv R_k(r)$ is a solution of

$$(3.5) \quad \frac{d^2 R_k}{dr^2} + \frac{p+1}{r} \frac{dR_k}{dr} + \left[\lambda - \frac{k(k+p)}{r^2} \right] R_k = 0,$$

$$(3.6) \quad \frac{dR_k}{dr}(1) \equiv R'_k(1) = 0, \quad R_k(0+) \text{ bounded.}$$

It is shown in [3], for example, that the numbers λ which satisfy (3.1), (3.3) for nonzero U are precisely the numbers $\lambda_{k\ell}$ for which (3.5), (3.6) has a nontrivial solution. Such solutions are $J_{k+p/2}(\omega_{k\ell}r)$, the Bessel functions (first kind) of order $k + p/2$, with $\lambda_{k\ell}$ identified by

$$\lambda_{k\ell} = \omega_{k\ell}^2,$$

where $\omega_{00} = 0$ and, for $k \geq 0$, $\omega_{k\ell}$ ($\ell = 1, 2, 3, \dots$) is the ℓ th positive root of the equation

$$0 = (-p/2)J_{k+p/2}(\omega) + \omega J'_{k+p/2}(\omega).$$

Thus for $k = 0$, ℓ assumes values $0, 1, 2, 3, \dots$, and for $k > 0$, the values $1, 2, 3, \dots$. The complete set of real normalized eigenfunction solutions of (3.5), (3.6) is

$$(3.7) \quad \begin{aligned} R_{00} &\equiv \sqrt{p+2} = \sqrt{n}, \quad k=0, \quad \ell=0, \\ R_{0\ell}(r) &= \left[\frac{2}{[J_{p/2}(\omega_{0\ell})]_2} \right]^{1/2} r^{-p/2} J_{p/2}(\omega_{0\ell}r), \quad k=0, \quad \ell=1, 2, \dots, \\ R_{k\ell}(r) &= \left[\frac{2\omega_{k\ell}^2}{[\omega_{k\ell}^2 - k(k+p)][J_{k+p/2}(\omega_{k\ell})]^2} \right]^{1/2} r^{-p/2} J_{k+p/2}(\omega_{k\ell}r), \\ &\quad k=1, 2, \dots, \quad \ell=1, 2, \dots. \end{aligned}$$

When we say that these solutions are normalized, we mean

$$\int_0^1 R_{k\ell}(r) R_{kj}(r) r^{n-1} dr = \delta_{\ell j}.$$

It can be shown [7] that at $r = 1$ these functions have the lower bound

$$(3.8) \quad |R_{k\ell}^{(p)}(1)| > |R_{k\ell}^{(0)}(1)| > \sqrt{2}, \quad p = 1, 2, \dots, \quad \ell = 1, 2, \dots$$

and that the $R_{k1}^{(p)}$ have the asymptotic property

$$(3.9) \quad |R_{k1}^{(p)}(1)|^2 > |R_{k1}^{(0)}(1)|^2 = \hat{C}k^{2/3} + O(1), \quad k \rightarrow \infty,$$

for some positive number \hat{C} .

A complete orthonormal system in $L^2(\Omega)$ consisting of eigenfunctions of the Laplacian Δ with domain V is obtained by substituting the $Y_{k,m}$ for Y and the $R_{k,\ell}$ for R in (3.4). We obtain

$$(3.10) \quad U_{010}(x) = U_{010}(r, \theta, \phi) = \sqrt{n} \left[\frac{\Gamma(n/2)}{2\pi^{n/2}} \right]^{1/2}, \quad k = 0, \quad m = 1, \quad \ell = 0,$$

$$(3.11) \quad U_{01\ell}(x) = U_{01\ell}(r, \theta, \phi) = R_{0\ell}(r) Y_0(\theta, \phi), \\ k = 0, \quad m = 1, \quad \ell = 1, 2, 3, \dots,$$

$$(3.12) \quad U_{km\ell}(x) = U_{km\ell}(r, \theta, \phi) = R_{k\ell}(r) Y_{km}(\theta, \phi), \\ k = 1, 2, 3, \dots, \quad m = 1, 2, \dots, h(k, p), \quad \ell = 1, 2, 3, \dots.$$

An arbitrary function $w \in L^2(\Omega)$ has an expansion, convergent in the $L^2(\Omega)$ norm, of the form

$$(3.13) \quad w(x) = w(r, \theta, \phi) = \sum_{\ell=0}^{\infty} w_{01\ell} U_{01\ell} + \sum_{k=1}^{\infty} \sum_{m=1}^{h(k,p)} \sum_{\ell=1}^{\infty} w_{km\ell} U_{km\ell},$$

where

$$(3.14) \quad w_{01\ell} = \int_{\Omega} w(r, \theta, \phi) U_{01\ell}(r, \theta, \phi) r^{n-1} dr ds, \\ k = 0, \quad m = 1, \quad \ell = 0, 1, 2, \dots, \\ w_{km\ell} = \int_{\Omega} w(r, \theta, \phi) U_{km\ell}(r, \theta, \phi) r^{n-1} dr ds, \\ k = 1, 2, 3, \dots, \quad m \equiv 1, 2, \dots, h(k, p), \quad \ell = 1, 2, 3, \dots.$$

Moreover, we have Parseval's identity,

$$\sum_{\ell=0}^{\infty} w_{01\ell}^2 + \sum_{k=1}^{\infty} \sum_{m=1}^{h(k,p)} \sum_{\ell=1}^{\infty} w_{km\ell}^2 = \|w\|_{L^2(\Omega)}^2.$$

We note, in closing this section, a point which will be useful in the sequel. It is known (see, e.g., [7]) that for all $k \geq 1, \ell \geq 1$,

$$(3.15) \quad \omega_{k\ell} > \sqrt{k(k+p)}.$$

This provides a lower bound for the $\omega_{k\ell}$. Actually, such an inequality is needed already in this section if $R_{k\ell}(r)$ is to represent a (finite) real number.

4. An equivalent moment problem. Let $w(x, t)$ be a solution of (1.1), (1.4), (1.6), as defined in § 2, the initial data having expansions in $L^2(\Omega)$:

$$(4.1) \quad w_0(x) = \sum_{\ell=0}^{\infty} w_{01\ell} U_{01\ell} + \sum_{k=1}^{\infty} \sum_{m=1}^{h(k,p)} \sum_{\ell=1}^{\infty} w_{km\ell} U_{km\ell},$$

$$(4.2) \quad v_0(x) = \sum_{\ell=0}^{\infty} v_{01\ell} U_{01\ell} + \sum_{k=1}^{\infty} \sum_{m=1}^{h(k,p)} \sum_{\ell=1}^{\infty} v_{km\ell} U_{km\ell}.$$

Let $T_1 > T$ (cf. end of § 2) and let $g = g(t) \in C^\infty[0, \infty]$ have support in $[T, T_1]$.

We define a function $g_{km\ell} \in C^\infty(\Omega \otimes [0, \infty))$ by

$$(4.3) \quad g_{km\ell}(x, t) = g(t)U_{km\ell}(x).$$

It is clear that the set of all such functions is complete in $L^2(\Omega \otimes [T, T_1])$.

Let $\varphi(x, t)$ be the solution of

$$\begin{aligned} \frac{\partial^2 \varphi}{\partial t^2} - \Delta \varphi &= g_{km\ell}(x, t), \\ \frac{\partial \varphi}{\partial \nu}(x, t) &\equiv 0, \quad x \in \Gamma, \quad t \in [0, T_1], \end{aligned}$$

with zero terminal data,

$$\varphi(x, T_1) \equiv \frac{\partial \varphi}{\partial t}(x, T_1) \equiv 0.$$

The form of $\varphi(x, t)$ is $\beta(t)U_{km\ell}(x)$, where $\beta(t)$ is the solution of

$$\beta'' + \omega_{k\ell}^2 \beta = g(t), \quad \beta(T_1) = \beta'(T_1) = 0.$$

It may be verified directly that

$$\begin{aligned} \varphi(x, t) &\equiv \left(\int_t^{T_1} \frac{1}{\omega_{k\ell}} \sin(\omega_{k\ell}(\sigma - t)) g_{km\ell}(\sigma) d\sigma \right) U_{km\ell}(x), \\ \frac{\partial \varphi}{\partial t} &\equiv \left(- \int_t^{T_1} \cos(\omega_{k\ell}(\sigma - t)) g_{km\ell}(\sigma) d\sigma \right) U_{km\ell}(x), \end{aligned}$$

or, using trigonometric identities and recalling that the support of $g_{km\ell}$ lies in $[T, T_1]$:

$$(4.4) \quad \varphi(x, t) = \left(\alpha_1 \cos \omega_{k\ell} t + \alpha_2 \frac{1}{\omega_{k\ell}} \sin \omega_{k\ell} t \right) U_{km\ell}(x),$$

$$(4.5) \quad \frac{\partial \varphi}{\partial t}(x, t) = (\alpha_2 \cos \omega_{k\ell} t - \omega_{k\ell} \alpha_1 \sin \omega_{k\ell} t) U_{km\ell}(x),$$

$$\alpha_1 = \int_T^{T_1} \frac{1}{\omega_{k\ell}} \sin(\omega_{k\ell} \sigma) g_{km\ell}(\sigma) d\sigma,$$

$$\alpha_2 = - \int_T^{T_1} \cos(\omega_{k\ell} \sigma) g_{km\ell}(\sigma) d\sigma.$$

(Here, and elsewhere in this paper, for $\omega_{00} = 0$, the conventions $\cos(\omega_{00}t) \equiv 1$, $(1/\omega_{00}) \sin(\omega_{00}t) \equiv t$, will be understood.) From this it is clear that $\varphi(x, t)$ satisfies the hypotheses of Theorem 2.2, in fact, $\varphi \in C^\infty(\Omega \otimes [0, T_1])$.

The solution $w(x, t)$ must then satisfy (2.11) for each $\varphi(x, t)$ of the type constructed above. This yields, upon substitution of (4.4) and (4.5) into (2.11),

$$\begin{aligned}
 (4.6) \quad & \int_{\Omega \otimes [T, T_1]} w(x, t) g_{km\ell}(x, t) dx dt \\
 &= \int_{\Omega} [\alpha_1 U_{km\ell}(x) v_0(x) - \alpha_2 U_{km\ell}(x) w_0(x)] dx \\
 &+ \int_{\Gamma \otimes [0, T]} \left[f(x, t) \left(\alpha_1 \cos(\omega_{k\ell} t) + \frac{\alpha_2}{\omega_{k\ell}} \sin(\omega_{k\ell} t) \right) U_{km\ell}(x) \right] ds dt.
 \end{aligned}$$

But the orthonormality of the $U_{km\ell}$ together with the expansions (4.1), (4.2) gives

$$\begin{aligned}
 (4.7) \quad & \int_{\Omega} U_{km\ell}(x) v_0(x) dx = v_{km\ell}, \quad k = 0, 1, 2, 3, \dots, \quad m = 1, 2, \dots, h(k, p), \\
 & \int_{\Omega} U_{km\ell}(x) w_0(x) dx = w_{km\ell}, \quad \ell = \begin{cases} 0, 1, 2, \dots, & k = 0, \\ 1, 2, 3, \dots, & k > 0. \end{cases}
 \end{aligned}$$

Then (4.6) becomes, for these values of $km\ell$,

$$\begin{aligned}
 & \alpha_1 \left(\int_{\Gamma \otimes [0, T]} f(x, t) \cos(\omega_{k\ell} t) U_{km\ell}(x) ds dt + v_{km\ell} \right) \\
 &+ \alpha_2 \left(\int_{\Gamma \otimes [0, T]} f(x, t) \frac{1}{\omega_{k\ell}} \sin(\omega_{k\ell} t) U_{km\ell}(x) ds dt - w_{km\ell} \right) \\
 &= \int_{\Omega \otimes [T, T_1]} w(x, t) g_{km\ell}(x, t) dx dt.
 \end{aligned}$$

From this it is clear that

$$\int_{\Omega \otimes [T, T_1]} w(x, t) g_{km\ell}(x, t) dx dt = 0,$$

for all functions $g_{km\ell}(x, t)$ as described in (4.3), and hence that (2.13) is satisfied, i.e.,

$$w(x, t) \equiv 0, \quad (x, t) \in \Omega \otimes [T, T_1],$$

if and only if the function $f(x, t)$ solves the moment problem consisting of the equations

$$(4.8) \quad \int_{\Gamma \otimes [0, T]} f(x, t) \cos(\omega_{k\ell} t) U_{km\ell}(x) ds dt = -v_{km\ell},$$

$$(4.9) \quad \int_{\Gamma \otimes [0, T]} f(x, t) \frac{1}{\omega_{k\ell}} \sin(\omega_{k\ell} t) U_{km\ell}(x) ds dt = w_{km\ell}$$

for those values of k, m, ℓ listed under (4.7).

If we expand $f(x, t)$ in terms of the hyperspherical harmonics:

$$(4.10) \quad f(x, t) = \sum_{k=0}^{\infty} \sum_{m=1}^{h(k,p)} f_{km}(t) Y_{km}(\theta, \phi), \quad t \in [0, T], \quad x = (1, \theta, \phi) \in \Gamma,$$

and use the orthonormality of the Y_{km} in $L^2(\Gamma)$ then (4.8), (4.9) reduces to an infinite collection of moment problems for functions $f_{km} \in L^2[0, T]$:

$$(4.11) \quad \int_0^T f_{km}(t) \cos(\omega_{k\ell} t) dt = -\frac{v_{km\ell}}{R_{k\ell}(1)},$$

$$(4.12) \quad \int_0^T f_{km}(t) \frac{1}{\omega_{k\ell}} \sin(\omega_{k\ell} t) dt = \frac{\omega_{km\ell}}{R_{k\ell}(1)},$$

$$\ell = 0, 1, 2, \dots, \quad \text{for } k = 0, \quad m = 1 = h(0, p),$$

$$\ell = 1, 2, 3, \dots, \quad \text{for } k = 1, 2, 3, \dots, \quad m = 1, 2, \dots, h(k, p).$$

We have, then, a doubly indexed set of moment problems; one moment problem for each pair of values (k, m) .

Equations (4.11), (4.12) can be put in the equivalent form:

$$(4.13) \quad \int_0^T f_{km}(t) \exp(i\omega_{k\ell} t) dt = \frac{-v_{km\ell} + i\omega_{k\ell} w_{km\ell}}{R_{k\ell}(1)},$$

$$(4.14) \quad \int_0^T f_{km}(t) \exp(-i\omega_{k\ell} t) dt = \frac{-v_{km\ell} - i\omega_{k\ell} w_{km\ell}}{R_{k\ell}(1)},$$

$$k = 0, 1, 2, \dots, \quad m = 1, 2, \dots, h(k, p), \quad \ell = 1, 2, \dots.$$

For $k = 0$, the value $\ell = 0$ yields the additional equations

$$(4.15) \quad \int_0^T f_{01}(t) dt = -\frac{v_{010}}{R_{00}(1)},$$

$$(4.16) \quad \int_0^T f_{01}(t) dt = \frac{w_{010}}{R_{00}(1)}.$$

5. Solution of abstract moment problems. An abstract moment problem may be posed as follows. Let H be a separable Hilbert space and let $\{p_k | k = 1, 2, 3, \dots\}$ be a sequence of elements of H . Let $\{C_k | k = 1, 2, 3, \dots\}$ be a sequence of scalars

$$(5.1) \quad \sum_{k=1}^{\infty} |C_k|^2 < \infty.$$

With $(\cdot, \cdot)_H$ denoting the scalar product in H , the moment problem consists in finding an element $r \in H$ such that

$$(5.2) \quad (r, p_k)_H = C_k, \quad k = 1, 2, 3, \dots.$$

THEOREM 5.1. *The moment problem (5.2) has a solution $r \in H$ with*

$$(5.3) \quad K_0 \sum_{k=1}^{\infty} |C_k|^2 \leq \|r\|^2 \leq K_1 \sum_{k=1}^{\infty} |C_k|^2$$

for positive constants K_0, K_1 determined by the sequence $\{p_k\}$ (and thus independent of $\{C_k\}$) if and only if

$$(5.4) \quad M_0 \sum_{k=1}^N |\alpha_k|^2 \leq \left\| \sum_{k=1}^N \alpha_k p_k \right\|^2 \leq M_1 \sum_{k=1}^N |\alpha_k|^2$$

for every positive integer N and collection $\alpha_1, \alpha_2, \dots, \alpha_N$ of scalars, where M_0 and M_1 are positive numbers independent of N and the collection $\alpha_1, \alpha_2, \dots, \alpha_N$.

Remark. This is by no means a new theorem. See, e.g., [2]. It is included here so that our presentation will be reasonably self-contained.

Proof. Let $S = S(\{p_k\})$ be the closed subspace of H spanned by $\{p_k\}$. The mapping

$$T: S \rightarrow \ell^2$$

defined on the dense subspace of S consisting of finite linear combinations $\sum_{k=1}^N \alpha_k p_k$ by

$$T\left(\sum_{k=1}^N \alpha_k p_k\right) = (\alpha_1, \alpha_2, \dots, \alpha_N, 0, 0, \dots) \in \ell^2,$$

is seen, from the first inequality in (5.4), to be bounded. Hence T may be extended to all of S and, as a mapping from S to ℓ^2 , satisfies

$$(5.5) \quad \|T\| \leq M_0^{1/2}.$$

The mapping $\hat{T}: \ell^2 \rightarrow S$ defined by

$$\hat{T}(\alpha_1, \alpha_2, \dots, \alpha_N, 0, 0, \dots) = \sum_{k=1}^N \alpha_k p_k$$

is likewise bounded, as we see from the second inequality in (5.4), and may be extended to a bounded linear transformation from ℓ^2 to S with

$$(5.6) \quad \|\hat{T}\| \leq M_1^{1/2}.$$

Since

$$T\hat{T}(\alpha_1, \alpha_2, \dots, \alpha_N, 0, 0, \dots) = (\alpha_1, \alpha_2, \dots, \alpha_N, 0, 0, \dots)$$

and

$$\hat{T}T\left(\sum_{k=1}^N \alpha_k p_k\right) = \sum_{k=1}^N \alpha_k p_k$$

for arbitrary positive N and scalars $\alpha_1, \alpha_2, \dots, \alpha_N$, we conclude that

$$T\hat{T} = I \quad \text{on } \ell^2,$$

$$\hat{T}T = I \quad \text{on } S.$$

Let adjoint operators $T^*: \ell^2 \rightarrow S$, $\hat{T}^*: S \rightarrow \ell^2$ be defined by

$$(r, T^*\alpha)_H = (Tr, \alpha)_{\ell^2},$$

$$(\hat{T}^*r, \alpha)_{\ell^2} = (r, \hat{T}\alpha)_H,$$

where r and α are arbitrary elements of S and ℓ^2 , respectively. It is a general theorem of functional analysis [5] that

$$(5.7) \quad \|T^*\| = \|T\|,$$

$$(5.8) \quad \|\hat{T}^*\| = \|\hat{T}\|.$$

For $k = 1, 2, \dots$, let

$$q_k = T^* e_k,$$

where $e_k = (0, \dots, 0, 1, 0, 0, \dots) \in \ell^2$, the “1” appearing in the k th position. Combining (5.5) and (5.7) we have

$$\|q_k\|_H \leq M_0^{1/2} \|e_k\|_{\ell^2} = M_0^{1/2}.$$

Let

$$r = \sum_{\ell=1}^{\infty} C_{\ell} q_{\ell} = T^*(C_1, C_2, C_3, \dots).$$

Since $(C_1, C_2, C_3, \dots) \in \ell^2$ by (5.3), r is well-defined. Then

$$\begin{aligned} (r, p_k)_H &= \sum_{\ell=1}^{\infty} C_{\ell} (q_{\ell}, p_k)_H \\ &= \sum_{\ell=1}^{\infty} C_{\ell} (T^* e_{\ell}, \hat{T} e_k)_H = \sum_{\ell=1}^{\infty} C_{\ell} (e_{\ell}, e_k)_{\ell^2} = C_k. \end{aligned}$$

Moreover,

$$(5.9) \quad \|r\|^2 \leq \|T^*\|^2 \|(C_1, C_2, C_3, \dots)\|_{\ell^2}^2 \leq M_0 \sum_{\ell=1}^{\infty} |C_{\ell}|^2,$$

so we may take $K_1 = M_0$. Since

$$(5.10) \quad (C_1, C_2, C_3, \dots) = \hat{T}^* r, \quad \sum_{\ell=1}^{\infty} |C_{\ell}|^2 \leq \|\hat{T}^*\|^2 \|r\|^2 \leq M_1 \|r\|^2$$

and we may take $K_0 = M_1$. This proves the “if” part of the theorem. The “only if” part is proved by interchanging the roles of $\{p_k\}$ and $\{q_k\}$, $\{C_k\}$ and $\{\alpha_k\}$, respectively, assuming (5.9) and (5.10) and arriving at (5.3).

Remark. The elements q_k constitute a *biorthogonal* set in S for the elements p_k in the sense that

$$(q_{\ell}, p_k) = \delta_{k\ell}, \quad k = 1, 2, 3, \dots, \quad \ell = 1, 2, 3, \dots$$

The requirement (5.3) establishes $\{p_k\}$ as a *Riesz basis* for S . The sequence $\{q_{\ell}\}$ is then the *adjoint* Riesz basis for S relative to $\{p_k\}$.

6. Solution of the moment problems (4.13)–(4.16). In this section we fix k, m and consider the moment problem

$$(6.1) \quad \int_0^T f_{km}(t) \exp(i\omega_{k\ell} t) dt = c_{km\ell},$$

$$(6.2) \quad \int_0^T f_{km}(t) \exp(-i\omega_{k\ell} t) dt = d_{km\ell},$$

where the definitions of $c_{km\ell}$ and $d_{km\ell}$ are clear from (4.13), (4.14). For $k = 0$, $m = 1$ we have two additional equations:

$$(6.3) \quad \int_0^T f_{01}(t) dt = c_{010},$$

$$(6.4) \quad \int_0^T f_{01}(t)t dt = d_{010}.$$

Whether or not such a moment problem is solvable depends decisively on the value of T and the properties of the nonnegative numbers $\omega_{k\ell}$ which, we recall from § 3, are the ℓ th positive roots of

$$(-p/2)J_{k+p/2}(\omega) + \omega J'_{k+p/2}(\omega) = 0.$$

A major section of [7], which is a more detailed treatment of the problem considered in this paper, is devoted to examination of the sequences $\{\omega_{k\ell}\}$. We refer the reader to that work¹ for proof of the following lemma.

LEMMA 6.1. *For fixed $p \geq 0$, and $k = 1, 2, 3, \dots$,*

$$(6.5) \quad \omega_{k,\ell+1} - \omega_{k\ell} > \pi, \quad \ell = 1, 2, 3, \dots,$$

and for $k = 0, 1, 2, 3, \dots$,

$$(6.6) \quad \lim_{\ell \rightarrow \infty} (\omega_{k,\ell+1} - \omega_{k\ell}) = \pi.$$

When $k = 0$, (6.5) holds for $\ell = 0, 1, 2, \dots$.

Lemma 6.1 implies that for $k = 0, 1, 2, 3, \dots$, the sequence $\{\omega_{k\ell}\}$ has an asymptotic gap π , as expressed by (6.6), and density

$$(6.7) \quad D \equiv \lim_{\ell \rightarrow \infty} \frac{\ell}{\omega_{k\ell}} = \frac{1}{\pi}.$$

Also, for $k \geq 0$, the gap $\omega_{k,\ell+1} - \omega_{k\ell}$ is uniformly bounded below by π , as expressed in (6.5).

Moment problems (6.1), (6.2) with $\omega_{k\ell}$ satisfying properties of the type expressed in Lemma 6.1 have been studied over a long period by many eminent mathematicians. (See [1, p. 75], [2], [4], [10], [11], [12], [16], [17], [18], [25].) These results are applied to a simpler version of the present problem in [19].

LEMMA 6.2. *If $T < 2$, the moment problem (6.1), (6.2) (or (6.1), (6.2), (6.3), (6.4) if $k = 0, m = 1$) has no solution in general.*

This is a special case of a result proved by Levinson [16, p. 3] who shows that the functions $\exp(i\omega_{k\ell}t)$, $\exp(-i\omega_{k\ell}t)$, $\ell = 1, 2, 3, \dots$, are linearly dependent (indeed, each function $\exp(i\omega_{k\ell}t)$ or $\exp(-i\omega_{k\ell}t)$ lies in the closed span of the other exponentials) in $L^2[0, T]$ if $T < 2\pi D$, D expressed as in (6.7). Since $D = 1/\pi$ in our case, Lemma 6.2 follows from this result.

¹ In [8], the same results are proved for the positive roots of

$$(-p/2 + \sigma)J_{k+p/2}(\omega) + \omega J'_{k+p/2}(\omega) = 0,$$

$0 \leq \sigma < \infty$. This condition defines the eigenvalues for the elastic boundary condition

$$\frac{\partial u}{\partial v} + \sigma u = 0.$$

LEMMA 6.3. Let $k \geq 1$, $1 \leq m \leq h(k, p)$, and suppose the sequences $\{c_{km\ell}|\ell = 1, 2, 3, \dots\}$, $\{d_{km\ell}|\ell = 1, 2, 3, \dots\}$ are square summable. Then, given $T > 2$, the moment problem (6.1), (6.2) has a solution $f_{km} \in L^2[0, T]$ with

$$(6.8) \quad \tilde{K}_0 \sum_{\ell=1}^{\infty} (|c_{km\ell}|^2 + |d_{km\ell}|^2) \leq \|f_{km}\|_{L^2[0, T]}^2 \leq \tilde{K}_1 \sum_{\ell=1}^{\infty} (|c_{km\ell}|^2 + |d_{km\ell}|^2),$$

where \tilde{K}_0 and \tilde{K}_1 are constants determined by the uniform lower bound π on the gap $\omega_{k, \ell+1} - \omega_{k\ell}$ and the positive number $T - 2$. (Hence \tilde{K}_0, \tilde{K}_1 are independent of k, m , and the particular sequences $\{c_{km\ell}\}, \{d_{km\ell}\}$.)

Proof. A result of Ingham [10] shows that when (6.5) holds and

$$T > 2 = 2\pi \left(\frac{1}{\pi} \right) \geq 2\pi \left(\frac{1}{\min_{\ell} (\omega_{k, \ell+1} - \omega_{k\ell})} \right),$$

there are positive numbers M_0 and M_1 , determined by the lower bound π for $\omega_{k, \ell+1} - \omega_{k\ell}$ and the positive number $T - 2$, such that for any scalars $\alpha_{-N}, \alpha_{-N+1}, \dots, \alpha_{-1}, \alpha_1, \dots, \alpha_{N-1}, \alpha_N$,

$$M_0 \sum_{\substack{\ell=-N \\ \ell \neq 0}}^N |\alpha_{\ell}|^2 \leq \int_0^T \left| \sum_{\ell=-N}^N \alpha_{\ell} \exp \left(\frac{\ell}{|\ell|} i \omega_{k|\ell|} t \right) \right|^2 dt \leq M_1 \sum_{\substack{\ell=-N \\ \ell \neq 0}}^N |\alpha_{\ell}|^2.$$

The lemma is then a direct consequence of Theorem 5.1 with $H = L^2[0, T]$ and S the subspace of $L^2[0, T]$ spanned by the functions $\exp(\pm i \omega_{k\ell} t)$, $\ell = 1, 2, 3, \dots$.

LEMMA 6.4. Let $k = 0, m = 1$ and suppose that the sequences $\{c_{01\ell}|\ell = 0, 1, 2, \dots\}$, $\{d_{01\ell}|\ell = 0, 1, 2, \dots\}$ are square summable. Then, given $T > 2$, the moment problem (6.1), (6.2), (6.3), (6.4) has a solution $f_{01} \in L^2[0, T]$ with

$$(6.9) \quad \hat{K}_0 \sum_{\ell=0}^{\infty} (|c_{01\ell}|^2 + |d_{01\ell}|^2) \leq \|f_{01}\|_{L^2[0, T]}^2 \leq \hat{K}_1 \sum_{\ell=0}^{\infty} (|c_{01\ell}|^2 + |d_{01\ell}|^2),$$

where \hat{K}_0, \hat{K}_1 are positive numbers determined by the asymptotic gap π (as expressed in (6.6)), and the positive number $T - 2$. (Hence \hat{K}_0, \hat{K}_1 are independent of the particular sequences $\{c_{01\ell}\}, \{d_{01\ell}\}$.)

The proof of this lemma is given in [19], with more detail being supplied in [7].

As a consequence of these lemmas, we are able to prove the following theorem concerning solutions of the moment problem (4.8), (4.9).

THEOREM 6.5. If $T > 2$ and

$$(6.10) \quad \sum_{\ell=0}^{\infty} |v_{01\ell}|^2 + \sum_{k=1}^{\infty} \sum_{m=1}^{h(k, p)} \sum_{\ell=1}^{\infty} |v_{km\ell}|^2 \equiv v^2 < \infty,$$

$$(6.11) \quad \sum_{\ell=0}^{\infty} |\omega_{0\ell} w_{01\ell}|^2 + \sum_{k=1}^{\infty} \sum_{m=1}^{h(k, p)} \sum_{\ell=1}^{\infty} |\omega_{k\ell} w_{km\ell}|^2 \equiv \mu^2 < \infty,$$

then the moment problem (4.8), (4.9) has a solution $f(x, t) \in L^2(\Gamma \otimes [0, T])$ and

$$(6.12) \quad \|f\|_{L^2(\Gamma \otimes [0, T])}^2 \leq K(v^2 + \mu^2),$$

where K is a positive constant independent of the coefficients $v_{km\ell}, w_{km\ell}$.

Proof. For $k = 0, 1, 2, \dots, m = 1, 2, \dots, h(k, p)$, let the functions $f_{km} \in L^2[0, T]$ be the solutions of the moment problems (6.1), (6.2) (and (6.3), (6.4) if $k = 0$) which we have seen to exist for $T > 2$. For each km we have, from (6.8), (4.13), (4.14),

$$(6.13) \quad \|f_{km}\|_{L^2[0, T]}^2 \leq 2\tilde{K}_1 \sum_{\ell=1}^{\infty} \frac{1}{|R_{k\ell}(1)|^2} (|v_{km\ell}|^2 + |\omega_{k\ell} w_{km\ell}|^2).$$

For $k = 0, m = 1$ the sum runs from $\ell = 0$ to ∞ and \tilde{K}_1 is replaced by \hat{K}_1 . Defining $f(x, t)$ by (4.10) and noting the definition (3.10), (3.11), (3.12) of the $U_{km\ell}$, we see immediately that the orthonormality of the hyperspherical harmonics $Y_{km}(\theta, \phi)$ implies that $f(x, t)$ is, at least formally, a solution of the moment problem (4.8), (4.9). The orthonormality of the $Y_{km}(\theta, \phi)$ also shows that for any fixed positive integers k_1, k_2 ,

$$(6.14) \quad \left\| \sum_{k=k_1}^{k_2} \sum_{m=1}^{h(k, p)} f_{km}(t) Y_{km}(\theta, \phi) \right\|_{L^2(\Gamma \otimes [0, T])}^2 = \sum_{k=k_1}^{k_2} \sum_{m=1}^{h(k, p)} \|f_{km}\|_{L^2[0, T]}^2 \\ \leq 2\tilde{K}_1 \sum_{k=k_1}^{k_2} \sum_{m=1}^{h(k, p)} \sum_{\ell=1}^{\infty} \frac{1}{|R_{k\ell}(1)|^2} (|v_{km\ell}|^2 + |\omega_{k\ell} w_{km\ell}|^2).$$

Since it is clear from (3.7) and (3.15) that

$$R_{k\ell}(1) \geq \sqrt{2},$$

we see that there is a positive number K such that, for all k, ℓ ,

$$2\tilde{K}_1 \frac{1}{|R_{k\ell}(1)|^2} \leq K.$$

Thus the left-hand side of (6.14) is bounded by

$$K \sum_{k=k_1}^{k_2} \sum_{m=1}^{h(k, p)} \sum_{\ell=1}^{\infty} (|v_{km\ell}|^2 + |\omega_{k\ell} w_{km\ell}|^2) \rightarrow 0 \quad \text{as } k_1, k_2 \rightarrow \infty,$$

as we see from (6.10), (6.11). Thus the series (4.10) converges in $L^2(\Gamma \otimes [0, T])$ and, in view of our previous remarks, represents a bona fide solution of the moment problem (4.8), (4.9). Letting $k_1 \rightarrow 0, k_2 \rightarrow \infty$ we obtain (6.12).

7. The controllable states: Proof of Theorem 1.1. Combining Theorem 6.5 with the results of § 4, we see that the set of initial states $w_0(x), v_0(x)$ which can be steered to the state 0, 0 with a control $f \in L^2[0, T]$, $T > 2$, includes those states with expansions (4.1), (4.2) whose coefficients satisfy (6.10), (6.11). To characterize these states more meaningfully, we will show that the requirements (6.10), (6.11) are equivalent to certain differentiability requirements. In the process we shall obtain the proof of Theorem 1.1.

The condition (6.10) simply indicates that v_0 has norm $v^2 < \infty$ in $L^2(\Omega)$ and hence lies in that space. No further analysis is required here. Thus we confine ourselves to the problem of determining which functions w_0 satisfy (6.11).

The lower bound (3.15) together with the well-known properties of the zeros $\omega_{0\ell}$ of $J_0(r)$ combine with (6.11) to imply

$$\sum_{\ell=0}^{\infty} |w_{01\ell}|^2 + \sum_{k=1}^{\infty} \sum_{m=1}^{h(k,p)} \sum_{\ell=1}^{\infty} |w_{km\ell}|^2 < \infty,$$

so that w_0 also lies in $L^2(\Omega)$.

LEMMA 7.1. *If $w_0 \in H^1(\Omega)$, then the inequality (6.11) is satisfied.*

Proof. If $w_0 \in H^1(\Omega)$, then $w_0 \in L^2(\Omega)$ and has an expansion (3.13). The orthonormality of the $U_{km\ell}$ in $L^2(\Omega)$ implies that (cf. (3.14))

$$(7.1) \quad w_{km\ell} = \int_{\Omega} w_0(x) U_{km\ell}(x) dx = -\frac{1}{\lambda_{k\ell}} \int_{\Omega} w_0(x) \Delta U_{km\ell}(x) dx,$$

the latter identity following from (3.1) and the definition of the $\lambda_{k\ell}$. Since $w_0 \in H^1(\Omega)$, the trace theorem [14] shows that the restriction of w_0 to $\Gamma = \partial\Omega$ lies in $H^{1/2}(\Gamma) \subset L^2(\Gamma)$. Since the eigenfunctions U_{kml} have derivatives of all orders we may compute (with ∇ denoting the gradient of the indicated function with respect to x)

$$(7.2) \quad \begin{aligned} & \int_{\Omega} [w_0(x) \Delta U_{km\ell}(x) + \nabla w_0(x) \cdot \nabla U_{km\ell}(x)] dx \\ &= \int_{\Omega} \operatorname{div} [w_0(x) \nabla U_{km\ell}(x)] dx \\ &= \int_{\Gamma} w_0(x) (\nabla U_{km\ell}(x) \cdot \nu(x)) ds \\ &= \int_{\Gamma} \left(w_0(x) \frac{\partial U_{km\ell}}{\partial \nu}(x) \right) ds = 0, \end{aligned}$$

where $\nu(x)$ denotes the unit exterior (with respect to Ω) normal vector to Γ at a point $x \in \Gamma$, ds is surface (volume) measure on Γ and $\partial/\partial \nu$ is the exterior normal derivative. The last equality follows from (3.1). Combining (7.1) and (7.2) we have

$$(7.3) \quad w_{km\ell} = \frac{1}{\lambda_{k\ell}} \int_{\Omega} \nabla w_0(x) \cdot \nabla U_{km\ell}(x) dx.$$

Let $L_n^2(\Omega)$ denote the space of n -dimensional vector functions

$$y(x) = \begin{pmatrix} y^1(x) \\ y^2(x) \\ \vdots \\ y^n(x) \end{pmatrix}, \quad y^i \in L^2(\Omega), \quad i = 1, 2, \dots, n.$$

With the inner product

$$\langle y, z \rangle_{L_n^2(\Omega)} \equiv \sum_{i=1}^n \langle y^i, z^i \rangle_{L^2(\Omega)},$$

$L_n^2(\Omega)$ is easily seen to be a Hilbert space.

We now show that the functions

$$\frac{1}{\omega_{k\ell}} \nabla U_{km\ell}(x),$$

with the exception of $k = 0, m = 1, \ell = 0$, form an orthonormal (but not complete) set in $L_n^2(\Omega)$ for

$$\begin{aligned} & \left\langle \frac{1}{\omega_{k\ell}} \nabla U_{km\ell}, \frac{1}{\omega_{\hat{k}\hat{\ell}}} \nabla U_{\hat{k}\hat{m}\hat{\ell}} \right\rangle_{L_n^2(\Omega)} \\ &= \frac{1}{\omega_{k\ell} \omega_{\hat{k}\hat{\ell}}} \int_{\Omega} [\operatorname{div} (U_{\hat{k}\hat{m}\hat{\ell}}(x) \nabla U_{km\ell}(x)) - U_{\hat{k}\hat{m}\hat{\ell}}(x) \Delta U_{km\ell}(x)] dx \\ &= \frac{1}{\omega_{k\ell} \omega_{\hat{k}\hat{\ell}}} \left\{ \int_{\Gamma} U_{\hat{k}\hat{m}\hat{\ell}}(x) \frac{\partial D_{km\ell}}{\partial \nu}(x) ds - \int_{\Omega} U_{\hat{k}\hat{m}\hat{\ell}} \Delta U_{km\ell}(x) dx \right\}. \end{aligned}$$

From (3.1) the integral over Γ is zero and

$$\Delta U_{km\ell}(x) \equiv -\lambda_{k\ell} U_{km\ell}(x), \quad x \in \Omega.$$

The orthonormality of the $U_{km\ell}$ in $L^2(\Omega)$ then gives

$$\begin{aligned} & \left\langle \frac{1}{\omega_{k\ell}} \nabla U_{km\ell}, \frac{1}{\omega_{\hat{k}\hat{\ell}}} \nabla U_{\hat{k}\hat{m}\hat{\ell}} \right\rangle_{L_n^2(\Omega)} \\ &= \frac{\lambda_{k\ell}}{\omega_{k\ell} \omega_{\hat{k}\hat{\ell}}} \int_{\Omega} U_{\hat{k}\hat{m}\hat{\ell}}(x) U_{km\ell}(x) dx = \frac{\delta_{\hat{k}\hat{m}\hat{\ell}}^{km\ell} \lambda_{k\ell}}{\omega_{k\ell} \omega_{\hat{k}\hat{\ell}}} \\ &= \begin{cases} 1 & \text{if } k = \hat{k}, m = \hat{m}, \ell = \hat{\ell}, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Now the “Fourier coefficients” of $\nabla w_0(x)$ with respect to the functions $\nabla U_{km\ell}(x)$ in $L_n^2(\Omega)$ must, by virtue of Bessel’s inequality, be square summable. Those coefficients are

$$\begin{aligned} a_{km\ell} &= \left\langle \nabla w_0(x), \frac{1}{\omega_{k\ell}} \nabla U_{km\ell}(x) \right\rangle_{L_n^2(\Omega)} = \frac{1}{\omega_{k\ell}} \int_{\Omega} \nabla w_0(x) \cdot \nabla U_{km\ell}(x) dx \\ &\quad \text{(from (7.3))} \\ &= \omega_{k\ell} w_{km\ell}. \end{aligned}$$

Thus the square summability of the $a_{km\ell}$ implies the finiteness of the sum (6.11) and the proof of Lemma 7.1 is complete.

The proof of Theorem 1.1 is also complete at this stage since we have shown that if the initial state w_0, v_0 satisfies (1.7), then the inequalities (6.10) and (6.11) are satisfied. Theorem 6.5 then shows that the moment problem (4.8), (4.9) is solvable for $f \in L^2(\Gamma \otimes [0, T])$, $T > 2$, and, as we have seen in § 4, this is equivalent to controllability.

8. Proof of Theorems 1.2 and 1.3. Before proving these theorems we present a lemma indicating that Theorem 1.1 is, indeed, all that can be obtained, in general, from the use of inequalities of the form (6.10), (6.11).

LEMMA 8.1. *If the inequalities (6.10), (6.11) are satisfied, then w_0, v_0 , as given by (4.1), (4.2), satisfy (1.7).*

Proof. As we have already noted in § 7, (6.10) is equivalent to

$$v_0 \in L^2(\Omega) = H^0(\Omega).$$

Now (6.11) clearly implies, since $\omega_{k\ell}^2 = \lambda_{k\ell}$,

$$\sum_{\ell=0}^{\infty} |\sqrt{\lambda_{0\ell} + 1} w_{01\ell}|^2 + \sum_{k=1}^{\infty} \sum_{m=1}^{h(k,p)} \sum_{\ell=1}^{\infty} |\sqrt{\lambda_{k\ell} + 1} w_{km\ell}|^2 < \infty,$$

which implies that w_0 , as given by (4.1), lies in the domain of the positive operator

$$(8.1) \quad (-\Delta + I)^{1/2}$$

in $L^2(\Omega)$. To study this domain we first consider the operator

$$(8.2) \quad -\Delta + I.$$

In [14, Chap. 2, § 6] we find a theorem which, in our case, specializes to the following statement: the operator (8.2) defines an isomorphism of V , as described in § 2, onto $L^2(\Omega) = H^0(\Omega)$. That is, given $\tilde{\omega} \in V$, there is exactly one $\hat{\omega} \in L^2(\Omega)$ such that

$$(8.3) \quad \hat{\omega} = (-\Delta + I)\tilde{\omega}$$

and, given $\hat{\omega} \in L^2(\Omega)$, there is exactly one $\tilde{\omega} \in V$ such that (8.3) holds. Moreover, there are constants L_0, L_1 such that

$$L_0 \|\tilde{\omega}\|_{H^2(\Omega)} \leq \|\hat{\omega}\|_{L^2(\Omega)} \leq L_1 \|\tilde{\omega}\|_{H^2(\Omega)}.$$

In [14], where the theory of interpolation spaces is developed, the domain of the operator (8.1) is identified as the space

$$[V, L^2(\Omega)]_{1/2} \subset [H^2(\Omega), L^2(\Omega)]_{1/2} = H^1(\Omega).$$

Indeed, $[V, L^2(\Omega)]_{1/2}$ is there shown to be the domain of $A^{1/2}$, where A is any positive self-adjoint operator on $L^2(\Omega)$ whose domain coincides with V and whose range is $L^2(\Omega)$.

Thus we have seen that (6.11) implies w_0 lies in the domain of (8.1) which in turn is contained in $H^1(\Omega)$. Hence $w_0 \in H^1(\Omega)$ and the proof of Lemma 8.1 is complete.

The point of Theorem 1.2 is that the inequalities (6.10), (6.11) are more restrictive than the inequality

$$\sum_{\ell=0}^{\infty} (|c_{01\ell}|^2 + |d_{01\ell}|^2) + \sum_{k=1}^{\infty} \sum_{m=1}^{h(k,p)} \sum_{\ell=1}^{\infty} (|c_{km\ell}|^2 + |d_{km\ell}|^2) < \infty$$

(cf. (4.13), (4.14), (4.15), (4.16), (6.1), (6.2), (6.3), (6.4)) which is required for solvability of the moment problem (4.8), (4.9) for $f \in L^2(\Gamma \otimes [0, T])$, $T > 2$. Consider, for example, the initial state given by

$$(8.4) \quad w_0(x) \equiv 0,$$

$$(8.5) \quad v_0(x) \equiv \sum_{k=1}^{\infty} \sum_{m=1}^{h(k,p)} -\alpha_k(p) R_{k1}(1) U_{km1}(x),$$

where

$$\alpha_k(p) = \left(\frac{1}{h(k,p)k^{1+\varepsilon}} \right)^{1/2}, \quad k = 1, 2, 3, \dots$$

The expansion coefficients $w_{km\ell}$ for w_0 are all zero and those of v_0 are given by

$$\begin{aligned} v_{01\ell} &= 0, & \ell &= 0, 1, 2, \dots, \\ v_{km1} &= -\alpha_k R_{k1}(1), & k &= 1, 2, \dots, \quad m = 1, 2, \dots, h(k,p), \\ v_{km\ell} &= 0, & k &= 1, 2, \dots, \quad m = 1, 2, \dots, h(k,p), \quad \ell = 2, 3, \dots \end{aligned}$$

For the moment problem (6.1), (6.2) we then have (cf. (4.13), (4.14))

$$\begin{aligned} c_{01\ell} &= d_{01\ell} = 0, & \ell &= 0, 1, 2, \dots, \\ c_{km1} &= d_{km1} = -\frac{v_{km1}}{R_{k1}(1)} = \alpha_k, & k &= 1, 2, 3, \dots, \quad m = 1, \dots, h(k,p), \\ c_{km\ell} &= d_{km\ell} = 0, & k &= 1, 2, 3, \dots, \quad m = 1, \dots, h(k,p), \\ & & \ell &= 2, 3, \dots \end{aligned}$$

Thus, for the control $f(x, t)$ to lie in $L^2(\Gamma \otimes [0, T])$, $T > 2$, we see from Lemma 6.4 and reasoning similar to that used in Theorem 6.5 that it is sufficient to have

$$2 \sum_{k=1}^{\infty} \sum_{m=1}^{h(k,p)} \alpha_k(p)^2 = 2 \sum_{k=1}^{\infty} \frac{1}{k^{1+\varepsilon}} < \infty,$$

which is true for any $\varepsilon > 0$. But then the coefficients in the expansion (8.5), i.e., the numbers $-\alpha_k(p) R_{k1}(1)$, have, using the asymptotic relationship (3.9), the form

$$-\alpha_k(p) R_{k1}(1) = O\left(\frac{1}{h(k,p)k^{1/3+\varepsilon}} \right).$$

Hence, for some constant $M > 0$,

$$\sum_{k=1}^{\infty} \sum_{m=1}^{h(k,p)} (\alpha_k(p) R_{k1}(1))^2 \geq M \sum_{k=1}^{\infty} \frac{1}{k^{2/3+2\varepsilon}} = \infty$$

if we take

$$\varepsilon \leq 1/6.$$

Thus, for such a choice of ε , (8.4), (8.5) represents a state w_0, v_0 for which $v_0 \notin L^2(\Omega)$ but which is controllable in the sense that the moment problem (4.8), (4.9) has a solution $f \in L^2(\Gamma \otimes [0, T])$. This then provides the proof of Theorem 1.2.

We pass, finally, to the proof of Theorem 1.3. Since

$$R_{k\ell}(1) = \left[\frac{\omega_{k\ell}^2}{\omega_{k\ell}^2 - k(k+p)} \right]^{1/2}$$

and since $\omega_{k\ell} > \omega_{k1}$ for $\ell > 1$, we have

$$(8.6) \quad R_{k\ell}(1) < R_{k1}(1), \quad k = 0, 1, 2, \dots, \quad \ell = 2, 3, \dots$$

Now it is known [26, p. 521], that for the case $p = 0$ (i.e., $n = 2$),

$$\omega_{k\ell} = k + \gamma k^{1/3} + O(k^{-1/3})$$

so that

$$(8.7) \quad |R_{k1}(1)|^2 = \frac{(k + \gamma k^{1/3} + O(k^{-1/3}))^2}{(k + \gamma k^{1/3} + O(k^{-1/3}))^2 - k^2} = \frac{k^2 + O(k^{4/3})}{2\gamma k^{4/3} + O(k)}.$$

Combining (8.6), (8.7) and (3.15) we have

$$|R_{k\ell}(1)|^2 \leq M k^{2/3} \leq M (\omega_{k\ell})^{2/3}$$

for all k, ℓ , where M is a fixed positive number. This means

$$\frac{1}{|R_{k\ell}(1)|^2} \geq \frac{1}{M} \omega_{k\ell}^{-2/3}.$$

Noting that the condition for controllability is the square summability of the coefficients appearing on the right-hand sides of (4.13)–(4.16), we see that for the control f to lie in $L^2(\Gamma \otimes [0, T])$, $T > 2$, $p = 0$ (i.e., $n = 2$) we must have

$$(8.8) \quad \sum_{k\ell m} (\omega_{k\ell})^{-2/3} |v_{km\ell}|^2 < \infty,$$

$$(8.9) \quad \sum_{k\ell m} (\omega_{k\ell})^{4/3} |w_{km\ell}|^2 < \infty.$$

Using the theory of the spaces $H^r(\Omega)$ as defined for all real r in [14] and the relationships between those spaces and the Laplacian operator Δ , also developed there, (8.8) and (8.9) are seen to imply

$$v_0 \in H^{-1/3}(\Omega), \quad w_0 \in H^{2/3}(\Omega),$$

and the proof of Theorem 1.3 is complete.

REFERENCES

- [1] S. BANACH, *Theorie des opérations linéaires*, Warsaw, 1932.
- [2] R. P. BOAS, JR., *A trigonometric moment problem*, J. London Math. Soc., 14 (1939), pp. 242–244.
- [3] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, vol. I, John Wiley, New York, 1953.
- [4] R. J. DUFFIN AND J. J. EACHUS, *Some notes on an expansion theorem of Paley and Wiener*, Bull. Amer. Math. Soc., 48 (1942), pp. 850–855.
- [5] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators*, vol. I, Interscience, New York, 1958.
- [6] A. ERDÉLYI, W. MAGNUS, F. OBERHETTINGER AND F. G. TRICOMI, *Higher Transcendental Functions*, vol. II, McGraw-Hill, New York, 1953.
- [7] K. D. GRAHAM, *On boundary value control of symmetric hyperbolic systems*, Ph.D. thesis, University of Minnesota, Minneapolis, 1973.

- [8] —, *Separation of eigenvalues of the wave equation for the unit ball in R^N* , Studies in Appl. Math., to appear.
- [9] C. J. HERGET, *On the controllability of distributed parameter systems*, Internat. J. Control, 11 (1970), no. 5, pp. 827–833.
- [10] A. E. INGHAM, *Some trigonometrical inequalities in the theory of series*, Math. Z., 41 (1936), pp. 367–379.
- [11] M. I. KADEC, *The exact value of the Paley-Wiener constant*, Sov. Math. Dokl., 5 (1964), no. 2, pp. 559–561.
- [12] N. LEVINSON, *Gap and Density Theorems*, Colloquium Publications, vol. 26, American Mathematical Society, New York, 1940.
- [13] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.
- [14] J. L. LIONS AND E. MAGENES, *Problèmes aux limites non homogènes et applications*, vol. 1, Dunod, Paris, 1968.
- [15] —, *Problèmes aux limites non homogènes et applications*, vol. 2, Dunod, Paris, 1968.
- [16] R. E. A. C. PALEY AND N. WIENER, *The Fourier Transform in the Complex Domain*, Colloquium Publications, vol. 19, American Mathematical Society, New York, 1934.
- [17] R. M. REDHEFFER, *Remarks on the incompleteness of $\{e^{i\lambda_n x}\}$, non-averaging sets, and entire functions*, Proc. Amer. Math. Soc., 2 (1951), pp. 365–369.
- [18] F. RIESZ AND B. SZ-NAGY, *Functional Analysis*, Frederick Ungar, New York, 1955.
- [19] D. L. RUSSELL, *Nonharmonic Fourier series in the control of distributed parameter systems*, J. Math. Anal. Appl., 18 (1967), pp. 542–560.
- [20] —, *On boundary-value controllability of linear symmetric hyperbolic systems*, Mathematical Theory of Control, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, New York, 1967.
- [21] —, *Boundary value control of the higher dimensional wave equation*, this Journal, 9 (1971), pp. 29–42.
- [22] —, *Boundary value control of the higher dimensional wave equation, Part II*, this Journal, 9 (1971), pp. 401–419.
- [23] —, *Control theory of hyperbolic equations related to certain questions in harmonic analysis and Spectral theory*, J. Math. Anal. Appl., 40 (1972), pp. 336–368.
- [24] —, *A unified boundary controllability theory for hyperbolic and parabolic partial differential equations*, Studies in Appl. Math., LII (1973), pp. 189–211.
- [25] L. SCHWARTZ, *Approximation d'une fonction quelconque*, Ann. Fac. Sci. Univ. Toulouse (43), 6 (1942), pp. 111–174.
- [26] G. N. WATSON, *A Treatise on the Theory of Bessel Functions*, 2nd ed., Cambridge Univ. Press, London, 1958.

PLANE MOTION OF A PARTICLE SUBJECT TO CURVATURE CONSTRAINTS*

E. J. COCKAYNE AND G. W. C. HALL†

Abstract. A particle P moves in the plane with constant speed and subject to an upper bound on the curvature of its path. This paper studies the classes of trajectories by which P can reach a given point in a given direction and obtains, for all t , the set $R(t)$ of all possible positions for P at time t , thus extending the results of several recent authors.

1. Introduction. An object P moves in the plane with constant speed v . If P is allowed to move freely, then its region of accessibility at time t , i.e., the set of possible positions for P at time t , is the closed disc $D(t)$ with center the initial position and radius vt . Suppose now that P cannot turn too sharply, i.e., there is an upper bound on the curvature of its path. Clearly this constraint will restrict P at time t to some proper subset of $D(t)$. This paper will be concerned with two main questions. First, given a point Q in the plane, at what times and by what kinds of trajectories may Q be reached by P ? Second, we shall determine the region of accessibility of P for all t .

We now give a more precise formulation of the problems. Let $Z(s)$ be a differentiable curve in the plane of arbitrary length $\sigma > 0$ parametrized by arc length, and let the tangent to $Z(s)$ at arc length s make angle $\phi(s)$ with the positive y -axis. (We assume $\phi(s)$ is positive if the rotation from the positive y -axis to a ray along the tangent is anticlockwise.) Then $Z(s)$ is an *admissible curve* if

- (i) $\phi(0) = 0$, $Z(0) = (0, 0)$,
- (ii) $\phi(s)$ is continuous on $[0, \sigma]$,
- (iii) $|\phi(s_2) - \phi(s_1)| \leq (1/\rho)|s_2 - s_1|$ for all $s_1, s_2 \in [0, \sigma]$, $\rho > 0$.

Condition (iii) imposes the curvature restriction on P . We note that admissible curves need not have curvature at every point. For example, a smooth union of a line segment and a circular arc of radius greater than or equal to ρ is admissible but does not have curvature at the point of tangency. Admissible curves do, however, possess curvature bounded by $1/\rho$ almost everywhere. The Lipschitz condition is used initially because certain closure properties of the class of curves will be needed. It will be seen later that for accessibility purposes, smaller and more practical families of curves are equivalent.

We shall need the following definitions and notations. $\mathcal{L}(z, \alpha, l)$ will denote the set of admissible curves of fixed length l which terminate at z in the direction α , i.e., the set of admissible curves for which ϕ is defined on $[0, l]$, $Z(l) = z$ and $\phi(l) = \alpha$. If any of the arguments is replaced by a $*$, then it is assumed that this argument is arbitrary. For example, $\mathcal{L}(z, \alpha, *)$ is the set of admissible curves of arbitrary length which terminate at z in direction α .

To illustrate this notation we note that at any point of any curve in $\mathcal{L}(z, \alpha, l)$, we may adjoin a circle of radius $r \geq \rho$ and thus construct a curve belonging to $\mathcal{L}(z, \alpha, l + 2\pi r)$.

* Received by the editors October 31, 1972, and in revised form August 30, 1973.

† Department of Mathematics, University of Victoria, British Columbia, Canada. This paper was completed while the first author was at the Mathematical Institute, Oxford University. The work of the first author was supported by the Canadian National Research Council under Grant NRC A 7544.

A *critical admissible curve* consists of unions of line segments and circular arcs of radius ρ . The following example and Fig. 1 will illustrate further notation.

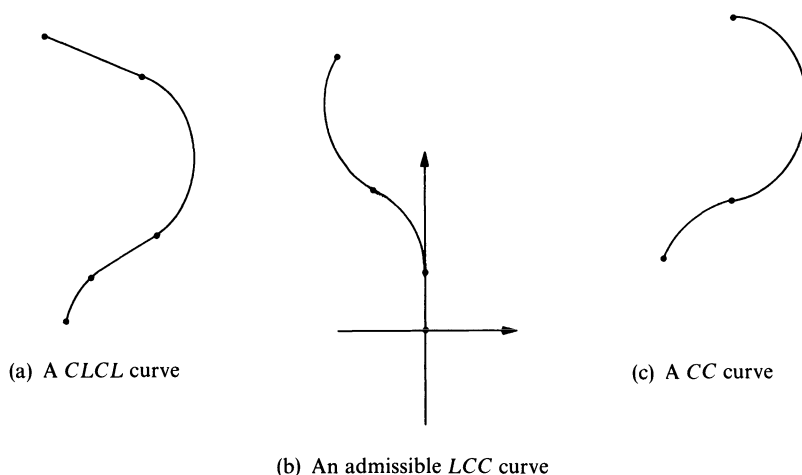


FIG. 1. Examples of the C-L notation. All circular arcs have radius ρ

A curve of type *CLCL* is a differentiable curve which is the union of two circular arcs of radius ρ and two line segments in the order (starting from its initial point) arc, segment, arc, segment. Such a curve may not be admissible as it need not satisfy (i) above. We emphasize that one or more arcs or segments may have zero length. For example, a *CC* curve may be thought of as a *CCC* curve with one circular arc having zero length.

$R(t)$ is the set of terminal points of all admissible curves of length vt , i.e., of all curves in $\mathcal{L}(*, *, vt)$. In the subsequent sections we establish a few properties of the \mathcal{L} classes and determine the region $R(t)$ for all t . Our methods enable us to program the CALCOMP computer plotter at the University of Victoria to draw the region for any t . Selections from the computer output are given in the Appendix.

L. E. Dubins [3], [4] has obtained many interesting results concerning the classes $\mathcal{L}(z, \alpha, *)$. In particular, he established the existence of a shortest curve in this class for all pairs z, α and showed that each shortest curve is either a *CLC* or a *CCC* curve. Further he has shown the existence of pairs z, α for which $\mathcal{L}(z, \alpha, *)$ is not arcwise connected. In this paper we have used notations, techniques and results from Dubins' papers.

Z. A. Melzak [8] has determined the regions of accessibility for $t \leq \pi\rho/v$ when trajectories are assumed to have piecewise continuous curvature bounded by $1/\rho$. His techniques will be useful in subsequent sections. H. G. Robertson [9] has shown that if P moves on a smooth curve whose curvature is bounded by $1/\rho$, then P is able to reach points inside his initial optimal turning circles if and only if $t > \pi\rho/v$. This result extends an earlier result of A. W. Goodman [5]. Finally, E. J. Cockayne [2] and R. Isaacs [6] have solved pursuit problems in which both pursuer and evader travel with constant speed under curvature constraints.

2. Properties of the \mathcal{L} classes. The principal purpose of this section is to establish Theorem 1 which states that any point in $R(t)$ can be reached by a critical curve of length vt having finitely many line segments and circular arcs. This theorem is of fundamental importance in our determination of the regions $R(t)$ in § 3. The proof of this simple property is surprisingly long and further research might well establish the theorem (and its generalization to other situations) more elegantly. In order to prove Theorem 1, several results of Dubins are used; also we have to obtain new properties of the \mathcal{L} classes. All such auxiliary results are termed "Propositions". Several new propositions are interesting in their own right as they extend Dubins' work. Some of the results proved for curves of length $\pi\rho/2$ might well hold for longer curves. Since $\pi\rho/2$ is sufficient for our purposes, i.e., the proof of Theorem 1, we have not investigated these possibilities.

We first introduce some notation used by Dubins in [3] and state some of his results which we shall need in our analysis.

PROPOSITION 1 (Dubins [3, Prop. 1]). $\mathcal{L}(z, \alpha, *)$ has a curve of minimum length for all z, α .

We note that the Lipschitz condition, which we used to define the curvature constraint, was used to prove this theorem. Following Dubins, we shall refer to the minimum length curves as ρ -geodesics.

Let $Z(s)$ be an admissible curve. At each point of Z there are two circles of radius ρ which are tangential to Z and on which opposite orientations are induced by the direction of Z . Let U_s, V_s be, respectively, the counterclockwise and clockwise oriented circles of radius ρ , tangent to the curve Z at the point $Z(s)$ (see Fig. 2).

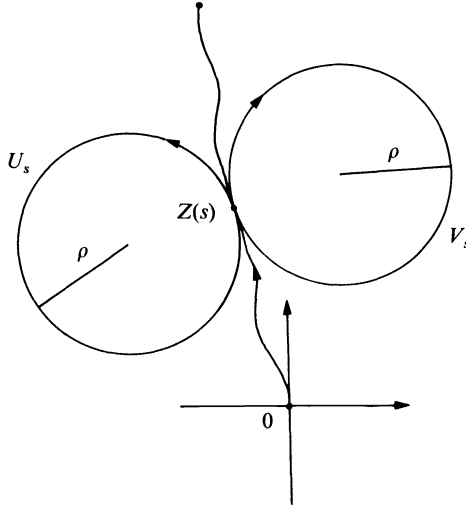


FIG 2. The circles U_s, V_s

PROPOSITION 2 (Dubins). Let Z be an admissible curve of length $\pi\rho/2$. For all $\sigma \in [0, \pi\rho/2]$ the circle V_σ is either disjoint from or tangential to the circle U_0 . Furthermore V_σ is tangential to U_0 if and only if $Z(s)$ on $[0, \sigma]$ is a CC critical admissible curve with the first critical arc coinciding with an arc of U_0 .

This is Proposition 6 of [3]. There is a similar result for circles U_σ and V_0 which we leave unstated but also refer to as Proposition 2.

PROPOSITION 3 (Dubins). *Let Z be an admissible curve of length $\leq \pi\rho/2$ with angle function $\phi(s)$. Then (i) for each $s \in [0, \pi\rho/2]$, $\mathcal{L}(Z(s), \phi(s), *)$ has ρ -geodesic of type CLC. (ii) If $m(s)$ is the length of the ρ -geodesic of $\mathcal{L}(Z(s), \phi(s), *)$, then $m(s)$ is a continuous function of s .*

Part (i) of this result is contained in Propositions 9, 11 and 12 of [3]. (Actually Dubins states his result for length $\leq \pi\rho/8$ but his proofs are valid for length $\leq \pi\rho/2$.) Part (ii) is simply deduced from the proof of Proposition 11 in [3]. We omit the details.

We now establish some new properties of the \mathcal{L} classes.

PROPOSITION 4. *Let $Z(s)$ be an admissible curve of length $\pi\rho/2$ and angle function $\phi(s)$. For each s in $[0, \pi\rho/2]$,*

(i) *if Z on $[0, s]$ is not of type CC, $\mathcal{L}(Z(s), \phi(s), *)$ contains precisely two CCC critical admissible curves with each circular arc having length strictly between zero and $\pi\rho$.*

(ii) *if Z on $[0, s]$ is of type CC, $\mathcal{L}(Z(s), \phi(s), *)$ contains precisely one CCC curve with each circular arc less than a semicircle, namely Z restricted to $[0, s]$.*

Proof. (i) Suppose Z restricted to $[0, s]$ is not of type CC (this implies $s > 0$). Any CCC curve in $\mathcal{L}(Z(s), \phi(s), *)$ has its first and third circular arcs either on V_0 and V_s respectively or on U_0 and U_s respectively. The existence of a unique CCC curve of the former type with the lengths of each circular arc strictly between zero and $\pi\rho$ will be established. The other proof is similar.

First, if $X(s)$ is the center of the circle V_s , then

$$|X(s) - X(0)| \leq |X(s) - Z(s)| + |Z(s) - Z(0)| + |Z(0) - X(0)| \leq 2\rho + s.$$

But $s \leq \rho\pi/2$, hence $|X(s) - X(0)| \leq (\pi/2 + 2)\rho < 4\rho$. Moreover, it follows from Proposition 4 of [3] (we emphasize a difference between notations in this reference and our present work) that $|X(s) - X(0)| > 0$. Therefore V_0, V_s are positioned so that there exist precisely two critical circles tangential to both. Let W_s be the circle, with center K , tangential to V_0, V_s at I and J respectively and lying to the left as one looks from $X(0)$ towards $X(s)$ (i.e., $X(0), K, X(s)$ is a clockwise ordering of vertices on the triangle). Then C_2 , the anticlockwise arc of W_s from I to J , has length strictly between zero and $\pi\rho$.

Suppose that the half-line starting at $X(0)$ in the direction $X(0)X(s)$ meets V_0 at L and the half-line starting at $X(s)$ in the direction of $X(s)X(0)$ meets V_s at M . We refer to Fig. 3. It will be shown that I is strictly between $Z(0)$ and L on the clockwise arc of V_0 from $Z(0)$ to L and that J is strictly between M and $Z(s)$ on the clockwise arc of V_s from M to $Z(s)$. Let $Y(0), Y(s)$ be the centers of U_0, U_s respectively. We deduce from Proposition 2 that the distances $|Y(s) - X(0)|$ and $|Y(0) - X(s)|$ are greater than 2ρ . From these inequalities and the equalities $|X(0) - Y(0)| = |X(s) - Y(s)| = 2\rho$ we deduce that K , which is an intersection point of circles of radius 2ρ with centers $X(0)$ and $X(s)$, lies "above" the line $Y(0)X(0)$ and "below" the line $Y(s)X(s)$. More precisely, $KX(0)Y(0)$ and $KY(s)X(s)$ are clockwise orders on the triangles they define. Therefore I , the intersection point of the segment $KX(0)$ with V_0 , lies strictly between $Z(0)$ and L on the clockwise arc of V_0 from $Z(0)$ to L as required and the similar result for J also follows.

Let C_1, C_3 respectively be the clockwise arcs of V_0 from $Z(0)$ to I and of V_s from J to $Z(s)$. The preceding paragraph has shown that the lengths of C_1 and C_3

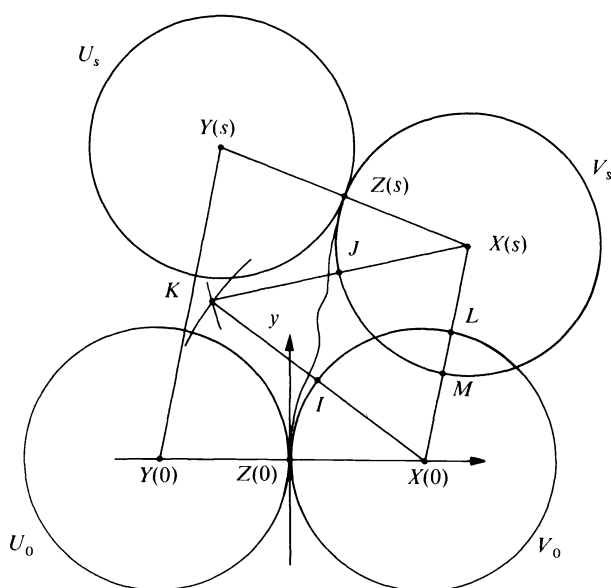


FIG. 3

are strictly positive. We now show that the length of C_1 is less than $\pi\rho$. A similar argument shows the same bound for C_3 .

Let $Z(s) = (x(s), y(s))$. The y -coordinate $\eta(s)$ of $X(s)$ is then equal to $y(s) - \rho \sin \phi(s)$. For each t in $[0, s]$, $|\phi(t)| \leq t/\rho$ and $s \leq \rho\pi/2$. Therefore

$$y(s) = \int_0^s \cos \phi(t) dt \geq \int_0^s \cos(t/\rho) dt = \rho \sin s/\rho$$

and

$$\eta(s) \geq \rho \sin(s/\rho) - \rho \sin \phi(s) \geq 0.$$

Thus the clockwise arc of V_0 from $Z(0)$ to L is no more than a semicircle and hence the length of C_1 is less than $\pi\rho$.

The union of C_1, C_2, C_3 is the desired CCC curve which we denote by $(ZV)_s$. Any other CCC curve in $\mathcal{L}(Z(s), \phi(s), *)$ with its first and third circular arcs on V_0, V_s respectively necessarily has at least one circular arc which is more than a semicircle.

A similar proof establishes the existence of a unique CCC curve $(ZU)_s$ in $\mathcal{L}(Z(s), \phi(s), *)$ with first and third circular arcs on U_0, U_s respectively. The strictly positive length property of the circular arcs enables us to assert that $(ZU)_s, (ZV)_s$ are distinct. This completes the proof of the first part of the proposition.

(ii) If Z restricted to $[0, s]$ is a CC curve (we re-emphasize this includes C curves), then as in part (i), we can construct $(ZU)_s$ and $(ZV)_s$ which are CCC curves of $\mathcal{L}(Z(s), \phi(s), *)$ with all circular arcs less than a semicircle. However, each has zero length circular arcs and each is precisely the original curve Z restricted to $[0, s]$. Details are left to the reader.

Let Z be an admissible curve of length $\pi\rho/2$ and let $|(ZU)_s|, |(ZV)_s|$ be the lengths of the two CCC curves of $\mathcal{L}(Z(s), \phi(s), *)$ whose existence was established by Proposition 4. Then $|(ZU)_s|, |(ZV)_s|$ are continuous functions of s on $[0, \pi\rho/2]$, and we have the following corollary.

COROLLARY 1. $M(s) = \max \{|(ZU)_s|, |(ZV)_s|\}$ is a continuous function of s on $[0, \pi\rho/2]$ and $M(0) = 0$.

Let Z_1, Z_2 be admissible curves of length $\pi\rho/2$ with angle functions $\phi_1(s), \phi_2(s)$. If

$$\|\phi_1(s) - \phi_2(s)\| = \sup_{s \in [0, \pi\rho/2]} |\phi_1(s) - \phi_2(s)|$$

is made sufficiently small, then for any $s \in [0, \pi\rho/2]$ the curves $(Z_1U)_s$ and $(Z_2U)_s$ and hence their lengths can be made arbitrarily close. A similar result is also true for the lengths $|(Z_1V)_s|$ and $|(Z_2V)_s|$.

COROLLARY 2. Let Z_1, Z_2 be admissible curves of length $\pi\rho/2$ with angle functions $\phi_1(s), \phi_2(s)$. For any $s \in [0, \pi\rho/2]$ and any $\varepsilon > 0$, there exists δ such that $\|\phi_1(s) - \phi_2(s)\| < \delta$ implies $|M_1(s) - M_2(s)| < \varepsilon$.

We now compare the lengths of the two CCC curves $(ZV)_s, (ZU)_s$ of Proposition 4 which we shall call *complementary CCC curves*.

PROPOSITION 5. Suppose that a CCC curve of length less than or equal to $\pi\rho$ has circular arcs of lengths $\rho a, \rho b, \rho c$ (in order from the initial point), and suppose that the complementary curve exists. Then a necessary and sufficient condition for the curve to be longer (shorter) than its complementary curve is

$$(1) \quad a - b + c < 0 \quad (> 0).$$

Proof. Without losing generality, we assume the curve is $(ZV)_s$. Let the lengths of the arcs of $(ZU)_s$ be $\rho p, \rho q, \rho r$ and let the center of the second arc of $(ZU)_s$ be Q . (See Fig. 4.) Summing the angles of quadrilateral $X(0)X(s)Y(s)Y(0)$ we obtain

$$a + c + p + r + (\pi - q) + (\pi - b) = 2\pi,$$

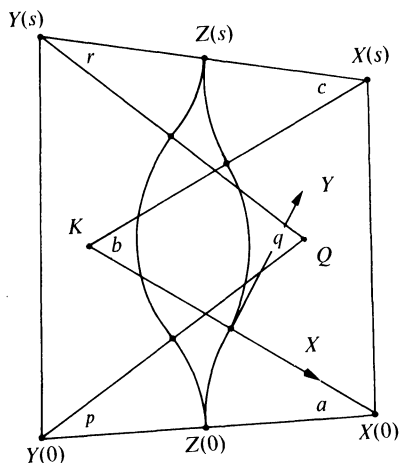


FIG. 4

i.e.,

$$(a + c) - b = q - (p + r).$$

Therefore

$$\begin{aligned} |(ZV)_s| - |(ZU)_s| &= \rho\{(a + b + c) - (p + q + r)\} \\ &= 2\rho\{a + c - q\}. \end{aligned}$$

Hence $(ZV)_s$ is longer (shorter) than its complementary curve $(ZU)_s$ if and only if $a + c > q$ ($a + c < q$), i.e., if and only if

$$(2) \quad \cos(a + c) < \cos q \quad (\cos(a + c) > \cos q).$$

Consider new Cartesian axes along the tangent and normal to $(ZV)_s$ at the point of tangency of the first and second arcs (see Fig. 4). Referred to these axes we have the following coordinates:

$$(3) \quad \begin{aligned} Y(0) &: (\rho(1 - 2\cos a), -2\rho \sin a), \\ Y(s) &: (\rho(2\cos b - 1 - 2\cos(b - c)), \rho(2\sin b - 2\sin(b - c))). \end{aligned}$$

From the isosceles triangle $Y(0)QY(s)$, $\sin(q/2) = |Y(0) - Y(s)|/4\rho$. Therefore from (3),

$$\begin{aligned} (4) \quad \cos q &= 1 - 2\sin^2(q/2) \\ &= \cos a + \cos b + \cos c + \cos(a - b + c) \\ &\quad - \cos(a - b) - \cos(b - c) - 1. \end{aligned}$$

Combining (2) and (4) we see that $(ZV)_s$ is longer than $(ZU)_s$ if and only if

$$\begin{aligned} \{\cos c + \cos(a - b + c)\} - \{\cos(a + c) + \cos(b - c)\} \\ + \cos a + \cos b - \cos(a - b) - 1 > 0. \end{aligned}$$

A little elementary trigonometry reduces the left-hand side to

$$-4 \sin\left(\frac{a}{2}\right) \sin\left(\frac{b}{2}\right) \sin\left(\frac{c}{2}\right) \sin\left(\frac{a - b + c}{2}\right).$$

Hence $(ZV)_s$ is longer (shorter) if and only if $a - b + c < 0$ (> 0) as required.

COROLLARY 3. *If a CCC curve with circular arcs of length ρa , ρb , ρc is longer than its complementary curve, then $a < b$ and $c < b$.*

PROPOSITION 6. *There exists $\beta > 0$ such that if Z is a critical admissible curve of length $\rho\beta$ which consists of critical circular arcs, then for each $s \in [0, \rho\beta]$, $M(s) = \max\{|(ZU)_s|, |(ZV)_s|\} \geq s$, with equality implying that Z restricted to $[0, s]$ is of type CCC.*

Proof. There exists $\eta > 0$ such that for any admissible curve of length $\rho\eta$ and all $s \in [0, \rho\eta]$, $M(s) \leq \pi\rho$. Let $\beta = \min\{\eta, \pi/2\}$. Then Proposition 5 is applicable to curves of length $\rho\beta$. Let Z have length $\rho\beta$ and let Z restricted to $[0, s]$ be the union of n circular arcs C_1, \dots, C_n , where C_i joins $Z(s_{i-1})$ and $Z(s_i)$ and $0 = s_0 < s_1 < \dots < s_{n-1} < s_n = s$. Suppose $M(s) < s$; then there is a smallest integer k so that $M(s_k) \geq s_k$ and $M(s_{k+1}) < s_{k+1}$. Clearly $3 \leq k \leq n - 1$. Without

losing generality we assume that C_{k+1} is an arc of $U_{s_{k+1}}$, i.e., $(ZU)_{s_k}$ is obtained by deleting C_{k+1} from $(ZU)_{s_{k+1}}$. Figure 5 shows the curves $(ZU)_{s_k}$, $(ZV)_{s_k}$,

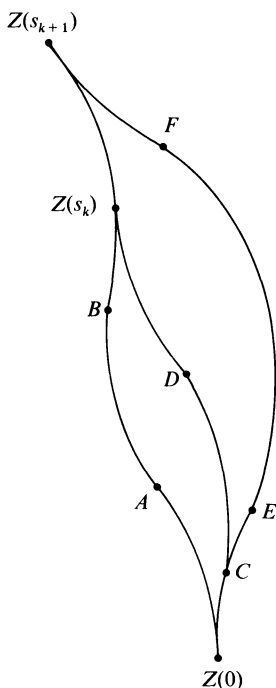


FIG. 5

$(ZU)_{s_{k+1}}$ and $(ZV)_{s_{k+1}}$; the curve Z itself is not shown. All labeled points are points of tangency of critical circles and, for example, we shall use the notations \overline{AB} , \overline{CEF} to denote lengths of the arcs AB , CEF .

Suppose $M(s_k) = |(ZU)_{s_k}|$. Then

$$\begin{aligned} M(s_{k+1}) &\geq |(ZU)_{s_{k+1}}| = |(ZU)_{s_k}| + \overline{Z(s_k)Z(s_{k+1})} \\ &= M(s_k) + (s_{k+1} - s_k) \\ &\geq s_k + (s_{k+1} - s_k) \\ &= s_{k+1}, \end{aligned}$$

which is a contradiction. On the other hand, if $M(s_k) = |(ZV)_{s_k}| > |(ZU)_{s_k}|$, then by Corollary 3,

$$(5) \quad \overline{DZ(s_k)} < \overline{DC}.$$

But the curves $CDZ(s_k)Z(s_{k+1})$ and $CEFZ(s_{k+1})$ are complementary CCC curves and (5) and Corollary 3 imply that

$$\overline{CEFZ(s_{k+1})} \geq \overline{CDZ(s_k)Z(s_{k+1})}.$$

Augmenting each of these by the arc $Z(0)C$, we obtain

$$|(ZV)_{s_{k+1}}| \geq |(ZV)_{s_k}| + \overline{Z(s_k)Z(s_{k+1})}.$$

Therefore

$$\begin{aligned} M(s_{k+1}) &\geq |(ZV)_{s_{k+1}}| \geq M(s_k) + (s_{k+1} - s_k) \\ &\geq s_k + (s_{k+1} - s_k) = s_{k+1}, \end{aligned}$$

once again a contradiction. Therefore $M(s) \geq s$.

Suppose $M(s) = s$ and that Z restricted to $[0, s]$ is not of type CCC, i.e., $n \geq 4$. Then $M(s_4) = s_4$ (otherwise $M(s_4) > s_4$ and analysis similar to the above successively establishes $M(s_k) > s_k$, $k = 5, \dots, n$, a contradiction). Suppose that the lengths of C_1, C_2, C_3, C_4 are $\rho\mu_1, \rho\mu_2, \rho\mu_3, \rho\mu_4$, respectively; then $\mu_2 \leq \mu_3$ or $\mu_3 \leq \mu_2$ (or both). Suppose, without losing generality, that the former is true. By Corollary 3, Z restricted to $[0, s_3]$ is shorter than its complementary curve Z' , and the union of Z' and C_4 which is either $(ZU)_{s_4}$ or $(ZV)_{s_4}$ is longer than s_4 , a contradiction. Hence Z restricted to $[0, s]$ is of type CCC.

COROLLARY 4. *There exists $\alpha > 0$ such that if Z is a critical curve of length $\rho\alpha$, then for all s in $[0, \rho\alpha]$, $M(s) \geq s$ with strict inequality if Z restricted to $[0, s]$ is not of type CCC.*

Proof. There is nothing further to prove if Z restricted to $[0, s]$ has no line segment.

Let Z be a critical admissible curve of length $\alpha\rho$, where α is defined by $\alpha + 4 \sin^{-1}(\alpha/4) = \beta$, and suppose that Z restricted to $[0, s]$ contains a line segment. We construct a new curve Z^* in $\mathcal{L}(Z(s), \phi(s), *)$ by replacing each line segment by a CCC curve tangent to the segment at its extremities. Z^* is a critical admissible curve consisting entirely of circular arcs of radius ρ and has the same pair of CCC curves as Z restricted to $[0, s]$. The increase in length caused by the replacement of a line segment of length λ by a CCC curve is $4\rho \sin^{-1}(\lambda/4\rho) - \lambda$. Hence, if the line segments of Z have lengths $\lambda_1, \dots, \lambda_N$, the length s^* of Z^* is

$$s - \sum_1^N \lambda_i + \sum_1^N 4\rho \sin^{-1} \left(\frac{\lambda_i}{4\rho} \right),$$

where

$$\lambda_i > 0 \quad \text{and} \quad \sum_1^N \lambda_i < s.$$

Therefore

$$\begin{aligned} s^* &< s - \sum_1^N \lambda_i + 4\rho \sin^{-1} \left(\frac{\sum_1^N \lambda_i}{4\rho} \right) \\ &< s + 4\rho \sin^{-1} (s/4\rho) \\ &\leq \alpha + 4\rho \sin^{-1} (\alpha/4) = \beta. \end{aligned}$$

Hence we can apply the proposition to Z^* , and $M(s) \geq s^* > s$ as required.

We note, without proof, that the constant β is at least $2/\sqrt{\pi}$ radians.

PROPOSITION 7. *Let Z be an admissible curve in $\mathcal{L}(z, \theta, \sigma)$. For any $\varepsilon > 0$ there exists a critical admissible curve W which consists entirely of critical circular arcs and satisfies*

- (i) $W \in \mathcal{L}(*, \theta, \sigma)$,
- (ii) for all $s \in [0, \sigma]$, $|W(s) - Z(s)| < \varepsilon$.

Proof. We construct, for each positive integer n , a function $\psi_n(s)$ on $[0, \sigma]$ from $\phi(s)$, the angle function of Z , as follows. Consider the graph of $\phi(s)$ restricted to $[k\sigma/n, (k+1)\sigma/n]$, where $0 \leq k \leq n-1$. We use the notation of Fig. 6. By the

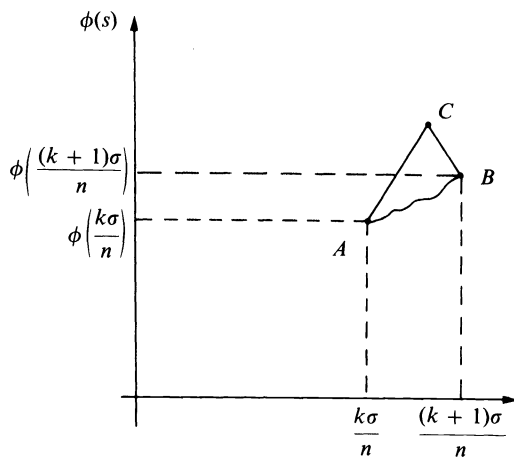


FIG. 6

Lipschitz condition, the slope of the chord AB is less than or equal to $1/\rho$ and hence the lines through A with slope $1/\rho$ and through B with slope $-1/\rho$ intersect at C , whose first coordinate is in the interval $[k\sigma/n, (k+1)\sigma/n]$. The graph of $\psi_n(s)$ restricted to $[k\sigma/n, (k+1)\sigma/n]$ consists of the two segments AC, CB . The graph of $\psi_n(s)$ on $[0, \sigma]$ is the union of such segments for $k = 0, \dots, n-1$. For each n , $\psi_n(0), \psi_n(\sigma) = \phi(\sigma) = \theta$ and $\psi_n(s)$ is the angle function of a critical W_n consisting of circular arcs. Moreover, by making n sufficiently large, $\|\psi_n(s) - \phi(s)\|$ can be made arbitrarily small. Therefore for sufficiently large n , W_n satisfies the requirements of the proposition.

Using this approximation result, we may now generalize Proposition 6 to all admissible curves of length less than or equal to $\rho\beta$ (β is the constant of Proposition 6). It is interesting to emphasize the connection with Dubin's work. He proves that the shortest curves in $\mathcal{L}(z, \alpha, *)$ (i.e., the ρ -geodesics of Proposition 1) are critical curves of type CLC or CCC . The following proposition shows that many classes $\mathcal{L}(z, \alpha, *)$ have local maxima (w.r.t. length) which are critical curves of type CCC .

PROPOSITION 8. *Let Z be an admissible curve of length $\rho\beta$. For each $s \in [0, \rho\beta]$, $M(s) \geq s$.*

Proof. Let $s \in [0, \rho\beta]$ and suppose contrary to the result for some $\varepsilon_1 > 0$,

$$(6) \quad M(s) = s - \varepsilon_1.$$

Let $\phi(s)$ be the angle function of Z . From the proof of Proposition 7 it follows that for all $\varepsilon_2 > 0$, there is a critical curve W of length s , consisting of circular arcs whose angle function $\psi(s)$ satisfies $\|\psi(s) - \phi(s)\| < \varepsilon_2$. Suppose

$$L(s) = \max \{|(WV)_s|, |(WU)_s|\}.$$

By Corollary 2 to Proposition 4, we may choose ε_2 sufficiently small so that

$$(7) \quad |L(s) - M(s)| \leq \varepsilon_1/2.$$

But by Proposition 6, $L(s) \geq s$. Therefore from (6),

$$L(s) - M(s) \geq s - (s - \varepsilon_1) = \varepsilon_1,$$

which contradicts (7). Therefore $M(s) \geq s$ as required.

In [3], Dubins gives an example of a class $\mathcal{L}(z, \theta, *)$ which is not arcwise connected. Using Proposition 8, we now exhibit an uncountable family of these classes, each member of which is not arcwise connected.

PROPOSITION 9. *There exists $\alpha > 0$ such that for all admissible curves Z with angle function $\phi(s)$, length $\rho\alpha$ and all $s \leq \rho\alpha$, $\mathcal{L}(Z(s), \phi(s), *)$ is not arcwise connected.*

Proof. Let $\eta > 0$ be a real number which is small compared with β , the constant of Proposition 6, and let α be such that for all admissible curves Z of length $\rho\alpha$ and $s \leq \rho\alpha$, $M(s) \leq \rho(\beta - \eta)$.

Now suppose that for some $s \leq \rho\alpha$ and ε satisfying $0 < \varepsilon \leq \rho\eta$, $\mathcal{L}(Z(s), \phi(s), *)$ contains a curve W of length $M(s) + \varepsilon$. Since this length is less than or equal to $\rho\beta$, we can apply Proposition 8 to the curve W and deduce that its length $M(s) + \varepsilon$ is less than or equal to the length L of its longer CCC curve. But L is precisely $M(s)$ since $W \in \mathcal{L}(Z(s), \phi(s), *)$, i.e., we have shown $M(s) + \varepsilon \leq M(s)$.

Therefore $\mathcal{L}(Z(s), \phi(s), *)$, which certainly contains curves of length $M(s)$ and of length $> M(s)$, contains no curve of length $M(s) + \varepsilon$ for $0 < \varepsilon \leq \rho\eta$. Hence $\mathcal{L}(Z(s), \phi(s), *)$ is not arcwise connected.

We are now able to prove the principal result of this section; that any point of $R(t)$ is attainable by a critical curve of finitely many arcs and segments.

THEOREM 1. *Let Z be an admissible curve with angle function $\phi(s)$ and length σ . Then $\mathcal{L}(Z(\sigma), \phi(\sigma), \sigma)$ contains a critical curve which is the union of finitely many line segments and circular arcs.*

Proof. Assume initially that $\sigma \leq \rho\beta$. For each s in $[0, \sigma]$, Proposition 3 asserts that there is a shortest curve $(ZG)_s$ which satisfies the Lipschitz condition, joins $Z(s)$ and $Z(\sigma)$ and is tangential to Z at these points. $(ZG)_s$ is a CLC curve. If its length is $a(s)$, then by definitions and Proposition 8,

$$(8) \quad \begin{aligned} a(\sigma) &= M(0) = 0, \\ a(0) &\leq \sigma, \quad M(\sigma) \geq \sigma. \end{aligned}$$

If $a(0) = \sigma$, $(ZG)_0$ is the required curve, and if $M(\sigma) = \sigma$, the longer of $(ZU)_\sigma$, $(ZV)_\sigma$ is the required curve, so we assume that the inequalities of (8) are strict. Consider the continuous function $g(s) = a(s) + M(s)$. By (8), $g(0) < \sigma$ and $g(\sigma) > \sigma$ and for some s^* in $(0, \sigma)$, $g(s^*) = \sigma$. Then the union of $(ZG)_{s^*}$ with the longer of $(ZU)_{s^*}$, $(ZV)_{s^*}$ is the required curve.

For $\sigma > \rho\beta$, we split up $[0, \sigma]$ into subintervals of length less than or equal to $\rho\beta$ and apply the above separately to Z restricted to each subinterval. The required curve is then the union of all the critical curves so constructed, and the proof is complete.

Since any point of $R(t)$ is accessible by a critical curve of length vt , we may change the Lipschitz condition (iii) in the definition of admissible curve to

(iii) $\phi'(s)$ is piecewise continuous on $[0, \sigma]$ and $|\phi'(s)| \leq 1/\rho$, without altering any of the regions of accessibility; i.e., the regions as defined in this paper are precisely the same as those defined by Melzak [8].

3. Determination of the regions $R(t)$. The steps which will determine the regions are as follows. We shall prove $R(t)$ is closed (Theorem 2). It is obviously bounded. Then, in § 3.1 we shall characterize the boundary of $R(t)$ with Theorems 3 and 4.

The former asserts that any point of the boundary is attainable by a CC or CL critical curve of length vt . Theorem 3 follows immediately from Theorem 1 (already proved), and the fact that any point of $R(t)$ attainable by a critical curve not of type CC or CL is either on the interior of $R(t)$ or is also attainable by a critical curve of type CC or CL (Propositions 10 and 11).

Theorem 4 shows that some CC and CL curves cannot terminate on the boundary of $R(t)$.

Each region is now determined (§ 3.2), as there is a unique way of drawing the closed boundary which is contained in the set of endpoints of those CC and CL critical curves not eliminated by Theorem 4 and such that the boundary changes continuously with t .

THEOREM 2. $R(t)$ is closed in E^2 for all t .

Proof. Let Φ_t be the set of angle functions of admissible curves of length vt . Then Φ_t is uniformly bounded and equicontinuous in $C[0, vt]$ (the space of all continuous functions on $[0, vt]$ with uniform norm). Hence by Arzela's theorem [7, p. 54], Φ_t is compact in $C[0, vt]$. But Φ_t is also closed [7, p. 29, Ex. (1)] and therefore is a compactum [7, p. 57]. The function from Φ_t to E^2 mapping the angle function of an admissible curve into its endpoint is continuous. Therefore $R(t)$, the image Φ_t under this function, is a compactum and hence closed [7, p. 59].

3.1. Which curves have endpoints in the boundary of $R(t)$? In this section we establish that most critical curves of $\mathcal{L}(*, *, vt)$ terminate in interior points of $R(t)$, using variations of a perturbation technique introduced by Melzak [8]. We denote the interior of $R(t)$ by $[R(t)]^\circ$. A critical curve is n -critical if it is the union of n circular arcs or straight segments each of which has nonzero length.

PROPOSITION 10. Let $\mathcal{L}(z, \theta, vt)$ contain an n -critical curve Z , where $n \geq 2$ and the last two critical arcs C_1, C_2 of Z form a curve of type LC . Either $Z \in [R(t)]^\circ$ or $\mathcal{L}(z, \theta, vt)$ contains an n -critical curve whose last two critical arcs form a curve of type CL .

Proof. Let C_1 be a straight segment of length l and C_2 be a critical circular arc of length $\rho\alpha$, where $l + \rho\alpha = vt$. We embed Z in a two-parameter family of curves in $\mathcal{L}(*, *, vt)$. $Z_{u,w}$ for $w \neq 0$ is obtained from Z by replacing C_1, C_2 with the union of a circular arc of radius $1/w$ and length $l + u$ and a circular arc of radius ρ and length $\rho\alpha - u$ (see Fig. 7). $Z_{u,0}$ is obtained from Z by replacing

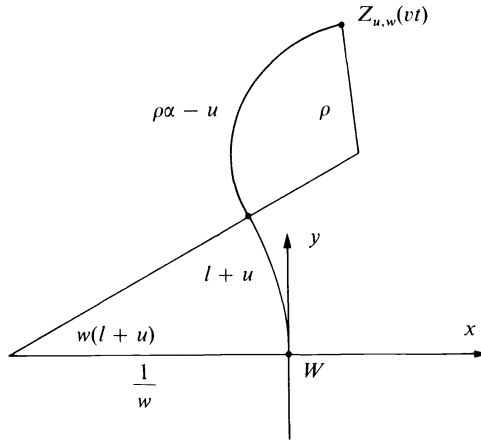


FIG. 7

C_1, C_2 with the union of a straight segment of length $l + u$ and a critical circular arc of length $\rho\alpha - u$.

We define $Z_\varepsilon = \{Z_{u,w} : |u| < \varepsilon, |w| < \varepsilon\}$, where ε is a positive constant which is small compared with $\min\{1/\rho, \rho\alpha\}$. Then $Z = Z_{0,0}$ and $Z_\varepsilon \subseteq \mathcal{L}(*, *, vt)$. Let $Z_{u,w}(vt) = (x(u, w), y(u, w))$ refer to axes along the normal and tangent to Z at the initial point W of C_1 , where

$$x(u, w) = \begin{cases} -\frac{1}{w} + \left(\frac{1}{w} + \rho\right) \cos(w(l+u)) - \rho \cos\left(\alpha - \frac{u}{\rho} - w(l+u)\right) & \text{if } w \neq 0, \\ \rho(1 - \cos(\alpha - u/\rho)) & \text{if } w = 0; \end{cases}$$

$$y(u, w) = \begin{cases} \left(\frac{1}{w} + \rho\right) \sin(w(l+u)) + \rho \sin\left(\alpha - \frac{u}{\rho} - w(l+u)\right) & \text{if } w \neq 0, \\ l + u + \rho \sin(\alpha - u/\rho) & \text{if } w = 0. \end{cases}$$

It is clear that $x(u, w), y(u, w)$ have continuous partial derivatives on the set $\{(u, w) : |u| < \varepsilon, |w| < \varepsilon, w \neq 0\}$. We omit the proof that the partials are also continuous for $w = 0$ and only calculate the Jacobian at $u = w = 0$. $(x_u)_{0,0}$ will mean $\partial x / \partial u$ evaluated at $u = w = 0$. It is left to the reader to verify that

$$(x_u)_{0,0} = -\sin \alpha,$$

$$(y_u)_{0,0} = 1 - \cos \alpha,$$

$$(x_w)_{0,0} = -l^2 - \rho l \sin \alpha,$$

$$(y_w)_{0,0} = \rho l (1 - \cos \alpha).$$

Hence the Jacobian J at $u = w = 0$ equals $2l^2 \sin^2(\alpha/2)$. If $\alpha \neq 2k\pi$, $J > 0$, and by the inverse function theorem [1, p. 144] the endpoints of some subfamily of Z_ε fill a neighborhood of $Z(vt)$. Hence $z \in [R(t)]^\circ$. If $\alpha = 2k\pi$, let Z' consist of the portion of Z from 0 to W followed by a circular arc of length $2k\pi$, followed

by a straight segment of length l . Z' is an n -critical curve whose last two critical arcs form a curve of type CL as required.

PROPOSITION 11. *Let $\mathcal{L}(z, \theta, vt)$ contain an n -critical curve Z , where $n \geq 3$ and the last three critical arcs C_1, C_2, C_3 of Z form a curve of one of the types (i) CCC , (ii) LCC , (iii) CCL , (iv) LCL . Then either $z \in [R(t)]^\circ$ or $\mathcal{L}(z, \theta, vt)$ contains an $(n - 1)$ -critical curve.*

Proof. We use a perturbation technique similar to that which established Proposition 10. Suppose that C_1, C_2, C_3 are circular arcs of radii a, b, c , respectively, and have lengths $\alpha\alpha, b\beta, c\gamma$ where $\alpha, \beta, \gamma > 0$ and such that their curvatures alternate in sign (as in a CCC curve). Let Y be the curve consisting of the union of C_1, C_2, C_3 . We embed Z in a two-parameter family of admissible curves in $\mathcal{L}(*, *, vt)$ as follows. We form $Z_{u,w}$ by replacing Y by the union of three circular arcs of radii a, b, c and having lengths $\alpha\alpha + u, b\beta + w, c\gamma - (u + w)$, respectively. $Z_\varepsilon = \{Z_{u,w} : |u| < \varepsilon, |w| < \varepsilon\}$, where ε is a positive real number satisfying $\varepsilon < \min\{\alpha\alpha, b\beta, c\gamma\}$. Then $Z = Z_{0,0}$ and $Z_\varepsilon \subseteq \mathcal{L}(*, *, vt)$.

Let $Z_{u,w}(vt) = (x(u, w), y(u, w))$ refer to coordinate axes along the normal and tangent at the initial point of C_1 . We omit the page of elementary trigonometry which shows that J , the Jacobian of $(x(u, w), y(u, w))$ evaluated at $u = w = 0$, is given by

$$(9) \quad J = \frac{4(a+b)(b+c)}{ab} \sin\left(\frac{\beta - \gamma}{2}\right) \sin\frac{\beta}{2} \sin\frac{\gamma}{2}.$$

(i) In this case $a = b = c = \rho$. We may assume β, γ are strictly between 0 and 2π . For if, say, $\beta = \beta' + 2k\pi$ for k a positive integer and $0 \leq \beta' < 2\pi$, a new critical admissible curve in $\mathcal{L}(z, \theta, vt)$ may be constructed by replacing C_1, C_2 by the union of critical circular arcs of length $\rho(\alpha + 2k\pi), \rho\beta'$ respectively. If $\beta' = 0$, the new curve is $(n - 1)$ -critical and there is nothing to prove.

From equation (9), for this case,

$$J = 16 \sin\left(\frac{\beta - \gamma}{2}\right) \sin\frac{\beta}{2} \sin\frac{\gamma}{2}.$$

Since $0 < \beta, \gamma < 2\pi$, $\sin(\beta/2) \sin(\gamma/2) \neq 0$. Hence $J \neq 0$ unless $\beta = \gamma$. Therefore for $\beta \neq \gamma$, by the inverse function theorem, $z \in [R(t)]^\circ$.

If $\beta = \gamma$, there is a tangent from $Z(vt)$ to the arc C_2 meeting C_2 at T (see Fig. 8). Let Y be the reflection of the portion of Z from T to $Z(vt)$ in the tangent.

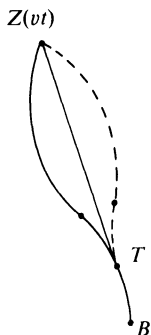


FIG. 8. The reflection process

Then Y united with the portion of Z from B to T is a CCC curve with the last two circular arcs having unequal lengths strictly between 0 and $2\pi\rho$. Applying the above we deduce $z \in [R(t)]^\circ$. We shall refer to this technique as the *reflection process*.

(ii) Let C_1 be a straight segment of length l and C_2, C_3 be critical circular arcs of length $\rho\beta, \rho\gamma$, respectively. If β or γ equals $2k\pi$, then $\mathcal{L}(z, \theta, vt)$ contains an $(n-1)$ -critical curve. If $\beta = \gamma + 2k\pi$, then the reflection process used in (i) may be applied to show $z \in [R(t)]^\circ$. Thus we assume $\sin(\beta/2) \sin((\beta - \gamma)/2) \sin \gamma/2 \neq 0$. We perturb C_1, C_2, C_3 as above, i.e., perturb the length of C_1 by u , C_2 by w and C_3 by $-(u + w)$. In this case J may be calculated from (9) with $b = c = \rho$ and allowing a to approach infinity. (The proof of this statement is left to the reader as is the verification that proceeding to the limit does not interfere with continuity of partial derivatives. This will also be the case in parts (iii) and (iv) of the proof of this proposition.)

$$\begin{aligned} J &= \lim_{a \rightarrow \infty} \frac{4(a + \rho)2\rho}{a\rho} \sin\left(\frac{\beta - \gamma}{2}\right) \sin\frac{\beta}{2} \sin\frac{\gamma}{2} \\ &= 8 \sin\left(\frac{\beta - \gamma}{2}\right) \sin\frac{\beta}{2} \sin\frac{\gamma}{2} \neq 0. \end{aligned}$$

Hence $z \in [R(t)]^\circ$ by the inverse function theorem.

(iii) Let C_1, C_2 be critical circular arcs of length $\rho\alpha, \rho\beta$, respectively, and let C_3 be a straight segment of length l . If $\beta = 2k\pi$, $\mathcal{L}(z, \theta, vt)$ has an $(n-1)$ -critical curve, so we may assume $\sin(\beta/2) \neq 0$. We perturb a as in the other cases and calculate J from (9) with $a = b = \rho$ and allowing c to approach infinity while keeping $c\gamma = l$. In fact, $J = (4l/\rho) \sin^2(\beta/2) \neq 0$ and $z \in [R(t)]^\circ$ as required.

(iv) Let C_1, C_3 be line segments of length l_1, l_3 and C_2 be a critical circular arc of length $\rho\beta$. If $\beta = 2k\pi$, then $\mathcal{L}(z, \theta, vt)$ contains an $(n-1)$ -critical curve, so we may assume $\sin(\beta/2) \neq 0$. We use a similar perturbation as in the other cases and calculate J from (9) with $b = \rho$ and allowing a, c to approach infinity while $c\gamma$ stays constant at l_3 . In this case, $J = (2l_3/\rho) \sin^2(\beta/2) \neq 0$ and $z \in [R(t)]^\circ$ as required. This completes the proof of Proposition 11.

The following theorem now follows immediately from Theorem 1 and Propositions 10 and 11. We note that it generalizes a result of Melzak [8, Thm. 1].

THEOREM 3. *If z is a boundary point of $R(t)$, then $\mathcal{L}(z, *, vt)$ contains a critical admissible curve of type CC or CL .*

It is possible to strengthen Theorem 3, since not all curves with length vt of type CC and CL have endpoints in the boundary of $R(t)$. We define the following two classes of curves.

$\mathcal{A}(t)$ consists of the CC admissible curves in $\mathcal{L}(*, *, vt)$ having endpoint z and circular arcs C_1, C_2 of length $\rho\alpha$ and $\rho\beta$ satisfying:

(i) $\alpha \leq \pi/2$,

(ii) $\beta \leq 2\pi$,

(iii) z and the center of C_1 are not on the same side of the y -axis.

$\mathcal{B}(t)$ consists of the CL admissible curves in $\mathcal{L}(*, *, vt)$ having endpoint z and circular arc C_1 of length $\rho\alpha$ satisfying:

(i) $\alpha \leq 2\pi$,

(ii) z and the center of C_1 are not on opposite sides of the y -axis.

THEOREM 4. *If z is a boundary point of $R(t)$, then $\mathcal{L}(z, *, vt)$ contains either a curve in $\mathcal{A}(t)$ or a curve in $\mathcal{B}(t)$.*

Proof. Any CC curve which does not satisfy (i) or (iii) in the definition of $\mathcal{A}(t)$ has its endpoint z in $[R(t)]^\circ$ by the reflection process described in part (i) of the proof of Proposition 11. If $\beta > 2\pi$, say, $\beta = 2k\pi + \beta'$ where $0 < \beta' < 2\pi$, then $\mathcal{L}(z, *, vt)$ contains another CC curve having circular arcs of length $\rho\alpha' = \rho(\alpha + 2k\pi)$ and $\rho\beta'$. But since $\alpha' > \pi/2$, $z \in [R(t)]^\circ$.

Let a CL admissible curve Z with endpoint z and circular arc C_1 with length $\rho\alpha$ have $\alpha > 2\pi$ (say $\alpha = \alpha' + 2k\pi$), where $0 < \alpha' < 2\pi$. Then $\mathcal{L}(z, *, vt)$ also contains a CCL curve where the circular arcs have lengths $\rho 2k\pi$ and $\rho\alpha'$. By Proposition 11, $z \in [R(t)]^\circ$.

One tangent from z to C_1 is the straight segment of Z itself. If z and the center of C_1 are on opposite sides of the y -axis, there is a second tangent τ from z to C_1 with point of contact T . Let Y be the reflection in τ of the portion of Z from T to z . The portion of Z from 0 to T united with Y is a CCL admissible curve in $\mathcal{L}(z, *, vt)$ and from what we have already proved, the second circular arc is not a multiple of 2π . Hence from Proposition 11, $z \in [R(t)]^\circ$.

The required result now follows from Theorem 3.

3.2. The regions. Let S_1 (S_2) denote the circle of radius ρ tangent to the y -axis at 0 from the right (left). The locus $E(t)$ of endpoints of admissible CL curves of length vt may be obtained as follows. Imagine a thin inextensible string of length vt with one end fixed at 0 and the other end Q at the point $(0, vt)$. Wrap the string around S_1 (S_2) keeping it taut until Q is on S_1 (S_2). Let $E_1(t)$ ($E_2(t)$) be the path traced by Q . Then $E(t)$ is the union of $E_1(t)$ and $E_2(t)$. $E_1(t)$ has parametric equations

$$(10) \quad \begin{aligned} x(\theta) &= \rho(1 - \cos \theta) + (vt - \rho\theta) \sin \theta, \\ y(\theta) &= \rho \sin \theta + (vt - \rho\theta) \cos \theta, \end{aligned}$$

where $0 \leq \theta \leq vt/\rho$. The locus $B(t)$ of endpoints of curves in $\mathcal{B}(t)$ is an easily identified subset of $E(t)$. Figure 9 shows $E(t)$ for $vt = 3\pi\rho/2$ and in this case $B(t) = E(t)$.

For all t , the locus $A(t)$ of endpoints of curves in $\mathcal{A}(t)$ is contained in the set of points on two congruent cardioids whose positions depend on t , but whose dimensions are independent of t . Let R_1 , R_2 be the endpoints of the admissible circular arcs of length vt on S_1 , S_2 , respectively. Roll the circle S_1 (S_2) without slipping around the circle S_2 (S_1). R_1 (R_2) traces a cardioid $\Sigma_1(t)$ ($\Sigma_2(t)$) whose vertex is R_2 (R_1) and whose axis is along the diameter of S_2 (S_1) through R_2 (R_1). $\Sigma_1(t)$ has parametric equations

$$(11) \quad \begin{aligned} x(\theta) &= \rho\{2 \cos \theta - 1 - \cos(2\theta - vt/\rho)\}, \\ y(\theta) &= \rho\{2 \sin \theta - \sin(2\theta - vt/\rho)\}. \end{aligned}$$

$A(t)$ is an easily identified subset of $\Sigma_1(t) \cup \Sigma_2(t)$.

An example is given in Fig. 10 in which we draw $\Sigma_1(2\pi\rho/3v)$. $A(t)$ is the union of the dotted portion of $\Sigma_1(2\pi\rho/3v)$ and the similar portion of $\Sigma_2(2\pi\rho/3v)$. (Alternatively, by symmetry, $A(t)$ is this dotted portion together with its reflection in the y -axis.)

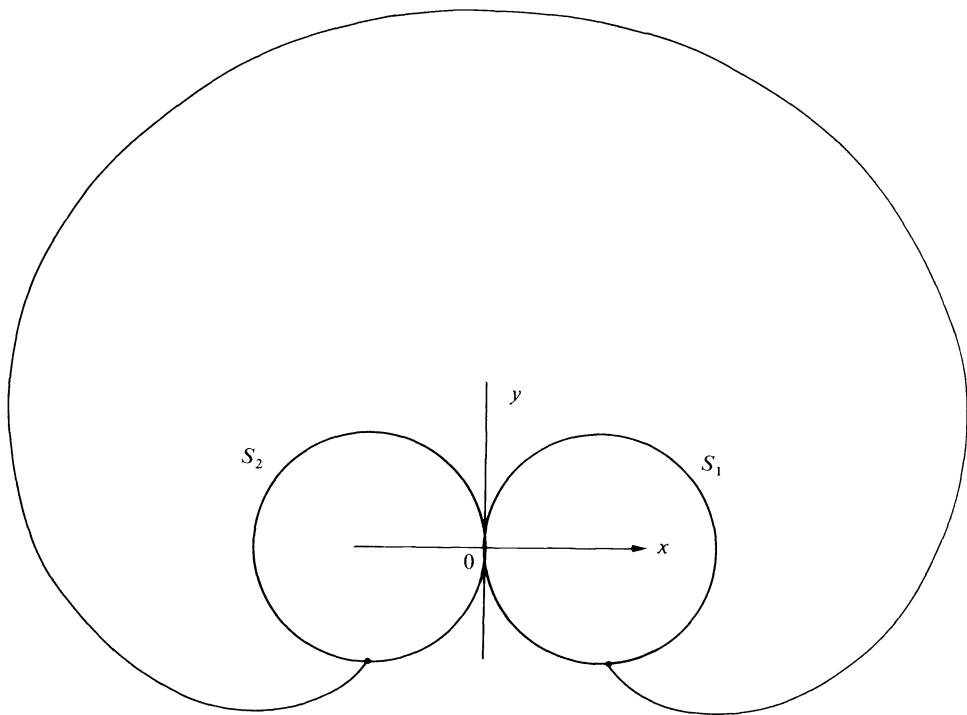


FIG. 9. $E(3\pi\rho/2v)$

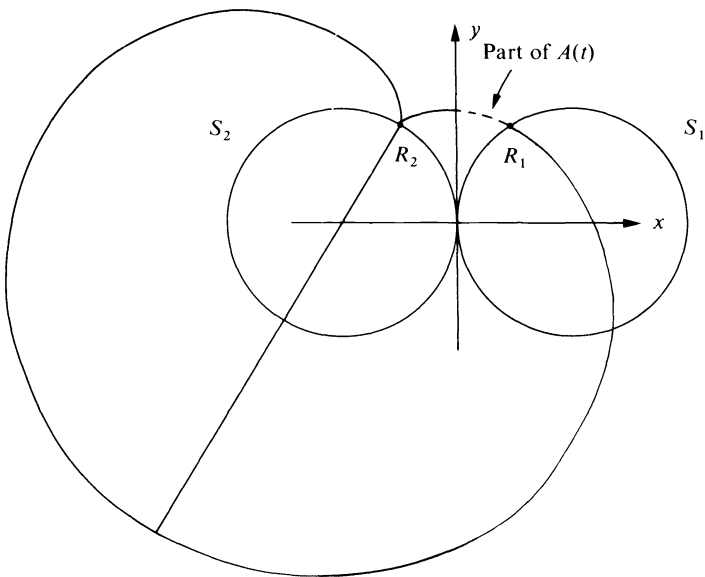


FIG. 10. $\Sigma_1(2\pi\rho/3v)$

The regions $R(t)$ are now uniquely determined by the following two facts:

(i) $R(t)$ is closed and bounded, with the boundary contained in $A(t) \cup B(t)$ and (ii) as t increases, $R(t)$ varies "continuously".

Using equations (10) and (11), the computer plotter was programmed to draw the regions for a selection of values of t . The results are shown in the Appendix. We have been compelled by the size of paper to reduce the scale as t increases. S_1 has been drawn in some of the diagrams for comparison. We note that the results for $t \leq \pi\rho/v$ are precisely those obtained by Melzak [8]. It is a simple matter to show that $t = \pi\rho/v$ is the first time at which $R(t)$ intersects $S_1 \cup S_2$ thus verifying Robertson's results [9].

3.2.1. Connectivity of the regions. For $0 < t < (\rho/v)(3\pi/2 + 1)$, $R(t)$ is simply connected. When $t = (\rho/v)(3\pi/2 + 1)$, $B(t)$ touches itself on the y -axis and $R(t)$ becomes doubly connected (see Appendix, Fig. A.5). By continuity, this situation remains until $t = 2\pi\rho/v$ and during the interval $[(\rho/v)(3\pi/2 + 1), 2\pi\rho/v]$ the internal boundary is formed by endpoints of both CC and CL admissible curves (for example, Fig. A.6). For times slightly greater than $2\pi\rho/v$, $R(t)$ is doubly connected but the internal boundary is completely comprised of endpoints of CC curves (e.g., Fig. A.8). The internal "hole" is shrinking as t increases and finally disappears at t^* when the cardioids $\Sigma_1(t^*)$ and $\Sigma_2(t^*)$ touch the y -axis. Figure 11 shows $\Sigma_1(t)$ at a time slightly less than t^* . The dotted portion and its reflection in the y -axis form the inner boundary. The curve drawn in heavily is a typical CC curve whose endpoint is on the inner boundary. For times greater than t^* , $R(t)$ is

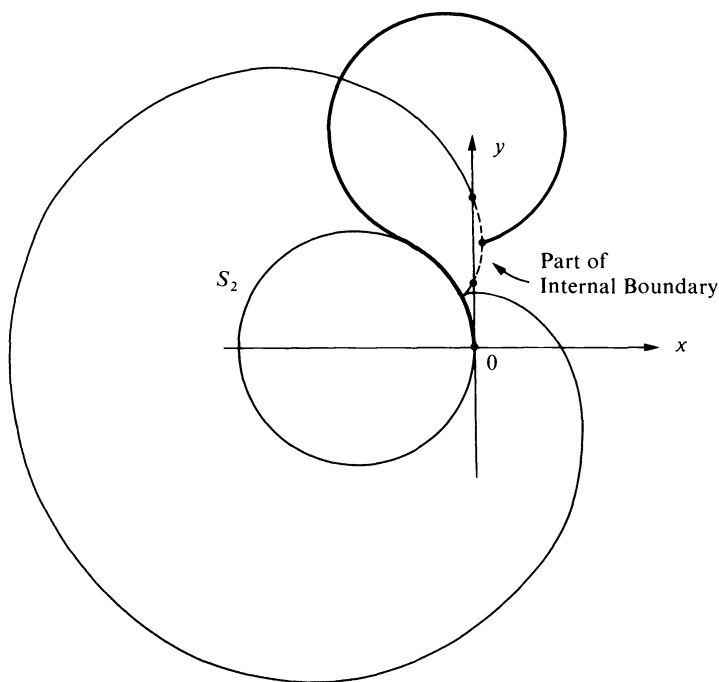


FIG. 11. Internal boundary of $R(t)$ for t slightly less than t^*

simply connected and, of course, approaches a circle of radius vt with center 0 as t gets large. We finally calculate t^* .

Using (11), $\Sigma_1(t)$ and $\Sigma_2(t)$ touch the y -axis, where

$$(12) \quad x(\theta) = \rho \left\{ 2 \cos \theta - 1 - \cos \left(2\theta - \frac{vt}{\rho} \right) \right\} = 0$$

and

$$(13) \quad \frac{d}{d\theta}(x(\theta)) = \rho \left\{ -2 \sin \theta + 2 \sin \left(2\theta - \frac{vt}{\rho} \right) \right\} = 0$$

are simultaneously true. From (13),

$$\sin \left(\frac{\theta}{2} - \frac{vt}{2\rho} \right) \cos \left(\frac{3\theta}{2} - \frac{vt}{2\rho} \right) = 0.$$

We are searching for a t^* satisfying $2\pi < vt^*/\rho$ and by Theorem 4, $\theta \leq \pi/2$ and $vt^*/\rho \leq 2\pi + \pi/2$. Moreover, $\theta = \pi/2$, $vt^*/\rho = 2\pi + \pi/2$ does not satisfy (12), so we may assume $\sin(\theta/2 - vt/\rho) \neq 0$ and hence deduce $\cos(3\theta/2 - vt/\rho) = 0$. Therefore $vt/\rho = 3\theta - (2n+1)\pi$ and the constraints on θ, t^* force the choice of $n = -1$, i.e., $vt/\rho = 3\theta + \pi$ or

$$(14) \quad vt/\rho = 2\pi + \lambda, \quad \text{where } \lambda = 3\theta - \pi.$$

Substituting (14) into (12) we obtain

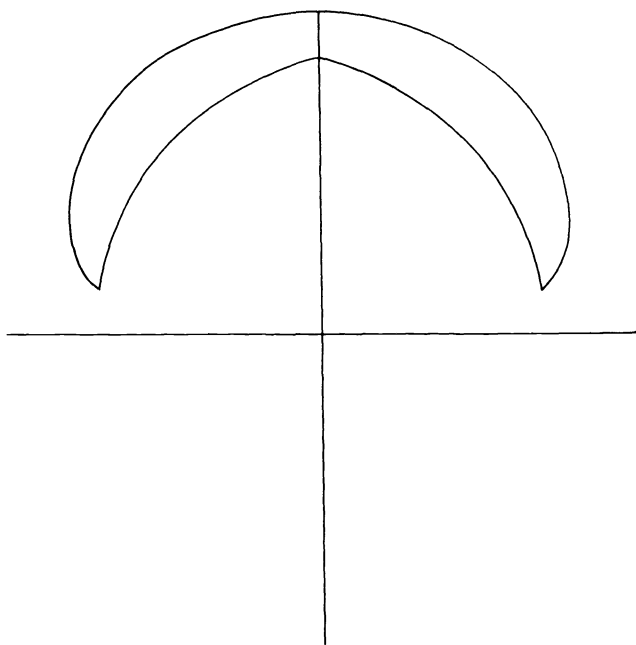
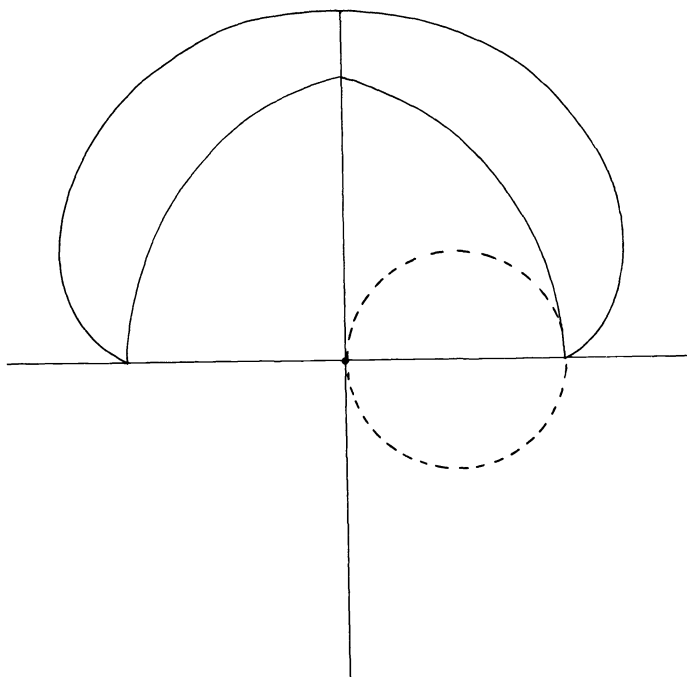
$$2 \cos \theta - 1 - \cos(2\theta - \lambda) = 0$$

or

$$2 \cos \theta - 1 - \cos(\pi - \theta) = 0.$$

Therefore $3 \cos \theta = 1$ and from (14), $\cos \lambda = -\cos 3\theta = 3 \cos \theta - 4 \cos^3 \theta = 23/27$. Thus t^* is given by $vt^*/\rho = 2\pi + \cos^{-1} 23/27$.

Acknowledgments. The authors wish to thank Professor Z. A. Melzak of University of British Columbia, Dr. Turan Onat, Yale University, and Dr. Brian Davies, Oxford University for stimulating and profitable conversations concerning the content of this paper.

Appendix. Diagrams of $R(t)$ drawn by computer plotter.FIG. A.1. *Region $R(t)$ for $vt/\rho = 7\pi/8$* FIG. A.2. *Region $R(t)$ for $vt/\rho = \pi$*

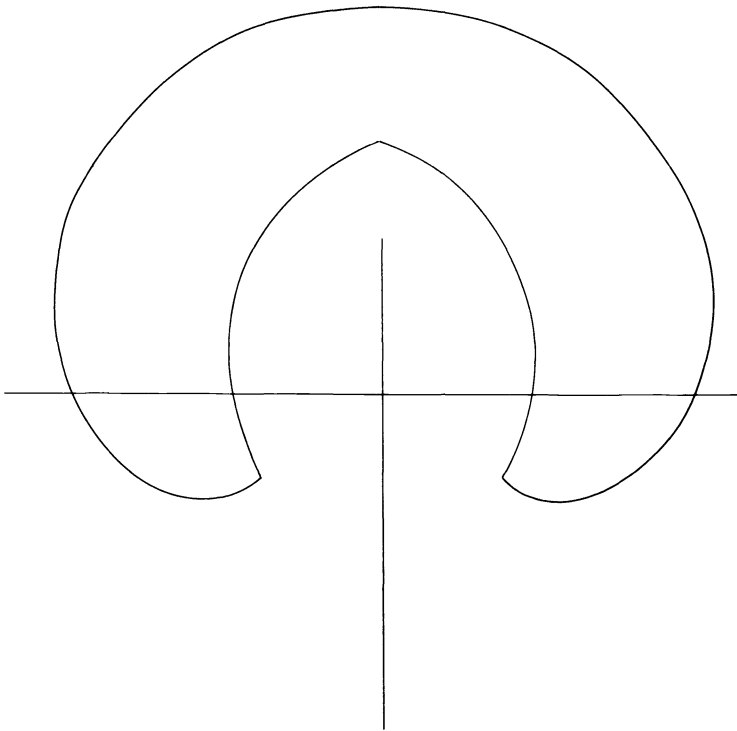


FIG. A.3. *Region $R(t)$ for $vt/\rho = 11\pi/8$*

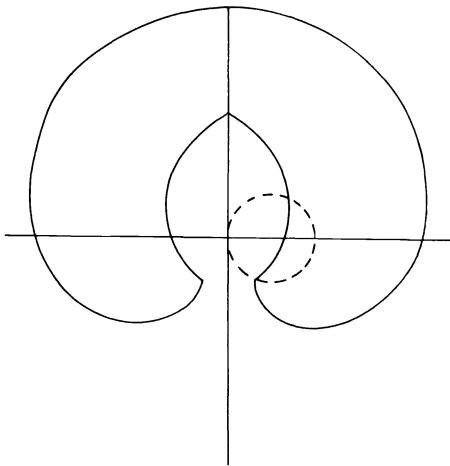
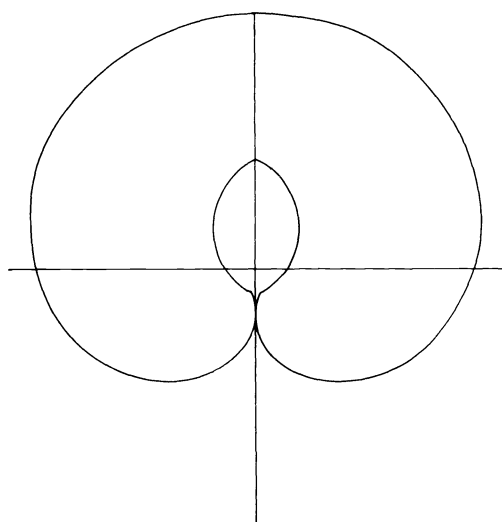
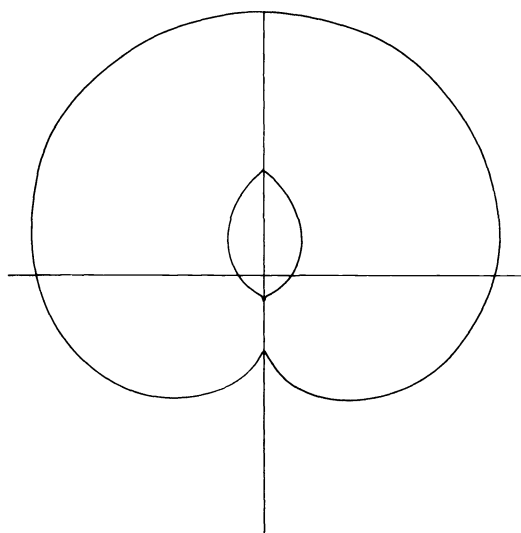


FIG. A.4. *Region $R(t)$ for $vt/\rho = 13\pi/8$*

FIG. A.5. *Region $R(t)$ for $vt/\rho = 3\pi/2 + 1$* FIG. A.6. *Region $R(t)$ for $vt/\rho = 15\pi/8$*

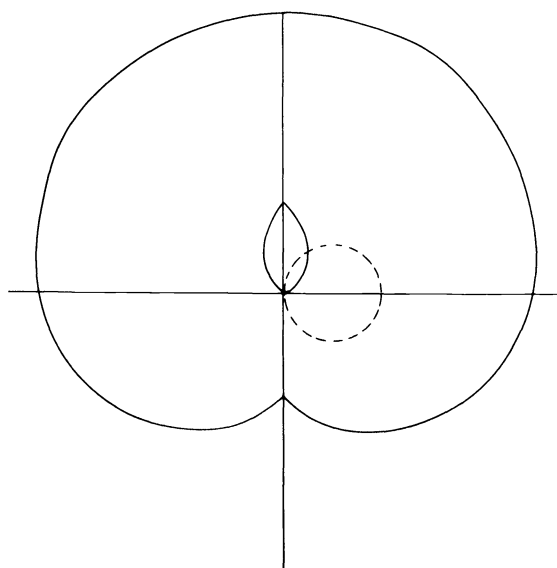


FIG. A.7. *Region $R(t)$ for $vt/\rho = 2\pi$*

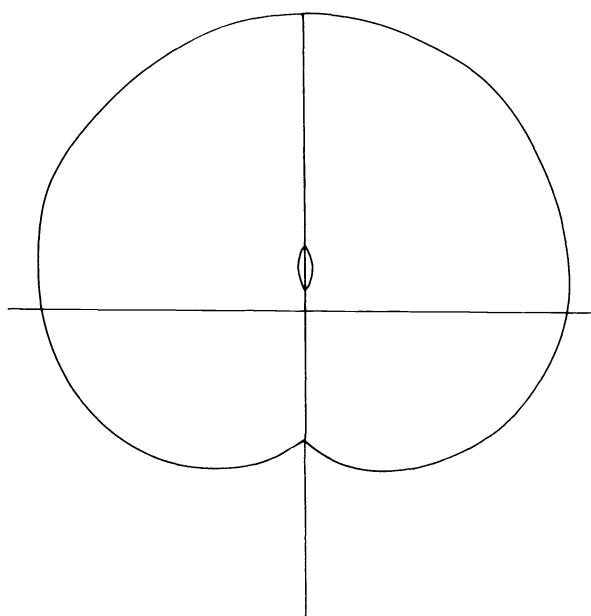
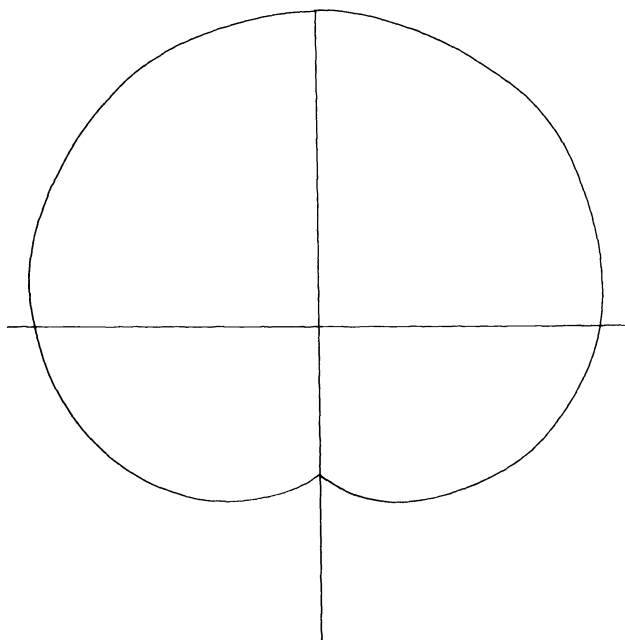


FIG. A.8. *Region $R(t)$ for $vt/\rho = 17\pi/8$*

FIG. A.9. Region $R(t)$ for $vt/\rho = 9\pi/4$

REFERENCES

- [1] T. M. APOSTOL, *Mathematical Analysis*, Addison-Wesley, Reading, Mass., 1957.
- [2] E. J. COCKAYNE, *Plane pursuit with curvature constraints*, SIAM J. Appl. Math., 15 (1967), pp. 1511–1516.
- [3] L. E. DUBINS, *On curves of minimal length with a constraint on average curvature and with prescribed initial and terminal positions and tangents*, Amer. J. Math., 79 (1957), pp. 497–516.
- [4] ———, *On plane curves with curvature*, Pacific J. Math., 11 (1961), pp. 471–482.
- [5] A. W. GOODMAN, *A partial differential equation and parallel plane curves*, Amer. Math. Monthly, 71 (1964), pp. 257–264.
- [6] R. ISAACS, *Differential Games*, John Wiley, New York, 1965.
- [7] A. N. KOLMOGOROV AND S. V. FOMIN, *Elements of the Theory of Functions and Functional Analysis*, vol. 1, Graylock Press, Rochester, N.Y., 1957.
- [8] Z. A. MELZAK, *Plane motion with curvature limitations*, J. Soc. Indust. Appl. Math., 9 (1961), pp. 422–432.
- [9] H. G. ROBERTSON, *Curvature and arc length*, SIAM J. Appl. Math., 19 (1970), pp. 697–699.

H^2 -FUNCTIONS AND INFINITE-DIMENSIONAL REALIZATION THEORY*

J. S. BARAS† AND R. W. BROCKETT‡

Abstract. In this paper the realization question for infinite-dimensional linear systems is examined for both bounded and unbounded operators. In addition to obtaining realizability criteria covering the basic cases, we discuss the relationship between canonical realizations of the same system. What one finds is that the set of transfer functions which are realizable by triples (A, b, c) with A bounded is related in a close way to the space of complex functions analytic and square integrable on the disk $|s| < 1$, and that the set of transfer functions which are realizable by triples (A, b, c) with A unbounded but generating a strongly continuous semigroup is related in a close way to functions analytic and square integrable on a half-plane. This relation makes possible a deeper study between the transfer function and the models which realize it. Some examples illustrate the results and their applications.

1. Preliminaries and notation. In this paper we study realization theory for a class of infinite-dimensional linear systems. On one hand our motivation comes from a desire to understand engineering problems involving transmission lines, elastic deformations, moving fluids, and related matters, where the assumption of finite-dimensionality is too restrictive; on the other hand, we want to see the finite-dimensional results themselves as part of a larger picture.

For the sake of definiteness we work in the most basic Hilbert space,

$$l_2(\mathbb{Z}^+) = \{\{a_i\}, i = 1, 2, 3, \dots, \text{ such that } \{a_i\} \text{ is a square summable sequence}\}.$$

This makes possible a fairly direct comparison with many well-known results concerning the finite-dimensional case. The problem is to express a given real function T defined on $[0, \infty)$ as $T(t) = c[e^{At}b]$, or to express its Laplace transform $\tilde{T}(s)$ as $c[(Is - A)^{-1}b]$ in some appropriately defined region of the complex plane.

We consider several distinct, but related cases. The first centers around the existence of realizations (A, b, c) with A a bounded operator on $l_2(\mathbb{Z}^+)$, b an element of $l_2(\mathbb{Z}^+)$ and c a bounded linear functional on $l_2(\mathbb{Z}^+)$. We call such triples *bounded realizations*. We call a triple (A, b, c) a *regular realization* if A is the infinitesimal generator of a strongly continuous semigroup of bounded operators $\{e^{At}\}$ on $l_2(\mathbb{Z}^+)$, b is an element of $l_2(\mathbb{Z}^+)$ and c is a bounded linear functional on $l_2(\mathbb{Z}^+)$. In both cases above the output can also be expressed, as is well known, as the inner product of $x(t)$ with some element of $l_2(\mathbb{Z}^+)$ which is uniquely determined by the functional c , and which we denote also by c ; i.e., we shall write $y(t) = c[x(t)] = \langle c, x(t) \rangle$.

We also consider cases where A is the infinitesimal generator of a strongly continuous semigroup of bounded operators on $l_2(\mathbb{Z}^+)$, b is restricted to belong to the domain of A (written $\mathcal{D}_0(A)$) but c is a linear functional defined on $\mathcal{D}_0(A)$ and such that $|c(x)| \leq k(\|Ax\| + \|x\|)$ for all $x \in \mathcal{D}_0(A)$ and some constant k . Such

* Received by the editors April 16, 1973, and in revised form November 27, 1973. This work was supported by the U.S. Office of Naval Research under the Joint Electronics Program by Contract N00014-67-A-0298-0006.

† Division of Engineering and Applied Physics, Harvard University, Cambridge, Massachusetts. Now at Department of Electrical Engineering, University of Maryland, College Park, Maryland 20742.

‡ Division of Engineering and Applied Physics, Harvard University, Cambridge, Massachusetts 02138.

realizations will be called *balanced realizations*. They have important properties not shared by regular realizations.

The triple (A, b, c) is called a realization for the weighting pattern $T(t)$ if and only if $T(t) = c[e^{At}b]$. The system theoretic interpretation of this equation in terms of “external” and “internal” descriptions of time-invariant linear systems with scalar input, scalar output is considered to be well known. The fact that we are using $l_2(\mathbb{Z}^+)$ as our state space is not very restrictive, since any separable Hilbert space is isometrically isomorphic to $l_2(\mathbb{Z}^+)$.

In order to describe what realizations realize what weighting patterns, we need to introduce some notation. The open disk of radius ρ is denoted by $\mathbb{D}_\rho = \{s \mid |s| < \rho\}$. We write \mathbb{D} for \mathbb{D}_1 . The boundary of \mathbb{D} , the unit circle, is denoted by \mathbb{T} . By $H^2(\mathbb{D})$ we mean the set of complex-valued functions which are holomorphic in \mathbb{D} and have a Taylor series about zero with square summable coefficients. The space $H^2(\mathbb{D}_\rho)$ is defined by saying that $\psi(s)$ belongs to $H^2(\mathbb{D}_\rho)$ if and only if $\psi(s/\rho)$ belongs to $H^2(\mathbb{D})$. By $L^2(\mathbb{T})$ we mean the set of complex-valued functions which are defined and square integrable, in the Lebesgue sense, on the unit circle. By $H^2(\mathbb{T})$ we mean the subspace of $L^2(\mathbb{T})$ of functions with vanishing negative Fourier coefficients. $H^2(\mathbb{D})$ and $H^2(\mathbb{T})$ are related by the fact that for any function in $H^2(\mathbb{D})$ the radial limits from within the disk $\lim_{r \rightarrow 1} \psi(re^{i\theta}) = \phi(\theta)$ exist for almost all θ and give an element ϕ of $H^2(\mathbb{T})$. This correspondence is, moreover, one-to-one and onto so that $H^2(\mathbb{D})$ and $H^2(\mathbb{T})$ are closely related indeed. In fact, the Fourier coefficients of ϕ are the Taylor coefficients of ψ . In addition, $H^2(\mathbb{D})$ is a Hilbert space with the inner product

$$\langle \psi_1, \psi_2 \rangle = \sum_{n=0}^{\infty} \bar{\alpha}_n \beta_n,$$

where $\psi_1(s) = \sum_{n=0}^{\infty} \alpha_n s^n$ and $\psi_2(s) = \sum_{n=0}^{\infty} \beta_n s^n$. This makes $H^2(\mathbb{D})$, $H^2(\mathbb{T})$ and $l_2(\mathbb{Z}^+)$ isomorphic as Hilbert spaces with the isomorphisms defined by

$$(a_0, a_1, a_2, \dots) \leftrightarrow \sum_{i=0}^{\infty} a_i s^i \leftrightarrow \sum_{n=0}^{\infty} a_n e^{in\theta}.$$

We denote by \prod_ρ^+ the half-plane $\operatorname{Re} s > \rho$. We understand by $H^2(\prod_\rho^+)$ the space of functions which are analytic in \prod_ρ^+ and square integrable along vertical lines in \prod_ρ^+ such that

$$\sup_{x > \rho} \int_{-\infty}^{+\infty} |\psi(x + iy)|^2 dy \leq M < \infty.$$

The relationship between $H^2(\mathbb{D})$ and $H^2(\prod^+)$ is this: $\phi(\cdot) \in H^2(\prod^+)$ if and only if ψ defined by

$$\psi(s) = \frac{1}{s-1} \phi\left(\frac{s+1}{s-1}\right)$$

belongs to $H^2(\mathbb{D})$. (See Hoffman [7, p. 130].)

We would like to recall some of the facts from Fourier transform theory that involve $H^2(\prod^+)$ and especially the Paley–Wiener theorem. We denote by \mathbb{I} the imaginary axis in the complex plane. It is well known that the Fourier transform

$$g(t) \xrightarrow{\mathcal{F}} \int_{-\infty}^{\infty} e^{-i\omega t} g(t) dt = G(i\omega)$$

is a unitary map between $L_2(-\infty, \infty)$ and $L_2(\mathbb{I}, d\omega/2\pi)$. Consider $L_2(0, \infty)$ as the subspace of $L_2(-\infty, \infty)$ of functions which vanish on $(-\infty, 0)$, and $L_2(-\infty, 0)$ as

the subspace of $L_2(-\infty, \infty)$ of functions which vanish on $(0, \infty)$. Then obviously $L_2(-\infty, 0) = L_2(0, \infty)^\perp$ in $L_2(-\infty, \infty)$. Moreover, if we let $H^2(\mathbb{I}) = \mathcal{F}L_2(0, \infty)$ and $\tilde{H}^2(\mathbb{I}) = \mathcal{F}L_2(-\infty, 0)$, we see that $H^2(\mathbb{I})^\perp = \tilde{H}^2(\mathbb{I})$. $H^2(\mathbb{I})$ consists exactly of the boundary values of the elements of $H^2(\mathbb{I}^+)$ (which exist for almost all ω). Moreover, if \mathcal{L} denotes the Laplace transform

$$g(t) \xrightarrow{\mathcal{L}} \int_0^\infty e^{-st} g(t) dt = G(s) \quad \text{for } g \in L_2(0, \infty),$$

$$f(t) \xrightarrow{\mathcal{L}} \int_{-\infty}^0 e^{-st} f(t) dt = F(s) \quad \text{for } f \in L_2(-\infty, 0),$$

the Paley–Wiener theorem says that $H^2(\mathbb{I}^+) = \mathcal{L}L_2(0, \infty)$. If we let \mathbb{I}^- denote the half-plane $\operatorname{Re} s < 0$, then also $H^2(\mathbb{I}^-) = \mathcal{L}L_2(-\infty, 0)$. Moreover, $\tilde{H}^2(\mathbb{I})$ consists exactly of the boundary values of the elements of $H^2(\mathbb{I}^-)$. The relation between $H^2(\mathbb{I}^+)$ and $H^2(\mathbb{I}^-)$ is simple. A function $f(s)$ belongs to $H^2(\mathbb{I}^+)$ if and only if $\overline{f(-\bar{s})}$ belongs to $H^2(\mathbb{I}^-)$. Then obviously we see that $\tilde{H}^2(\mathbb{I}) = \overline{H^2(\mathbb{I})}$.

As we were preparing the original version of this paper we received from Paul Fuhrmann a manuscript [13] which analyzes the bounded case and obtains a number of the results described here with certain small changes due to the fact that he works with discrete time systems. Helton [14] also investigates some questions of this type but emphasizes a different class of ideas. A result similar to our Theorem 4 appears in Balakrishnan [23].

2. Realizability criteria, bounded case. In this section we characterize the class of weighting patterns which admit bounded realizations.

Let $T: [0, \infty) \rightarrow \mathbb{R}^1$ be a continuous function of time. When can it be written as

$$T(t) = \langle c, e^{At}b \rangle,$$

where $b, c \in l_2(\mathbb{Z}^+)$ and $A: l_2(\mathbb{Z}^+) \rightarrow l_2(\mathbb{Z}^+)$ is bounded? As is well known such a representation is possible for T with $[A, b, c]$ all finite-dimensional if and only if T is of the exponential order and its transform

$$\tilde{T}(s) = \int_0^\infty e^{-st} T(t) dt, \quad \operatorname{Re} s > \sigma_0,$$

is rational. In the present case A is bounded; $\{e^{At}\}$ defines a uniformly continuous semigroup of operators (see [1, p. 626]), and since b and c belong to $l_2(\mathbb{Z}^+)$ we have

$$\langle c, e^{At}b \rangle \leq \|b\| \cdot \|c\| \cdot M \cdot e^{\|A\| |t|}, \quad t \in \mathbb{R}^1,$$

where $\|e^{At}\| \leq M e^{\|A\| |t|}$, and the norms are $l_2(\mathbb{Z}^+)$ and induced $l_2(\mathbb{Z}^+)$ respectively. Thus the class we are looking for includes only functions of exponential order. Moreover, since A is bounded, $\langle c, e^{At}b \rangle$ is an entire function.

The following two theorems characterize in the time and frequency domain the set of realizable input–output maps.

THEOREM 1. $T: [0, \infty) \rightarrow \mathbb{R}^1$ has a bounded realization if and only if T is an entire function of exponential order.¹

¹ This is a standard engineering term; in the mathematical literature this is called “exponential type”.

Proof. The necessity follows from the above. For the sufficiency, since T is entire it has a power series expansion

$$T(t) = \sum_{n=0}^{\infty} c_n t^n$$

valid in the finite complex plane. Let σ_0 be the exponential order of $T(\cdot)$. Then $\overline{\lim}_{n \rightarrow \infty} (n!|c_n|)^{1/n} = \sigma_0$ (see [20, p. 95]). So for $k > \sigma_0$ we have

$$n!|c_n|/k^n \leq (\sigma_0/k)^n$$

and consequently the sequence $\{n!|c_n|/k^n\}_{n=0}^{\infty} \in l_2(\mathbb{Z}^+)$. Now take

$$A = k \begin{bmatrix} 0 & & & & \\ 1 & 0 & & & \\ 0 & 1 & 0 & & \\ & 0 & 1 & 0 & \\ & & 0 & 1 & 0 \ddots \\ & & & 0 & \ddots \ddots \end{bmatrix},$$

$$b = \{1, 0, 0, \dots\},$$

$$c = \{c_0, c_1/k, \dots, n!c_n/k^n, \dots\}$$

and this completes the proof.

Now using Laplace transforms in the complex domain we pass from the equation $T(t) = \langle c, e^{At}b \rangle$ to the equation $\tilde{T}(s) = \langle c, (Is - A)^{-1}b \rangle$ for $\operatorname{Re} s > \|A\|$. Since A is bounded using an elementary analytic continuation argument we see that $\tilde{T}(s)$ is analytic for $|s| > \|A\|$ and also that $\tilde{T}(\infty) = 0$. Hence $\tilde{T}(s) = \langle c, b \rangle s^{-1} + \langle c, Ab \rangle s^{-2} + \langle c, A^2b \rangle s^{-3} + \dots$ for $|s| > \|A\|$.

THEOREM 2. *The function $T: [0, \infty) \rightarrow \mathbb{R}^1$ has a bounded realization if and only if the Laplace transform $\tilde{T}(\cdot)$ of $T(\cdot)$ is analytic at infinity and vanishes there.*

Proof. The necessity follows clearly from the above. For the sufficiency since $\tilde{T}(s)$ is analytic at infinity and vanishes there, it has a power series expansion

$$\tilde{T}(s) = \sum_{i=0}^{\infty} a_i s^{-(i+1)} \quad \text{for } |s| > c$$

for some finite c . Then for $k > c$ we have that the sequence $\{|a_i|/k^i\} \in l_2(\mathbb{Z}^+)$. So again take

$$A = k \begin{bmatrix} 0 & & & & \\ 1 & 0 & & & \\ 0 & 1 & 0 & & \\ & 0 & 1 & 0 & \\ & & 0 & 1 & 0 \ddots \\ & & & 0 & \ddots \ddots \end{bmatrix},$$

$$b = \{1, 0, 0, \dots\},$$

$$c = \{a_0, a_1/k, a_2/k^2, \dots\},$$

and this completes the proof.

Remark 2.1. It is clear from the above that the singularities of $\tilde{T}(\cdot)$ must be contained in a compact set of the complex plane. Hence if the singularities include branch points, in order for $\tilde{T}(\cdot)$ to have a bounded realization it must be possible to consider the branch cuts in the finite plane (since any $\tilde{T}(\cdot)$ which has a bounded realization is analytic at infinity). For example, a transfer function with a single branch point does not have a bounded realization. On the other hand, there are many cases of complex-valued functions with branch points for which the branch cuts can be taken either in the finite plane or through infinity (e.g., $1/\sqrt{s^2 + 1}$ is typical).

Suppose we have a physical system (and perhaps a model based on the underlying physical theories) with a transfer function which has this property, i.e., we can consider the branch cuts either in the finite plane or through infinity. We consider the question of whether or not this particular transfer function has a bounded realization. Now in most physical problems the position of the branch cuts is implied by physical or asymptotic conditions (see the first few pages of Noble's book [25] on the Wiener–Hopf technique for an example). Therefore if these conditions exclude the possibility of considering the branch cuts in the finite plane, then one concludes immediately that there is no bounded realization for this transfer function.

Remark 2.2. It is apparent from the above that if $T(\cdot)$ has any bounded realization, it can be realized by a multiple of the unilateral shift in $l_2(\mathbb{Z}^+)$ or by a multiple of the bilateral shift in $l_2(\mathbb{Z})$. To see the latter take

$$A = k \begin{bmatrix} \ddots & & & & \\ & 0 & & & \\ \ddots & & 1 & 0 & \\ & \ddots & & 0 & 1 & 0 \\ & & & 0 & 1 & 0 \\ & & & & 0 & 1 & 0 \\ & & & & & 0 & 1 \\ & & & & & & 0 \end{bmatrix},$$

$$b = \{\cdots, 0, 1, 0, 0, \cdots\},$$

$$c = \{\cdots, 0, 0, c_0, c_1/k, \cdots, n!c_n/k^n\}$$

(with $\{c_i\}$ and k as in Theorem 1). This is not surprising in view of the fact that the shift can be considered as a “universal model” for bounded operators in Hilbert space (cf. [9]).

We would like to discuss in more detail the relation $\tilde{T}(s) = \langle c, (Is - A)^{-1}b \rangle$ for $\operatorname{Re} s > \gamma$, where γ is large enough, and draw some conclusions from it. Since A here is a bounded operator the point at infinity is in the resolvent set of A , denoted $\rho(A)$. We denote by $\rho_0(A)$ the connected component of $\rho(A)$ containing the point at infinity, by $\sigma(A)$ the spectrum of A and by $\sigma_0(A)$ the complement in \mathbb{C} (the complex plane) of $\rho_0(A)$. The function $\langle c, (Is - A)^{-1}b \rangle$ is obviously analytic

for $|s| > \|A\|$ and thus for $\operatorname{Re} s > \|A\|$. $T(\cdot)$ is of exponential order, say $|T(t)| \leq M e^{\sigma_0 t}$; then $\tilde{T}(s)$ is analytic for $\operatorname{Re} s > \sigma_0$ (and also for $|s| > \sigma_0$) (see [20, p. 95]). Since $T(t) = \langle c, e^{At}b \rangle$ we see that $\sigma_0 \leq \|A\|$. So from the equation $\tilde{T}(s) = \langle c, (Is - A)^{-1}b \rangle$ valid for $\operatorname{Re} s > \|A\|$ we deduce by analytic continuation that $\tilde{T}(\cdot)$ is analytic for all $s \in \rho_0(A)$.

If we let $\sigma(\tilde{T}) = \{s \in \mathbb{C} | \tilde{T}(\cdot) \text{ is not analytic at } s\}$ we deduce that for any bounded realization (A, b, c) of $T(\cdot)$ we must have

$$\sigma(\tilde{T}) \subseteq \sigma_0(A).$$

This relation will be referred in the sequel as the *spectral inclusion property*. For example, the function $T(t) = e^{t/2}$ can obviously be realized by the unilateral shift as above. Then for A being the unilateral shift we have $\sigma(A) = \mathbb{D} \cup \mathbb{T}$ (i.e., the closed disk). But $\tilde{T}(\cdot)$ has just a pole at $s = \frac{1}{2}$. Consider now

$$T(t) = \sum_{n=0}^{\infty} \frac{t^{2^n - 1}}{(2^n - 1)!}.$$

Then we know that $\tilde{T}(s) = \sum_{n=0}^{\infty} s^{-2^n}$ has \mathbb{T} as its natural boundary. Obviously we can realize T using the construction of Theorems 1 or 2 with any $k > 1$.

Remark 2.3. The realization constructed in Theorems 1 and 2 uses the operator kU , where U is the unilateral shift on $l_2(\mathbb{Z}^+)$. The spectrum of kU is the closed disk of radius k , and hence its resolvent set is connected. The value of k we used is big enough so that the singularities of $\tilde{T}(\cdot)$ are included in the disk of radius k . In other words, we used an operator with spectrum big enough to include all the singularities of $\tilde{T}(\cdot)$. It is therefore of interest to know how small k can be taken for a given realizable weighting pattern $T(\cdot)$. It follows from a theorem in Widder [20, p. 95], that if σ_0 is the exponential order (or "type" in Widder's terminology) of the entire function $T(\cdot)$, then $\tilde{T}(\cdot)$ will be analytic for $|s| > \sigma_0$ and will vanish at infinity, and conversely. Hence k in Theorems 1 and 2 must satisfy $k \geq \sigma_0$.

The connectedness of the resolvent set of the infinitesimal generator A has important implications as far as the relationship to frequency response methods for system identification is concerned. The values of \tilde{T} for s purely imaginary are often empirically determined by letting $u(t) = \sin \omega t$ and looking at the periodic solution which results. If the periodic component of the response is $M(\omega) \sin(\omega t + \phi(\omega))$, then

$$\tilde{T}(i\omega) = M(\omega) e^{i\phi(\omega)}.$$

However, if the domain of analyticity of \tilde{T} is such that the entire imaginary axis does not belong to a single component, then there is no way that experimental data taken in different components can be pieced together and we must regard the system as consisting of several unrelated parts.

Remark 2.4. There are two typical sets of examples of systems with uniformly continuous state-transition operators, i.e., examples for which the operator A is bounded. The first comes from systems governed by parabolic and certain hyperbolic partial differential equations with constant coefficients, where the spatial domain is infinite or semi-infinite, after semidiscretization with uniform spatial mesh (see Birkhoff and Varga [24] and Brockett and Willems [17]).

The second comes from systems governed by certain particular classes of partial differential equations. The ideas involved are best illustrated by the following example. Consider the system

$$\frac{\partial}{\partial t} \frac{\partial}{\partial z} x(t, z) + \left(\frac{\partial}{\partial t} - \frac{\partial}{\partial z} \right) x(t, z) = b(z)u(t),$$

$$y(t) = x(t, 1),$$

where $x(t, \cdot) \in L_2(0, 1)$ is absolutely continuous and $x(t, 0) = 0$; $x(\cdot, z)$ and $(\partial/\partial z)x(\cdot, z)$ are differentiable; $b(\cdot) \in L_2(0, 1)$. The domain of $-\partial/\partial z$ is $\mathcal{D}_0(-\partial/\partial z) = \{h \in L_2(0, 1) \text{ such that } h \text{ is absolutely continuous and } h(0) = 0\}$. If we let $\xi(t, z) = (\partial/\partial z + I)x(t, z)$, we obtain the system

$$\frac{\partial}{\partial t} \xi(t, z) = \frac{\partial}{\partial z} \left(\frac{\partial}{\partial z} + I \right)^{-1} \xi(t, z) + b(z)u(t),$$

$$y(t) = \left(\frac{\partial}{\partial z} + I \right)^{-1} \xi(t, z)|_{z=1}.$$

We can study the first system via this second one and in the latter the operator $(\partial/\partial z)(\partial/\partial z + I)^{-1}$ is a bounded operator on $L_2(0, 1)$ (indeed is given by $\xi(t, z) \mapsto \xi(t, z) - \int_0^z e^{-(z-\sigma)} \xi(t, \sigma) d\sigma$).

Other examples can be found in problems of infinite queues and in systems governed by certain classes of integro-differential equations (compare with previous example).

3. Realizability criteria, general case. In this section we investigate the realizability problem when A is the infinitesimal generator of a C_0 semigroup of bounded operators on a Hilbert space \mathcal{H} . By the Hille–Yosida theorem [18] a necessary and sufficient condition that a closed linear operator A with domain $\mathcal{D}_0(A)$ dense be the infinitesimal generator of a C_0 semigroup, is that there exist positive real numbers M and β such that for every real $\lambda > \beta$, λ is in the resolvent set of A and

$$\|(I\lambda - A)^{-n}\| \leq \frac{M}{(\lambda - \beta)^n}, \quad n = 1, 2, \dots$$

If these conditions hold for all $\lambda > \beta$, then $(Is - A)^{-1}$ exists for all complex s with $\operatorname{Re} s > \beta$ and is given by $(Is - A)^{-1}x = \int_0^\infty e^{-st} e^{At} x dt$ for all $x \in \mathcal{H}$, $\|(Is - A)^{-n}\| \leq M/(\operatorname{Re} s - \beta)^n$ for $\operatorname{Re} s > \beta$, and $\|e^{At}\| \leq Me^{\beta t}$.

In case of a *regular realization*, $b \in \mathcal{H}$ and c is a bounded linear functional on \mathcal{H} . Then the observation procedure (i.e., $y(t) = c[x(t)]$) is somehow restricted since we cannot have point evaluations, or point evaluations of derivatives as $c(\cdot)$. Moreover, since b is just an element of \mathcal{H} we can regard in general the equation

$$(1) \quad \frac{d}{dt} x(t) = Ax(t) + bu(t)$$

only in the weak sense (i.e., $x(\cdot)$ satisfies the integral equation given by the variation of constants formula, but not the differential equation). On the other hand, in a regular realization the properties of b and c are symmetric, a fact which has some implications for the desired *duality* in systems theory.

In case of a *balanced realization*, $b \in \mathcal{D}_0(A)$ and c is a linear functional defined on $\mathcal{D}_0(A)$ and such that $|c(x)| \leq k(\|Ax\| + \|x\|)$ for all $x \in \mathcal{D}_0(A)$ and some constant k . Here we can regard equation (1) in the strong sense. Moreover, we can allow point evaluations, or point evaluations of derivatives as $c(\cdot)$; (for example, with A being $\partial^2/\partial z^2$ on $L_2[0, \infty)$ and $c(\cdot)$ being $[\partial/\partial z(\cdot)]_0$, or with A being $\partial/\partial z$ on $L_2[0, \infty)$ and c being evaluated at 0). However, in this case c and b do not have symmetric properties.

Remark 3.1. If c is a closed linear functional on \mathcal{H} with $\mathcal{D}(c) \supseteq \mathcal{D}_0(A)$, then c satisfies the conditions stated above in the case of a balanced realization. To see this we have that $\mathcal{D}_0(A)$ with the norm $\|x\|_1 = \|Ax\| + \|x\|$ becomes a Banach space since A is closed. Then the restriction of c to $\mathcal{D}_0(A)$ is a closed linear operator, defined everywhere, and hence by the closed graph theorem is bounded. Hence there exists a k such that

$$|c(x)| \leq k\|x\|_1 = k(\|Ax\| + \|x\|) \quad \text{for all } x \in \mathcal{D}_0(A).$$

The following theorem proves that in our setting (more specifically when the state space is a Hilbert space) the class of weighting patterns which admit balanced realizations is identical with the class of weighting patterns which admit regular realizations.

THEOREM 3. *A weighting pattern $T(\cdot)$ has a balanced realization if and only if it has a regular one. Moreover, the infinitesimal generators in the two cases can be taken to be the same.*

Proof. Suppose $T(\cdot)$ has a regular realization. Then there exist c_1, b_1 , elements of \mathcal{H} , and a linear operator A generating a C_0 semigroup e^{At} on \mathcal{H} such that

$$T(t) = \langle c_1, e^{At}b_1 \rangle.$$

By the Hille–Yosida theorem there exists a positive real number β such that for every real $\lambda > \beta$, λ is in the resolvent set of A . Choose such a $\lambda > 1$. Then $(\lambda I - A)^{-1}$ is an everywhere defined bounded operator, and it maps the whole \mathcal{H} onto $\mathcal{D}_0(A)$ since A is closed (see [21, p. 209]). Let

$$b = (\lambda I - A)^{-1}b_1.$$

Then $b \in \mathcal{D}_0(A)$ and $b_1 = (\lambda I - A)b$. Hence

$$T(t) = \langle c_1, e^{At}(\lambda I - A)b \rangle = \langle c_1, (\lambda I - A)e^{At}b \rangle.$$

Define the linear functional $c(\cdot)$ via

$$c(x) = \langle c_1, (\lambda I - A)x \rangle \quad \text{for } x \in \mathcal{D}_0(A).$$

Then

$$\begin{aligned} |c(x)| &\leq \|c_1\| \|\lambda x - Ax\| \leq \|c_1\|(\lambda\|x\| + \|Ax\|) \\ &\leq \lambda\|c_1\|(\|Ax\| + \|x\|). \end{aligned}$$

Therefore $T(t) = c[e^{At}b]$ and this is a balanced realization.

Conversely, assume that $T(\cdot)$ has a balanced realization. Then there exist A, b, c as in the definition of a balanced realization so that $T(t) = c[e^{At}b]$. Consider $\mathcal{D}_0(A)$ with the inner product $\langle x, y \rangle_A = \langle Ax, Ay \rangle + \langle x, y \rangle$ for $x, y \in \mathcal{D}_0(A)$. This inner product induces the norm $\|x\|_A = (\|Ax\|^2 + \|x\|^2)^{1/2}$. Since A is closed,

$\mathcal{D}_0(A)$ is complete under the norm $\|\cdot\|_A$ and hence it is a Hilbert space with the inner product $\langle \cdot, \cdot \rangle_A$. For $x \in \mathcal{D}_0(A)$ we have

$$|c(x)| \leq k(\|Ax\| + \|x\|) \leq 2k(\|Ax\|^2 + \|x\|^2)^{1/2} = 2k\|x\|_A.$$

Thus $c(\cdot)$ is a bounded linear functional on the Hilbert space $\mathcal{D}_0(A)$ (with the above inner product). Hence by the Riesz representation theorem there exists $d \in \mathcal{D}_0(A)$ such that

$$c(x) = \langle d, x \rangle_A \quad \text{for all } x \in \mathcal{D}_0(A).$$

Hence $T(t) = \langle Ad, Ae^{At}b \rangle + \langle d, e^{At}b \rangle$.

Since the space we are working with is a Hilbert space, A^* also generates a C_0 semigroup which is exactly $(e^{At})^*$. Hence if we choose a real $\lambda > \beta$ (β from the Hille–Yosida theorem), then both $(\lambda I - A)$ and $(\lambda I - A^*)^{-1}$ are everywhere defined bounded operators.

We then have

$$\begin{aligned} T(t) &= \langle Ad, (A - \lambda I)e^{At}b \rangle + \lambda \langle Ad, e^{At}b \rangle + \langle d, e^{At}b \rangle \\ &= \langle Ad, e^{At}(A - \lambda I)b \rangle + \langle \lambda Ad + d, e^{At}b \rangle. \end{aligned}$$

If we let $(A - \lambda I)b = b_1 \in \mathcal{H}$, then $b = (A - \lambda I)^{-1}b_1$. Therefore

$$\begin{aligned} T(t) &= \langle Ad, e^{At}b_1 \rangle + \langle \lambda Ad + d, e^{At}(A - \lambda I)^{-1}b_1 \rangle \\ &= \langle Ad + (A^* - \lambda I)^{-1}(\lambda Ad + d), e^{At}b_1 \rangle. \end{aligned}$$

Let $c_1 = Ad + (A^* - \lambda I)^{-1}(\lambda Ad + d)$. Then

$$T(t) = \langle c_1, e^{At}b_1 \rangle$$

and obviously A, b_1, c_1 is a regular realization for $T(\cdot)$.

The last statement in the theorem is obvious from the above construction.

This theorem motivates the following definition.

DEFINITION. A weighting pattern $T(\cdot)$ is *realizable* if and only if it has a balanced realization.

We now give a preliminary description of the realizable weighting patterns.

THEOREM 4. A necessary condition for $T(\cdot)$ to be realizable is that it be continuous and of exponential order. A sufficient condition is that it be locally absolutely continuous (i.e., absolutely continuous, on each bounded closed interval) and that $\dot{T}(t)$ (which then exists as an a.e. defined function) be of exponential order (i.e., $\text{ess sup } |\dot{T}(t)| \leq Ke^{\alpha t}$ for some positive K, α).

Proof. Necessity. Since $T(\cdot)$ has a balanced realization, and hence by Theorem 3 it has a regular one, $T(t) = \langle c, e^{At}b \rangle$. Since e^{At} is strongly continuous we get that $T(\cdot)$ is continuous. Since $\|e^{At}\| \leq Me^{\beta t}$, by the Hille–Yosida theorem we get that $T(\cdot)$ is of exponential order.

Sufficiency. Let $T(\cdot)$ be as in the hypothesis. Then for large enough σ , $e^{-\sigma t}\dot{T}(t) \in L_2(0, \infty)$. Hence the function $e^{-\sigma t}T(t)$ is in $L_2(0, \infty)$, it is locally absolutely continuous and its derivative belongs to $L_2[0, \infty)$. Take as b the function $e^{-\sigma t}T(t)$, and as Hilbert space \mathcal{H} the space $L_2(0, \infty)$. The differentiation operator $A = \partial/\partial z$ on $L_2(0, \infty)$ is a closed operator with domain dense, generates the semigroup of

left translations (restricted to $[0, \infty)$ of course) and its spectrum is the closed left half-plane (i.e., $\sigma(A) = \{s \in \mathbb{C} | \operatorname{Re} s \leq 0\}$) (see [18]). Its domain consists of elements of $L_2(0, \infty)$ which are locally absolutely continuous and whose derivatives also belong to $L_2(0, \infty)$. Consider as c the linear functional whose action on a function f is described by

$$c[f] = f(0) \quad (\text{i.e., evaluation at } 0).$$

Then c is defined on $\mathcal{D}_0(A)$. Moreover, for $x \in \mathcal{D}_0(A)$ we have

$$\begin{aligned} |c(x)|^2 &= |x(0)|^2 \leq \int_0^\infty 2|x(z)| |\dot{x}(z)| dz \\ &\leq \int_0^\infty |x(z)|^2 dz + \int_0^\infty |\dot{x}(z)|^2 dz. \end{aligned}$$

So $|c(x)| \leq (\|Ax\| + \|x\|)$. Hence b, c satisfy our requirements. Now $c[e^{At}b] = c[e^{-\sigma(t+z)}T(t+z)] = e^{-\sigma t}T(t)$ and therefore $T(t) = c[e^{(A+\sigma I)t}b]$. This is a balanced realization.

From the equation $T(t) = \langle c, e^{At}b \rangle$ we get via Laplace transform the equation

$$\tilde{T}(s) = \langle c, (Is - A)^{-1}b \rangle \quad \text{for } \operatorname{Re} s > \beta,$$

where the β comes from the Hille–Yosida theorem. On the other hand, since $T(\cdot)$ is realizable we know it is of exponential order, say σ_0 . Hence $\tilde{T}(\cdot)$ is analytic in $\operatorname{Re} s > \sigma_0$. Moreover, from Theorem 4 we have that $\sigma_0 \leq \beta$ and by the Hille–Yosida theorem the function $\langle c, (Is - A)^{-1}b \rangle$ is analytic in $\operatorname{Re} s > \beta$. Let $\rho_0(A)$ be the connected component of $\rho(A)$ which contains the half-plane $\operatorname{Re} s > \beta$. Then by analytic continuation we see that $\tilde{T}(\cdot)$ is analytic for all $s \in \rho_0(A)$. So again (as in the bounded case) we arrive at the conclusion, that for any realization A, b, c of $T(\cdot)$ we must have the *spectral inclusion property*

$$\sigma(\tilde{T}) \subseteq \sigma_0(A),$$

where $\sigma_0(A)$ is the complement of $\rho_0(A)$ in \mathbb{C} .

The corresponding (to Theorem 4) conditions in the complex domain are described below.

THEOREM 5. *A necessary condition for $T(\cdot)$ to be realizable is that its Laplace transform $\tilde{T}(s)$ belong to $H^2(\prod_\rho^+) \cap H^\infty(\prod_\rho^+)$ for some $\rho > 0$. A sufficient condition is that $\tilde{T}(s) \in H^2(\prod_\rho^+)$ and $(s\tilde{T}(s) - T(0)) \in H^2(\prod_\rho^+)$ for some $\rho > 0$.*

Proof. This is an immediate consequence of Theorem 4, the Paley–Wiener theorem [7] and the Hille–Yosida theorem.

Example. The delayed step whose transform is e^{-s}/s is not realizable, whereas the delayed ramp e^{-s}/s^2 is realizable.

Remark 5.1. Suppose $T(\cdot)$ is continuous and of exponential order. Let $b \in L_2(0, \infty)$ be the function $e^{-\sigma t}T(t)$, where σ is large enough. Let $c_\lambda \in L_2(0, \infty)$ be the function $c_\lambda(z) = (2/\pi)\lambda/(\lambda^2 + z^2)$ and A be the differentiation operator. Then by Theorem 9.9 in [19] we have

$$e^{-\sigma t}T(t) = \lim_{\lambda \rightarrow 0} \langle c_\lambda, e^{At}b \rangle = \lim_{\lambda \rightarrow 0} \int_0^\infty e^{-\sigma(t+z)}T(t+z)c_\lambda(z) dz.$$

Hence $T(t) = \lim_{\lambda \rightarrow 0} \langle c_\lambda, e^{(A+\sigma I)t} b \rangle$. Therefore $T(\cdot)$ is the pointwise limit of a one-parameter family of realizable functions.

In order to give some better sufficient conditions for realizability we need the following well-known result [22] from the theory of $H^p(\mathbb{T}^+)$ functions: If $f \in H^p(\mathbb{T}^+)$, $1 \leq p < \infty$, then it is represented by the proper Cauchy integral of its boundary values. That is, for $\operatorname{Re} s > 0$ we have the representation

$$f(s) = \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} \frac{f(i\omega)}{s - i\omega} d(i\omega).$$

THEOREM 6. *Let $T \in L_2(0, \infty)$ and be continuous. If $\tilde{T}(i\omega) = \overline{F_1(i\omega)} F_2(i\omega)$, where F_1, F_2 belong to $H^2(\mathbb{T})$, then T is realizable.*

Proof. Certainly $\tilde{T}(i\omega) \in L_1(\mathbb{T}; d\omega/2\pi)$. Hence since $T \in L_2(0, \infty)$ we have that

$$T(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \tilde{T}(i\omega) e^{i\omega t} d\omega \quad \text{a.e.}$$

But since both sides are continuous the equality holds everywhere. Thus

$$T(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \overline{F_1(i\omega)} e^{i\omega t} F_2(i\omega) d\omega.$$

But this equality says that if we take as Hilbert space $H^2(\mathbb{T})$, as b the function F_2 , as c the function F_1 and as A the operator induced on $H^2(\mathbb{T})$, by multiplication by $i\omega$ we have

$$T(t) = \langle c, e^{At} b \rangle$$

(where the inner product is that of $L^2(\mathbb{T}, d\omega/2\pi)$). Hence A, b, c is a regular realization for T , and by Theorem 3, T is realizable.

Let us note that since the Fourier transform is a unitary map between $L_2(0, \infty)$ and $H^2(\mathbb{T})$, and multiplication by $e^{i\omega t}$ on $H^2(\mathbb{T})$ corresponds to left translation on $L_2(0, \infty)$, we can give also a realization of T in $L_2(0, \infty)$ by the left translation semigroup. Indeed, if we let

$$f_1 = \mathcal{F}^{-1}(F_1), \quad f_2 = \mathcal{F}^{-1}(F_2)$$

and e^{At} = left translation semigroup restricted to $L_2(0, \infty)$, we have $T(t) = \langle f_1, e^{At} f_2 \rangle$. Moreover, we can give a realization in terms of the right translation semigroup on $L_2(0, \infty)$ since we also have

$$T(t) = \langle f_2, e^{A^*t} f_1 \rangle$$

with A, f_1, f_2 as above. (Note that $L_2(0, \infty)$ is invariant under right translations.)

Note. If T satisfies the conditions of Theorem 6, then by the Paley–Wiener theorem $\tilde{T} \in H^2(\mathbb{T}^+)$. Hence

$$\tilde{T}(s) = \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} \frac{\tilde{T}(i\omega) d(i\omega)}{s - i\omega} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\overline{F_1(i\omega)} F_2(i\omega) d\omega}{s - i\omega}$$

and we could have used this approach in the proof.

COROLLARY 6.1. *Suppose T is continuous and of exponential order. If for some α the function $T_1(t) = e^{-\alpha t} T(t)$ satisfies the conditions of Theorem 6, then T is realizable.*

Proof. Of course if α is bigger than the exponential order of T then $e^{-\alpha t}T(t)$ belongs to $L_2(0, \infty)$. So we really have to check for the factorization only. Now, if T_1 satisfies Theorem 6, then $T_1(t) = \langle c, e^{At}b \rangle$. So $T(t) = \langle c, e^{(A+\alpha I)t}b \rangle$.

Remark 6.2. We see from the spectral inclusion property and from the Hille–Yosida theorem that the singularities of $\tilde{T}(\cdot)$, for any realizable T , are in some left half-plane. In all our constructions we used operators with spectrum large enough (in fact with spectrum some left half-plane) to include the singularities of a large class of realizable functions.

Remark 6.3. The conditions of Corollary 6.1 are weaker than those of Theorem 4. To see this observe first of all that continuity is required in both. Theorem 4 implies that for large enough σ the function $e^{-\sigma t}T(t) = T_1(t)$ belongs to $L_2(0, \infty)$, is locally absolutely continuous and its derivative belongs to $L_2(0, \infty)$. Hence $\tilde{T}_1(i\omega)$ and $i\omega\tilde{T}_1(i\omega)$ belong to $H^2(\mathbb{I})$ by the Paley–Wiener theorem. But $(1 - i\omega)\tilde{T}_1(i\omega) = g(i\omega)$ also belongs to $H^2(\mathbb{I})$. Hence

$$\tilde{T}_1(i\omega) = \frac{1}{1 - i\omega} g(i\omega) = \overline{\frac{1}{1 + i\omega}} g(i\omega)$$

and since $1/(1 + i\omega)$ belongs to $H^2(\mathbb{I})$ we see that T satisfies the conditions of Corollary 6.1.

We give another sufficient condition for realizability.

THEOREM 7. *If $T \in L_2(0, \infty)$ is continuous and $\tilde{T} \in H^1(\prod^+)$, then T is realizable.*

Proof. Since T is continuous, belongs to $L_2(0, \infty)$ and $\tilde{T}(i\omega) \in L^1(\mathbb{I}; d\omega/2\pi)$ we have that

$$T(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \tilde{T}(i\omega) e^{i\omega t} d\omega,$$

the equality holding everywhere. We know that $f \in H^1(\prod^+)$ if and only if $f = f_1 f_2$, where $f_1, f_2 \in H^2(\prod^+)$ (see [7, p. 134]). Hence there exist $F_3, F_2 \in H^2(\prod^+)$ such that $\tilde{T}(i\omega) = F_3(i\omega)F_2(i\omega) = \overline{F_1(i\omega)}F_2(i\omega)$, where $F_1 = \overline{F_3} \in H^2(\prod^-)$. Hence

$$T(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \overline{F_1(i\omega)} e^{i\omega t} F_2(i\omega) d\omega.$$

Thus by taking as Hilbert space \mathcal{H} the space $L_2(\mathbb{I}; d\omega/2\pi)$, as A multiplication by $i\omega$; as c the function F_1 and as b the function F_2 , we get

$$T(t) = \langle c, e^{At}b \rangle$$

and T is realizable.

Again using Fourier transforms we can give in the above case a realization of T on $L_2(-\infty, \infty)$ using the left translation semigroup or the right translation semigroup.

COROLLARY 7.1. *Let T be continuous and of exponential order. If for some α , $\tilde{T}(\cdot) \in H^1(\prod_{\alpha}^+)$, then T is realizable.*

Proof. By taking σ big enough, then, we can make the function $e^{-\sigma t}T(t)$ satisfy the conditions of Theorem 7.

Remark 7.2. The realization constructed in Theorem 7 has as infinitesimal generator the differentiation operator on $L_2(-\infty, \infty)$, whose spectrum is just the imaginary axis. So in this model we do not have a connected resolvent set.

On the other hand, in the realization of Theorem 6 we do have a connected resolvent set, but instead the spectrum becomes very large.

This last theorem indicates some other classes of realizable functions. We need first a standard definition for fractional derivatives in the L_2 -sense, or equivalently for the Sobolev spaces of fractional order.

DEFINITION. Let $0 \leq \gamma \leq 1$. Then $f \in L_2(0, \infty)$ has an L_2 -derivative of fractional order γ if and only if there exists $g \in L_2(0, \infty)$ such that $s^\gamma \tilde{f}(s) = \tilde{g}(s)$, where we always choose the branch of s^γ so that $\operatorname{Re} s^\gamma > 0$ for $\operatorname{Re} s > 0$. The space of all those f is usually denoted by H_γ^2 .

COROLLARY 7.3. If T is continuous and belongs to H_γ^2 for $\frac{1}{2} < \gamma \leq 1$, then T is realizable.

Proof. We have that $\tilde{T}(s)$ and $s^\gamma \tilde{T}(s) = g(s) \in H^2(\mathbb{C}^+)$. Hence $\tilde{T}(s) = (1/s^\gamma)g(s)$. Since for $\frac{1}{2} < \gamma \leq 1$ and for all $\alpha > 0$ we have trivially that $1/(s + \alpha)^\gamma \in H^2(\mathbb{C}^+)$, we get finally that for all $\alpha > 0$, $\tilde{T} \in H^1(\mathbb{C}^+)$ and the result follows from Corollary 7.1.

Finally we have the obvious generalization of Corollary 7.3.

COROLLARY 7.4. If T is continuous and for some $\alpha > 0$, $e^{-\alpha t}T(t) \in H_\gamma^2$ with $\frac{1}{2} < \gamma \leq 1$, then T is realizable.

4. Canonical realizations. In the rest of this paper we shall restrict our study to weighting patterns T with bounded realizations. Moreover, we assume that T is realizable by the unilateral shift itself (i.e., we can take $k = 1$ in the construction of Theorems 1 or 2), or equivalently that $(1/s)\tilde{T}(1/s) \in H^2(\mathbb{D})$. This does not harm the generality of the discussion, since we can reduce the general case by a simple change of variable, to the above case. Indeed, if we define

$$\tilde{T}_k(s) = \sum_{i=0}^{\infty} \frac{a_i}{k^i} s^{-(i+1)} = k \tilde{T}(ks) = \mathcal{L} \left[T \left(\frac{t}{k} \right) \right],$$

where \mathcal{L} denotes Laplace transform, then \tilde{T}_k satisfies the above for some finite $k > 0$.

It is obvious that if a weighting pattern has one realization it has many. An element ϕ of a separable Hilbert space \mathcal{H} is called a *cyclic vector* for a bounded operator A if and only if the linear span of $\phi, A\phi, A^2\phi, A^3\phi, \dots$ is dense in \mathcal{H} . One calls a realization (A, b, c) *canonical* whenever b is a cyclic vector for A , and c is a cyclic vector for A^* . However, we avoid the term “minimal” because many of the implications of this term are absent in the present setting. (Some authors prefer to call such a realization “controllable and observable” or “ ε -controllable and ε -observable” or “weakly controllable and weakly observable”.) If a weighting pattern T has a finite-dimensional realization, it has one with minimal dimension of the state space, which is called minimal. As is well known (see [2, pp. 105–115]) a finite-dimensional system is minimal if and only if it is controllable and observable (canonical). Moreover, any two minimal realizations differ by a change of basis in the state space, and the spectral properties of A in any minimal realization are *uniquely* determined by the weighting pattern. Here the situation is much more complicated. It happens that a canonical realization is much more loosely specified by the weighting pattern.

We start with a construction of a canonical realization starting from a given one.

THEOREM 8. Let $T(t) = \langle c, e^{At}b \rangle$. Let M be the closed linear span of $c, A^*c, A^{*2}c, \dots$ in \mathcal{H} (a separable Hilbert space). Let P_M be the orthogonal projection on M . Then (i) $T(t) = \langle c, e^{P_M A P_M t} P_M b \rangle$.

Now let N be the closed linear span of $P_M b, \dots, (P_M A P_M)^i P_M b, \dots$ in M and let P_N be the orthogonal projection on N . Then (ii) $T(t) = \langle P_N c, e^{(P_N A P_N)t} P_M b \rangle$. Moreover, N is the closed linear span of $P_M b, \dots, (P_N A P_N)^i P_M b, \dots$ and the closed linear span of $P_N c, \dots, (P_N A P_N)^* P_N c$.

Proof. It is obvious that M is the smallest closed subspace of \mathcal{H} which contains c and is invariant under A^* . Hence M^\perp is invariant under A . Hence $A(I - P_M)x \in M^\perp$ for all $x \in \mathcal{H}$. So $P_M A(I - P_M)x = 0$ for all $x \in \mathcal{H}$. Hence

$$(2) \quad P_M A = P_M A P_M.$$

Using (2) we get $(P_M A P_M)^i P_M b = P_M A^i b$.

Hence $\langle c, e^{P_M A P_M t} P_M b \rangle = \langle c, P_M e^{At} b \rangle = \langle c, e^{At} b \rangle = T(t)$ and this proves (i). Similarly, N is the smallest closed subspace of M which contains $P_M b$ and is invariant under $P_M A P_M$. Then for every $x \in \mathcal{H}$, $(I - P_N)P_M A P_M P_N x = 0$. Thus

$$(3) \quad P_N A P_N = P_N P_M A P_M P_N = P_M A P_M P_N = P_M A P_N.$$

Using (2) and (3) we obtain

$$(4) \quad \begin{aligned} (P_N A P_N)^i P_M b &= P_M A P_N (P_N A P_N)^{i-1} P_M b = P_M A P_N P_N A P_N (P_N A P_N)^{i-2} P_M b \\ &= P_M A P_M A P_N (P_N A P_N)^{i-2} P_M b = P_M A^2 P_N (P_N A P_N)^{i-2} P_M b \\ &= P_M A^i P_N P_M b = P_M A^i P_M b = (P_M A P_M)^i P_M b. \end{aligned}$$

Thus

$$\begin{aligned} \langle P_N c, e^{(P_N A P_N)t} P_M b \rangle &= \langle c, P_N e^{(P_N A P_N)t} P_M b \rangle = \langle c, e^{(P_M A P_M)t} P_M b \rangle \\ &= \langle c, P_M e^{At} b \rangle = \langle c, e^{At} b \rangle = T(t), \end{aligned}$$

and this proves (ii).

From (3) we get

$$(5) \quad (P_N A P_N)^* P_N c = P_N A^* c.$$

The first assertion in the last statement is proved by (4), i.e., by the fact that $(P_N A P_N)^i P_M b = (P_M A P_M)^i P_M b$ for all i . The second is an easy consequence of (5) and of the cyclicity of c for A^* .

Here, as we assumed in the beginning of this section, if $T(\cdot)$ is realizable, it can be realized by the shift (unilateral or bilateral). Some important questions which arise naturally are the following. It is obvious that the realization given by Theorems 1 and 2 is controllable. Also we know that the spectrum of the unilateral shift is the closed unit disk. Given a weighting pattern T , how simple can the spectrum of the infinitesimal generator A of a realization be? How small can the spectrum be? If we take a canonical realization (A, b, c) , is the spectrum of A uniquely determined by T ? How are all canonical realizations of a given T related to each other? When can we make the resolvent set of the infinitesimal generator A connected?

An immediate observation, which gives, however, some indication of the interplay of the notions described in these questions is the following: We can realize any such T by the bilateral shift. Such a realization is obviously non-

canonical. On the other hand, since the spectrum of the bilateral shift is just \mathbb{T} , the spectrum can be considered as “simple”. However, the resolvent set is not connected.

Given a weighting pattern T we have the *shift realization* as described in Theorems 1 and 2:

$$\frac{d}{dx} x(t) = Ux(t) + bu(t),$$

$$y(t) = \langle c, x(t) \rangle,$$

where $x(t) \in l_2(\mathbb{Z}^+)$ for all t , $b = \{1, 0, 0, \dots\}$, U is the unilateral shift and $c = \{T(0), T^{(1)}(0), T^{(2)}(0), \dots\}$. Here b is obviously a cyclic vector for U . It is immediately seen as a consequence of Theorem 8, that if we let M be the closed linear span of $c, U^*c, \dots, U^{*i}c, \dots$ and P_M the projection on M , then $(P_M U P_M, P_M b, c)$ is a canonical realization of T , with state space M . We can write the “shift realization” in terms of H^2 functions as follows:

$$\frac{d}{dt} x(t, s) = sx(t, s) + u(t),$$

$$y(t) = \int_{\mathbb{T}} s \tilde{T}(s) x(t, s) d\mu(s),$$

where $x(t, \cdot) \in H^2(H^2(\mathbb{D}) \text{ or } H^2(\mathbb{T}))$. (Compare with [17] where similar equations are used.) Under the isomorphism between $l_2(\mathbb{Z}^+)$ and $H^2(\mathbb{D})$, c corresponds to $(1/s)\tilde{T}(1/s)$ which equals $s\tilde{T}(s)$ on \mathbb{T} (since $\tilde{T}(\cdot)$ has real Taylor coefficients). U corresponds to multiplication by s , U^* corresponds to the mapping:

$$f(s) \mapsto \frac{f(s) - f(0)}{s} \quad \text{on } H^2(\mathbb{D}).$$

We need a few well-known facts from the theory of H^2 -functions and Toeplitz operators. The reader is referred to [7], [8] and [10] for further details. A function $f \in H^2(\mathbb{D})$ is called *inner* if $|f(e^{i\theta})| = 1$ a.e. A function $f \in H^2(\mathbb{D})$ is called *outer* if it is a cyclic vector for the shift in $H^2(\mathbb{D})$ (i.e., the linear span of the functions f, sf, s^2f, \dots is dense in $H^2(\mathbb{D})$). A *Blaschke product* is a function of the form

$$B(s) = s^k \prod_{j=1}^{\infty} \frac{a_j - s}{1 - \overline{a_j}s} \frac{\overline{a_j}}{|a_j|},$$

where k is a nonnegative integer and the a_j are complex numbers (not necessarily distinct) such that $0 < |a_j| < 1$, $\sum_{j=1}^{\infty} (1 - |a_j|) < \infty$. A *singular function* is a function of the form

$$S(s) = \exp \left(- \int \frac{e^{i\theta} + s}{e^{i\theta} - s} d\mu(\theta) \right),$$

where μ is any positive finite measure on $[0, 2\pi]$ which is singular with respect to the normalized Lebesgue measure. Every $f \in H^2(\mathbb{D})$ has a factorization $f = \phi \cdot h$, where ϕ is *inner* and h is *outer*. The factors are unique up to constant factors of modulus one. Any inner function has a factorization $\phi = cBS$, where c is a constant

of modulus one, B is a Blachke product and S is a singular function. An *inner* function is *normalized* if we choose $c = 1$, or equivalently if we require the first nonzero Taylor coefficient to be real and positive. Beurling showed that to every closed subspace M of $H^2(\mathbb{D})$ which is invariant under the shift (i.e., under multiplication by s) there corresponds a unique normalized inner function ϕ such that $M = \phi H^2(\mathbb{D})$ and conversely. We also have the corresponding facts for $H^2(\mathbb{T})$.

A Laurent operator on $l_2(\mathbb{Z})$ has a matrix representative which is constant on diagonals (i.e., $\alpha_{ij} = a_{i-j}$) and corresponds to multiplication by $\phi(s) = \sum_{i=-\infty}^{\infty} a_i s^i$ on $L^2(\mathbb{T})$ (where $a_i = \alpha_{i+k,k}$). A Toeplitz operator A on $l_2(\mathbb{Z}^+)$ has a similar matrix representative (which is infinite in only one direction). If $P: L^2(\mathbb{T}) \rightarrow H^2(\mathbb{T})$ is the associated projection, then for all $f \in H^2(\mathbb{T})$ we have

$$Af = P(\phi \cdot f).$$

The only way the “shift realization” can be canonical is if c is a cyclic vector for U^* (i.e., for the backward shift) or equivalently if $(1/s)\tilde{T}(1/s)$ is a cyclic vector for the backward shift on $H^2(\mathbb{D})$. (See also Fuhrmann [13, Thm. 2.6].) In [5] the authors studied cyclic vectors of the backward shift very extensively. We are going to use some of their results and we refer to [5] for further details. There exist many cyclic vectors for the backward shift on $H^2(\mathbb{D})$, as well as noncyclic ones. The rational functions are noncyclic. The authors give several ways of constructing cyclic vectors. Any H^2 -function with isolated branch points on \mathbb{T} is a cyclic vector and any function with lacunary Taylor series and square summable Taylor coefficients is also a cyclic vector. Since $f(s) \in H^2(\mathbb{D})$ is a cyclic vector for the backward shift if and only if $sf(s)$ is one, we have two cases to consider: namely, the case when $\tilde{T}(1/s)$ is a cyclic vector for the backward shift and the case when $\tilde{T}(1/s)$ is noncyclic.

We would like to close this section with some important remarks about the cyclic and noncyclic case. Let Q be the subset of the realizable transfer functions, for elements of which the set of real numbers k , such that $\tilde{T}(k/s)$ is in $H^2(\mathbb{D})$, is open. Let G be the complement of Q in the set of realizable functions. Theorem 2.2.4 in [5] reads as follows: *If f is holomorphic in $|s| < R$ for some $R > 1$, then f is either cyclic or a rational function.* Since $k > \sigma_0$ for elements of Q , an immediate consequence of the above theorem is that the elements of Q are either cyclic or rational functions. Then in G we have either cyclic or noncyclic but not rational functions, as illustrated in Fig. 1.

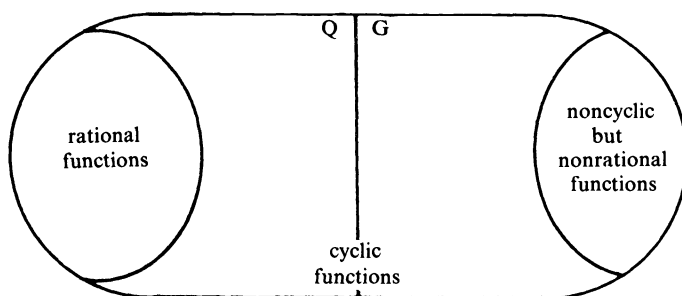


FIG. 1

Also from [5] we have that the set of cyclic vectors is dense in $H^2(\mathbb{D})$ as is the set of noncyclic vectors. However, the set of noncyclic vectors is a set of the first category, whereas the set of cyclic vectors is not. Hence the noncyclic vectors are somehow much more rare than the cyclic ones. Moreover, an element of $H^2(\mathbb{D})$ is noncyclic if and only if there exists a sequence of rational functions (satisfying special conditions [5, Thm. 4.1.1]) which converges to it in the $L^2(\mathbb{T})$ -norm, a fact which indicates that the noncyclic case is very much like the rational functions whereas the cyclic situation is new, harder, and potentially more interesting.

5. The noncyclic case. Now consider the case where $\tilde{T}(1/s)$ is not a cyclic vector for the backward shift. This case is treated by Fuhrmann [13] in detail; however, there are some additional facts given here about the spectrum of A . (H^2 stands for $H^2(\mathbb{D})$ or $H^2(\mathbb{T})$.)

To proceed we need the following theorem from [5, p. 56].

THEOREM 9 ([5]). *$f \in H^2(\mathbb{D})$ is noncyclic if and only if there exist $g \in H^2(\mathbb{D})$ and an inner function ϕ such that $f(e^{i\theta}) = e^{-i\theta}g(e^{i\theta})\phi(e^{i\theta})$ a.e. on \mathbb{T} . Moreover, if we require that ϕ be normalized and relatively prime to the inner factor of g , then ϕ and g are uniquely determined. In this case the closed subspace generated by $U^{*n}f$, $n = 0, \dots, \infty$, is precisely $(\phi H^2(\mathbb{D}))^\perp$.*

The normalized inner function ϕ thus uniquely associated with each noncyclic (for the backward shift) vector f is called the *associated inner function* of f .

We see immediately that the subspace M of $l^2(\mathbb{Z}^+)$ which is the state space for the canonical realization $(P_M U P_M, P_M b, c)$ derived from the “shift realization” corresponds to the closed subspace of H^2 generated by $\{U^{*n}(1/s)\tilde{T}(1/s)\}_{n=0}^\infty$ which we also call M . Hence applying Theorem 9 we get that

$$M = (\phi H^2)^\perp,$$

where $\phi(e^{i\theta})\overline{g(e^{i\theta})} = \tilde{T}(e^{-i\theta}) = \overline{\tilde{T}(e^{i\theta})}$ a.e. on \mathbb{T} (since $\tilde{T}(e^{i\theta})$ has real Fourier coefficients), and ϕ and g are uniquely determined by Theorem 9.

We need another theorem now from [6].

THEOREM 10 [6]. *Let $K = \phi H^2$, i.e., K is a closed subspace of H^2 invariant under the shift U . Let $M = (\phi H^2)^\perp$. Then the spectrum of U restricted on M is the set s_ϕ which consists of*

- (i) *all the points in \mathbb{C} with $|\lambda| < 1$, where $\phi(\lambda) = 0$,*
- (ii) *all the points in \mathbb{C} with $|\lambda| = 1$, where $\phi(\cdot)$ is not continuable analytically across the boundary \mathbb{T} of \mathbb{D} at λ .*

Using Theorems 9 and 10 we see that the spectrum of the infinitesimal generator of the canonical realization $(P_M U P_M, P_M b, c)$ is uniquely determined by T . Namely, the spectrum consists of the zeros of ϕ in \mathbb{D} (which coincide with the zeros of the Blaschke product part of ϕ) and the points of \mathbb{T} through which ϕ is not continuable analytically outside the unit circle (which coincide with the union of the support of the measure of \mathbb{T} which is associated with the singular part of ϕ and the set of points of \mathbb{T} which are accumulation points of the sequence of zeros of ϕ (see [7, pp. 68–69])).

Recall now that $\tilde{T}(e^{i\theta}) = \overline{\phi(e^{i\theta})}g(e^{i\theta})$ a.e. on \mathbb{T} , where $g \in H^2(\mathbb{T})$. Since ϕ is inner we get $\tilde{T}(e^{i\theta}) = g(e^{i\theta})/\phi(e^{i\theta})$ a.e. on \mathbb{T} . When \tilde{T} has a meromorphic continuation in \mathbb{D} we have $\tilde{T}(s) = g(s)/\phi(s)$ in \mathbb{D} . Since g, ϕ are analytic on \mathbb{D} and ϕ is

relatively prime to the inner factor of g , it follows that the singularities of $\tilde{T}(\cdot)$ in \mathbb{D} are exactly the zeros of $\phi(\cdot)$ in \mathbb{D} (with the same multiplicity as well). On the other hand, since $e^{-i\theta}\tilde{T}(e^{-i\theta}) \in (\phi H^2(\mathbb{T}))^\perp$ and $e^{-i\theta}\tilde{T}(e^{-i\theta})$ is a *noncyclic vector* for the backward shift by assumption, we know [5, pp. 58–59, Cors. 3.1.8 and 3.1.10] that the set of points of \mathbb{T} , through which ϕ is analytically continuable, coincides with the set of points of \mathbb{T} , through which $(1/s)\tilde{T}(1/s)$ is analytically continuable. Hence the set of points of \mathbb{T} , through which ϕ is analytically continuable, coincides with the set of points of \mathbb{T} , through which $\tilde{T}(1/s)$ is analytically continuable, which is the same as the set of points of \mathbb{T} , through which \tilde{T} is analytically continuable (in the reverse direction). So in this case we arrived at the conclusion that the spectrum of $P_M U P_M$ consists of the set of points of \mathbb{D} which are singularities of \tilde{T} and of points of \mathbb{T} through which \tilde{T} cannot be continued analytically. Obviously the last set is what we have defined as $\sigma(\tilde{T})$. Hence we obtain

$$\sigma(\tilde{T}) = \sigma(P_M U P_M).$$

We have thus proved the following theorem.

THEOREM 11. *Let T be given weighting pattern with $(1/s)\tilde{T}(1/s) \in H^2(\mathbb{D})$, such that $\tilde{T}(1/s)$ is not a cyclic vector for the backward shift on $H^2(\mathbb{D})$. Then there exists a canonical realization of T with the spectrum of the infinitesimal generator of the realization being exactly s_ϕ , where ϕ is the associated inner factor of $(1/s)\tilde{T}(1/s)$. If $M = (\phi H^2)^\perp$, this realization is constructed by taking as c the function $(1/s)\tilde{T}(1/s)$, as b the projection of 1 on M and as A the restriction of the forward shift on M . Moreover, if T has a meromorphic continuation in \mathbb{D} , this spectrum is just $\sigma(\tilde{T})$.*

We see that in the above case the “spectral inclusion property” becomes in fact an equality, i.e., the spectrum of the infinitesimal generator of the realization described in Theorem 11 is *minimal*. This motivates the following definition.

DEFINITION. A canonical realization (A, b, c) of a weighting pattern T is called *S-minimal* (S from spectrum) if $\sigma(A) = \sigma(\tilde{T})$ (multiplicities counted whenever possible).

These considerations lead us to a trivial corollary of Theorem 11.

COROLLARY 11.1. *Any T which has the “shift realization” and is such that $\tilde{T}(1/s)$ is not a cyclic vector of the backward shift on H^2 , and where \tilde{T} has a meromorphic continuation in \mathbb{D} , has an S-minimal realization, with A having a connected resolvent set.*

We do not have a complete picture for the relation between canonical (resp. S-minimal) realizations of the same weighting pattern T , in this case. However, a partial analysis indicates that the noncyclic case is very similar to the rational case.

6. The cyclic case. The cyclic case is very interesting since it reflects a number of physically interesting phenomena; for example, transfer functions with branch points and branch cuts. Transfer functions like these arise in systems governed by partial differential equations. Hence an understanding of the cyclic case should undoubtedly shed some light towards the realization problem for distributed systems.

This case is more difficult, since the associated inner factor of \tilde{T} which proves so crucial in the noncyclic case is now trivial. That is, the shift realization for cyclic

transfer functions is already canonical. However, the spectrum of this realization is far from being equal to $\sigma(\tilde{T})$, unless we have a pathological transfer function with branch points on a dense subset of \mathbb{T} . Hence canonical by no means implies S -minimal. (Again compare with Fuhrmann [13, Cor. 2.7] who observes the nonuniqueness of the spectrum.)

It is apparent from the spectral inclusion property that all the points on the branch cuts (if the transfer function has branch points) are included in the spectrum of any infinitesimal generator A with connected resolvent set which realizes the transfer function. However, branch cuts are not uniquely defined (e.g., for $1/\sqrt{s^2 + 1}$ any curve between i and $-i$ can be a branch cut provided it has no self intersection). Hence the set $\sigma(\tilde{T})$ is not uniquely determined and consequently there is not a unique "minimal spectrum" for the infinitesimal generators of the realizations. A reasonable expectation is that the spectrum of an S -minimal realization (provided one exists) will be unique if there are no branch points and otherwise will be unique modulo the branch cuts.

We conclude this section with an example of a realization for the Bessel function of zeroth order $\mathcal{T}_0(\cdot)$ which is S -minimal. It is easy to verify that

$$\mathcal{T}_0(t) = \sum_{m=0}^{\infty} \frac{(-1)^m (t/2)^{2m}}{(m!)^2}$$

satisfies our realizability criteria. $\tilde{\mathcal{T}}_0(s) = 1/\sqrt{s^2 + 1}$ has branch points on $\pm i$, hence $\tilde{\mathcal{T}}_0(\cdot)$ is a cyclic vector. We must take the branch cut in the finite plane. We are after a canonical realization whose infinitesimal generator has spectrum exactly the line between i and $-i$. Recalling that $\exp(\frac{1}{2}(s - 1/s)t)$ is a generating function for the Bessel functions of integral order, i.e., that

$$\exp\left(\frac{1}{2}\left(s - \frac{1}{s}\right)t\right) = \sum_{n=-\infty}^{\infty} \mathcal{T}_n(t)s^n,$$

and using Laurent operators we see that for

$$A = \frac{1}{2} \begin{bmatrix} \ddots & & & & & \\ & -1 & & & & 0 \\ & \ddots & 0 & -1 & & \\ & \ddots & 1 & 0 & -1 & \ddots \\ & & & 1 & 0 & \ddots \\ 0 & & & & 1 & \ddots \end{bmatrix},$$

$$e^{At} = \begin{bmatrix} \vdots & \vdots & \vdots & \vdots & & \\ \cdot & \mathcal{T}_1(t) & \mathcal{T}_0(t) & \mathcal{T}_{-1}(t) & \mathcal{T}_{-2}(t) & \cdot & \cdot \\ \cdot & \cdot & \mathcal{T}_1(t) & \mathcal{T}_0(t) & \mathcal{T}_{-1}(t) & \cdot & \cdot \\ \cdot & \cdot & \mathcal{T}_2(t) & \mathcal{T}_1(t) & \mathcal{T}_0(t) & \mathcal{T}_{-1}(t) & \cdot \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \end{bmatrix}.$$

Hence the above A along with $b = c = [\cdots 0 \ 0 \ 1 \ 0 \ 0 \cdots]$ gives a realization for $\mathcal{T}_0(\cdot)$ in $l_2(\mathbb{Z})$. That the spectrum of A is exactly $[i, -i]$ is a well-known fact from [10]. However, this realization is *not* canonical (it is easy to verify that the vector $[\cdots 0 \ 0 \ 1 \ 0 \ 1 \ 0 \ 0 \cdots]$ is orthogonal to $A^i b$ for all i). We are going to use Theorem 8 to reduce the above realization to a canonical one. So let M be the closed linear span of $c, A^*c, A^{*2}c, \dots$ in $l_2(\mathbb{Z})$. Then M is A^* invariant. But since $A^* = -A$ it is also A invariant, i.e., M reduces A . Since $b = c$, we get by Theorem 8 that $A_1 = A$ restricted to M, b, c is a realization of \mathcal{T}_0 which is obviously canonical. Let $\lambda \in \rho(A)$. Then $\lambda I - A$ has a bounded inverse. But since M reduces $A, \lambda I - A_1$ has also a bounded inverse. Hence

$$\rho(A) \subseteq \rho(A_1).$$

Therefore $\rho(A_1)$ is connected. Using the spectral inclusion property, the fact that $\sigma(A) = \sigma(\tilde{\mathcal{T}})$ and the above relation, we have

$$\sigma(\tilde{\mathcal{T}}) \subseteq \sigma(A_1) \subseteq \sigma(A) = \sigma(\tilde{\mathcal{T}}).$$

Thus

$$\sigma(A_1) = \sigma(\tilde{\mathcal{T}})$$

and A_1, b, c is an S -minimal realization.

This example shows that S -minimal realizations can exist for cyclic functions as well. It also shows that there exists no Hilbert space analogue of the finite-dimensional state space isomorphism theorem between two canonical realizations of the same T unless further assumptions are made. (See Helton [14].)

Notice also that nearly the same realization will work for the Bessel function \mathcal{T}_n , where

$$\tilde{\mathcal{T}}_n(s) = \frac{1}{\sqrt{s^2 + 1}} \left(\frac{1}{s + \sqrt{s^2 + 1}} \right)^n$$

provided we keep A, b as above and take $c = \{\cdots 0 \ 0 \ 1 \ 0 \ 0 \cdots\}$, with the 1 in the n th place.

REFERENCES

- [1] N. DUNFORD AND J. SCHWARTZ, *Linear Operators*, John Wiley, New York, 1958.
- [2] R. W. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.
- [3] A. V. BALAKRISHNAN, *Introduction to Optimization Theory in a Hilbert Space*, Springer-Verlag, New York, 1970.
- [4] H. O. FATTORINI, *On complete controllability of linear systems*, J. Differential Equations, 3 (1967), pp. 391–402.
- [5] R. G. DOUGLAS, H. S. SHAPIRO AND A. L. SHIELDS, *Cyclic vectors and invariant subspaces for the backward shift operator*, Ann. Inst. Fourier (Grenoble), 20 (1971), no. 1, pp. 37–76.
- [6] J. W. MOELLER, *On the spectra of some translation invariant spaces*, J. Math. Anal. Appl., 4 (1962), pp. 276–296.
- [7] K. HOFFMAN, *Branch Spaces of Analytic Functions*, Prentice-Hall, Englewood Cliffs, N.J., 1962.
- [8] H. HELSON, *Lectures on Invariant Subspaces*, Academic Press, New York, 1964.
- [9] GIAN-CARLO ROTA, *On models for linear operators*, Comm. Pure Appl. Math., 13 (1960), pp. 468–472.
- [10] P. HALMOS, *A Hilbert Space Problem Book*, Van Nostrand, Princeton, 1967.

- [11] BELA SZ-NAGY AND C. FOIAS, *Harmonic Analysis of Operators on Hilbert Space*, North Holland and American Elsevier, Amsterdam and New York, 1970.
- [12] ———, *Vecteurs cycliques et quasi-affinités*, *Studia Math.*, 31 (1968), pp. 35–42.
- [13] P. FUHRMANN, *On realization of linear systems and applications to some questions of stability*, *Math. Systems Theory*, to appear.
- [14] J. W. HELTON, *Discrete time systems, operator models, and scattering theory*, to appear.
- [15] B. SZ-NAGY AND C. FOIAS, *Opérateurs sans multiplicité*, *Acta. Sci. Math.*, 30 (1960), pp. 1–18.
- [16] ———, *Modèle de Jordan pour une classe d'opérateurs de l'espace de Hilbert*, *Ibid.*, 31 (1970), pp. 93–117.
- [17] R. W. BROCKETT AND J. L. WILLEMS, *Least squares optimization for stationary linear partial difference equations*, *Proc. IFAC Symp. on the Control of Distributed Parameter Systems*, Banff, Canada, 1971.
- [18] A. FRIEDMAN, *Partial Differential Equations*, Holt, Rinehart and Wilson, New York, 1969.
- [19] W. RUDIN, *Real and Complex Analysis*, McGraw-Hill, New York, 1966.
- [20] D. WIDDER, *The Laplace Transform*, Princeton Univ. Press, Princeton, N.J., 1946.
- [21] K. YOSIDA, *Functional Analysis*, Springer-Verlag, New York, 1965.
- [22] E. HILLE AND J. D. TAMARKIN, *On the absolute integrability of Fourier transforms*, *Fund. Math.*, 25 (1935).
- [23] A. V. BALAKRISHNAN, *System theory and stochastic control*, NATO Advanced Study Institute on Network and Signal Theory, England, 1972.
- [24] G. BIRKHOFF AND R. S. VARGA, *Discretization errors for well-set Cauchy problems, I*, *J. Math. and Phys.*, 44 (1965).
- [25] BENJAMIN NOBLE, *Methods Based on the Wiener-Hopf Technique for the Solution of Partial Differential Equations*, Pergamon Press, New York, 1958.

STATISTICAL MECHANICS AND SYSTEMS OF LARGE NUMBERS OF ELEMENTS WITH RANDOM INTERACTION*

YULIA SCHMOOKLER†

Abstract. In this work, the evident analogy between problems of statistical mechanics and large systems with chance interaction is developed into a one-to-one correspondence. As an element of a system, a stochastic automaton described by an ergodic Markov chain is taken, so that we can receive “gas of automata”, “solid state of automata”, etc. The formula linking the physical notions of energy and temperature with the matrix of transient probabilities of Markov chain is given. The limits of implication of this formula are determined by the existence of a “detailed balance” in the ergodic finite Markov chain. Detailed balance by itself gives the opportunity to solve some complicated problems in Markov chain theory which have been unsolved for rather a long time.

The mathematical description of the behavior of systems consisting of a large number of elements is extremely complicated. Every step and achievement in this direction gives new approaches to many problems of biology, economics, sociology etc., which are of great interest and importance.

The main difficulty in this field lies in the fact that because the number of elements is large we cannot directly predict the behavior of the whole system, even if the behavior of a single element is very simple. Such a system presents in its behavior laws of statistical nature, which have nothing in common with the rules of behavior of the single element but depend on them in a rather complex way.

In dealing with such systems, one first of all considers the simplest case when the elements of the system are the same, interacting according to a set of given rules. It is natural to divide the systems in regard to a type of interaction as follows:

- (i) Systems with a random interaction.
- (ii) Systems where the interaction between the elements depends on the structure of the whole system—for example, the geometry of the position of the elements.

The easiest way to achieve progress in this field lies in the solution of some model problems, simple enough to be solved but complicated enough for all the basic features of the real problem to be present. The first and extremely important set of such model problems was suggested by the outstanding Soviet mathematician the late Michael Tsetlin [1]–[5]. His problems dealt with systems with random interaction. As an element of the system Tsetlin took a stochastic automaton, and this choice luckily determined the adequate mathematical language for the description of the system’s behavior—Markov chains. Later Prof. Pyatetskii-Shapiro stated problems dealing with systems with a structure [6]–[10].

However, even these very simple model problems had one great drawback—it was difficult to solve them analytically. From our point of view, an

* Received by the editors September 16, 1971, and in final revised form February 20, 1974. This translation into English was prepared by Dr. Yuri Meckler, Tel-Aviv University. Translation and publication of this article was supported by the National Science Foundation under Grant GN-870.

† Institute of Control Sciences, Academy of Sciences of the U.S.S.R., Moscow, U.S.S.R. Now at Department of Mathematical Sciences, Tel-Aviv University, Ramat-Aviv, Tel-Aviv, Israel.

analytical solution here is really essential. The model problem is only the rough image, caricature, of the reality, crudely distorted to make some features more evident. The justification for such a crude approximation—and at the same time its strength—lies in the opportunity to find an analytical solution to problems where our intuition generally fails. In most problems of collective behavior of automata the evolution of the system in time is described by ergodic Markov chains, so that there exists a final distribution. But it is hard to find the exact form of this distribution in transition probability matrices of extremely high order except in some very simple cases.

The idea which was used in this paper consists in applying to such systems the methods and representations of another science—statistical mechanics. The analogy between some stochastic problems and statistical mechanics had been pointed out and thoroughly discussed by Professor Mark Kac [11], [12]. Here the attempt was made to develop this analogy further and to turn it into a one-to-one correspondence between the problems of statistical mechanics and the problems of the collective behavior of automata, so that we could say, “Here is the ‘gas’ of the automata, here is the ‘solid state’ of the automata”, etc.

As is well known in statistical mechanics, the properties of a system in a state of thermodynamic equilibrium—the distribution of the microstates, the mean values, etc.—do not depend on the characteristics of the interaction between the physical systems, as long as this interaction may be considered sufficiently small, i.e., we can neglect the energy of interaction in comparison with the Hamiltonian of the system. The properties of the system in the state of equilibrium are determined only by the Hamiltonian and such macroparameters as the number of particles, the temperature or the mean energy of the system. This is the reason why the analogy between the physical and cybernetic systems can be used notwithstanding the differences in the mechanism of the interaction between the elements of the system.

Of course it is important to know how far we can follow this analogy, to what limits, because, generally speaking, in cybernetic systems there can be characteristics which have no analogy in statistical mechanics.

But even intuitive consideration permits us to see the analogy between the first type of system—many similar elements with random interaction—and the classical model of statistical mechanics, an ideal gas; and the analogy between the second type of system—many similar particles with a rigid structure—and the solid state model of statistical mechanics.

So there arises a problem of the correspondence between the notions of statistical mechanics and the Markov chain theory, in which we now describe the complex system behavior. Having done this, we should be able to use the ideas and methods of statistical mechanics in our attempts to understand the behavior of collectives much more complex than the collective of interacting molecules.

In this work the one-to-one correspondence between the problems of Markov chain theory and models of ideal gas and solid state models with different Hamiltonians is established. As a result a formula is obtained connecting the temperature and the energy of statistical physics with the matrix of transition probabilities of Markov chains, which permits us to find the exact solution of

several Markov chain problems unsolved for a considerable time. Limits of the validity of this formula are given which are connected with the concept of “reversibility” in Markov chains [13] (“detailed balance” in the terminology of this paper).

1. The model of “exchange”: the analogy of the “ideal gas” model in statistical mechanics.

1.1. The model. Let us consider a system of N automata ($N \gg 1$), each of the automata having a finite number of states ($k = 0, 1, 2, \dots, \infty$). The state of the whole system is described by a vector, with the components as the ordinal numbers of the states of all the automata from the first to the N th, (k_1, k_2, \dots, k_N) .

Then a state of the system is a point in the N -dimensional space of the system, with the numbers of the automata states being plotted along the axes.

The quantity K is given from the beginning—the sum of the numbers of the states of all the automata.

Let us determine also the rules of interaction between the automata (the model of “exchange”) as follows:

(a) At each step of the system’s functioning only two automata interact; for any pair the probabilities of interaction are equal to

$$P_{\text{int}} = 1 / \binom{N}{2}.$$

(b) Assume that before the interaction the pair of automata i and j were in the states k_i and k_j . The conditional transition probabilities are as follows:

$$\begin{aligned} P(k_i, k_j \rightarrow k_{i-1}, k_{j+1}) &= \rho, \\ P(k_i, k_j \rightarrow k_{i+1}, k_{j-1}) &= \rho, \\ P(k_i, k_j \rightarrow k_i, k_j) &= 1 - 2\rho, & k_i, k_j > 0, \\ P(0, k_j \rightarrow 1, k_{j-1}) &= \rho, \\ P(0, k_j \rightarrow 0, k_j) &= 1 - \rho, \\ P(0, 0 \rightarrow 0, 0) &= 1. \end{aligned}$$

During the subsequent interactions of pairs of automata the point which describes the state of the whole system in the N -dimensional space will travel along some trajectory. According to the rules of interaction each state of the system can pass into not more than $2\binom{N}{2}$ states (the zero components lessen the number of possibilities). It is easy to see that the transition matrix for the whole system, $P_{\alpha\beta}$ is ergodic. Therefore, there exist the final probabilities of the system’s states, $W(k_1, \dots, k_N)$. At the same time after long enough interaction with the other $(N - 1)$ automata, for each of the automata its own final distribution over the states of the automata, w_k , will be established. The problem is: to find $W(k_1, \dots, k_N)$ and w_k and to determine the connection between them.

1.2. The physical interpretation. As K and N are very large numbers, the matrix $P_{\alpha\beta}$ is practically boundless. Therefore, obtaining an explicit form of the

$W(k_1, \dots, k_N)$ and w_k is connected with overcoming immense mathematical difficulties—as it seems from first glance.

Let us interpret the problem in terms of statistical mechanics. Let us consider the ensemble with N automata as the ideal gas with N molecules, the states of the ensemble as the microscopic states of the system, the k th state of the automata as the discrete level of energy $\varepsilon_k = k$. It is easy to see that the quantity K retains its value for the states where the system passes (as is seen from the rules of interaction). This is the analogue of the “conservation energy” law under each step of the interaction: if before the interaction the sum of the ordinal numbers of the states of two automata were $k_i + k_j$, after the interaction it would be the same. The automata only exchange the units of energy. Therefore the quantity K determines the surface of constant energy in the space of the states of the system, along which the system moves.

It is easy to see that if the state α goes over to the state β with a certain probability $(\rho, (1 - \rho), 1)$, so the state β according to the rules of interaction will go over to the state α with the same probability. Thus the elements of the matrix $P_{\alpha\beta}$ are symmetrical. This symmetry is the analogue of the principle of microscopic reversibility and provides the equality of the final probabilities of the microscopic states (as a result of the twice stochastic matrix $P_{\alpha\beta}$)

$$(1.1) \quad W(k_1, \dots, k_N) = [\Omega]^{-1},$$

where Ω is the number of microscopic states, which correspond to the given energy of the system K (cf. (1.11)).

Thus we find the microcanonical (uniform) distribution for the states of the system which corresponds to the thermodynamical equilibrium of the system with the strictly given energy K .

As is well known, the statistical distribution for a small subsystem as a part of a big reservoir (thermostat) being in a state of equilibrium, is a Gibbs distribution, corresponding to some fixed temperature T :

$$(1.2) \quad w_k = \frac{1}{z} e^{-\varepsilon_k/T},$$

where ε_k is the energy of the k th level of the subsystem, and z the normalized coefficient, so-called “statistical sum”, $z = \sum_k e^{-\varepsilon_k/T}$.

In our case the subsystem is a single automaton, the thermostat is the other $(N - 1)$ automata, interacting with this single automaton. The interaction can be considered as the random heat disturbance leading to the Brownian motion of the automaton through its states.

So for the single automaton the final distribution over its states is the Gibbs distribution:

$$(1.3) \quad w_k = \frac{1}{z} e^{-k/T},$$

$$(1.4) \quad z = \sum_{k=0}^K e^{-k/T}.$$

For $N \rightarrow \infty, K \rightarrow \infty$,

$$\lim \frac{K}{N} = \bar{k},$$

where \bar{k} is a finite number, and the statistical sum tends to the limit

$$(1.5) \quad \lim_{K \rightarrow \infty} z = \sum_{k=0}^{\infty} e^{-k/T} = [1 - e^{-1/T}]^{-1}.$$

Thus we have

$$(1.6) \quad w_k = (1 - e^{-1/T}) e^{-k/T}.$$

Now we can find the “temperature” of the gas of the automata T . Bearing in mind that the mean energy of a single automaton by reason of symmetry is $\bar{k} = K/N$, we can find T from the condition

$$(1.7) \quad \bar{k} = \sum_{k=0}^{\infty} \varepsilon_k w_k.$$

Thus, using (1.6), we find that

$$\bar{k} = (1 - e^{-1/T}) \sum_{k=0}^{\infty} k e^{-k/T} = (1 - e^{-1/T}) e^{-1/T} \sum_{k=0}^{\infty} k e^{-(k-1)/T}.$$

Set

$$e^{-1/T} = \alpha.$$

Then

$$\bar{k} = (1 - \alpha) \alpha \frac{d}{d\alpha} \sum_{k=0}^{\infty} \alpha^k = \frac{\alpha}{1 - \alpha},$$

that is,

$$[e^{1/T} - 1]^{-1} = \bar{k}.$$

Thus

$$(1.8) \quad \frac{1}{T} = \log \left(1 + \frac{1}{\bar{k}} \right).$$

Hence,

$$(1.9) \quad z = (1 + \bar{k}),$$

and, finally,

$$(1.10) \quad w_k = (1 + \bar{k})^{-1} \left(\frac{\bar{k}}{\bar{k} + 1} \right)^k.$$

1.3. The combinatorial derivation. The solution given above was based on the general theorem of statistical physics on the connection between canonical and microcanonical distributions. But in this case it is not so difficult to derive the distribution w_k directly from the distribution $W(k_1, \dots, k_N)$ using combinatorial considerations.

The number of the microscopic states Ω , corresponding to the given energy K , is determined as the number of different ways in which the sum K can be

decomposed into N components [16]:

$$(1.11) \quad \Omega = \binom{N + K - 1}{N - 1}.$$

(Each component can be an integer from 0 to K , which corresponds to the Boze–Einstein statistics. Distributions which differ in the order of integers are considered as different.)

If some automaton is in the k th state, the sum of the other $(N - 1)$ automata's ordinal numbers of the states is $(K - k)$, which can be realized in Ω_k ways, where

$$\Omega_k = \binom{N + K - k - 2}{N - 2}.$$

Hence, because of the uniformity of the microscopic states,

$$(1.12) \quad w_k = \frac{\Omega_k}{\Omega} = \frac{\binom{N + K - k - 2}{N - 2}}{\binom{N + K - 1}{N - 1}}.$$

This is the strict formula for the determination of w_k for any N and K . For large N and K , where Stirling's formula can be used, we find

$$(1.13) \quad w_k = \frac{N}{N + K} \left(\frac{K}{N + K} \right)^k,$$

which coincides with (1.10). The exponential character of the expression (1.13) is due to the fact that the smaller k , the ordinal number of the state of the given automaton, the larger is the quantity $(K - k)$ which must be decomposed into $(N - 1)$ components and the greater the number of ways in which it can be decomposed.

1.4. The transition probabilities between the states of a single automaton. In the state of thermodynamical equilibrium in a system after many collisions between the automata, the transition probabilities between the states of a single automaton are established (so to speak, the “final” transition probabilities). The general course of reasoning in this case is as follows: the transition matrix of the whole system $P_{\alpha\beta} \rightarrow$ the final distribution for the whole system $W(k_1, \dots, k_N) \rightarrow$ the final distribution for the states of a single automaton $w_k \rightarrow$ the transition matrix for a single automaton $p_{k,k-1}$.

After establishing the final probabilities of being in its states, the automaton in the ensemble will collide with others with probabilities which are proportional to the final. Let us calculate for a single automaton the probability of going over from the k th state to the $(k - 1)$ st state during one elementary act of interaction. Such a transition will take place with probability ρ notwithstanding the state of the other automaton with which our automaton collides:

$$(1.14) \quad P(k \rightarrow k - 1) = \sum_{i=0}^K \rho w_i = \rho.$$

The transition from the state $(k - 1)$ to the state k will occur with probability ρ if the colliding automaton is in the states $1, 2, \dots, K$, but will not happen if it is in the state 0:

$$(1.15) \quad P(k - 1 \rightarrow k) = \sum_{i=1}^K \rho w_i = \rho(1 - w_0).$$

The fact of the inequality of the transition probabilities $P(k \rightarrow k - 1)$ and $P(k - 1 \rightarrow k)$ can be interpreted as the "potential field of probabilities". So the behavior of the automaton is the random Brownian walk over its states with transition probabilities $P(k \rightarrow k - 1)$ and $P(k - 1 \rightarrow k)$ in the potential field of the probabilities.

On the other hand, from the cybernetic point of view, the interaction of the automaton with the other $(N - 1)$ automata, pushing it "up" and "down", can be considered as the interaction with the medium (surroundings), giving the "stimulation" and "punishment". If we consider the probability of the stimulation as $q = P(k \rightarrow k - 1)$ and the probability of the punishment (the "fine") as $p = P(k - 1 \rightarrow k)$, we have the Tsetlin automaton with linear tactics [1], in which the probability of the stimulation for the first action is equal to the probability of the fine for the second action and vice versa.

Let us calculate the logarithm of the ratio of the transition probabilities:

$$\log \frac{q}{p} = \log \frac{1}{1 - w_0}.$$

According to (1.9),

$$w_0 = z^{-1} = (1 + \bar{k})^{-1}.$$

Hence,

$$\log \frac{q}{p} = \log \left(1 + \frac{1}{\bar{k}} \right).$$

Taking into account (1.8), we find that

$$(1.16) \quad \frac{1}{T} = \log \frac{q}{p}.$$

So the single automaton "feels" the temperature of the ensemble through the ratio of its transition probabilities.

As the final probabilities of its states for a certain automaton depend only on the ordinal number of the state and temperature, it is clear that they depend only on the ratio of the transition probabilities, and not on their absolute values. The absolute value influences the relaxation time of the system—the time of the final distribution being established.

In a more general form, the expression (1.16) can be written as follows:

$$(1.17) \quad \frac{\Delta E_{ij}}{T} = \log \frac{p_{ij}}{p_{ji}},$$

where ΔE_{ij} is the energy difference between the i th and j th levels (the states

between which the transition is possible)—in our case $\Delta E_{ij} = 1$; p_{ij} is the probability of transition from the i th to j th level, and p_{ji} is the probability of transition from the j th level to the i th level.

Formula (1.17) connects the description of the behavior of the automaton, determined by its transition probabilities, with the description of the automaton as a single “molecule” of the ideal gas, with the given energy K , being in a state of thermodynamical equilibrium. This formula will play an important role in the further exposition.

2. The ballot problem with random error “one-by-one”: the analogy of the one-dimensional Ising model in solid state physics. The ballot problem with random error, which was stated by Prof. Pyatetskii-Shapiro [10], is an example of the second type of system (“with structure”), where the interaction of the elements is determined by the geometry of their position in the system. It is the new class of problem, where obtaining the solution is extremely difficult. Because of this the problem was not solved for a considerable time.

The systems “with structure” interest us especially because in their behavior we can expect the “phase transition” as it takes place in the models of solid state physics. Because of this it was natural to try to solve the ballot problem as a purely physical problem, and then to prove that the solution obtained is really valid.

As a result the exact analytic solution of the “one-by-one” ballot problem was obtained and the identity of this problem with the one-dimensional Ising model, well-known in the theory of phase transitions [6], is established. There is no phase transition in this problem as would be expected in the one-dimensional case.

2.1. Statement of the problem. Here we deal with the investigation of a finite medium, consisting of n points situated on a circumference. (Below we shall refer to such a medium as “a ring”.) Each point of the ring can be in one of two states: $+1$ and -1 . Suppose that the state of each point is fixed at the moment of time t . Then at the moment of time $t + 1$ one point of the medium chosen at random (i.e., with probability $1/n$) will, with probability $(1 - \varepsilon)$, be in the same state that the majority of its neighbors, including the point itself, was in at the moment of time t , and with probability ε it will be in the other state, “achieves error”. We want to find the final probability distribution of the states of such a medium (for nonzero probability of error ε the system is ergodic).

2.2. The physical interpretation. As was done in the model of exchange the approach to the problem is based on the idea of finding the exact analogy for our problem among the models of statistical mechanics. It should be done in such a way that the matrix of transition probabilities of the system describes the process of relaxation (transition to the equilibrium) in the physical subsystem with a certain spectrum of energy levels in contact with a thermostat.

The role of the thermostat is played by an ensemble of N identical systems (N is a large number), with exactly similar data, and the role of the random thermal perturbations arising from the contact of the system with the thermostat is managed by a given form of interaction between the systems in the ensemble. Then for each system chosen from the ensemble the probabilities of transition

from state to state (assigned in the conditions of the problem) are determined as the result of interaction under the condition that the ensemble is in a state of thermodynamic equilibrium.

With this approach the difficulties arising in finding final probabilities for stochastic matrices with an arbitrarily large number of states are avoided by the appropriate choice of the Hamiltonian of the system (determining the energy as a function of the state of the system) and appropriate organization of the interaction between the systems in the ensemble.

Energy spectrum of the system. The system (ring) can be in 2^n states, each of which is an n -valued sequence of $+1$'s and -1 's. This situation recalls the one-dimensional Ising model known in statistical physics, in which the projection of the spin of each of n atoms can take on two distinct values $\sigma_i = \pm 1, i = 1, \dots, n$.

Therefore, for the Hamiltonian of the system we first take the Hamiltonian of the Ising model

$$(2.1) \quad H(\sigma_1, \dots, \sigma_n) = - \sum_{i=1}^n \sigma_i \sigma_{i+1}, \quad \sigma_1 = \sigma_{n+1}.$$

It is easy to see that, due to the symmetry of the matrix of the transition probabilities of a ring, the probability of finding $+1$ or -1 at any point of the ring is equal to $1/2$, which is equivalent to the absence of an external magnetic field in the Ising model. In accordance with the Hamiltonian (2.1) the energy state of the ring (of a certain configuration of spins) can take on values from $E_{\min} = -n$ (all spins have the same direction) to

$$E_{\max} = \begin{cases} +n & (n \text{ even}), \\ +n - 2 & (n \text{ odd}) \end{cases}$$

(neighboring spins are of opposite direction).

We note that the energy of the configuration depends only on the number of transition points l —conceptual points separating spins of opposite orientations, i.e., on the number of sign changes in the configuration ($l = 2m$, an even number). Each transition point increases the energy of the configuration in comparison with the minimal ($l = 0$) by 2, and therefore, the energy of a configuration with $l = 2m$ transition points is equal to

$$(2.2) \quad E(2m) = -(n - 4m).$$

It is not difficult to see that the number of different states with the same number of transition points $2m$ (the statistical weight) is $\Gamma(2m) = 2\binom{n}{2m}$.

Analogously to what was done in the exchange model, we form the ensemble of N rings (gas) and assign a number K to the sum of energies of all the rings (energy of the gas). We also assign a rule of interaction between rings which guarantees first of all preservation of the given K in the process of interaction (which corresponds to motion of the gas along a surface of constant energy K) and secondly, microcanonical equiprobable distribution of the probabilities of the states of the gas, determined as an ordered set of the states of all the rings forming the gas. (Microcanonical distribution is the result of reversibility of interaction.)

Then the statistical distribution for a small subsystem (one ring $l = 2m$) found in a state of equilibrium with the gas, having microcanonical distribution, is a Gibbs distribution, corresponding to some fixed temperature T ($1/T = \beta$):

$$P(2m) = \alpha e^{-\beta[(n-4m)]},$$

where α is a normalizing constant (statsum). Letting

$$(2.3) \quad e^{-2\beta} = a,$$

we obtain

$$(2.4) \quad P(2m) = \gamma a^{2m}.$$

The total probability of all states with given energy is

$$(2.5) \quad w(2m) = 2\gamma \binom{n}{2m} a^{2m}.$$

From the normalization condition $\sum_{m=0}^{n/2} w(2m) = 1$ it follows that

$$(2.6) \quad \gamma = [(1+a)^n + (1-a)^n]^{-1}.$$

Interaction of rings in a gas. According to (2.1) each point, depending on the state of its neighbors and itself, can be found in one of three positions determined by its contribution to the Hamiltonian of the system. The sum of two terms of the Hamiltonian, depending on the state of the given point, can take one of three values:

	σ_{i-1}	σ_i	σ_{i+1}	
(i)	+1	+1	+1	$\xi_i = \sigma_{i-1}\sigma_i + \sigma_i\sigma_{i+1} = 2$ (position of the index +2),
	-1	-1	-1	
(ii)	+1	-1	+1	$\xi_i = \sigma_{i-1}\sigma_i + \sigma_i\sigma_{i+1} = -2$ (position of the index -2),
	-1	+1	-1	
(iii)	+1	+1	-1	$\xi_i = \sigma_{i-1}\sigma_i + \sigma_i\sigma_{i+1} = 0$ (position of the index 0).
	+1	-1	-1	
	-1	-1	+1	
	-1	+1	+1	

For a fixed number l of transition points in the configuration, the probability that a given point is in position $\xi = 2$ is

$$q(\xi = 2) = \frac{\binom{n-2}{l}}{\binom{n}{l}}.$$

Analogously,

$$q(\xi = -2) = \binom{n-2}{l-2} / \binom{n}{l}, \quad q(\xi = 0) = 2 \binom{n-2}{l-1} / \binom{n}{l}.$$

The unconditional probability of finding a point in the ring with position $\xi = 2$ is

$$(2.7) \quad Q(\xi = 2) = \sum_{m=0}^{[n/2]} q(\xi = 2) w(2m) = \gamma \frac{[(1+a)^{n-2} + (1-a)^{n-2}]}{2}.$$

Analogously,

$$(2.8) \quad Q(\xi = -2) = \gamma a^2 \frac{[(1+a)^{n-2} + (1-a)^{n-2}]}{2},$$

$$(2.9) \quad Q(\xi = 0) = 2\gamma a \frac{[(1+a)^{n-2} + (1-a)^{n-2}]}{2}.$$

We organize the following interaction between rings in the ensemble:

(i) at each moment of time only two rings interact, and for any pair the probability of interaction is $1/\binom{N}{2}$;

(ii) in each of the two interacting rings a point is chosen at random (i.e., with probability $1/n$) and its position determined;

(iii) if the positions are $\xi_i^1 = 2$ and $\xi_i^2 = -2$ (and vice versa), then with probability ρ the points simultaneously change sign and consequently the positions change indices to the opposite ones ($\xi_i^1 = -2$, $\xi_i^2 = 2$);

(iv) if these positions are $\xi_i^1 = 0$ and $\xi_i^2 = 0$, then the points change sign simultaneously with probability ρ , and the positions remain the same ($\xi_i^1 = 0$, $\xi_i^2 = 0$);

(v) for other combinations of indices the points always retain their signs.

It is easy to verify that under the given rules the matrix of transition probabilities for the whole gas of rings is symmetric and hence doubly stochastic.

Now we determine the transition probabilities between different positions of some one point of a ring after sufficiently long interaction with the remaining $N - 1$ rings. The probability of transition at one time of an arbitrary point of a ring found at position $\xi_i = 2$ to position $\xi_i = -2$ is:

$$P(\xi_i^1 = +2 \rightarrow \xi_i^1 = -2) = Q(\xi_i^2 = -2)\rho,$$

$$P(\xi_i^2 = -2 \rightarrow \xi_i^2 = +2) = Q(\xi_i^2 = +2)\rho.$$

According to the ballot problem statement, we must have the relationship

$$\frac{P(\xi_i = +2 \rightarrow \xi_i = -2)}{P(\xi_i = -2 \rightarrow \xi_i = +2)} = \frac{\varepsilon}{1 - \varepsilon}.$$

But, taking (2.7) and (2.8) into account, we have

$$\frac{P(\xi_i = +2 \rightarrow \xi_i = -2)}{P(\xi_i = -2 \rightarrow \xi_i = +2)} = a^2.$$

Hence,

$$(2.10) \quad \frac{\varepsilon}{1 - \varepsilon} = a^2.$$

When (2.10) is satisfied, the final probability of the state of the system is

$$(2.11) \quad P(2m) = \gamma \left(\frac{\varepsilon}{1 - \varepsilon} \right)^m.$$

The summed probability of all the states with the given number of sign changes is

$$(2.12) \quad w(2m) = 2\gamma \binom{n}{2m} \left(\frac{\varepsilon}{1 - \varepsilon} \right)^m,$$

where γ is a normalizing coefficient

$$(2.13) \quad \gamma = \left[\left(1 + \sqrt{\frac{\varepsilon}{1 - \varepsilon}} \right)^n + \left(1 - \sqrt{\frac{\varepsilon}{1 - \varepsilon}} \right)^n \right]^{-1}.$$

The probability of error ε plays the role of a parameter giving the temperature of the system. As in the model of exchange, the logarithm of the ratio of the probabilities of transition between two states of the system is equal to the ratio of the difference of energies of the corresponding states to the temperature:

$$(2.14) \quad \frac{\Delta E}{T} = \frac{4}{T} = \log \frac{1 - \varepsilon}{\varepsilon}.$$

We can see that if in this case we just used the formula (1.17) directly, as if it were valid for all cases, we would immediately come to (2.14) and then directly to the result (2.11), (2.12), (2.13), omitting all the preceding calculations. Really, as a result of sign changing at the arbitrarily chosen point, the index of the position can change from $\xi = +2$ to $\xi = -2$ (or vice versa), which means that the energy of the configuration can change by 4 units only. But according to the statement of the problem, the ratio of the transition probabilities in one elementary act of interaction is $\varepsilon/(1 - \varepsilon)$, which immediately leads to (2.14).

In particular,

- (i) $T = 0$, $\varepsilon = 0$, $2m = 0$ —spins of the same directions in all the rings,
- (ii) $T = \infty$, $\varepsilon = \frac{1}{2}$ —all possible configurations are equiprobable,
- (iii) $T = -0$, $\varepsilon = 1$, $2m = n$ —in all rings neighboring spins are of opposite direction.

2.3. Proof. We prove the obtained formula (2.11) directly by substitution in the equation for the final probabilities of the original system. Suppose we have a configuration with $2m$ transition points containing k points of position index $+2$, r points of position index -2 , and s points of position index 0 , so that $n = k + r + s$. Since the balloting each time takes place at one point, the given configuration can be obtained by translation from n other configurations differing from the given one by only one position, and also by transition from the given configuration to itself. Configurations differing from the given one by one position of index 0 have the same number of transition points, $2m$; configurations differing from the given one by one position of index 2 (in place of $\xi = -2$ in the given one) contain $2m - 2$ transition points, and, finally, configurations differing from the given one by a position index -2 (in place of $\xi = +2$) contain $2m + 2$ transition points.

Hence,

$$(2.15) \quad P(2m) = \frac{1}{n} \left[s\delta P(2m) + s(1 - \delta)P(2m) + r\varepsilon P(2m) + k(1 - \varepsilon)P(2m) + r\varepsilon P(2m - 2) + k(1 - \varepsilon)P(2m + 2) \right],$$

where δ is the probability of "error" in a position of index 0, $\delta = wQ(\xi = 0)$. The quantity δ obviously has no influence on the form of the final distribution. It is easy to prove that substitution in (2.15) of the values of $P(2m)$ according to (2.11) (for $l = 2m, 2m - 2, 2m + 2$) transforms it into an identity.

Thus the final distribution in the ballot problem with random error corresponds to a state of thermodynamic equilibrium in the one-dimensional Ising model. As is known from statistical physics, in the one-dimensional Ising model with infinite growth of the dimensions of the system ($n \rightarrow \infty$) the correlation between directions of spins positioned far from each other disappears. Indeed, the probability that a point separated from a given one by f intervals has the same sign as the given one is

$$w\left(-1 / \underbrace{\dots}_f -1\right) = \frac{[(1+a)^{n-f} + (1-a)^{n-f}][(1+a)^f + (1-a)^f]}{2[(1+a)^n + (1-a)^n]},$$

$$\lim_{n \rightarrow \infty, f \rightarrow \infty} w\left(-1 / \underbrace{\dots}_f -1\right) \rightarrow \frac{1}{2}.$$

By the same token, as was to be expected, phase transitions are lacking in the system.

3. The generalized "one-by-one" ballot problem with random error: the analogy of the solid state model with a general form additive Hamiltonian.

3.1. Statement of the problem. As has been shown in § 2, the one-by-one ballot problem with random error is identical to the unidimensional Ising problem with the Hamiltonian $H = -\sum_i \sigma_i \sigma_{i+1}$.

Obviously there is some correspondence between the form of the Hamiltonian of a physical system and the statement of a generalized ballot problem defined by the following parameters:

- (i) the number and the position of the neighbors of the "voting" point;
- (ii) the probability table of sign change by the voting point depending on the state of the neighbors and its own state;
- (iii) the "way" of balloting (simultaneous—if at each moment all points of a ring are voting; one-by-one—if at each moment one point, chosen at random, is voting).

In this section only one-by-one ballot problems will be discussed. There arises a problem of assigning to each physical system with a general form Hamiltonian such a statement of a generalized ballot problem that the final probabilities

of the ballot problem would be determined by the system Hamiltonian H in such a way that $w = (1/z) e^{-H/T}$, where T is the temperature of the system.

Let us consider, as in § 2, an ensemble consisting of N rings ($N \gg 1$). On each ring there are n points each one of which can be in one of two states: $\sigma_i = +1$ and $\sigma_i = -1$, $i = 1, \dots, n$. Correspondingly every ring can be at each moment in 2^n states.

The energy of every ring equals the value of the Hamiltonian for a given state of the ring. The sum of the energies of all the rings K , is also given.

Further, we shall consider the Hamiltonian of the form

$$(3.1) \quad H = - \sum_i \phi(\sigma_i, \sigma_{i+1}, \dots, \sigma_{i+h}),$$

where h is the integer constant, smaller than n .

The function ϕ may actually depend on some of the arguments $\sigma_i, \dots, \sigma_{i+h}$. The condition $\phi(\sigma_i, \dots, \sigma_{i+j}, \dots, \sigma_{i+h}) = -\phi(\sigma_i, \dots, -\sigma_{i+j}, \dots, \sigma_{i+h})$ should be satisfied for every σ_{i+j} , $0 \leq j \leq h$, on which ϕ depends.

3.2. The solution. The Hamiltonian being additive, those of its members which include a contribution of the i th point can be set apart. The index of the i th point is determined as the sum of the values of these members. The points entering this sum together with the i th point are its neighbors.

For example, the contribution of the i th point in the Hamiltonian

$$H = - \sum_i \sigma_i \sigma_{i+1}$$

considered in § 2, is determined by the members $\xi_i = \sigma_i(\sigma_{i-1} + \sigma_{i+1})$. Consequently the neighbors of the i th point are the point to the right ($i+1$) and the point to the left ($i-1$).

Depending on the sign of the i th point, its index can have a set of values consisting of pairs of values equal in absolute magnitude and different in sign, $\pm \xi_\mu(i)$, $\mu = 1, \dots, v$.

Let us now arrange an interaction between the rings of the ensemble which would have the following properties:

- (a) ergodicity (the matrix of the transition probabilities of the ensemble of rings must be ergodic);
- (b) reversibility (the matrix of the transition probabilities of the ensemble of rings must be symmetrical);
- (c) conservation of the summary energy of the ensemble, K .

In particular, the interaction described below has the above properties:

1. At each moment of time two rings are picked at random (with probability $1/\binom{N}{2}$) out of the ensemble.
2. In each one of the two rings a point is picked randomly (with a probability $1/n$) and its index is determined.
3. If the indices of the two points are equal at absolute magnitude and have different signs, the points interact with each other. If not, the states of the ring do not change.

4. As a result of the interaction the two points simultaneously change their signs with probability ρ , and the indices of the points are therefore changed to the opposite.

It is easily checked that the above rules guarantee the ergodicity, the symmetry of the transition probabilities matrix of the ensemble and the conservation of the energy K with time, and thus guarantee the microcanonical distribution for the whole system of rings and Gibbs distribution for the configurations of a single ring.

As a consequence of an elementary interaction, the energy of a ring can change only by the double index value of the interacting points, and the energies of the states are quantified correspondingly by $2\xi_\mu$, $\mu = 1, \dots, v$, units of energy.

Further, we shall use formula (1.17), according to which the relation

$$(3.2) \quad \frac{2\xi_\mu}{T} = \log \frac{P(-\xi_\mu)}{P(+\xi_\mu)}$$

must hold for every μ , where $P(\xi)$ is the probability of sign change by the point with the index ξ .

The relation (3.2) holds in particular if $P(\xi) = p^\xi$ while $\xi > 0$ and $P(-\xi) = q^{-\xi}$ while $\xi < 0$, where $0 < p < 1$, $0 < q < 1$.

From here we obtain the expression for the temperature of the system

$$(3.3) \quad \frac{2}{T} = \log \frac{q}{p}.$$

The values of $P(\xi)$ obtained are the values of the probability table for sign change by a voting point depending on its state and the state of its neighbors. This table, in addition to indicating the neighbors, describes the statement of the one-by-one ballot problem whose solution is determined by the Hamiltonian given.

It remains to check the solution by substituting it into the final equations of the problem. Let us show in some examples the use of the concepts introduced above.

I. The Hamiltonian of a ring:

$$(3.4) \quad H = -\sum_i \sigma_{i-1} \sigma_{i+1}.$$

The index of the i th point is

$$\xi_i = \sigma_i(\sigma_{i-2} + \sigma_{i+2}).$$

Thus, the i th point has two neighbors: one on the left ($i-2$) and one on the right ($i+2$).

According to the Hamiltonian, ξ can take on the values $\pm \xi_1 = 2$, $\pm \xi_2 = 0$.

According to (3.2) we obtain

$$\frac{2\xi_1}{T} = \frac{4}{T} = \log \frac{\delta}{\epsilon},$$

$$\frac{2\xi_2}{T} = 0 = \log \frac{\alpha}{\beta},$$

for any $\alpha = \beta$.

The probability table for sign changes by a voting point is as follows:

Probability of sign change						δ				
ε										
Types of positions	σ_{i-2}	σ_{i-1}	σ_i	σ_{i+1}	σ_{i+2}	σ_{i-2}	σ_{i-1}	σ_i	σ_{i+1}	σ_{i+2}
	+1	.	+1	.	+1	+1	.	-1	.	+1
	-1	.	-1	.	-1	-1	.	+1	.	-1

Probability of sign change						∞				
Types of positions	+1	.	+1	.	-1	-1	.	+1	.	+1
	+1	.	-1	.	-1	-1	.	-1	.	+1

The form of the Hamiltonian given shows that the final probability of a state depends only on $2g$, the number of sign changes in the configuration if it is considered by skipping every other symbol: $\sigma_1, \sigma_3, \sigma_5, \dots$.

The energy of this configuration equals

(3.5)
$$E(2g) = -(n - 4g).$$

Thus the final probability of a state of the ring with $2g$ sign changes in a configuration considered by skipping every other symbol is

(3.6)
$$P(2g) = \gamma \left(\frac{\varepsilon}{\delta} \right)^g.$$

Let us check this solution by substitution in the equation for the final probabilities in the ballot problem. Assume a configuration with $2g$ transition points, skipping every other symbol, which contains k points in the position index $+2$, r points in the position index -2 , and s points in the position index 0 (all considered skipping every other symbol), so that $n = k + r + s$. As the voting occurs in one point only each time, the configuration given can be obtained by transition from n other configurations differing from it by only one position and also by transition of the configuration into itself. The configuration differing from the given one by a position indexed $\xi = 0$ has the same number of transition points, $2g$, the configuration, differing from the given by the position indexed $\xi = +2$ (instead of $\xi = -2$ in the given one) has $2g - 2$ transition points, and finally, the configuration, differing from the given one by the position indexed $\xi = -2$ (instead of $\xi = +2$) contains $2g + 2$ transition points.

Hence

$$P(2g) = \frac{1}{n} [s\alpha P(2g) + s(1 - \alpha)P(2g) + r(1 - \delta)P(2g) + k(1 - \varepsilon)P(2g) + r\varepsilon P(2g - 2) + k\delta P(2g + 2)].$$

It is easy to check that substituting into the equation of the solution (3.6) transforms it into an identity.

II. The Hamiltonian of a ring:

(3.7)
$$H = -\sum_i \sigma_{i-1} \sigma_i \sigma_{i+1}.$$

The index of the *i*th point is

$$\xi_i = \sigma_i(\sigma_{i-2}\sigma_{i-1} + \sigma_{i-1}\sigma_{i+1} + \sigma_{i+1}\sigma_{i+2}).$$

Thus the *i*th point has four neighbors: two on the right—(*i* + 1), (*i* + 2), and two on the left—(*i* − 2), (*i* − 1).

According to the Hamiltonian (3.7) the index of the point can have the values $\pm \xi_1 = 1$ and $\pm \xi_2 = 3$. Opposite to what it had been before, here the sign variants obtained by substituting +1 for −1 (and vice versa) have the indices of opposite signs (but of the same absolute value).

Using (3.2) we obtain

$$\begin{aligned} \frac{2\xi_1}{T} &= \frac{2}{T} = \log \frac{\delta}{\varepsilon}, \\ \frac{2\xi_2}{T} &= \frac{6}{T} = \log \frac{\delta^3}{\varepsilon^3}. \end{aligned}$$

The probability table for sign changes by a voting point is as follows:

Probability of sign change						ε^3					δ^3				
Types of position	σ_{i-2}	σ_{i-1}	σ_i	σ_{i+1}	σ_{i+2}	σ_{i-2}	σ_{i-1}	σ_i	σ_{i+1}	σ_{i+2}	σ_{i-2}	σ_{i-1}	σ_i	σ_{i+1}	σ_{i+2}
	+1	+1	+1	+1	+1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
	+1	-1	-1	+1	-1	+1	+1	-1	+1	+1	+1	+1	-1	+1	+1
	-1	-1	+1	-1	-1	+1	-1	+1	+1	+1	+1	-1	+1	-1	-1
	-1	+1	-1	-1	+1	-1	+1	+1	-1	+1	-1	+1	-1	+1	+1
etc.						etc.									
Index						$\xi = +3$					$\xi = -3$				

Probability of sign change						ε					δ							
Types of position		-1	+1	+1	+1	+1		-1	+1	-1	+1	+1		-1	+1	-1	+1	+1
		-1	+1	-1	+1	-1		-1	+1	+1	+1	-1		-1	+1	+1	+1	-1
		etc.						etc.						etc.				
Index						$\xi = +1$					$\xi = -1$							

To determine the final probabilities of configurations let us match to each configuration the succession of Hamiltonian members $\psi_i, \psi_i = \pm 1$, calculated for each point from 1 to *n*, according to the rule $\psi_i = \sigma_{i-1}\sigma_i\sigma_{i+1}$.

Then the energy of the configuration for which the corresponding succession ψ_i contains *d* negative numbers is

(3.8)
$$E(d) = -[n - 2d]$$

and the final probability of such a state is

$$(3.9) \quad P(d) = \gamma \left(\frac{\varepsilon}{\delta} \right)^d.$$

It is easy to check that substituting into the final equation of the solution (3.9) turns this equation into an identity.

III. The Hamiltonian of the ring:

$$(3.10) \quad H = - \left[\sum_i \sigma_i \sigma_{i+1} + \sum \sigma_i \right]$$

(Ising model with external field).

The index of the i th point is

$$\xi_i = \sigma_i(\sigma_{i-1} + \sigma_{i+1} + 1).$$

Thus the point has two neighbors: one on the left ($i - 1$) and one on the right ($i + 1$).

The probability table of sign changes by a voting point is as follows:

Probability of sign change	ε^3	δ^3	ε	δ
Types of position			+1 +1 -1	+1 -1 -1
+1 +1 +1	+1 -1 +1	-1 +1 +1	-1 -1 -1	-1 +1 -1
Index	$\xi = +3$	$\xi = -3$	$\xi = +1$	$\xi = -1$

The final probability of the configuration having l negative members of the Hamiltonian is

$$(3.11) \quad P(l) = \gamma \left(\frac{\varepsilon}{\delta} \right)^l.$$

Note. There arises a question: why does relation (3.2) for the probabilities of sign change by the voting point guarantee the existence of a solution of the form $w = 1/z e^{-H/T}$?

A careful analysis of all similar situations discloses a general law: the transition probabilities matrix for the states of the ring turns out to be reversible in these cases. We shall discuss the meaning of this fact in detail in the next section.

4. The detailed balance in finite Markov chains. Ballot problems with simultaneous transition. Assume a discrete ergodic Markov chain with a finite number of states n , $S = \{s_1, s_2, \dots, s_n\}$ and a probability matrix $P = \|p_{ij}\|$ for transition from the state s_i into state s_j , $i = 1, \dots, n$; $j = 1, \dots, n$. For such a Markov chain there exist final probabilities of states satisfying the equation system

$$w_j = \sum_i w_i p_{ij}.$$

DEFINITION 1. If for two states of a Markov chain, s_i and s_j , the relation

$$(4.1) \quad w_i p_{ij} = w_j p_{ji}$$

holds, we shall say that there exists between these states a *detailed balance*. If a detailed balance exists between any two states, then the Markov chain is reversible [13]. (We shall say that in a reversible chain there exists a *strong detailed balance*).

Let us assign each Markov chain a full oriented graph with the apices of the graph corresponding to the states of the chain; the oriented rib (i, j) is assigned the probability p_{ij} and also the value $w_i p_{ij}$ which we shall call the *probability flow* from the i th state to the j th. Then the detailed balance between a pair of states s_i and s_j means that the summary probability flow on the rib (i, j)

$$(w_i p_{ij} - w_j p_{ji})$$

equals zero.

THEOREM 1. For a strong detailed balance (reversibility) to exist in an ergodic Markov chain with a finite number of states it is necessary and sufficient that for any series of states of the chain $\{s_{i_1}, s_{i_2}, \dots, s_{i_k}\}$ the following condition be satisfied:

$$(4.2) \quad p_{i_1 i_2} p_{i_2 i_3} \cdots p_{i_{k-1} i_k} p_{i_k i_1} = p_{i_1 i_k} p_{i_k i_{k-1}} \cdots p_{i_2 i_1}.$$

Proof. 1. *Necessity.* According to (4.1),

$$w_{i_k} p_{i_k i_1} \prod_{l=1}^{k-1} w_{i_l} p_{i_l i_{l+1}} = w_{i_1} p_{i_1 i_k} \prod_{l=1}^{k-1} w_{i_{l+1}} p_{i_{l+1} i_l},$$

from which (4.2) follows immediately.

2. *Sufficiency.* Let $p_{i_2 i_1} \neq 0$. From (4.2),

$$(4.3) \quad \frac{p_{i_1 i_2}}{p_{i_2 i_1}} = \frac{p_{i_1 i_k} \prod_{l=2}^{k-1} p_{i_l i_{l+1}}}{p_{i_k i_1} \prod_{l=2}^{k-1} p_{i_l i_{l+1}}}.$$

Since (4.3), according to the hypothesis, holds for any series of states $\{s_{i_l}\}$, $l = 1, 2, \dots, k$, then

$$(4.4) \quad \frac{p_{i_1 i_2}}{p_{i_2 i_1}} = \frac{\sum_{i_3} \cdots \sum_{i_k} p_{i_1 i_k} p_{i_k i_{k-1}} \cdots p_{i_3 i_2}}{\sum_{i_3} \cdots \sum_{i_k} p_{i_2 i_3} \cdots p_{i_{k-1} i_k} p_{i_k i_1}} = \frac{(P)_{i_1 i_2}^{k-1}}{(P)_{i_2 i_1}^{k-1}}.$$

For an ergodic Markov chain,

$$\lim_{k \rightarrow \infty} (P^k)_{ij} = w_j.$$

Hence,

$$(4.5) \quad \frac{p_{i_1 i_2}}{p_{i_2 i_1}} = \lim_{k \rightarrow \infty} \frac{(P^{k-1})_{i_1 i_2}}{(P^{k-1})_{i_2 i_1}} = \frac{w_{i_2}}{w_{i_1}}.$$

Suppose now $p_{i_2 i_1} = 0$. Then, by virtue of the ergodicity of the chain, there exists a sequence of states s_{i_l} , $l = 3, 4, \dots, k$, such that $p_{i_2 i_3} \cdot p_{i_3 i_4} \cdots p_{i_k i_1} \neq 0$. For

this sequence of states, the equality (4.2) holds only in the case when $P_{i_1 i_2} = 0$. But then (4.1) is satisfied trivially. The theorem is proved.

Thus, a necessary and sufficient condition of a strong detailed balance is constituted by the probabilities of going round a cycle in two opposite directions being equal in the transition graph of a Markov chain. (By the probability of going round a cycle in any given direction is meant the product of all the transition probabilities corresponding to the ribs of the cycle.)

In contrast to the criterion of reversibility of a Markov chain cited in [13] (Theorem 5.3.3) the above criterion does not require knowing the final probabilities. On the contrary, whenever (4.2) holds, the final probabilities can be simply deduced from (4.1) and the normalization condition.

COROLLARY 1. *There always exists a strong detailed balance in an ergodic Markov chain for which the graph, obtained from the transition graph by omitting all the ribs (i, j) with $p_{ij} = 0$, is a tree.*

COROLLARY 2. *In an ergodic Markov chain for which the transition graph is complete (all $p_{ij} > 0$), a necessary and sufficient condition of a strong detailed balance is for the condition (4.2) to hold for any cycle of three states. (This condition is noted in [13].)*

Now let the set of states S of the Markov chain be broken, by virtue of some a priori considerations, into m classes S^1, S^2, \dots, S^m of states whose final probabilities are equal, so that

$$(4.6) \quad w_i = w^l$$

if

$$i \in S^l, \quad |S^l| = K_l, \quad \sum_{l=1}^m K_l = n.$$

Then the final probability of the system to be in one of the states of the class S^l equals

$$(4.7) \quad W^l = \sum_{i \in S^l} w_i = K_l w^l.$$

In this case the probabilities of transitions between classes are as follows:

$$(4.8) \quad P^{lr} = \frac{\sum_{i \in S^l} \sum_{j \in S^r} w_i p_{ij}}{\sum_{i \in S^l} w_i} = \frac{1}{K_l} \sum_{i \in S^l} \sum_{j \in S^r} p_{ij},$$

that is, they do not depend on final probabilities of states and can be expressed immediately through transition probabilities of the original Markov chain. Thus, whenever (4.6) holds it is possible to *merge* the states of the original Markov chain belonging to the same class and obtain a new Markov chain with a set of states $\{S^l\}$, transition probabilities P^{lr} and final probabilities W^l .

DEFINITION 2. We shall say that in the original Markov chain for which (4.6) is satisfied, a *weak detailed balance* exists whenever there is a strong detailed balance in the Markov chain generated by the classes of states of the original

chain, that is, if

$$(4.9) \quad W^l P^{lr} = W^r P^{rl}$$

or, otherwise, if

$$(4.9') \quad w^l \sum_{i \in S^l} \sum_{j \in S^r} p_{ij} = w^r \sum_{j \in S^r} \sum_{i \in S^l} p_{ji}.$$

Theorem 1 implies Theorem 2.

THEOREM 2. *A necessary and sufficient condition of a weak detailed balance is given by the condition that for any sequence of classes of states $\{S^{lg}\}$, $g = 1, 2, \dots, h$, the equality*

$$(4.10) \quad p^{i_1 i_2} p^{i_2 i_3} \dots p^{i_{h-1} i_h} p^{i_h i_1} = p^{i_1 i_h} p^{i_h i_{h-1}} \dots p^{i_2 i_1}$$

holds true.

By using the concepts of strong and weak detailed balance the exact solution can be found for various versions of the ballot problem with simultaneous transition of all the points—which has not been found for a long time. From the point of view of a detailed balance there is no difference now between problems with simultaneous transition and transition “one-by-one”—although these two kinds of problem are fundamentally different from the point of view of statistical mechanics, and the Hamiltonians of the physical problems analogous to the ballot problems with simultaneous transition look somewhat strange.

Below we shall state several theorems concerning two of the simultaneous transition ballot problems and the generalized problem of “one-by-one” ballot.

THEOREM 3. *Suppose in the “one-by-one” ballot problem the probabilities of sign change by a voting point depending on the state of its neighbors (one on the left and one on the right) and its own state are given by the table:*

The probability of sign change			κ			ρ			δ		
Types of position			σ_{i-1}	σ_i	σ_{i+1}						
	+1	+1	+1			+1	-1	+1	+1	+1	-1
	-1	-1	-1			-1	+1	-1	-1	-1	+1

Then for any values of the parameters κ, ρ, δ not equal to 0 or 1, in the Markov chain, the states of which are various configurations of signs of the points, there is a strong detailed balance. The final probability of every configuration for which the number of sign changes equals $2m$, $m = 0, 1, \dots, [(n+1)/2]$, is

$$(4.11) \quad w(2m) = \gamma \left(\frac{\kappa}{\rho} \right)^m,$$

where $\gamma = [(1 + \sqrt{\kappa/\rho})^n + (1 - \sqrt{\kappa/\rho})^n]^{-1}$ is a normalized constant. The number of configurations having the same number of sign changes (the statistical weight) is $\Gamma(2m) = 2 \binom{n}{2m}$.

Proof. It is necessary to prove that the condition (4.2) is satisfied for every cycle of the chain.

1. Only states differing from the original one not more than in one point can be obtained in one step in this problem; therefore, the length of the state sequence s_0, \dots, s_l , where s_l differs from s_0 in k points (and $\prod_{i=1}^{l-1} p_{i,i+1} \neq 0$), can not be less than $(k+1)$ (that is the number of steps $l \geq k$). We shall prove that it is sufficient to consider only cycles in which the states differing in k points are exactly k steps apart from each other.

Assume a cycle $s_0, s_1, \dots, s_l, s_{l+1}, \dots, s_{l+m}, s_0$ in which the states s_0 and s_l differ in k points, where $l > k$ and $m+1 > k$. Then a state sequence $s_0, s'_1, s'_2, \dots, s'_{k-1}, s_l$ exists which can lead from the state s_0 to state s_l in k steps exactly (see Fig. 1). If (4.2) holds for the cycles $s_0, s_1, \dots, s_l, s'_{k-1}, s'_{k-2}, \dots, s'_1, s_0$ and

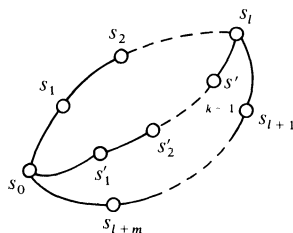


FIG. 1

$s_0 s'_1, \dots, s'_{k-1}, s_l, s_{l+1}, \dots, s_{l+m}, s_0$ then, as is easy to check, it is true for the original cycle also. Thus, it is sufficient to consider two shorter cycles instead of the original one, in which the states s_0 and s_l are exactly k steps apart. This implies that it is sufficient to prove that relation (4.2) is satisfied for cycles of length $2k \leq 2n$ having the form $s_0, s_1, \dots, s_{k-1}, s_k, s'_{k-1}, \dots, s'_1, s_0$, where the states s_i and s'_i , $i = 1, 2, \dots, k$, are different from s_0 in i points.

2. As any state in a cycle (see Fig. 2) can be assumed to be an original one, it is clear that the states s_i and s'_{k-i} differ in k points as well as s_0 and s_k , and that in transition from the states s'_{k-i} to s'_{k-i-1} the same point changes its sign as in the transition from s_i to s_{i+1} .

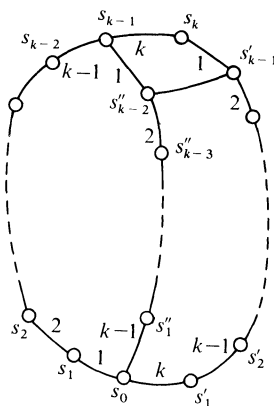


FIG. 2

Let us number the points in the order of changing their signs. Then the cycle considered will assume the form shown in Fig. 2 where the numbers at the ribs denote the number of the point whose sign is changed in the corresponding transition.

Using induction, let us prove that relation (4.2) is satisfied for the cycles of the form described.

3. It can be checked immediately that (4.2) holds true for cycles of the length $2k = 4$. Suppose it to be true for all cycles with a number of states not more than $2(k - 1)$. The cycle $s_0, s_1, \dots, s_{k-1}, s_k, s'_{k-1}, \dots, s'_1, s_0$ with length $2k$ can be broken into 3 cycles of lesser length $s_0, s_1, \dots, s_{k-1}, s''_{k-2}, \dots, s''_1, s_0; s_0, s'_1, \dots, s'_{k-2}, s'_{k-1}, \dots, s'_1, s_0$ and $s_{k-1}, s_k, s'_{k-1}, s''_{k-2}, s_{k-1}$ (see Fig. 2) by introducing the states $s''_i, i = 1, \dots, k - 2$, differing from the state s_0 in the points numbered $k - 1, k - 2, \dots, k - i$. As (4.2) holds for these three cycles according to the hypothesis, it also holds for the original cycle of length $2k$.

4. Having established the existence of a detailed balance for all cycles, it is easy to obtain an explicit expression for the final probabilities (4.11). The theorem is proved.

Note that the probability of sign change by a voting point has been used essentially only in the proof of the detailed balance in all cycles of 4 states. Hence we have the following corollary.

COROLLARY. *In all "one-by-one" ballot problems with a random error in which there is a detailed balance in every cycle of 4 states, there is a strong detailed balance.*

THEOREM 4. *In the problem analogous to the one discussed in Theorem 3 but with simultaneous ballot, the necessary and sufficient condition of a strong detailed balance takes the form*

$$(4.12) \quad \frac{\kappa}{1 - \kappa} \cdot \frac{\rho}{1 - \rho} \cdot \frac{(1 - \delta)^2}{\delta^2} = 1$$

(κ, ρ, δ do not equal 0 or 1).

Under the condition of a strong detailed balance the final probabilities of the configurations depend only on the number of positions of various types. The final probability of every configuration containing k positions of the type

$$r \text{ positions of the type } \begin{matrix} +1 & -1 & +1 \\ -1 & +1 & -1 \end{matrix} \text{ and } s = 2g \text{ positions of the type } \begin{matrix} +1 & +1 & -1 \\ -1 & -1 & +1 \end{matrix},$$

$$\begin{matrix} +1 & -1 & -1 \\ -1 & +1 & +1 \end{matrix}, \quad k + r + s = n, \text{ is}$$

$$(4.13) \quad w(k, r, s) = \beta [\kappa(1 - \kappa)]^{-k/2} [\rho(1 - \rho)]^{-r/2} [\delta(1 - \delta)]^{-s/2},$$

where β is a normalized constant.

The statistical weight of a state with given numbers of positions k, r, s equals

$$\Gamma(k, r, s) = \frac{n!}{g!} \binom{r + g - 1}{g - 1} \binom{n - r - g - 1}{g - 1}.$$

Proof. As in the problem with simultaneous ballot, all $p_{ij} > 0$, it is sufficient for proving the theorem, according to Corollary 2, to find a necessary and sufficient condition for relation (4.2) holding true for every cycle of 3 states.

1. Consider a cycle of 3 states, in which two states differ only by the sign of one point, the third state being arbitrary. (We shall call such a cycle elementary.) We shall prove that relation (4.2) holding true for every elementary cycle is a necessary condition of a strong detailed balance. The proof will be based on induction.

Consider any cycle of three states (see Fig. 3) s_1, s_k, s^0, s_1 , where the states

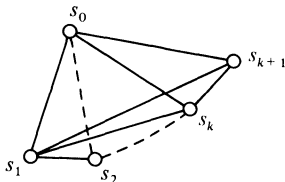


FIG. 3

s_1 and s_k differ in $(k - 1)$ points, the state s^0 being arbitrary.

With $k = 2$ the cycle is an elementary one.

Assume (4.2) to hold true for all $k' \leq k$, in every cycle of this form. Consider the cycle s_1, s_{k+1}, s^0, s_1 , where the states s_1 and s_{k+1} differ in k points. Obviously a state s_k exists which differs from the state s_1 in $k - 1$ points and from the state (s_{k+1}) only in one point. The cycle s_1, s_k, s^0, s_1 and the elementary cycle s^0, s_k, s_{k+1}, s^0 satisfy relation (4.2) by hypothesis. Consequently the relation is satisfied by the cycle $s_1, s_k, s_{k+1}, s^0, s_1$ also. Hence, as the cycle s_1, s_k, s_{k+1}, s_1 is an elementary one and satisfies relation (4.2), the relation is also satisfied by the cycle s_1, s_{k+1}, s^0, s_1 . Thus, it is proved that relation (4.2) being satisfied by the elementary cycles is sufficient for the relation holding true in any cycle of 3 states, and, consequently, for a detailed balance in all the chain.

2. Let us find a condition for satisfying relation (4.2) for an elementary cycle, where the states s_1 and s_2 differ in one point, the state s_3 being arbitrary (see Fig. 4). Note, that a sign change in one point changes the probability of

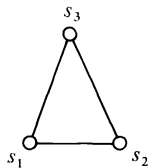


FIG. 4

transition only in three points—the one changed and its two neighbors. Let us denote by π_{ij} the product of transition probabilities for these three points and by p'_{ij} the product of transition probabilities of all other points. Then

$$(4.14) \quad p_{ij} = p'_{ij} \cdot \pi_{ij}, \quad i, j = 1, 2, 3.$$

Evidently,

$$(4.15) \quad p'_{12} = p'_{21}, \quad p'_{13} = p'_{23}, \quad p'_{31} = p'_{32}.$$

Hence, relation (4.2) for the cycle considered,

$$p_{12}p_{23}p_{31} = p_{13}p_{32}p_{21},$$

is reduced to the condition

$$(4.16) \quad \pi_{12}\pi_{23}\pi_{31} = \pi_{13}\pi_{32}\pi_{21}.$$

The probabilities of transition π_{ij} depend on the combinations of types of positions in these three points. Allowing for the sign and inversion (with relation to the middle point) symmetry, there exist 10 different combinations of position types in the original state s_1 .

Considering all the possible cases, it is easy to ascertain that (4.16) is either satisfied identically or it can be reduced to one and the same condition (4.12).

3. Let us now prove that the final probabilities are given by formula (4.13). By virtue of the detailed balance (reversibility) it is sufficient to prove that for any two states

$$(4.17) \quad \frac{w_i}{w_j} = \frac{p_{ji}}{p_{ij}} = [\chi(1 - \chi)]^{(k_j - k_i)/2} \cdot [\rho(1 - \rho)]^{(r_j - r_i)/2} \cdot [\delta(1 - \delta)]^{(s_j - s_i)/2}.$$

Let relation (4.17) hold for any two states differing in one point only. Then, making use of the detailed balance, it is easy to prove by induction that the relation holds true for any two states. Considering the abovementioned 10 different combinations of position types for states differing in one point and taking into account relation (4.12), we ascertain that expression (4.17) does indeed give the relation of transition probabilities, which proves that (4.13) holds. The theorem is proved.

Let us consider further the problem of simultaneous ballot on a circumference, in which the probability of sign change by a point depends on the state of itself and only one (left or right) of its neighbors. A similar problem on an infinite line has been discussed in [17].

THEOREM 5. *Suppose that in the simultaneous ballot problem with a finite number of points n , placed on a circle, the probabilities of sign change by a voting point depending on its state and the state of its right neighbor are given by the table:*

The probability of sign change	λ		μ	
Types of position	+1	+1	+1	-1
(left point votes)	-1	-1	-1	+1

Then a strong detailed balance exists only when $\lambda = \mu = \frac{1}{2}$ (in this case the final probabilities of all the configurations are equal). A weak detailed balance between configuration classes differing by a shift along the circle (translation variants) and by changing the signs to opposite (sign variants), exists for any value, not 0 or 1, of the parameters λ and $\mu = \frac{1}{2}$.

The final probability of the configuration depends on the number of sign changes in the configuration $2m$ and equals

$$(4.18) \quad w(2m) = \alpha[4\lambda(1 - \lambda)]^m,$$

where α is a normalized constant. The statistical weight of a configuration with a given number of sign changes is $\Gamma(2m) = 2\binom{n}{2m}$.

Proof. 1. By deductions analogous to those used in the proof of Theorem 4 it is easy to establish that strong detailed balance exists only when $\lambda = \mu = \frac{1}{2}$.

2. Consider the case of a weak detailed balance. Let S^1, S^2, \dots , be the classes of translational variants of various configurations. A weak detailed balance exists if, for any sequence of classes of translational variants, relation (4.10) holds true. With $\mu \neq \frac{1}{2}$ examples can be cited when (4.10) does not hold.

Assume $\mu = \frac{1}{2}$ and let us prove that for any λ for every S^1, S^2 ,

$$(4.19) \quad \frac{\sum_{i \in S^1} \sum_{j \in S^2} p_{ji}}{\sum_{i \in S^1} \sum_{j \in S^2} p_{ij}} = [4\lambda(1 - \lambda)]^{m_1 - m_2},$$

where m_k is the number of sign changes in states of the class S^k , $k = 1, 2$.

Relation (4.19) immediately implies the validity of relation (4.10) and formula (4.18).

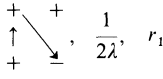
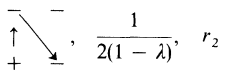
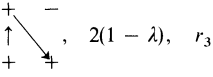
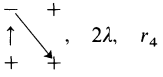
3. To prove that (4.19) holds it is sufficient to prove that for every pair of states $s_i \in S^1$, $s_j \in S^2$ a state $s_{j'}$ $\in S^2$ exists such that

$$(4.20) \quad \frac{p_{ji}}{p_{ij'}} = [4\lambda(1 - \lambda)]^{m_1 - m_2},$$

so that different s_i correspond to different $s_{j'}$.

We shall show that the state $s_{j'}$ satisfying condition (4.20) can be obtained from the state s_j by a shift of one position to the left. (This idea originated with P. I. Wasilewski.) Consider the factors in the transition probabilities p_{ji} and $p_{ij'}$, related to the point having the number l in the states s_i and s_j . The expression for the probability p_{ji} includes a factor, corresponding to the transition of the l th point of the configuration s_j into the l th point of the configuration s_i , while the probability $p_{ij'}$ includes the factor, corresponding to the transition of the l th point in the configuration s_i into the $(l + 1)$ st point of the configuration s_j .

Note that if in the configurations s_i and s_j the type of the position of the l th point is the same, then the corresponding factors in the relation $p_{ji}/p_{ij'}$ are cancelled. There remains to consider the cases when the l th point in the configurations s_i and s_j enters into positions of different types. The contribution to the relation $p_{ji}/p_{ij'}$ of the factors related to transitions of the l th point depends also on whether the signs of the l th point in the configurations s_i and s_j coincide (that is, whether the sign change in one of these configurations leads to the signs of the succeeding points being different) or not. The four cases possible are given by the following table.

	Transition from identical to opposite signs in the configurations	Transition from opposite to identical signs in the configurations
Sign change in the configuration	 $\frac{1}{2\lambda}, r_1$	 $\frac{1}{2(1-\lambda)}, r_2$
Sign change in the configuration	 $2(1-\lambda), r_3$	 $2\lambda, r_4$

In each compartment of the table, an example (one of the sign variants) of the situation under consideration is given, and also the contribution of transition probabilities of the given point into the relation $p_{ji}/p_{ij'}$ and the number r_v of such situations in the sequence of the points in the configurations s_i, s_j , $v = 1, 2, 3, 4$. Thus

$$(4.21) \quad \frac{p_{ji}}{p_{ij'}} = (2\lambda)^{r_4 - r_1} [2(1 - \lambda)]^{r_3 - r_2}.$$

According to the hypothesis,

$$(4.22) \quad \begin{aligned} r_1 + r_2 &= 2m_1, \\ r_3 + r_4 &= 2m_1. \end{aligned}$$

As the sequences of the points in the configurations s_i, s_j form a cycle, the number of transitions from identical to opposite signs equals the number of transitions from opposite to identical signs, that is,

$$(4.23) \quad r_1 + r_3 = r_2 + r_4.$$

Taking into account (4.22) and (4.23), we obtain (4.19), which proves the theorem.

Note that in the original problem of simultaneous ballot [10], the condition of strong detailed balance is not satisfied. It would be all the more interesting to obtain its solution.

From the point of view of the detailed balance concept, the meaning of relations (1.17), (3.2), appearing in all the problems discussed above, becomes clear. Indeed, the probability of sign change by a point indexed $P(\xi)$ is just the probability of transition from the original state into the state differing from it by the sign of this point. Thus the right side of relation (3.2) represents the logarithm of the relation of the transition probabilities, while the left side is the logarithm of the relation of final probabilities, given in the form of Gibbs distribution.

Naturally there arises the question of what is to be done in cases lacking a detailed balance. At present it appears to be possible to set apart the case of "weak" violation of detailed balance. A detailed description of this case will be given later.

Acknowledgments. In conclusion I express my deep gratitude to Professor Pyatetskii-Shapiro, A. Leontovich, N. Vasil'ev, A. Toom, O. Stavskaya, P. I. Wasilewski and L. Levitin for numerous and helpful discussions of this work.

REFERENCES

- [1] M. L. TSETLIN, *Finite automata and the simulation of the simplest forms of behavior*, Russian Math. Surveys, 18 (1963), no. 4.
- [2] I. M. GEL'FAND, I. I. PYATETSKII-SHAPIRO AND M. L. TSETLIN, *On certain classes of games and robot games*, Soviet Math. Dokl., 4 (1963), no. 4.
- [3] V. YU. KRYLOV AND M. L. TSETLIN, *On games for automata*, Automat. Remote Control, 24 (1963), no. 7.
- [4] V. A. VOLKONSKII, *Asymptotic properties of the behavior of the elementary automata*, Problemy Peredachi Informatsii, 1 (1965), no. 2.
- [5] L. ROSENOER, *Random logical nets*, Automat. Remote Control, 30 (1969), no. 5, 6.
- [6] O. STAVSKAYA AND I. I. PYATETSKII-SHAPIRO, *On the uniform networks from the active elements*, Problemy Kibernet. (1968), no. 20.
- [7] A. TOOM, *A family of uniform nets of formal neurons*, Soviet Math. Dokl., 9 (1968).
- [8] N. VASIL'EV, *Limit behavior of one random medium*, Problemy Peredachi Informatsii, 5 (1969), no. 4.
- [9] L. VASERSHTEIN AND A. LEONTOVICH, *Invariant measures of certain Markov operators describing a homogeneous random medium*, Ibid., 6 (1970), no. 1.
- [10] N. VASIL'EV, M. PETROVSKAYA AND I. PYATESKII-SHAPIRO, *Model of voting with random error*, Automat. Remote Control, 30 (1969), no. 10.
- [11] M. KAC, *Probability and Related Topics in Physical Sciences*, Proc. Summer Seminar, Boulder, Colo., 1957, vol. 1, Interscience, New York, 1959.
- [12] ———, *Probability in classical physics*, Proc. Symp. Appl. Math., vol. 7, McGraw-Hill, New York, 1957, pp. 73–85.
- [13] J. KEMENY AND J. SNELL, *Finite Markov Chains*, The University Series in Undergraduate Mathematics, Van Nostrand, Princeton, N.J., 1960.
- [14] YU. SCHMOOKLER, *Thermodynamic model of adaptation*, Soviet Physics Dokl., 13 (1969), no. 10.
- [15] ———, *The ballot problem with random error*, Soviet Math. Dokl., 12 (1971), no. 1.
- [16] L. D. LANDAU AND E. M. LIFSHITZ, *Statistical Physics*, Addison-Wesley, Reading, Mass., 1958.
- [17] YU. K. BELYAEV, YU. GROMAK AND V. MALYSHEV, *Invariant random Boolean fields*, Matem. Zametki, 6 (1969), no. 5.
- [18] YU. SCHMOOKLER, *One-dimensional and two-dimensional Ghur game*, Automat. Remote Control, 31 (1970), no. 10.

AN APPLICATION OF ERROR BOUNDS FOR CONVEX PROGRAMMING IN A LINEAR SPACE*

STEPHEN M. ROBINSON†

Abstract. We extend to general spaces an error-bounding technique for convex programming, used by Fiacco and McCormick. This technique is then applied to prove a version of Hoffman's theorem for K -convex inequalities.

1. Introduction. In [1, Thm. 29], Fiacco and McCormick showed how one could obtain an effective upper bound for the optimal objective value of a convex program in \mathbb{R}^n satisfying the Slater constraint qualification [7], [4, § 5.4.3], starting from an infeasible point. The bound is obtained by constructing a feasible linear combination of the given point and the point at which the Slater condition is satisfied. It has the desirable property that as the infeasible point approaches the feasible region, the feasible point thus constructed becomes arbitrarily close to it.

The purpose of this note is to point out that the same technique, slightly modified, can be used for cone-convex programming problems in very general spaces, and to show how it can be applied to obtain a generalization to convex constraints of a very useful theorem of Hoffman on linear inequalities.

2. Constructing a feasible point. Let X be a linear space and let Y be a normed linear space containing a nonempty closed convex cone K . Suppose that θ is a real-valued convex function on a convex subset $X_0 \subset X$, and that g is a K -convex function from X_0 into Y : that is, a function such that for each $x_1, x_2 \in X_0$ and each $\lambda \in [0, 1]$ we have

$$\lambda g(x_1) + (1 - \lambda)g(x_2) \in g[\lambda x_1 + (1 - \lambda)x_2] + K.$$

One may then consider the generalized convex programming problem

$$(1) \quad \text{minimize} \quad \{\theta(x) | 0 \in g(x) + K, x \in X_0\}.$$

If g satisfies a generalized Slater condition (that is, if there exist a point x_s and a real number $\delta > 0$ such that

$$(2) \quad \delta S \subset g(x_s) + K,$$

where S is the closed unit ball in Y), then an optimality condition of Kuhn-Tucker type may be derived for (1) [5, Thms. 3.1, 3.3].

We shall suppose that we are given a point $x \in X_0$ for which, in general, $0 \notin g(x) + K$. We wish to construct from x a point feasible for (1) and in some sense "better" than x_s (which is, of course, also feasible). Accordingly, we set $\lambda := \rho/(\rho + \delta)$, where

$$\rho := d(0, g(x) + K) := \inf \{\|y\| \mid y \in g(x) + K\},$$

* Received by the editors March 22, 1973, and in revised form September 12, 1973.

† Mathematics Research Center, The University of Wisconsin, Madison, Wisconsin 53706. This work was sponsored by the United States Army under Contract DA-31-124-ARO-D-462.

and where δ is as in (2). Clearly $0 \leq \lambda < 1$. We now assert that the point $x_\lambda := (1 - \lambda)x + \lambda x_s$ is feasible for (1). Indeed, by the K -convexity of g we have

$$g(x_\lambda) + K \supset (1 - \lambda)[g(x) + K] + \lambda[g(x_s) + K].$$

Also, for arbitrary $\varepsilon > 0$ there is a point $w \in g(x) + K$ with $\|w\| \leq \rho + \varepsilon$, and so the point $\delta(\rho + \varepsilon)^{-1}(-w)$ lies in δS , which in turn is contained in $g(x_s) + K$. Hence

$$g(x_\lambda) + K \ni (1 - \lambda)w + \lambda\delta(\rho + \varepsilon)^{-1}(-w) = \varepsilon(\rho + \varepsilon)^{-1}(1 - \lambda)w.$$

Therefore $g(x_\lambda) + K$ contains points of arbitrarily small norm, and since K is closed we must actually have $0 \in g(x_\lambda) + K$. Finally, $x_\lambda \in X_0$ by convexity, which completes the proof.

The feasibility of x_λ then implies, as noted in [1], that $(1 - \lambda)\theta(x) + \lambda\theta(x_s)$ is an upper bound for the optimal objective value in (1) (so, of course, is $\theta(x_\lambda)$). We remark that the distance of x_λ from x (measured along the line segment from x to x_s , since we have not metrized X) is nearly proportional to $d(0, g(x) + K)$ for small values of the latter quantity. Hence, if x almost satisfies the constraints, then x_λ will be nearly equal to x . We apply this observation in the next section.

3. Hoffman's theorem and convex inequalities. In 1952, Hoffman [2] proved a fundamental result on linear inequalities, one form of which can be stated as follows. Let A be any $m \times n$ matrix. Then there exists a constant α , depending on A , with the following property: for any $b \in \mathbb{R}^m$ such that the set

$$F := \{x \in \mathbb{R}^n | Ax \leq b\}$$

is nonempty, and for any $x \in \mathbb{R}^n$, we have

$$d(x, F) \leq \alpha \| (Ax - b)^+ \|,$$

where the norms involved are arbitrary, though fixed (the constant α depends on them). Here we have used c^+ to denote the vector obtained by replacing each negative component of a vector c by zero. Thus, Hoffman's theorem may be used to estimate the distance from a point to the solution set of a system of linear inequalities in terms of the extent to which the observed point fails to satisfy the system, or, to use a term frequently employed in numerical linear algebra, in terms of the size of the residuals. We indicate here how the result of § 2 can be applied to prove a result of this type for K -convex inequalities. We observe first that for any l_p -norm and any $x \in \mathbb{R}^n$,

$$\|x^+\|_p = d(0, x + \mathbb{R}_+^n),$$

where \mathbb{R}_+^n is the nonnegative orthant. Hence the quantity $d(0, g(x) + K)$ is an appropriate generalization to arbitrary cones K of the expression $\|g(x)^+\|$ in the simpler case in which $K = \mathbb{R}_+^n$ and an l_p -norm is used.

Now assume the notation of § 2, with the additional assumption that X is a normed linear space. Then we have $x - x_\lambda = \rho(\rho + \delta)^{-1}(x - x_s)$, so letting $Q := \{x | 0 \in g(x) + K, x \in X_0\}$, the feasible set for (1), we have

$$(3) \quad d(x, Q) \leq \|x - x_\lambda\| = \rho(\rho + \delta)^{-1} \|x - x_s\|.$$

Since

$$\rho(\rho + \delta)^{-1} \leq \delta^{-1} \rho = \delta^{-1} d(0, g(x) + K),$$

the expression (3) is a bound of the same type as that in Hoffman's theorem, except for the factor $\|x - x_s\|$. Examples can be constructed to show that this factor cannot be removed in the general (nonlinear) case. However, since x_s is usually known, both δ and $\|x - x_s\|$ can be obtained or estimated; hence if $d(0, g(x) + K)$ can be estimated, then the bound in (3) will actually be computable.

If the set Q is bounded, we can modify (3) to remove the factor $\|x - x_s\|$. From (3), we obtain

$$\delta \|x - x_\lambda\| = \rho(\|x - x_s\| - \|x - x_\lambda\|),$$

so

$$\begin{aligned} d(x, Q) &\leq \delta^{-1} \rho(\|x - x_s\| - \|x - x_\lambda\|) \\ (4) \quad &\leq \delta^{-1} \rho \|x_\lambda - x_s\| \\ &\leq \delta^{-1} \Delta d(0, g(x) + K), \end{aligned}$$

where we have denoted by Δ the diameter of Q , and have used the triangle inequality.

We observe that (3) shows that if x lies in a bounded set, then K -convex constraints are correct in the sense of [3, p. 4]: that is, if $d(0, g(x_k) + K)$ converges to zero for a bounded sequence $\{x_k\} \subset X_0$, then $d(x_k, Q) \xrightarrow{k \rightarrow \infty} 0$. If Q itself is bounded, then we appeal to (4) and the boundedness of $\{x_k\}$ is not (explicitly) required. In fact, (3) and (4) imply more than this: namely, that the R -order [6, § 9.2] associated with the convergence of $d(x_k, Q)$ to zero is at least as large as that of the sequence $\{d(0, g(x_k) + K)\}$. Roughly speaking, $\{x_k\}$ approaches Q at least as fast as $d(0, g(x_k) + K)$ approaches zero.

REFERENCES

- [1] A. V. Fiacco and G. P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley, New York, 1968.
- [2] A. J. Hoffman, *On approximate solutions of systems of linear inequalities*, J. Res. Nat. Bur. Standards, 49 (1952), pp. 263–265.
- [3] E. S. Levitin and B. T. Polyak, *Constrained minimization methods*, Ž. Vyčisl. Mat. i Mat. Fiz., 6 (1966), no. 5, pp. 787–823 = U.S.S.R. Comput. Math. and Math. Phys., 6 (1966), no. 5, pp. 1–50.
- [4] O. L. Mangasarian, *Nonlinear Programming*, McGraw-Hill, New York, 1969.
- [5] L. W. Neustadt, *Sufficiency conditions and a duality theory for mathematical programming problems in arbitrary linear spaces*, Nonlinear Programming, J. B. Rosen, O. L. Mangasarian and K. Ritter, eds., Academic Press, New York, 1970, pp. 323–348.
- [6] J. M. Ortega and W. C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*, Academic Press, New York, 1970.
- [7] M. Slater, *Lagrange multipliers revisited: a contribution to nonlinear programming*, Cowles Commission Discussion Paper No. Math. 403, 1950; also RAND Report RM-676, The RAND Corporation, Santa Monica, Calif., 1951.

OPTIMAL CONTROL OF PARABOLIC SYSTEMS WITH BOUNDARY CONDITIONS INVOLVING TIME DELAYS*

P. K. C. WANG†

Abstract. Various optimal control problems with quadratic cost functionals for parabolic systems with Neumann boundary conditions involving time delays are considered. Necessary and sufficient conditions which the optimal controls must satisfy are derived. Estimates and a sufficient condition for the boundedness of solutions are obtained for systems with specified forms of feedback controls. Similar problems for systems with more complex boundary conditions involving time delays are discussed briefly.

1. Introduction. Various mathematical problems associated with the control of parabolic systems with time delays appearing in the equations have been studied recently [1]–[7]. Here, we consider parabolic systems in which the time delays appear in the boundary conditions. Such systems arise physically in the control of diffusion processes in which time-delayed feedback signals are introduced at the boundary of a system's spatial domain. For example, in the area of plasma control, it is of interest to confine a plasma in a given bounded spatial domain Ω by introducing a finite electric potential barrier or a "magnetic mirror" surrounding Ω . For a collision-dominated plasma, its particle density is describable by a parabolic equation. Due to particle inertia and finiteness of electric potential or the magnetic-mirror field strength, the particle reflection at the domain boundary is not instantaneous. Consequently, the particle flux at the boundary of Ω at any time depends on the flux of particles which escaped earlier and reflected back into Ω at a later time. This leads to boundary conditions involving time delays.

In this paper, we consider various optimal control problems for parabolic systems with Neumann boundary conditions involving time delays. Necessary and sufficient conditions which the optimal controls must satisfy are derived. Estimates and a sufficient condition for the boundedness of solutions are obtained for systems with specified forms of feedback controls.

2. Preliminaries. Let Ω be a bounded open set in R^n with an infinitely differentiable boundary Γ and I denote a given finite time interval $]0, T[$. We consider a system described by the following parabolic equation:

$$(1) \quad \frac{\partial y}{\partial t} + A(t)y = f \quad \text{in } Q = \Omega \times I,$$

where

$$(2) \quad A(t)y = - \sum_{i,j=1}^n \frac{\partial}{\partial x_i} \left(a_{ij}(x, t) \frac{\partial y}{\partial x_j} \right).$$

* Received by the editors June 25, 1973, and in revised form September 26, 1973.

† Department of System Science, School of Engineering and Applied Science, University of California, Los Angeles, California 90024. This work was supported by the U.S. Air Force Office of Scientific Research under Grants AFOSR-72-2303 and AFOSR-74-2662.

The coefficients a_{ij} are real C^∞ -functions defined on \bar{Q} (closure of Q), and they satisfy the ellipticity condition

$$(3) \quad \sum_{i,j=1}^n a_{ij}(x, t) \xi_i \xi_j \geq \alpha \sum_{i=1}^n \xi_i^2$$

for all $(x, t) \in \bar{Q}$, $\xi = (\xi_1, \dots, \xi_n) \in R^n$ and some $\alpha > 0$. The function f corresponds to either a distributed control or a specified function defined on Q .

Let Γ^1 and Γ^2 be given disjoint subsets of Γ such that $\Gamma = \Gamma^1 \cup \Gamma^2$. Let $\Sigma = \Gamma \times I$; $\Sigma^i = \Gamma^i \times I$, $i = 1, 2$. We consider the following Neumann boundary condition involving a time delay:

$$(4) \quad \frac{\partial y}{\partial v_A}(x, t) = \sum_{i,j=1}^n a_{ij}(x, t) \cos(\eta, x_i) \frac{\partial y}{\partial x_j}(x, t) = q(x, t) \quad \text{on } \Sigma,$$

where

$$(5) \quad q(x, t) = \Phi(x) \{b(x, t)y(w(x), t - \tau) + u(x, t)\},$$

where Φ is a given C^∞ -function defined on Γ with compact support in Γ^1 ; $\cos(\eta, x_i)$ is the i th directional cosine of the outward normal η at a point $x \in \Gamma$; b is a given real C^∞ -function defined on Σ ; u represents either a boundary control or a given function defined in Σ ; the time delay τ is a specified positive number; w is a continuously differentiable bijection of Γ onto Γ such that $w(x) = x$ if $x \in \Gamma^2$ and $w(x) \in \Gamma^2$ if $x \in \Gamma^1$, and whose Jacobian does not vanish on Γ . Note that the simplest boundary condition $(\partial y / \partial v_A)(x, t) = y(x, t - \tau) + u(x, t)$ on Σ is included in the above description; i.e., we let Γ^2 be empty, $\Phi(x) \equiv 1$ and $w(x) = x$ on Γ .

The initial data for (1) are given by

$$(6) \quad \begin{aligned} y(x, 0) &= y_0(x), \quad x \in \Omega; \\ y(x, t') &= \phi_0(x, t'), \quad (x, t') \in \Gamma \times [-\tau, 0], \end{aligned}$$

where y_0 and ϕ_0 are specified functions. Note that only $\tilde{\phi}_0$, the restriction of ϕ_0 to $\Gamma^2 \times [-\tau, 0]$, is of importance here. In the sequel, we shall first establish sufficient conditions for the existence of a unique solution of the mixed initial-boundary value problem (1), (4) and (6). Then, various optimal control problems will be considered.

3. Existence and uniqueness of solutions. For simplicity, let the final time $T = K\tau$, where K is a given positive integer. We introduce the following notations: $I_j =](j-1)\tau, j\tau[$, $Q_j = \Omega \times I_j$, $\Sigma_j = \Gamma \times I_j$ and $\Sigma_j^i = \Gamma^i \times I_j$ for $i = 1, 2$ and $j = 0, 1, \dots, K$. Let $H^r(\Omega)$, $r \geq 0$, denote the Sobolev space of order r on Ω . For any pair of real numbers $r, s \geq 0$, the Sobolev space $H^{r,s}(Q)$ is defined by

$$(7) \quad H^{r,s}(Q) = H^0(I; H^r(\Omega)) \cap H^s(I; H^0(\Omega)), \quad Q = \Omega \times I,$$

which is a Hilbert space normed by

$$(8) \quad \left\{ \int_0^T \|y(t)\|_{H^r(\Omega)}^2 dt + \|y\|_{H^s(0,T;H^0(\Omega))}^2 \right\}^{1/2},$$

where $H^s(I; X)$ denotes the Sobolev space of order s of functions defined on I and taking values in X .

The existence of a unique solution for the mixed initial-boundary value problem (1), (4) and (6) on Q can be established by first solving the problem on Q_1 . Then, the existence of a unique solution on Q_2 is established by using the solution on Q_1 to generate the initial data at $t = \tau$. This advancing process is repeated for Q_3, Q_4, \dots until the final cylinder set Q_K is reached. Hereafter, the solution on Q_j will be denoted by $y_j, j = 1, \dots, K$. Now, the existence of a unique solution y_j can be established by making use of the results of Lions and Magenes [9] specialized to the case of a second order parabolic equation (1) with boundary condition (4) and initial data at $t = (j - 1)\tau$:

$$(9) \quad y_j(x, (j - 1)\tau) = y_{j-1}(x, (j - 1)\tau), \quad x \in \Omega.$$

Note that once the solution y_{j-1} defined on Q_{j-1} is determined, the right-hand side of (4) becomes a known function given by

$$(10) \quad q_j(x, t) = \Phi(x) \{b(x, t)y_{j-1}(w(x), t - \tau) + u(x, t)\}.$$

For optimal control problems, it is of importance to consider the cases where the control f or u belongs to $L^2(Q)$ or $L^2(\Sigma)$ respectively.

Case 1. $f \in L^2(Q)$. In this case, we make use of the following result.

LEMMA 1 (Lions and Magenes [9, p. 33]). *Let $f, y_{j-1}(\cdot, (j - 1)\tau)$ and q_j be given with*

$$(11a) \quad f \in L^2(Q),$$

$$(11b) \quad y_{j-1}(\cdot, (j - 1)\tau) \in H^1(\Omega),$$

$$(11c) \quad q_j \in H^{1/2, 1/4}(\Sigma_j).$$

Then, there exists a unique solution $y_j \in H^{2,1}(Q_j)$ for the mixed initial-boundary value problem (1), (4) and (9).

Evidently for $j = 1$, $y_{j-1}(w(x), t - \tau) = \phi_0(w(x), t - \tau)$ and conditions (11b) and (11c) can be satisfied if we assume that $u \in H^{1/2, 1/4}(\Sigma)$ and $\tilde{\phi}_0 \in H^{1/2, 1/4}(\Sigma_0^2)$. Thus, under the above assumptions, the existence of a unique solution $y_1 \in H^{2,1}(Q_1)$ follows from Lemma 1 if $y_0 \in H^1(\Omega)$.

In order to extend the result to Q_2 , it is sufficient to verify that

$$(11b') \quad y_2(\cdot, \tau) = y_1(\cdot, \tau) \in H^1(\Omega),$$

$$(11c') \quad q_2 \in H^{1/2, 1/4}(\Sigma_2).$$

To verify (11b'), we note that $y_1 \in H^{2,1}(Q_1)$ implies $y_1 \in L^2(I_1, H^2(\Omega))$ and $dy_1/dt \in L^2(I_1; H^0(\Omega))$. Then, it follows from a result of Lions and Magenes [8, p. 19, Thm. 3.1] that the mapping $t \rightarrow y_1(\cdot, t)$ is continuous from $[0, \tau] \rightarrow H^1(\Omega)$. Consequently, condition (11b') is satisfied. From the trace theorem [9, p. 9, Thm. 2.1], $y_1 \in H^{2,1}(Q_1)$ implies that $y_1 \rightarrow y_1|_{\Sigma_1}$ is a continuous linear mapping of $H^{2,1}(Q_1) \rightarrow H^{1/2, 1/4}(\Sigma_1)$. From (10) with $j = 2$ and the assumption that Φ and b are C^∞ -functions, if $u \in H^{1/2, 1/4}(\Sigma)$, then $q_2 \in H^{1/2, 1/4}(\Sigma_2)$. By Lemma 1, there exists a unique solution $y_2 \in H^{2,1}(Q_2)$. Evidently, we can now extend the result to any $Q_j, 2 < j \leq K$. The foregoing result is now summarized.

THEOREM 1. *Let y_0, ϕ_0, u and f be given with $y_0 \in H^1(\Omega)$, $\tilde{\phi}_0 \in H^{1/2, 1/4}(\Sigma_0^2)$, $u \in H^{1/2, 1/4}(\Sigma)$ and $f \in L^2(Q)$. Then, there exists a unique solution $y \in H^{2,1}(Q)$ for*

problem (1), (4) and (6). Moreover, $y(\cdot, j\tau) \in H^1(\Omega)$ for $j = 1, \dots, K$.

Case 2. $u \in L^2(\Sigma)$. For this case, the following result is applicable.

LEMMA 2 (Lions and Magenes [9, p. 81]). Let

$$(12a) \quad f \in (H^{1/2, 1/4}(Q))', \quad u \in L^2(\Sigma),$$

$$(12b) \quad q_j \in L^2(\Sigma_j),$$

$$(12c) \quad y_{j-1}(\cdot, (j-1)\tau) \in H^{1/2}(\Omega),$$

where X' denotes the dual of X . Then, there exists a unique solution $y_j \in H^{3/2, 3/4}(Q_j)$ for the mixed initial-boundary value problem (1), (4) and (9) defined on Q_j .

For $j = 1$, conditions (12b) and (12c) can be satisfied if we assume that $y_0 \in H^{1/2}(\Omega)$ and $\tilde{\phi}_0 \in L^2(\Sigma_0^2)$. These assumptions are sufficient to ensure the existence of a unique solution $y_1 \in H^{3/2, 3/4}(Q_1)$. To extend the result to Q_j , $1 < j \leq K$, by means of Lemma 2, it is sufficient to verify that $y_1(\cdot, \tau) \in H^{1/2}(\Omega)$ and $y_1|_{\Sigma_1} \in L^2(\Sigma_1)$. Since $y_1 \in H^{3/2, 3/4}(Q_1)$ implies that the mapping $t \rightarrow y_1(\cdot, t)$ is continuous from $[0, \tau] \rightarrow H^{3/4}(\Omega)$ [8, p. 19, Thm. 3.1], hence $y_1(\cdot, \tau) \in H^{3/4}(\Omega) \subset H^{1/2}(\Omega)$. Again, from the trace theorem [9, p. 9, Thm. 2.1], $y_1 \in H^{3/2, 3/4}(Q_1)$ implies that $y_1 \rightarrow y_1|_{\Sigma_1}$ is a continuous linear mapping of $H^{3/2, 3/4}(Q_1) \rightarrow H^{1, 1/2}(\Sigma_1)$. Thus, $y_1|_{\Sigma_1} \in L^2(\Sigma_1)$. We now summarize the foregoing result.

THEOREM 2. Let y_0, ϕ_0, u and f be given with $y_0 \in H^{1/2}(\Omega)$, $\tilde{\phi}_0 \in L^2(\Sigma_0^2)$, $u \in L^2(\Sigma)$ and $f \in (H^{1/2, 1/4}(Q))'$. Then, there exists a unique solution $y \in H^{3/2, 3/4}(Q)$ for problem (1), (4) and (6) with $y(\cdot, j\tau) \in H^{1/2}(\Omega)$ for $j = 1, \dots, K$.

Remark. We may also consider the mixed initial-boundary value problem for (1) with the following Dirichlet boundary condition involving a time delay:

$$(13) \quad y(x, t) = q(x, t) \quad \text{on } \Sigma,$$

where q is given by (5). Here, if $f \in L^2(Q)$, the result of Lions and Magenes [9, p. 33] ensures the existence of a unique solution $y_1 \in H^{2, 1}(Q_1)$ provided that $y_0 \in H^1(\Omega)$, $q_1 \in H^{3/2, 3/4}(\Sigma)$ and the following compatibility condition for the initial data is satisfied:

$$(14) \quad y_0(x) = q_1(x, 0) \quad \text{on } \Gamma.$$

In order to extend the result to Q_2 using the same approach, it is necessary to impose the compatibility condition $y_1(x, \tau) = q_2(x, \tau)$ on Γ , which is unnatural for the problem. The same situation arises for the case where $u \in L^2(\Sigma)$.

4. Optimal control. We shall consider various optimal control problems for system (1), (4) and (6) in which the control corresponds to either f or u belonging to a specified closed convex set $\mathcal{U}_Q \subseteq L^2(Q)$ or $\mathcal{U}_\Sigma \subseteq L^2(\Sigma)$ respectively.

Problem 1a. Let y_0, ϕ_0 and u be given functions satisfying the hypothesis of Theorem 1. Let $y(x, t; f)$ denote the solution of (1), (4) and (6) at (x, t) corresponding to a given control $f \in \mathcal{U}_Q$. The problem is to find an $f^\circ \in \mathcal{U}_Q$ such that $J(f^\circ) \leq J(f)$ for all $f \in \mathcal{U}_Q$, where J is a cost functional given by

$$(15) \quad \begin{aligned} J(f) = & \lambda_1 \int_Q |y(x, t; f) - y_d|^2 dx dt + \lambda_2 \int_\Omega |y(x, T; f) - y_{dT}|^2 dx \\ & + \lambda_3 \int_Q (Nf)f dx dt, \end{aligned}$$

where $\lambda_i \geq 0$ and $\lambda_1 + \lambda_2 + \lambda_3 > 0$; y_d and y_{dT} are given in $L^2(Q)$ and $L^2(\Omega)$ respectively; and N is a positive linear operator on $L^2(Q)$ into $L^2(Q)$. We note from Theorem 1 that for any $f \in \mathcal{U}_Q$, $J(f)$ is well-defined since $y(f) \in H^{2,1}(Q) \subset L^2(Q)$ and $y(\cdot, T; f) \in H^1(\Omega) \subset L^2(\Omega)$.

For the above problem, it is well known [3] that for $\lambda_3 > 0$, a unique optimal control f° exists; moreover, f° is characterized by

$$(16) \quad J'(f^\circ) \cdot (f - f^\circ) \geq 0 \quad \text{for all } f \in \mathcal{U}_Q,$$

or in explicit form,

$$(17) \quad \begin{aligned} & \lambda_1 \int_Q (y(f^\circ) - y_d)(y(f) - y(f^\circ)) dx dt \\ & + \lambda_2 \int_\Omega (y(x, T; f^\circ) - y_{dT})(y(x, T; f) - y(x, T; f^\circ)) dx \\ & + \lambda_3 \int_Q (Nf^\circ)(f - f^\circ) dx dt \geq 0 \end{aligned} \quad \text{for all } f \in \mathcal{U}_Q.$$

To simplify (17), we introduce the following adjoint equation. For every $f \in \mathcal{U}_Q$, we define $p = p(f) = p(x, t; f)$ as the solution of

$$(18) \quad -\frac{\partial p(f)}{\partial t} + A^*(t)p(f) = \lambda_1(y(f) - y_d) \quad \text{in } Q,$$

with terminal condition

$$(19) \quad p(x, T; f) = \lambda_2(y(x, T; f) - y_{dT}), \quad x \in \Omega,$$

and boundary conditions

$$(20) \quad \frac{\partial p(f)}{\partial v_{A^*}}(x, t) = 0 \quad \text{for } (x, t) \in ([\Gamma - w(\text{supp}(\Phi))] \times I) \cup (w(\text{supp}(\Phi)) \times I_K),$$

$$(21) \quad \frac{\partial p(f)}{\partial v_{A^*}}(x, t) = \Phi(w^{-1}(x))b(w^{-1}(x), t + \tau)|J_w(x)|p(w^{-1}(x), t + \tau; f)$$

$$\text{for } (x, t) \in w(\text{supp}(\Phi)) \times]0, T - \tau[,$$

where J_w denotes the Jacobian of w ; $w(\text{supp}(\Phi))$ is the image of the support of Φ under the mapping w and

$$(22) \quad \begin{aligned} \frac{\partial p(f)}{\partial v_{A^*}}(x, t) &= \sum_{i,j=1}^n a_{ji}(x, t) \cos(\eta, x_i) \frac{\partial p(f)}{\partial x_j}(x, t), \\ A^*(t)p &= - \sum_{i,j=1}^n \frac{\partial}{\partial x_j} \left(a_{ij}(x, t) \frac{\partial p}{\partial x_i} \right). \end{aligned}$$

We observe that for given y_d, y_{dT} and f , problem (18)–(21) can be solved backward in time starting from $t = T$ by first obtaining the solution $p = p_K$ on Q_K with terminal condition (19) and boundary condition

$$(23) \quad \frac{\partial p_K(f)}{\partial v_{A^*}}(x, t) = 0 \quad \text{for } (x, t) \in \Sigma_K.$$

Having found p_K , we may proceed to solve the problem on Q_{K-1} backward in time starting with terminal data at $t = (K - 1)\tau$:

$$(24) \quad p_{K-1}(x, (K - 1)\tau) = p_K(x, (K - 1)\tau), \quad x \in \Omega,$$

and with boundary conditions

$$(25) \quad \frac{\partial p_{K-1}(f)}{\partial v_{A^*}}(x, t) = 0 \quad \text{for } (x, t) \in [\Gamma - w(\text{supp}(\Phi))] \times I_{K-1},$$

$$(26) \quad \frac{\partial p_{K-1}(f)}{\partial v_{A^*}}(x, t) = \Phi(w^{-1}(x))b(w^{-1}(x), t + \tau)|J_w(x)|p_K(w^{-1}(x), t + \tau; f) \\ \text{for } (x, t) \in w(\text{supp}(\Phi)) \times I_{K-1}.$$

Note that the right-hand side of (26) is completely determined once p_K is known. This backward process is repeated until the initial cylinder set Q_1 is reached.

To establish the existence of a unique solution p_K on Q_K , we recall from Theorem 1 that for any $f \in L^2(Q)$, there exists a unique solution $y(f) \in H^{2,1}(Q)$ and $y(\cdot, T; f) \in H^1(\Omega)$. Thus, if $y_d \in L^2(Q)$ and $y_{dT} \in H^1(\Omega)$, the right-hand sides of (18) and (19) are functions in $L^2(Q)$ and $H^1(\Omega)$ respectively. Now, we may apply Theorem 1 (with obvious change of variables) to problem (18)–(21) (with reversed sense of time, i.e., $t' = T - t$) to establish the existence of a unique solution $p_K(f) \in H^{2,1}(Q_K)$ with $p_K(\cdot, (K - 1)\tau) \in H^1(\Omega)$. The result can be extended to Q_{K-1} and for any Q_j , $1 \leq j \leq K - 1$, in the same way, since the right-hand side of (26) is in $H^{1/2, 1/4}(\Sigma_{K-1})$ (by the trace theorem). Thus, we have the following result.

LEMMA 3. *Let the hypothesis of Theorem 1 be satisfied. Then, for given $y_d \in L^2(Q)$, $y_{dT} \in H^1(\Omega)$ and any $f \in L^2(Q)$, there exists a unique solution $p(f) \in H^{2,1}(Q)$ to the problem (18)–(21).*

Now, in view of Lemma 3, we can proceed to simplify (17) using the adjoint equation. Setting $f = f^\circ$ in (18)–(21), multiplying both sides of (18) by $(y(f) - y(f^\circ))$, then integrating over Q and using (19) lead to the identity

$$(27) \quad \begin{aligned} & \lambda_1 \int_Q (y(f^\circ) - y_d)(y(f) - y(f^\circ)) \, dx \, dt \\ &= \int_Q \left(-\frac{\partial p(f^\circ)}{\partial t} + A^*(t)p(f^\circ) \right) (y(f) - y(f^\circ)) \, dx \, dt \\ &= -\lambda_2 \int_\Omega (y(x, T; f^\circ) - y_{dT})(y(x, T; f) - y(x, T; f^\circ)) \, dx \\ & \quad + \int_Q p(f^\circ) \frac{\partial}{\partial t} (y(f) - y(f^\circ)) \, dx \, dt + \int_Q A^*(t)p(f^\circ)(y(f) - y(f^\circ)) \, dx \, dt. \end{aligned}$$

The last term in (27), in view of Green's formula, can be rewritten as

$$(28) \quad \begin{aligned} & \int_Q A^*(t)p(f^\circ)(y(f) - y(f^\circ)) \, dx \, dt \\ &= \int_Q p(f^\circ)[A(t)(y(f) - y(f^\circ))] \, dx \, dt + \int_0^T \int_\Gamma p(f^\circ) \left\{ \frac{\partial y(f)}{\partial v_A} - \frac{\partial y(f^\circ)}{\partial v_A} \right\} d\Gamma \, dt \\ & \quad - \int_0^T \int_\Gamma \frac{\partial p(f^\circ)}{\partial v_{A^*}} (y(f) - y(f^\circ)) \, d\Gamma \, dt. \end{aligned}$$

From boundary condition (4), the second integral in the right-hand side of (28) becomes

$$\begin{aligned}
 & \int_0^T \int_{\Gamma} p(f^\circ) \left\{ \frac{\partial y(f)}{\partial v_A} - \frac{\partial y(f^\circ)}{\partial v_A} \right\} d\Gamma dt \\
 (29) \quad &= \int_0^T \int_{\text{supp}(\Phi)} p(x, t; f^\circ) \Phi(x) b(x, t) [y(w(x), t - \tau; f) \\
 &\quad - y(w(x), t - \tau; f^\circ)] d\Gamma dt \\
 &= \int_{-\tau}^{T-\tau} \int_{\text{supp}(\Phi)} p(x, t' + \tau; f^\circ) \Phi(x) b(x, t' + \tau) [y(w(x), t'; f) \\
 &\quad - y(w(x), t'; f^\circ)] dx dt'.
 \end{aligned}$$

Since w is a continuously differentiable bijection of Γ onto Γ with non-vanishing Jacobian J_w on Γ , (29) can be rewritten as

$$\begin{aligned}
 (30) \quad & \int_{-\tau}^{T-\tau} \int_{w(\text{supp}(\Phi))} p(w^{-1}(x), t' + \tau; f^\circ) \Phi(w^{-1}(x)) b(w^{-1}(x), t' + \tau) \\
 & \cdot [y(x, t'; f) - y(x, t'; f^\circ)] |J_w(x)| d\Gamma dt'.
 \end{aligned}$$

Now, the last term in (28) can be split into two integrals:

$$\begin{aligned}
 (31) \quad & \int_0^T \int_{\Gamma} \left\{ \frac{\partial p(f^\circ)}{\partial v_{A^*}} (y(f) - y(f^\circ)) \right\} d\Gamma dt \\
 &= \int_0^T \int_{w(\text{supp}(\Phi))} \{\dots\} d\Gamma dt + \int_0^T \int_{\Gamma - w(\text{supp}(\Phi))} \{\dots\} d\Gamma dt.
 \end{aligned}$$

Substituting (30) and (31) into (28), and then the result into (27) gives

$$\begin{aligned}
 & \lambda_1 \int_Q (y(f^\circ) - y_d)(y(f) - y(f^\circ)) dx dt \\
 & \quad + \lambda_2 \int_{\Omega} (y(x, T; f^\circ) - y_{dT})(y(x, T; f) - y(x, T; f^\circ)) dx \\
 &= \int_Q p(f^\circ) \left(\frac{\partial}{\partial t} + A(t) \right) (y(f) - y(f^\circ)) dx dt \\
 (32) \quad & - \int_{T-\tau}^T \int_{w(\text{supp}(\Phi))} \frac{\partial p(f^\circ)}{\partial v_{A^*}} (y(f) - y(f^\circ)) d\Gamma dt \\
 & - \int_0^T \int_{\Gamma - w(\text{supp}(\Phi))} \frac{\partial p(f^\circ)}{\partial v_{A^*}} (y(f) - y(f^\circ)) d\Gamma dt \\
 & - \int_0^{T-\tau} \int_{w(\text{supp}(\Phi))} \left(\frac{\partial p}{\partial v_{A^*}}(x, t; f^\circ) - \Phi(w^{-1}(x)) |J_w(x)| b(w^{-1}(x), t + \tau) \right. \\
 & \quad \left. \cdot p(w^{-1}(x), t + \tau; f^\circ) \right) (y(x, t; f) - y(x, t; f^\circ)) d\Gamma dt \\
 & + \int_{-\tau}^0 \int_{w(\text{supp}(\Phi))} \Phi(w^{-1}(x)) b(w^{-1}(x), t + \tau) |J_w(x)| \\
 & \quad \cdot p(w^{-1}(x), t + \tau; f^\circ) (y(x, t; f) - y(x, t; f^\circ)) d\Gamma dt.
 \end{aligned}$$

The last term in (32) vanishes, since $y(x, t; f) = y(x, t; f^\circ) = \phi_0(x, t)$ for $(x, t) \in \Gamma \times [-\tau, 0[$. Finally, from (1) and boundary conditions (20)–(21), (32) reduces to

$$(33) \quad \begin{aligned} & \lambda_1 \int_Q (y(f^\circ) - y_d)(y(f) - y(f^\circ)) dx dt \\ & + \lambda_2 \int_\Omega (y(x, T; f^\circ) - y_{dT})(y(x, T; f) - y(x, T; f^\circ)) dx = \int_Q p(f^\circ)(f - f^\circ) dx dt. \end{aligned}$$

Thus, (17) simplifies to

$$(34) \quad \int_Q (p(f^\circ) + \lambda_3 N f^\circ)(f - f^\circ) dx dt \geq 0 \quad \text{for all } f \in \mathcal{U}_Q.$$

We now summarize the foregoing result.

THEOREM 3. *For Problem 1a with cost functional given by (15) with $y_d \in L^2(Q)$, $y_{dT} \in H^1(\Omega)$ and $\lambda_3 > 0$, there exists a unique optimal control f° which is determined by the solution of (1), (18) with boundary conditions (4), (20)–(21), initial condition (6) and terminal condition (19) (all with $f = f^\circ$). Moreover, f° satisfies (34).*

For the special case where $\mathcal{U}_Q = L^2(Q)$, (34) is satisfied when

$$(35) \quad f^\circ = -\lambda_3^{-1} N^{-1} p(f^\circ).$$

In particular, if N is the identity operator on $L^2(Q)$, then, in view of Lemma 2, we have $f^\circ \in H^{2,1}(Q)$. In order to obtain the optimal control (35) in feedback form, we follow Lions' approach [3] by first considering the following set of equations with $s \in I$:

$$(36) \quad \begin{aligned} & \frac{\partial y}{\partial t} + A(t)y + \lambda_3^{-1} N^{-1} p = 0, \quad (x, t) \in \Omega \times]s, T[, \\ & -\frac{\partial p}{\partial t} + A^*(t)p - \lambda_1 y = -\lambda_1 y_d, \quad (x, t) \in \Omega \times]s, T[, \end{aligned}$$

with boundary conditions

$$(37) \quad \frac{\partial y}{\partial v_A}(x, t) = \begin{cases} \Phi(x) \{b(x, t)y(w(x), t - \tau) + u(x, t)\} & \text{if } t - \tau \geq s, \\ \Phi(x) \{b(x, t)\phi_s(w(x), t - \tau) + u(x, t)\} & \text{if } t - \tau < s, \end{cases}$$

$$(x, t) \in \Omega \times]s, T[,$$

$$(38) \quad \frac{\partial p}{\partial v_{A^*}}(x, t) = \begin{cases} 0 & \text{if } (x, t) \in \hat{\Sigma}_s \triangleq (\Gamma - w(\text{supp}(\Phi)) \\ & \quad \times]s, T[) \cup (w(\text{supp}(\Phi)) \times I_K), \\ \Phi(w^{-1}(x))b(w^{-1}(x), t + \tau)|J_w(x)|p(w^{-1}(x), t + \tau) & \\ & \text{if } (x, t) \in w(\text{supp}(\Phi)) \times]s, T - \tau[, \end{cases}$$

and with initial and terminal conditions

$$(39) \quad \begin{aligned} & y(x, s) = y_s(x), \quad x \in \Omega, \\ & p(x, T) = \lambda_2(y(x, T) - y_{dT}), \quad x \in \Omega, \end{aligned}$$

where y_s is given in $H^1(\Omega)$ and ϕ_s is a given function defined in $\Gamma \times [s - \tau, s[$, whose restriction $\tilde{\phi}_s$ to $\Gamma^2 \times [s - \tau, s[$ is in $H^{1/2,1/4}(\Gamma^2 \times [s - \tau, s[)$. Note that (36)–(39) provide the solution to the optimal control problem associated with (1) for $t \in]s, T[$, $\mathcal{U}_Q = L^2(Q)$ and with a cost functional given by

$$(40) \quad \begin{aligned} J_s(f) = & \lambda_1 \int_s^T \int_{\Omega} |y(x, t; f) - y_d|^2 dx dt + \lambda_2 \int_{\Omega} |y(x, T; f) - y_{dT}|^2 dx \\ & + \lambda_3 \int_s^T \int_{\Omega} (Nf)f dx dt. \end{aligned}$$

The problem with $\lambda_3 > 0$ has a unique optimal control in the form of (35). Consequently, (36)–(39) has a unique solution $\{y, p\}$ also. In fact, for any given pair $(y_s, \tilde{\phi}_s) \in H^1(\Omega) \times H^{1/2,1/4}(\Gamma^2 \times [s - \tau, s[)$, the solution $y, p \in H^{2,1}(\Omega \times]s, T[)$. Moreover, the following property can be readily established.

PROPOSITION. *Let $\{y, p\}$ be the solution of (36)–(39) with $s = 0$. Define σ_s , the system “state” at time s , by the pair $(y(\cdot, s), \tilde{\phi}_s)$, where*

$$(41) \quad \tilde{\phi}_s(\cdot, t') = \begin{cases} \tilde{\phi}_0(\cdot, t') & \text{for } t' \in \hat{I}_s = [-\tau, 0[\cap [s - \tau, s[, \\ y(\cdot, t')|_{\Gamma^2} & \text{for } t' \in [s - \tau, s[-\hat{I}_s. \end{cases}$$

Then, for all pairs $s \leq t$ in I ,

$$(42) \quad p(\cdot, t) = P(t, s)\sigma_s + r_s(\cdot, t),$$

where $P(t, s)$ and $r_s(\cdot, t)$ are defined as follows:

(i) We solve the equations

$$(43) \quad \begin{aligned} \frac{\partial \beta}{\partial t} + A(t)\beta + \lambda_3^{-1}N^{-1}\gamma &= 0, & (x, t) \in \Omega \times]s, T[, \\ -\frac{\partial \gamma}{\partial t} + A^*(t)\gamma - \lambda_1\beta &= 0, & (x, t) \in \Omega \times]s, T[, \end{aligned}$$

with boundary conditions

$$(44) \quad \frac{\partial \beta}{\partial v_A}(x, t) = \begin{cases} \Phi(x)b(x, t)\beta(w(x), t - \tau) & \text{if } t - \tau \geq s, \\ \Phi(x)b(x, t)\tilde{\phi}_s(w(x), t - \tau) & \text{if } t - \tau < s, \end{cases}$$

$$(x, t) \in \Omega \times]s, T[,$$

$$(45) \quad \frac{\partial \gamma}{\partial v_{A^*}}(x, t) = \begin{cases} 0 & \text{if } (x, t) \in \hat{\Sigma}_s, \\ \Phi(w^{-1}(x))b(w^{-1}(x), t + \tau)|J_w(x)|\gamma(w^{-1}(x), t + \tau) & \text{if } (x, t) \in w(\text{supp}(\Phi)) \times]s, T - \tau[, \end{cases}$$

and initial and terminal conditions

$$(46) \quad \begin{aligned} \beta(x, s) &= y(x, s), & x \in \Omega, \\ \gamma(x, T) &= \lambda_2\beta(x, T), & x \in \Omega; \end{aligned}$$

then

$$(47) \quad P(t, s)\sigma_s = \gamma(\cdot, t).$$

(ii) We solve the equations

$$(48) \quad \begin{aligned} \frac{\partial \eta}{\partial t} + A(t)\eta + \lambda_3^{-1}N^{-1}\xi &= 0, \quad (x, t) \in \Omega \times]s, T[, \\ -\frac{\partial \xi}{\partial t} + A^*(t)\xi - \lambda_1\eta &= -\lambda_1 y_d, \quad (x, t) \in \Omega \times]s, T[, \end{aligned}$$

with boundary conditions

$$(49) \quad \frac{\partial \eta}{\partial v_A}(x, t) = \begin{cases} \Phi(x)\{b(x, t)\eta(w(x), t - \tau) + u(x, t)\} & \text{if } t - \tau \geq s, \\ \Phi(x)u(x, t), & \text{if } t - \tau < s, \end{cases}$$

$$(x, t) \in \Gamma \times]s, T[,$$

$$(50) \quad \frac{\partial \xi}{\partial v_{A^*}}(x, t) = \begin{cases} 0 & \text{if } (x, t) \in \hat{\Sigma}_s, \\ \Phi(w^{-1}(x))b(w^{-1}(x), t + \tau)J_w(x)\xi(w^{-1}(x), t + \tau) & \text{if } (x, t) \in w(\text{supp } (\Phi)) \times]s, T - \tau[, \end{cases}$$

and with initial and terminal conditions

$$(51) \quad \begin{aligned} \eta(x, s) &= 0, & x \in \Omega, \\ \xi(x, T) &= \lambda_2(\eta(x, T) - y_{dT}), & x \in \Omega; \end{aligned}$$

then

$$(52) \quad r_s(x, t) = \xi(x, t).$$

Now, the optimal feedback control can be obtained by setting $s = t$ in (42) and substituting the result into (35):

$$(53) \quad f^\circ(\cdot, t) = -\lambda_3^{-1}N^{-1}(P(t, t)\sigma_t + r_t(\cdot, t)), \quad t \in I.$$

By making use of Schwartz's kernel theorem [10], it can be verified that the optimal feedback control (53) (with N being the identity operator on $L^2(Q)$) can be represented in the form

$$(54) \quad \begin{aligned} f^\circ(x, t) = & -\lambda_3^{-1} \left\{ \int_{\Omega} K_0(x, x', t)y(x', t) dx' \right. \\ & \left. + \int_{t-\tau}^t \int_{\Gamma^2} K_1(x, x', t, t')\phi_t(x', t') d\Gamma^2 dt' + r_t(x, t) \right\}, \end{aligned}$$

where $\{K_0, K_1\}$ is the kernel of $P(t, t)$.

Problem 1b. Let y_0 , ϕ_0 and f be given functions satisfying the hypothesis of Theorem 2. Let $y(x, t; u)$ denote the solution of (1), (4) and (6) at (x, t) corresponding to a given control $u \in \mathcal{U}_\Sigma$. We wish to find a $u \in \mathcal{U}_\Sigma$ such that the following cost functional is minimized over \mathcal{U}_Σ :

$$(55) \quad \begin{aligned} J(u) = & \lambda_1 \int_Q |y(x, t; u) - y_d|^2 dx dt + \lambda_2 \int_{\Omega} |y(x, T; u) - y_{dT}|^2 dx \\ & + \lambda_3 \int_0^T \int_{\text{supp } (\Phi)} (\tilde{N}u)u d\Gamma dt, \end{aligned}$$

where the λ_i 's, y_d and y_{dT} are as in Problem 1a, and \tilde{N} is a positive linear operator on $L^2(\Sigma)$ into $L^2(\Sigma)$. If y_0 , ϕ_0 and f satisfy the conditions of Theorem 2, then for each $u \in \mathcal{U}_\Sigma$, $y(u) \in H^{3/2, 3/4}(Q)$ and $y(\cdot, T; u) \in H^{1/2}(\Omega)$. Hence $J(u)$ is defined. Similar to Problem 1a, this problem has a unique optimal control $u^\circ \in \mathcal{U}_\Sigma$ if $\lambda_3 > 0$. Also, u° can be characterized by

$$(56) \quad \begin{aligned} & \lambda_1 \int_Q (y(u^\circ) - y_d)(y(u) - y(u^\circ)) dx dt \\ & + \lambda_2 \int_\Omega (y(x, T; u^\circ) - y_{dT})(y(x, T; u) - y(x, T; u^\circ)) dx \\ & + \lambda_3 \int_0^T \int_{\text{supp}(\Phi)} (\tilde{N}u^\circ)(u - u^\circ) d\Gamma dt \geq 0 \quad \text{for all } u \in \mathcal{U}_\Sigma. \end{aligned}$$

The above inequality can be simplified by introducing an adjoint equation whose form is identical to (18)–(22). From Theorem 2, for any $u \in L^2(\Sigma)$, there exists a unique solution $y(u) \in H^{3/2, 3/4}(Q)$ with $y(\cdot, T; u) \in H^{1/2}(\Omega)$. If $y_d \in L^2(Q)$ and $y_{dT} \in H^{1/2}(\Omega)$, then the right-hand sides of (18) and (19) are in $L^2(Q)$ and $H^{1/2}(\Omega)$ respectively. Similar to Problem 1a, we can establish the existence of a unique solution $p(u) \in H^{3/2, 3/4}(Q)$ for (18)–(22). Moreover, (56) can be simplified as

$$(57) \quad \int_0^T \int_{\text{supp}(\Phi)} (p(u^\circ)\Phi(x) + \lambda_3 \tilde{N}u^\circ)(u - u^\circ) d\Gamma dt \geq 0 \quad \text{for all } u \in \mathcal{U}_\Sigma.$$

In Problems 1a and 1b, the cost functionals involve only deviations of the solutions from their desired values averaged over the interior of the spatial domain Ω . In what follows, we shall consider an optimal control problem in which the spatial averaging in the cost functional is taken over both Ω and its boundary Γ .

Problem 2. This problem is identical to Problem 1b except that the cost functional (55) is replaced by

$$(58) \quad \begin{aligned} \hat{J}(u) = & \lambda_1 \int_\Sigma |y(u)|_\Sigma - y_{\Sigma d}|^2 d\Gamma dt + \lambda_2 \int_\Omega |y(x, T; u) - y_{dT}|^2 dx \\ & + \lambda_3 \int_0^T \int_{\text{supp}(\Phi)} (\tilde{N}u)u d\Gamma dt, \end{aligned}$$

where $y_{\Sigma d}$ and y_{dT} are given in $L^2(\Sigma)$ and $L^2(\Omega)$ respectively. From Theorem 2 and the trace theorem [9, p. 9], for each $u \in \mathcal{U}_\Sigma$, there exists a unique solution $y(u) \in H^{3/2, 3/4}(Q)$ with $y(u)|_\Sigma \in L^2(\Sigma)$ and $y(\cdot, T; u) \in H^{1/2}(\Omega)$. Thus, $\hat{J}(u)$ is defined. Moreover, the optimal control u° is characterized by

$$(59) \quad \begin{aligned} & \lambda_1 \int_\Sigma (y(u^\circ)|_\Sigma - y_{\Sigma d})(y(u)|_\Sigma - y(u^\circ)|_\Sigma) d\Gamma dt \\ & + \lambda_2 \int_\Omega (y(x, T; u) - y_{dT})(y(x, T; u) - y(x, T; u^\circ)) dx \\ & + \lambda_3 \int_0^T \int_{\text{supp}(\Phi)} (\tilde{N}u^\circ)(u - u^\circ) d\Gamma dt \geq 0 \quad \text{for all } u \in \mathcal{U}_\Sigma. \end{aligned}$$

Here, we introduce the adjoint equation

$$(60) \quad -\frac{\partial p(u^\circ)}{\partial t} + A^*(t)p(u^\circ) = 0 \quad \text{in } Q,$$

with terminal condition

$$(61) \quad p(x, T; u^\circ) = \lambda_2(y(x, T; u^\circ) - y_{dT}), \quad x \in \Omega,$$

and boundary conditions

$$(62) \quad \frac{\partial p(u^\circ)}{\partial \nu_{A^*}}(x, t) = \begin{cases} \lambda_1(y(u^\circ)|_\Sigma(x, t) - y_{\Sigma d}) & \text{if } (x, t) \in \hat{\Sigma}_0, \\ \phi(w^{-1}(x))b(w^{-1}(x), t + \tau)|J_w(x)|p(w^{-1}(x), t + \tau; u^\circ) \\ \quad + \lambda_1(y(u^\circ)|_\Sigma(x, t) - y_{\Sigma d}) & \text{if } (x, t) \in w(\text{supp}(\Phi)) \times]0, T - \tau[, \end{cases}$$

where $\hat{\Sigma}_0$ is defined in (38).

Again, as in Problem 1b, Problem (59)–(62) has a unique solution $p(u^\circ) \in H^{3/2, 3/4}(Q)$. Also, condition (59) can be rewritten in the form of (57).

5. Estimates for solutions of feedback systems. The explicit expressions for the kernels of the optimal feedback controls such as (54) are generally quite complex. This motivates the consideration of suboptimal feedback controls with prescribed kernels having simple forms. In this section, estimates for the solutions of (1) with various specified forms of feedback controls introduced at the domain boundary Γ will be derived.

5.1. Feedback system with Neumann boundary condition. Consider system (1) with $f = 0$ and with Neumann boundary condition given by (4) and (5). The results of § 4 suggest a feedback control u which is a linear function $\mathcal{F}(t)$ of the state σ_t (defined by (41)) and has a representation of the form

$$(63) \quad \begin{aligned} u(x, t) &= (\mathcal{F}(t)\sigma_t)(x) = (\mathcal{F}_0(t)y(\cdot, t) + \mathcal{F}_1(t)\sigma_t)(x) \\ &= \int_{\Omega} F_0(x, x', t)y(x', t) dx' + \int_{t-\tau}^t \int_{\Gamma^2} F_1(x, x', t, t')\tilde{\phi}_t(x', t') d\Gamma^2 dt', \end{aligned}$$

where $\{F_0, F_1\}$ is the kernel of $\mathcal{F}(t)$ and $\tilde{\phi}_t$ is defined in (41). Formally, the solution of (1)–(6) (with feedback control u given by (63)) in any cylinder set $Q_j, j = 1, 2, \dots$, can be written as

$$(64) \quad \begin{aligned} y_j(\cdot, t) &= \mathcal{G}_{0j}(t, (j-1)\tau)y_{j-1}(\cdot, (j-1)\tau) + \mathcal{G}_{1j}(t, (j-1)\tau)q_j \\ &= \mathcal{G}_{0j}(t, (j-1)\tau)y_{j-1}(\cdot, (j-1)\tau) \\ &\quad + \mathcal{G}_{1j}(\tau, (j-1)\tau)[\Phi\{by_{j-1}|_{\Gamma}(w(\cdot), \cdot - \tau) + \mathcal{F}_0(t)y_{j-1}(\cdot, (j-1)\tau) \\ &\quad + \mathcal{F}_1(t)\tilde{\phi}_{j,t}\}], \quad t \in I_j, \quad j = 1, 2, \dots, \end{aligned}$$

where

$$(65) \quad \begin{aligned} \tilde{\phi}_{j,t}(\cdot, t') &= \begin{cases} y_j(\cdot, t')|_{\Gamma^2}, & t' \in [t - \tau, t] \cap I_j, \\ y_{j-1}(\cdot, t')|_{\Gamma^2}, & t' \in [t - \tau, t] \cap I_{j-1}; \end{cases} \\ y_0(\cdot, t) &= \phi_0(\cdot, t), \quad t \in I_0, \end{aligned}$$

and where \mathcal{G}_{0j} and \mathcal{G}_{1j} are linear integral operators. Evidently, once the \mathcal{G}_{ij} 's are known, estimates for the solutions of the feedback system can be obtained recursively. In what follows, explicit results will be obtained only for a simple one-dimensional case for illustrative purposes.

Let $\Omega =]0, 1[$ and let the system equation be given by

$$(66) \quad \frac{\partial y}{\partial t} = \frac{\partial^2 y}{\partial x^2}, \quad t > 0, \quad x \in \Omega,$$

with boundary conditions

$$(67) \quad \begin{aligned} \frac{\partial y}{\partial x}(0, t) &= q(t) \triangleq by(1, t - \tau) + u(t), \\ \frac{\partial y}{\partial x}(1, t) &= 0 \end{aligned}$$

and initial data

$$(68) \quad \begin{aligned} y(x, 0) &= y_0(x), \quad x \in \Omega, \\ y(1, t') &= \tilde{\phi}_0(t'), \quad t' \in [-\tau, 0]. \end{aligned}$$

The feedback control u is given by

$$(69) \quad u(t) = \int_0^1 F_0(x, t)y(x, t) dx + \int_{t-\tau}^t F_1(t, t')y(1, t') dt',$$

where F_0 and F_1 are specified continuous functions defined on $\Omega \times [0, \infty[$ and $[0, \infty[\times [-\tau, \infty[$ respectively. We assume that there exist two nonnegative numbers ρ_0 and ρ_1 such that

$$(70) \quad \int_0^1 |F_0(x, t)| dx \leq \rho_0 \quad \text{and} \quad \int_{t-\tau}^t |F_1(t, t')| dt' \leq \rho_1$$

for all $t \in [0, \infty[$. It can be shown that if $y_0 \in C^0(\bar{\Omega})$, then the solution to (66)–(68) with a given $q \in C^0([0, \infty[)$ has a representation of the form [11]

$$(71) \quad y(x, t) = \int_0^1 G_0(x, x', t)y_0(x') dx' + \int_0^t G_1(x, t - t')q(t') dt', \quad t > 0, \quad x \in]0, 1],$$

where

$$(72) \quad \begin{aligned} G_0(x, x', t) &= \sum_{m=-\infty}^{\infty} \{K(x, t; (2m + x'), 0) + K(-x, t; (2m + x'), 0)\}, \\ G_1(x, t - t') &= 2 \sum_{m=-\infty}^{\infty} K(x + 2m, t; 0, t'), \\ K(x, t; \xi, t') &= \frac{1}{2}(\pi(t - t'))^{-1/2} \exp\left(-\frac{(x - \xi)^2}{4(t - t')}\right). \end{aligned}$$

Using (67), (69) and (71), we have an integral equation for y_j in the form of (64):

$$(73) \quad \begin{aligned} y_j(x, t) = & \int_0^1 G_0(x, x', t - (j-1)\tau) y_{j-1}(x, (j-1)\tau) dx' \\ & + \int_{(j-1)\tau}^t dt' G_1(x, t-t') \left\{ b y_{j-1}(1, t' - \tau) + \int_0^1 F_0(x', t') y_j(x', t') dx' \right. \\ & \left. + \int_{t'-\tau}^{t'} F_1(t', t'') \tilde{\phi}_{j,t'}(t'') dt'' \right\}, \quad (x, t) \in]0, 1] \times I_j, \quad j = 1, 2, \dots, \end{aligned}$$

where

$$(74) \quad \begin{aligned} \tilde{\phi}_{j,t'}(t') &= \begin{cases} y_j(1, t'), & t' \in [t - \tau, t[\cap I_j, \\ y_{j-1}(1, t'), & t' \in [t - \tau, t[\cap I_{j-1}; \end{cases} \\ y_0(1, t) &= \tilde{\phi}_0(t'), \quad t' \in \bar{I}_0. \end{aligned}$$

The existence of a unique solution to (73) can be established by embedding the problem in the Banach space $C^0(Q_j)$ with the sup norm, and using the contraction mapping principle. Now, estimates for y_j will be derived. Let

$$(75) \quad \|y_j(\cdot, t)\| = \sup_{x \in \Omega} |y_j(x, t)|, \quad \|y_j\|_{Q_j} = \sup_{Q_j} |y_j(x, t)|.$$

From (70) and (73), we have

$$(76) \quad \begin{aligned} \|y_j(\cdot, t)\| \leq & \delta_{0j}(t) \|y_{j-1}(\cdot, (j-1)\tau)\| + \delta_{1j}(t) \left\{ |b| \sup_{(j-1)\tau < t' < t} |y_{j-1}(1, t' - \tau)| \right. \\ & \left. + \rho_0 \sup_{(j-1)\tau < t' < t} \|y_j(\cdot, t')\| + \rho_1 \sup_{(j-1)\tau < t' < t} \left\{ \sup_{t'-\tau < t'' < t'} |\phi_{j,t'}(t'')| \right\} \right\}, \end{aligned}$$

where

$$(77) \quad \delta_{0j}(t) = \sup_{x \in \Omega} \int_0^1 G_0(x, x', t - (j-1)\tau) dx',$$

$$(78) \quad \delta_{1j}(t) = \sup_{x \in \Omega} \int_{(j-1)\tau}^t G_1(x, t-t') dt'.$$

By straightforward calculations (see Appendix), it can be shown that

$$(79) \quad \delta_{0j}(t) \leq 2,$$

$$(80) \quad \delta_{1j}(t) \leq 2(t - (j-1)\tau)^{1/2} (\pi^{-1/2} + \sqrt{\tau})$$

for all $t \in \bar{I}_j$. The last term on the right-hand side of (76) satisfies

$$(81) \quad \begin{aligned} & \sup_{(j-1)\tau < t' < t} \left\{ \sup_{t'-\tau < t'' < t'} |\tilde{\phi}_{j,t'}(t'')| \right\} \\ & \leq \max \left\{ \sup_{t' \in I_{j-1}} |y_{j-1}(1, t')|, \sup_{(j-1)\tau < t' < t} |y_j(1, t')| \right\} \\ & \leq \sup_{t' \in I_{j-1}} |y_{j-1}(1, t')| + \sup_{(j-1)\tau < t' < t} |y_j(1, t')|, \quad t \in I_j. \end{aligned}$$

Thus, it follows from (76), (79)–(81) that

$$(82) \quad \|y_j\|_{Q_j} \leq 2\|y_{j-1}(\cdot, (j-1)\tau)\| + \hat{\delta}_{1j}\{(|b| + \rho_1) \sup_{t \in I_{j-1}} |y_{j-1}(1, t)| + \rho_0\|y_j\|_{Q_j} + \rho_1 \sup_{t \in I_j} |y_j(1, t)|\},$$

where

$$(83) \quad \sup_{t \in I_j} \delta_{1j}(t) \leq 2((\tau/\pi)^{1/2} + \tau) = \hat{\delta}_{1j}.$$

Using (73) and (A.5) in the Appendix, we can derive the following estimate for $y_j(1, t)$ in a similar manner:

$$(84) \quad \sup_{t \in I_j} |y_j(1, t)| \leq \frac{3}{2}\|y_{j-1}(\cdot, (j-1)\tau)\| + \hat{\delta}_{1j}\left\{(|b| + \rho_1) \sup_{t \in I_{j-1}} |y_{j-1}(1, t)| + \rho_0\|y_j\|_{Q_j} + \rho_1 \sup_{t \in I_j} |y_j(1, t)|\right\}.$$

Estimates (82) and (84) can be combined as:

$$(85) \quad \begin{bmatrix} 1 - \hat{\delta}_{1j}\rho_0 & -\hat{\delta}_{1j}\rho_1 \\ -\hat{\delta}_{1j}\rho_0 & 1 - \hat{\delta}_{1j}\rho_1 \end{bmatrix} \begin{bmatrix} \|y_j\|_{Q_j} \\ \sup_{t \in I_j} |y_j(1, t)| \end{bmatrix} \leq \begin{bmatrix} 2 & \hat{\delta}_{1j}(|b| + \rho_1) \\ 3/2 & \hat{\delta}_{1j}(|b| + \rho_1) \end{bmatrix} \begin{bmatrix} \|y_{j-1}(\cdot, (j-1)\tau)\| \\ \sup_{t \in I_{j-1}} |y_{j-1}(1, t)| \end{bmatrix}, \quad j = 1, 2, \dots.$$

Under the assumption

$$(86) \quad \hat{\delta}_{1j}(\rho_0 + \rho_1) < 1,$$

(85) implies that

$$(87) \quad \|y_j\|_{Q_j} \leq (1 - \hat{\delta}_{1j}(\rho_0 + \rho_1))^{-1} \left\{ (2 - \rho_1\hat{\delta}_{1j}/2) \|y_{j-1}(\cdot, (j-1)\tau)\| + \hat{\delta}_{1j}(|b| + \rho_1) \sup_{t \in I_{j-1}} |y_{j-1}(1, t)| \right\},$$

$$(88) \quad \sup_{t \in I_j} |y_j(1, t)| \leq (1 - \hat{\delta}_{1j}(\rho_0 + \rho_1))^{-1} \left\{ (5/2 - \rho_0\hat{\delta}_{1j}/2) \|y_{j-1}(\cdot, (j-1)\tau)\| + \hat{\delta}_{1j}(|b| + \rho_1) \sup_{t \in I_{j-1}} |y_{j-1}(1, t)| \right\}, \quad j = 1, 2, \dots.$$

Evidently, estimates for the solutions can be obtained recursively using (87) and (88) starting with $j = 1$.

5.2. Feedback system with Dirichlet boundary condition. Now, we consider system (1) with $f = 0$ and with Dirichlet boundary condition given by (13) and

(5). Again, let the feedback control be of the form (63) so that the boundary condition is given by

$$(89) \quad y(x, t) = \Phi(x) \left\{ b(x, t)y(w(x), t - \tau) + \int_{\Omega} F_0(x, x', t)y(x', t) dx' + \int_{t-\tau}^t \int_{\Gamma^2} F_1(x, x', t, t')\phi_t(x', t') d\Gamma^2 dt' \right\}, \quad (x, t) \in \Gamma \times [0, \infty[.$$

We shall establish a sufficient condition for the boundedness of solutions of this feedback system.

Let Q_{t_1} denote the cylinder set $\Omega \times]0, t_1[$ for $0 < t_1 < \infty$. Let $y_0 \in C^0(\bar{\Omega})$ and let y be a continuous function in \bar{Q}_{t_1} satisfying (1) in $Q_{t_1} \cup S_{t_1}$, where $S_t = \Omega \times \{t\}$. Then the weak maximum principle [12] asserts that the maximum of $|y|$ in \bar{Q}_{t_1} is attained on the part $\bar{S}_0 \cup (\Gamma \times]0, t_1])$ of the boundary of Q_{t_1} , i.e.,

$$(90) \quad \max_{\bar{Q}_{t_1}} |y(x, t)| \leq \max \left\{ \max_{\bar{\Omega}} |y_0(x)|, \sup_{\Gamma \times]0, t_1]} |y(x, t)| \right\}.$$

We shall show by contradiction that if

$$(91) \quad \eta(x, t) \triangleq |\Phi(x)| \left\{ |b(x, t)| + \int_{\Omega} |F_0(x, x', t)| dx' + \int_{t-\tau}^t \int_{\Gamma^2} |F_1(x, x', t, t')| d\Gamma^2 dt' \right\} < 1$$

for all $(x, t) \in \Gamma \times [0, \infty[$, then the maximum of $|y|$ is attained on \bar{S}_0 or

$$(92) \quad \max_{x \in \bar{\Omega}} |y(x, t)| \leq \max_{x \in \bar{\Omega}} |y_0(x)| \quad \text{for all } t \geq 0.$$

First, assume that there exists a point $(x^*, t^*) \in \Gamma \times]\tau, t_1]$ such that

$$(93) \quad |y(x^*, t^*)| = \max_{\bar{Q}_{t_1}} |y(x, t)|.$$

Then, in view of (89), we have

$$(94) \quad |y(x^*, t^*)| \leq |\Phi(x^*)| \left\{ |b(x^*, t^*)| |y(w(x^*), t^* - \tau)| + \int_{\Omega} |F_0(x^*, x, t^*)| |y(x, t^*)| dx + \int_{t^*-\tau}^{t^*} \int_{\Gamma^2} |F_1(x^*, x, t^*, t)| |y(x, t)| dx dt \right\} \leq \eta(x^*, t^*) |y(x^*, t^*)|,$$

where η is defined in (91). It is evident that if (91) is satisfied, then (94) leads to a contradiction. Consequently, such a point (x^*, t^*) does not exist under condition (91).

Now, assume that a point $(x^*, t^*) \in \Gamma \times [0, \tau]$ exists such that (93) holds. Then, from (89),

$$(95) \quad |y(x^*, t^*)| \leq |\Phi(x^*)| \left\{ |b(x^*, t^*)| |\phi_0(w(x^*), t^* - \tau)| + \left(\int_{\Omega} |F_0(x^*, x, t^*)| dx + \int_{t^*-\tau}^{t^*} \int_{\Gamma^2} |F_1(x^*, x, t^*, t)| dx dt \right) |y(x^*, t^*)| \right\}.$$

If we impose the condition that ϕ_0 is continuous on Σ_0 and satisfies

$$(96) \quad \sup_{\Sigma_0} |\phi_0(x, t)| \leq \max_{\bar{\Omega}} |y_0(x)|,$$

then

$$(97) \quad |\phi_0(w(x^*), t^* - \tau)| \leq \max_{\bar{\Omega}} |y_0(x)| \leq |y(x^*, t^*)|$$

and

$$(98) \quad |y(x^*, t^*)| \leq \eta(x^*, t^*)|y(x^*, t^*)|.$$

Again, under condition (91), (98) leads to a contradiction. Thus, such an (x^*, t^*) does not exist. Finally, we note that the foregoing results remain valid for arbitrarily large t_1 . Consequently, (92) holds. Thus, we have established the following theorem.

THEOREM 4. *Let y be a classical solution of (1) with $f = 0$ satisfying boundary condition (89) and initial data (6). Let $y_0 \in C^0(\bar{\Omega})$ and $\phi_0 \in C^0(\Sigma_0)$ such that (96) is satisfied. Then, under condition (91), $\max_{x \in \bar{\Omega}} |y(x, t)| \leq \max_{\bar{\Omega}} |y_0(x)|$ for all $t \geq 0$.*

Remark. The conditions in Theorem 4 represent restrictions on the initial data, the parameters in the boundary conditions and the kernels of the feedback control operators. They are independent of the parameters a_{ij} in the elliptic operator $A(t)$. Physically speaking, the result simply states that if the feedback gain and the maximum magnitude of the past boundary data are sufficiently small, then the solution will not grow with time.

6. Concluding remarks. In this paper, only parabolic systems with the simplest form of boundary condition involving a time delay have been considered. One may consider optimal control problems for parabolic system (1) with a variety of more complex boundary conditions involving time delays. A few examples are given below.

Example 1. Boundary condition involving multiple time delays. This is a generalization of the Neumann boundary condition (4):

$$(99) \quad \begin{aligned} \frac{\partial y}{\partial v_A}(x, t) + \gamma(x, t)y(x, t) = \Phi(x) & \left\{ \sum_{m=0}^M b_m(x, t)y(w(x), t - \tau_m) \right. \\ & \left. + \sum_{m=1}^M c_m(x, t) \frac{\partial y}{\partial v_A}(w(x), t - \tau_m) + u(x, t) \right\} \quad \text{on } \Sigma, \end{aligned}$$

where $0 \leq \tau_0 < \tau_1 < \cdots < \tau_M$; γ , b_m and c_m are specified coefficients. If γ , c_m , $m = 1, \dots, M$, are identically zero on Γ , the results of this paper can be extended to this case without difficulty.

Example 2. Boundary condition with indirect control. Here, the boundary condition is identical to (4) and (5) except that u is generated indirectly by

$$(100) \quad u(x, t) = g(x)^T z(t), \quad (x, t) \in \Sigma,$$

where $(\cdot)^T$ denotes transposition; g is a given mapping from Γ into R^r ; $z(t) \in R^r$ is the solution of the following system of linear ordinary differential-difference

equations:

$$(101) \quad \frac{dz(t)}{dt} = \sum_{m=0}^M F_m(t)z(t - \tau_m) + G(t)v(t)$$

with given initial data

$$(102) \quad z(t') = \omega(t') \in R^r, \quad t' \in [-\tau_m, 0],$$

where $0 \leq \tau_0 < \tau_1 < \dots < \tau_M$; $F_m(t)$ and $G(t)$ are given matrices. The control $v(\cdot) \in \mathcal{V}$ is a specified closed convex subset of $L^2(I; R^s)$. One may consider the optimal control problem involving the minimization of a convex cost functional of the solution y and the control v . Also, instead of (101), one may replace it by a functional differential equation. Finally, one may consider similar problems for hyperbolic systems with boundary conditions involving time delays. These problems will be discussed elsewhere.

Appendix. To establish estimate (79), we note that $G_0 \geq 0$ and

$$\begin{aligned} & \int_0^1 G_0(x, x', t - (j-1)\tau) dx' \\ (A.1) \quad & \leq \int_0^1 \left\{ K(x, t; x', (j-1)\tau) + K(-x, t; x', (j-1)\tau) \right. \\ & \quad \left. + 2 \int_0^\infty \{ K(x, t; (2\alpha + x'), (j-1)\tau) + K(-x, t; (2\alpha + x'), \right. \\ & \quad \left. (j-1)\tau) \} d\alpha \right\} dx' \\ & = \mathcal{J}_1(x, t) + \mathcal{J}_2(x, t), \end{aligned}$$

where

$$\begin{aligned} (A.2) \quad \mathcal{J}_1(x, t) &= \int_0^1 \{ K(x, t; x', (j-1)\tau) + K(-x, t; x', (j-1)\tau) \} dx' \\ &= \frac{1}{2} (\pi(t - (j-1)\tau))^{-1/2} \int_{x-1}^{x+1} \exp \left(-\frac{\hat{x}^2}{4(t - (j-1)\tau)} \right) d\hat{x}, \\ \mathcal{J}_2(x, t) &= 2 \int_0^1 \int_0^\infty \{ K(x, t; (2\alpha + x'), (j-1)\tau) + K(-x, t; (2\alpha + x'), (j-1)\tau) \} d\alpha dx' \\ (A.3) \quad &= \frac{1}{2} (\pi(t - (j-1)\tau))^{-1/2} \int_0^1 \left\{ \int_{x'+x}^\infty \exp \left(-\frac{\alpha^2}{4(t - (j-1)\tau)} \right) d\alpha \right. \\ & \quad \left. + \int_{x'-x}^\infty \exp \left(-\frac{\alpha^2}{4(t - (j-1)\tau)} \right) d\alpha \right\} dx' \\ &\leq (\pi(t - (j-1)\tau))^{-1/2} \int_0^\infty \exp \left(-\frac{\alpha^2}{4(t - (j-1)\tau)} \right) d\alpha = 1. \end{aligned}$$

Thus,

$$\begin{aligned}
 \delta_{0j}(t) &\leq 1 + \sup_{x \in \Omega} \mathcal{J}_1(x, t) \\
 (A.4) \quad &\leq 1 + \frac{1}{2}(\pi(t - (j - 1)\tau))^{-1/2} \int_{-1}^1 \exp\left(-\frac{\hat{x}^2}{4(t - (j - 1)\tau)}\right) d\hat{x} \\
 &\leq 1 + (\pi(t - (j - 1)\tau))^{-1/2} \int_0^\infty \exp\left(-\frac{\hat{x}^2}{4(t - (j - 1)\tau)}\right) d\hat{x} = 2.
 \end{aligned}$$

Note that for the case where $x = 1$,

$$(A.5) \quad \mathcal{J}_1(x, t) \leq \frac{1}{2}(\pi(t - (j - 1)\tau))^{-1/2} \int_0^\infty \exp\left(-\frac{\hat{x}^2}{4(t - (j - 1)\tau)}\right) d\hat{x} = \frac{1}{2}.$$

To derive estimate (80), consider

$$\begin{aligned}
 \int_{(j-1)\tau}^t G_1(x, t - t') dt' &= \int_{(j-1)\tau}^t 2 \sum_{m=-\infty}^{\infty} K(x + 2m, t; 0, t') dt' \\
 &\leq \int_{(j-1)\tau}^t (\pi(t - t'))^{-1/2} dt' \left\{ \sum_{m=-\infty}^{\infty} \exp\left(-\frac{(x + 2m)^2}{4\tau}\right) \right\} \\
 (A.6) \quad &\leq 2 \left(\frac{t - (j - 1)\tau}{\pi} \right)^{1/2} \left\{ \exp\left(-\frac{x^2}{4\tau}\right) + 2 \int_0^\infty \exp\left(-\frac{(x + 2\alpha)^2}{4\tau}\right) d\alpha \right\} \\
 &\leq 2 \left(\frac{t - (j - 1)\tau}{\pi} \right)^{1/2} \left\{ \exp\left(-\frac{x^2}{4\tau}\right) + 2 \int_0^\infty \exp\left(-\frac{\alpha^2}{\tau}\right) d\alpha \right\} \\
 &= 2 \left(\frac{t - (j - 1)\tau}{\pi} \right)^{1/2} \left\{ \exp\left(-\frac{x^2}{4\tau}\right) + \sqrt{\pi\tau} \right\}, \quad (x, t) \in \bar{Q}_j.
 \end{aligned}$$

The desired estimate (80) follows trivially from (A.6).

REFERENCES

- [1] M. ARTOLA, *Equation paraboliques à retardement*, C.R. Acad. Sci. Paris, 264 (1967), pp. 668–671.
- [2] C. BAIocchi, *Sulle equazioni differenziale astratte lineari del primo e del secondo ordine negli spazi di Hilbert*, Ann. Mat. Pura Appl., 76 (1967), pp. 233–304.
- [3] J. L. LIONS, *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*, Dunod, Paris, 1968.
- [4] M. N. OĞUZTÖRELİ, *A class of nonlinear integro-differential equations of parabolic type with a delayed argument*, Bul. Inst. Politehn. Iași, 14 (1968), pp. 43–49.
- [5] ———, *Un problema misto concenente un'equazione integro-differenziale di tipo parabolico co argomento ritardato*, Rend. Mat., 2 (1969), pp. 245–294.
- [6] P. K. C. WANG AND M. L. BANDY, *Stability of distributed-parameter systems with time delays*, J. Electronics and Control, 15 (1963), pp. 342–362.
- [7] P. K. C. WANG, *Asymptotic stability of a diffusion system with time-delays*, J. Appl. Mech. Ser. E, 30 (1963), pp. 500–504.
- [8] J. L. LIONS AND E. MAGENES, *Non-homogeneous Boundary Value Problems and Applications*, vol. 1, Springer-Verlag, New York, 1972.
- [9] ———, *Non-homogeneous Boundary Value Problems and Applications*, vol. 2, Springer-Verlag, New York, 1972.
- [10] L. SCHWARTZ, *Theorie des noyaux*, Proc. Internat. Congress of Mathematicians, vol. 1, Amer. Math. Soc., Providence, R.I., 1950, pp. 220–230.

- [11] P. HARTMAN AND A. WINTNER, *On the solutions of the equations of heat conduction*, Amer. J. Math., 72 (1950), pp. 367–395.
- [12] P. A. LADYŽENSKAJA, V. A. SOLONNIKOV AND N. N. URAL'CEVA, *Linear and Quasilinear Equations of Parabolic Type*, Translations of Math. Monographs No. 23, Amer. Math. Soc., Providence, R.I., 1968.

LYAPUNOV VECTOR MEASURES*

GREGORY KNOWLES†

Abstract. Consider the following infinite-dimensional extension of the linear control system discussed in Hermes and La Salle [9].

Let T be a set (time interval), \mathcal{S} a σ -algebra of subsets of T , X a quasi-complete locally convex topological vector space, and $\mathbf{m} = (m_i)$ a sequence of vector measures $m_i: \mathcal{S} \rightarrow X$, $i = 1, 2, \dots$. For each $i = 1, 2, \dots$ a bounded real-valued \mathcal{S} -measurable function f_i represents the effect of the i th control on the system. The total effect of all these controls is given by $\sum_{i=1}^{\infty} \int_T f_i dm_i$. If the controls are restricted so that $(f_i(t)) \in \mathcal{F}(t)$, $\mathcal{F}(t)$ a subset of the product of countably many copies of the real line, for each $t \in T$, then the set of all values of the series above is the attainable set of this system, denoted by $A_{\mathcal{F}}(\mathbf{m})$. The general bang-bang principle for this system is considered and conditions given for $A_{\mathcal{F}}(\mathbf{m}) = A_{\text{ex } \mathcal{F}}(\mathbf{m})$, where $\text{ex } \mathcal{F}(t)$ is the set of extreme points of $\mathcal{F}(t)$, $t \in T$. This problem is closely related to the study of Lyapunov vector measures, that is, vector measures whose restriction to any subset of \mathcal{S} has a compact, convex range.

The first part of the paper gives necessary and sufficient conditions for a vector measure to be Lyapunov. Then conditions on \mathbf{m} are derived for $A_{\mathcal{F}}(\mathbf{m}) = A_{\text{ex } \mathcal{F}}(\mathbf{m})$, in particular, these conditions are satisfied if each m_i is nonatomic, and X is finite-dimensional. If each m_i is just assumed nonatomic, we show in general that $A_{\mathcal{F}}(\mathbf{m})$ need only be equal to the weak closure of $A_{\text{ex } \mathcal{F}}(\mathbf{m})$.

Introduction. In considering a linear control system steered by a sequence of independently operating controls, we look at the following model. A sequence $\mathbf{m} = (m_i)$, $m_i: \mathcal{S} \rightarrow X$, \mathcal{S} a σ -algebra of subsets of a time interval T , X a quasi-complete locally convex topological vector space, is given. For each $i = 1, 2, \dots$, m_i represents the effect of the i th control f_i on the system. The total effect is given by

$$(1) \quad \sum_{i=1}^{\infty} \int_T f_i dm_i.$$

If at each time $t \in T$ the controls $(f_i(t))$ are constrained to belong to a set $\mathcal{F}(t)$ contained in the countable product of the real line, the set of all values of (1) is called the attainable set of this system and is denoted by $A_{\mathcal{F}}(\mathbf{m})$. We consider the bang-bang principle for this system, namely, under what conditions does $A_{\mathcal{F}}(\mathbf{m}) = A_{\text{ex } \mathcal{F}}(\mathbf{m})$, where $\text{ex } \mathcal{F}(t)$ is the set-valued function taking values in the extreme points of $\mathcal{F}(t)$, $t \in T$.

It has been found that this problem is closely related to the study of Lyapunov vector measures, that is, vector measures with compact, convex range. Accordingly, § 2 is devoted to Lyapunov vector measures. In Theorem 1 we give conditions for a vector measure to be Lyapunov, and in Theorem 2 we consider certain permanence properties of Lyapunov vector measures.

In Theorem 3 we give conditions on \mathbf{m} for $A_{\mathcal{F}}(\mathbf{m}) = A_{\text{ex } \mathcal{F}}(\mathbf{m})$ for certain set-valued functions \mathcal{F} , and prove that this result holds if each m_i is nonatomic and X is finite-dimensional.

* Received by the editors November 24, 1972, and in revised form June 25, 1973.

† School of Mathematical Sciences, The Flinders University of South Australia, Bedford Park, South Australia 5042.

Finally in Theorem 4 we show that if the assumptions on the m_i are weakened, in general it is only possible to recover $A_{\mathcal{F}}(\mathbf{m})$ from the weak closure of $A_{\text{ex } \mathcal{F}}(\mathbf{m})$. This result follows a method in [8] and extends Theorems 5.1, 5.2 in [6].

1. Definitions. Let X be a quasi-complete locally convex topological vector space (l.c.t.v.s.) with dual X' . For a set $H \subset X$, $\text{co}H$ and $\overline{\text{co}}H$ denote the convex hull and closed convex hull of H respectively. If H is convex, $\text{ex } H$ denotes the extreme points of H .

Let \mathcal{S} be a σ -algebra of subsets of some abstract set T , $\mathcal{M}(\mathcal{S})$ the set of all bounded \mathcal{S} -measurable functions on T . $\mathcal{M}^\infty(\mathcal{S})$ denotes the set of all sequences (f_i) of functions in $\mathcal{M}(\mathcal{S})$ with $\sup \{\|f_i\|_\infty : i = 1, 2, \dots\} < \infty$. We can also interpret elements of $\mathcal{M}^\infty(\mathcal{S})$ as functions on T with values in R^∞ , the countable product of the real line treated as a l.c.t.v.s. under the product topology.

A set-valued function \mathcal{F} defined on T whose values are subsets of R^∞ will be called bounded if there exists a compact set $H \subset R^\infty$ such that $\mathcal{F}(t) \subset H$ for every $t \in T$. For such a set-valued function \mathcal{F} , we put

$$\mathcal{M}_{\mathcal{F}}^\infty(\mathcal{S}) = \{f : f \in \mathcal{M}^\infty(\mathcal{S}), f(t) \in \mathcal{F}(t), t \in T\}.$$

Similarly, if K is a subset of the real numbers, define

$$\mathcal{M}_K(\mathcal{S}) = \{f : f \in \mathcal{M}(\mathcal{S}), f(t) \in K, t \in T\}.$$

Denote by CCR^∞ the family of compact, convex subsets of R^∞ . For a set-valued function $\mathcal{F} : T \rightarrow CCR^\infty$ we define the set-valued function $\text{ex } \mathcal{F}$ on T by

$$(\text{ex } \mathcal{F})(t) = \text{ex } (\mathcal{F}(t)), \quad t \in T.$$

Following Valadier [8], we call the set-valued function $\mathcal{F} : T \rightarrow CCR^\infty$ scalarly measurable if, for every $x' \in (R^\infty)'$, the function

$$t \rightarrow \sup \{\langle x', x \rangle : x \in \mathcal{F}(t)\},$$

$t \in T$, is \mathcal{S} -measurable.

By a vector measure on \mathcal{S} we mean a map $m : \mathcal{S} \rightarrow X$ which is countably additive. For $E \in \mathcal{S}$ we define

$$\mathcal{S}_E = \{F : F \subseteq E, F \in \mathcal{S}\},$$

the restriction of \mathcal{S} to E , and

$$\mathcal{R}(m, E) = \{m(F) : F \in \mathcal{S}_E\}.$$

To simplify we set $\mathcal{R}(m) = \mathcal{R}(m, T)$. The restriction of m to \mathcal{S}_E will again be denoted by m . A set $E \in \mathcal{S}$ is m -negligible if $\mathcal{R}(m, E) = 0$.

If $\mathcal{R}(m, E)$ is convex for every $E \in \mathcal{S}$, and $\mathcal{R}(m)$ is weakly compact, we call m a Lyapunov vector measure.

If X is a quasi-complete l.c.t.v.s., $m : \mathcal{S} \rightarrow X$ a vector measure, for each $x' \in X'$ we define a measure $\langle x', m \rangle : \mathcal{S} \rightarrow R$ by $\langle x', m \rangle(E) = \langle x', m(E) \rangle$, $E \in \mathcal{S}$. A real-valued \mathcal{S} -measurable function f will be called m -integrable if it is integrable with respect to every measure $\langle x', m \rangle$, $x' \in X'$, and if for every $E \in \mathcal{S}$, there is an

element $x_E \in X$ such that

$$(2) \quad \langle x', x_E \rangle = \int_E f d\langle x', m \rangle.$$

We write $x_E = \int_E f dm$, and $x_T = \int_T f dm = \int f dm$. The properties of this integral have been discussed in [4] and [6]. Any function $f \in \mathcal{M}(\mathcal{S})$ is m -integrable. In fact, f will be the uniform limit of \mathcal{S} -measurable simple functions. The existence of an $x_E \in X$ satisfying (2), $E \in \mathcal{S}$, now follows from the quasi-completeness of X .

Let A be an index set directed by the relation \leq . Let $\{E_\alpha\}_{\alpha \in A}$ be a net of sets in \mathcal{S} . The net $\{E_\alpha\}_{\alpha \in A}$ is said to be m -convergent to a set E (m -Cauchy) if, for every neighborhood \mathcal{U} of 0 in X , there exists $\alpha_{\mathcal{U}} \in A$ such that $\mathcal{R}(m, E_\alpha \Delta E) \subseteq \mathcal{U}$, for every $\alpha \in A$ with $\alpha_{\mathcal{U}} \leq \alpha$ (such that $\mathcal{R}(m, E_\alpha \Delta E_\beta) \subset \mathcal{U}$ for every $\alpha \in A$, $\beta \in B$ with $\alpha_{\mathcal{U}} \leq \alpha$, $\beta_{\mathcal{U}} \leq \beta$).

A vector measure is said to be closed if \mathcal{S} is m -complete, i.e., if every m -Cauchy net of sets in \mathcal{S} is m -convergent to a member of \mathcal{S} .

The basic properties of closed vector measures that will be needed in this note are proved in [3] and [6]. Namely, the indefinite integral of an m -integrable function with a closed vector measure is closed, m is closed if and only if $\overline{\text{co}} \mathcal{R}(m) = \{\int_T f dm : f \in \mathcal{M}_{[0,1]}(\mathcal{S})\}$, and any vector measure taking values in a metrizable space is closed.

The set of all finite σ -additive real-valued measures on \mathcal{S} is denoted by $\text{ca}(\mathcal{S})$, and by $|\lambda|$ we mean the variation of $\lambda \in \text{ca}(\mathcal{S})$. For a vector measure $m : \mathcal{S} \rightarrow X$ set $\Theta = \{\langle x', m \rangle : x' \in X'\}$.

Two functions f, g in $\mathcal{M}(\mathcal{S})$ will be called m -equivalent if $\int_T |f - g| d|\theta| = 0$ for all $\theta \in \Theta$. The equivalence class consisting of all functions m -equivalent to $f \in \mathcal{M}(\mathcal{S})$ is denoted by $[f]_m$. Set $\mathcal{M}(\mathcal{S}, m) = \{[f]_m : f \in \mathcal{M}(\mathcal{S})\}$, $M_K(\mathcal{S}, m) = \{[f]_m : f \in \mathcal{M}_K(\mathcal{S})\}$, K a set of real numbers.

If $\Gamma = \{\mu \in \text{ca}(\mathcal{S}) : \mu \ll \theta, \theta \in \Theta\}$, then the assignment $[f]_m \rightarrow \int f d\mu$, $f \in \mathcal{M}(\mathcal{S})$, gives an unambiguously defined functional on $M(\mathcal{S}, m)$ for each $\mu \in \Gamma$. We let $\sigma(m)$ denote the weakest topology that makes all these functionals continuous.

A sequence of closed vector measures $m_i : \mathcal{S} \rightarrow X$, $i = 1, 2, \dots$, will be called a control system if $\sum_{i=1}^\infty x_i$ is convergent for each $x_i \in \mathcal{R}(m_i)$, $i = 1, 2, \dots$. As $0 \in \mathcal{R}(m_i)$, $i = 1, 2, \dots$, this convergence is unconditional. If each m_i is Lyapunov, $i = 1, 2, \dots$, it will be called a Lyapunov control system. Set $\mathbf{m} = (m_i)$.

The next lemma will be used often in the text without reference. The proof is given in [5, Lemma 3].

LEMMA 1. If $\mathbf{m} = (m_i)$ is a control system, and $f = (f_i) \in \mathcal{M}^\infty(\mathcal{S})$, then the series $\sum_{i=1}^\infty \int_E f_i dm_i$ converges for each $E \in \mathcal{S}$.

For a bounded set-valued function \mathcal{F} on T whose values are subsets of R^∞ , we define

$$A_{\mathcal{F}}(\mathbf{m}) = \left\{ \sum_{i=1}^\infty \int_T f_i dm_i : f = (f_i) \in \mathcal{M}_{\mathcal{F}}^\infty(\mathcal{S}) \right\}.$$

If $\mathcal{F} : T \rightarrow CCR^\infty$ is a bounded, scalarly measurable set-valued function, then by [5, Thm. 1], $A_{\mathcal{F}}(\mathbf{m})$, the attainable set, is a convex, weakly compact subset of X . It is the aim of this note to study the relation between $A_{\mathcal{F}}(\mathbf{m})$ and $A_{\text{ex } \mathcal{F}}(\mathbf{m})$

for various control systems, in particular, to give a necessary and sufficient condition on \mathbf{m} for $A_{\mathcal{F}}(\mathbf{m}) = A_{\text{ex } \mathcal{F}}(\mathbf{m})$. It can be seen that if $A_{\mathcal{F}}(\mathbf{m}) = A_{\text{ex } \mathcal{F}}(\mathbf{m})$ for every scalarly measurable, bounded set-valued function $\mathcal{F}: T \rightarrow CCR^\infty$, then \mathbf{m} must be a Lyapunov control system. Accordingly we start this note with a study of Lyapunov vector measures.

2. Lyapunov vector measures. The first result gives a necessary and sufficient condition for a closed vector measure to be Lyapunov. It generalizes the concept of “thinness” introduced by Kingman and Robertson in [2].

THEOREM 1. *If $n: \mathcal{S} \rightarrow X$ is a closed vector measure, the following properties are equivalent:*

- (i) *For every non- n -negligible subset Z of \mathcal{S} there exists a bounded \mathcal{S} -measurable function v , with $v \neq 0$ on Z , $v = 0$ outside Z , and $\int_T v \, dn = 0$.*
- (ii) *For every $[u]_n \in M(\mathcal{S}, n)$, with $[u]_n \neq 0$, there exists a bounded measurable function v , with $[uv]_n \neq 0$, and $\int_T uv \, dn = 0$.*
- (iii) *n is a Lyapunov vector measure.*

Proof. Clearly (i) and (ii) are equivalent. In the proof we follow the scheme $\sim(i) \Rightarrow \sim(\text{iii})$, $(\text{ii}) \Rightarrow (\text{iii})$.

Suppose (i) is false. There exists a non- n -negligible set Z , and consequently a σ -algebra \mathcal{S}_Z such that the map $J: M(\mathcal{S}_Z, n) \rightarrow X$ defined by $J(f) = \int_Z f \, dn$ is one-to-one. Thus $\mathcal{R}(n, Z) = J(M_{\{0,1\}}(\mathcal{S}_Z, n)) \subsetneq J(M_{\{0,1\}}(\mathcal{S}_Z, n)) \subset \overline{\text{co}} \mathcal{R}(n, Z)$. Consequently n cannot be Lyapunov.

Consider $(\text{ii}) \Rightarrow (\text{iii})$. Set $H = M_{\{0,1\}}(\mathcal{S}, n)$. The map $J: M(\mathcal{S}, n) \rightarrow X$ defined by $J(f) = \int f \, dn$ is continuous if we give $M(\mathcal{S}, n)$ the $\sigma(n)$ topology and X its weak (i.e., $\sigma(X, X')$) topology. The proof follows if

$$H \subset M_{\{0,1\}}(\mathcal{S}, n) + J^{-1}(0).$$

Let $[u]_n \in H$. As n is closed, H is $\sigma(n)$ compact [3, Cor. 2 to Lemma 8], and so $H_0 = H \cap ([u]_n + J^{-1}(0))$ is $\sigma(n)$ compact and convex. Let $[u_0]_n \in \text{ex } H_0$. It only remains to show that $[u_0]_n \in M_{\{0,1\}}(\mathcal{S}, n)$. By [5, Lemma 5], $M_{\{0,1\}}(\mathcal{S}, n) = \text{ex } H$, and hence suppose $[u_0]_n \notin \text{ex } H$. Then there exists a $[v]_n \in M(\mathcal{S}, n)$, with $[v]_n \neq 0$, and $[u_0]_n \pm [v]_n \in H$. Applying (ii) there exists a function w , which can be chosen with $w(t) \in [-1, 1]$, $t \in T$, with $[vw]_n \neq 0$ and $\int_T vw \, dn = 0$. Now $[u_0]_n \pm [vw]_n \in H_0$, which contradicts our choice of u_0 .

We can now answer the question of the existence of nontrivial infinite-dimensional Lyapunov vector measures. The examples below give Lyapunov vector measures which cannot be written as the direct sum of finite-dimensional vector measures.

Example 1. Suppose $T = [0, 1] \times [0, 1]$, \mathcal{S} is the σ -algebra of Borel sets on T , and λ is Lebesgue measure on \mathcal{S} . Define functions $\varphi_n(s, t) = t^n$, $0 \leq s, t \leq 1$, $n = 1, 2, \dots$. Let $E \in \mathcal{S}$ be a set which is not λ -negligible. Then the L^2 -closed linear span of the functions $\{\varphi_n: n = 1, 2, \dots\}$ cannot cover $L^2(E)$. Consequently we can find a bounded measurable function f , nonzero on E with

$$\int_E f(s, t) \varphi_n(s, t) \, ds \, dt = 0 \quad \text{for } n = 1, 2, \dots$$

Define measures $\mu_n: \mathcal{S} \rightarrow R$ by

$$\mu_n(E) = \int_E \varphi_n(s, t) ds dt, \quad E \in \mathcal{S}.$$

Then the vector measure $m: \mathcal{S} \rightarrow c_0$ given by $m(E) = (\mu_1(E), \mu_2(E), \dots)$, $E \in \mathcal{S}$, will be Lyapunov by Theorem 1.

We can extend this construction to more general function spaces, as in the next example.

Example 2. Let $(T, \mathcal{S}, \lambda)$ be any finite measure space and let P be a set. For each $r \in P$, let a bounded measurable function $\varphi_r: T \rightarrow R$ be given. Assume that, for every set E in \mathcal{S} with $\lambda(E) \neq 0$, the system $\{\varphi_r: r \in P\}$ is incomplete, in the sense that there exists a bounded measurable function f , not vanishing λ -almost everywhere on E such that $\int_E f \varphi_r d\lambda = 0$ for every $r \in R$. Let X be a vector space of real-valued functions on P equipped with a quasi-complete l.c.t.v.s. topology making the application $m: \mathcal{S} \rightarrow X$, defined by $m(E)(r) = \int_E \varphi_r d\lambda$, $E \in \mathcal{S}$, $r \in P$, a vector measure. Then m is closed [6, Lemma 3.2] and, by Theorem 1, it is Lyapunov.

The following lemma which will be used later is a simple application of Theorem 1.

LEMMA 2. Suppose $m: \mathcal{S} \rightarrow X$ is a vector measure, $u \in \mathcal{M}(\mathcal{S})$ and $n: \mathcal{S} \rightarrow X$ is a vector measure defined by $n(E) = \int_E u dm$, $E \in \mathcal{S}$. If m is Lyapunov, then n is Lyapunov, and conversely, if u is bounded below by some positive constant, and if n is Lyapunov, then m is Lyapunov.

Proof. Suppose m is Lyapunov. Then

$$\left\{ \int f dm : f \in \mathcal{M}_{[0,1]}(\mathcal{S}) \right\} = \overline{\text{co}} \mathcal{R}(m),$$

and so m is closed by [3, Cor. 3 to Thm. 3]. Let $[w]_n \in M(\mathcal{S}, n)$. Then $[wu]_m \neq 0$, and so by Theorem 1 there exists an $h \in \mathcal{M}(\mathcal{S})$ with $[wuh]_m \neq 0$ and $\int_T wuh dm = 0$. Hence $[wh]_n \neq 0$, and $\int_T wh dn = 0$. Theorem 1 gives the result.

In the second part $1/u \in \mathcal{M}(\mathcal{S})$ and so it is n -integrable, and $\int_E (1/u) dn = m(E)$, $E \in \mathcal{S}$.

The conditions on the density function u in this lemma can be relaxed considerably. The next lemma will be used for this.

LEMMA 3. Suppose $m_i: \mathcal{S} \rightarrow X$, $i = 1, 2, \dots$, is a control system of Lyapunov vector measures for which there exists a sequence $(E_i)_{i=1}^\infty$ of pairwise disjoint subsets of \mathcal{S} with $\bigcup_{i=1}^\infty E_i = T$, and each m_i is zero outside E_i , $i = 1, 2, \dots$. Then the vector measure $m: \mathcal{S} \rightarrow X$ defined by $m(E) = \sum_{i=1}^\infty m_i(E)$, $E \in \mathcal{S}$, is also Lyapunov.

Proof. The definition of a control system implies that m is a well-defined vector measure. We show first that m is closed. Suppose A is some index set and $(E_\alpha)_{\alpha \in A}$ is an m -Cauchy set in \mathcal{S} . Then for each $i = 1, 2, \dots$, $(E_\alpha \cap E_i)_{\alpha \in A}$ is m_i -Cauchy, and since each m_i is closed, $(E_\alpha \Delta E_i)_{\alpha \in A}$ is m_i -convergent to a set $F_i \in \mathcal{S}$, where $F_i \subset E_i$. Then $(E_\alpha)_{\alpha \in A}$ is m -convergent to $F = \bigcup_{i=1}^\infty F_i \in \mathcal{S}$.

Let $f \in \mathcal{M}_{[0,1]}(\mathcal{S})$. Then

$$\int_T f dm = \sum_{i=1}^\infty \int_{E_i} f dm_i,$$

and since each m_i is Lyapunov we can find a set $H_i \in \mathcal{S}$, $H_i \subset E_i$, such that

$$\int_{E_i} f dm_i = m_i(H_i), \quad i = 1, 2, \dots$$

If $H = \bigcup_{i=1}^{\infty} H_i$, then $H \in \mathcal{S}$, and $\int_T f dm = m(H)$. Since m is closed we have [3, Thm. 2]

$$\overline{\text{co}} \mathcal{R}(m) = \left\{ \int_T f dm : f \in \mathcal{M}_{[0,1]}(\mathcal{S}) \right\} = \mathcal{R}(m).$$

THEOREM 2. Suppose $m: \mathcal{S} \rightarrow X$ is a vector measure, u an m -integrable real-valued function and $n: \mathcal{S} \rightarrow X$ a vector measure given by

$$n(E) = \int_E u dm, \quad E \in \mathcal{S}.$$

If m is Lyapunov, then n is Lyapunov. Conversely, suppose $E_0 = \{t: u(t) = 0\}$ and m restricted to \mathcal{S}_{E_0} is Lyapunov. Then if n is Lyapunov, m is Lyapunov.

Proof. Suppose m is Lyapunov. Set $E_i = \{t: i \leq u(t) < i + 1\}$, $i = 0, \pm 1, \pm 2, \dots$. Then $(E_i)_{i=-\infty}^{\infty}$ are pairwise disjoint and $\bigcup_{i=-\infty}^{\infty} E_i = T$. Define the sequence of vector measures $n_i: \mathcal{S} \rightarrow X$ by

$$n_i(F) = \int_F X_{E_i} u dm, \quad i = 0, \pm 1, \pm 2, \dots, \quad F \in \mathcal{S}.$$

Each n_i is Lyapunov by Lemma 2, and

$$n(F) = \sum_{i=-\infty}^{\infty} n_i(F), \quad F \in \mathcal{S}.$$

Lemma 3 gives the result.

Conversely, suppose n is Lyapunov. First consider the case $u(t) > 0$ for all $t \in T$. Set $E_1 = \{t: u(t) > 1\}$, $E_i = \{t: 1/i < u(t) \leq 1/(i-1)\}$, $i = 2, 3, \dots$. The sets $(E_i)_{i=1}^{\infty}$ are pairwise disjoint, and $\bigcup_{i=1}^{\infty} E_i = T$. Define $\mathcal{S}_i = \{F: F \subseteq E_i, F \in \mathcal{S}\}$ and let n_i, m_i be the restrictions of n, m to \mathcal{S}_i respectively, $i = 1, 2, \dots$. Each n_i will be Lyapunov from the definition. Since

$$n_i(F) = \int_F u dm_i, \quad F \in \mathcal{S}_i,$$

each m_i will be Lyapunov by Lemma 2. If we define a sequence of vector measures $m'_i: \mathcal{S} \rightarrow X$ by

$$m'_i(G) = m_i(G \cap E_i), \quad i = 1, 2, \dots, \quad G \in \mathcal{S},$$

then

$$\mathcal{R}(m'_i, T) = \mathcal{R}(m_i),$$

and so each m'_i is Lyapunov, $i = 1, 2, \dots$. Since

$$m(G) = \sum_{i=1}^{\infty} m'_i(G), \quad G \in \mathcal{S},$$

the result follows by Lemma 3.

In the general case we define $E_1 = \{t: u(t) > 0\}$, $E_2 = \{t: u(t) < 0\}$, and E_0 as given. The result follows as above by applying Lemma 3 to the restrictions of n and m to the σ -algebras \mathcal{S}_{E_i} , $i = 0, 1, 2$.

COROLLARY. *If $\{t: u(t) = 0\}$ is m -negligible in the previous theorem, then m is Lyapunov if and only if n is Lyapunov.*

3. The attainable set. We define $\text{lca}(\mathcal{S})$ to be the set of all sequences (μ_i) with $\mu_i \in \text{ca}(\mathcal{S})$, $i = 1, 2, \dots$, and $\sum_{i=1}^{\infty} |\mu_i|(T) < \infty$.

Suppose $\mathbf{m} = (m_i)$, where $m_i: \mathcal{S} \rightarrow X$ is a control system. Set

$$\Theta = \{(\langle x', m_i \rangle) : x' \in X'\}.$$

We call $f, g \in \mathcal{M}^\infty(\mathcal{S})$ \mathbf{m} -equivalent if

$$\sum_{i=1}^{\infty} \int |f_i - g_i| d|\theta_i| = 0 \quad \text{for all } \theta = (\theta_i) \in \Theta.$$

The set of all $g \in \mathcal{M}^\infty(\mathcal{S})$ \mathbf{m} -equivalent to $f \in \mathcal{M}^\infty(\mathcal{S})$ is denoted by $[f]_{\mathbf{m}}$. We define

$$M^\infty(\mathcal{S}, \mathbf{m}) = \{[f]_{\mathbf{m}} : f \in \mathcal{M}^\infty(\mathcal{S})\}.$$

For a bounded set-valued function \mathcal{F} mapping T into the subsets of R^∞ , we put

$$M_{\mathcal{F}}^\infty(\mathcal{S}, \mathbf{m}) = \{[f]_{\mathbf{m}} : f \in \mathcal{M}_{\mathcal{F}}^\infty(\mathcal{S})\}.$$

Set Γ as the set of all sequences $\mu = (\mu_i) \in \text{lca}(\mathcal{S})$ with $\mu_i \ll \theta_i$, $i = 1, 2, \dots$, for some $\theta = (\theta_i) \in \Theta$. Then the maps

$$[f]_{\mathbf{m}} \rightarrow \sum_{i=1}^{\infty} \int f_i d\mu_i, \quad [f]_{\mathbf{m}} \in M^\infty(\mathcal{S}, \mathbf{m}),$$

are well-defined for each $\mu = (\mu_i) \in \Gamma$. The weakest topology that makes all these maps continuous is called the $\sigma(\mathbf{m})$ topology.

For a bounded set-valued function \mathcal{F} taking values in the subsets of R^∞ we define a map $J: M_{\mathcal{F}}^\infty(\mathcal{S}, \mathbf{m}) \rightarrow X$ by

$$J([f]_{\mathbf{m}}) = \sum_{i=1}^{\infty} \int f_i dm_i, \quad [f]_{\mathbf{m}} \in M_{\mathcal{F}}^\infty(\mathcal{S}, \mathbf{m}).$$

Then J is well-defined, and continuous if $M_{\mathcal{F}}^\infty(\mathcal{S}, \mathbf{m})$ is given the $\sigma(\mathbf{m})$ topology and X its weak (i.e., $\sigma(X, X')$) topology.

We are now able to give conditions for $A_{\mathcal{F}}(\mathbf{m}) = A_{\text{ex } \mathcal{F}}(\mathbf{m})$.

THEOREM 3. *If $\mathbf{m} = (m_i)$ is a control system, then $A_{\mathcal{F}}(\mathbf{m}) = A_{\text{ex } \mathcal{F}}(\mathbf{m})$ for every scalarly measurable, bounded, set-valued function $\mathcal{F}: T \rightarrow CCR^\infty$, if and only if*

- (A) *for every $[u]_{\mathbf{m}} \in M^\infty(\mathcal{S}, \mathbf{m})$ with $[u]_{\mathbf{m}} \neq 0$, there exists a $v \in \mathcal{M}(\mathcal{S})$ with $[uv]_{\mathbf{m}} \neq 0$ and $\sum_{i=1}^{\infty} \int u_i v dm_i = 0$.*

Proof. Suppose condition (A) holds. We must now show that

$$M_{\mathcal{F}}^\infty(\mathcal{S}, \mathbf{m}) \subset M_{\text{ex } \mathcal{F}}^\infty(\mathcal{S}, \mathbf{m}) + J^{-1}(0).$$

Set $H = M_{\mathcal{F}}^\infty(\mathcal{S}, \mathbf{m})$. As $\text{ex } M_{\mathcal{F}}^\infty(\mathcal{S}, \mathbf{m}) = M_{\text{ex } \mathcal{F}}^\infty(\mathcal{S}, \mathbf{m})$ (see [5, Lemma 5]), the

proof follows in the same way the implication (ii) \Rightarrow (iii) was shown in Theorem 1.

Conversely, for any $u \in M^\infty(\mathcal{S}, \mathbf{m})$, $[u]_{\mathbf{m}} \neq 0$, consider the set-valued function defined by

$$\mathcal{F}(t) = \{au(t) : 0 \leq a \leq 1\}, \quad t \in T.$$

Then $\mathcal{F} : T \rightarrow CCR^\infty$ is bounded and scalarly measurable, and

$$M_{\text{ex}}^\infty(\mathcal{F}) = \{X_E u : E \in \mathcal{S}\}.$$

$$A_{\text{ex}}(\mathcal{F})(\mathbf{m}) = \left\{ \sum_{i=1}^{\infty} \int_E u_i dm_i ; E \in \mathcal{S} \right\}$$

is convex and weakly compact by hypothesis. Define a vector measure $w : \mathcal{S} \rightarrow X$ by

$$w(E) = \sum_{i=1}^{\infty} \int_E u_i dm_i, \quad E \in \mathcal{S}.$$

Then $\mathcal{R}(w) = A_{\text{ex}}(\mathcal{F})(\mathbf{m})$, and so w is Lyapunov. If T is w -negligible, we choose $v = 1$; otherwise, a function v satisfying property (A) can be chosen by Theorem 1.

COROLLARY 1. *If $\mathbf{m} = (m_i)$ is a Lyapunov control system in Theorem 3, and the space X has the property:*

(B) *for any Lyapunov control system $\mathbf{w} = (w_i)$, $w_i : \mathcal{S} \rightarrow X$, $i = 1, 2, \dots$, the vector measure $w : \mathcal{S} \rightarrow X$ defined by $w(E) = \sum_{i=1}^{\infty} w_i(E)$, $E \in \mathcal{S}$, is also Lyapunov;*

then $A_{\mathcal{F}}(\mathbf{m}) = A_{\text{ex}}(\mathcal{F})(\mathbf{m})$.

Proof. We shall show that condition (A) holds. Let $[u]_{\mathbf{m}} \in M^\infty(\mathcal{S}, \mathbf{m})$ and $[u]_{\mathbf{m}} \neq 0$. Define a vector measure $w : \mathcal{S} \rightarrow X$ by

$$w(E) = \sum_{i=1}^{\infty} \int_E u_i dm_i.$$

By (B) and Lemma 2, w is Lyapunov. As before we can now select a $v \in \mathcal{M}(\mathcal{S})$ satisfying property (A) from Theorem 1.

COROLLARY 2. *If $\mathbf{m} = (m_i)$, $m_i : \mathcal{S} \rightarrow \mathbb{R}^n$, n a positive integer, $i = 1, 2, \dots$, is a control system of nonatomic vector measures, and $\mathcal{F} : T \rightarrow CCR^\infty$ is a scalarly measurable, bounded, set-valued function, then $A_{\mathcal{F}}(\mathbf{m}) = A_{\text{ex}}(\mathcal{F})(\mathbf{m})$.*

Proof. We shall show that $X = \mathbb{R}^n$ satisfies condition (B). Let $\mathbf{w} = (w_i)$ be any control system of n -dimensional Lyapunov (i.e., nonatomic) vector measures, and define $w : \mathcal{S} \rightarrow \mathbb{R}^n$ by

$$w(E) = \sum_{i=1}^{\infty} w_i(E), \quad E \in \mathcal{S}.$$

If $n = 1$, and each of the measures w_i is positive, then clearly w is nonatomic and so Lyapunov.

In the general case, for each vector measure w_i we can find a finite, positive, real-valued measure λ_i such that $\lambda_i(E) \rightarrow 0$, $E \in \mathcal{S}$, implies $w_i(E) \rightarrow 0$ (see [1]). These (λ_i) can be chosen to form a control system. Then the measure $\lambda : \mathcal{S} \rightarrow \mathbb{R}$

defined by

$$\lambda(E) = \sum_{i=1}^{\infty} \lambda_i(E), \quad E \in \mathcal{S},$$

will be finite, positive and nonatomic from our earlier remarks. Since $\lambda(E) \rightarrow 0$, $E \in \mathcal{S}$, implies $w(E) \rightarrow 0$, by [3, Lemma 3] w is nonatomic and so Lyapunov.

A vector measure $m: \mathcal{S} \rightarrow X$ will be called scalarly nonatomic if $\langle x', m \rangle: \mathcal{S} \rightarrow \mathbb{R}$ is nonatomic for each $x' \in X'$.

For a general quasi-complete l.c.t.v.s. X the problem of whether $A_{\mathcal{F}}(\mathbf{m}) = A_{\text{ex } \mathcal{F}}(\mathbf{m})$ for a Lyapunov control system $\mathbf{m} = (m_i)$, $m_i: \mathcal{S} \rightarrow X$, $i = 1, 2, \dots$, is as yet unsolved. However, under the weaker assumption that each of the control system measures m_i , $i = 1, 2, \dots$, is scalarly nonatomic, we can recover $A_{\mathcal{F}}(\mathbf{m})$ from the weak closure of $A_{\text{ex } \mathcal{F}}(\mathbf{m})$. The counterexample of Uhl [7] with \mathcal{F} constant and taking values in the product of intervals shows that this result is the best possible. For completeness we give some sufficient conditions for a vector measure to be scalarly nonatomic.

Suppose $m: \mathcal{S} \rightarrow X$ is a vector measure. Two sets $E, F \in \mathcal{S}$ are termed m -equivalent if $E \Delta F = (E - F) \cup (F - E)$ is m -negligible. The class of sets in \mathcal{S} which are m -equivalent to $E \in \mathcal{S}$ is denoted by $[E]_m$. The σ -algebra \mathcal{S} is called m -essentially countably generated if there exists a countably generated σ -algebra $\mathcal{S}_0 \subset \mathcal{S}$ such that, for every $E \in \mathcal{S}$, there is an $F \in \mathcal{S}_0$ with $E \in [F]_m$. The following result is proved in [3, Lemmas 2, 4].

LEMMA 4. *Let $m: \mathcal{S} \rightarrow X$ be a nonatomic vector measure. If either X is metrizable or \mathcal{S} is m -essentially countably generated, then m is scalarly nonatomic.*

THEOREM 4. *If $\mathbf{m} = (m_i)$ is a control system of scalarly nonatomic vector measures, and $\mathcal{F}: T \rightarrow CCR^\infty$ is a bounded scalarly measurable set-valued function, then $A_{\mathcal{F}}(\mathbf{m})$ coincides with the weak closure of $A_{\text{ex } \mathcal{F}}(\mathbf{m})$.*

Proof. Since \mathcal{F} is scalarly measurable, $M_{\mathcal{F}}^\infty(\mathcal{S}, \mathbf{m}) \neq \emptyset$ [8, Prop. 7]. By [5, Lemma 5], $\text{ex}(M_{\mathcal{F}}^\infty(\mathcal{S}, \mathbf{m})) = M_{\text{ex } \mathcal{F}}^\infty(\mathcal{S}, \mathbf{m})$, and so $M_{\text{ex } \mathcal{F}}^\infty(\mathcal{S}, \mathbf{m}) \neq \emptyset$. If $M_{\text{ex } \mathcal{F}}^\infty(\mathcal{S}, \mathbf{m})$ is a singleton, the result is clear from the Krein–Millman theorem. Otherwise, suppose $[f]_m, [g]_m \in M_{\text{ex } \mathcal{F}}^\infty(\mathcal{S}, \mathbf{m})$ are distinct. Define a vector measure $n: \mathcal{S} \rightarrow X$ by

$$n(E) = \sum_{i=1}^{\infty} \int_E (g_i - f_i) dm_i, \quad E \in \mathcal{S}.$$

Following the proof of Theorem 3, Corollary 2 we see that n is scalarly nonatomic, and so $\overline{\mathcal{R}(n)}$ is convex (e.g., [3, Thm. 1]), where the overbar denotes weak closure. Now

$$\begin{aligned} \sum_{i=1}^{\infty} \int_T f_i dm_i &= n(\phi) + \sum_{i=1}^{\infty} \int_T f_i dm_i, \\ \sum_{i=1}^{\infty} \int_T g_i dm_i &= n(T) + \sum_{i=1}^{\infty} \int_T f_i dm_i; \end{aligned}$$

consequently, for any $0 \leq \lambda \leq 1$,

$$\lambda \left(\sum_{i=1}^{\infty} \int_T f_i dm_i \right) + (1 - \lambda) \left(\sum_{i=1}^{\infty} \int_T g_i dm_i \right) = \lambda n(\phi) + (1 - \lambda) n(T) + \sum_{i=1}^{\infty} \int_T f_i dm_i.$$

Since $\overline{\mathcal{R}(n)}$ is convex, $\lambda n(\phi) + (1 - \lambda)n(T) \in \overline{\mathcal{R}(n)}$, and so there exists some directed set A , and a net $(Z_\alpha)_{\alpha \in A}$, $Z_\alpha \in \mathcal{S}$, with

$$n(Z_\alpha) \xrightarrow{w} \lambda n(\phi) + (1 - \lambda)n(T), \quad \alpha \in A.$$

Set

$$n(Z_\alpha) + \sum_{i=1}^{\infty} \int_T f_i dm_i = \sum_{i=1}^{\infty} \int_T h_i^\alpha dm_i$$

with

$$h^\alpha(t) = \begin{cases} g(t), & t \in Z^\alpha, \\ f(t), & t \notin Z^\alpha. \end{cases}$$

With this choice of h^α , $\alpha \in A$, we have $[h^\alpha]_{\mathbf{m}} \in M_{\text{ex } \mathcal{F}}^\infty(\mathcal{S}, \mathbf{m})$, $\alpha \in A$, and

$$\sum_{i=1}^{\infty} \int_T h_i^\alpha dm_i \xrightarrow{w} \lambda \left(\sum_{i=1}^{\infty} \int_T f_i dm_i \right) + (1 - \lambda) \left(\sum_{i=1}^{\infty} \int_T g_i dm_i \right).$$

For $x, y \in \overline{A_{\text{ex } \mathcal{F}}(\mathbf{m})}$, we can find a directed set B and nets x_β, y_β , $\beta \in B$, such that $x_\beta \xrightarrow{w} x$, $y_\beta \xrightarrow{w} y$, $\beta \in B$. Since $\lambda x_\beta + (1 - \lambda)y_\beta \in \overline{A_{\text{ex } \mathcal{F}}(\mathbf{m})}$ by the above, for each $0 \leq \lambda \leq 1$, $\overline{A_{\text{ex } \mathcal{F}}(\mathbf{m})}$ must be convex. Thus by the Krein–Millman theorem, $A_{\mathcal{F}}(\mathbf{m}) = \overline{\text{co } A_{\text{ex } \mathcal{F}}(\mathbf{m})} = \overline{A_{\text{ex } \mathcal{F}}(\mathbf{m})}$.

REFERENCES

- [1] R. G. BARTLE, N. DUNFORD AND J. T. SCHWARTZ, *Weak compactness and vector measures*, Canad. J. Math., 7 (1955), pp. 289–305.
- [2] J. F. C. KINGMAN AND A. P. ROBERTSON, *On a theorem of Lyapunov*, J. London Math. Soc., 47 (1968), pp. 347–351.
- [3] I. KLUVANEK, *The range of a vector-valued measure*, Math. Systems Theory, 7 (1973), pp. 44–54.
- [4] ———, *Fourier transforms of vector-valued functions*, Studia Math., 37 (1970), pp. 1–12.
- [5] I. KLUVANEK AND G. KNOWLES, *Attainable sets in infinite dimensional spaces*, Math. Systems Theory, 7 (1973).
- [6] G. KNOWLES, *Vector integration of set-valued functions*, Ibid., to appear.
- [7] J. J. UHL, *The range of a vector-valued measure*, Proc. Amer. Math. Soc., 23 (1969), pp. 158–163.
- [8] M. VALADIER, *Multi-applications à valeurs convexes et compacts*, J. Math. Pures Appl., 50 (1971), pp. 265–297.
- [9] H. HERMES AND J. P. LASALLE, *Functional Analysis and Time Optimal Control*, Academic Press, New York, 1969.

ESTIMATION AND FEEDBACK IN LINEAR TIME-VARYING SYSTEMS: A DETERMINISTIC THEORY*

MASAO IKEDA,[†] HAJIME MAEDA[‡] AND SHINZO KODAMA[‡]

Abstract. The problems of (i) state-variable feedback, (ii) deterministic state estimation, and (iii) state estimation and feedback are considered for a finite-dimensional linear time-varying system. The following are the questions; they can be regarded as natural and logical extensions to a time-varying case of the pole assignment problem [2]–[6] for a time-invariant system.

(i) What is the relation between controllability and reachability of an open loop system and the possibility of constructing a linear state-variable feedback law which realizes the prescribed stability and instability degrees of the closed loop system?

(ii) What is the relation between observability and reconstructibility of the system to be estimated and the possibility of constructing a linear state estimator which realizes the prescribed stability and instability degrees of the estimate error?

(iii) What is the relation between controllability, reachability, observability and reconstructibility of an open loop system and the possibility of constructing a linear feedback compensator composed of a state estimator and an amplifier which realizes the prescribed stability and instability degrees of the closed loop system?

Here, the stability and instability degrees mean the indices of exponential functions which bound the (zero-input) response from above and below respectively. This paper gives some principal answers to the questions.

1. Introduction. This paper considers a finite-dimensional linear time-varying system and is concerned with the synthesis problem of a linear feedback compensator to realize specified stability properties of the closed loop system. It is assumed that the compensator consists of a dynamic state estimator and an amplifier. The object is to clarify the relations between the existence of such a compensator and controllability, reachability, observability and reconstructibility [1] of the open loop system. For this purpose, three subproblems—problems of state-variable feedback, deterministic state estimation, and state estimation and feedback—are considered, which themselves are very important problems.

First, in § 3, the state-variable feedback problem is considered from a viewpoint of designing the internal stability property (i.e., the zero-input behavior) of the closed loop system which can be obtained by means of linear feedback of the state. The question is as follows: Under what condition on the open loop system does the feedback gain exist which actualizes the prescribed stability and instability degrees of the closed loop system? Here, stability and instability degrees mean the indices of exponential functions which bound the zero-input response from above and below respectively. For a time-invariant case, this problem is

* Received by the editors January 22, 1973, and in revised form October 6, 1973.

[†] Department of Systems Engineering, Faculty of Engineering, Kobe University, Rokko-Dai, Nada, Kobe 657, Japan.

[‡] Department of Communication Engineering, Faculty of Engineering, Osaka University, Yamada-Kami, Suita, Osaka 565, Japan.

completely solved [2]–[6], i.e., it is well known that controllability of the open loop system is equivalent to the possibility of assigning an arbitrary set of the closed loop poles. For a time-varying case, Wolovich [7] considers this problem for a class of bounded systems whose parameters have bounded derivatives of higher order, and obtains a result which implies that a slightly more restrictive condition than uniform controllability is a sufficient condition for the closed loop system to have the desired stability and instability degrees. For a more general class of time-varying systems, a partial answer is known, i.e., the problem of realizing the prescribed stability degree by means of linear state-variable feedback has been studied extensively [8]–[10]; it is known that uniform complete controllability is a sufficient condition for such stabilization, and in bounded systems it is also necessary as well as sufficient.

Section 3 solves the problem completely. It is shown that uniform complete controllability of the open loop system is a sufficient condition for specifying any stability and instability degrees (of course, upper index \geq lower index) of the closed loop system simultaneously, and that in case of a bounded open loop system and a bounded feedback gain, uniform complete controllability is also necessary. Thus the results of § 3 can be regarded as logical and natural extensions to a time-varying case of those for a time-invariant system.

The deterministic state estimation problem is considered in § 4, where it is formulated as a design problem of a dynamic state estimator whose dimension is equal to that of the open loop system to be estimated. The relations are clarified between uniform complete observability of the open loop system and the existence of an estimator which realizes the prescribed stability and instability degrees of the estimate error. For this, it is shown that such an estimation problem can be reduced to the state-variable feedback problem for the dual system which is well known for a time-invariant case [6]. Thus we obtain the results corresponding to those attained in § 3.

When state estimate is employed for feedback, a question arises as to whether we are still able to specify the stability and instability degrees of the closed loop system. In § 5, it is shown that uniform complete controllability and uniform complete observability of the open loop system guarantee the existence of a feedback compensator (composed of a dynamic estimator and an amplifier) which actualizes the prescribed stability and instability degrees of the total closed loop system, and that in case of a bounded open loop system and a bounded compensator, the converse is also valid. However, the difference is that the upper index can be assigned any value provided it is strictly greater than the lower index, whereas in § 3 they can be equal. This result is also well known for a time-invariant system [6], and has been obtained for a class of time-varying systems [18], [19].

Throughout the paper, all vectors and matrices are assumed to have real elements. For a vector x and a matrix A , x' and A' are respectively the transposes. The norm of x denoted by $\|x\|$ is the Euclidean norm, i.e., $\|x\| \triangleq (x'x)^{1/2}$, and $\|A\|$ denotes the norm of A compatible with $\|x\|$, i.e., $\|A\| \triangleq \sup_{\|x\|=1} \|Ax\|$. For symmetric matrices P and Q , $P > 0$ ($P \geq 0$) means that P is positive (nonnegative) definite, and $P > Q$ ($P \geq Q$) means $P - Q > 0$ ($P - Q \geq 0$). I is the identity matrix.

2. System description and controllability, reachability, observability, reconstructibility. We assume that the open loop system has a representation

$$(S_o) \quad \begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)u(t), \\ y(t) &= C(t)x(t) + D(t)u(t), \end{aligned}$$

where $x(\cdot)$ is the state n -vector, $u(\cdot)$ is the input r -vector, $y(\cdot)$ is the output m -vector, and $A(\cdot)$, $B(\cdot)$, $C(\cdot)$ and $D(\cdot)$ are matrix functions with appropriate dimension. In addition, $u(\cdot)$, $A(\cdot)$, $B(\cdot)$, $C(\cdot)$ and $D(\cdot)$ are assumed to be measurable and bounded on every finite subinterval of time. The transition matrix associated with $\dot{x}(t) = A(t)x(t)$ is denoted by $\Phi(\cdot, \cdot)$. In view of the assumptions on $A(\cdot)$, a unique, continuous, nonsingular $\Phi(\cdot, \cdot)$ exists on $(-\infty, \infty)$ [12].

Concerning controllability and reachability, we introduce the following definitions. Let the controllability Gramian and the reachability Gramian of (S_o) be denoted by $W(\cdot, \cdot)$ and $Y(\cdot, \cdot)$ respectively, i.e.,

$$(1) \quad W(t, s) = \int_t^s \Phi(t, \tau) B(\tau) B'(\tau) \Phi'(t, \tau) d\tau, \quad t < s,$$

$$(2) \quad Y(s, t) = \int_s^t \Phi(t, \tau) B(\tau) B'(\tau) \Phi'(t, \tau) d\tau, \quad s < t.$$

DEFINITION 2.1. (S_o) is said to be *uniform with respect to complete controllability* if there are positive numbers σ , α_1 and α_2 such that

$$(3) \quad 0 < \alpha_1 I \leq W(t, t + \sigma) \leq \alpha_2 I$$

holds for all t .

The concept of uniformity with respect to complete controllability means that we can transfer the state from any x at any t to the origin 0 in a finite time σ , and moreover that the minimum input energy required for this transference is neither larger than $\alpha_1^{-1} \|x\|^2$ nor smaller than $\alpha_2^{-1} \|x\|^2$ (see [13]).

DEFINITION 2.2. (S_o) is said to be *uniform with respect to complete reachability* if there are positive numbers σ , α_3 and α_4 such that

$$(4) \quad 0 < \alpha_3 I \leq Y(t - \sigma, t) \leq \alpha_4 I$$

holds for all t .

Like uniformity with respect to complete controllability, the concept of uniformity with respect to complete reachability means the uniformity of the minimum input energy which is required for transferring the state from 0 to x in a finite time σ [13].

DEFINITION 2.3 [8]. (S_o) is said to be *uniformly completely controllable* if there are positive numbers σ and α_i , $i = 1, 2, 3, 4$, such that

$$(5a) \quad 0 < \alpha_1 I \leq W(t, t + \sigma) \leq \alpha_2 I,$$

$$(5b) \quad 0 < \alpha_3 I \leq Y(t - \sigma, t) \leq \alpha_4 I$$

hold for all t .

Remark 2.1. It is obvious from the above definitions that if (S_0) is uniformly completely controllable, then it is both uniform with respect to complete controllability and uniform with respect to complete reachability. Besides, the converse is also valid. (The proof of this statement is omitted.) Therefore, uniform complete controllability is the conjunction of uniformity with respect to complete controllability and uniformity with respect to complete reachability.

Remark 2.2. It is shown in [8] that if (5a) and (5b) hold for some σ , then there is a scalar function $\gamma(\cdot)$ such that

$$(5c) \quad \|\Phi(t, s)\| \leq \gamma(|t - s|) \quad \text{for all } t \text{ and } s,$$

and that any two of the three conditions (5a), (5b) and (5c) imply the remainder. This fact indicates that uniformity with respect to complete controllability, uniformity with respect to complete reachability and uniform complete controllability are equivalent for a bounded system, since (5c) is immediately implied by the boundedness of $A(\cdot)$.

Remark 2.3. For an unbounded system, the above three concepts are not equivalent. This is shown by the following examples.

Example 2.1. Consider a scalar system

$$(6) \quad \dot{x}(t) = 3t^2x(t) + \sqrt{6}tu(t).$$

The controllability Gramian with $\sigma = 1$ is

$$W(t, t + 1) = 1 - \exp[-6t^2 - 6t - 2],$$

and hence

$$1 - \exp[-1/2] \leq W(t, t + 1) \leq 1 \quad \text{for all } t.$$

Therefore, (6) is uniform with respect to complete controllability. But, it is not uniform with respect to complete reachability since the reachability Gramian is

$$Y(t - \sigma, t) = \exp[6\sigma t^2 - 6\sigma^2t + 2\sigma^3] - 1 \quad \text{for } \sigma > 0.$$

Naturally, it is not uniformly completely controllable.

Example 2.2. A scalar system

$$(7) \quad \dot{x}(t) = -3t^2x(t) + \sqrt{6}tu(t)$$

is uniform with respect to complete reachability. But it is not uniform with respect to complete controllability.

The definitions relative to observability and reconstructibility are the following. Let the observability Gramian and reconstructibility Gramian of (S_0) be denoted by $M(\cdot, \cdot)$ and $N(\cdot, \cdot)$ respectively, i.e.,

$$(8) \quad M(t, s) = \int_t^s \Phi'(\tau, t)C'(\tau)C(\tau)\Phi(\tau, t) d\tau, \quad t < s,$$

$$(9) \quad N(s, t) = \int_s^t \Phi'(\tau, t)C'(\tau)C(\tau)\Phi(\tau, t) d\tau, \quad s < t.$$

DEFINITION 2.4. (S_o) is said to be *uniform with respect to complete observability* if there are positive numbers σ , β_1 and β_2 such that

$$(10) \quad 0 < \beta_1 I \leq M(t, t + \sigma) \leq \beta_2 I$$

holds for all t .

DEFINITION 2.5. (S_o) is said to be *uniform with respect to complete reconstructibility* if there are positive numbers σ , β_3 and β_4 such that

$$(11) \quad 0 < \beta_3 I \leq N(t - \sigma, t) \leq \beta_4 I$$

holds for all t .

DEFINITION 2.6 [8]. (S_o) is said to be *uniformly completely observable* if there are positive numbers σ and β_i , $i = 1, 2, 3, 4$, such that

$$(12a) \quad 0 < \beta_1 I \leq M(t, t + \sigma) \leq \beta_2 I,$$

$$(12b) \quad 0 < \beta_3 I \leq N(t - \sigma, t) \leq \beta_4 I$$

hold for all t .

Remark 2.4. Similar to the case of controllability and reachability, uniform complete observability is the conjunction of uniformity with respect to complete observability and uniformity with respect to complete reconstructibility, and these three concepts are equivalent for a bounded system.

It is well known that observability and reconstructibility of (S_o) are closely related to reachability and controllability of its dual system [1], [8], [11]

$$(D_u) \quad \begin{aligned} \dot{x}(t) &= A'(-t)x(t) + C'(-t)u(t), \\ y(t) &= B'(-t)x(t) + D'(-t)u(t). \end{aligned}$$

The following lemmas are immediately implied by the relation between the transition matrices associated with $\dot{x}(t) = A(t)x(t)$ and $\dot{x}(t) = A'(-t)x(t)$.

LEMMA 2.1. (S_o) is uniform with respect to complete reconstructibility if and only if (D_u) is uniform with respect to complete controllability.

LEMMA 2.2. (S_o) is uniform with respect to complete observability if and only if (D_u) is uniform with respect to complete reachability.

LEMMA 2.3. (S_o) is uniformly completely observable if and only if (D_u) is uniformly completely controllable.

3. State-variable feedback. In this section, we deal with only the state equation

$$(S'_o) \quad \dot{x}(t) = A(t)x(t) + B(t)u(t)$$

and assume that the state $x(\cdot)$ can be measured directly. Consider linear state-variable feedback of the form

$$(13) \quad u(t) = K(t)x(t) + v(t),$$

where the feedback gain $K(\cdot)$ is $r \times n$, measurable and bounded on every finite subinterval of time, and $v(\cdot)$ is the new input. Then the closed loop system has a representation

$$(S'_c) \quad \dot{x}(t) = \{A(t) + B(t)K(t)\}x(t) + B(t)v(t).$$

In view of the assumptions on $A(\cdot)$, $B(\cdot)$ and $K(\cdot)$, a unique, continuous and nonsingular transition matrix associated with $\dot{x}(t) = \{A(t) + B(t)K(t)\}x(t)$ exists on $(-\infty, \infty)$ [12].

Now, we introduce the concepts of stabilizability, anticausal-stabilizability and designability, which are concerned with the stability and instability degrees that can be obtained by the state-variable feedback.

DEFINITION 3.1. (S'_0) is said to be *uniformly completely stabilizable* if, for any real number M , there are a positive number b and a feedback gain $K(\cdot)$ such that any zero-input response of (S'_0) satisfies

$$(14) \quad \|x(t_2)\| \leq b\|x(t_1)\| \exp [M(t_2 - t_1)] \quad \text{for all } t_1 \text{ and } t_2 \geq t_1.$$

DEFINITION 3.2. (S'_0) is said to be *uniformly completely anticausal-stabilizable*¹ if, for any real number m , there are a positive number a and a feedback gain $K(\cdot)$ such that any zero-input response of (S'_0) satisfies

$$(15) \quad a\|x(t_1)\| \exp [m(t_2 - t_1)] \leq \|x(t_2)\| \quad \text{for all } t_1 \text{ and } t_2 \geq t_1.$$

DEFINITION 3.3. (S'_0) is said to be *uniformly completely designable* if, for any pair of real numbers m and M such that $m \leq M$, there are positive numbers a , b and a feedback gain $K(\cdot)$ such that any zero-input response of (S'_0) satisfies

$$(16) \quad a\|x(t_1)\| \exp [m(t_2 - t_1)] \leq \|x(t_2)\| \leq b\|x(t_1)\| \exp [M(t_2 - t_1)]$$

for all t_1 and $t_2 \geq t_1$.

Remark 3.1. Definition 3.1 is equivalent to Definition 4 of [10] in which M is restricted to be nonpositive; if some b and $K(\cdot)$ for a nonpositive M are found, then they also work for any positive M .

Remark 3.2. It is obvious from the definitions that uniform complete designability implies both uniform complete stabilizability and uniform complete anticausal-stabilizability. However, it remains to show that the converse is also true or not.

Remark 3.3. There is a system which is not uniformly completely stabilizable, but uniformly completely anticausal-stabilizable. The opposite case also exists.

Example 3.1. Consider a scalar system

$$(17) \quad \dot{x}(t) = 2t \cdot x(t) + 1(t)u(t),$$

where $1(\cdot)$ is the unit step function. Let the feedback gain be defined by

$$K(t) = (-2t + M)1(t) \quad \text{for any } M.$$

Then any zero-input response of the closed loop system satisfies

$$|x(t_2)| \leq b|x(t_1)| \exp [M(t_2 - t_1)], \quad b \triangleq \exp [M^2/4]$$

for all t_1 and $t_2 \geq t_1$. Therefore, (17) is uniformly completely stabilizable. However, it is not uniformly completely anticausal-stabilizable; for any feedback gain the closed loop zero-input response at $t_2 = 0$ for the initial time $t_1 < 0$ is

$$x(0) = x(t_1) \exp [-t_1^2],$$

¹ The terminology "anticausal" means that the time-ordering is the opposite of the usual one [17].

and there is no positive number a for a fixed real number m such that

$$a \exp [m(0 - t_1)] \leq \exp [-t_1^2] \quad \text{for all } t_1 < 0.$$

Example 3.2. A scalar system

$$(18) \quad \dot{x}(t) = -2t \cdot x(t) + 1(t)u(t)$$

is uniformly completely anticausal-stabilizable, but is not uniformly completely stabilizable.

3.1. Stabilization. It is known that (S'_c) can be made uniformly asymptotically stable with the prescribed stability degree if (S'_o) is uniformly completely controllable [9], [10]. Theorem 3.1 improves this; a weaker sufficient condition for such stabilization is given.

THEOREM 3.1. *If (S'_o) is uniform with respect to complete controllability, then it is uniformly completely stabilizable.*

THEOREM 3.2. *A bounded system (S'_o) is uniformly completely stabilizable by a bounded feedback gain if and only if it is uniform with respect to complete controllability.*

Remark 3.4. The converse of Theorem 3.1 is not valid, though its condition is weaker than that of Theorem 2 of [10]. This fact is shown by the following example.

Example 3.3. Consider the scalar system (17) of Example 3.1 again. It is uniformly completely stabilizable as shown there. But it is not uniform with respect to complete controllability since, for any $\sigma > 0$, its controllability Gramian is zero for all $t < -\sigma$.

Proof of Theorem 3.1. Assume that (S'_o) is uniform with respect to complete controllability. For a real number M , define a feedback gain $K(\cdot)$ by

$$(19) \quad K(t) = -\frac{1}{2}B'(t)\tilde{W}^{-1}(t, t + \sigma),$$

where

$$\tilde{W}(t, t + \sigma) \triangleq \int_t^{t+\sigma} \Phi(t, \tau)B(\tau)B'(\tau)\Phi'(t, \tau) \exp [-2M(t - \tau)] d\tau$$

and σ is the positive number for which (3) holds for all t . Note that

$$0 < \alpha_1 \exp [-2|M|\sigma]I \leq \tilde{W}(t, t + \sigma) \leq \alpha_2 \exp [2|M|\sigma]I$$

holds for all t . Let a scalar function be defined by

$$(20) \quad V(x, t) = x'\tilde{W}^{-1}(t, t + \sigma)x.$$

Then, it satisfies

$$0 < \alpha_2^{-1} \exp [-2|M|\sigma]\|x\|^2 \leq V(x, t) \leq \alpha_1^{-1} \exp [2|M|\sigma]\|x\|^2$$

at every t , and its time derivative $\dot{V}(\cdot, \cdot)$ along any zero-input response of (S'_c) satisfies

$$\dot{V}(x(t), t) \leq 2MV(x(t), t)$$

at almost every t . Hence, by means of Lyapunov's second method, we can show that there is a positive number b such that any zero-input response of (S'_0) satisfies (14) for all t_1 and $t_2 \geq t_1$. Q.E.D.

Proof of Theorem 3.2. Necessity. When (S'_0) is bounded, its controllability Gramian is always bounded from above independently of t [14]. Hence, a bounded system (S'_0) is uniform with respect to complete controllability if and only if there are positive numbers σ and α_1 such that

$$0 < \alpha_1 I \leq W(t, t + \sigma)$$

holds for all t . With this fact, we can show in the same way as in the proof of the necessity part of Theorem 3 of [10] that (S'_0) is uniform with respect to complete controllability in case it is uniformly completely stabilizable by a bounded feedback gain.

Sufficiency. The proof of Theorem 3.1 applies here. Since $B(\cdot)$ is bounded in this case, $K(\cdot)$ defined by (19) is bounded. Q.E.D.

From the definitions, uniformity with respect to complete controllability is implied by uniform complete controllability, and for a bounded system, these concepts are equivalent as mentioned in Remarks 2.1 and 2.2. The following corollaries follow from this relation.

COROLLARY 3.1 (Theorem 2 of [10]). *If (S'_0) is uniformly completely controllable, then it is uniformly completely stabilizable.*

COROLLARY 3.2 (Theorem 3 of [10]). *A bounded system (S'_0) is uniformly completely stabilizable by a bounded feedback gain if and only if it is uniformly completely controllable.*

Remark 3.5. Theorem 3.2 follows immediately from Theorem 3 of [10] and the fact that uniformity with respect to complete controllability is equivalent to uniform complete controllability for a bounded system. Nevertheless, the proof of Theorem 3.2 is notable; it does not take a round about way through uniform complete controllability (which is the concept concerned with reachability as well as controllability) and hence it indicates that stabilizability is related to controllability only.

3.2. Instabilization (specification of instability degree). From the analogy of the relations between stabilizability and controllability, it is natural to expect from the physical meanings that anticausal-stabilizability is closely related to reachability.

THEOREM 3.3. *If (S'_0) is uniform with respect to complete reachability, it is uniformly completely anticausal-stabilizable.*

THEOREM 3.4. *A bounded system (S'_0) is uniformly completely anticausal-stabilizable by a bounded feedback gain if and only if it is uniform with respect to complete reachability.*

Remark 3.6. The converse of Theorem 3.3 is not valid. An example which shows this fact is the scalar system (18) of Example 3.2.

The proof of Theorem 3.3 is analogous to that of Theorem 3.1. Define $K(\cdot)$ for any real number m by

$$(21) \quad K(t) = \frac{1}{2} B'(t) \tilde{Y}^{-1}(t - \sigma, t),$$

where

$$\tilde{Y}(t - \sigma, t) \triangleq \int_{t-\sigma}^t \Phi(t, \tau) B(\tau) B'(\tau) \Phi'(t, \tau) \exp[-2m(t - \tau)] d\tau$$

and σ is the positive number for which (4) holds for all t . Let a scalar function be defined by

$$(22) \quad V(x, t) = x' \tilde{Y}^{-1}(t - \sigma, t) x,$$

and we can show by means of Lyapunov's second method that there is a positive number a such that any zero-input response of (S'_c) satisfies (15) for all t_1 and $t_2 \geq t_1$.

The proof of Theorem 3.4 is analogous to that of Theorem 3.2. In this case, we must note the following three facts to prove the necessity part. First, the transition matrices $\Phi(\cdot, \cdot)$ and $\hat{\Phi}(\cdot, \cdot)$ which are associated with $\dot{x}(t) = A(t)x(t)$ and $\dot{x}(t) = \{A(t) + B(t)K(t)\}x(t)$ respectively satisfy

$$\Phi(t_2, t_1) \hat{\Phi}(t_1, t_2) = I - \int_{t_1}^{t_2} \Phi(t_2, \tau) B(\tau) K(\tau) \hat{\Phi}(\tau, t_2) d\tau$$

for all t_1 and t_2 . Second, any zero-input response of (S'_c) satisfies (15) for all t_1 and $t_2 \geq t_1$ if and only if

$$\|\hat{\Phi}(t_1, t_2)\| \leq (1/a) \exp[-m(t_2 - t_1)]$$

for all t_1 and $t_2 \geq t_1$. Third, a bounded system (S'_o) is uniform with respect to complete reachability if and only if there are positive numbers σ and α_3 such that

$$0 < \alpha_3 I \leq Y(t - \sigma, t).$$

holds for all t .

The following corollaries immediately follow from Theorems 3.3 and 3.4, and Remarks 2.1 and 2.2.

COROLLARY 3.3. *If (S'_o) is uniformly completely controllable, then it is uniformly completely anticausal-stabilizable.*

COROLLARY 3.4. *A bounded system (S'_o) is uniformly completely anticausal-stabilizable by a bounded feedback gain if and only if it is uniformly completely controllable.*

3.3. Specification of stability and instability degrees. It has been shown that uniformity with respect to complete controllability implies uniform complete stabilizability, uniformity with respect to complete reachability implies uniform complete anticausal-stabilizability, and both converses are valid in the case of a bounded open loop system and a bounded feedback gain. From these facts, we expect that analogous relations should hold between uniform complete controllability and uniform complete designability.

THEOREM 3.5. *If (S'_o) is uniformly completely controllable, then it is uniformly completely designable.*

THEOREM 3.6. *A bounded system (S'_o) is uniformly completely designable by a bounded feedback gain if and only if it is uniformly completely controllable.*

Remark 3.7. The converse of Theorem 3.5 is not valid. This fact is shown by the following example.

Example 3.4. Consider a scalar system

$$(23) \quad \dot{x}(t) = x(t) + \exp[2t] \cdot u(t).$$

For any pair of real numbers m and M such that $m \leq M$, define

$$K(t) = \{-1 + (m + M)/2\} \exp[-2t].$$

Then, any zero-input response of (S'_c) satisfies

$$x(t_2) = x(t_1) \exp[\tfrac{1}{2}(m + M)(t_2 - t_1)] \quad \text{for all } t_1 \text{ and } t_2.$$

This implies that (23) is uniformly completely designable. However, since the controllability Gramian and the reachability Gramian of (23) are

$$W(t, t + \sigma) = \tfrac{1}{2}(\exp[2\sigma] - 1) \exp[4t],$$

$$Y(t - \sigma, t) = \tfrac{1}{2}(1 - \exp[-2\sigma]) \exp[4t]$$

for any $\sigma > 0$, (23) is not uniformly completely controllable.

In the proofs of the theorems, the following two lemmas are necessary.

LEMMA 3.1. *If (S'_0) is uniformly completely controllable, then there are positive numbers c_1 and c_2 such that*

$$(24) \quad \int_{t_1}^{t_2} \|B(\tau)\|^2 d\tau \leq c_1 + c_2(t_2 - t_1)$$

for all t_1 and $t_2 \geq t_1$.

Proof. Note that $\|B(t)\|^2 = \|B'(t)\|^2 = \lambda_{\max}(B(t)B'(t)) = \|B(t)B'(t)\|$, where $\lambda_{\max}(\cdot)$ denotes the maximum eigenvalue. For any t_1 ,

$$(25) \quad \begin{aligned} & \int_{t_1}^{t_1 + \sigma} \|B(\tau)\|^2 d\tau \\ &= \int_{t_1}^{t_1 + \sigma} \|\Phi(\tau, t_1)\Phi(t_1, \tau)B(\tau)B'(\tau)\Phi'(t_1, \tau)\Phi'(\tau, t_1)\| d\tau \\ &\leq \sup_{t_1 \leq \tau \leq t_1 + \sigma} \|\Phi(\tau, t_1)\|^2 \int_{t_1}^{t_1 + \sigma} \|\Phi(t_1, \tau)B(\tau)B'(\tau)\Phi'(t_1, \tau)\| d\tau \\ &\leq \sup_{0 \leq \tau \leq \sigma} (\gamma(\tau))^2 n \alpha_2 \triangleq c_1 \end{aligned}$$

holds, where σ is a positive number for which (5a) and (5b) hold for all t . Therefore, for any t_1 and t_2 such that $t_1 + i\sigma \leq t_2 \leq t_1 + (i + 1)\sigma$, $i = 0, 1, 2, \dots$,

$$\begin{aligned} \int_{t_1}^{t_2} \|B(\tau)\|^2 d\tau &\leq \int_{t_1}^{t_1 + (i+1)\sigma} \|B(\tau)\|^2 d\tau \\ &= \int_{t_1}^{t_1 + \sigma} \|B(\tau)\|^2 d\tau + \int_{t_1 + \sigma}^{t_1 + 2\sigma} \|B(\tau)\|^2 d\tau \\ &\quad + \dots + \int_{t_1 + i\sigma}^{t_1 + (i+1)\sigma} \|B(\tau)\|^2 d\tau \end{aligned}$$

$$\begin{aligned}
&\leq (i+1)c_1 \\
&\leq c_1\{1+(t_2-t_1)/\sigma\} \\
&\triangleq c_1+c_2(t_2-t_1). \qquad \text{Q.E.D.}
\end{aligned}$$

LEMMA 3.2. Let $L(\cdot)$ be an $n \times n$ measurable and bounded matrix function. If (S'_o) is uniformly completely controllable, then (S'_c) of the form

$$(26) \quad \dot{x}(t) = \{A(t) + B(t)B'(t)L(t)\}x(t) + B(t)v(t)$$

is also uniformly completely controllable.

Proof. From Theorem 4 of [15], it suffices to show that there is a positive number c_3 such that

$$\int_t^{t+\sigma} \|B'(\tau)L(\tau)\|^2 d\tau \leq c_3 \quad \text{for all } t.$$

Let $\|L(t)\| \leq k_1$ for all t . Then, from (25),

$$\begin{aligned}
\int_t^{t+\sigma} \|B'(\tau)L(\tau)\|^2 d\tau &\leq k_1^2 \int_t^{t+\sigma} \|B(\tau)\|^2 d\tau \\
&\leq k_1^2 c_1 \triangleq c_3 \quad \text{for all } t. \qquad \text{Q.E.D.}
\end{aligned}$$

Proof of Theorem 3.5. Assume that (S'_o) is uniformly completely controllable. Then, from Corollary 3.3, for any real number M and any positive number ε , there are a feedback gain $K_1(\cdot)$ and a positive number a_1 such that any zero-input response of

$$(27) \quad \dot{x}(t) = \{A(t) + B(t)K_1(t)\}x(t) + B(t)v(t)$$

satisfies

$$a_1\|x(t_1)\| \exp[(M+\varepsilon)(t_2-t_1)] \leq \|x(t_2)\|$$

for all t_1 and $t_2 \geq t_1$, or equivalently the transition matrix $\Phi_1(\cdot, \cdot)$ associated with $\dot{x}(t) = \{A(t) + B(t)K_1(t)\}x(t)$ satisfies

$$(28) \quad \|\Phi_1(t_1, t_2)\| \leq (1/a_1) \exp[-(M+\varepsilon)(t_2-t_1)]$$

for all t_1 and $t_2 \geq t_1$. Moreover, from Lemma 3.2, (27) can be made uniformly completely controllable by adopting the feedback gain defined as (21); there are positive numbers σ_1 , $\tilde{\alpha}_1$ and $\tilde{\alpha}_2$ such that

$$0 < \tilde{\alpha}_1 I \leq W_1(t, t + \sigma_1) \leq \tilde{\alpha}_2 I$$

holds for all t , where $W_1(\cdot, \cdot)$ is the controllability Gramian of (27), i.e.,

$$W_1(t, t + \sigma_1) = \int_t^{t+\sigma_1} \Phi_1(t, \tau) B(\tau) B'(\tau) \Phi_1'(t, \tau) d\tau.$$

We shall first consider the case $m < M$. Define a feedback gain for real numbers M and $m < M$ by

$$(29) \quad K(t) = -\frac{1}{2}B'(t)\tilde{W}_1^{-1}(t, t + \sigma'),$$

where

$$\tilde{W}_1(t, t + \sigma') \triangleq \int_t^{t+\sigma'} \Phi_1(t, \tau) B(\tau) B'(\tau) \Phi_1'(t, \tau) \exp[-2M(t - \tau)] d\tau$$

and

$$\sigma' \triangleq \max \left\{ \sigma_1, \frac{1}{2\varepsilon} \log \frac{c_2 \exp[2|M|\sigma_1]}{2a_1^2 \tilde{\alpha}_1 (M - m)} \right\}.$$

Here it is notable that for any $\sigma' (\geq \sigma_1)$,

$$(30) \quad \begin{aligned} 0 < c_4 I &\leq \tilde{W}_1(t, t + \sigma') \leq c_5 I, \\ c_4 &\triangleq \tilde{\alpha}_1 \exp[-2|M|\sigma_1], \quad c_5 \triangleq c_1 / (a_1^2 (1 - \exp[-2\varepsilon\sigma])) \end{aligned}$$

holds for all t , which will be shown in the last part of this proof. Performing state-variable feedback on (27) by the feedback gain (29) yields the closed loop system

$$(31) \quad \dot{x}(t) = \{A(t) + B(t)K_1(t) + B(t)K_2(t)\}x(t) + B(t)v(t).$$

Let a scalar function be defined by

$$(32) \quad V(x, t) = x' \tilde{W}_1^{-1}(t, t + \sigma') x.$$

Then

$$(33) \quad \frac{1}{c_5} \|x\|^2 \leq V(x, t) \leq \frac{1}{c_4} \|x\|^2 \quad \text{for all } t,$$

and the time derivative along any zero-input response of (31) is

$$\begin{aligned} \dot{V}(x(t), t) &= x'(t) \{2M \tilde{W}_1^{-1}(t, t + \sigma') - \tilde{W}_1^{-1}(t, t + \sigma') \Phi_1(t, t + \sigma') \\ &\quad \cdot B(t + \sigma') B'(t + \sigma') \Phi_1'(t, t + \sigma') \tilde{W}_1^{-1}(t, t + \sigma') \\ &\quad \cdot \exp[2M\sigma']\} x(t) \quad \text{at almost every } t. \end{aligned}$$

According to (28) and (30), this implies that

$$\left\{ 2M - \frac{\exp[-2\varepsilon\sigma']}{a_1^2 c_4} \|B(t + \sigma')\|^2 \right\} V(x(t), t) \leq \dot{V}(x(t), t) \leq 2M V(x(t), t)$$

holds for almost all t . From the second inequality and (33), it is obvious that any zero-input response of (31) satisfies

$$\|x(t_2)\| \leq \sqrt{(c_5/c_4)} \|x(t_1)\| \exp[M(t_2 - t_1)]$$

for all t_1 and $t_2 \geq t_1$. On the other hand, the first inequality and Lemma 3.1 imply that

$$\begin{aligned} V(x(t_2), t_2) &\geq V(x(t_1), t_1) \exp \left[2M(t_2 - t_1) - \frac{\exp[-2\varepsilon\sigma']}{a_1^2 c_4} \int_{t_1 + \sigma'}^{t_2 + \sigma'} \|B(\tau)\|^2 d\tau \right] \\ &\geq \exp \left[-\frac{c_1 \exp[-2\varepsilon\sigma']}{a_1^2 c_4} \right] V(x(t_1), t_1) \end{aligned} \quad (\text{continued})$$

$$\cdot \exp \left[2 \left(M - \frac{c_2 \exp [-2\varepsilon\sigma']}{2a_1^2 c_4} \right) (t_2 - t_1) \right]$$

holds for all t_1 and $t_2 \geq t_1$. Since the definition of σ' means that

$$M - \frac{c_2 \exp [-2\varepsilon\sigma']}{2a_1^2 c_4} \geq m,$$

the above inequality is reduced to

$$V(x(t_2), t_2) \geq \exp \left[-\frac{c_1 \exp [-2\varepsilon\sigma']}{a_1^2 c_4} \right] V(x(t_1), t_1) \exp [2m(t_2 - t_1)].$$

Consequently, any zero-input response of (31) satisfies

$$\sqrt{\frac{c_4}{c_5}} \exp \left[-\frac{c_1 \exp [-2\varepsilon\sigma']}{a_1^2 c_4} \right] \|x(t_1)\| \exp [m(t_2 - t_1)] \leq \|x(t_2)\|$$

for all t_1 and $t_2 \geq t_1$.

Next, we consider the case $m = M$. In this case, $K_2(\cdot)$ is defined by

$$(34) \quad K_2(t) = -\frac{1}{2}B'(t)\tilde{W}_1^{-1}(t, \infty)$$

instead of (29). Since $\tilde{W}_1(t, s) \leq c_5 I$ for all $s \geq t$ and $\tilde{W}_1(t, s_1) \leq \tilde{W}_1(t, s_2)$ for $s_1 \leq s_2$, $\tilde{W}_1(t, \infty)$ exists and

$$0 < c_4 I \leq \tilde{W}_1(t, \infty) \leq c_5 I \quad \text{for all } t.$$

Let a scalar function be defined by

$$(35) \quad V(x, t) = x' \tilde{W}_1^{-1}(t, \infty) x$$

instead of (32). Then, (33) holds for all t , and the time derivative along any zero-input response of (31) is

$$\dot{V}(x(t), t) = 2MV(x(t), t) = 2mV(x(t), t)$$

at almost every t . Hence, by means of Lyapunov's second method, any zero-input response of (31) satisfies

$$\begin{aligned} \sqrt{(c_4/c_5)} \|x(t_1)\| \exp [m(t_2 - t_1)] &\leq \|x(t_2)\| \\ &\leq \sqrt{(c_5/c_4)} \|x(t_1)\| \exp [M(t_2 - t_1)] \end{aligned}$$

for all t_1 and $t_2 \geq t_1$.

Therefore, if (S'_0) is uniformly completely controllable, it is uniformly completely designable.

Now, we must show that (30) holds for all t . Since $\sigma' \geq \sigma_1$,

$$\tilde{W}_1(t, t + \sigma') \geq \tilde{W}_1(t, t + \sigma_1) \geq \tilde{\alpha}_1 \exp [-2|M|\sigma_1] I;$$

the lower bound is obtained. On the other hand, from (28),

$$\begin{aligned}\tilde{W}_1(t, t + \sigma') &\leq \int_t^\infty \|B(\tau)\|^2 \|\Phi_1(t, \tau)\|^2 \exp[-2M(t - \tau)] d\tau I \\ &\leq \int_t^\infty \frac{1}{a_1^2} \|B(\tau)\|^2 \exp[-2\varepsilon(\tau - t)] d\tau I.\end{aligned}$$

Note (25), and we can obtain the upper bound as follows:

$$\begin{aligned}\tilde{W}_1(t, t + \sigma') &\leq \frac{1}{a_1^2} \left\{ \int_t^{t+\sigma} \|B(\tau)\|^2 \exp[-2\varepsilon(\tau - t)] d\tau \right. \\ &\quad + \int_{t+\sigma}^{t+2\sigma} \|B(\tau)\|^2 \exp[-2\varepsilon(\tau - t)] d\tau \\ &\quad + \int_{t+2\sigma}^{t+3\sigma} \|B(\tau)\|^2 \exp[-2\varepsilon(\tau - t)] d\tau \\ &\quad \left. + \dots \right\} I \\ &\leq \frac{1}{a_1^2} \left\{ \int_t^{t+\sigma} \|B(\tau)\|^2 d\tau \right. \\ &\quad + \exp[-2\varepsilon\sigma] \int_{t+\sigma}^{t+2\sigma} \|B(\tau)\|^2 d\tau \\ &\quad + \exp[-4\varepsilon\sigma] \int_{t+2\sigma}^{t+3\sigma} \|B(\tau)\|^2 d\tau \\ &\quad \left. + \dots \right\} I \\ &\leq \frac{c_1}{a_1^2} \{1 + \exp[-2\varepsilon\sigma] + \exp[-4\varepsilon\sigma] + \dots\} I \\ &= \frac{c_1}{a_1^2(1 - \exp[-2\varepsilon\sigma])} I.\end{aligned}$$

Here, note that the upper and lower bounds are independent of σ' . The proof is completed. Q.E.D.

Proof of Theorem 3.6. Necessity. If a bounded system (S'_0) is uniformly completely designable by a bounded feedback gain, then, evidently, it is uniformly completely stabilizable by a bounded feedback gain. Therefore, from Corollary 3.2, it is uniformly completely controllable.

Sufficiency. The proof of Theorem 3.5 applies here. In this case, since $B(\cdot)$ is bounded, $K_1(\cdot)$ defined as (21) and $K_2(\cdot)$ defined by (29) or (34) are bounded. Hence $K(\cdot) \triangleq K_1(\cdot) + K_2(\cdot)$ is also bounded. Q.E.D.

4. Deterministic state estimation. In this section, the state of the open loop system

$$(S_o) \quad \begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)u(t), \\ y(t) &= C(t)x(t) + D(t)u(t) \end{aligned}$$

is estimated by a dynamical estimator. We assume that the estimator belongs to the class of n -dimensional linear time-varying systems of the form

$$(E_s) \quad \dot{z}(t) = F(t)z(t) + G(t)u(t) + H(t)y(t),$$

where $z(\cdot)$ is the state estimate n -vector, and matrix functions $F(\cdot)$, $G(\cdot)$ and $H(\cdot)$ are respectively $n \times n$, $n \times r$ and $n \times m$, measurable and bounded on every finite subinterval of time. It is known [16] that $F(\cdot)$ and $G(\cdot)$ should be defined by

$$(36) \quad F(t) = A(t) - H(t)C(t),$$

$$(37) \quad G(t) = B(t) - H(t)D(t).$$

This is due to the natural requirement that if the estimate error $e(\cdot) \triangleq x(\cdot) - z(\cdot)$ is 0 at some time t_0 , it should continue to be 0 at every $t \geq t_0$ independently of $x(\cdot)$ and $u(\cdot)$. In consequence, the design parameter of (E_s) is only $H(\cdot)$, and the error behavior is governed by the homogeneous equation

$$(S_e) \quad \dot{e}(t) = \{A(t) - H(t)C(t)\}e(t).$$

That is, the estimate property of (E_s) is determined by the stability and instability degrees of (S_e) .

The definitions of detectability, anticausal-detectability and estimatability follow the above argument.

DEFINITION 4.1. (S_o) is said to be *uniformly completely detectable* if, for any real number M , there are a positive number b and an estimator gain $H(\cdot)$ such that any solution of (S_e) satisfies

$$(38) \quad \|e(t_2)\| \leq b\|e(t_1)\| \exp [M(t_2 - t_1)] \quad \text{for all } t_1 \text{ and } t_2 \geq t_1.$$

DEFINITION 4.2. (S_o) is said to be *uniformly completely anticausal-detectable* if, for any real number m , there are a positive number a and an estimator gain $H(\cdot)$ such that any solution of (S_e) satisfies

$$(39) \quad a\|e(t_1)\| \exp [m(t_2 - t_1)] \leq \|e(t_2)\| \quad \text{for all } t_1 \text{ and } t_2 \geq t_1.$$

DEFINITION 4.3. (S_o) is said to be *uniformly completely estimatable* if, for any pair of real numbers m and M such that $m \leq M$, there are positive numbers a, b and an estimator gain $H(\cdot)$ such that any solution of (S_e) satisfies

$$(40) \quad a\|e(t_1)\| \exp [m(t_2 - t_1)] \leq \|e(t_2)\| \leq b\|e(t_1)\| \exp [M(t_2 - t_1)]$$

for all t_1 and $t_2 \geq t_1$.

Among detectability, anticausal-detectability and estimatability, the relations analogous to those of stabilizability, anticausal-stabilizability and designability hold. (See Remarks 3.2 and 3.3.)

4.1. Relation between deterministic state estimation and state-variable feedback. As considered in § 3, the state-variable feedback problem is the problem of constructing a feedback gain $K(\cdot)$ for given $A(\cdot)$ and $B(\cdot)$ such that

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

has the prescribed stability property. We may expect that this problem has close relevance with the deterministic state estimation problem, in which the design parameter is the estimator gain $H(\cdot)$ of

$$\dot{e}(t) = \{A(t) - H(t)C(t)\}e(t)$$

and $A(\cdot)$ and $C(\cdot)$ are given. Indeed, it is known that these problems are equivalent for a time-invariant case [6]; detectability, anticausal-detectability and estimatability of

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t) \end{aligned} \quad (\tilde{S}_o)$$

respectively correspond to stabilizability, anticausal-stabilizability and designability of the dual system

$$\begin{aligned} \dot{x}(t) &= A'x(t) + C'u(t), \\ y(t) &= B'x(t) + D'u(t) \end{aligned} \quad (\tilde{D}_u)$$

since the stability property of

$$\dot{e}(t) = (A - HC)e(t)$$

is identical with that of its dual system

$$\dot{x}(t) = (A' - C'H')x(t)$$

when A , C and H are constant.

For a time-varying case, such equivalence is generally invalid because the stability property of $\dot{x}(t) = S(t)x(t)$ does not have direct relevance with that of its dual system $\dot{x}(t) = S'(-t)x(t)$ as shown below in Example 4.1. However, concerning our problems considered in this paper, it is valid; the estimation problem in which the stability property of the estimate error is specified by exponential functions can be reduced to the state-variable feedback problem for the dual system

$$\begin{aligned} \dot{x}(t) &= A'(-t)x(t) + C'(-t)u(t), \\ y(t) &= B'(-t)x(t) + D'(-t)u(t). \end{aligned} \quad (D_u)$$

This is implied by the following lemma.

LEMMA 4.1. *Every solution of $\dot{x}(t) = S(t)x(t)$ satisfies*

$$\|x(t_2)\| \leq b\|x(t_1)\| \exp [M(t_2 - t_1)]$$

for all t_1 and $t_2 \geq t_1$ if and only if every solution of $\dot{x}(t) = S'(-t)x(t)$ satisfies the same inequality for all t_1 and $t_2 \geq t_1$. Analogously, every solution of $\dot{x}(t) = S(t)x(t)$

satisfies

$$a\|x(t_1)\| \exp[m(t_2 - t_1)] \leq \|x(t_2)\|$$

for all t_1 and $t_2 \geq t_1$ if and only if every solution of $\dot{x}(t) = S'(-t)x(t)$ satisfies the same inequality for all t_1 and $t_2 \geq t_1$.

Proof. It immediately follows from the relation between the transition matrices of $\dot{x}(t) = S(t)x(t)$ and $\dot{x}(t) = S'(-t)x(t)$ (see [11]). Q.E.D.

This lemma indicates that the stability and instability degrees of

$$(S_e) \quad \dot{e}(t) = \{A(t) - H(t)C(t)\}e(t)$$

are identical with those of its dual system

$$(41) \quad \dot{x}(t) = \{A'(-t) - C'(-t)H'(-t)\}x(t)$$

respectively. Thus we obtain the following lemmas.

LEMMA 4.2. (S_o) is uniformly completely detectable if and only if (D_u) is uniformly completely stabilizable.

LEMMA 4.3. (S_o) is uniformly completely anticausal-detectable if and only if (D_u) is uniformly completely anticausal-stabilizable.

LEMMA 4.4. (S_o) is uniformly completely estimatable if and only if (D_u) is uniformly completely designable.

Example 4.1. Consider a linear time-varying system

$$(42a) \quad \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ e^{2t} & -4 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}$$

and its dual system

$$(42b) \quad \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} -1 & e^{-2t} \\ 0 & -4 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}.$$

The transition matrices $\Psi(\cdot, \cdot)$ and $\tilde{\Psi}(\cdot, \cdot)$ associated with (42a) and (42b) respectively are

$$\Psi(t, \tau) = \begin{bmatrix} e^{-(t-\tau)} & 0 \\ \frac{1}{5}e^{t+\tau} - \frac{1}{5}e^{-4t+6\tau} & e^{-4(t-\tau)} \end{bmatrix},$$

$$\tilde{\Psi}(t, \tau) = \begin{bmatrix} e^{-(t-\tau)} & \frac{1}{5}e^{-t-\tau} - \frac{1}{5}e^{-6t+4\tau} \\ 0 & e^{-4(t-\tau)} \end{bmatrix}.$$

This means that (42b) is asymptotically stable while (42a) is not. Therefore, in case of time-varying systems, the stability property of $\dot{x}(t) = S(t)x(t)$ does not imply that the similar property holds with respect to the dual system $\dot{x}(t) = S'(-t)x(t)$.

Remark 4.1. This example implies that the nonexponential state estimation problem [20], i.e., the estimation problem in which the stability property of the estimate error is not specified by exponential functions, cannot be reduced to the state-variable feedback problem for the *dual system*. Instead, it can be reduced to

the state-variable feedback problem for the *nonhomogeneous adjoint system* [11]

$$(A_d) \quad \begin{aligned} \dot{x}(t) &= -A'(t)x(t) + C'(t)u(t), \\ y(t) &= B'(t)x(t) + D'(t)u(t). \end{aligned}$$

In this case, stabilizing the estimate error is equivalent to unstabilizing the closed loop system, and in contrast, unstabilizing the estimate error is equivalent to stabilizing the closed loop system. This fact follows from the relation between the transition matrices of

$$(S_e) \quad \dot{e}(t) = \{A(t) - H(t)C(t)\}e(t)$$

and its adjoint system

$$\dot{x}(t) = \{-A'(t) + C'(t)H'(t)\}x(t)$$

(which is the closed loop system obtained from (A_d)).

4.2. Results about the deterministic state estimation problem. As discussed above, we have clarified the relation between the deterministic state estimation problem and the state-variable feedback problem (Lemmas 4.2–4.4). As a consequence, the following theorems about the state estimation problem are immediately obtained by reference to the results about the state-variable feedback problem (Theorems 3.1–3.6) and Lemmas 2.1–2.3.

THEOREM 4.1. *If (S_o) is uniform with respect to complete reconstructibility, then it is uniformly completely detectable.*

THEOREM 4.2. *A bounded system (S_o) is uniformly completely detectable by a bounded estimator if and only if it is uniform with respect to complete reconstructibility.*

THEOREM 4.3. *If (S_o) is uniform with respect to complete observability, then it is uniformly completely anticausal-detectable.*

THEOREM 4.4. *A bounded system (S_o) is uniformly completely anticausal-detectable by a bounded estimator if and only if it is uniform with respect to complete observability.*

THEOREM 4.5. *If (S_o) is uniformly completely observable, then it is uniformly completely estimatable.*

THEOREM 4.6. *A bounded system (S_o) is uniformly completely estimatable by a bounded estimator if and only if it is uniformly completely observable.*

Remark 4.2. Theorems 4.1–4.4 indicate that the uniform asymptotic state estimation problem has closer relevance to uniformity with respect to complete reconstructibility than uniformity with respect to complete observability. This is due to the facts that the asymptotic state estimation problem is the problem of estimating the present state from past data (i.e., the past inputs and outputs of the system to be estimated) and reconstructibility is the concept which guarantees the possibility of determining the present state from past data, while observability is the concept which guarantees the possibility of determining the present state from future data [1].

The proofs of these theorems are straightforward. For example, the proof of Theorem 4.1 is as follows. Assume that (S_o) is uniform with respect to complete reconstructibility. Then (D_u) is uniform with respect to complete controllability

(Lemma 2.1), and hence (D_u) is uniformly completely stabilizable (Theorem 3.1). This implies that (S_o) is uniformly completely detectable (Lemma 4.2).

The other theorems can be proved analogously. In the proofs of Theorems 4.2, 4.4 and 4.6 where (S_o) is bounded, note that the boundedness of the estimator gain $H(\cdot)$ implies the estimator (E_s) is bounded since $F(\cdot)$ and $G(\cdot)$ are defined by (36) and (37). (It is obvious that the boundedness of (E_s) implies that of $H(\cdot)$.)

Also, corresponding to Corollaries 3.1–3.4, the following corollaries are obtained.

COROLLARY 4.1. *If (S_o) is uniformly completely observable, then it is uniformly completely detectable.*

COROLLARY 4.2. *A bounded system (S_o) is uniformly completely detectable by a bounded estimator if and only if it is uniformly completely observable.*

COROLLARY 4.3. *If (S_o) is uniformly completely observable, then it is uniformly completely anticausal-detectable.*

COROLLARY 4.4. *A bounded system (S_o) is uniformly completely anticausal-detectable by a bounded estimator if and only if it is uniformly completely observable.*

5. Feedback of the state estimate. The state estimate $z(\cdot)$ produced by the estimator (E_s) is now substituted for the real state $x(\cdot)$ into the control law (13). In this case, a question arises as to whether the stability and instability degrees of the total closed loop system can still be specified. Thus, the problem we shall consider is that of designing the internal stability property of the resultant $2n$ -dimensional system

$$(S_c) \quad \begin{bmatrix} \dot{x}(t) \\ \dot{z}(t) \end{bmatrix} = \begin{bmatrix} A(t) & B(t)K(t) \\ H(t)C(t) & A(t) - H(t)C(t) + B(t)K(t) \end{bmatrix} \begin{bmatrix} x(t) \\ z(t) \end{bmatrix} + \begin{bmatrix} B(t) \\ B(t) \end{bmatrix} v(t),$$

$$y(t) = [C(t) \quad D(t)K(t)] \begin{bmatrix} x(t) \\ z(t) \end{bmatrix} + D(t)v(t).$$

For a time-invariant case, this problem has been solved. That is, if (S_o) is (uniformly completely) controllable and (uniformly completely) observable, then n poles of (S_c) can be placed arbitrarily in complex-conjugate pairs and real numbers by the feedback gain $K(\cdot)$ and other n poles can be similarly assigned by the estimator gain $H(\cdot)$; the converse is also valid.

For a time-varying case, analogous results are obtained.

THEOREM 5.1. *If (S_o) is both uniformly completely controllable and uniformly completely observable, then for any pair of real numbers m and M such that $m < M$, there exist positive numbers a and b , a feedback gain $K(\cdot)$ and an estimator gain $H(\cdot)$ such that any zero-input state response of (S_c) satisfies*

$$(43) \quad a \left\| \begin{bmatrix} x(t_1) \\ z(t_1) \end{bmatrix} \right\| \exp [m(t_2 - t_1)] \leq \left\| \begin{bmatrix} x(t_2) \\ z(t_2) \end{bmatrix} \right\|$$

$$\leq b \left\| \begin{bmatrix} x(t_1) \\ z(t_1) \end{bmatrix} \right\| \exp [M(t_2 - t_1)]$$

for all t_1 and $t_2 \geq t_1$.

THEOREM 5.2. *Let (S_o) be a bounded system. Then for any pair of real numbers m and M such that $m < M$, there exist positive numbers a and b , a bounded feedback gain $K(\cdot)$ and a bounded estimator gain $H(\cdot)$ such that any zero-input state response of (S_o) satisfies (43) for all t_1 and $t_2 \geq t_1$ if and only if (S_o) is both uniformly completely controllable and uniformly completely observable.*

Remark 5.1. Note that the case $m = M$ is excluded in Theorems 5.1 and 5.2. (Compare with Theorems 3.5 and 3.6.) This situation can be easily resolved on investigating a time-invariant case. Consider a linear time-invariant controllable and observable system whose dimension n is odd. If we desire to attain $m = M$ in (S_o) , two poles among $2n$ poles should be placed at $s = M$ on the real axis in the complex s -plane, thereby producing a multiple pole. Hence it may have a basis function $t \exp [Mt]$ which decays more slowly than $\exp [Mt]$ in case $M < 0$ or grows faster than $\exp [Mt]$ in case $M \geq 0$.

Proof of Theorem 5.1. First, note that the homogeneous system

$$(44) \quad \begin{bmatrix} \dot{x}(t) \\ \dot{z}(t) \end{bmatrix} = \begin{bmatrix} A(t) & B(t)K(t) \\ H(t)C(t) & A(t) - H(t)C(t) + B(t)K(t) \end{bmatrix} \begin{bmatrix} x(t) \\ z(t) \end{bmatrix}$$

can be reduced to a unilateral composite system

$$(45) \quad \begin{bmatrix} \dot{x}(t) \\ \dot{e}(t) \end{bmatrix} = \begin{bmatrix} A(t) + B(t)K(t) & -B(t)K(t) \\ 0 & A(t) - H(t)C(t) \end{bmatrix} \begin{bmatrix} x(t) \\ e(t) \end{bmatrix}$$

by the nonsingular transformation

$$(46) \quad \begin{bmatrix} x(t) \\ e(t) \end{bmatrix} = \begin{bmatrix} I & 0 \\ I & -I \end{bmatrix} \begin{bmatrix} x(t) \\ z(t) \end{bmatrix}.$$

Since (46) is a Lyapunov transformation, the stability and instability degrees of (44) are identical with those of (45). So, we shall consider (45) because of its simpler structure. Let the transition matrices associated with

$$(S_c'') \quad \dot{x}(t) = \{A(t) + B(t)K(t)\}x(t)$$

and

$$(S_e) \quad \dot{e}(t) = \{A(t) - H(t)C(t)\}e(t)$$

be denoted by $\hat{\Phi}(\cdot, \cdot)$ and $\tilde{\Phi}(\cdot, \cdot)$ respectively. Then, the transition matrix $\Phi_T(\cdot, \cdot)$ associated with (45) can be represented as

$$(47) \quad \Phi_T(t_2, t_1) = \begin{bmatrix} \hat{\Phi}(t_2, t_1) & -\int_{t_1}^{t_2} \hat{\Phi}(t_2, \tau)B(\tau)K(\tau)\tilde{\Phi}(\tau, t_1) d\tau \\ 0 & \tilde{\Phi}(t_2, t_1) \end{bmatrix}$$

and its norm satisfies

$$(48) \quad \begin{aligned} \|\Phi_T(t_2, t_1)\| &\leq \|\hat{\Phi}(t_2, t_1)\| + \|\tilde{\Phi}(t_2, t_1)\| \\ &+ \left\| \int_{t_1}^{t_2} \hat{\Phi}(t_2, \tau)B(\tau)K(\tau)\tilde{\Phi}(\tau, t_1) d\tau \right\| \end{aligned}$$

for all t_1 and t_2 .

Assume that (S_0) is both uniformly completely controllable and uniformly completely observable. Then, Theorems 3.5 and 4.5 guarantee that for any pair of real numbers m and M such that $m < M$ and any positive number $\varepsilon \leq (M - m)/2$, there exist positive numbers a_1 , a_2 , b_1 and b_2 , a feedback gain $K(\cdot)$ and an estimator gain $H(\cdot)$ such that any solution $x(\cdot)$ of (S'_e) and any solution $e(\cdot)$ of (S_e) satisfy

$$\begin{aligned} a_1 \|x(t_1)\| \exp[(m + \varepsilon)(t_2 - t_1)] &\leq \|x(t_2)\| \\ &\leq b_1 \|x(t_1)\| \exp[(M - \varepsilon)(t_2 - t_1)] \end{aligned}$$

and

$$\begin{aligned} a_2 \|e(t_1)\| \exp[(m + \varepsilon)(t_2 - t_1)] &\leq \|e(t_2)\| \\ &\leq b_2 \|e(t_1)\| \exp[(M - \varepsilon)(t_2 - t_1)] \end{aligned}$$

for all t_1 and $t_2 \geq t_1$ respectively, or equivalently,

$$(49a) \quad \|\hat{\Phi}(t_1, t_2)\| \leq (1/a_1) \exp[-(m + \varepsilon)(t_2 - t_1)],$$

$$(49b) \quad \|\hat{\Phi}(t_2, t_1)\| \leq b_1 \exp[(M - \varepsilon)(t_2 - t_1)],$$

$$(49c) \quad \|\tilde{\Phi}(t_1, t_2)\| \leq (1/a_2) \exp[-(m + \varepsilon)(t_2 - t_1)],$$

$$(49d) \quad \|\tilde{\Phi}(t_2, t_1)\| \leq b_2 \exp[(M - \varepsilon)(t_2 - t_1)]$$

hold for all t_1 and $t_2 \geq t_1$. Moreover, as adopted in the proof of Theorem 3.5, we can employ a feedback gain of the form $K(\cdot) = B'(\cdot)L(\cdot)$, where $L(\cdot)$ is bounded. Let $\|L(t)\| \leq k_2$ for all t . Then, from Lemma 3.1,

$$\begin{aligned} &\left\| \int_{t_1}^{t_2} \hat{\Phi}(t_2, \tau) B(\tau) K(\tau) \tilde{\Phi}(\tau, t_1) d\tau \right\| \\ &\leq b_1 b_2 k_2 \exp[(M - \varepsilon)(t_2 - t_1)] \int_{t_1}^{t_2} \|B(\tau)\|^2 d\tau \\ &\leq b_1 b_2 k_2 \exp[M(t_2 - t_1)] \exp[-\varepsilon(t_2 - t_1)] \{c_1 + c_2(t_2 - t_1)\} \\ &\leq b_1 b_2 k_2 (c_1 + c_2 \varepsilon^{-1} e^{-1}) \exp[M(t_2 - t_1)] \\ &\triangleq b_3 \exp[M(t_2 - t_1)] \end{aligned}$$

for all t_1 and $t_2 \geq t_1$. Consequently, from (48), (49b) and (49d),

$$(50) \quad \|\Phi_T(t_2, t_1)\| \leq (b_1 + b_2 + b_3) \exp[M(t_2 - t_1)]$$

holds for all t_1 and $t_2 \geq t_1$. This and (46) imply that any solution of (44) satisfies

$$(51) \quad \left\| \begin{bmatrix} x(t_2) \\ z(t_2) \end{bmatrix} \right\| \leq 4(b_1 + b_2 + b_3) \left\| \begin{bmatrix} x(t_1) \\ z(t_1) \end{bmatrix} \right\| \exp[M(t_2 - t_1)]$$

for all t_1 and $t_2 \geq t_1$.

On the other hand, we can show analogously that there exists a positive number a_3 such that

$$\left\| \int_{t_2}^{t_1} \hat{\Phi}(t_1, \tau) B(\tau) K(\tau) \tilde{\Phi}(\tau, t_2) d\tau \right\| \leq \frac{1}{a_3} \exp[-m(t_2 - t_1)]$$

holds for all t_1 and $t_2 \geq t_1$. Then, from (48), (49a) and (49c),

$$(52) \quad \|\Phi_T(t_1, t_2)\| \leq \left(\frac{1}{a_1} + \frac{1}{a_2} + \frac{1}{a_3} \right) \exp[-m(t_2 - t_1)]$$

for all t_1 and $t_2 \geq t_1$, and hence, from (46), any solution of (44) satisfies

$$(53) \quad \frac{1}{4} \left(\frac{1}{a_1} + \frac{1}{a_2} + \frac{1}{a_3} \right)^{-1} \left\| \begin{bmatrix} x(t_1) \\ z(t_1) \end{bmatrix} \right\| \exp[m(t_2 - t_1)] \leq \left\| \begin{bmatrix} x(t_2) \\ z(t_2) \end{bmatrix} \right\|$$

for all t_1 and $t_2 \geq t_1$.

Q.E.D.

Proof of Theorem 5.2. Necessity. Assume that, for any pair of real numbers m and M such that $m < M$, there exist positive numbers a and b , a bounded feedback gain $K(\cdot)$ and a bounded estimator gain $H(\cdot)$ such that any zero-input response of (S_c) satisfies (43) for all t_1 and $t_2 \geq t_1$. Then, from (46),

$$\|\Phi_T(t_2, t_1)\| \leq 4b \exp[M(t_2 - t_1)],$$

and hence, from (47),

$$\|\hat{\Phi}(t_2; t_1)\| \leq 4b \exp[M(t_2 - t_1)],$$

$$\|\tilde{\Phi}(t_2, t_1)\| \leq 4b \exp[M(t_2 - t_1)]$$

for all t_1 and $t_2 \geq t_1$. This implies that (S_o) is uniformly completely stabilizable by a bounded feedback gain and uniformly completely detectable by a bounded estimator. Therefore, from Corollaries 3.2 and 4.2, (S_o) is uniformly completely controllable and uniformly completely observable.

Sufficiency. The proof of Theorem 5.1 applies here. In this case, we can employ a bounded feedback gain and a bounded estimator gain since (S_o) is bounded (Theorems 3.6 and 4.6).

Q.E.D.

The following theorems are concerned with the upper bound and the lower bound of the zero-input state response of (S_c) respectively. The proofs are similar to that of Theorem 5.1 and are therefore omitted.

THEOREM 5.3. Suppose that (24) holds for all t_1 and $t_2 \geq t_1$ in (S_o) . If (S_o) is uniform with respect to complete controllability and uniform with respect to complete reconstructibility, then for any real number M , there exist a positive number b , a feedback gain $K(\cdot)$ and an estimator gain $H(\cdot)$ such that any zero-input state response of (S_c) satisfies

$$(54) \quad \left\| \begin{bmatrix} x(t_2) \\ z(t_2) \end{bmatrix} \right\| \leq b \left\| \begin{bmatrix} x(t_1) \\ z(t_1) \end{bmatrix} \right\| \exp[M(t_2 - t_1)]$$

for all t_1 and $t_2 \geq t_1$.

THEOREM 5.4. Suppose that (24) holds for all t_1 and $t_2 \geq t_1$ in (S_o) . If (S_o) is uniform with respect to complete reachability and uniform with respect to complete

observability, then for any real number m , there exist a positive number a , a feedback gain $K(\cdot)$ and an estimator gain $H(\cdot)$ such that any zero-input state response of (S_e) satisfies

$$(55) \quad a \left\| \begin{bmatrix} x(t_1) \\ z(t_1) \end{bmatrix} \right\| \exp [m(t_2 - t_1)] \leq \left\| \begin{bmatrix} x(t_2) \\ z(t_2) \end{bmatrix} \right\|$$

for all t_1 and $t_2 \geq t_1$.

Remark 5.2. In Theorems 5.3 and 5.4, the assumption that (24) holds for all t_1 and $t_2 \geq t_1$ is made to suppress the influence of the estimate error on the stability property. Such an assumption is not necessary in Theorems 5.1 and 5.2 because uniform complete controllability of (S_o) implies (24).

REFERENCES

- [1] R. E. KALMAN, P. L. FALB AND M. A. ARBIB, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1969.
- [2] R. W. BROCKETT, *Poles, zeros, and feedback: State space interpretation*, IEEE Trans. Automatic Control, AC-10 (1965), pp. 129–135.
- [3] V. M. POPOV, *Hyperstability and optimality of automatic systems with several control functions*, Rev. Roumaine. Sci. Tech. Electrotech. Energet., 9 (1964), pp. 629–690.
- [4] B. D. O. ANDERSON AND D. G. LUENBERGER, *Design of multivariable feedback systems*, Proc. IEE, 114 (1967), pp. 395–399.
- [5] W. M. WONHAM, *On pole assignment in multi-input controllable systems*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 660–665.
- [6] J. D. SIMON AND S. K. MITTER, *A theory of modal control*, Information and Control, 13 (1968), pp. 316–353.
- [7] W. A. WOLOVICH, *On the stabilization of controllable systems*, IEEE Trans. Automatic Control, AC-13 (1968), pp. 569–572.
- [8] R. E. KALMAN, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana (2), 5 (1960), pp. 102–119.
- [9] B. D. O. ANDERSON AND J. B. MOORE, *Linear Optimal Control*, Prentice-Hall, Englewood Cliffs, N.J., 1971.
- [10] M. IKEDA, H. MAEDA AND S. KODAMA, *Stabilization of linear systems*, this Journal, 10 (1972), pp. 716–729.
- [11] H. D'ANGELO, *Linear Time-Varying Systems: Analysis and Synthesis*, Allyn and Bacon, Boston, 1970.
- [12] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.
- [13] R. E. KALMAN, Y. C. HO AND K. S. NARENDRA, *Controllability of linear dynamical systems*, Contributions to Differential Equations, 1 (1962), pp. 189–213.
- [14] L. M. SILVERMAN AND B. D. O. ANDERSON, *Controllability, observability and stability of linear systems*, this Journal, 6 (1968), pp. 121–130.
- [15] B. D. O. ANDERSON AND J. B. MOORE, *New results in linear system stability*, this Journal, 7 (1969), pp. 398–414.
- [16] G. W. JOHNSON, *A deterministic theory of estimation and control*, IEEE Trans. Automatic Control, AC-14 (1969), pp. 380–384.
- [17] L. WEISS AND R. E. KALMAN, *Contributions to linear system theory*, Internat. J. Engrg. Sci., 3 (1965), pp. 141–171.
- [18] W. A. WOLOVICH, *On state estimation of observable systems*, Preprints of the 1968 JACC, pp. 210–220.
- [19] Y. Ö. YÜKSEL AND J. J. BONGIORNO, JR., *Observers for linear multivariable systems with applications*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 603–613.
- [20] M. IKEDA, Y. SASANO AND S. KODAMA, *On the state estimators for linear time-varying systems*, Trans. Inst. Electronics Comm. Engrs., Japan, 55-D (1972), pp. 448–455. (In Japanese.)

THE SINGULARLY PERTURBED LINEAR STATE REGULATOR PROBLEM. II*

R. E. O'MALLEY, JR.[†] AND C. F. KUNG[‡]

Abstract. We consider regulator problems for the singularly perturbed system

$$\begin{aligned}\frac{dx}{dt} &= A_1(t, \varepsilon)x + A_2(t, \varepsilon)z + B_1(t, \varepsilon)u, \\ \varepsilon \frac{dz}{dt} &= A_3(t, \varepsilon)x + A_4(t, \varepsilon)z + B_2(t, \varepsilon)u\end{aligned}$$

on $0 \leq t \leq 1$, seeking their asymptotic solution as $\varepsilon \rightarrow 0$. The hypotheses used are simple. Results are related to the current literature.

1. Introduction and main results. Let us consider the linear time-varying state regulator problem consisting of the system of differential equations

$$\begin{aligned}(1.1) \quad \frac{dx}{dt} &= A_1(t, \varepsilon)x + A_2(t, \varepsilon)z + B_1(t, \varepsilon)u, \\ \varepsilon \frac{dz}{dt} &= A_3(t, \varepsilon)x + A_4(t, \varepsilon)z + B_2(t, \varepsilon)u\end{aligned}$$

on the interval $0 \leq t \leq 1$ (or any other closed, bounded interval), the initial states

$$(1.2) \quad x(0, \varepsilon) \quad \text{and} \quad z(0, \varepsilon)$$

being prescribed, and the scalar cost functional

$$\begin{aligned}(1.3) \quad J(\varepsilon) &= \frac{1}{2} \begin{pmatrix} x(1, \varepsilon) \\ z(1, \varepsilon) \end{pmatrix}' \pi(\varepsilon) \begin{pmatrix} x(1, \varepsilon) \\ z(1, \varepsilon) \end{pmatrix} \\ &+ \frac{1}{2} \int_0^1 \left[\begin{pmatrix} x(t, \varepsilon) \\ z(t, \varepsilon) \end{pmatrix}' Q(t, \varepsilon) \begin{pmatrix} x(t, \varepsilon) \\ z(t, \varepsilon) \end{pmatrix} + u'(t, \varepsilon)u(t, \varepsilon) \right] dt\end{aligned}$$

which is to be minimized by selection of the control $u(t, \varepsilon)$. Here x, z and u are vectors of dimension n, m and r , respectively, the prime denotes transposition, π and Q are symmetric, nonnegative definite matrices having the block forms

$$(1.4) \quad Q = \begin{pmatrix} Q_1 & Q_2 \\ Q_2' & Q_3 \end{pmatrix} \quad \text{and} \quad \pi = \begin{pmatrix} \pi_1 & \varepsilon\pi_2 \\ \varepsilon\pi_2' & \varepsilon\pi_3 \end{pmatrix}$$

and ε is a small positive parameter.

Such singular perturbation problems have been of considerable interest in the recent literature and are of significance in practical situations when ε represents

* Received by the editors June 25, 1973, and in revised form October 15, 1973. This research was supported by the Air Force Office of Scientific Research under Grant AFOSR-71-2013 at the Courant Institute, New York University.

[†] Department of Mathematics, University of Arizona, Tucson, Arizona 85721.

[‡] Department of Mathematics, Ladycliff College, Highland Falls, New York 10928.

certain often neglected "parasitic" parameters (cf., for example, Kokotović and Sannuti [8] and Wilde and Kokotović [16]). The object is to obtain asymptotic series expansions for the optimal solution as $\varepsilon \rightarrow 0$. In order to do so, we shall assume that the matrices A_i , B_i , $x(0, \varepsilon)$, $z(0, \varepsilon)$, $\pi_i(\varepsilon)$, and Q_i all have asymptotic series expansions as $\varepsilon \rightarrow 0$ with the expansions for A_i , B_i and Q_i being uniformly valid in $0 \leq t \leq 1$ and with infinitely differentiable coefficients. The discussion follows that of O'Malley [13] with somewhat simpler hypotheses and further discussion of the results, assumptions and related problems.

Our main conclusions are contained in the following theorem.

THEOREM. *Suppose*

- (i) *the matrix $A_4(t, 0)$ is invertible, and*
- (ii) *all eigenvalues of the matrix*

$$(1.5) \quad G(t) = \begin{pmatrix} A_4(t, 0) & -B_2(t, 0)B_2'(t, 0) \\ -Q_3(t, 0) & -A_4'(t, 0) \end{pmatrix}$$

have nonzero real parts throughout $0 \leq t \leq 1$.

- (iii) *Suppose*

$$(1.6) \quad T(t) = \begin{pmatrix} T_{11}(t) & T_{12}(t) \\ T_{21}(t) & T_{22}(t) \end{pmatrix}$$

is a nonsingular matrix such that

$$(1.7) \quad T^{-1}GT = \begin{pmatrix} -\Lambda & 0 \\ 0 & \Lambda \end{pmatrix}$$

with all eigenvalues of $\Lambda(t)$ having positive real parts, and such that the matrices

$$(1.8) \quad T_{11}(0)$$

and

$$(1.9) \quad T_{22}(1) - \pi_3(0)T_{12}(1)$$

are both nonsingular.

- (iv) *Suppose the matrix*

$$(1.10) \quad \begin{aligned} \mathcal{Q}(t, \varepsilon) \equiv & Q_1 - Q_2 A_4^{-1} A_3 - A_3' A_4'^{-1} Q_2' + A_3' A_4'^{-1} Q_3 A_4^{-1} A_3 \\ & - (Q_2 - A_3' A_4'^{-1} Q_3) A_4^{-1} B_2 (I + B_2' A_4'^{-1} Q_3 A_4^{-1} B_2)^{-1} \\ & B_2' A_4'^{-1} (Q_2' - Q_3 A_4^{-1} A_3) \end{aligned}$$

is positive semidefinite throughout $0 \leq t \leq 1$ for $\varepsilon = 0$.

Then the optimal control, the corresponding trajectories, and the optimal cost $J^(\varepsilon)$ satisfy*

$$(1.11) \quad \begin{aligned} u(t, \varepsilon) &= U_0(t) + O(\varepsilon), & 0 < t < 1, \\ x(t, \varepsilon) &= X_0(t) + O(\varepsilon), & 0 \leq t \leq 1, \\ z(t, \varepsilon) &= Z_0(t) + O(\varepsilon), & 0 < t < 1, \\ J^*(\varepsilon) &= J_0^* + O(\varepsilon), \end{aligned}$$

where $U_0(t)$ is the optimal control; $X_0(t)$ and $Z_0(t)$, the corresponding trajectories; and J_0^* the optimal cost for the reduced problem

$$\begin{aligned} \frac{dX_0}{dt} &= A_1(t, 0)X_0 + A_2(t, 0)Z_0 + B_1(t, 0)U_0, \\ (1.12) \quad 0 &= A_3(t, 0)X_0 + A_4(t, 0)Z_0 + B_2(t, 0)U_0, \\ X_0(0) &= x(0, 0) \end{aligned}$$

with

$$\begin{aligned} J(0) &= \frac{1}{2}X_0'(1)\pi_1(0)X_0(1) \\ &+ \frac{1}{2}\int_0^1 \left[\begin{pmatrix} X_0(t) \\ Z_0(t) \end{pmatrix}' Q(t, 0) \begin{pmatrix} X_0(t) \\ Z_0(t) \end{pmatrix} + U_0'(t)U_0(t) \right] dt \end{aligned}$$

to be minimized. This reduced problem has a unique solution. Further, asymptotic expansions of the unique solution of the full problem (1.1)–(1.4) can be obtained which are uniformly valid throughout $0 \leq t \leq 1$.

Remark 1. Asking that A_4 be invertible allows us to obtain Z_0 as a linear function of X_0 and U_0 and, thereby, to identify the limiting solution within $0 < t < 1$ as the solution of a lower order regulator problem. Note that if some eigenvalues of A_4 have positive real parts, practical instability problems could result in implementing this asymptotic analysis (cf. Wilde and Kokotović [17]). The invertibility of A_4 is not necessary, however, to obtain a solution (cf. O'Malley [13]).

Remark 2. We note that $-\lambda$ is an eigenvalue of G whenever λ is (cf. O'Donnell [10] and Anderson and Moore [1]). Thus assumption (ii) implies the essential result that G is nonsingular. Likewise, assumption (ii) would be implied by the existence of G^{-1} provided G had no purely imaginary eigenvalues (cf. exercise 15.2–4 of Anderson and Moore). Finally, we observe that Wilde and Kokotović achieved assumption (ii) by their “boundary layer controllability and observability” hypotheses and that the results of Hoppensteadt [6] indicate that assumption (ii) might be slightly weakened.

Remark 3. The matrices (1.8) and (1.9) are somewhat familiar from the time-invariant regulator problem (cf. Anderson and Moore [1, p. 352], where T_{11} is always nonsingular). Since the columns of T are (generalized) eigenvectors of G , the hypotheses will either hold (or fail to hold) for all choices of T (i.e., whatever the normalization). We note that assumption (iii) corresponds to (H4) of O'Malley [13] and, finally, that the matrix $\Lambda(t)$ is not necessarily diagonal.

Remark 4. It is our conjecture that assumption (iv) follows from the non-negative definiteness of Q and the invertibility of A_4 . Indeed, when Q_3 is nonsingular, we have

$$\begin{aligned} \mathcal{Q} &= Q_1 - Q_2 Q_3^{-1} Q_2' + (A_3' A_4'^{-1} Q_3 - Q_2) \\ &\cdot (Q_3^{-1} + A_4^{-1} B_2 B_2' A_4'^{-1})^{-1} Q_3^{-1} (Q_3 A_4^{-1} A_3 - Q_2') \end{aligned}$$

positive semidefinite. Likewise, the assumption holds when either $Q_3 = 0$, $B_2 B_2'$ is invertible, or $B_2 = 0$. Kokotović and Yackel [9] show that (iv) follows from

the existence of $(A'_{40}K + Q_{30})^{-1}$, where $K = T_{21}T_{11}^{-1}$.

Remark 5. In O'Malley [13], an additional hypothesis eliminating turning points was introduced. This allowed us to use a convenient set of asymptotic solutions. This technical assumption is eliminated in this presentation.

2. The reduced problem. We note that the reduced problem (1.12) which would be naturally considered by a control engineer who neglects small parasitics is also the natural reduced (or "outer") problem of singular perturbation theory (cf. Wasow [16] or O'Malley [15]).

Since $A_{40} = A_4(t, 0)$ is invertible, we have

$$(2.1) \quad Z_0(t) = -A_{40}^{-1}A_{30}X_0 - A_{40}^{-1}B_{20}U_0.$$

Then introducing

$$(2.2) \quad V_0(t) = U_0 - \bar{R}^{-1}\bar{C}'X_0,$$

where $\bar{R} = I + B'_{20}A_{40}^{-1}Q_{30}A_{40}^{-1}B_{20}$ and $\bar{C} = (Q_{20} - A'_{30}A_{40}^{-1}Q_{30})A_{40}^{-1}B_{20}$, (X_0, V_0) will satisfy the linear regulator problem

$$(2.3) \quad \frac{dX_0}{dt} = \alpha(t)X_0 + \bar{B}(t)V_0, \quad X_0(0) = x(0, 0)$$

with

$$J(0) = \frac{1}{2}X'_0(1)\pi_1(0)X_0(1) + \frac{1}{2} \int_0^1 [X'_0(t)\mathcal{Q}(t, 0)X_0(t) + V'_0(t)\bar{R}(t)V_0(t)] dt,$$

where $\alpha = \bar{A} + \bar{B}\bar{R}^{-1}\bar{C}'$, $\bar{A} = A_{10} - A_{20}A_{40}^{-1}A_{30}$, $\bar{B} = B_{10} - A_{20}A_{40}^{-1}B_{20}$, and \mathcal{Q} is the positive semidefinite matrix of (1.10). Since $\pi_1(0)$ is positive semidefinite and $\bar{R}(t)$ positive definite, it is well known that this problem has a unique solution $(V_0(t), X_0(t), J_0^*)$ (cf. Kalman [7] and, for example, Anderson and Moore [1]). We have thus obtained the following lemma.

LEMMA. *The solution (U_0, X_0, Z_0, J_0^*) of the reduced problem (1.12) is uniquely determined by the solution (V_0, X_0, J_0^*) of the linear regulator problem (2.3) while U_0 and Z_0 are determined by the linear equations (2.2) and (2.1).*

We note that Haddad and Kokotović [4] previously examined the reduced problem by the Riccati matrix approach.

3. The asymptotic expansions.

3.1. Reformulation of the problem. To obtain necessary and sufficient conditions for an optimal control, we introduce the Hamiltonian

$$(3.1) \quad H(x, z, u, p_1, p_2, t, \varepsilon) = \frac{1}{2}(x'Q_1x + 2x'Q_2z + z'Q_3z + u'u) + p'_1(A_1x + A_2z + B_1u) + p'_2(A_3x + A_4z + B_2u),$$

where the costate vectors p_1 and εp_2 satisfy the terminal value problem consisting of the linear system

$$(3.2) \quad \frac{dp_1}{dt} = -\frac{\partial H}{\partial x}, \quad \varepsilon \frac{dp_2}{dt} = -\frac{\partial H}{\partial z}$$

on $0 \leq t \leq 1$ and the terminal conditions

$$(3.3) \quad \begin{aligned} p_1(1, \varepsilon) &= \pi_1(\varepsilon)x(1, \varepsilon) + \varepsilon\pi_2(\varepsilon)z(1, \varepsilon), \\ p_2(1, \varepsilon) &= \pi'_2(\varepsilon)x(1, \varepsilon) + \pi_3(\varepsilon)z(1, \varepsilon). \end{aligned}$$

Using elementary calculus of variations (cf. Anderson and Moore [1]), we find that along an optimal trajectory we have $\partial H/\partial u = 0$ and this implies the control relation

$$(3.4) \quad u(t, \varepsilon) = -(B'_1 p_1 + B'_2 p_2).$$

Note that the extremum will be a minimum since $\partial^2 H/\partial u^2 = I$ is positive definite.

Substituting into the state equations (1.1) and the costate equations (3.2) yields a singularly perturbed linear boundary value problem for the linear system

$$(3.5) \quad \begin{aligned} \frac{dx}{dt} &= A_1 x + A_2 z - S_1 p_1 - S p_2, \\ \frac{dp_1}{dt} &= -Q_1 x - Q_2 z - A'_1 p_1 - A'_3 p_2, \\ \varepsilon \frac{dz}{dt} &= A_3 x + A_4 z - S' p_1 - S_2 p_2, \\ \varepsilon \frac{dp_2}{dt} &= -Q'_2 x - Q_3 z - A'_2 p_1 - A'_4 p_2, \end{aligned}$$

where

$$S_1 = B_1 B'_1, \quad S = B_1 B'_2, \quad \text{and} \quad S_2 = B_2 B'_2.$$

This is subject to the initial conditions (1.2) for the state vectors and the terminal conditions (3.3) linking the state and costate vectors.

It was shown in O'Malley [13] that this singular perturbation problem has (under somewhat different hypotheses) a unique asymptotic solution of the form

$$(3.6) \quad \begin{aligned} x(t, \varepsilon) &= X(t, \varepsilon) + \varepsilon m_1(\kappa, \varepsilon) + \varepsilon n_1(\sigma, \varepsilon), \\ z(t, \varepsilon) &= Z(t, \varepsilon) + m_2(\kappa, \varepsilon) + n_2(\sigma, \varepsilon), \\ p_1(t, \varepsilon) &= P_1(t, \varepsilon) + \varepsilon \rho_1(\kappa, \varepsilon) + \varepsilon \gamma_1(\sigma, \varepsilon), \\ p_2(t, \varepsilon) &= P_2(t, \varepsilon) + \rho_2(\kappa, \varepsilon) + \gamma_2(\sigma, \varepsilon), \end{aligned}$$

where the outer expansion (X, Z, P_1, P_2) satisfies the system (3.5) and has an asymptotic series expansion in ε ; the initial boundary layer correction $(\varepsilon m_1, m_2, \varepsilon \rho_1, \rho_2)$ has an asymptotic series expansion in ε whose terms tend to zero as

$$(3.7) \quad \kappa = t/\varepsilon$$

tends to infinity; and $(\varepsilon n_1, n_2, \varepsilon \gamma_1, \gamma_2)$, the terminal boundary layer correction, tends to zero as

$$(3.8) \quad \sigma = (1 - t)/\varepsilon$$

tends to infinity.

The asymptotic correctness of these expansions follows from the well-established theory (cf. O'Malley [11] and Harris [5]) for such linear singular perturbation problems.

3.2. The outer expansion. Since the outer expansion satisfies (3.5), its leading term $(X_0, Z_0, P_{10}, P_{20})$ must satisfy the limiting system (3.5) with $\varepsilon = 0$. Noting that $\mu = A'_{40} + Q_{30}A_{40}^{-1}S_{20}$ is invertible because G is, we obtain

$$\begin{aligned} Z_0 &= -A_{40}^{-1}(A_{30} + S_{20}\mu^{-1}(Q'_{20} - Q_{30}A_{40}^{-1}A_{30}))X_0 \\ (3.9) \quad &+ A_{40}^{-1}(S'_0 - S_{20}\mu^{-1}(A'_{20} + Q_{30}A_{40}^{-1}S'_0))P_{10}, \\ P_{20} &= -\mu^{-1}(Q'_{20} - Q_{30}A_{40}^{-1}A_{30})X_0 - \mu^{-1}(A'_{20} + Q_{30}A_{40}^{-1}S'_0)P_{10} \end{aligned}$$

and there remains a linear system for X_0 and P_{10} . Moreover, (3.6) implies that X_0 and P_{10} satisfy the limiting boundary conditions for $x(0, \varepsilon)$ and $p_1(1, \varepsilon)$.

After considerable algebraic manipulation, we find that X_0 and P_{10} will satisfy the two-point problem

$$\begin{aligned} (3.10) \quad \frac{dX_0}{dt} &= aX_0 - \bar{B}\bar{R}^{-1}\bar{B}'P_{10}, \quad X_0(0) = x(0, 0), \\ \frac{dP_{10}}{dt} &= -\mathcal{Q}X_0 - a'P_{10}, \quad P_{10}(1) = \pi_1(0)X_0(1) \end{aligned}$$

which is equivalent to the reduced problem (2.3). Thus, X_0 and P_{10} are uniquely determined and Z_0 and P_{20} follow from (3.9). Since these leading terms of the outer expansion do not involve the initial condition for z or the terminal condition for p_2 , we can expect nonuniform convergence at both endpoints.

We note that the outer expansion must satisfy the system (3.5) to all orders ε^j . Higher order terms of the outer expansion will therefore satisfy a nonhomogeneous form of the limiting system with successively known nonhomogeneous terms. The boundary values $X_j(0)$ and $P_{1j}(0) - \pi_1(0)X_j(0)$ can likewise be recursively determined by lower order terms in the full expansion (3.6). Thus, the Fredholm alternative implies that the terms of the expansion can be uniquely generated successively.

3.3. The boundary layer correction at $t = 0$. Since the full expansion (3.6) and the outer expansion both satisfy the system (3.5), linearity and the fact that a boundary layer correction at one endpoint is asymptotically negligible at the other endpoint imply that both boundary layer corrections must also satisfy (3.5). In particular, the initial boundary layer correction will satisfy the linear system

$$\begin{aligned} (3.11) \quad \frac{dm_1}{d\kappa} &= \varepsilon A_1(\varepsilon\kappa, \varepsilon)m_1 + A_2(\varepsilon\kappa, \varepsilon)m_2 - \varepsilon S_1(\varepsilon\kappa, \varepsilon)\rho_1 - S(\varepsilon\kappa, \varepsilon)\rho_2, \\ \frac{d\rho_1}{d\kappa} &= -\varepsilon Q_1(\varepsilon\kappa, \varepsilon)m_1 - Q_2(\varepsilon\kappa, \varepsilon)m_2 - \varepsilon A'_1(\varepsilon\kappa, \varepsilon)\rho_1 - A'_3(\varepsilon\kappa, \varepsilon)\rho_2, \\ \frac{dm_2}{d\kappa} &= \varepsilon A_3(\varepsilon\kappa, \varepsilon)m_1 + A_4(\varepsilon\kappa, \varepsilon)m_2 - \varepsilon S'(\varepsilon\kappa, \varepsilon)\rho_1 - S_2(\varepsilon\kappa, \varepsilon)\rho_2, \end{aligned}$$

$$\frac{d\rho_2}{d\kappa} = -\varepsilon Q_2(\varepsilon\kappa, \varepsilon)m_1 - Q_3(\varepsilon\kappa, \varepsilon)m_2 - \varepsilon A'_2(\varepsilon\kappa, \varepsilon)\rho_1 - A'_4(\varepsilon\kappa, \varepsilon)\rho_2$$

for $\kappa \geq 0$. The leading term $(m_{10}, \rho_{10}, m_{20}, \rho_{20})$ must satisfy this system with $\varepsilon = 0$. In particular, since

$$(3.12) \quad \frac{d}{d\kappa} \begin{pmatrix} m_{20} \\ \rho_{20} \end{pmatrix} = G(0) \begin{pmatrix} m_{20} \\ \rho_{20} \end{pmatrix},$$

equation (1.7) of assumption (iii) implies the general decaying solution

$$m_{20}(\kappa) = T_{11}(0) e^{-\Lambda(0)\kappa} c_1,$$

$$\rho_{20}(\kappa) = T_{21}(0) e^{-\Lambda(0)\kappa} c_1.$$

Thus, hypothesis (iii) implies that

$$c_1 = T_{11}^{-1}(0)(z(0, 0) - Z_0(0)).$$

Integrating, then, this uniquely determines the decaying solutions

$$(3.13) \quad \begin{aligned} m_{10}(\kappa) &= -(A_{20}(0)T_{11}(0) - S_0(0)T_{21}(0))\Lambda^{-1}(0)T_{11}^{-1}(0)m_{20}(\kappa), \\ \rho_{10}(\kappa) &= (Q_{20}(0)T_{11}(0) + A'_{30}(0)T_{21}(0))\Lambda^{-1}(0)T_{11}^{-1}(0)m_{20}(\kappa). \end{aligned}$$

Higher order terms in this boundary layer correction can be obtained by equating higher order terms in the system (3.11) and in the initial condition for z . Proceeding analogously, we obtain a nonhomogeneous version of the system (3.12) with a successively known, exponentially decaying nonhomogeneous term. The initial value for $m_{2j}(0)$ will be determined by $Z_{j-1}(0)$. The terms of the expansion can therefore be uniquely determined recursively as exponentially decaying vectors.

3.4. The boundary layer correction at $t = 1$. The terminal boundary layer correction must satisfy

$$(3.14) \quad \begin{aligned} \frac{dn_1}{d\sigma} &= -\varepsilon A_1(1 - \varepsilon\sigma, \varepsilon)n_1 - A_2(1 - \varepsilon\sigma, \varepsilon)n_2 + \varepsilon S_1(1 - \varepsilon\sigma, \varepsilon)\gamma_1 \\ &\quad + S(1 - \varepsilon\sigma, \varepsilon)\gamma_2, \\ \frac{d\gamma_1}{d\sigma} &= \varepsilon Q_1(1 - \varepsilon\sigma, \varepsilon)n_1 + Q_2(1 - \varepsilon\sigma, \varepsilon)n_2 + \varepsilon A'_1(1 - \varepsilon\sigma, \varepsilon)\gamma_1 \\ &\quad + A'_3(1 - \varepsilon\sigma, \varepsilon)\gamma_2, \\ \frac{dn_2}{d\sigma} &= -\varepsilon A_3(1 - \varepsilon\sigma, \sigma)n_1 - A_4(1 - \varepsilon\sigma, \varepsilon)n_2 + \varepsilon S'(1 - \varepsilon\sigma, \varepsilon)\gamma_1 \\ &\quad + S_2(1 - \varepsilon\sigma, \varepsilon)\gamma_2, \\ \frac{d\gamma_2}{d\sigma} &= \varepsilon Q'_2(1 - \varepsilon\sigma, \varepsilon)n_1 + Q_3(1 - \varepsilon\sigma, \varepsilon)n_2 + \varepsilon A'_2(1 - \varepsilon\sigma, \varepsilon)\gamma_1 \\ &\quad + A'_4(1 - \varepsilon\sigma, \varepsilon)\gamma_2 \end{aligned}$$

for $\sigma \geq 0$. When $\varepsilon = 0$, then, we have

$$\frac{d}{d\sigma} \begin{pmatrix} n_{20} \\ \gamma_{20} \end{pmatrix} = -G(1) \begin{pmatrix} n_{20} \\ \gamma_{20} \end{pmatrix}$$

and this system has the general decaying solution

$$\begin{aligned} n_{20}(\sigma) &= T_{12}(1) e^{-\Lambda(1)\sigma} k_2, \\ \gamma_{20}(\sigma) &= T_{22}(1) e^{-\Lambda(1)\sigma} k_2. \end{aligned}$$

The terminal condition for p_2 (cf. (3.3)) and the representation (3.6) imply that

$$(3.15) \quad \gamma_2(0, \varepsilon) - \pi_3(\varepsilon)n_2(0, \varepsilon) = -P_2(1, \varepsilon) + \pi'_2(\varepsilon)(X(1, \varepsilon) + \varepsilon n_1(0, \varepsilon)) + \pi_3(\varepsilon)Z(1, \varepsilon).$$

So, in particular, $\gamma_{20}(0) - \pi_3(0)n_{20}(0)$ is determined by the solution of the reduced problem, i.e., the leading term of the outer expansion. Thus, we have

$$k_2 = (T_{22}(1) - \pi_3(0)T_{12}(1))^{-1}(-P_{20}(1) + \pi'_2(0)X_0(1) + \pi_3(0)Z_0(1))$$

(using (1.9) of assumption (iii)). This also uniquely determines the decaying terms

$$(3.16) \quad \begin{aligned} n_{10}(\sigma) &= (A_{20}(1)T_{12}(1) - S_0(1)T_{22}(1))\Lambda^{-1}(1) e^{-\Lambda(1)\sigma} k_2, \\ \gamma_{10}(\sigma) &= -(Q_{20}(1)T_{12}(1) + A'_{30}(1)T_{22}(1))\Lambda^{-1}(1) e^{-\Lambda(1)\sigma} k_2 \end{aligned}$$

and higher order terms follow in the usual fashion.

3.5. Asymptotic expansions for the optimal control and cost. Knowing the expansions (3.6) for the costates, the control relation (3.4) implies that the corresponding optimal control has the form

$$(3.17) \quad u(t, \varepsilon) = U(t, \varepsilon) + v(\kappa, \varepsilon) + w(\sigma, \varepsilon),$$

where U , v and w have asymptotic series expansions such that the boundary layer corrections v and w tend to zero as κ and σ , respectively, tend to infinity and $U(t, 0)$ is the optimal control $U_0(t)$ for the reduced problem (1.12). Further, the optimal cost $J^*(\varepsilon)$ takes the form

$$(3.18) \quad J^*(\varepsilon) = \frac{1}{2}\lambda(\varepsilon) + \frac{1}{2}\int_0^1 L_1(t, \varepsilon) dt + \frac{\varepsilon}{2}\int_0^\infty L_2(\kappa, \varepsilon) d\kappa + \frac{\varepsilon}{2}\int_0^\infty L_3(\sigma, \varepsilon) d\sigma,$$

where λ and the L_i 's have asymptotic series expansions as $\varepsilon \rightarrow 0$ with integrable coefficients. Thus $J^*(\varepsilon)$ has an expansion whose leading term J_0^* is the optimal cost of the reduced problem.

4. Related problems.

4.1. The fixed endpoint problem. Instead of our previous problem (1.1)–(1.4), suppose we wish to solve (1.1)–(1.4) with the additional condition that the terminal states

$$(4.1) \quad x(1, \varepsilon) \quad \text{and} \quad z(1, \varepsilon)$$

are also prescribed. We would again introduce a Hamiltonian and obtain necessary and sufficient conditions for an optimal control as before. Specifically, the

states and the costates would again satisfy the linear system (3.5) and the control would be given by the control relation (3.4), but the terminal conditions (3.3) would now be replaced by the fixed endpoint condition (4.1).

Let us again seek an asymptotic solution of the form (3.6). Under hypotheses (i) and (iii) of our theorem, we find that the leading term $(X_0, Z_0, P_{10}, P_{20})$ of the outer solution will now satisfy the two-point problem

$$(4.2) \quad \begin{aligned} \frac{dX_0}{dt} &= aX_0 - \bar{B}\bar{R}^{-1}\bar{B}'P_{10}, & X_0(0) &= x(0, 0), \\ \frac{dP_{10}}{dt} &= -\mathcal{Q}X_0 - a'P_{10}, & X_0(1) &= x(1, 0) \end{aligned}$$

(cf. (3.10)), with Z_0 and P_{20} given by (3.9). Under hypothesis (iv) that \mathcal{Q} is non-negative definite, it is well known (cf. Bucy [2]) that problem (4.2) has a unique solution if the system is observable. By making these assumptions, the terms of the outer expansion can again be determined recursively. Moreover, by assuming, as in (1.8), that $T_{11}(0)$ is nonsingular, the initial boundary layer correction can be calculated as before. Likewise, the terminal boundary layer correction can be readily obtained if, instead of (1.9), we ask that $T_{12}(1)$ be nonsingular. Thus, the fixed endpoint problem can be solved like the free endpoint problem (cf. also Wilde and Kokotović [17]). We note that the limiting outer solution can again be associated with the solution of the natural reduced problem which is equivalent to a familiar regulator problem (cf. Lemma 1).

4.2. Problems with a more significant terminal cost. The reader may have already felt that the special form of the matrix $\pi(\varepsilon)$ in (1.4) was somewhat arbitrarily imposed. In particular, note that (1.4) does not allow $\pi(0)$ to be positive definite.

Let us then consider the control problem (1.1)–(1.3) with \mathcal{Q} as in (1.4) but with

$$(4.3) \quad \tilde{\pi} = \begin{pmatrix} \tilde{\pi}_1(\varepsilon) & \tilde{\pi}_2(\varepsilon) \\ \tilde{\pi}_2'(\varepsilon) & \tilde{\pi}_3(\varepsilon) \end{pmatrix}$$

symmetric and nonnegative definite. We shall examine only the case where $\tilde{\pi}_3(0)$ is nonsingular. (Note that if $\tilde{\pi}_3(0)$ were zero, we would also have $\tilde{\pi}_2(0) = 0$ and (1.4) would apply.)

Introducing costate vectors as before, we find that the linear system (3.5) must again be satisfied subject to the initial conditions (1.2) and the terminal conditions

$$(4.4) \quad \begin{aligned} p_1(1, \varepsilon) &= \tilde{\pi}_1(\varepsilon)x(1, \varepsilon) + \tilde{\pi}_2(\varepsilon)z(1, \varepsilon), \\ \varepsilon p_2(1, \varepsilon) &= \tilde{\pi}_2'(\varepsilon)x(1, \varepsilon) + \tilde{\pi}_3(\varepsilon)z(1, \varepsilon). \end{aligned}$$

Now, if we seek an asymptotic solution of the form (3.6), we find that (under hypotheses (i) and (ii)) the leading term of the outer expansion must satisfy the

linear problem

$$(4.5) \quad \begin{aligned} \frac{dX_0}{dt} &= \alpha X_0 - \bar{B}\bar{R}^{-1}\bar{B}'P_{10}, & X_0(0) &= x(0, 0), \\ \frac{dP_{10}}{dt} &= -\mathcal{Q}X_0 - \alpha'P_{10}, & P_{10}(1) &= (\tilde{\pi}_1(0) - \tilde{\pi}_2(0)\tilde{\pi}_3^{-1}(0)\tilde{\pi}_2'(0))X_0(1) \end{aligned}$$

with Z_0 and P_{20} given by (3.9). Noting that $\tilde{\pi}_1 - \tilde{\pi}_2\tilde{\pi}_3^{-1}\tilde{\pi}_2'$ is nonnegative definite, we have a unique solution to (4.5) provided \mathcal{Q} is nonnegative definite (cf. Anderson and Moore [1] or Bucy [2]). Likewise, the boundary layer corrections can be determined provided the matrices $T_{11}(0)$ and $T_{12}(1)$ are nonsingular. The relationship between the limiting solution and the solution of the reduced problem now seems more complicated than under condition (1.4). This, no doubt, explains the persistence of this condition.

4.3. Other generalizations. First, note that the problem (1.1)–(1.4) could also be solved using the matrix Riccati equation approach. Note, in particular, that that approach has certain advantages with regard to feedback and computation and that it can be readily extended to analogous time-invariant problems on semi-infinite intervals. The authors have studied such techniques and will report them elsewhere.

Second, the technique developed could be further extended to certain quasi-linear problems (cf. O'Malley [14]) and to state systems involving several small parameters (cf. O'Malley [12]). More challenging extensions would be to problems with bounded controls (cf. Collins [3]).

REFERENCES

- [1] B. D. O. ANDERSON AND J. B. MOORE, *Linear Optimal Control*, Prentice-Hall, Englewood Cliffs, N.J., 1971.
- [2] R. S. BUCY, *Two-point boundary value problems of linear Hamiltonian systems*, SIAM J. Appl. Math., 15 (1967), pp. 1385–1389.
- [3] W. D. COLLINS, *Singular perturbations of linear time-optimal control problems*, Recent Mathematical Developments in Control, D. J. Bell, ed., Academic Press, London, 1973, pp. 123–136.
- [4] A. H. HADDAD AND P. V. KOKOTOVIĆ, *A note on singular perturbation of linear state regulators*, IEEE Trans. Automatic Control, 16 (1971), pp. 279–281.
- [5] W. A. HARRIS, JR., *Singularly perturbed boundary value problems revisited*, Symposium on Ordinary Differential Equations (Minneapolis, 1972), W. A. Harris, Jr. and Y. Sibuya, eds., Lecture Notes in Mathematics 312, Springer-Verlag, Berlin, 1973, pp. 54–64.
- [6] F. HOPPENSTEADT, *Properties of solutions of ordinary differential equations with a small parameter*, Comm. Pure Appl. Math., 24 (1971), pp. 807–840.
- [7] R. E. KALMAN, *Contributions to the theory of optimal control*, Bull. Soc. Math. Mexicana, 5 (1960), pp. 102–119.
- [8] P. V. KOKOTOVIĆ AND P. SANNUTI, *Singular perturbation method for reducing the model order in optimal control design*, IEEE Trans. Automatic Control, AC.13 (1968), pp. 377–384.
- [9] P. V. KOKOTOVIĆ AND R. A. YACKEL, *Singular perturbation of linear regulators: Basic theorems*, Ibid., AC.17 (1972), pp. 29–37.
- [10] J. J. O'DONNELL, *Asymptotic solution of the matrix Riccati equation of optimal control*, Proc. Fourth Allerton Conference on Circuit and System Theory, University of Illinois, Urbana, 1966, pp. 577–586.
- [11] R. E. O'MALLEY, JR., *Singular perturbations of a boundary value problem for a system of nonlinear differential equations*, J. Differential Equations, 8 (1970), pp. 431–447.

- [12] ———, *On initial value problems for nonlinear systems of differential equations with two small parameters*, Arch Rational Mech. Anal., 40 (1971), pp. 209–222.
- [13] ———, *The singularly perturbed linear state regulator problem*, this Journal, 10 (1972), pp. 399–413.
- [14] ———, *Boundary layer methods for certain nonlinear singularly perturbed optimal control problems*, J. Math. Anal Appl., 45 (1974).
- [15] ———, *Introduction to Singular Perturbations*, Academic Press, New York, 1974.
- [16] W. WASOW, *Asymptotic Expansions for Ordinary Differential Equations*, Interscience, New York, 1965.
- [17] R. R. WILDE AND P. V. KOKOTOVIĆ, *Optimal open and closed loop control of singularly perturbed linear systems*, IEEE Trans. Automatic Control, 18 (1973), pp. 616–626.

DYNAMIC PROGRAMMING AND MINIMUM PRINCIPLES FOR SYSTEMS WITH JUMP MARKOV DISTURBANCES*

RAYMOND RISHEL†

Abstract. This paper studies optimum control of random differential equations of the form

$$\dot{x} = f^{r(t)}(t, x, u)$$

in which $r(t)$ is a jump Markov process. Optimality conditions of dynamic programming type and stochastic minimum principles are given. The problems posed involve terminal conditions, and transversality conditions are shown to hold.

Introduction. The aim of this paper is to study the optimal control of stochastic systems of the form

$$(1) \quad \dot{x} = f^{r(t)}(t, x, u).$$

In (1), $r(t)$ is a finite state Markov process and the derivative changes from $f^i(t, x, u)$ to $f^j(t, x, u)$ as $r(t)$ jumps from i to j . Stated very imprecisely, the optimization problem will be to find in a given class \mathcal{U} of control functions $u^i(t, x)$ a control which simultaneously for each initial condition (t, x, i) minimizes the conditional expectations

$$(2) \quad E\{\varphi(\tau, x(\tau)) | x(t) = x, r(t) = i\}.$$

In (2), τ denotes the first time that the solution $x(t)$ of (1) corresponding to the control $u^i(t, x)$ reaches a terminal set M and $\varphi(t, x)$ is the cost incurred if the terminal time and state are (t, x) .

In corresponding deterministic control problems, optimal controls may be discontinuous functions of (t, x) . Since a similar situation is to be expected in the stochastic situation, the optimization problem should be formulated so as to include discontinuous controls in the class \mathcal{U} of controls.

If a given fixed discontinuous control $u^i(t, x)$ is substituted for u in (1), the equation

$$(3) \quad \dot{x} = f^{r(t)}(t, x, u^{r(t)}(t, x))$$

results. This is a stochastic differential equation with discontinuous right-hand side.

Part 1 of the paper studies this type of stochastic differential equation. The differentiability properties with respect to initial conditions of the conditional expectation of a cost function of the type given in (2) are described. It is shown

*Received by the editors July 5, 1973, and in revised form October 29, 1973.

† Department of Mathematics, University of Kentucky, Lexington, Kentucky 40506.

roughly that the performance function (2) of a control is differentiable except on a set where the control may not be differentiable. Knowing the differentiability properties of a performance function allows a rigorous presentation of a dynamic programming argument in Part 2. The conditional expectations (2) and their partial derivatives are shown to satisfy integral equations. One set of integral equations is given which contains terminal conditions which are the transversality conditions found in optimal control theory. The conditional expectations (2) are generalized solutions of systems of hyperbolic partial differential equations. A maximal principle is established for this type of partial differential operator.

Part 2 states the optimal control problem and applies the results of Part 1 to obtain optimality conditions. Four different sets of necessary and sufficient conditions for optimality are given. Satisfaction of the system of partial differential equations of dynamic programming except on certain lower-dimensional sets is a necessary and sufficient condition for optimality. Two minimum principle sets of conditions are obtained from the dynamic programming argument. Another minimum principle which is necessary and sufficient for optimality is obtained through use of the deterministic Pontryagin principle. While an attempt was made to avoid imposing restrictive a priori conditions on the class of controls, the conditions of this paper do have the deficiency common in many dynamic programming arguments that some of the a priori conditions imposed on the class of controls are restrictive and the optimal controls for certain examples will lie outside of this class. Perhaps the most restrictive of these assumptions is that the expression (8) be bounded away from zero on certain cells.

Problems similar to the optimization problem described above were considered in a sequence of papers by Krassovskii and Lidskii. References [8] and [9] are typical of these. In [8], [9], the control problem is on an infinite interval and terminal conditions are not imposed. In [17], Wonham considered the special case of the linear regulator version of this problem. He explicitly computes the optimal control law for this case. In a slightly different setting Kushner [7] defined a stochastic maximal principle. For fixed terminal time linear problems without terminal conditions, Szwed [15] gave a maximal principle. The minimum principles given in this paper are different from those of [7] or [15]. In particular, the adjoint equations are deterministic equations while those of [7] and [15] are stochastic differential equations.

Random differential equations of the type discussed in Part 1 are a special case of a class of stochastic processes called random evolutions by Griego and Hersh [5]. A survey of work on generalizations of this type of process is given by Pinsky [11]. Stochastic differential equations with continuous coefficients similar to those considered in Part 1 were studied by Heath in [6].

This present paper could be considered as an extension to the stochastic case of corresponding results in deterministic control theory given by Berkovitz [1], [2], Boltyanskii [3] and Mirica [10].

In [13], it was shown that a sufficient condition for optimality could be expressed in terms of a generalized solution of the partial differential equation of dynamic programming. The hypothesis of the corresponding sufficiency condition of this paper is slightly less restrictive and the proof is simpler.

Part 1
Stochastic Differential Equations with Discontinuous
Right-Hand Side and their Conditional Expectations

The assumptions which will be imposed on the quantities involved in the differential equations which will be described are quite complex. Therefore, we shall begin by stating somewhat imprecisely the situation to be considered and later give precise conditions on these quantities.

Let G denote an open subset of E^{n+1} . Let I denote a finite set of integers. For each $i \in I$ let $f^i(t, x)$,

$$(4) \quad f^i: G \rightarrow E^n,$$

be a piecewise continuous piecewise smooth function. Let $r(t)$ be a Markov process with state space I . Consider the stochastic process $x(t)$, where $x(t)$ is a solution of

$$(5) \quad \dot{x} = f^{r(t)}(t, x).$$

Let M be a closed subset of G . M will be called the terminal set. Let τ be the first time $x(t)$ belongs to M . Let $\varphi(t, x)$ be a continuously differentiable real-valued function defined on G . Let

$$(6) \quad \psi^i(t, x) = E\{\varphi(\tau, x(\tau)) | x(t) = x, r(t) = i\}$$

denote the conditional expectation of $\varphi(\tau, x(\tau))$ given $x(t) = x$ and $r(t) = i$. We shall study the continuity and differentiability properties of $\psi^i(t, x)$.

1. Deterministic differential equations with discontinuous right-hand side. The way in which a solution of (5) is defined needs to be precisely described. The differential equation will be interpreted as holding for almost every sample function. In order to define this type of solution of (5) it is necessary to have a concept of solution of an ordinary differential equation with discontinuous right-hand side.

In deterministic differential equations with discontinuous right-hand side, there are several differing concepts of solution. The concept we shall use is due to Boltyanskii [3], [16]. The definitions given below are a minor modification of those of [3], [16]. The statement of the hypotheses needed to define this concept of solution is quite long; however, these hypotheses will play an important role in studying the stochastic differential equation (5). For this reason we shall give a detailed discussion of these ideas. There is a difference in the definitions given below and those given by Boltyanskii in that surfaces on which the differential equation can have discontinuities are described implicitly rather than parametrically. This leads to a convenient formulation of transversality conditions and jump conditions for adjoint equations.

To begin the discussion of deterministic differential equations with discontinuous right-hand side, let G and M be as defined previously and $f(t, x)$ be a mapping

$$f: G \rightarrow E^n.$$

Consider the deterministic differential equation

$$(7) \quad \dot{x} = f(t, x).$$

A solution of (7) defined on an interval $[t, \tau]$ having the value x at an initial time t will be denoted by $x(s; t, x)$.

The conditions (A), (B), (C) given below are assumptions which will be made on the mapping f . When they are satisfied f will be said to have an *admissible set of discontinuities*. The set \mathcal{S}_n described below is to be interpreted as a set which contains the discontinuities of f .

(A) There are sets

$$\mathcal{S}_0 \subset \mathcal{S}_1 \subset \cdots \subset \mathcal{S}_n \subset G.$$

Each set \mathcal{S}_i is closed in the relative topology of G . Call the components of $G - \mathcal{S}_n$ or of $\mathcal{S}_i - \mathcal{S}_{i-1}$, $i = 1, \dots, n$, cells. For each nonempty cell of $\mathcal{S}_i - \mathcal{S}_{i-1}$ there is a continuously differentiable mapping

$$\theta: G \rightarrow E^{n+1-i}$$

such that the cell is a relatively open subset of

$$\{(t, x): \theta(t, x) = 0\}.$$

We shall call this function θ the determining function of the cell. For each cell C there is a continuously differentiable mapping $f_c(t, x)$ defined on a neighborhood of the closure of the cell such that $f_c(t, x) = f(t, x)$ on the cell C . Only a finite number of cells intersect any compact subset of G . The terminal set M is contained in \mathcal{S}_n and is the union of a finite number of cells.

(B) Cells can be classified into three categories called type (i), type (ii) or type (iii). Each cell of $G - \mathcal{S}_n$ is required to be of type (i). If C is a cell of type (i), there is a unique corresponding cell $\pi(C)$ whose associated function θ has one more component than that of C (has one component if C is a cell of $G - \mathcal{S}_n$). For each $(t, x) \in C$ there is a necessarily unique trajectory of (7) starting at x at time t , which after a finite time leaves C by reaching $\pi(C)$. Each trajectory upon leaving C strikes $\pi(C)$ at a positive angle, and on $\pi(C)$ this angle is bounded below. That is, if θ_i is the additional component of the function θ associated with $\pi(C)$ over that of the function θ associated with C and the trajectory exits from C at $(t, x) \in \pi(C)$, then there is an $\eta > 0$ such that

$$(8) \quad \theta_{ii}(t, x) + \theta_{ix}(t, x)f_c(t, x) \geq \eta.$$

If C is a cell of type (ii), there is a cell $\Sigma(C)$ of type (i) whose associated function θ has one less component than that of C . From each $(t, x) \in C$ there starts a unique trajectory of (7) going into $\Sigma(C)$ which has only the point (t, x) in common with C . The function $f(t, x)$ is continuously differentiable on $C \cup \Sigma(C)$. Cells of M are of type (iii).

It follows a trajectory of (7) may be continued from cell to cell until it hits M 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). Furthermore it is assumed that:

- (C) (i) Every trajectory remains in G and reaches M in a finite time.
- (ii) Every trajectory goes through a finite number of cells to reach M .
- (iii) Let $\tau(t, x)$ be the time and $x(t, x)$ the position that the trajectory of (7) starting at (t, x) hits M . Then $\tau(t, x)$ and $x(t, x)$ are continuous functions of (t, x) on G .

The following are consequences of assumptions (A)–(C).

For any $(t, x) \in G$ there will be a unique trajectory passing through a finite number of type (i) cells C_1, C_2, \dots, C_q to reach M . For any $i, 1 \leq i < q$, if $\pi(C_i)$ is of type (i), then $C_{i+1} = \pi(C_i)$ and the trajectory goes directly from C_i 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 $\pi(C_i)$ in a single point.

Let $t_i(t, x), i = 1, 2, \dots, q$, denote the times at which the trajectory reaches $\pi(C_i)$. That is, $t_i(t, x), i = 1, \dots, q - 1$, are the times the trajectory crosses from cell to cell and $t_q(t, x)$ is the time the trajectory reaches M . Let

$$(9) \quad x_i(t, x) = x(t_i(t, x); t, x).$$

Then $(t_i(t, x), x_i(t, x)) \in \pi(C_i)$ and $(s, x(s; t, x)) \in C_{i+1}$ for $t_i(t, x) \leq s < t_{i+1}(t, x)$ if $\pi(C_i)$ is of type (i) or for $t_i(t, x) < s < t_{i+1}(t, x)$ if $\pi(C_i)$ is of type (ii).

Theorem 1 describes the differentiability properties with respect to the initial conditions (t, x) of the trajectory $x(s; t, x)$ and the crossing times and positions $t_j(t, x), x_j(t, x)$. Theorem 1 is very similar to corresponding work in [10, Lemma 3.1, pp. 296–302]. A proof of Theorem 1 is given in Appendix A.

Recall that $f_{c_j}(t, x)$ is a continuously differentiable function defined on a neighborhood of \bar{C}_j which agrees with $f(t, x)$ on C_j and that $x(s; t, x)$ belongs to C_j for $\{s: t_{j-1}(t, x) < s < t_j(t, x)\}$.

THEOREM 1. *Let C be a cell of $G - \mathcal{G}_n$ and consider a trajectory $x(s; t, x)$ of (7) as a function of its initial conditions (t, x) on C . Then:*

- (a) $x(s; t, x)$ is continuously differentiable in (s, t, x) on

$$\{(s, t, x): (t, x) \in C, t_{j-1}(t, x) < s < t_j(t, x)\}$$

for each j . Denote by $\delta x(s; t, x)$ the matrix of partial derivatives of $x(s; t, x)$ with respect to x . Then for fixed $(t, x) \in C$, $\delta x(s; t, x)$ satisfies the matrix differentiable equation

$$(10) \quad \delta \dot{x}(s; t, x) = f_{c_j x}(s, x(s; t, x)) \delta x(s; t, x)$$

on

$$\{s: t_{j-1}(t, x) < s < t_j(t, x)\}.$$

If $\delta x(s; t, x)$ denotes the vector of partial derivatives of $x(s; t, x)$ with respect to t , the vector differential equation denoted symbolically by (10) holds on the same set for the vector $\delta x(s; t, x)$ of partial derivatives with respect to t . For any compact set K of G , $\delta x(s; t, x)$ is bounded on the set $\{(s, t, x): (t, x) \in C \cap K, t \leq s \leq t_q(t, x)\}$.

(b) The times $t_j(t, x)$ and positions $x_j(t, x)$ at which $x(s; t, x)$ leaves the respective cells C_j are continuously differentiable on C . On the intersection of any compact set $K \subset G$ with C these partial derivatives are bounded.¹

¹ In examples in which (8) does not hold, $\delta x(s; t, x)$ and the partial derivatives of $t_j(t, x)$ may not be locally bounded. For an example see [16, p. 32].

We shall say a function defined on $G - \mathcal{S}_n$ is locally bounded if it is bounded on $G - \mathcal{S}_n \cap K$ for each compact subset K of G .

(c) At the times $t_{j-1} = t_{j-1}(t, x)$ and $t_j = t_j(t, x)$, if $\delta x(s)$ denotes the matrix of partial derivatives of $x(s; t, x)$ with respect to x , then $\delta x(s)$ has the right- and left-hand limits.

$$(11) \quad \delta x(t_{j-1}^+) = -f_{c_j}(t_{j-1}, x_{j-1}) \frac{\partial t_{j-1}}{\partial x} + \frac{\partial x_{j-1}}{\partial x}$$

and

$$(12) \quad \delta x(t_j^-) = -f_{c_j}(t_j, x_j) \frac{\partial t_j}{\partial x} + \frac{\partial x_j}{\partial x}.$$

If $\delta x(s)$ denotes the vector of partial derivatives of $x(s; t, x)$ with respect to t , equations (11) and (12) hold with the matrices of partial derivatives with respect to x , $\partial t_j / \partial x$ and $\partial x_j / \partial x$ replaced by the vectors $\partial t_j / \partial t$ and $\partial x_j / \partial t$.

2. Definition of solution of the stochastic differential equation. We shall assume for each function $f^i(t, x)$, $i \in I$, that $f^i(t, x)$ has an admissible set of discontinuities. Recall that the sample functions of a jump Markov process are step functions with probability one. For each step function sample function $r(s)$ there is a decomposition of the interval $[t, \infty)$ by a divergent sequence

$$t = s_0 < s_1 < \cdots < s_m < \cdots$$

such that $r(s) = j_m$ if $s_{m-1} \leq s < s_m$, where j_m takes values in I . Each deterministic differential equation

$$(13) \quad \dot{x} = f^i(s, x)$$

has a solution which extends until it hits M . Define a function $x(s)$ inductively as follows. Define $x(t) = x$. Suppose $x(s)$ has been defined for $t \leq s \leq s_{m-1}$ and $x(s)$ does not belong to M for any s in this interval. Then let $x(s)$ be the solution of

$$\dot{x} = f^{j_m}(s, x)$$

on $s_{m-1} \leq s \leq s_m$ if $x(s) \notin M$ for $s_{m-1} \leq s \leq s_m$ or on $s_{m-1} \leq s \leq \tau < s_m$ if $x(s) \notin M$ for $s_{m-1} \leq s < \tau$ and $x(\tau) \in M$.

Define the solution $x(s, \omega)$ of (5) to be the function $x(s)$ just defined if $r(s, \omega)$ is a step function. If $r(s, \omega)$ is not a step function, let $x(s, \omega)$ be an arbitrary continuous function joining x and M . It is seen that this procedure defines a solution $x(t, \omega)$ of (5) on an interval $[t, \tau]$, where τ is the first time $x(t, \omega)$ hits M , or on the interval $[t, \infty]$ if $x(t, \omega)$ does not hit M .

An argument can be given to show that this process defines a measurable stochastic process and that the time τ is a stopping time. In fact, it can be shown that $(x(s), r(s))$ defines a strong Markov process killed at the time τ . (This last assertion will not be used in this paper.)

2.1. Assumptions on $f^i(t, x)$. As mentioned we shall assume for each i that $f^i(t, x)$ has an admissible set of discontinuities. Let $G \supset \mathcal{S}_n^i \supset \mathcal{S}_{n-1}^i \supset \cdots \supset \mathcal{S}_0^i$ be the sets of condition (A) for the function $f^i(t, x)$. The cells corresponding to

these sets will be called i -cells. In addition, we shall assume conditions (D) and (E) given below.

(D) There are sets $G \supset \mathcal{S}_n \supset \mathcal{S}_{n-1} \supset \cdots \supset \mathcal{S}_0$ such that

$$(14) \quad \bigcup_{i \in I} \mathcal{S}_n^i \subset \mathcal{S}_n$$

and these sets are such that conditions (A), (B), (C) of the definition of an admissible set of discontinuities are satisfied with respect to them for each function $f^i(t, x)$, $i \in I$. The connected components of $G - \mathcal{S}_n$ or $\mathcal{S}_k - \mathcal{S}_{k-1}$, $k = 1, \dots, n$, will be called cells of the common decomposition of G .

(E) Each solution $x^i(s; t, x)$ of

$$(15) \quad \dot{x} = f^i(s, x), \quad x^i(t; t, x) = x$$

meets the sets \mathcal{S}_n^j , $j \neq i$, at a finite set of points.

Define $S(t, x)$ to be the set of vector functions $x(t)$ satisfying:

(a) $x(s)$ is defined and continuous on some interval $[t, \tau]$ whose left endpoint is t .

(b) $x(t) = x$; either $\tau = \infty$ or $x(\tau) \in M$.

(c) There is a decomposition of the interval $[t, \infty)$ by a divergent sequence $\{s_m\}$ and a sequence of integers $j_m \in I$,

$$t = s_0 < \cdots < s_m < \cdots,$$

so that if $s_{m-1} < s < s_m < \tau$ or $s_{m-1} < s < \tau < s_m$, $x(s)$ is a solution of

$$\dot{x} = f^{j_m}(s, x)$$

on either the interval

$$s_{m-1} < s < s_m \quad \text{or} \quad s_{m-1} < s < \tau.$$

Define the *domain of influence of a point* (t, x) to be the set

$$\{(s, y) : x(s) = y \text{ for some } x(\cdot) \in S(t, x)\}.$$

Let the *domain of influence of a set* be the union of domains of influence of points in the set. Notice that with probability one the sample functions of (5) are in $S(t, x)$ and these sample functions lie in the domain of influence of (t, x) . Notice that the definition of the set $S(t, x)$ and domain of influence of (t, x) involve the functions $f^i(t, x)$, but do not involve probability considerations. Thus conditions placed on $S(t, x)$ or the domain of influence of sets are conditions on the functions $f^i(t, x)$. We shall also assume the condition:

(F) For each compact set K the domain of influence of K is compact.

Notice that (F) implies that with probability one solutions of (5) starting with initial conditions in a compact set K reach M before some bounded time.

Let $t^i(t, x)$ denote the time and $x^i(t, x)$ the position at which the solution $x^i(s; t, x)$ of (13) hits M . By assumption (C) (iii) we are assuming that these are continuous functions on G for each i . Notice that if $t, x \in M$,

$$(16) \quad (t^i(t, x), x^i(t, x)) = (t, x).$$

Hence the continuity of these functions implies that, for any neighborhood of a point of (t, x) , there is a smaller neighborhood such that, for any i , a solution of $\dot{x} = f^i(s, x)$ which begins in the smaller neighborhood will hit M within the larger neighborhood. This is a special type of boundary behavior which is a definite restriction on the functions $f^i(s, x)$.

3. Properties of $\psi^i(t, x)$. In this section we determine the continuity and differentiability properties of the conditional expectation

$$(17) \quad \psi^i(t, x) = E\{\varphi(\tau, x(\tau)) | x(t) = x, r(t) = i\}$$

and derive integral equations which $\psi^i(t, x)$ and its partial derivatives satisfy.

3.1. Representation of $\psi^i(t, x)$. Recall that a finite state Markov process $r(t)$ is characterized by a matrix (λ_{ij}) of real numbers such that if $i \neq j$,

$$(18) \quad \lambda_{ij} \geq 0, \quad \sum_j \lambda_{ij} = 0$$

and

$$(19) \quad \begin{aligned} P(r(t+h) = j | r(t) = i) &= \lambda_{ij}h + o(h), \\ P(r(t+h) = i | r(t) = i) &= 1 + \lambda_{ii}h + o(h), \end{aligned}$$

where $o(h)$ is a quantity such that

$$\lim_{h \downarrow 0} \frac{o(h)}{h} = 0.$$

Given $r(t) = i$, the probability density that the first jump of $r(t)$ after time t is from i to j and occurs at time s is given by

$$(20) \quad \lambda_{ij} e^{\lambda_{ii}(s-t)}.$$

Given $r(t) = i$, the probability that there are no jumps of $r(t)$ in the interval $[t, s]$ is given by

$$e^{\lambda_{ii}(s-t)}.$$

See [4, p. 350] for a discussion of these statements.

The conditional expectation (17) can be computed by a "renewal-like" method. Consider the events that $r(t)$ has exactly n jumps, $n = 0, 1, \dots$, in the interval (t, τ) . Assumption (F) implies τ is bounded, with probability one; thus the event " $r(t)$ has more than a finite number of jumps in (t, τ) " has probability zero. Let $\chi_n(t, \omega)$ be the characteristic function of the set of ω for which $r(t, \omega)$ has exactly n jumps in $(t, \tau(\omega))$. Then

$$(21) \quad \psi^i(t, x) = \sum_{n=0}^{\infty} E\{\chi_n(t) \varphi(\tau, x(\tau)) | x(t) = x, r(t) = i\}.$$

Notice using the formula (20) for the probability density of a jump from i to j at s that

$$(22) \quad \begin{aligned} &E\{\chi_n(t) \varphi(\tau, x(\tau)) | x(t) = x, r(t) = i\} \\ &= \sum_{j \neq i} \int_t^{\tau^i(t, x)} \lambda_{ij} e^{\lambda_{ii}(s-t)} E\{\chi_{n-1}(s) \varphi(\tau, x(\tau)) | x(s) = x^i(s; t, x), r(s) = j\} ds. \end{aligned}$$

Hence the terms of (21) can be written by induction starting with

$$(23) \quad E\{\chi_0(t)\varphi(\tau, x(t)) | x(t) = x, r(t) = i\} = \varphi(t^i(t, x), x) e^{\lambda_{ii}(t^i(t, x) - t)}.$$

The formula (23) follows because if $r(t)$ does not jump, $x(s) \equiv x^i(s; t, x)$, $\tau = t^i(t, x)$ and the probability of no jumps in (t, τ) is

$$e^{\lambda_{ii}(t^i(t, x) - t)}.$$

3.2. Continuity of $\psi^i(t, x)$. Since by assumption (C)(iii), $t^i(t, x)$ and $x^i(t, x)$ are continuous as functions of (t, x) on G , (23) is continuous as a function of (t, x) . Arguing inductively using (22), we see that each term of the series (21) is continuous as a function of (t, x) on G . For each compact subset of G , we shall show the series (21) converges uniformly on the domain of influence of that compact subset; hence we conclude that as a function of (t, x) , $\psi_i(t, x)$ is continuous.

To see that (21) converges uniformly on the domains of influence of each compact subset of G , we shall proceed as follows. Let \mathcal{E} denote the set of continuous k -dimensional vector-valued functions on G . Let H denote the mapping of \mathcal{E} into \mathcal{E} for which the i th component of $H[g]$ is defined by

$$(24) \quad H[g]^i(t, x) = \sum_{j \neq i} \int_t^{t^i(t, x)} \lambda_{ij} e^{\lambda_{ii}(s-t)} g^j(s, x^i(s; t, x)) ds.$$

For each compact set $K \subset G$, let $I(K)$ denote the domain of influence of K . Define the seminorms

$$(25) \quad \|g\|_K = \max_i \sup_{(t, x) \in I(K)} |g^i(t, x)|.$$

Let

$$(26) \quad T_K = \max_i \sup_{(t, x) \in I(K)} |t^i(t, x)| \quad \text{and} \quad S_K = \min \{t : (t, x) \in K\}.$$

Then recalling that $\sum_{j \neq i} \lambda_{ij} = -\lambda_{ii}$ we see that (24) implies

$$(27) \quad \|H[g]\|_K \leq [1 - e^{\lambda_{ii}(T_K - S_K)}] \|g\|_K.$$

Since $1 - e^{\lambda_{ii}(T_K - S_K)} < 1$, H contracts $\|\cdot\|_K$. Let H^k denote the k th iterate of the operator H . For brevity let

$$(28) \quad y^i(t, x) = \varphi(t^i(t, x), x^i(t, x)) e^{\lambda_{ii}(t^i(t, x) - t)}.$$

Then (21)–(24) and (28) imply

$$(29) \quad \psi^i(t, x) = y^i(t, x) + \sum_{k=1}^{\infty} H^k[y]^i(t, x).$$

Each term of the series is continuous. Since H contracts $\|\cdot\|_K$, comparison of the seminorms of the terms of the series with the geometric series implies the series converges uniformly on the domain of influence of each compact set K .

3.3. Integral equation for $\psi^i(t, x)$. Notice that since the series in (29) converges uniformly on compact subsets of G , we have that $\psi^i(t, x)$ satisfies the operator equation

$$\psi^i(t, x) = y^i(t, x) + H[\psi]^i(t, x).$$

Written out this becomes the following theorem.

THEOREM 2. *The function $\psi^i(t, x)$ satisfies the integral equation*

$$(30) \quad \begin{aligned} \psi^i(t, x) = & \varphi(t^i(t, x), x^i(t, x)) e^{\lambda_{ii}[t^i(t, x) - t]} \\ & + \sum_{j \neq i} \int_t^{t^i(t, x)} \lambda_{ij} e^{\lambda_{ii}(s - t)} \psi^j(s, x^i(s; t, x)) ds. \end{aligned}$$

COROLLARY. *The function $\psi^i(t, x)$ satisfies the integral equation*

$$(31) \quad \psi^i(t, x) = \varphi(t^i(t, x), x^i(t, x)) + \sum_j \int_t^{t^i(t, x)} \lambda_{ij} \psi^j(s, x^i(s; t, x)) ds.$$

Proof. Notice that if $\psi^i(t, x)$ satisfies (30), an interchange of integration shows that

$$(32) \quad \begin{aligned} & \int_t^{t^i(t, x)} \lambda_{ii} \psi^i(v, x^i(v, t, x)) dv \\ &= \varphi(t^i(t, x), x^i(t, x)) \int_t^{t^i(t, x)} \lambda_{ii} e^{\lambda_{ii}[t^i(t, x) - v]} dv \\ &+ \sum_{j \neq i} \int_t^{t^i(t, x)} \int_v^{t^i(t, x)} \lambda_{ii} \lambda_{ij} e^{\lambda_{ii}(s - v)} \psi^j(s, x^i(s; t, x)) ds dv \\ &= \psi^i(t, x) - \varphi(t^i(t, x), x^i(t, x)) - \sum_{j \neq i} \int_t^{t^i(t, x)} \lambda_{ij} \psi^j(s, x^i(s; t, x)) ds \end{aligned}$$

which gives (31).

3.4. Differentiability of $\psi^i(t, x)$. Let K be an arbitrary compact subset of G . We shall show each term of the series (29) is continuously differentiable on $G - \mathcal{S}_n^i$ and that the series of partial derivatives converges uniformly on $K \cap (G - \mathcal{S}_n^i)$. Thus $\psi^i(t, x)$ will be continuously differentiable on $G - \mathcal{S}_n^i$ and have locally bounded partial derivatives. Notice from Theorem 1 that

$$\varphi(t^i(t, x), x^i(t, x)) e^{\lambda_{ii}(t^i(t, x) - t)}$$

is continuously differentiable in (t, x) on $G - \mathcal{S}_n^i$ and its partial derivatives are bounded on $(G - \mathcal{S}_n^i) \cap K$.

Let D denote the space of continuous k -dimensional vector functions $d(t, x)$ defined on G such that, for each i , the i th component $d^i(t, x)$ of $d(t, x)$ is continuously differentiable with respect to x on the set $G - \mathcal{S}_n^i$ and these partial derivatives of $d^i(t, x)$ are locally bounded. Let us show that H maps D into D .

Let C denote a cell of $G - \mathcal{S}_n$, and $C = C_1, C_2, \dots, C_j$ denote the succession of type (i) cells of the common decomposition of G through which $x^i(s; t, x)$ passes to reach M , when the initial conditions (t, x) belong to C . Let $t_0(t, x) = t$ and $t_k(t, x), k = 1, \dots, q$, denote the times at which $x^i(s; t, x)$ reaches $\pi(C_k)$.

Now

$$\begin{aligned}
 H[d]^i(t, x) &= \sum_{j \neq i} \int_t^{t_i(t, x)} \lambda_{ij} e^{\lambda_{ij}(s-t)} d^j(s, x^i(s; t, x)) dx \\
 (33) \qquad &= \sum_{j \neq i} \sum_{k=1}^q \int_{t_{k-1}(t, x)}^{t_k(t, x)} \lambda_{ij} e^{\lambda_{ij}(s-t)} d^j(s, x^i(s; t, x)) ds.
 \end{aligned}$$

Let us check that the individual terms in the integrals of the sum are differentiable and conditions for "differentiating under the integral sign" are satisfied. Since by assumption (E), $x^i(s; t, x)$ meets \mathcal{S}_n^j , $j \neq i$, at only a finite number of points and $\mathcal{S}_n^j \subset \mathcal{S}_n$, these points must be points of type (ii) cells for $x^i(s; t, x)$ of the common decomposition of G . Hence each of the type (i) cells C_k through which $x^i(s; t, x)$ passes is contained in some j -cell of $G - \mathcal{S}_n^j$ on which $d^j(s, x)$ is continuously differentiable. For (t, x) in a compact neighborhood N contained in C the set

$$(34) \quad \{(s, y) : (s, y) = (s, x^i(s; t, x)) \text{ for some } (t, x) \in N \text{ and } t_{k-1}(t, x) \leq s \leq t_k(t, x)\}$$

is compact, so the partial derivatives of $d^j(s, x)$, $j \neq i$, are bounded on this set. From Theorem 1 the trajectory $x^i(s; t, x)$ has partial derivatives $\delta x^i(s; t, x)$ with respect to x if $t_{k-1}(t, x) < s < t_k(t, x)$ which are solutions of (10). These partial derivatives are bounded on (34). By Theorem 1, $t_k(t, x)$ is differentiable with respect to (t, x) . Hence the integrals of the sum in (33) satisfy conditions of formulas for differentiation of integrals with respect to parameters. Carrying out this differentiation term by term, we see that our assumptions imply the resulting partial derivatives of each term are continuous in (t, x) on C . Summing and canceling terms with opposite signs gives for the partial derivative with respect to x

$$\begin{aligned}
 \frac{\partial}{\partial x} H[d]^i(t, x) &= \sum_{j \neq i} \int_t^{t_i(t, x)} \lambda_{ij} e^{\lambda_{ij}(s-t)} d_{x^i}^j(s, x^i(s; t, x)) \delta x^i(s; t, x) ds \\
 (35) \qquad &+ \sum_{j \neq i} \lambda_{ij} e^{\lambda_{ij}(t_i(t, x)-t)} d^j(t_i(t, x), x^i(t, x)) \frac{\partial t_i(t, x)}{\partial x}.
 \end{aligned}$$

Hence H maps D into D .

Recall that the continuity of $t^i(t, x)$ and $x^i(t, x)$ imply that

$$(t^i(t, x), x^i(t, x)) = (t, x) \quad \text{for } (t, x) \in M.$$

Hence

$$(36) \qquad H^k[y]^i(t, x) \equiv 0 \quad \text{for } (t, x) \in M, \quad k \geq 1,$$

and

$$\begin{aligned}
 (37) \qquad y^i(t, x) &= \varphi(t^i(t, x), x^i(t, x)) e^{\lambda_{ii}(t^i(t, x)-t)} = \varphi(t, x) \\
 &\text{for } (t, x) \in M \text{ and all } i = 1, \dots, k.
 \end{aligned}$$

Differentiating the terms of (29) with respect to x , using (35), canceling a term from $\partial y^i(t, x)/\partial x$ with terms from $\partial H[y]^i(t, x)/\partial x$ through use of $\lambda_{ii} = -\sum_{j \neq i} \lambda_{ij}$ and

(37) gives the formal series

$$(38) \quad \varphi_t t_x^i + \varphi_x x_x^i + \sum_{k=0}^{\infty} \int_t^{t^i} \sum_{j \neq i} \lambda_{ij} e^{\lambda_{ij}(s-t)} H_x^k[y]^j(s, x^i(s; t, x)) \delta x^i(s; t, x) ds.$$

Let us show that this series converges uniformly on $(G - \mathcal{S}_n^i) \cap K$ for each compact subset K of G .

Let $\|\delta x\|_K$ denote the supremum of the sum of the absolute values of the elements of the matrix $\delta x^i(s; t, x)$, where the supremum is taken over elements in

$$\{(t, x, s, i) : (t, x) \in K, t \leq s \leq t^i(t, x), i \in I\}.$$

For a compact set K it is seen from (10)–(12) that $\|\delta x\|_K$ is finite. Let r denote the number of integers in I . Let B denote a number so that $\max \{\lambda_{ij}, \|\delta x\|_K\} \leq B/r$.

Let $\|(\partial/\partial x)H^n[y]^i(t, x)\|$ be the maximum of the absolute value of the elements in the matrix of partial derivatives

$$\frac{\partial}{\partial x} H^n[y]^i(t, x).$$

Let

$$T_K = \sup_{(t, x) \in I(K)} \max_i |t^i(t, x)|.$$

Let

$$\left\| \frac{\partial}{\partial x} H^n[y] \right\|_{tK} = \max_j \sup_{\{(t, x) \in I(K) \cap G - S_n^j\}} \left\| \frac{\partial}{\partial x} H^n[y]^j(t, x) \right\|$$

with the convention that the supremum is zero if it is taken over the empty set. Now if $n > 1$, by (35) and (36),

$$(39) \quad \frac{\partial}{\partial x} H^n[y]^i(t, x) = \sum_{j \neq i} \int_t^{t^i(t, x)} \lambda_{ij} e^{\lambda_{ij}(s-t)} \frac{\partial}{\partial x} H^{n-1}[y]^j(s, x^i(s; t, x)) \delta x^i(s; t, x) ds.$$

Since $x^i(s; t, x)$ belongs to $I(K) \cap G - S_n^j$ for all but finitely many s ,

$$(40) \quad \left\| \frac{\partial}{\partial x} H^n[y] \right\|_{tK} \leq B \int_t^{T_k} \left\| \frac{\partial}{\partial x} H^{n-1}[y] \right\|_{sK} ds.$$

Iterating formula (40) gives

$$(41) \quad \left\| \frac{\partial}{\partial x} H^n[y] \right\|_{tK} \leq B^{n-1} \int_t^{T_k} \int_{s_1}^{T_k} \cdots \int_{s_{n-1}}^{T_k} \left\| \frac{\partial}{\partial x} H[y] \right\|_{sK} ds ds_1 \cdots ds_{n-1}.$$

Therefore,

$$(42) \quad \left\| \frac{\partial}{\partial x} H^n[y] \right\|_{tK} \leq B^{n-1} \frac{(T_k - t)^{n-1}}{(n-1)!} \left[\sup_{t \leq s \leq T_k} \left\| \frac{\partial}{\partial x} H[y] \right\|_{sK} \right].$$

Define $S_K = \inf \{t : (t, x) \in K\}$; (42) implies

$$\max_j \sup_{\{(t, x) : (t, x) \in I(K) \cap G - S_n^j\}} \left\| \frac{\partial}{\partial x} H^n[y]^j(t, x) \right\| \leq B^n \frac{(T_k - S_K)^{n-1}}{(n-1)!} \sup_{S_K \leq s \leq T_k} \left\| \frac{\partial}{\partial x} H[y] \right\|_{sK}.$$

Hence comparison with the exponential series implies the i th component of each term of the series (38) converges uniformly on $I(K) \cap (G - \mathcal{S}_n^i)$. Hence $\psi^i(t, x)$ is continuously differentiable on $G - \mathcal{S}_n^i$ and has locally bounded partial derivatives.

Since we have just shown that $\psi^i(t, x)$ belongs to D , the integral equations (31) may be differentiated with respect to x using (35) to show that $\psi_x^i(t, x)$ satisfies the integral equations

$$(43) \quad \begin{aligned} \psi_x^i(t, x) = & \varphi_t(t^i(t, x), x^i(t, x)) \frac{\partial t^i(t, x)}{\partial x} + \varphi_x(t^i(t, x), x^i(t, x)) \frac{\partial x^i(t, x)}{\partial x} \\ & + \int_t^{t^i(t, x)} \sum_{j \neq i} \lambda_{ij} e^{\lambda_{ij}(s-t)} \psi_x^j(s, x^j(s; t, x)) \delta x^i(s; t, x) ds \end{aligned}$$

in the sense that for each i , (43) holds on $G - \mathcal{S}_n^i$.

Notice that an integral equation for $\psi_x^i(t, x)$ may also be obtained by differentiating (30) with respect to t . Next we shall show that (43) has a unique solution. We shall gather together the equations involved in (43).

For a point (t, x) of $G - \mathcal{S}_n^i$, let $x^i(s; t, x)$ be the solution of

$$(44) \quad \dot{x} = f^i(s, x)$$

with initial condition $x^i(t; t, x) = x$. Let $t_0(t, x) = t$ and $t_k(t, x)$, $k = 1, \dots, q-1$, be the times of crossing of $x^i(s; t, x)$ from i -cell to i -cell. Let $\delta x^i(s; t, x)$ be the solution of the matrix differential equation

$$(45) \quad \delta x^i(s) = f_{cx}^i(s, x^i(s; t, x)) \delta x^i(s)$$

on the interval $t_{k-1}(t, x) < s < t_k(t, x)$ which has the value $\delta x^i(t; t, x) = I$, where I is the identity matrix, and at the times $t_k(t, x)$, $k = 1, \dots, q-1$, has the jumps

$$(46) \quad \delta x^i(t_k^+; t, x) - \delta x^i(t_k^-; t, x) = [f_{cx}^i(t_k, x_k) - f_{cx}^i(t_{k-1}, x_k)] \frac{\partial t_k(t, x)}{\partial x}.$$

For each $i \in I$, let $p^i(t, x)$ be a continuous locally bounded real-valued function defined on $G - \mathcal{S}_n^i$ satisfying

$$(47) \quad \begin{aligned} p^i(t, x) = & \varphi_t(t^i(t, x), x^i(t, x)) \frac{\partial t^i(t, x)}{\partial x} + \varphi_x(t^i(t, x), x^i(t, x)) \frac{\partial x^i(t, x)}{\partial x} \\ & + \int_t^{t^i(t, x)} \left[\sum_{j \neq i} \lambda_{ij} e^{\lambda_{ij}(s-t)} p^j(s, x^j(s; t, x)) \delta x^i(s; t, x) \right] ds. \end{aligned}$$

THEOREM 3. *Let $p^i(t, x)$ be a solution of the system of equations (44)–(47). Then*

$$(48) \quad p^i(t, x) = \psi_x^i(t, x) \quad \text{on } G - \mathcal{S}_n^i \quad \text{for each } i \in I.$$

Proof. Let K be a compact subset of G and let $I(K)$ denote the domain of influence of K . Let V denote the space of vector functions $\{h^i(t, x)\}$ such that for each $i \in I$, $h^i(t, x)$ is a continuous locally bounded real-valued function defined on $I(K) - \mathcal{S}_n^i$. Let $S_K = \min \{t : (t, x) \in K\}$ and $T_K = \max \{t : (t, x) \in I(K)\}$. Define

$T: V \rightarrow V$ by

$$(49) \quad T[h]^i(t, x) = \int_t^{t^i(t, x)} \sum_{j \neq i} \lambda_{ij} e^{\lambda_{ii}(s-t)} h^i(s, x^i(s; t, x)) \delta x^i(s; t, x) ds.$$

It follows from Theorem 1(a) that the image of T is again in V . Let η be a number such that

$$(50) \quad \max_i \sup_{\substack{(t, x) \in K \cap G - \mathcal{S}_n^i \\ t \leq s \leq t_q(t, x)}} |\delta x^i(s; t, x) \sum_{j \neq i} \lambda_{ij}| \leq \eta.$$

On V define seminorms by

$$(51) \quad \|h\|_K = \max_{i \in I} \sup_{(t, x) \in I(K) - S_n^i} |h^i(t, x) e^{\eta(t - S_K)}|.$$

Notice that if $h^i(t, x)$ is a solution of

$$(52) \quad h^i(t, x) = T[h]^i(t, x),$$

then

$$\begin{aligned} |h^i(t, x) e^{\eta(t - S_K)}| &\leq e^{\eta(t - S_K)} \int_t^{t^i(t, x)} \eta |h^i(s, x^i(s; t, x))| ds \\ (53) \quad &= e^{\eta(t - S_K)} \int_t^{t^i(t, x)} \eta e^{-\eta(s - S_K)} |h^i(s; x^i(s; t, x)) e^{\eta(s - S_K)}| ds \\ &= [1 - e^{\eta(t^i(t, x) - t)}] \|h\|_K \end{aligned}$$

or

$$(54) \quad \|h\|_K \leq [1 - e^{\eta(T_K - S_K)}] \|h\|_K.$$

Now since our assumptions imply $1 - e^{\eta(T_K - S_K)} < 1$, (54) implies that $\|h\|_K = 0$ for any solution of (52). If $p^i(t, x)$ is a solution of (47), since $\psi_x^i(t, x)$ is also a solution of (47), $p^i(t, x) - \psi_x^i(t, x)$ is a solution of (52). Hence the conclusion of the theorem follows from the assertion that

$$(55) \quad \|p^i(t, x) - \psi_x^i(t, x)\|_K = 0.$$

4. Systems of hyperbolic partial differential equations with discontinuous coefficients. In § 3.4 we have proved that $\psi^i(t, x)$ is continuously differentiable on $G - \mathcal{S}_n^i$. Thus if $(s, x^i(s; t, x)) \in G - \mathcal{S}_n^i$, $\psi^i(s, x^i(s; t, x))$ is differentiable in s . Using formula (31) we have

$$(56) \quad \frac{d}{ds} \psi^i(s, x^i(s; t, x)) = \psi_t^i + \psi_x^i \frac{dx^i(s)}{dt} = - \sum_j \lambda_{ij} \psi^j.$$

Evaluating this at $s = t$, we have the following theorem.

THEOREM 4. *The functions $\psi^i(t, x)$ satisfy the system of partial differential equations*

$$(57) \quad \psi_t^i(t, x) + \psi_x^i(t, x) f^i(t, x) + \sum_j \lambda_{ij} \psi^j(t, x) = 0$$

for $(t, x) \in G - \mathcal{S}_n^i, i \in I$. The boundary condition

$$(58) \quad \psi^i(t, x) = \varphi(t, x)$$

for $(t, x) \in M$ is satisfied.

Thus a solution of the integral equations (30) satisfies the partial differential equations (57) with boundary conditions (58). We could ask conversely if each such solution of (57) and (58) is a solution of (30). Theorem 5 gives an affirmative answer to this question.

THEOREM 5. *Let G be an open subset of E^{n+1} . Let $\mathcal{R}^i, i \in I$, be subsets of G . It will be assumed for each i and each compact subset K of G that $\mathcal{R}^i \cap K$ is contained in the zero sets of a finite number of continuously differentiable functions $\gamma_{ik}(t, x)$ whose gradient vectors $(\gamma_{ikt}^i, \text{grad}_x \gamma_{ik}^i)$ do not vanish on \mathcal{R}^i . That is,*

$$(59) \quad \mathcal{R}^i \cap K \subset \bigcup_k \{(t, x) : \gamma_{ik}(t, x) = 0\}.$$

For each i let $h^i(t, x)$ be a continuous function on G which is continuously differentiable on $G - \mathcal{R}^i$ whose partial derivatives are bounded on $(G - \mathcal{R}^i) \cap K$ for any compact set K . For each i , let the partial differential equation

$$(60) \quad h_t^i(t, x) + h_x^i(t, x)f^i(t, x) + \sum_j \lambda_{ij}h^j(t, x) = 0$$

be satisfied on $G - \mathcal{R}^i$. Let the functions $f^i(t, x)$ and constants λ_{ij} satisfy the assumptions (A)–(F) and (18) and (19) made in §§ 1 and 2. Let

$$(61) \quad h^i(t, x) = \varphi(t, x) \quad \text{on } M \quad \text{for each } i \in I.$$

Then the functions $h^i(t, x)$ satisfy the system (30) of integral equations

$$h^i(t, x) = \varphi(t^i(t, x), x^i(t, x))e^{\lambda_{ii}(s-t)} + \sum_{j \neq i} \int_t^{t^i(t, x)} \lambda_{ij} e^{\lambda_{ii}(s-t)} h^j(s, x^i(s; t, x)) ds.$$

Remark. Notice that the integral equations (30) have a unique solution and the conditional expectation $\psi^i(t, x)$ satisfies the integral equations. Thus a consequence of the theorem is that any solution of the system of partial differential equations (57) with boundary condition (58) can be interpreted as a conditional expectation. Notice also that the discontinuities of the partial derivatives of $\psi^i(t, x)$ are contained in \mathcal{S}_n^i , the set on which $f^i(t, x)$ may have discontinuities. Thus the theorem implies that it may always be assumed that $\mathcal{R}^i \subset \mathcal{S}_n^i$ or that roughly a solution of a system of partial differential equations of the type (57), (58) will be continuously differentiable on regions where its coefficient functions are continuously differentiable.

To prove Theorem 5 a lemma is needed. This lemma can be proved in a manner entirely analogous to a corresponding treatment in Boltyanskii [3, Lemma 1, p. 330].

LEMMA 1. *Let (t_0, x_0) be a point of G and $x^i(s; t_0, x_0)$ a solution of (5) on $[t_0, t^i(t_0, x_0)]$. Let C_0, \dots, C_{q-1} be the succession of type (i), i -cells through which $x^i(s; t_0, x_0)$ passes, let t_1, \dots, t_{q-1} be the points of crossing of $x^i(s; t_0, x_0)$ from*

i -cell to i -cell, and let $t_q = t^i(t_0, x_0)$. Considering these times as fixed define

$$(62) \quad \tilde{f}^i(s, x) = f_{c_j}^i(s, x) \quad \text{if } t_{j-1} \leq s < t_j.$$

Then there is some neighborhood N of (t_0, x_0) such that for $t, x \in N$ the solution $\tilde{x}^i(s; t, x)$ of

$$(63) \quad \dot{x} = \tilde{f}^i(s, x) \quad \text{with } \tilde{x}^i(t; t, x) = x$$

exists on $[t, t_q]$ and in each subneighborhood of N there is some t, x so that $\tilde{x}^i(s; t, x)$ intersects $R^i \cup \mathcal{S}_n^i$ in a finite or empty set of points.

Theorems on continuity of solutions of (63) with respect to initial conditions imply

$$(64) \quad \lim_{(t, x) \rightarrow (t_0, x_0)} \sup_{t_0 \leq s \leq t_q} |\tilde{x}^i(s; t, x) - x^i(s; t_0, x_0)| = 0.$$

Proof of Theorem 5. By Lemma 1 there is a sequence of points (t_n, x_n) approaching (t, x) such that $(s, \tilde{x}^i(s; t_n, x_n))$ intersects R^i at a finite number of points on $t_n \leq s \leq t_q$.

Then by the chain rule,

$$h^i(s, \tilde{x}^i(s; t_n, x_n))$$

is piecewise continuously differentiable in s with locally bounded derivative given by

$$(65) \quad h_t^i(s, \tilde{x}^i(s; t_n, x_n)) + h_x^i(s, \tilde{x}^i(s; t_n, x_n)) f^i(s, \tilde{x}^i(s; t_n, x_n))$$

except at the finitely many points at which

$$(s, \tilde{x}^i(s; t_n, x_n)) \in R^i \cup \mathcal{S}_n^i;$$

hence $h^i(s, \tilde{x}^i(s; t_n, x_n))$ is absolutely continuous. Substituting $\tilde{x}^i(s; t_n, x_n)$ for x in (60), multiplying by the integrating factor $e^{\lambda_{ii}(s-t_0)}$ and integrating from t_0 to t_q gives

$$(66) \quad e^{\lambda_{ii}(t_q-t_0)} h^i(t_q, \tilde{x}^i(t_q; t_n, x_n)) - h^i(t_0, \tilde{x}^i(t_0; t_n, x_n)) \\ = \int_{t_0}^{t_q} \sum_{j \neq i} \lambda_{ij} e^{\lambda_{ii}(s-t_0)} h^j(s, \tilde{x}^i(s; t_n, x_n)) ds.$$

Since the $h^i(s, x)$ are continuous on G , taking limits as n becomes infinite using (64) gives

$$(67) \quad h^i(t_0, x_0) = e^{\lambda_{ii}(t_q-t_0)} h^i(t_q, x^i(t_q; t_0, x_0)) \\ + \int_{t_0}^{t_q} \sum_{j \neq i} \lambda_{ij} e^{\lambda_{ii}(s-t_0)} h^j(s, x^i(s; t_0, x_0)) ds.$$

Recalling the meaning of the notations in (30) and (67), we see that they are identical expressions and hence the theorem is proved.

The following theorem is a maximal principle for partial differential operators of the form (60).

THEOREM 6. Let sets $R^i \subset G$ be as in Theorem 5. Let $h^i(t, x)$, $i \in I$, be continuous

functions defined on G such that $h^i(t, x)$ is continuously differentiable on $G - R^i$ and has locally bounded partial derivatives. If for $i \in I$,

$$(68) \quad h_t^i(t, x) + h_x^i(t, x)f^i(t, x) + \sum_j \lambda_{ij}h^j(t, x) \geq 0 \quad \text{on } G - R^i$$

and

$$(69) \quad h^i(t, x) \leq 0 \quad \text{if } (t, x) \in M,$$

then

$$(70) \quad h^i(t, x) \leq 0 \quad \text{on } G.$$

Proof. Define t_n, x_n and $\tilde{x}^i(s; t_n, x_n)$ as in the proof of Theorem 5. Then

$$(71) \quad \begin{aligned} \frac{d}{ds} [e^{\lambda_{ii}(s-t)} h^i(s, \tilde{x}^i(s; t_n, x_n))] &= e^{\lambda_{ii}(s-t)} [h_t^i(s, \tilde{x}^i(s; t_n, x_n)) \\ &\quad + h_x^i(s, \tilde{x}^i(s; t_n, x_n)) + \lambda_{ii}h^i(s, \tilde{x}^i(s; t_n, x_n))] \\ &\geq -e^{\lambda_{ii}(s-t)} \sum_{j \neq i} \lambda_{ij}h^j(s, \tilde{x}^i(s; t_n, x_n)) \end{aligned}$$

except for the finitely many values of s for which $\tilde{x}^i(s; t_n, x_n) \in R^i$. Integrating both sides of (71) gives

$$(72) \quad \begin{aligned} h^i(t, \tilde{x}(t; t_n, x_n)) - e^{\lambda_{ii}(t_q-t)} h^i(t_q, \tilde{x}(t_q; t_n, x_n)) \\ \leq \sum_{j \neq i} \int_t^{t_q} \lambda_{ij} e^{\lambda_{ii}(s-t)} h^j(s, \tilde{x}^i(s; t_n, x_n)) ds. \end{aligned}$$

Passing to the limit as n becomes infinite, we have

$$(73) \quad h^i(t, x) \leq \sum_{j \neq i} \int_t^{t^i(t, x)} \lambda_{ij} e^{\lambda_{ii}(s, t)} h^j(s, x^i(s; t, x)) ds.$$

Formula (73) can be rewritten as

$$(74) \quad h^i(t, x) \leq H[h]^i(t, x).$$

Notice that if $g^i(t, x) \geq 0$ that $H[g]^i(t, x) \geq 0$. Hence applying the operator H to the inequality (74), we obtain

$$(75) \quad H^2[h]^i(t, x) \geq H[h]^i(t, x) \geq h^i(t, x).$$

Proceeding inductively we obtain

$$(76) \quad H^n[h]^i(t, x) \geq h^i(t, x).$$

Now (27) implies that $H^n[h]^i(t, x)$ converges uniformly to zero on each compact subset of G as n becomes infinite. Hence we obtain (70).

Part 2

Optimal Control of Differential Equations with Jump Disturbances

Let U denote a closed subset of E^m and let $f^i(t, x, u)$, $i \in I$,

$$f^i: G \times U \rightarrow E^n$$

be continuously differentiable functions.

Let \mathcal{U} be a class of control functions $u^i(t, x)$, $i \in I$,

$$u^i: G \rightarrow U$$

such that for each control function the functions

$$(77) \quad f^i(t, x, u^i(t, x)), \quad i \in I,$$

satisfy conditions (A) through (F) of Part 1.

With regard to condition (A) we shall assume for each $i \in I$ that there are sets \mathcal{S}_n^i as described in (A) and that for each cell C there is a continuously differentiable function $u_c^i(t, x)$ defined on a neighborhood of the closure of the cell such that $u_c^i(t, x) = u^i(t, x)$ on the cell C .

Consider the system of controlled stochastic processes

$$(78) \quad \dot{x}(t) = f^{r(t)}[t, x(t), u^{r(t)}(t, x(t))];$$

as in Part 1, $r(t)$ is a Markov process with state space I .

Let M be a terminal set as defined in Part 1, let τ_u be the first time the solution of (78) reaches M , and let $\varphi(t, x)$ be a continuously differentiable function defined on G .

The optimization problem is to find the control $u^i(t, x) \in \mathcal{U}$ so that simultaneously for each initial condition (t, x, i) ,

$$(79) \quad E\{\varphi(\tau_u, x(\tau_u)) | x(t) = x, r(t) = i\}$$

is a minimum. In (79), $x(t)$ is the solution of (78), with initial condition $x(t) = x$, corresponding to the control $u^i(t, x)$.

A control $u^i(t, x) \in \mathcal{U}$ is called an optimal feedback control (or more briefly optimum) if it simultaneously minimizes (79) over the class \mathcal{U} for each (t, x) in G and $i \in I$.

Remark. Notice that since we are assuming that for each control $u^i(t, x)$ of \mathcal{U} the conditions (A)–(F) of Part 1 are satisfied for the function

$$f^i(t, x, u^i(t, x)),$$

the results of Part 1 apply if in each of the theorems given there we make the replacements

$$f^i(t, x) \quad \text{by} \quad f^i(t, x, u^i(t, x))$$

and

$$f_{cx}(t, x) \quad \text{by} \quad f_x^i(t, x, u^i(t, x)) + f_u^i(t, x, u^i(t, x))u_{cx}(t, x).$$

When we speak in this section of a statement or theorem given in Part 1, we will always understand that this replacement has been made.

THEOREM 7. *A necessary and sufficient condition that a control $u^i(t, x) \in \mathcal{U}$ be optimum is that for each $i \in I$ its performance function*

$$(80) \quad \psi^i(t, x) = E\{\varphi(\tau_u, x(\tau_u)) | x(t) = x, r(t) = i\}$$

satisfy the partial differential equations

$$(81) \quad \min_{u \in U} \{ \psi_t^i(t, x) + \psi_x^i(t, x) f^i(t, x, u) + \sum_j \lambda_{ij} \psi^j(t, x) \} \\ = \psi_t^i(t, x) + \psi_x^i(t, x) f^i(t, x, u^i(t, x)) + \sum_j \lambda_{ij} \psi^j(t, x) = 0$$

on $G - \mathcal{S}_n^i$, and the boundary condition

$$(82) \quad \psi^i(t, x) = \varphi(t, x) \quad \text{if } (t, x) \in M.$$

Proof. Necessity. Let (t, x) be on a point of $G - \mathcal{S}_n^i$. Then there is a neighborhood of (t, x) which is contained in $G - \mathcal{S}_n^i$. Let $u^i(s, x)$ be an optimum control, u a point of U , and \tilde{u} the control defined by

$$(83) \quad \tilde{u}^i(s, x) = \begin{cases} u & \text{if } t \leq s \leq t + h, \\ u^i(s, x) & \text{otherwise.} \end{cases}$$

Let $\psi^i(t, x)$ denote the performance function (80) corresponding to $u^i(t, x)$, and $\tilde{\psi}^i(t, x)$ the performance function corresponding to $\tilde{u}^i(t, x)$ denote the solution of

$$(84) \quad \dot{x} = f^i(s, x, \tilde{u}^i(s, x))$$

with initial condition $\tilde{x}^i(t; t, x) = x$ and $\tilde{t}^i(t, x)$ the first time that $\tilde{x}^i(t, x)$ reaches M . By the corollary of Theorem 2,

$$(85) \quad \tilde{\psi}^i(t, x) = \varphi(\tilde{t}^i(t, x), \tilde{x}^i(t, x)) e^{\lambda_{ii}(\tilde{t}^i(t, x) - t)} + \int_t^{\tilde{t}^i(t, x)} \sum_j \lambda_{ij} \tilde{\psi}^j(s, \tilde{x}^i(s; t, x)) ds.$$

Since (t, x) is an interior point of $G - \mathcal{S}_n^i$ and $M \subset \mathcal{S}_n^i$, there is an $h > 0$ such that $\tilde{x}^i(s; t, x)$ lies in $G - M$ for $t \leq s \leq t + h$. Since for $s \geq t + h$,

$$\tilde{x}^i(s; t, x) = x^i(s; t + h, \tilde{x}^i(t + h; t, x)),$$

we have

$$(\tilde{t}^i(t, x), \tilde{x}^i(t, x)) = (t^i(t + h, \tilde{x}^i(t + h; t, x)), x^i(t + h, \tilde{x}^i(t + h; t, x))).$$

Hence

$$(86) \quad \tilde{\psi}^i(t, x) = \psi^i(t + h, \tilde{x}^i(t + h; t, x)) + \int_t^{t+h} \sum_j \lambda_{ij} \tilde{\psi}^j(s, \tilde{x}^i(s; t, x)) ds.$$

Since $\tilde{\psi}^j(s, x)$ is locally bounded, ψ^i is continuous at (t, x) , and $\tilde{x}^i(t + h; t, x)$ converges to x as h decreases to zero, (86) implies that as h converges to zero $\tilde{\psi}^i(t, x)$ converges uniformly on compact subsets of G to $\psi^i(t, x)$. Since $u^i(t, x)$ is an optimal control, (86) implies that

$$(87) \quad \psi^i(t, x) \leq \psi^i(t + h, \tilde{x}^i(t + h; t, x)) + \int_t^{t+h} \sum_j \lambda_{ij} e^{\lambda_{ii}(s-t)} \tilde{\psi}^j(s, \tilde{x}^i(s; t, x)) ds.$$

Since $\psi^i(t, x)$ is continuously differentiable on $G - \mathcal{S}_n^i$, and $\tilde{\psi}^i(t, x)$ converges to $\psi^i(t, x)$, (87) implies that

$$(88) \quad 0 \leq \psi_t^i(t, x) + \psi_x^i(t, x) f^i(t, x, u) + \sum_j \lambda_{ij} \psi^j(s, x).$$

Combining (88) with Theorem 4 of Part 1 gives (81).

Sufficiency. Let $\tilde{u}^i(t, x)$ be any other control in \mathcal{U} and $\tilde{\psi}^i(t, x)$ its corresponding performance. Let \mathcal{S}_n^i and $\tilde{\mathcal{S}}_n^i$ be the sets of condition (A) of Part 1 for $u^i(t, x)$ and $\tilde{u}^i(t, x)$. Let $R^i = \mathcal{S}_n^i \cup \tilde{\mathcal{S}}_n^i$. By Theorem 4 of Part 1 and (81),

$$\begin{aligned} \psi^i(t, x) - \tilde{\psi}^i(t, x) + f^i(t, x, \tilde{u}^i(t, x))[\psi_x^i(t, x) - \tilde{\psi}_x^i(t, x)] \\ + \sum_j \lambda_{ij}[\psi_x^i(t, x) - \tilde{\psi}_x^i(t, x)] \geq 0 \end{aligned}$$

on $G - R^i$, and $\psi^i(t, x) - \tilde{\psi}^i(t, x)$ satisfies the conditions of the maximal principle of Theorem 7. Hence

$$\psi^i(t, x) \leq \tilde{\psi}^i(t, x) \quad \text{on } G,$$

which is the asserted optimality statement.

Instead of considering controls of the form $u^i(t, x)$, controls which had one functional form until a jump of $r(\cdot)$ and then another functional form after each jump could be considered. A simple example of this type which can be treated by the methods we have already developed is the following. Let K be a positive integer and

$$\mathcal{M} = \{u^{ki}(s, x) : 0 \leq k \leq K, i \in I\}$$

be a doubly indexed family of control functions so that the functions

$$f^i(s, x, u^{ki}(s, x))$$

satisfy conditions (A)–(F) of Part 1 with the double index ki replacing the single index i . Define

$$n(s) = \min \{K, \text{number of jumps of } r(s) \text{ on } [0, s]\}.$$

Then $(n(s), r(s))$ is a finite state Markov process. The results of Part 1 go through (suitably modified) for the stochastic differential equation

$$\dot{x}(s) = f^{r(s)}(s, x(s), u^{n(s), r(s)}(s, x(s))).$$

Solutions of this differential equation depend on the past of $r(\cdot)$. Controls of the form $u^i(t, x)$ are in the class \mathcal{M} ; they are constant in the index k .

It can be seen from Theorem 7 modified for the class \mathcal{M} of controls that the following theorem holds.

THEOREM 8. *If there is an optimal control in \mathcal{U} for the control problem formulated for the differential equation*

$$\dot{x} = f^{r(s)}(s, x, u^{r(s)}(s, x)),$$

that control is also optimal in the class \mathcal{M} for the corresponding control problem formulated for the differential equation

$$\dot{x} = f^{r(s)}(s, x, u^{n(s), r(s)}(s, x)).$$

Thus if there is an optimum in the class \mathcal{U} , no advantage can be gained by going to the wider class \mathcal{M} .

Theorem 3 of Part 1 asserts that functions $p^i(t, x)$ which are solutions of the system of equations (44)–(47) agree on $G - \mathcal{S}_n^i$ with $\psi_x^i(t, x)$. This allows us to

recast Theorem 7 in the following form.

THEOREM 9. *A necessary and sufficient condition that a control $u^i(t, x)$ be optimum is that for solutions $x^i(s; t, x)$ of*

$$(89) \quad \dot{x}^i = f^i(t, x, u^i(t, x))$$

with initial condition $x^i(t; t, x) = t$, for solutions $\delta x^i(s; t, x)$ of

$$(90) \quad \begin{aligned} \delta \dot{x}^i = & [f_x^i(s, x^i(s; t, x), u^i(s, x^i(s; t, x))) \\ & + f_u^i(s, x^i(s; t, x), u^i(s, x^i(s; t, x)))u_{c_k x}(s, x^i(s; t, x))] \delta x^i \end{aligned}$$

with initial condition $\delta x^i(s; t, x) = 1$ and jump conditions

$$(91) \quad \begin{aligned} \delta x^i(t_k^+; t, x) - \delta x^i(t_k^-; t, x) = & [f^i(t_k, x_k; u_{c_k}^i(t_k, x_k)) \\ & - f^i(t_k, x_k, u_{c_{k+1}}^i(t_k, x_k))] \frac{\partial t_k(t, x)}{\partial x} \end{aligned}$$

at the times t_k of crossing of $x^i(s; t, x)$ from type (i) i -cell C_k to type (i) i -cell C_{k+1} of $x^i(s; t, x)$ and for solutions $p^i(t, x)$ of the system of integral equations

$$(92) \quad \begin{aligned} p^i(t, x) = & \varphi_t(t^i(t, x), x^i(t, x)) \frac{\partial t^i(t, x)}{\partial x} + \varphi_x(t^i(t, x), x^i(t, x)) \frac{\partial x^i(t, x)}{\partial x} \\ & + \int_t^{t^i(t, x)} \left[\sum_{j \neq i} \lambda_{ij} e^{\lambda_{ij}(s-t)} p^j(s, x^i(s; t, x)) \delta x^i(s; t, x) \right] ds, \end{aligned}$$

that

$$(93) \quad \min_{u \in U} p^i(t, x) f^i(t, x, u) = p^i(t, x) f^i(t, x, u^i(t, x)) \quad \text{on } G - \mathcal{S}_n^i.$$

Proof. By Theorem 3,

$$p^i(t, x) = \psi_x^i(t, x).$$

Thus (93) implies

$$(94) \quad \begin{aligned} \min_{u \in U} \{ & \psi_t^i(t, x) + \psi_x^i(t, x) f^i(t, x, u) + \sum_j \lambda_{ij} \psi^j(t, x) \} \\ = & \{ \psi_t^i(t, x) + \psi_x^i(t, x) f^i(t, x, u^i(t, x)) + \sum_j \lambda_{ij} \psi^j(t, x) \}. \end{aligned}$$

Theorem 4 implies that the right-hand side of (94) equals zero. Hence the theorem follows from Theorem 7.

It is natural to ask if there is a theorem analogous to Theorem 9 which more closely resembles the maximum principle of deterministic control theory. Proceeding toward this goal, consider the following systems of equations.

For a point (t, x) of $G - \mathcal{S}_n^*$, let $x^i(s; t, x)$ be the solution of

$$(95) \quad \dot{x} = f^i(s, x, u^i(s, x))$$

with initial condition $x^i(t; t, x) = x$. Let $t_0(t, x) = t$ and $t_k(t, x)$, $k = 1, \dots, q$, denote the times at which $x^i(s; t, x)$ reaches $\pi(C_k)$. Let $\theta_k(t, x)$, $k = 1, \dots, q$, denote the additional component of the determining function of $\pi(C_k)$ over that of C_k .

Let $\mu_q(t_q, x_q)$ be a solution of

$$(96) \quad \begin{aligned} & \mu_q(t_q, x_q) [\theta_{qx}(t_q, x_q) f^i(t_q, x_q, u^i(t_q, x_q)) + \theta_{qt}(t_q, x_q)] \\ & + \varphi_x(t_q, x_q) f^i(t_q, x_q, u^i(t_q, x_q)) + \varphi_t(t_q, x_q) = 0 \end{aligned}$$

on $\pi(C_q)$. Let

$$\begin{aligned} p^i(t_k, x_k)^+ &= \lim_{s \downarrow t_k} p^i(s, x^i(s; t, x)), \\ p^i(t_k, x_k)^- &= \lim_{s \uparrow t_k} p^i(s, x^i(s; t, x)). \end{aligned}$$

Let $\mu_k(t_k, x_k)$ be a solution of

$$(97) \quad \begin{aligned} & p^i(t_k, x_k)^+ [f^i(t_k, x_k, u_{c_k}^i(t_k, x_k)) - f^i(t_k, x_k, u_{c_{k+1}}^i(t_k, x_k))] \\ & = \mu(t_k, x_k) [\theta_{kx}(t_k, x_k) f^i(t_k, x_k, u_{c_k}^i(t_k, x_k)) + \theta_{kt}(t_k, x_k)]. \end{aligned}$$

Let $p^i(t, x)$ satisfy the system of integral equations

$$(98) \quad \begin{aligned} p^i(t, x) &= \varphi_x(t_q, x_q) + \sum_{k=1}^q \mu_k(t_k, x_k) \theta_{kx}(t_k, x_k) \\ &+ \int_t^{t^i(t, x)} [p^i(s, x^i(s; t, x)) f_x^i(s, x^i(s; t, x)) u^i(s, x^i(s; t, x))] \\ &+ \sum_j \lambda_{ij} p^j(s, x^i(s; t, x))] ds \\ &+ \sum_{k=1}^q \int_{t_{k-1}(t, x)}^{t_k(t, x)} p^i(s, x^i(s; t, x)) f_u^i(s, x^i(s; t, x), u^i(s, x^i(s; t, x))) \\ &\quad \cdot u_{c_k x}^i(s, x^i(s; t, x)) ds. \end{aligned}$$

THEOREM 10. *Equations (95)–(98) have a solution. Any solution $p^i(t, x)$ of (95)–(98) is also a solution of (89)–(92).*

Theorem 10 is proved in Appendix B.

The next theorem follows immediately from Theorems 9 and 10.

THEOREM 11. *A necessary and sufficient condition that a control $u^i(t, x) \in \mathcal{U}$ be optimum is that there exist functions $p^i(t, x)$ which are solutions of (95)–(98) and (93) holds on $G - \mathcal{S}_n^i$.*

Remark. Conditions (96) and (98) imply that

$$(99) \quad p^i(t_q, x_q) = \varphi_x(t_q, x_q) + \mu_q(t_q, x_q) \theta_{qx}(t_q, x_q)$$

and

$$(100) \quad p^i(t_q, x_q) f_{c_q}^i(t_q, x_q) = -[\varphi_t(t_q, x_q) + \mu_q(t_q, x_q) \theta_{qt}(t_q, x_q)].$$

The conditions (99), (100) imply the usual statement of the transversality conditions. Conditions (97) and (98) imply that

$$(101) \quad p^i(t_k, x_k)^+ - p^i(t_k, x_k)^- = \mu_k(t_k, x_k) \theta_{kx}(t_k, x_k)$$

and

$$(102) \quad p^i(t_k, x_k)^+ f_{c_{k+1}}^i(t_k, x_k) - p^i(t_k, x_k)^- f_{c_k}^i(t_k, x_k) = -\mu_k(t_k, x_k) \theta_{kt}(t_k, x_k).$$

These conditions (101), (102) imply the jump conditions [10, Formula (4.10), p. 303] given by S. Miraca.

Remarks. Theorem 11 is analogous to Theorems 6.1 and 6.2 of [10] in the deterministic case. Its derivation uses basically a dynamic programming or field theory argument. If one wishes to use Theorems 9 or 11 to compute optimal controls, condition (93) only determines the control on $G - \mathcal{S}_n^i$. According to Theorems 7, 9 or 11 any admissible control which satisfies respectively (81) or (93) on $G - \mathcal{S}_n^i$ and for which the terminal conditions are satisfied is optimal. This implies that the structure of the sets \mathcal{S}_n^i determines the control on that set. This is a difficulty of these conditions in that the determination of the sets \mathcal{S}_n^i and the value of the controls on them is not made explicit by the optimality conditions.

The deterministic Pontryagin principle can be used to determine another set of necessary and sufficient conditions for optimality. To avoid some technical difficulties, assume:

- (i) The terminal set M of the optimization problem contains

$$\{(t, x) \in G : t = T\}.$$

That is, if the terminal set has not been reached before T , the problem stops at time T .

- (ii) For any piecewise continuous open loop control $u(s)$ the trajectories of

$$\dot{x} = f^i(s, x, u(s))$$

remain in G for $t \leq s \leq T$.

Let $u^i(t, x)$ be an optimal control for the stochastic optimization problem and $\psi^i(t, x)$ the corresponding conditional expectations of the performance. Formulate the deterministic optimal control problem of minimizing the performance function

$$(103) \quad \varphi(t^i, x^i) e^{\lambda_{ii}[t-t^i]} + \sum_{j \neq i} \int_{t^i}^t \lambda_{ij} e^{\lambda_{ii}(s-t^i)} \psi^j(s, x^i(s; t, x)) ds$$

subject to $x^i(s; t, x)$ being a solution of

$$(104) \quad \dot{x} = f^i(t, x, u)$$

with initial condition $x^i(t; t, x) = x$ and terminal condition

$$(105) \quad (t^i, x^i(t^i; t, x)) = (t^i, x^i) \in M.$$

Here the class of control laws is the set of piecewise continuous, piecewise continuously differentiable functions $u(t)$ such that $u(t) \in U$.

Notice that $u(s) = u^i(s, x^i(s; t, x))$ must be an optimal control for this deterministic problem. Suppose it was not and there were some $u^*(s)$, piecewise continuous and piecewise continuously differentiable, which gave (103) a smaller value.

Let \bar{u} be an arbitrary control value in U . Define a control $u^{ki}(s, x)$, $0 \leq k \leq 1$, by letting

$$(106) \quad \begin{aligned} u^{0i}(s, x) &= \begin{cases} u^*(s) & \text{if } t \leq s \leq t^i(t, x), \\ \bar{u} & \text{if } s < t \text{ or } t^i(t, x) < s, \end{cases} \\ u^{1j}(s, x) &= u^j(s, x), \quad j \in I. \end{aligned}$$

Because of (i), (ii) this control is in \mathcal{M} . If $u^*(s)$ gave a smaller value to the deterministic optimization problem, we would have a control of \mathcal{M} which for the particular initial conditions (t, x, i) gave a performance smaller than that of the optimal in the class \mathcal{U} for the stochastic problem, contradicting Theorem 8.

The deterministic optimization problem described in (103)–(105) does not satisfy the usual hypothesis of Pontryagin's principle in that in the performance (103) the functions $\psi^i(t, x)$ are only continuous and piecewise continuously differentiable. Assumption (E) implies that $x^i(s; t, x)$ meets these discontinuities in a finite set of points. The discontinuities are in the integrand of the performance function and this integrand does not involve the control. It can be shown that under these circumstances the usual statement of Pontryagin's principle does hold. Thus we have asserted the necessity half of the following theorem.

THEOREM 12. *Let conditions (A)–(F) of Part 1 and conditions (i), (ii) just described hold. Let $\theta(t, x)$ be an l -dimensional vector-valued function which is the determining function of the cell of the terminal set M at which the solution $x^i(s; t, x)$ of (104) terminates. Let the matrix θ_k have rank l on this cell of M . Necessary and sufficient conditions that a control $u^i(t, x) \in \mathcal{U}$ be optimal for the stochastic optimization problem are:*

For each $i \in I$ and $(t, x) \in G$, there is an absolutely continuous vector function $p(s)$ defined on $t \leq s \leq t^i(t, x)$, a constant γ_0 , and a vector γ . These quantities satisfy $\gamma_0 \geq 0$, $(\gamma_0, p(s)) \neq 0$,

$$(107) \quad p(s) = -p(s)f_x^i(s, x^i(s; t, x), u^i(s, x^i(s; t, x))) - \gamma_0 \sum_{j \neq i} \lambda_{ij} e^{\lambda_{ij}(s-t)} \psi_x^j(s, x^i(s; t, x))$$

except at a finite number of points $[t, t^i]$. The boundary conditions

$$(108) \quad p(t^i) = \gamma_0 \varphi_x(t^i, x^i) + \gamma \theta_x(t^i, x^i)$$

and

$$(109) \quad p(t^i)f^i(t^i, x^i, u^i(t^i, x^i)) = \gamma_0 \psi_t(t^i, x^i) + \gamma \theta_t(t^i, x^i)$$

are satisfied. The Hamiltonian $p(s)f^i(s, x^i(s; t, x), u^i(s, x^i(s; t, x)))$ is continuous as a function of s on $[t, t^i(t, x)]$ and

$$(110) \quad \min_{u \in U} p^i(s)f^i(s, x^i(s; t, x), u) = p^i(s)f^i(s, x^i(s; t, x), u^i(s, x^i(s; t, x)))$$

holds except at a finite number of points of $[t, t^i(t, x)]$.

Remark. Notice the differences between the $p(s)$ of this theorem and the $p^i(t, x)$ of Theorem 11. For each i and (t, x) , $p(s)$ is a solution of (107), (108). There may be several trajectories $x^i(s; t, x)$ starting out at different initial points (t, x) which follow along the same geometric curve for a portion of their trajectory

near the same terminal point (t^i, x^i) . In order to have the functions $p(s)$ for each of these trajectories be continuous and (110) to hold, the $p(s)$ for each of the different trajectories must be different along the curve which the trajectories have in common. Thus these $p(s)$ cannot be expressed as a function of the coordinates of the curve, as the functions $p^i(s, x)$ of Theorem 11 are.

To prove the sufficiency of Theorem 12, the following lemma is needed.

LEMMA 2. Let (t, x) be a point of $G - \mathcal{S}_n^i$. If on $t_{k-1}(t, x) < s < t_k(t, x)$, $p(s)$ is a vector function such that

$$(111) \quad \min_{u \in U} p(s) f^i(s, x^i(s; t, x), u) = p(s) f^i(s, x^i(s; t, x), u_{c_k}^i(s, x^i(s; t, x)))$$

and $\delta x^i(s; t, x)$ is as in (90), (91), then

$$(112) \quad p(s) f_u u_{c_k x} \delta x^i(s; t, x) = 0.$$

Proof. Let (v, y) denote a point in a neighborhood N of (t, x) contained in $G - \mathcal{S}_n^i$. The expression

$$p(s) f^i(s, x^i(s; t, x), u_{c_k}^i(s, x^i(s; v, y)))$$

is a differentiable function of v, y , and from (111) it attains its minimum at (t, x) . Hence its differential with respect to y which is given by (112) must be zero at (t, x) .

Proof of Theorem 12. Sufficiency. Let $p(s)$ be a solution of (107) on $[t, t^i(t, x)]$ with terminal condition (108), where $(\gamma_0, \gamma) \neq 0$. Let $\delta x^i(s; t, x)$ be a solution of (45) on $t_{k-1}(t, x) < s < t_k(t, x)$ with conditions (11), (12) holding at $t_k(t, x)$, $k = 1, \dots, q-1$, and (12) holding at t_q . Multiplying equation (107) by $\delta x^i(s; t, x)$ and multiplying (45) by $p(s)$, then adding and subtracting the term (112) of Lemma 2 gives

$$(113) \quad \frac{d}{ds} [p(s) \delta x^i(s; t, x)] = -\gamma_0 \sum_{j \neq i} \lambda_{ij} e^{\lambda_{ii}(s-t)} \psi_x^j(s, x^i(s; t, x), \delta x^i(s; t, x))$$

on $t_{k-1}(t, x) \leq t_k(t, x)$. Hence

$$(114) \quad \begin{aligned} & p(t_{k-1}) \delta x^i(t_{k-1}; t, x) - p(t_k) \delta x^i(t_k; t, x) \\ &= \int_{t_{k-1}}^{t_k} \gamma_0 \sum_{j \neq i} \lambda_{ij} e^{\lambda_{ii}(s-t)} \psi_x^j(s, x^i(s; t, x)) \delta x^i(s; t, x) ds. \end{aligned}$$

Adding (114) from $k = 1$ to q gives

$$(115) \quad \begin{aligned} & p(t) + \sum_{k=1}^{q-1} p(t_k) [\delta x^i(t_k; t, x)^- - \delta x^i(t_k; t, x)^+] \\ &= p(t_q) \delta x^i(t_q; t, x) + \gamma_0 \int_t^{t_q} \sum_{j \neq i} \lambda_{ij} e^{\lambda_{ii}(s-t)} \psi_x^j(s, x^i(s; t, x)) \delta x^i(s; t, x) ds. \end{aligned}$$

From (12), (107) and (108),

$$(116) \quad p(t_q) \delta x^i(t_q; t, x) = p(t^i) \delta x^i(t^i; t, x) = p(t^i) \left[\frac{\partial x^i}{\partial x} - f(t^i, x^i, s^i(t^i, x^i)) \frac{\partial t^i}{\partial x} \right]$$

(cont'd)

$$\begin{aligned}
 (116 \text{ cont'd}) \quad &= \gamma_0 \varphi_x(t^i, x^i) \frac{\partial x^i}{\partial x} + \gamma \theta_x(t^i, x^i) \frac{\partial x^i}{\partial x} \\
 &\quad + \gamma_0 \varphi_t(t^i, x^i) \frac{\partial x^i}{\partial t} + \gamma \theta_t(t^i, x^i) \frac{\partial x^i}{\partial t} \\
 &= \gamma_0 \left[\varphi_x(t^i, x^i) \frac{\partial x^i}{\partial x} + \varphi_t(t^i, x^i) \frac{\partial t^i}{\partial x} \right].
 \end{aligned}$$

The last equality follows because

$$\theta(t^i(t, x), x^i(t, x)) \equiv 0,$$

which implies

$$\gamma \left[\theta_x \frac{\partial x^i}{\partial x} + \theta_t \frac{\partial t^i}{\partial t} \right] = 0.$$

From (11), (12) and the continuity of $p(s)f^i(s, x^i(s; (t, x), u^i(s, x^i(s; t, x))))$,

$$\begin{aligned}
 &p(t_k) [\delta x^i(t_k; t, x)^- - \delta x^i(t_k; t, x)^+] \\
 &= [p(t_k)f(t_k, x_k, u_{c_k}(t_k, x_k)) - f(t_k, x_k, u_{c_{k+1}}(t_k, x_k))] \frac{\partial t_k}{\partial x} = 0.
 \end{aligned}$$

Hence

$$\begin{aligned}
 (117) \quad p(t) &= \gamma_0 \left[\varphi_x(t^i, x^i) \frac{\partial x^i}{\partial x} + \varphi_t(t^i, x^i) \frac{\partial t^i}{\partial x} \right] \\
 &\quad + \gamma_0 \int_t^{t^i(t, x)} \sum_{j \neq i} \lambda_{ij} e^{\lambda_{ij}(s-t)} \psi_x^j(s, x^i(s; t, x)) \delta x^i(s; t, x) ds \\
 &= \gamma_0 \psi_x^i(t, x).
 \end{aligned}$$

If γ_0 were zero, we would have $(\gamma_0, p(t)) = 0$, contradicting Pontryagin's principle; hence $\gamma_0 > 0$. Now (t, x) was an arbitrary point of $G - \mathcal{S}_n^i$, hence (111) and (117) imply (81) holds on $G - \mathcal{S}_n^i$. Hence by Theorem 7, $u^i(t, x)$ is optimal.

Notice since $\gamma_0 > 0$ and the conditions of Theorem 12 are unchanged if $P(t)$ is multiplied by a positive constant that γ_0 may always be taken as one in Theorem 12.

Concluding remarks. Some comments are in order on the usefulness of the conditions obtained to compute optimal control laws. Consider the case in which the transition probabilities of the finite state Markov process are such that the state space of the process can be divided into an ordered set of disjoint classes and transitions are possible only from a state in a higher class to one in a lower class. Then the states of the lowest class are all absorbing. For these states equations (107)–(110) reduce to the statement of the ordinary Pontryagin principle because $\lambda_{ij} = 0, j \neq i$. If a synthesis of the control for these states can be obtained through use of Pontryagin's principle and the performance starting in each of them computed as a function of the initial conditions, these performances could

be used in equations (107)–(110) to attempt to obtain a synthesis of the control's corresponding states in the second class. Proceeding by induction in this manner, a control computation somewhat more complicated than the control computation using Pontryagin's principle in the deterministic case is obtained.

Use of Theorem 11 to compute optimal controls is complicated by two things. The minimum principle (93) only determines the control on $G - \mathcal{S}_n^i$ and the equations (98) involve the partial derivatives of some extension $u_{ck}^i(t, x)$ of the control on the cell C_k . It is an interesting question to ask if the necessity portion of Theorem 11 could be strengthened to the statement that for some solution of (95)–(98), the minimum principle (93) must hold everywhere on G . If this were the case, an argument using Lemma 2 in the proof of Theorem 11 could be given to show the terms involving $u_{ckx}^i(t, x)$ in (98) may be dropped. Equation (98) might be solved together with (93) to provide a synthesis or it might be used as the basis of a successive approximation technique. Equation (98) does have the advantage that it involves only the partial derivatives of the optimal performance function and is independent of the optimal performance function itself.

Early work on using dynamic programming conditions to compute optimal controls is reported in [8], [9]. It appears that many of the difficulties in using dynamic programming to compute optimal controls in the deterministic case again appear in this stochastic case.

Use of Theorem 9 to compute optimal controls has been investigated very little. However, of all the conditions it does appear that this one involves the control in the least explicit manner.

As mentioned in the Introduction, the requirement that (8) be bounded away from zero on each cell $\pi(C)$ corresponding to each type (i) cell C is restrictive. If this is relaxed to merely requiring (8) positive, the arguments of § 3.4 of Part 1 will go through if the operator H maps the function $y^i(t, x)$ defined in (3.7) into the space D and if the formula (3.5) for differentiating under the integral sign holds for the function $y^i(t, x)$. The convergence of the series (38) would be uniform on compact subsets of the cells of $G - \mathcal{S}_n^i$. This would enable most of the development of this paper to be carried through under these assumptions. Since at present we don't know conditions expressed in terms of the controls which would imply these conditions, we have preferred to carry out the development under the restrictive assumption rather than impose conditions of the type mentioned.

The adjoint equations of optimality conditions given by Kushner [7] or Sworder [15] involve stochastic adjoint equations. The hypotheses in Kushner's work [7] are different from those of this paper. However, one can ask, if these conditions are formally applied to the present problem, what is the relationship between the stochastic adjoint equations of [7], [15] and the corresponding deterministic equations of this paper? This question is still open.

Appendix A. Proof of Theorem 1. Let C be a cell of $G - \mathcal{S}_n^i$. Let $C = C_1, \dots, C_q$ denote the succession of type (i) cells through which a trajectory $x(s; t, x)$ beginning at $(t, x) \in C$ passes to reach M . Let $(t_j(t, x), x_j(t, x))$ denote the time and position at which $x(s; t, x)$ leaves C_j .

As a convention to begin the induction, define $t_0(t, x)$ and $x_0(t, x)$ on C by

$$t_0(t, x) = t, \quad x_0(t, x) = x.$$

Then as an induction hypothesis, suppose for an integer $j - 1$ it has been shown that the conclusion of (A) is valid provided that $t \leq s \leq t_{j-1}(t, x)$, and that condition (B) and (C) are valid for integers $\leq j - 1$.

With this hypothesis up to $j - 1$ it is desired to show that the induction hypothesis is satisfied up to j .

To begin this argument recall the definition of $t_j(t, x)$, $x_j(t, x)$ and note that $x(s; t, x)$ is a solution of

$$(A.1) \quad \dot{x} = f(s, x)$$

defined on $t_{j-1}(t, x) \leq s \leq t_j(t, x)$ passing through $x_{j-1}(t, x)$ at time $t_{j-1}(t, x)$ and $(s, x(s; t, x))$ is contained in C_j for $t_{j-1}(t, x) < s < t_j(t, x)$. Now $f_{c_j}(s, x)$ is a continuously differentiable function defined on a neighborhood of the closure of C_j which agrees with $f(s, x)$ on C_j . For any fixed (t_0, x_0) of C , theorems on continuous dependence of solutions on initial conditions and theorems on continuation of solutions imply that for $(w; z)$ in a neighborhood A_{j-1} of $(t_{j-1}(t_0, x_0), x_{j-1}(t_0, x_0))$ and some $\varepsilon_j > 0$ there is a solution $\beta_j(s; w, z)$ of

$$(A.2) \quad \dot{\beta} = f_{c_j}(s, \beta)$$

defined on $t_{j-1}(t_0, x_0) - \varepsilon_j \leq s \leq t_j(t_0, x_0) + \varepsilon_j$ such that $\beta_j(w; w, z) = z$. (For consistency we always assume $|w - t_{j-1}(t_0, x_0)| < \varepsilon_j$ if $(w, z) \in A_{j-1}$.)

Theorems on the differentiability of solutions of differential equations with respect to initial conditions imply that A_j may be selected so that the mapping

$$(s, w, z) \rightarrow \beta_j(s; w, z)$$

is continuously differentiable on

$$[t_{j-1}(t_0, x_0) - \varepsilon_j \leq s \leq t_j(t_0, x_0) + \varepsilon_j] \times A_j.$$

If $\delta\beta_j(s)$ denotes the matrix of partial derivatives with respect to z of $\beta_j(s; w, z)$, these theorems imply that the differential equation

$$(A.3) \quad \delta\dot{\beta}_j(s) = f_{c_j x}(s, \beta_j(s; w, z)) \delta\beta_j(s)$$

and initial condition $\delta\beta_j(w) = I$, where I is the identity matrix, are satisfied by $\delta\beta_j(s)$.

If $\delta\beta_j(s)$ denotes the vector of partial derivatives of $\beta_j(s; w, z)$ with respect to w , the vector equation denoted by the same formula as (A.3) holds for $\delta\beta_j(s)$. The initial condition $\delta\beta_j(w) = -f_{c_j}(w, z)$ holds in this case. Since solutions of (A.2) are unique and $f_{c_j}(s, x) = f(s, x)$ for (s, x) in C_j , we must have from assumption (B) that

$$(A.4) \quad \beta_j(s, t_{j-1}(t, x), x_{j-1}(t, x)) = x(s; t, x)$$

for $\{(s, t, x) : (t, x) \in (t_{j-1}, x_{j-1})^{-1}[A_j], t_{j-1}(t, x) \leq s \leq t_j(t, x)\}$.

Let $\theta_j(t, x)$ denote the extra component that the determining function $\theta(t, x)$ of $\pi(C_j)$ has in addition to the components of the determining function of C_j . Considered as an equation in s ,

$$(A.5) \quad \theta_j[s, \beta_j(s, t_{j-1}(t, x), x_{j-1}(t, x))] = 0$$

has a solution $t_j(t, x)$ defined on $(t_{j-1}, x_{j-1})^{-1}A_j$. This follows because for (t, x) belonging to this set,

$$\begin{aligned}\theta_j[t_j(t, x), \beta_j(t_j(t, x); t_{j-1}(t, x), x_{j-1}(t, x))] &= \theta_j[t_j(t, x), x(t_j(t, x); t, x)] \\ &= \theta_j(t_j(t, x), x_j(t, x)) = 0.\end{aligned}$$

The function $\theta_j[s, \beta_j(s; t_{j-1}(t, x), x_{j-1}(t, x))]$ is continuously differentiable on

$$\{(s, t, x) : (t, x) \in (t_{j-1}, x_{j-1})^{-1}A_j, t_{j-1}(t_0, x_0) - t_j < s < t_{j-1}(t_0, x_0) + \varepsilon\}$$

and (8) holds at $(t_j(t, x), x_j(t, x))$. Thus the implicit function theorem implies that $t_j(t, x)$ is continuously differentiable on $(t_{j-1}, x_{j-1})^{-1}[A_j]$. Since (t_0, x_0) was an arbitrary point of C and $(t_{j-1}, x_{j-1})^{-1}[A_j]$ is a neighborhood of t_0, x_0 , we conclude that $t_j(t, x)$ must be continuously differentiable on C .

Since $t_j(t, x)$ is continuously differentiable on C and $x_j(t, x) = x(t_j(t, x), t, x) = \beta_j(t_j(t, x); t_{j-1}(t, x), x_j(t, x))$, it follows from the differentiability of $\beta_j(s; w, z)$ and of $t_{j-1}(t, x)$ and $x_{j-1}(t, x)$ that $x_j(t, x)$ is continuously differentiable on C .

For $t_{j-1}(t, x) < s < t_j(t, x)$,

$$(A.6) \quad x(s; t, x) = \beta_j(s, t_{j-1}(t, x), x_{j-1}(t, x)).$$

The right-hand side of (A.6) is differentiable in (t, x) and has matrices and vectors of partial derivatives with respect to x and t represented by

$$(A.7) \quad \delta x(s; t, x) = \beta_{jw}(s) \frac{\partial t_{j-1}}{\partial x} + \beta_{jz}(s) \frac{\partial x_{j-1}}{\partial x}$$

and

$$(A.8) \quad \delta x(s; t, x) = \beta_{jw}(s) \frac{\partial t_{j-1}}{\partial t} + \beta_{jz}(s) \frac{\partial x_{j-1}}{\partial t}$$

on the interval $t_{j-1}(t, x) < s < t_j(t, x)$. For a compact subset K of G ,

$$\{(s, t, x) : (t, x) \in K \cap C, t_{j-1}(t, x) \leq s \leq t_j(t, x)\}$$

is compact. Hence bounds on $\delta x(s; t, x)$ follow from the continuous differentiability of β_j and bounds on the partial derivatives of $t_{j-1}(t, x)$ and $x_{j-1}(t, x)$. Since β_{jw} and β_{jz} satisfy the differential equations (10), it follows that the differential equations (10) hold for the partial derivatives $\delta x(s; t, x)$ of $x(s; t, x)$. Since β_j is continuously differentiable in (w, z) , for $t_{j-1}(t_0, x_0) - \varepsilon_j < s < t_{j-1}(t_0, x_0) + \varepsilon_j$, (A.7) and (A.8) imply that

$$\begin{aligned}\lim_{s \downarrow t_j(t, x)} \delta x(s; t, x) &= \beta_{jw}(t_{j-1}, t_{j-1}, x_{j-1}) t_{j-1, x} + \beta_{jz}(t_{j-1}; t_{j-1}, x_{j-1}) x_{j-1, x}, \\ \lim_{s \uparrow t_j(t, x)} \delta x(s; t, x) &= \beta_{jw}(t_j, t_{j-1}, x_{j-1}) t_{j-1, x} + \beta_{jz}(t_j; t_{j-1}, x_{j-1}) x_{j-1, x}.\end{aligned}$$

The equation

$$x_{j-1}(t, x) = \beta_j(t_{j-1}(t, x), t_{j-1}(t, x), x_{j-1}(t, x))$$

implies that

$$(A.9) \quad \begin{aligned} x_{j-1,x} &= f_{c_j}(t_{j-1}, x_{j-1})t_{j-1,x} + \beta_{jw}t_{j-1,x} + \beta_{jz}x_{j-1,x} \\ &= f_{c_j}(t_{j-1}, x_{j-1})t_{j-1,x} + \delta x(t_{j-1})^+. \end{aligned}$$

Similarly, the equation

$$x_j(t, x) = \beta_j(t_j(t, x), t_{j-1}(t, x), x_{j-1}(t, x))$$

implies that

$$(A.10) \quad \begin{aligned} x_{jx} &= f_{c_j}(t_j, x_j)t_{jx} + \beta_{jw}t_{j-1,x} + \beta_{jz}x_{j-1,x}, \\ \frac{\partial x_j}{\partial x} &= f_{c_j}(t_j, x_j)\frac{\partial t_j}{\partial x} + \frac{\partial \beta_j}{\partial w}\frac{\partial t_{j-1}}{\partial x} + \frac{\partial \beta_j}{\partial z}\frac{\partial x_{j-1}}{\partial w} \\ &= f_{c_j}(t_j, x_j)t_{jx} + \delta x(t_j)^-. \end{aligned}$$

Similar relationships hold for the partial derivatives of x_{j-1} and x_j with respect to t .

A bound for the partial derivatives of $t_j(t, x)$ on the intersection of C with any compact set can be obtained from relationships such as

$$(A.11) \quad t_{jx}(t, x) = \frac{\theta_{jx}[\beta_{jw}t_{j-1,x} + \beta_{jz}x_{j-1,x}]}{\theta_{jt} + \theta_{jx}f_{c_j}},$$

using the bounds on $t_{j-1,x}$, $x_{j-1,x}$ and the relationship (8).

The induction from $j-1$ to j has now been completed and we may assert by induction that the statement must hold for $j = q$. This is the conclusion of the theorem.

Appendix B. Proof of Theorem 10. In the proof of Theorem 10 the control $u^i(t, x)$ will be fixed throughout. Since this is the case, to save space we shall revert to the notation of Part 1 replacing

$$f^i(t, x, u^i(t, x)) \quad \text{by} \quad f^i(t, x)$$

and

$$f_x^i(t, x, u_{c_k}^i(t, x)) + f_u^i(t, x, u_{c_k}^i(t, x))u_{c_kx}^i(t, x) \quad \text{by} \quad f_{c_kx}^i(t, x).$$

We shall first show that if $p^i(t, x)$ is a solution of the system (95)–(98), then $p^i(t, x)$ is a solution of the system (89)–(92). This will follow if we establish (92). Since $p^i(s, x^i(s; t, x))$ is absolutely continuous on each interval $t_{k-1} \leq s \leq t_k$, differentiating this expression gives that on these intervals:

$$(B.1) \quad \begin{aligned} &\frac{d}{ds}p^i(s, x^i(s; t, x)) \\ &= -p^i(s, x^i(s; t, x))f_{c_kx}^i(s, x^i(s; t, x)) - \sum_{j=1}^k \lambda_{ij}p^j(s, x^i(s; t, x)). \end{aligned}$$

Now from (10),

$$(B.2) \quad \frac{d}{ds} \delta x^i(s; t, x) = f_{c_k x}(s, x^i(s; t, x)) \delta x^i(s; t, x)$$

on these intervals. Multiplying (B.1) on the right by $\delta x^i(s; t, x)$ and (B.2) on the left by $p^i(s, x^i(s; t, x))$, then adding the two equations and multiplying the result by $e^{\lambda_{ii}(s-t)}$ gives

$$(B.3) \quad \frac{d}{ds} e^{\lambda_{ii}(s-t)} p^i(s, x^i(s; t, x)) \delta x^i(s; t, x) = - \sum_{j \neq i} \lambda_{ij} e^{\lambda_{ii}(s-t)} p^j(s, x^i(s; t, x)) \delta x^i(s; t, x).$$

Integrating this expression from $t_{k-1}(t, x)$ to $t_k(t, x)$ and summing these integrals gives

$$(B.4) \quad \begin{aligned} p^i(t, x) &= \sum_{k=1}^{q-1} e^{\lambda_{ii}(t_k-t)} [p^i[t_k, x_k]^- \delta x^i(t_k; t, x)^- - p^i[t_k, x_k]^+ \delta x^i(t_k; t, x)^+] \\ &+ e^{\lambda_{ii}(t_q-t)} p^i[t_q, x_q]^- \delta x^i(t_q; t, x)^- \\ &+ \int_t^{t^i(t, x)} \sum_{j \neq i} \lambda_{ij} e^{\lambda_{ii}(s-t)} p^j[s, x^i(s; t, x)] \delta x^i(s; t, x) ds. \end{aligned}$$

Equations (98), (96) and (12) imply that

$$(B.5) \quad \begin{aligned} p^i[t_q, x_q]^- \delta x^i(t_q; t, x)^- &= [\varphi_x(t_q, x_q) + \mu_q(t_q, x_q) \theta_{qx}(t_q, x_q)] \delta x^i(t_q; t, x)^- \\ &= [\varphi_x(t_q, x_q) + \mu_q(t_q, x_q) \theta_{qx}(t_q, x_q)] \left[\frac{\partial x_q}{\partial x} - f_{c_q}(t_q, x_q) \frac{\partial t_q}{\partial x} \right] \\ &= \varphi_x(t_q, x_q) \frac{\partial x_q}{\partial x} + \varphi_t(t_q, x_q) \frac{\partial t_q}{\partial x} + \mu_q(t_q, x_q) \left[\theta_{qx}(t_q, x_q) \frac{\partial x_q}{\partial x} + \theta_{qt}(t_q, x_q) \frac{\partial t_q}{\partial x} \right] \\ &= \varphi_x(t^i(t, x), x^i(t, x)) \frac{\partial x^i(t, x)}{\partial x} + \varphi_t(t^i(t, x), x^i(t, x)) \frac{\partial t^i(t, x)}{\partial t}. \end{aligned}$$

The last equality in (B.5) holds because

$$(B.6) \quad \theta_q(t_q(t, x), x_q(t, x)) \equiv 0$$

and hence

$$(B.7) \quad \theta_{qx} \frac{\partial x_q}{\partial x} + \theta_{qt} \frac{\partial t_q}{\partial x} = 0.$$

Equations (98), (11), (12) and (97) imply that

$$\begin{aligned} &p^i(t_k, x_k)^+ \delta x^i(t_k; t, x)^+ - p^i(t_k, x_k)^- \delta x^i(t_k; t, x)^- \\ &= p^i(t_k, x_k)^+ [\delta x^i(t_k; t, x)^+ - \delta x^i(t_k; t, x)^-] \\ &\quad + [p^i(t_k, x_k)^+ - p^i(t_k, x_k)^-] \delta x^i(t_k; t, x)^- \\ &= \mu_k(t_k, x_k) [\theta_{kx}(t_k, x_k) f_{c_k}^i(t_k, x_k) + \theta_{kt}(t_k, x_k)] \frac{\partial t_k}{\partial x} \\ &\quad + \mu_k(t_k, x_k) \theta_{kx}(t_k, x_k) \left[-f_{c_k}^i(t_k, x_k) \frac{\partial t_k}{\partial x} + \frac{\partial x_k}{\partial x} \right] \end{aligned}$$

(cont'd)

$$= \mu_k(t_k, x_k) \left[\theta_{kt}(t_k, x_k) \frac{\partial t_k}{\partial x} + \theta_{kx}(t_k, x_k) \frac{\partial x_k}{\partial x} \right] = 0.$$

The last equality holds because

$$(B.8) \quad \theta_k(t_k(t, x), x_k(t, x)) \equiv 0.$$

Hence we have, since $(t^i(t, x), x^i(t, x)) \equiv (t_q(t, x), x_q(t, x))$, that

$$(B.9) \quad \begin{aligned} p^i(t, x) &= \varphi_x(t^i(t, x), x^i(t, x)) \frac{\partial x^i(t, x)}{\partial x} + \varphi_t(t^i(t, x), x^i(t, x)) \frac{\partial t^i(t, x)}{\partial x} \\ &+ \int_t^{t^i(t, x)} \sum_{j \neq i} \lambda_{ij} e^{\lambda_{ij}(s-t)} p^j[s, x^i(s; t, x)] \delta x^i(s; t, x) ds. \end{aligned}$$

Thus $p^i(t, x)$ is a solution of the system (95)–(98) of equations.

Next let us show that (95)–(98) have a solution. Let $h^i(s; t, x)$ denote the solution of the differential equation

$$(B.10) \quad h^i(\dot{s}) = -h^i(s) f_{c_k x}(s, x^i(s; t, x)) - \lambda_{ii} h^i(s) + \sum_{i \neq j} \lambda_{ij} p_x^j(s, x^i(s; t, x))$$

on the succession of intervals $t_{k-1}(t, x) < s < t_k(t, x)$. The solution $h^i(s; t, x)$ satisfies the terminal condition at t_q :

$$(B.11) \quad h^i(t_q; t, x) = \mu_q(t_q, x_q) \theta_{qx}(t_q, x_q) + \varphi_x(t_q, x_q),$$

where $\mu_q(t_q, x_q)$ is a solution of (96). It also satisfies the jump conditions at t_k :

$$(B.12) \quad h^i(t_k^+, t, x) - h^i(t_k^-, t, x) = \mu_k(t_k, x_k) \theta_{kx}(t_k, x_k),$$

where

$$(B.13) \quad \begin{aligned} h^i(t_k^+; t, x) &[f_{c_k}^i(t_k, x_k) - f_{c_{k+1}}^i(t_k, x_k)] \\ &= \mu_k(t_k, x_k) [\theta_{kx}(t_k, x_k) f_{c_k}^i(t_k, x_k) + \theta_{kt}(t_k, x_k)], \end{aligned}$$

where $\mu(t_k, x_k)$ is a solution of (97). Formula (8) of assumption (B) implies that (96) and (97) have unique solutions for $\mu_q(t_q, x_q)$ and $\mu_k(t_k, x_k)$. Equations (B.10) are a system of linear differential equations with bounded piecewise continuous coefficients on the interval $t \leq s \leq t_q(t, x)$; hence there is a solution of (B.10)–(B.13).

Notice since $x^i(v; t, x) = x^i(v; s, x^i(s; t, x))$ if $v \geq s$, that the coefficients of equations (B.10)–(B.13) defining $h^i(v; t, x)$ and $h^i(v; s, x^i(s; t, x))$ agree. Hence we must have

$$(B.14) \quad h^i(s; t, x) = h^i(s; s, x^i(s; t, x)).$$

Let

$$(B.15) \quad H^i(s, x) = h^i(s; s, x).$$

Then using (B.14) and (B.15) we see that

$$(B.16) \quad \frac{d}{ds} H^i(s, x^i(s; t, x)) = -H^i(s, x^i(s; t, x)) f_{c_k x}(s, x^i(s; t, x)) \\ - \lambda_{ii} H^i(s, x^i(s; t, x)) - \sum_{i \neq j} \lambda_{ij} \psi_x^j(s, x^i(s; t, x))$$

on the interval $t_{k-1}(t, x) < s < t_k(t, x)$. Multiplying (B.16) on the right by $\delta x^i(s; t, x)$ and (90) on the left by $H^i(s, x^i(s; t, x))$, adding these two equations and then multiplying the result by $e^{\lambda_{ii}(s-t)}$ gives

$$(B.17) \quad \frac{d}{ds} e^{\lambda_{ii}(s-t)} H^i(s, x^i(s; t, x)) \delta x^i(s; t, x) \\ = - \sum_{j \neq i} \lambda_{ij} e^{\lambda_{ii}(s-t)} \psi_x^j(s, x^i(s; t, x)) \delta x^i(s; t, x).$$

An argument identical to that of equations (B.4)–(B.9) now shows that

$$(B.18) \quad H^i(t, x) = \psi_x^i(t, x) \quad \text{on } G - \mathcal{S}_n^i.$$

Since $x^i(s; t, x)$ meets $G - \mathcal{S}_n^i$ in only a finite number of points, replacing $\psi_x^j(s, x^i(s; t, x))$ by $H^j(s, x^i(s; t, x))$ in (B.16), integrating (B.16) from $t_{k-1}(t, x)$ to $t_k(t, x)$ and setting

$$p(t_q, x_k) = \varphi_x(t_q, x_q) + \eta_q(t_q, x_q) \theta_{qx}(t_q, x_q),$$

and for $k = 1, \dots, q-1$,

$$p(t_k, x_k)^+ - p(t_k, x_k)^- = \eta_k(t_k, x_k) \theta_{kx}(t_k, x_k),$$

gives that $H^i(s, x)$ is a solution of the system (95)–(98).

REFERENCES

- [1] L. D. BERKOVITZ, *Variational methods in control and programming*, J. Math. Anal. Appl., 3 (1961), pp. 145–161.
- [2] ———, *A Hamilton–Jacobi theory for a class of control problems*, Colloque sur la Théorie Mathématique du Contrôle Optimal (Bruxelles, 1969), Centre Belge de Recherches Mathématiques Vaneler, Lauvain, 1970.
- [3] V. G. BOLTJANSKII, *Sufficient conditions for optimality and the justification of the dynamic programming method*, this Journal, 4 (1966), pp. 326–361.
- [4] I. I. GIKHMAN AND A. V. SKOROHOD, *Introduction to the Theory of Random Process*, W. B. Saunders, Philadelphia, 1969.
- [5] R. GRIEGO AND R. HERSH, *Theory of random evolutions with applications to partial differential equations*, Trans. Amer. Math. Soc., 156 (1971), pp. 405–418.
- [6] D. C. HEATH, *Probabilistic analysis of hyperbolic systems of partial differential equations*, Thesis, University of Illinois, Urbana, 1969.
- [7] H. J. KUSHNER, *On the stochastic maximum principle: Fixed time of control*, J. Math. Anal. Appl., 11 (1965), pp. 78–92.
- [8] N. N. KRASSOVSKII AND E. A. LIDSKII, *Analytical design of controllers in stochastic systems with velocity limited controlling action*, Appl. Math. Mech., 25 (1961), pp. 627–643.
- [9] E. A. LIDSKII, *Optimal control of systems with random properties*, Ibid., 27 (1963), pp. 33–45.
- [10] S. MIRICA, *On the admissible synthesis in optimal control theory and differential games*, this Journal, 7 (1969), pp. 292–316.
- [11] M. PINSKY, *Multiplicative operator functionals and their asymptotic properties*, to appear.

- [12] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISCHENKO, *The Mathematical Theory of Optimal Processes*, John Wiley, New York, 1962.
- [13] R. W. RISHEL, *Optimality of controls for systems with jump Markov disturbances*, Optimization Techniques. A. V. Balakrishnan, ed., Academic Press, New York, 1972.
- [14] P. ROBINSON, J. Franklin Institute, to appear.
- [15] D. D. SWORDER, *Feedback control of a class of linear systems with jump parameters*, IEEE Trans. Automatic Control, AC-14 (1969), pp. 9–14.
- [16] V. G. BOLTYANSKII, *Mathematical Methods of Optimal Control*, Holt, Rinehart and Winston, New York, 1971.
- [17] W. M. WONHAM, *Random differential equations in control theory*, Probabilistic Methods in Applied Mathematics, A. T. Bharucha-Reid, ed., Academic Press, New York, 1970, pp. 191–199.

SECOND ORDER NECESSARY CONDITIONS FOR PROBLEMS WITH STATE INEQUALITY CONSTRAINTS*

I. BERT RUSSAK†

Abstract. A set of second order necessary conditions is obtained for solution to a basic problem in optimal control involving the state constraints

$$\psi^\alpha(t, x) \leq 0, \quad \alpha = 1, \dots, m.$$

For the case of a normal solution arc, the methods of Hestenes are used to show that the second variation of a certain functional is nonnegative on a class of variations of the solution arc. The results for this problem are basic in the sense that they can be extended to include analogous results for much more complicated problems.

1. Introduction. A general problem in optimal control can be stated as follows:

Let C be the class of arcs

$$\begin{aligned} a: \quad & x^i(t), \quad u^k(t), \quad b^\sigma, \quad t^0 \leq t \leq t^1, \\ & i = 1, \dots, N, \quad k = 1, \dots, K, \quad \sigma = 1, \dots, r, \end{aligned}$$

whose elements $(t, x(t), u(t), b)$ lie in a region R in $t \times ub$ space and which in addition have $u(t)$ piecewise continuous and satisfy the constraints

$$\begin{aligned} \dot{x}^i &= f^i(t, x, u, b), \quad i = 1, \dots, N, \\ \psi^\alpha(t, x, b) &\leq 0, \quad \alpha = 1, \dots, m', \quad \psi^\alpha(t, x, b) = 0, \quad \alpha = m' + 1, \dots, m, \\ \theta^\eta(t, x, u, b) &\leq 0, \quad \eta = 1, \dots, L', \quad \theta^\eta(t, x, u, b) = 0, \quad \eta = L' + 1, \dots, L, \\ I_\gamma(a) &\leq 0, \quad \gamma = 1, \dots, p', \quad I_\gamma(a) = 0, \quad \gamma = p' + 1, \dots, p, \\ x^i(t^s) &= X^{is}(b), \quad i = 1, \dots, N, \quad s = 0, 1, \end{aligned}$$

where

$$I_\gamma(a) = g_\gamma(b) + \int_{t^0}^{t^1} L_\gamma(t, x(t), u(t), b) dt, \quad \gamma = 1, \dots, p.$$

It is desired to minimize the integral

$$I_0(a) = g_0(b) + \int_{t^0}^{t^1} L_0(t, x(t), u(t), b) dt$$

on the class C .

With the dimensions of our terms allowed to assume new values, the above problem is expressible as one in which the constraints have the simpler basic form shown below and also an assumption made on the rank of the matrix $(\psi_x^\alpha f_u)$ in the general problem is transformed into a stronger condition in the basic problem.

* Received by the editors July 5, 1973, and in revised form January 18, 1974. This work was supported by an NPS Foundation Grant.

† Naval Postgraduate School, Monterey, California 93940.

The basic constraints are:

$$\begin{aligned}\dot{x}^i &= f^i(t, x, u), & i &= 1, \dots, N, \\ \psi^\alpha(t, x) &\leq 0, & \alpha &= 1, \dots, m, \\ I_\gamma(a) &\leq 0, & \gamma &= 1, \dots, p', & I_\gamma(a) &= 0, & \gamma &= p' + 1, \dots, p, \\ x^i(t^s) &= X^{is}(b), & s &= 0, 1, & i &= 1, \dots, N.\end{aligned}$$

The methods of Hestenes are used to show that when a solution arc to the basic problem is normal, then the second variation of a certain functional is nonnegative on a class of variations of that solution arc. Extension of this result to the general problem is the subject of a succeeding paper.

2. A basic problem. In what follows we shall assume familiarity with [2] and [3] and shall use quantities defined there. In addition, unless otherwise specified, the conventions and notations of those papers will also apply here.

A slight generalization of the problem of §4 of [2] appears in the following statement.

Let C be the class of arcs

$$\begin{aligned}a: \quad & x^i(t), \quad u^k(t), \quad b^\sigma, \quad t^0 \leq t \leq t^1, \\ & i = 1, \dots, N, \quad k = 1, \dots, K, \quad \sigma = 1, \dots, r,\end{aligned}$$

whose¹ elements $(t, x(t), u(t))$ and b lie respectively in open sets R in $t \times u$ space and b in B space and which in addition have $u(t)$ piecewise continuous. The terms x^i are called state variables. The terms u^k, b^σ are called control variables and control parameters respectively. We require these arcs to satisfy the constraints

$$\begin{aligned}(1a) \quad & \dot{x}^i = f^i(t, x, u), & i &= 1, \dots, N, \\ (1b) \quad & \psi^\alpha(t, x) \leq 0, & \alpha &= 1, \dots, m, \\ (1c) \quad & I_\gamma(a) \leq 0, & \gamma &= 1, \dots, p', & I_\gamma(a) &= 0, & \gamma &= p' + 1, \dots, p, \\ (1d) \quad & x^i(t^s) = X^{is}(b), & s &= 0, 1,\end{aligned}$$

where

$$I_\gamma(a) = g_\gamma(b) + \int_{t^0}^{t^1} L_\gamma(t, x(t), u(t)) dt, \quad \gamma = 1, \dots, p.$$

It is desired to minimize the integral

$$(2) \quad I_0(a) = g_0(b) + \int_{t^0}^{t^1} L_0(t, x(t), u(t)) dt$$

on the class C .

We suppose that the arc

$$a_0: \quad x_0(t), \quad u_0(t), \quad b_0, \quad t^0 \leq t \leq t^1,$$

¹ Unless otherwise noted the indices i, k, σ, α will have the respective ranges $i = 1, \dots, N, k = 1, \dots, K, \sigma = 1, \dots, r, \alpha = 1, \dots, m$.

is a solution to our problem. In addition we set² $\phi^\alpha = \psi_t^\alpha + \psi_{x_i}^\alpha f^i$ and except for the functions $X^{js}(b)$, $j = 1, \dots, N$, $s = 0, 1$, which are assumed to be C^1 on B , we assume all other functions to have the continuity properties introduced in § 4 of [2].³ The set R_0 is defined as the set of points (t, x, u) in R satisfying

$$(3a) \quad \psi^\alpha \leq 0,$$

$$(3b) \quad \phi^\alpha \geq 0 \quad \text{for all } \alpha \text{ with } \psi^\alpha = 0 \quad \text{or} \quad \phi^\alpha \leq 0 \quad \text{for all } \alpha \text{ with } \psi^\alpha = 0.$$

Finally let \bar{D} be the set of points $(t, x_0(t), u)$ in R_0 with $u = u_0(t)$ or for arbitrary u , with t interior to an interval of continuity of $u_0(t)$. Then we assume that the matrix

$$(4) \quad (\phi_{u_k}^\alpha), \quad \alpha = 1, \dots, m,$$

has rank m on \bar{D} .

In stating the above problem, the primary modification we have made to the problem of § 4 of [2] was in allowing the functions ψ^α to have arbitrary form.

An examination of the proof of Theorem 6.1 of [2] will show that with the following minor modifications the statements of that theorem apply to the problem introduced above: the symbols $N + m$, $K + m$, $r + 2m$ are replaced by N , K , r respectively and the function $\mathcal{G}(b)$ has the form

$$(5) \quad \mathcal{G}(b) = \lambda_0 g_0 + \lambda_\gamma g_\gamma + \lambda_{p+i} X^{i0}, \quad \gamma = 1, \dots, p, \quad i = 1, \dots, N.$$

We require two properties of the multipliers $\mu_\alpha(t)$ not proved in Theorem 6.1 of [2]. The first of these properties is established in the following lemma.

LEMMA 2.1. *The multipliers $\mu_\alpha(t)$ are constant on intervals upon which $\psi^\alpha(t) < 0$.*

Proof. The proof of this is a simple modification of a proof in [2]. Fix $\alpha = \bar{\alpha}$, $1 \leq \bar{\alpha} \leq m$, and let $[t', t'']$ be an interval such that $\psi^\alpha(t) < 0$ there. Referring to § 13 of [2], let w be an arc in the class W satisfying

$$w^\Gamma(t) = 0, \quad \Gamma = 1, \dots, K, \quad \Gamma \neq \bar{\alpha}, \quad t^0 \leq t \leq t^1,$$

$$w^{\bar{\alpha}}(t) = 0, \quad t \in [t^0, t'] \cup [t'', t^1].$$

Referring to the argument involving (169) and (170) of [2] we see that those relations hold now also so that corresponding to (171) of [2] we obtain

$$(6) \quad \int_{t'}^{t''} -\mu_{\bar{\alpha}} \dot{w}^{\bar{\alpha}} dt \geq 0,$$

where according to the definition of the class W , the sign of $w^{\bar{\alpha}}$ is arbitrary on $[t', t'']$. Then by the Fundamental lemma in the calculus of variations, the multiplier $\mu_{\bar{\alpha}}(t)$ is constant on $[t', t'']$. While the construction of the proof has been with $[t', t'']$ interior to $[t^0, t^1]$, nevertheless, continuity properties of $\mu_{\bar{\alpha}}(t)$ establish it also for all subintervals $[t', t'']$ of $[t^0, t^1]$, thus proving the lemma.

The second property is that

$$\mu_\alpha(t^1) = 0 \quad \text{if } \psi^\alpha(t^1) < 0, \quad \alpha = 1, \dots, m.$$

² Where for a function $M(t, x, u, b)$, the notations M_{x^i} , M_{u^k} , $M_{b^{\sigma}}$, M_i denote first partial derivatives with respect to the indicated variable. Also unless otherwise specified, repeated indices will be summed.

³ We shall at a later point strengthen these assumptions for proving the second order theorems to follow.

This property is established by an argument directly analogous to that used in [1, p. 372] and will not be repeated here.

3. Normality. Let $q_{\gamma j}(t)$ and $z_{ij}(t)$ ($\gamma = 0, 1, \dots, p$; $i, j = 1, \dots, N$) be the functions of (66) in [2] and as in (70) of [2] define the functionals

$$(7a) \quad J_{\gamma}(a) = q_{\gamma j}(t^0)[x^j(t^0) - X^{j0}(b)] + I_{\gamma}(a), \quad \gamma = 0, 1, \dots, p,$$

$$(7b) \quad J_{p+i}(a) = z_{ij}(t^0)[X^{j0}(b) - x^j(t^0)] + x^i(t^1) - X^{i1}(b),$$

$$(7c) \quad J_{p+N+i}(a) = X^{i0}(b) - x^i(t^0), \quad i, j = 1, \dots, N.$$

Next, denote by δa , the triple of vectors

$$\delta a: \quad \delta x^i(t), \quad \delta u^k(t), \quad \delta b^{\sigma}, \quad t^0 \leq t \leq t^1,$$

(with $\delta x(t)$, $\delta u(t)$ respectively continuous and piecewise continuous functions).

We call δa a variation of the arc a_0 and for any function $M(t, x, u, b)$ which possesses first partial derivatives with respect to x, u, b along a_0 , we call⁴

$$\delta M(t) = M_{x^i}(t)\delta x^i(t) + M_{u^k}(t)\delta u^k(t) + M_{b^{\sigma}}(t)\delta b^{\sigma}$$

the variation of M along a_0 due to the variation δa .

We call a variation δa admissible if it satisfies⁵

$$\delta \dot{x}^i = \delta f^i, \quad i = 1, \dots, N,$$

$\delta \psi^{\alpha}(t) \leq 0$ on a neighborhood of s^{α} , $\alpha = 1, \dots, m$, where $s^{\alpha} \equiv \{t | \psi^{\alpha}(t, x_0(t)) = 0\}$.

Next, define the expressions

$$(8a) \quad J'_{\gamma}(a_0, \delta a) = q_{\gamma j}(t^0)[\delta x^j(t^0) - \delta X^{j0}] + I'_{\gamma}(a_0, \delta a),$$

$$(8b) \quad J'_{p+i}(a_0, \delta a) = z_{ij}(t^0)[\delta X^{j0} - \delta x^j(t^0)] + \delta x^i(t^1) - \delta X^{i1},$$

$$(8c) \quad J'_{p+N+i}(a_0, \delta a) = \delta X^{i0} - \delta x^i(t^0)$$

as variations of the functionals (7) due to the variation δa of a_0 and where

$$I'_{\gamma}(a_0, \delta a) \equiv \delta g_{\gamma} + \int_{t^0}^{t^1} \delta L_{\gamma} dt, \quad \gamma = 0, 1, \dots, p.$$

Now let γ_k be those indices for which

$$(9) \quad I_{\gamma}(a_0) < 0, \quad 1 \leq \gamma \leq p',$$

and let \bar{p} be the number of such indices. According to Theorem 6.1 of [2], then with λ_{γ} as the multipliers of that theorem we have

$$\lambda_{\gamma_k} = 0, \quad k = 1, \dots, \bar{p}.$$

Next let $\mu_{\alpha}(t)$ be the functions of Theorem 6.1 of [2] and define the solution arc a_0 to be a normal solution if there are $\Omega = p - \bar{p} + 2N$ (where \bar{p} is the con-

⁴ Unless otherwise noted, for a function $Z(t, x, u, b)$ the notation $Z(t)$ will denote $Z(t, x_0(t), u_0(t), b)$ for the arc a_0 under consideration.

⁵ See §11 of [2].

stant introduced above) admissible variations $\delta_\omega a$ ($\omega = 1, \dots, \Omega$) of a_0 which satisfy the conditions

$$(10a) \quad \delta_\omega \psi^\alpha(t^0) = 0 \quad \text{if } \mu_\alpha(t^0) \neq 0,$$

$$(10b) \quad \delta_\omega \psi^\alpha(t) = 0 \quad \text{on a neighborhood of } s^\alpha, \quad \alpha = 1, \dots, m,$$

in addition to providing that the matrix

$$(10c) \quad (J'_\rho(a_0, \delta_\omega a)), \quad \rho = 1, \dots, p + 2N, \quad \rho \neq \gamma_k, \quad \omega = 1, \dots, \Omega,$$

is nonsingular.

Define the class Y of admissible variations

$$y: \quad y^i(t), \quad v^k(t), \quad z^\sigma, \quad t^0 \leq t \leq t^1,$$

of a_0 which in addition satisfy⁶

$$(11a) \quad \delta_y \psi^\alpha(t^c) = 0 \quad \text{if } \mu_\alpha(t^c) \neq 0, \quad c = 0, 1,$$

$$(11b) \quad \delta_y \psi^\alpha(\bar{t}) = 0 \quad \text{if } \mu_\alpha(t) \text{ is not constant on a neighborhood of } \bar{t},$$

$$(11c) \quad \delta_y \psi^\alpha(t^0) = 0 \quad \text{if } \psi^\alpha(t^0) = 0, \quad 1 \leq \alpha \leq m,$$

$$(11d) \quad J'_\gamma(a_0, y) = 0 \quad \text{if } \lambda_\gamma \neq 0, \quad 1 \leq \gamma \leq p',$$

$$(11e) \quad J'_\gamma(a_0, y) \leq 0 \quad \text{if } \lambda_\gamma = 0, \quad \gamma \neq \gamma_k, \quad 1 \leq \gamma \leq p',$$

$$(11f) \quad J'_\rho(a_0, y) = 0, \quad p' < \rho \leq p + 2N,$$

where γ_k are the indices of (9b) and where the multipliers $\lambda_\gamma, \mu_\alpha(t)$ are those of Theorem 6.1 of [2].

At this point we include in our assumptions the following conditions: (i) the functions ψ^α are of class c^3 and c^1 respectively with respect to x and t and ψ_t^α is of class c^2 with respect to x while the functions $f^i, L_\gamma, g_\gamma, X^{ic}$ are of class c^2 ; (ii) the function $u_0(t)$ is continuous on $[t^0, t^1]$. For this problem we will prove the following theorem.

THEOREM 3.1. *Let a_0 be a normal solution to our problem. Then the multipliers $\lambda_\rho, \mu_\alpha(t), p_i(t), k^\alpha, \rho = 0, 1, \dots, p + N$, of Theorem 6.1 of [2] satisfy: (i) $\lambda_0 \neq 0$, and (ii) with $\lambda_0 = 1$, these multipliers are unique. Furthermore with \mathcal{G} and H as the terms of Theorem 6.1 of [2], then for each variation y in the class Y , we have that*

$$(12) \quad 0 \leq [(p_i(t^c)X_{b^\sigma b^\tau}^{ic})_{c=0}^1 + \mathcal{G}_{b^\sigma b^\tau} + K^\alpha \psi_{x^i x^j}^\alpha(t^0)X_{b^\sigma}^{i0}X_{b^\tau}^{j0}]z^\sigma z^\tau \\ - \int_{t^0}^{t^1} [H_{x^i x^j} y^i y^j + 2H_{x^i u^k} y^i v^k + H_{u^h u^k} v^h v^k] dt,$$

$$i, j = 1, \dots, N, \quad h, k = 1, \dots, K, \quad \sigma, \tau = 1, \dots, r, \quad \alpha = 1, \dots, m.$$

4. An imbedding lemma. In order to prove this result we will first prove the following lemma with the aid of the proof of Lemma 11.1 of [2].

LEMMA 4.1. *Let $a(\beta)$ be the \bar{h} parameter family of arcs*

$$a(\beta): \quad x(t, \beta), \quad u(t, \beta), \quad b(\beta), \quad |\beta| \leq \beta^0, \quad h = 1, \dots, \bar{h},$$

⁶ Where the notation $\delta_y \psi^\alpha(t)$ denotes the variation in the functions $\psi^\alpha(t)$ due to y . Similar notation will also be used for other functions.

associated with the admissible variations

$$\delta_h \omega: \quad \delta_h x(t), \quad \delta_h u(t), \quad \delta_h b, \quad h = 1, \dots, \bar{h},$$

as constructed in Lemma 11.1 of [2]. Then

$$(13a) \quad \frac{\partial x^i}{\partial \beta_h}(t, 0) = \delta_h x^i(t),$$

$$(13b) \quad \frac{\partial u^k}{\partial \beta_h}(t, 0) = \delta_h u^k(t),$$

$$(13c) \quad \frac{\partial b^\sigma(0)}{\partial \beta_h} = \delta_h b^\sigma, \quad h = 1, \dots, \bar{h}, \quad t^0 \leq t \leq t^1.$$

In addition, the functions

$$(14a) \quad \frac{\partial^2 x^i(t, \beta)}{\partial \beta_h \partial \beta_s},$$

$$(14b) \quad \frac{\partial^2 u^k(t, \beta)}{\partial \beta_h \partial \beta_s},$$

$$(14c) \quad \frac{\partial^2 b^\sigma(\beta)}{\partial \beta_h \partial \beta_s}, \quad h, s = 1, \dots, \bar{h}, \quad t' \leq t \leq t'', \quad |\beta| \leq \beta^0,$$

exist and are continuous where: (i) β^0 is the constant referred to above and (ii) $[t', t'']$ is an interval of continuity of $\delta_h u(t)$ ($1 \leq h \leq \bar{h}$).

Proof. As a first step, we note according to the continuity properties of the functions ψ^α and their derivatives together with the definition of the functions⁷ ϕ^α , that these latter functions together with their mixed partial derivatives up through the second order with respect to x and u are continuous on R . Hence these statements hold also for the functions $\phi^{m+1}, \dots, \phi^k$ introduced in § 8 of [2]. Referring to Lemma 10.1 of [2] and using these additional continuity properties of ϕ^Γ we see that the functions $U(t, x, s)$ of that lemma are continuous and have continuous first and second order mixed partial derivatives with respect to x , and s on a neighborhood of the arc α_0 .

Furthermore, by referring to the remarks following (121) of [2], the existence and continuity of the functions

$$(15) \quad \frac{\partial x^i(t, \beta)}{\partial \beta_h}, \quad \frac{\partial u^k(t, \beta)}{\partial \beta_h}, \quad |\beta| \leq \beta^0, \quad t' \leq t \leq t'',$$

(where $u(t, \beta)$ is defined as $U(t, x(t, \beta), s(t, \beta))$ as in [2]) is established, with $[t', t'']$ as an interval of continuity of $\delta_h u$, ($1 \leq h \leq \bar{h}$).

Now define the $N\bar{h}$ functions

$$(16) \quad K_h^j(t, k, \beta) = f_{x^i}^j k_h^i + f_{u^\tau}^j \left[U_{x^i k_h^i}^\tau + U_{s^\Gamma}^\tau \frac{\partial s^\Gamma}{\partial \beta_h}(t, \beta) \right],$$

$$i, j = 1, \dots, N, \quad h = 1, \dots, \bar{h}, \quad \tau, \Gamma = 1, \dots, K,$$

⁷ See remarks below (2).

where (i) the functions U are those of Lemma 10.1 of [2] and (ii) the arguments of the derivatives of f^j and U^r are $(t, x(t, \beta), s(t, \beta))$ as described in (115), (118) and (119–1) of [2].

By the properties of the functions $f, x(t, \beta), U, s(t, \beta)$, we may solve the system of differential equations⁸

$$(17a) \quad \begin{aligned} k_h^i &= K_h^i, \\ \dot{\beta} &= 0, \end{aligned}$$

with

$$(17b) \quad \begin{aligned} k_h^i(t^0, \beta) &= \frac{\partial x^i(t^0, \beta)}{\partial \beta_h}, \\ \bar{\beta}(t^0) &= \beta, \end{aligned} \quad h = 1, \dots, \bar{h},$$

for $|\beta| \leq \beta^0$, where this is the constant referred to above. We thus obtain the family of functions

$$(18) \quad k_h^i(t, \beta), \quad h = 1, \dots, \bar{h}, \quad t^0 \leq t \leq t^1, \quad |\beta| \leq \beta^0.$$

Now according to the definition of $x^i(t^0, \beta)$ given in (119–2) of [2] together with Lemma 10.4 of [2] we see that the functions $(\partial x^i(t^0, \beta)/\partial \beta_h)$ are of class C^1 . Then by the theorem referred to above, we see that the functions $k_h^j(t, \beta)$ and $(\partial k_h^j(t, \beta)/\partial \beta_s)$, $h, s = 1, \dots, \bar{h}$, are continuous on intervals $[t', t'']$ upon which $\delta_1 u(t), \dots, \delta_{\bar{h}} u(t)$ are continuous.

Next, by referring to the functions K_h^j of (16), we see that the functions $(\partial x^j(t, \beta)/\partial \beta_h)$ of (15) also satisfy the system (17). Then by the uniqueness of solutions to (17) we obtain

$$(19) \quad k_h^i(t, \beta) = \frac{\partial x^i(t, \beta)}{\partial \beta_h}, \quad h = 1, \dots, \bar{h}.$$

Furthermore by the properties of the functions $k_h^j(t, \beta)$ already established, we see that

$$(20a) \quad \frac{\partial^2 x^i(t, \beta)}{\partial \beta_h \partial \beta_s}, \quad h, s = 1, \dots, \bar{h}, \quad |\beta| \leq \beta^0, \quad t' \leq t \leq t'',$$

also exist and are continuous, where $[t', t'']$ is an interval upon which $\delta_1 u(t), \dots, \delta_{\bar{h}} u(t)$ are continuous.

By (20a) together with the definition of $s(t, \beta)$ and the properties of the functions U of Lemma 10.1 of [2] as described in the remarks above (15), we see that the functions

$$(20b) \quad \frac{\partial^2 u^k(t, \beta)}{\partial \beta_h \partial \beta_s}, \quad h, s = 1, \dots, \bar{h}, \quad |\beta| \leq \beta^0, \quad [t', t''],$$

are also continuous (where $[t', t'']$ is as described below (20a)).

⁸ See for example [1, Chap. 1, Thm. 14.2]. Also we have included the parameter β as the variable $\bar{\beta}$ in this system.

Finally, by using the functions $\theta^\sigma(\beta)$ of (121) of [2], we see that the above statement also holds for

$$(20c) \quad \frac{\partial^2 b(\beta)}{\partial \beta_h \partial \beta_s}, \quad h, s = 1, \dots, \bar{h}, \quad |\beta| \leq \beta^0, \quad [t', t''].$$

Thus statement (14) is proved.

In order now to establish statement (13), we observe that (119–2) and (114) of [2] yield

$$(21) \quad \frac{\partial x^i(t^0, 0)}{\partial \beta_h} = \delta_h x^i(t^0), \quad h = 1, \dots, \bar{h}.$$

Furthermore, by the definition of $u(t, \beta)$ together with $s(t, \beta)$ defined as

$$(22) \quad s^\Gamma(t, \beta) = \phi^\Gamma(t) + \beta_h [\phi_{x^i}^\Gamma(t) \delta_h x^i(t) + \phi_{u^k}^\Gamma(t) \delta_h u^k(t)],$$

$$h = 1, \dots, \bar{h}, \quad \Gamma = 1, \dots, K,$$

in (115–2) and (115–3) of [2], we obtain

$$(23) \quad \frac{\partial u^k(t, 0)}{\partial \beta_h} = U_{x^i}^k \frac{\partial x^i(t, 0)}{\partial \beta_h} + U_{s^\Gamma}^k [\phi_{x^i}^\Gamma \delta_h x^i + \phi_{u^\tau}^\Gamma \delta_h u^\tau],$$

$$k, \Gamma, \tau = 1, \dots, K, \quad h = 1, \dots, \bar{h},$$

where the bracketed term has argument t and represents the partial derivative of (22) with respect to β_h and where the argument of the partial derivatives of U are $(t, x(t, 0), s(t, 0))$.

Since $s^\Gamma(t, 0) = \phi^\Gamma(t)$ and $x(t, 0) = x_0(t)$, then by (102) of [2] we can write the relation (23) as

$$(24) \quad \frac{\partial u^k(t, 0)}{\partial \beta_h} = -\zeta_\Gamma^k \phi_{x^i}^\Gamma \frac{\partial x^i(t, 0)}{\partial \beta_h} + \zeta_\Gamma^k [\phi_{x^i}^\Gamma \delta_h x^i + \phi_{u^\tau}^\Gamma \delta_h u^\tau],$$

$$h = 1, \dots, \bar{h}, \quad \Gamma, k, \tau = 1, \dots, K,$$

where the matrix $(\zeta_\Gamma^k(t))$, $k, \Gamma = 1, \dots, K$, is the inverse of the matrix $(\phi_{u^k}^\Gamma(t))$ as introduced in (64) of [2]. Thus there results

$$(25) \quad \frac{\partial u^k(t, 0)}{\partial \beta_h} = \zeta_\Gamma^k \phi_{x^i}^\Gamma \left[\delta_h x^i - \frac{\partial x^i(t, 0)}{\partial \beta_h} \right] + \delta_h u^k,$$

$$h = 1, \dots, \bar{h}, \quad \Gamma, k = 1, \dots, K.$$

Now writing $x(t, \beta)$ for x in the right-hand side of (119–1) of [2], then differentiating with respect to β_h at $\beta = 0$ and using (25) results in

$$(26) \quad \frac{d}{dt} \left(\frac{\partial x^i(t, 0)}{\partial \beta_h} \right) = f_{x^j}^i \frac{\partial x^j}{\partial \beta_h} + f_{u^k}^i \left[\zeta_\Gamma^k \phi_{x^\rho}^\Gamma \left(\delta_h x^\rho - \frac{\partial x^\rho}{\partial \beta_h} \right) + \delta_h u^k \right],$$

$$k, \Gamma = 1, \dots, K, \quad i, j, \rho = 1, \dots, N, \quad h = 1, \dots, \bar{h},$$

where the unlisted arguments in the terms $\partial x / \partial \beta$ are $(t, 0)$ and in the other terms is t . Then according to the definition of an admissible variation as listed above (8), together with (21), we see by substituting the terms $\delta_h x$ for $(\partial x(t, 0) / \partial \beta_h)$ in

(26) that these functions both satisfy the same set of differential equations and pass through the same initial point. Then by the uniqueness of solutions to (26) and (21) we see that (13a) is proved. The relations (13b) and (13c) follow respectively from (13a) together with (25) and from the definition of the functions $\theta^\sigma(\beta)$ of (121) of [2]. Thus Lemma 4.1 is proved.

5. Proof of the first two statements of Theorem 3.1. By the construction of Lemma 11.1 of [2] our family $a(\beta)$ of Lemma 4.1, for $\beta_h \geq 0$, $h = 1, \dots, \bar{h}$, is in the class \mathcal{A} described above (71) of [2].

Now let $\tilde{\lambda}_\rho$ ($\rho = 0, 1, \dots, p + 2N$) be the constants of Theorem 8.2 of [2] and let $\delta_\omega a$ ($\omega = 1, \dots, \Omega$) be the variations described in (7). Then by (70), in addition to Lemmas 11.1 and 15.1 and the inequality (155-2) all from [2] together with Lemma 4.1 and the definition of the terms $J'_\rho(a_0, \delta_\omega a)$ given in (8), we have that

$$(27) \quad \tilde{\lambda}_\rho J'_\rho(a_0, \delta_\omega a) \geq 0, \quad \rho = 0, 1, \dots, p + 2N, \quad \omega = 1, \dots, \Omega,$$

where $J'_\rho(a_0, \delta_\omega a)$ is interpreted as a variation of J_ρ from the family $a(\beta)$ with $\beta_h \geq 0$ ($1 \leq h \leq \bar{h}$) of Lemma 4.1.

Furthermore, according to our definition of an admissible variation, we see that the variations $\delta_\omega a$, $\omega = 1, \dots, \Omega$, are also admissible. Thus by an argument analogous to that used above we see that (27) holds with $-\delta_\omega a$ replacing $\delta_\omega a$. Since

$$J'_\rho(a_0, -\delta_\omega a) = -J'_\rho(a_0, \delta_\omega a),$$

then we must have that

$$(28) \quad \tilde{\lambda}_\rho J'_\rho(a_0, \delta_\omega a) = 0, \quad \rho = 0, 1, \dots, p + 2N, \quad \omega = 1, \dots, \Omega,$$

(where Ω is always the constant introduced above (10)).

Next, according to § 7 of [2], the multipliers $p_i(t)$, $\mu_\alpha(t)$, K^α of Theorems 6.1 and 8.2 of [2] are the terms of (76) of [2] and the relationship between the multipliers $\tilde{\lambda}$, λ respectively of Theorems 8.2 and 6.1 of [2] is by (92) and (76) of [2]:

$$\lambda_\gamma = \tilde{\lambda}_\gamma, \quad \gamma = 0, 1, \dots, p, \quad \lambda_{p+i} = \tilde{\lambda}_{p+N+i} = K^\alpha \psi_{\alpha i}^*(t^0), \quad i = 1, \dots, N.$$

According to the definition of the multiplier $p_i(t)$ given in (76) and (66) of [2] we see that

$$(29a) \quad p_i(t^1) = -\tilde{\lambda}_{p+i},$$

and by the relationships indicated above (29) we see that with γ_k as the indices referred to in (9b),

$$(29b) \quad 0 = \lambda_{\gamma_k} = \tilde{\lambda}_{\gamma_k}, \quad k = 1, \dots, \bar{p}.$$

Thus in (28) we may restrict ρ not to be any of the indices γ_k , $k = 1, \dots, \bar{p}$, and obtain

$$(30) \quad \tilde{\lambda}_\rho J'_\rho(a_0, \delta_\omega a) = 0, \quad \rho = 0, 1, \dots, p + 2N, \quad \rho \neq \gamma_k, \quad \omega = 1, \dots, \Omega.$$

Now assume that $\tilde{\lambda}_0 = 0$. Then by (29b) together with (30) and the non-singularity of the matrix of (10c) we must have

$$(31) \quad \tilde{\lambda}_\rho = 0, \quad \rho = 0, 1, \dots, p + 2N,$$

According to the definition of $\mu_\alpha(t)$ (as given in (76) of [2]) and the relationships indicated above (29) and the nonsingularity of the matrix of (108) of [2] we obtain

$$(32a) \quad \begin{aligned} K^\alpha &= 0, \\ \mu_\alpha(t) &= 0, \quad t^0 \leq t \leq t^1. \end{aligned}$$

Thus by (31), (30) and (29a) there results in particular that

$$(32b) \quad \tilde{\lambda}_\gamma = 0, \quad (0 \leq \gamma \leq p), \quad K^\alpha = 0, \quad p_i(t^1) = 0, \quad \mu_\alpha(t^1) = 0,$$

which violates condition (81) of Theorem 8.2 of [2].

We have thus proved that if a_0 is a normal solution, then the multiplier $\tilde{\lambda}_0$ of Theorem 8.2 of [2] cannot vanish. In order to now prove the first statement of Theorem 3.1 we need only refer to the relationships above (29) to see that $\lambda_0 = \tilde{\lambda}_0$. Thus the first statement of Theorem 3.1 is established.

In order to prove the second statement of Theorem 3.1, let z be the $\Omega + 1$ -dimensional vector

$$\begin{aligned} z^0 &= J'_0(a_0, \delta a), \\ z^l &= J'_{\gamma_l}(a_0, \delta a), \quad 1 \leq \gamma_l \leq p, \quad \gamma_l \neq \gamma_k, \quad k = 1, \dots, \bar{p}, \\ z^{p-\bar{p}+j} &= J'_{p+j}(a_0, \delta a), \quad j = 1, \dots, 2N, \end{aligned}$$

for an admissible variation δa (where $\gamma_k, \bar{p}, \Omega$ are the terms introduced in (9) and in the remarks below (9) and $J'(a_0, \delta a)$ are the terms of (8)). In addition, let Λ be the $\Omega + 1$ -dimensional vector with components $\tilde{\lambda}_\rho, \rho = 0, 1, \dots, p + 2N, \rho \neq \gamma_k$. Then by (30) and the nonsingularity of the matrix of (10c) we see that the vector z belongs to an Ω -dimensional space Z satisfying

$$(33) \quad (\Lambda, z) = 0, \quad z \in Z.$$

Thus Λ must belong to a one-dimensional subset of vectors. Hence fixing $\tilde{\lambda}_0 = 1$ fixes all components of the vector Λ and hence also all of $\tilde{\lambda}_\rho, \rho = 1, \dots, p + 2N, \rho \neq \gamma_k$. By using (29) and the definition of $p_i(t)$ and $\mu_\alpha(t)$ together with the relations indicated above (29) and the nonsingularity of the matrix (108) from [2], we see that fixing $\tilde{\lambda}_\rho, \rho = 0, 1, \dots, p + 2N, \rho \neq \gamma_k$, then also fixes $p_i(t), \mu_\alpha(t)$ and K^α . Thus fixing $\tilde{\lambda}_0$ fixes all the $\tilde{\lambda}_\rho$ multipliers in addition to fixing $p_i(t), \mu_\alpha(t)$ and K^α . The proof of the second statement of Theorem 3.1 then follows from the relation above (29) between the multipliers of Theorems 6.1 and 8.2 of [2]. Thus the second statement of Theorem 3.1 is established.

6. Proof of the inequality (12). As a next step, consider the admissible variations $\delta_\omega a, \omega = 1, \dots, \Omega$, introduced in (10) together with an admissible variation y in the class Y introduced above (11). Thus we have the variations

$$\begin{aligned} \delta_\omega a: \quad & \delta_\omega x(t), \quad \delta_\omega u(t), \quad \delta_\omega b, \quad \omega = 1, \dots, \Omega, \\ y: \quad & y(t), \quad v(t), \quad z, \quad t^0 \leq t \leq t^1. \end{aligned}$$

With the vector $\beta = (\beta_1, \dots, \beta_\Omega)$ and π as parameters and referring to the proof of Lemma 11.1 of [2] we construct the family of arcs

$$a(\beta, \pi): \quad x(t, \beta, \pi), \quad u(t, \beta, \pi), \quad b(t, \beta, \pi), \quad |\beta| \leq \bar{\beta}, \quad |\pi| \leq \bar{\pi},$$

(where $\bar{\beta}$ and $\bar{\pi}$ are certain positive constants) which satisfy the system of differential equations (119) of [2] but for the present case. Furthermore, an examination of the proof of that lemma will show that because of property (10b) of the variations $\delta_\omega \alpha$, then the family so constructed satisfies

$$(34) \quad \psi^\alpha(t, \beta, \pi) \leq 0, \quad |\beta| \leq \bar{\beta}, \quad 0 \leq \pi \leq \bar{\pi}, \quad t^0 \leq t \leq t^1,$$

(i.e., with β not restricted to nonnegative values). Thus for the last indicated ranges of β, π , the family $\alpha(\beta, \pi)$ is in the class \mathcal{A} defined above (71) of [2]. We shall henceforth understand the term, the family $\alpha(\beta, \pi)$ to be that family with the parameter ranges of (34) unless otherwise specified.

As a next step, we see according to the definition of J_γ in (7a) that $J_\gamma(a_0) = I_\gamma(a_0)$, $\gamma = 0, 1, \dots, p$. Thus with γ_k as the indices (9) we have

$$(35) \quad J_{\gamma_k}(a_0) < 0, \quad k = 1, \dots, \bar{p}.$$

Furthermore, according to the continuity properties of the functions g_γ and L_γ defining I_γ , there is a neighborhood O of the arc a_0 such that a in O implies that $J_{\gamma_k}(a) < 0$, $k = 1, \dots, \bar{p}$.

By the above argument we then see that an arc a in the family $\alpha(\beta, \pi)$ will satisfy all the constraints of (1) if $a \in O$ and if

$$(36) \quad \begin{aligned} J_\gamma(\alpha(\beta, \pi)) &\leq 0, & \gamma \neq \gamma_k, 1 \leq \gamma \leq p', \\ J_\rho(\alpha(\beta, \pi)) &= 0, & \rho = p' + 1, \dots, p + 2N. \end{aligned}$$

We shall henceforth refer to an arc which satisfies all the conditions of (1) to be totally admissible.

In order to show that there is a one-parameter family of totally admissible arcs which is contained in the family $\alpha(\beta, \pi)$, consider the system of equations

$$(37) \quad J_\rho(\alpha(\beta, \pi)) - \pi J'_\rho(a_0, y) = 0, \quad \rho = 1, \dots, p + 2N, \quad \rho \neq \gamma_k,^9$$

which is defined for $|\beta| \leq \bar{\beta}$, $|\pi| \leq \bar{\pi}$. According to Lemma 4.1, the functional determinant of (37) with respect to $\beta_1, \dots, \beta_\Omega$ at $\beta = \pi = 0$, is the matrix (10c). The system (37) has the initial solution point $(\beta, \pi) = (0, 0)$. Then by the properties of the family $\alpha(\beta, \pi)$ for $|\beta| \leq \bar{\beta}$, $|\pi| \leq \bar{\pi}$, together with the implicit function theorem and the nonsingularity of the matrix of (10c), there is the solution to (37)

$$(38) \quad \beta_\omega = \beta_\omega(\pi), \quad \omega = 1, \dots, \Omega, \quad |\pi| \leq \hat{\pi},$$

(where $\hat{\pi}$ is a suitable positive constant) of class C^2 and satisfying $\beta_\omega(0) = 0$.

By reducing if necessary the value of $\hat{\pi}$, we can guarantee that $\alpha(\beta(\pi), \pi) \in O$ for $|\pi| \leq \hat{\pi}$, where O is the neighborhood introduced after (35).

Thus according to the properties of the variation y and the statement of (36) the family

$$(39) \quad \alpha(\beta(\pi), \pi), \quad 0 \leq \pi \leq \hat{\pi},$$

(where $\hat{\pi}$ has been further reduced if necessary) is a totally admissible family of arcs with $\alpha(\beta(0), 0) = a_0$.

⁹ Here, as always, γ_k are the indices of (9).

Furthermore, by substituting the solution $\beta_\omega(\pi)$ into (37) and differentiating with respect to π at $\pi = 0$, we obtain

$$(40) \quad J'_\rho(a_0, \delta_\omega a) \beta'_\omega(0) = 0, \quad \rho = 1, \dots, p + 2N, \quad \rho \neq \gamma_k, \quad \omega = 1, \dots, \Omega,$$

where $\beta'_\omega(0)$ denotes $d\beta_\omega(0)/d\pi$. By the nonsingularity of the matrix of (10c) we must then have

$$(41) \quad \beta'_\omega(0) = 0, \quad \omega = 1, \dots, \Omega.$$

Thus by Lemma 4.1 we obtain

$$(42) \quad \begin{aligned} \frac{dx(t, \beta(0), 0)}{d\pi} &= \delta_\omega x(t) \beta'_\omega(0) + y(t) = y(t), \\ \frac{du(t, \beta(0), 0)}{d\pi} &= \delta_\omega u(t) \beta'_\omega(0) + v(t) = v(t), \quad t^0 \leq t \leq t^1, \\ \frac{db(\beta(0), 0)}{d\pi} &= \delta_\omega b \beta'_\omega(0) + z = z, \end{aligned}$$

where the terms other than $\beta'_\omega(0)$ in the right-hand side of (42) are those introduced above (34).

Now let G be the function of (76) of [2] and define the function J on the family $\alpha(\beta(\pi), \pi)$ as

$$(43) \quad \begin{aligned} J(\pi) &= [p_i(t^c) X^{ic}(b(\pi))]_{c=0}^c + \tilde{\lambda}_\gamma g_\gamma(b(\pi)) + \int_{t^0}^{t^1} G(s, \pi) ds \\ &\quad + \tilde{\lambda}_{p+N+i} [X^{i0}(b(\pi)) - x^i(t^0, \pi)], \quad \gamma = 0, 1, \dots, p, \end{aligned}$$

where: (i) $p(t)$ are the multipliers previously referred to, (ii) X^{ic} and g_γ are defined in the problem statement in (1) and (iii) the argument (π) means $(\beta(\pi), \pi)$ for the family of (39).

Differentiating (43) at $\pi = 0$ and using (42) yields

$$(44) \quad \begin{aligned} \frac{dJ(0)}{d\pi} &= [p_i(t^c) \delta_\gamma X^{ic}]_{c=0}^c + \tilde{\lambda}_\gamma \delta_\gamma g_\gamma + \int_{t^0}^{t^1} [G_{x^i} y^i + G_{u^k} v^k] ds \\ &\quad + \tilde{\lambda}_{p+N+i} [\delta_\gamma X^{i0} - y^i(t^0)], \end{aligned}$$

where, for example, the term $\delta_\gamma X^{ic}$ is defined as $X_{b\sigma z}^{ic}$, i.e., the variation in the functions X^{ic} due to the variation γ and where in (44) we have used (42).

Furthermore by (80) and (89) of [2] we have

$$(45) \quad G_{x^i} = 0, \quad G_{u^k} = 0,$$

so that we may eliminate the integral in (44) and write that relation as

$$(46) \quad \begin{aligned} \frac{dJ(0)}{d\pi} &= [p_i(t^c) \delta_\gamma X^{ic}]_{c=0}^c + \tilde{\lambda}_\gamma \delta_\gamma g_\gamma + \tilde{\lambda}_{p+N+i} [\delta_\gamma X^{i0} - y^i(t^0)] \\ &= -\tilde{\lambda}_{p+N+i} y^i(t^0), \end{aligned}$$

the last equality following from the transversality condition (77) of [2].

According to the properties of our family of (37) as established in Lemma 4.1 together with the properties of the solution $\beta_\omega(\pi)$ of (41) we can differentiate (43) a second time to get¹⁰

$$\begin{aligned}
 \frac{d^2 J(0)}{d\pi^2} = & \left[p^i(t^c) \left(X_{b^\sigma b^\tau}^{ic} z^\sigma z^\tau + X_{b^\sigma}^{ic} \frac{d^2 b^\sigma}{d\pi^2} \right) \right]_{c=0}^{c=1} + \tilde{\lambda}_\gamma \left[g_{\gamma b^\sigma b^\tau} z^\sigma z^\tau + g_{\gamma b^\sigma} \frac{d^2 b^\sigma}{d\pi^2} \right] \\
 & + \int_{t^0}^{t^1} \left[G_{x^i x^j} y^i y^j + 2G_{x^i u^h} y^i v^h + G_{u^h u^k} v^h v^k + G_{x^i} \frac{d^2 x^i}{d\pi^2} \right. \\
 (47) \quad & \left. + G_{u^h} \frac{d^2 u^h}{d\pi^2} \right] ds + \tilde{\lambda}_{p+N+i} \left[X_{b^\sigma b^\tau}^{i0} z^\sigma z^\tau + X_{b^\sigma}^{i0} \frac{d^2 b^\sigma}{d\pi^2} - \frac{d^2 x^i}{d\pi^2}(t^0) \right], \\
 & i, j = 1, \dots, N, \quad \sigma, \tau = 1, \dots, r, \quad h, k = 1, \dots, K, \quad \gamma = 0, 1, \dots, P,
 \end{aligned}$$

where all derivatives with respect to b, x, u , are formed along α_0 while all derivatives with respect to π are formed at $(\beta(\pi), \pi) = (\beta(0), 0)$, i.e., at $\pi = 0$ on the family of (39).

As a next step we recall from the construction of Lemma 11.1 of [2] that the initial points $x(t^0, \beta(\pi), \pi)$ were selected with

$$x(t^0, \beta(\pi), \pi) = \bar{X}(\beta(\pi), \pi),$$

where the function \bar{X} was constructed such that

$$\begin{aligned}
 (48) \quad & \psi^\alpha(t^0, \bar{X}(\beta(\pi), \pi)) - [\psi^\alpha(t^0) + \psi_{x^i}^\alpha(t^0)(\beta_\omega(\pi)\delta_\omega x^i(t^0) + \pi y^i(t^0))] = 0, \\
 & \alpha = 1, \dots, m, \quad \omega = 1, \dots, \Omega,
 \end{aligned}$$

where the functions \bar{X} are those of Lemma 10.4 of [2]. According to the properties of the functions involved, we can differentiate twice with respect to π at $\pi = 0$ to obtain

$$\begin{aligned}
 (49) \quad & \psi_{x^i}^\alpha(t^0) \left[\frac{d^2 \bar{X}^i}{d\pi^2} - \beta_\omega'' \delta_\omega x^i(t^0) \right] + \psi_{x^i x^j}^\alpha(t^0) \frac{d\bar{X}^i}{d\pi} \frac{d\bar{X}^j}{d\pi} = 0, \\
 & i, j = 1, \dots, N, \quad \omega = 1, \dots, \Omega,
 \end{aligned}$$

where the unlisted arguments are at $\pi = 0$ and where $\beta_\omega''(0)$ denotes $(d^2 \beta_\omega(0)/d\pi^2)$. According to (77) from [2] together with (42) evaluated at t^0 , (45) and the relations above (29), we may use (49) to rewrite (47) as

$$\begin{aligned}
 \frac{d^2 J(0)}{d\pi^2} = & [[p_i(t^c) X_{b^\sigma b^\tau}^{ic}]_{c=0}^{c=1} + \tilde{\lambda}_\gamma g_{\gamma b^\sigma b^\tau} z^\sigma z^\tau + K^\alpha \psi_{x^i}^\alpha(t^0) [X_{b^\sigma b^\tau}^{i0} z^\sigma z^\tau - \beta_\omega''(0) \delta_\omega x^i(t^0)] \\
 (50) \quad & + K^\alpha \psi_{x^i x^j}^\alpha(t^0) y^i(t^0) y^j(t^0) + \int_{t^0}^{t^1} [G_{x^i x^j} y^i y^j + 2G_{x^i u^h} y^i v^h + G_{u^h u^k} v^h v^k] dt, \\
 & h, k = 1, \dots, K, \quad i, j = 1, \dots, N, \quad \sigma, \tau = 1, \dots, r, \quad \gamma = 0, 1, \dots, p.
 \end{aligned}$$

¹⁰ The terms $G_{x^i x^j}$ denote the second order partial derivative of G with respect to x^i, x^j and the other terms have similar definitions.

Recalling the definition of the function \mathcal{G} of Theorem 6.1 of [2] as listed in (5) and using the relations indicated above (29) together with property (11f) for the variation y , then (50) may be written as

$$(51) \quad \frac{d^2 J(0)}{d\pi^2} = [[p_i(t^c)X_{b^\sigma b^\tau}^{ic}]_{c=0}^{c=1} + \mathcal{G}_{b^\sigma b^\tau} + K^\alpha \psi_{x^i x^j}^\alpha(t^0)X_{b^\sigma}^{i0}X_{b^\tau}^{j0}]z^\sigma z^\tau \\ - K^\alpha \psi_{x^i}^\alpha(t^0)\beta_\omega''(0)\delta_\omega x^i(t^0) + \int_{t^0}^{t^1} [G_{x^i x^j} y^i y^j + 2G_{x^i u^h} y^i v^h + G_{u^h u^k} v^h v^k] dt.$$

Referring to (43), we see that according to the definition of G in (76) of [1] and J_ρ , $\rho = 0, 1, \dots, p + 2N$, in (8), there results

$$(52) \quad J(\pi) = J_0(\pi) + \tilde{\lambda}_\rho J_\rho(\pi) + \Lambda(\pi), \quad \rho = 1, \dots, p + 2N,$$

on the family of (39), where

$$\Lambda(\pi) \equiv \Lambda(\beta(\pi), \pi) = \int_{t^0}^{t^1} \mu_\alpha(s) \phi^\alpha(s, \beta(\pi), \pi) ds, \quad \alpha = 1, \dots, m,$$

and the functions μ_α, ϕ^α are as used in § 1. In addition, according to the properties (11b) for the variation y together with Lemma 2.1 and the property (10b) for the variations $\delta_\omega \alpha$, we obtain with α not summed,

$$(53) \quad \lim_{t \rightarrow \bar{t}} \left[\delta_\omega \psi^\alpha(\bar{t}) \frac{\mu_\alpha(t) - \mu_\alpha(\bar{t})}{t - \bar{t}} \right] = 0, \quad \omega = 1, \dots, \Omega, \\ \lim_{t \rightarrow \bar{t}} \left[\delta_y \psi^\alpha(\bar{t}) \frac{\mu_\alpha(t) - \mu_\alpha(\bar{t})}{t - \bar{t}} \right] = 0, \quad t^0 \leq \bar{t} \leq t^1, \quad \alpha \text{ not summed}.$$

Thus according to the admissibility of¹¹ the variations $\delta_\omega \alpha, y$ and the continuity properties of μ_α as expressed in Theorem 6.1 of [2], we have

$$(54a) \quad \frac{d}{dt} [\mu_\alpha(\bar{t}) \delta_\omega \psi^\alpha(\bar{t})] = \lim_{t \rightarrow \bar{t}} \frac{\mu_\alpha(t) \delta_\omega \psi^\alpha(t) - \mu_\alpha(\bar{t}) \delta_\omega \psi^\alpha(\bar{t})}{t - \bar{t}} \\ = \lim_{t \rightarrow \bar{t}} \mu_\alpha(t) \frac{\delta_\omega \psi^\alpha(t) - \delta_\omega \psi^\alpha(\bar{t})}{t - \bar{t}} + \lim_{t \rightarrow \bar{t}} \delta_\omega \psi^\alpha(\bar{t}) \frac{\mu_\alpha(t) - \mu_\alpha(\bar{t})}{t - \bar{t}} \\ = \mu_\alpha(\bar{t}) \delta_\omega \phi^\alpha(\bar{t}), \quad \omega = 1, \dots, \Omega, \quad \alpha \text{ not summed},$$

and similarly for y so that

$$(54b) \quad \frac{d}{dt} [\mu_\alpha(\bar{t}) \delta_y \psi^\alpha(\bar{t})] = \mu_\alpha(\bar{t}) \delta_y \phi^\alpha(\bar{t}), \\ \alpha = 1, \dots, m, \quad t^0 \leq t \leq t^1, \quad \alpha \text{ not summed}.$$

In order now to evaluate the term $\Lambda(\pi)$ introduced above, we recall from the construction of Lemma 11.1 of [2] but for the present case that

$$(55) \quad \phi^\alpha(t, \beta(\pi), \pi) = \phi^\alpha(t) + \beta_\omega(\pi) \delta_\omega \phi^\alpha(t) + \pi \delta_y \phi^\alpha(t), \\ \alpha = 1, \dots, m, \quad t^0 \leq t \leq t^1.$$

¹¹ See (11.1) of [2].

Then as the family $\alpha(\beta(\pi), \pi)$ contains α_0 for $\pi = 0$ we must have

$$(56) \quad \Lambda(\pi) - \Lambda(0) = \int_{t^0}^{t^1} \mu_\alpha(s) [\beta_\omega(\pi) \delta_\omega \phi^\alpha(s) + \pi \delta_\gamma \phi^\alpha(s)] ds.$$

However, by (54) this becomes

$$(57) \quad \begin{aligned} \Lambda(\pi) - \Lambda(0) &= \int_{t^0}^{t^1} \pi \left[\frac{d}{ds} [\mu_\alpha(s) \delta_\gamma \psi^\alpha(s)] + \beta_\omega(\pi) \frac{d}{ds} [\mu_\alpha(s) \delta_\omega \psi^\alpha(s)] \right] ds \\ &= [\mu_\alpha(t^c) (\beta_\omega(\pi) \delta_\omega \psi^\alpha(t^c) + \pi \delta_\gamma \psi^\alpha(t^c))]_{t^0}^{t^1} = 0, \end{aligned}$$

the last equality following from properties (11a) of the variation γ and from properties (10a) and (10b) of the variations $\delta_\omega \alpha$ together with the property listed below (6) for the multipliers $\mu_\alpha(t)$. Thus by (52) we have

$$(58) \quad J(\pi) - J(0) = J_0(\pi) = J_0(0) + \tilde{\lambda}_\rho [J_\rho(\pi) - J_\rho(0)], \quad \rho = 0, 1, \dots, p + 2N.$$

Next, by (29), together with the definition of the family of (39) under consideration, the relation (37) and the properties of the variation γ , there results

$$(59) \quad \tilde{\lambda}_\rho J_\rho(\pi) = 0, \quad \rho = 1, \dots, p + 2N, \quad \rho \text{ not summed.}$$

Then (58) becomes

$$(60) \quad J(\pi) - J(0) = J_0(\pi) - J_0(0)$$

for the family of (39).

However since the arc α_0 is a solution to our problem and as the family of (39) is totally admissible, then we must have that

$$(61) \quad J(\pi) - J(0) \geq 0.$$

Now according to the relation between the function H of Theorem 6.1 of [2] and the function G of (51)

$$(62) \quad H = -\dot{p}_i(t)x^i - G$$

as given in (93) of [2], we see that in (51) the derivatives of G may be replaced by the derivatives of H . Then by (46), (51) and these remarks, the inequality (12) will be proved if we prove that

$$(63a) \quad \tilde{\lambda}_{p+N+i} y^i(t^0) = 0,$$

$$(63b) \quad \tilde{\lambda}_{p+N+i} \beta''_\omega(0) \delta_\omega x^i(t^0) = K^\alpha \psi_{x^i}^\alpha(t^0) \beta''_\omega(0) \delta_\omega x^i(t^0) = 0, \quad \omega = 1, \dots, \Omega,$$

the first equality in (63b) following from the relations above (29). This result will be established in the following section.

7. An auxiliary lemma.

LEMMA 7.1. Let $\alpha_j, j = 1, \dots, \hat{m}$, be those indices such that $\psi^\alpha(t^0) = 0$. Then given the N -dimensional vector $D = (d^1, \dots, d^N)$ satisfying

$$(64) \quad \begin{aligned} \psi_{x^i}^\alpha(t^0) d^i &= 0, & j &= 1, \dots, \hat{m}, \\ \psi_{x^i}^\beta(t^0) d^i &= 0 \quad \text{if } \mu_\beta(t^0) \neq 0, \end{aligned}$$

the relation

$$(65) \quad \tilde{\lambda}_{p+N+i} d^i = 0$$

is also true.

Proof. According to the above definitions together with the properties of the functions ψ^α , we may select a positive constant δ so small that

$$(66) \quad \psi^\beta(t) < 0, \quad \beta \neq \alpha_j, \quad j = 1, \dots, \hat{m}, \quad t^0 \leq t \leq t^0 + \delta.$$

Define the K -dimensional arc w satisfying

$$(67a) \quad w^\alpha(t^0) = \psi_{x^i}^\alpha(t^0) d^i,$$

$$(67b) \quad \dot{w}^\alpha(t) = \begin{cases} (-\psi_{x^i}^\alpha(t^0) d^i) \frac{2}{\delta}, & t^0 \leq t \leq t^0 + \delta/2, \\ 0, & t^0 + \delta/2 \leq t \leq t^1, \quad \alpha = 1, \dots, m, \end{cases}$$

$$(67c) \quad w^\Gamma(t) \equiv 0, \quad \Gamma = m+1, \dots, K, \quad t^0 \leq t \leq t^1.$$

Then w is in the class W of § 13 of [2]. By Lemma 13.1 of [2], we can find an admissible variation

$$\delta a: \quad \delta x(t), \quad \delta u(t), \quad \delta b, \quad t^0 \leq t \leq t^1,$$

satisfying

$$(68a) \quad \delta x^{j_s}(t^0) = d^{j_s}, \quad j_s \neq i_\rho, \quad s = 1, \dots, N-m,$$

$$(68b) \quad \delta b = 0,$$

where i_ρ are the indices of (108) of [2] and such that

$$(68c) \quad \psi_{x^i}^\alpha(t) \delta x^i(t) \equiv \delta \psi^\alpha(t) = w^\alpha(t), \quad \alpha = 1, \dots, m,$$

$$(68d) \quad \delta \phi^\Gamma(t) = w^\Gamma(t), \quad \Gamma = m+1, \dots, K, \quad t^0 \leq t \leq t^1.$$

According to the above, and due to the admissibility of δa we have that

$$(69a) \quad \delta \phi^\alpha(t) = \frac{d}{dt} \delta \psi^\alpha(t) = \dot{w}^\alpha(t) = \begin{cases} (-\psi_{x^i}^\alpha(t^0) d^i) \frac{2}{\delta}, & t^0 \leq t \leq t^0 + \delta/2, \\ 0, & t^0 + \delta/2 \leq t \leq t^1, \end{cases}$$

and by (67c) and (68d) also,

$$(69b) \quad \delta \phi^\Gamma(t) = 0, \quad \Gamma = m+1, \dots, K, \quad t^0 \leq t \leq t^1.$$

Furthermore, by (67a) and (68a) together with (68c) evaluated at t^0 we obtain

$$(70) \quad \frac{\partial \psi^\alpha(t^0)}{\partial x^{i_\rho}} [d^{i_\rho} - \delta x^{i_\rho}(t^0)] = 0, \quad \rho, \alpha = 1, \dots, m,$$

where i_ρ are the indices of (68a). Then by the nonsingularity of the matrix $(\partial \psi^\alpha(t^0)/\partial x^{i_\rho})$ (see (108) of [2]) we see that $\delta x^{i_\rho}(t^0) = d^{i_\rho}$, $\rho = 1, \dots, m$, so that together with (68a) this becomes

$$(71) \quad \delta x^i(t^0) = d^i, \quad i = 1, \dots, N.$$

Next, by (155-2) and Lemmas 11.1 and 15.1 all of [2] together with (69) and (68b) we get by computing the variations of our functionals J_ρ with the above formed variation, that

$$(72) \quad \tilde{\lambda}_\rho \int_{t^0}^{t^0+\delta/2} F_{\rho u^k} \zeta_\alpha^k (-\psi_{x^i}^\alpha(t^0) d^i) \frac{2}{\delta} dt - \tilde{\lambda}_{p+N+i} d^i \geq 0, \quad \rho = 0, 1, \dots, p+N,$$

(where the functions F_ρ and ζ_α^k are defined respectively in (67) and (64) of [2]). Recalling the definition of the multipliers $\mu_\alpha(t)$ given in (74) and (76), both of [2], we see that (72) may be written in the form

$$(73) \quad (-\psi_{x^i}^\alpha(t^0) d^i) \frac{2}{\delta} \int_{t^0}^{t^0+\delta/2} \mu_\alpha dt - \tilde{\lambda}_{p+N+i} d^i \geq 0.$$

Now if $\alpha = \alpha_j$, $1 \leq j \leq \hat{m}$, (where these are the indices of (64)) then the first term of (73) vanishes, while if $\alpha \neq \alpha_j$, $j = 1, \dots, \hat{m}$, then by Lemma 2.1 we know that $\mu_\alpha(t)$ is constant on $[t^0, t^0 + \delta/2]$ so that the first term of (73) becomes for each such α ,

$$(74) \quad -\mu_\alpha(t^0) \psi_{x^i}^\alpha(t^0) d^i, \quad \alpha \text{ not summed},$$

which according to the second part of (64) also vanishes. Thus the first term of (73) vanishes for $\alpha = 1, \dots, m$ and we get

$$(75) \quad -\tilde{\lambda}_{p+N+i} d^i \geq 0.$$

Now replace the vector D of the hypothesis by $-D$. The above argument holds in an analogous fashion for $-D$ with the analogue of (75) being

$$(76) \quad -\tilde{\lambda}_{p+N+i} (-d^i) \geq 0.$$

Thus we obtain

$$(77) \quad \tilde{\lambda}_{p+N+i} d^i = 0,$$

and the lemma is proved.

By referring to properties (10a) and (10b) for the variations $\delta_\omega a$ and to properties (11a) and (11c) for the variation y , we see that the hypotheses of Lemma 7.1 is satisfied by these variations. Thus (63) is proved and Theorem 3.1 is established.

REFERENCES

- [1] M. R. HESTENES, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.
- [2] I. B. RUSSAK, *On problems with bounded state variables*, J. Optimization Theory Appl., 5 (1970), pp. 114-157.
- [3] ———, *On general problems with bounded state variables*, Ibid., 6 (1970), pp. 425-452.

ON THE CONCEPT OF SYMMETRY IN PONTRYAGIN'S MAXIMUM PRINCIPLE*

MINEO IKEDA AND KENTARO SAKAMOTO†

Abstract. From the standpoint of Lie group theory, the concept of symmetry (invariance) is introduced for a class of dynamical control problems which are treated on the basis of Pontryagin's maximum principle. The so-called linear first integrals of a holonomic dynamical system are extended to the framework of the maximum principle, leading to the concept of special linear first integrals. The symmetry group of a control system can be generated by a set of these integrals. Particular attention is paid to the case where the dynamical system is a natural one, and it is shown that the kinematical symmetry group of a natural system is locally isomorphic to the symmetry group of the corresponding control system. These results are applied to the problem of optimum trajectories for space navigation.

1. Introduction. The group-theoretical method is extremely useful when natural phenomena are studied with particular reference to their symmetry (invariance under a transformation group of some kind). For instance, many research workers have discussed the symmetry problem in the physics of elementary particles and atomic nuclei, and a great many contributions have been made to the unified description of experimental results obtained using cosmic rays and enormous accelerators.¹ In this connection, the symmetries in both classical and quantum mechanics have been reexamined from the standpoint of group theory (e.g., [6], [17]).

It may be expected that such symmetries also exist in various branches other than the fundamental sciences and that their clarification would bring about outstanding progress in the studies of these branches. In the present work, a heuristic introduction of the symmetry concept will be performed for a class of dynamical control problems which are treated on the basis of Pontryagin's maximum principle [16]. Particular attention will be paid to the case where a holonomic dynamical system is controlled by an external or internal force.

It will be found in § 2 how the conjugate variables transform when a coordinate transformation is carried out in the configuration space of a holonomic dynamical system. In § 3 and § 4 the so-called linear first integrals of the dynamical system will be extended to the framework of the maximum principle, leading to the concept of the special linear first integral. The symmetry group of a control system can be generated by a set of special linear first integrals. § 5 will be devoted to the case where the dynamical system is a natural one [14], and it will be shown in § 6 that the kinematical symmetry group of a natural system is locally isomorphic to the symmetry group of the corresponding control system. In § 7 these results will be applied to the problem of optimum rocket trajectories in a central force field.

2. Transformation law of conjugate variables. We shall consider a holonomic dynamical system with N degrees of freedom and describe it within the frame-

* Received by the editors July 16, 1973.

† Department of Applied Mathematics and Physics, Faculty of Engineering, Kyoto University, Kyoto, Japan.

¹ Though there are innumerable papers concerning this problem, we cite only [5] here.

work of Pontryagin's maximum principle. We begin by studying how the conjugate variables are transformed when the coordinates are changed in the configuration space of the dynamical system.

For a holonomic dynamical system with N degrees of freedom, the kinetic energy is a positive definite quadratic form in the velocities \dot{x}^i , namely,

$$T = \frac{1}{2}g_{ij}(x)\dot{x}^i\dot{x}^j,$$

where the coefficients g_{ij} are functions of the coordinates x^i . Latin indices take the values $1, 2, \dots, N$ and the summation convention is adopted. We note that the discussions are local throughout the present paper and that functions are assumed to be differentiable up to any necessary order.

The configuration space V_N of the dynamical system can be regarded as an N -dimensional Riemannian space endowed with the fundamental tensor g_{ij} . If we use the Christoffel symbols $\left\{ \begin{smallmatrix} i \\ jk \end{smallmatrix} \right\}$ formed from g_{ij} , the equations of motion are written as

$$(2.1) \quad \begin{aligned} \dot{x}^i &= y^i, \\ \dot{y}^i &= f^i - \left\{ \begin{smallmatrix} i \\ jk \end{smallmatrix} \right\} y^j y^k, \end{aligned}$$

where the f^i , the components of the force, are functions of x^i .

Let us describe this dynamical system in the framework of Pontryagin's maximum principle [16]. We first introduce the Hamiltonian defined by

$$(2.2) \quad H = y^i \varphi_i + \left(f^i - \left\{ \begin{smallmatrix} i \\ jk \end{smallmatrix} \right\} y^j y^k \right) \psi_i,$$

where φ_i and ψ_i denote the variables conjugate to x^i and y^i respectively. Making use of this Hamiltonian, we can derive (2.1) as part of the canonical equations,

$$(2.3) \quad \dot{x}^i = \frac{\partial H}{\partial \varphi_i}, \quad \dot{y}^i = \frac{\partial H}{\partial \psi_i}.$$

The remaining part becomes

$$(2.4) \quad \begin{aligned} \dot{\varphi}_i &= -\frac{\partial H}{\partial x^i} = -\left(\frac{\partial f^i}{\partial x^i} - \frac{\partial}{\partial x^i} \left\{ \begin{smallmatrix} i \\ jk \end{smallmatrix} \right\} y^j y^k \right) \psi_i, \\ \dot{\psi}_i &= -\frac{\partial H}{\partial y^i} = -\varphi_i + 2 \left\{ \begin{smallmatrix} i \\ lj \end{smallmatrix} \right\} y^j \psi_i. \end{aligned}$$

In what follows, we shall call the system with the Hamiltonian (2.2) a *free* system in the sense that no control force is applied.

Now, it is evident that the velocities y^i and the force f^i are transformed as contravariant vectors under a coordinate transformation in the configuration space V_N . Let us study the transformation law of the conjugate variables φ_i and ψ_i . For this purpose, we assume that the Hamiltonian H is a scalar, since it is reasonable to consider that H is independent of the choice of coordinates.

LEMMA 2.1. *In order that the Hamiltonian defined by (2.2) should be a scalar, it is necessary and sufficient that the conjugate variables have the transformation characters*

$$(2.5) \quad \varphi'_j = \frac{\partial x^i}{\partial x^{j'}} \varphi_i + \frac{\partial x^{k'}}{\partial x^i} \frac{\partial^2 x^i}{\partial x^{j'} \partial x^{k'}} y^{l'} \psi_l,$$

$$(2.6) \quad \psi'_j = \frac{\partial x^i}{\partial x^{j'}} \psi_i.$$

Proof. We substitute in (2.2) the transformation formulas for y^i , f^i and $\{^i_{jk}\}$. We have

$$H = y^{j'} \frac{\partial x^i}{\partial x^{j'}} \varphi_i + f^{j'} \frac{\partial x^i}{\partial x^{j'}} \psi_i - \left(\left\{ j \right\}'_{kl} \frac{\partial x^i}{\partial x^{j'}} - \frac{\partial^2 x^i}{\partial x^{k'} \partial x^{l'}} \right) y^{k'} y^{l'} \psi_i,$$

which must be equal, independently of $y^{i'}$ and $f^{i'}$, to the Hamiltonian in the primed coordinate system. Thus we obtain (2.5) and (2.6). The converse is obvious.

Remark. (2.6) means that ψ_i is a covariant vector.

3. Special linear first integrals. In the symmetry theory of holonomic dynamical systems, an important role is played by the so-called linear first integrals, that is, first integrals which are linear forms in the momenta p_i . These integrals have the property that they generate point transformations in the configuration space V_N . We shall study, in the framework of the maximum principle, first integrals which have this same property.

We first consider an infinitesimal point transformation in the configuration space,

$$(3.1) \quad 'x^i = x^i + \varepsilon \xi^i(x),$$

where ε denotes an infinitesimal parameter and the ξ^i , functions of the coordinates x^i , are the components of a contravariant vector. It is to be noted that we use the notation $'x^i$ for the transformed point and $x^{i'}$ for the transformed coordinates (cf. § 2). In order to obtain the transformation formula for the velocities y^i , we assume x^i and $'x^i$ to be functions of time t . We differentiate (3.1) with respect to t and replace \dot{x}^i by y^i . Thus we have

$$(3.2) \quad 'y^i = y^i + \varepsilon \frac{\partial \xi^i}{\partial x^j} y^j.$$

This formula is usually used when one discusses the extension of a group in the theory of Lie transformation groups (cf. § 6) [4].

(3.1) and (3.2) give an infinitesimal point transformation in the space of the state variables (x^i, y^i) . These equations can be obtained as a canonical transformation with the generating function

$$(3.3) \quad L = \xi^i \varphi_i + \frac{\partial \xi^i}{\partial x^j} y^j \psi_i.$$

Concerning this function we have the following lemma.

LEMMA 3.1. *If the conjugate variables φ_i and ψ_i have the transformation characters (2.5) and (2.6) under a coordinate transformation in the configuration space V_N , the function L defined by (3.3) is a scalar.*

Proof. We substitute (2.5), (2.6) and the transformation formulas for y^i and ξ^i into the expression for L in the primed coordinates. We find $L' = L$ by simple calculation.

We next study a condition for L to be a first integral of the free system with the Hamiltonian (2.2). This condition is written as $\{L, H\} = 0$, where $\{\cdot, \cdot\}$ denotes the Poisson bracket defined by

$$(3.4) \quad \{f_1, f_2\} = \frac{\partial f_1}{\partial \varphi_i} \frac{\partial f_2}{\partial x^i} + \frac{\partial f_1}{\partial \psi_i} \frac{\partial f_2}{\partial y^i} - \frac{\partial f_1}{\partial x^i} \frac{\partial f_2}{\partial \varphi_i} - \frac{\partial f_1}{\partial y^i} \frac{\partial f_2}{\partial \psi_i}.$$

THEOREM 3.2. *For a free system, the function L of (3.3) is a first integral if and only if*

$$(3.5) \quad \mathcal{L}_\xi \left\{ \begin{smallmatrix} i \\ jk \end{smallmatrix} \right\} \equiv \frac{\partial^2 \xi^i}{\partial x^j \partial x^k} + \xi^l \frac{\partial}{\partial x^l} \left\{ \begin{smallmatrix} i \\ jk \end{smallmatrix} \right\} + \left\{ \begin{smallmatrix} i \\ lk \end{smallmatrix} \right\} \frac{\partial \xi^l}{\partial x^j} + \left\{ \begin{smallmatrix} i \\ jl \end{smallmatrix} \right\} \frac{\partial \xi^l}{\partial x^k} - \left\{ \begin{smallmatrix} l \\ jk \end{smallmatrix} \right\} \frac{\partial \xi^i}{\partial x^l} = 0,$$

$$(3.6) \quad \mathcal{L}_\xi f^i \equiv \xi^j \frac{\partial f^i}{\partial x^j} - f^j \frac{\partial \xi^i}{\partial x^j} = 0,$$

where \mathcal{L}_ξ denotes the Lie derivative with respect to the vector ξ^i [19].

Proof. If we substitute (2.2) and (3.3) into $\{L, H\} = 0$ and rearrange the result, we obtain

$$\{L, H\} = -\psi_i y^j y^k \mathcal{L}_\xi \left\{ \begin{smallmatrix} i \\ jk \end{smallmatrix} \right\} + \psi_i \mathcal{L}_\xi f^i.$$

This equation must vanish independently of y^i and ψ_i , and we have (3.5) and (3.6).

Remark. (3.5) and (3.6) show that the transformation (3.1) is an affine motion which leaves the force f^i invariant.

As is seen from (3.3), the function L is a linear form in the conjugate variables φ_i and ψ_i , the coefficients being particular functions of the state variables x^i and y^i . When ξ^i satisfy (3.5) and (3.6), we shall call the function L a *special linear first integral* within the framework of the maximum principle.

4. A class of control systems. In this section, we shall discuss a problem of minimum-time control in the case where an external control force is applied to the system treated in the last section. In particular, we shall find a condition for L of (3.3) to be a first integral of this case.

Let us assume that the magnitude of the control force is a function $k(t)$ of time t such that $0 \leq k(t) \leq k_0$, k_0 being a given positive constant. Then the second equation of (2.1) is replaced by

$$(4.1) \quad \dot{y}^i = f^i - \left\{ \begin{smallmatrix} i \\ jk \end{smallmatrix} \right\} y^j y^k + k \theta^i,$$

where θ^i denotes the contravariant unit vector representing the direction of the control force. The corresponding Hamiltonian is given by (cf. (2.2))

$$(4.2) \quad H = y^i \varphi_i + \left(f^i - \left\{ \begin{smallmatrix} i \\ jk \end{smallmatrix} \right\} y^j y^k + k \theta^i \right) \psi_i.$$

It is evident from the form of (4.2) that Pontryagin's maximum principle is valid for the minimum-time control problem. When the Hamiltonian (4.2) is maximum with respect to the control force, θ^i has the direction of ψ^i and k takes its maximum value k_0 , that is,

$$(4.3) \quad \theta^i = \psi^i (g^{jk} \psi_j \psi_k)^{-1/2}, \quad k = k_0.$$

Throughout the present paper, indices are lowered and raised by means of g_{ij} and its conjugate g^{ij} , respectively. On using (4.3) the Hamiltonian (4.2) is reduced to

$$(4.4) \quad H_o = y^i \varphi_i + \left(f^i - \left\{ \begin{matrix} i \\ jk \end{matrix} \right\} y^j y^k \right) \psi_i + k_0 (g^{ij} \psi_i \psi_j)^{1/2}.$$

We shall call a system with the Hamiltonian (4.4) an *optimum system*. The following statements hold good for this system.

LEMMA 4.1. *Under the assumption of Lemma 3.1, the Hamiltonian H_o for the optimum system is a scalar.*

This is obvious, since the new term of (4.4), $k_0 (g^{ij} \psi_i \psi_j)^{1/2}$, is a scalar.

THEOREM 4.2. *For the optimum system, a necessary and sufficient condition for L of (3.3) to be a first integral is given by (3.6) and the Killing equations*

$$(4.5) \quad \mathcal{L}_\xi g_{ij} \equiv \nabla_i \xi_j + \nabla_j \xi_i = 0,$$

where ∇_i denotes covariant differentiation with respect to g_{ij} .

Proof. In the same way as in the proof of Theorem 3.2 we have

$$\{L, H_o\} = -\psi_i y^j y^k \mathcal{L}_\xi \left\{ \begin{matrix} i \\ jk \end{matrix} \right\} + \psi_i \mathcal{L}_\xi f^i + k_0 (g^{kl} \psi_k \psi_l)^{1/2} \psi_i \psi_j \mathcal{L}_\xi g^{ij}.$$

Since this must vanish for arbitrary y^i and ψ_i , we have

$$\mathcal{L}_\xi \left\{ \begin{matrix} i \\ jk \end{matrix} \right\} = 0, \quad \mathcal{L}_\xi f^i = 0, \quad \mathcal{L}_\xi g^{ij} = 0.$$

As is well known, the first equation follows from the third, while the latter equation is equivalent to $\mathcal{L}_\xi g_{ij} = 0$. This proves the theorem.

Remark. This theorem means that the transformation (3.1) together with (3.6) and (4.5) is an isometry in V_N which leaves the force f^i invariant.

5. The case associated with a natural dynamical system. In the following discussions, we shall assume that the original holonomic system is a natural one² in the sense of Pars [14]. The purpose of this section is to clarify the relation among first integrals of the three corresponding systems, i.e., natural, free (cf. § 2) and optimum (cf. § 4) systems.

For a natural dynamical system, there is a scalar function $U(x)$ representing the potential energy, from which the force f^i is derived as

$$(5.1) \quad f_i = -\nabla_i U.$$

² Synge and Griffith call this a simple dynamical system [18].

The coordinates x^i and the momenta p_i are conjugate to each other. The point transformation (3.1) in the configuration space V_N has a generating function

$$(5.2) \quad L_n = \xi^i p_i.$$

This is a first integral of the natural system, if and only if the ξ^i satisfy the Killing equations (4.5) and

$$(5.3) \quad \mathcal{L}_\xi U \equiv \xi^i \nabla_i U = 0.$$

Linear first integrals of the natural system have the following relations to special linear first integrals within the framework of Pontryagin's maximum principle.

PROPOSITION 5.1. *If the function L_n in (5.2) is a first integral of the natural dynamical system, then L defined by (3.3) with the same ξ^i as L_n is a first integral of the corresponding optimum system.*

Proof. We differentiate (5.3) covariantly and use the commutability of covariant and Lie differentiations. The result is given by (3.6) together with (5.1), and the proposition follows from Theorem 4.2.

PROPOSITION 5.2. *If the function L in (3.3) is a first integral of the optimum system, it is also a first integral of the corresponding free system.*

Proof. The proposition is obvious from Theorem 3.2, since (3.5) can be derived from (4.5).

COROLLARY 5.3. *Under the assumption of Proposition 5.1, the function L defined by (3.3) with the same ξ^i as L_n is a first integral of the corresponding free system.*

This corollary is obtained by combining Propositions 5.1 and 5.2.

6. Symmetry in the maximum principle. We are interested in the case where canonical transformations generated by a set of first integrals constitute a finite continuous group in the sense of Lie [4]. We shall study this case in the following.

The concept of a kinematical (geometrical) symmetry group is well known for a natural dynamical system [6], [17]. It is locally isomorphic to a group of isometries in the configuration space V_N which leave the potential function U invariant. This group will be referred to as a *scalar-preserving isometry group* in V_N [8]. We denote the generators of the group by

$$(6.1) \quad X_\alpha = \xi_\alpha^i \frac{\partial}{\partial x^i},$$

where Greek indices take the values $1, 2, \dots, r$. These generators are associated with infinitesimal transformations of the form (3.1) together with (4.5) and (5.3). They satisfy the commutation relations

$$(6.2) \quad [X_\alpha, X_\beta] = C_{\alpha\beta}^\gamma X_\gamma \quad (C_{\alpha\beta}^\gamma = \text{const.}),$$

where the left-hand side denotes the Lie bracket of X_α and X_β .

We consider the extension of a scalar-preserving isometry group. The extended group is generated by infinitesimal transformations of the form (3.1) and (3.2). For these transformations let us introduce the operators

$$(6.3) \quad Y_\alpha = \xi_\alpha^i \frac{\partial}{\partial x^i} + \frac{\partial \xi_\alpha^i}{\partial x^j} y^j \frac{\partial}{\partial y^i}.$$

The following fact is well known.

PROPOSITION 6.1. *The operators Y_α satisfy the same commutation relations as X_α , namely, $[Y_\alpha, Y_\beta] = C_{\alpha\beta}^\gamma Y_\gamma$.*

We next take up the formalism of Pontryagin's maximum principle. If a free or optimum system admits a special linear first integral (3.3), the infinitesimal canonical transformation generated by (3.3) is given by (3.1), (3.2) and

$$\begin{aligned} \varphi_i &= \varphi_i - \varepsilon \frac{\partial L}{\partial x^i} = \varphi_i - \varepsilon \left(\frac{\partial \xi^j}{\partial x^i} \varphi_j + \frac{\partial^2 \xi^j}{\partial x^i \partial x^k} y^k \psi_j \right), \\ \psi_i &= \psi_i - \varepsilon \frac{\partial L}{\partial y^i} = \psi_i - \varepsilon \frac{\partial \xi^j}{\partial x^i} \psi_j. \end{aligned} \quad (6.4)$$

For this transformation we define the operators

$$Z = \xi^i \frac{\partial}{\partial x^i} + \frac{\partial \xi^j}{\partial x^i} y^j \frac{\partial}{\partial y^i} - \left(\frac{\partial \xi^j}{\partial x^i} \varphi_j + \frac{\partial^2 \xi^j}{\partial x^i \partial x^k} y^k \psi_j \right) \frac{\partial}{\partial \varphi_i} - \frac{\partial \xi^j}{\partial x^i} \psi_j \frac{\partial}{\partial \psi_i}, \quad (6.5)$$

which have the following properties.

LEMMA 6.2. *If a special linear first integral L and an operator Z have the same ξ^i , we obtain*

$$Zf = \{L, f\}, \quad (6.6)$$

where f is a function of x^i, y^i, φ_i and ψ_i .

Proof. Making use of the generating function (3.3) for the canonical transformation (3.1), (3.2) and (6.4), we can express Z as

$$Z = \frac{\partial L}{\partial \varphi_i} \frac{\partial}{\partial x^i} + \frac{\partial L}{\partial \psi_i} \frac{\partial}{\partial y^i} - \frac{\partial L}{\partial x^i} \frac{\partial}{\partial \varphi_i} - \frac{\partial L}{\partial y^i} \frac{\partial}{\partial \psi_i},$$

which completes the proof.

LEMMA 6.3. *We have the relation*

$$[Z_\alpha, Z_\beta]f = \{\{L_\alpha, L_\beta\}, f\}, \quad (6.7)$$

where L_α and Z_α are obtained from (3.3) and (6.5) respectively by replacing ξ^i with ξ_α^i .

This can be proved by applying (6.6) to the left-hand side of (6.7) and using Jacobi's identities for the Poisson bracket.

Remark. According to this lemma, the Lie and the Poisson brackets correspond to each other if an operator and a special linear first integral are related by (6.6).

Now, it is possible to generalize Proposition 6.1 as follows.

PROPOSITION 6.4. *The operators Z_α satisfy the same commutation relations as X_α or Y_α , namely, $[Z_\alpha, Z_\beta] = C_{\alpha\beta}^\gamma Z_\gamma$.*

Proof. The Poisson bracket of L_α and L_β is calculated as

$$\{L_\alpha, L_\beta\} = \left(\xi_\alpha^j \frac{\partial \xi_\beta^i}{\partial x^j} - \xi_\beta^j \frac{\partial \xi_\alpha^i}{\partial x^j} \right) \varphi_i + \frac{\partial}{\partial x^k} \left(\xi_\alpha^j \frac{\partial \xi_\beta^i}{\partial x^j} - \xi_\beta^j \frac{\partial \xi_\alpha^i}{\partial x^j} \right) y^k \psi_i.$$

From (6.2) we have

$$\xi_\alpha^j \frac{\partial \xi_\beta^i}{\partial x^j} - \xi_\beta^j \frac{\partial \xi_\alpha^i}{\partial x^j} = C_{\alpha\beta}^\gamma \xi_\gamma^i,$$

and accordingly,

$$\{L_\alpha, L_\beta\} = C_{\alpha\beta}^\gamma \left(\xi_\gamma^i \varphi_i + \frac{\partial \xi_\gamma^i}{\partial x^k} y^k \psi_i \right) = C_{\alpha\beta}^\gamma L_\gamma.$$

It follows from (6.7) that

$$[Z_\alpha, Z_\beta]f = \{C_{\alpha\beta}^\gamma L_\gamma, f\} = C_{\alpha\beta}^\gamma Z_\gamma f.$$

As a corollary of this proposition we obtain the following theorem.

THEOREM 6.5. *Let a natural dynamical system admit a kinematical symmetry group G and let the associated scalar-preserving isometry group have the generators X_α . Then the corresponding special linear first integrals L_α generate a transformation group which is locally isomorphic to G .*

This theorem shows how important the special linear first integrals are for the symmetry problem in the maximum principle. On the other hand, one of the present authors (M.I.), together with Fujitani [7] and Nishino [8], studied natural dynamical systems which admit symmetry groups of higher orders.³ If we combine these results with (3.3), we can readily obtain the forms of special linear first integrals for the corresponding free or optimum systems. However, we omit the results.

7. Application to optimum control problem. In the remainder of the paper, we shall illustrate the application of our result, taking the problem of optimum trajectories for space navigation.

The object is to minimize the flying hours of a rocket in a central force field. This problem has been studied by many research workers [10], [11, pp. 103–146] and it is known, among other things, that there are first integrals corresponding to the angular momenta [1], [2], [12], [15]. Moyer [12] derived the integrals on the basis of Noether's theorem [13], and Burns [2] discussed their relation to the angular momenta in classical mechanics. It will be shown in the following that these integrals are easily found by a proper use of our result.

First of all, it is to be remembered that a flying rocket emits part of its mass as gas. Therefore, the mass m is not a constant, but a decreasing function of time t . This means that m should be regarded as one of the state variables. The kinetic energy of the rocket is no longer a quadratic form in the velocities, and the foregoing result cannot be applied directly.

In order to supply this gap we assume that the configuration space is a 3-dimensional Euclidean space, in which the motion of the rocket takes place. This is different from the treatment in the foregoing sections, where the Riemannian structure of the configuration space is associated with the kinetic energy.

Now, we make the following two assumptions for the sake of simplicity:

- (i) The emitted gas has a constant speed c relative to the rocket;
- (ii) The magnitude of the thrust is a function $k(t)$ of t such that $0 \leq k \leq k_0$ (k_0 being a positive constant) and that k is proportional to $-\dot{m}$.

³ Linear first integrals of the dynamical system were recently studied independently by Iliev [9]

We are now in a position to derive the equations of motion for the rocket. During the time interval Δt the momentum of the rocket changes by

$$f^i \Delta t = (m - \Delta m)(v^i + \Delta v^i) + \Delta m u^i - m v^i = m \Delta v^i - \Delta m (v^i - u^i),$$

where f^i , v^i and u^i are contravariant vectors representing the external force, the rocket velocity and the gas velocity respectively. Thus we have, in general curvilinear coordinates,

$$m \frac{\delta v^i}{\delta t} = f^i - \dot{m} c^i, \quad c^i = u^i - v^i,$$

where $\delta/\delta t$ denotes absolute differentiation and c^i means the relative velocity of the gas. From the assumption (i), $c = (c_i c^i)^{1/2}$ is a constant, and from (ii) we may put $\dot{m} = -k/c$ by a suitable choice of units. Thus the equations of motion are

$$(7.1) \quad \begin{aligned} \dot{x}^i &= y^i, \quad \dot{m} = -k/c, \\ \dot{y}^i &= \frac{1}{m}(f^i + k\theta^i) - \left\{ \begin{matrix} i \\ jk \end{matrix} \right\} y^j y^k, \end{aligned}$$

where we put $\theta^i = c^i/c$, the unit vector in the direction of c^i .

The corresponding Hamiltonian takes the form

$$(7.2) \quad H = y^i \varphi_i + \left[\frac{1}{m}(f^i + k\theta^i) - \left\{ \begin{matrix} i \\ jk \end{matrix} \right\} y^j y^k \right] \psi_i - \frac{k\lambda}{c},$$

where λ is the variable conjugate to m . The discussion made in § 4 can be followed starting with this Hamiltonian. We state only an outline of the result in what follows.

It is easily seen that the first equation of (4.3) holds as it is, and the Hamiltonian (7.2) is reduced to

$$(7.3) \quad H = y^i \varphi_i + \left(\frac{1}{m} f^i - \left\{ \begin{matrix} i \\ jk \end{matrix} \right\} y^j y^k \right) \psi_i + k \left(\frac{\psi}{m} - \frac{\lambda}{c} \right).$$

Since the mass m is added to the state variables, it is reasonable to consider a first integral of the form

$$(7.4) \quad L = \xi^i \varphi_i + \frac{\partial \xi^i}{\partial x^j} y^j \psi_i + \eta \lambda,$$

where η is a function of x^i , y^i and m . The study of (7.4) will be carried out according to the value of $\psi/m - \lambda/c$, which appears in the last term of (7.3).

In the case $\psi/m - \lambda/c \geq 0$, k must take the maximum value k_0 , and (7.4) is a first integral if and only if

$$\mathcal{L}_\xi \left\{ \begin{matrix} i \\ jk \end{matrix} \right\} = 0, \quad \mathcal{L}_\xi f^i = 0, \quad \mathcal{L}_\xi g^{ij} = 0, \quad \eta = 0,$$

where use is made of the equations $f_i = -m \nabla_i U$ corresponding to (5.1). The first integral (7.4) takes the same form as (3.3) and the condition for a first integral

coincides with that in Theorem 4.2. Therefore the result obtained in the foregoing sections is applicable.

As is well known, the central potential problem admits the angular momenta $M_{ij} = x^i y^j - x^j y^i$ as linear first integrals. The corresponding special linear first integrals can be easily obtained from (3.3), that is,

$$(7.5) \quad L_{ij} = x^i \varphi_j - x^j \varphi_i + y^i \psi_j - y^j \psi_i,$$

which were previously derived by tedious calculations [1], [2], [12], [15]. These integrals may be regarded as angular momenta within the framework of the maximum principle. They generate a transformation group locally isomorphic to the 3-dimensional rotation group.

In the case $\psi/m - \lambda/c \leq 0$, k must take the minimum value 0, and accordingly we have $m = \text{const}$. The system under consideration reduces to the free system discussed in § 3, and we have only to study a first integral of the form (3.3). According to Theorem 3.2 the condition for a first integral is given by (3.5) and (3.6), which admit solutions corresponding to the angular momenta L_{ij} .

In the case $\psi/m - \lambda/c \equiv 0$, the maximum principle is no longer applicable, but we mention it briefly for the sake of completeness. It can be shown that the condition for a first integral is

$$(7.6) \quad \begin{aligned} \mathcal{L}_\xi g^{ij} &= 0, & -\frac{1}{m} \mathcal{L}_\xi f^i + \frac{kc}{m^2} g^{ij} \frac{\partial \eta}{\partial y^j} &= 0, \\ \frac{\partial \eta}{\partial x^i} y^i + \frac{\partial \eta}{\partial y^i} \left(\frac{1}{m} f^i - \left\{ \begin{smallmatrix} i \\ jk \end{smallmatrix} \right\} y^j y^k \right) - \frac{k}{c} \frac{\partial \eta}{\partial m} + \frac{k}{mc} \eta &= 0. \end{aligned}$$

As is easily seen, the angular momenta L_{ij} are first integrals also in this case. However, since $\eta = m$ and $\xi^i = 0$ give a solution of (7.6), $L = m\lambda = c\psi$ is another first integral, which is not admitted by the other cases treated above.

From these discussions it can be concluded that *all cases admit a symmetry group, which is generated by the angular momenta L_{ij} and is locally isomorphic to the 3-dimensional rotation group.*

Acknowledgment. The authors wish to express their cordial thanks to Professor H. Fukawa who offered some valuable comments from the standpoint of optimum control theory. Thanks are also due to Mr. D. A. Forrester for carefully reading the manuscript and improving the English of the text.

REFERENCES

- [1] T. N. EDELBAUM AND S. PINES, *Fifth and sixth integrals for optimum rocket trajectories in a central force*, AIAA J., 8 (1970), pp. 1201–1204.
- [2] I. F. BURNS, *A parallel between Keplerian integrals and integrals of the adjoint equations*, Ibid., 8 (1970), pp. 809–810.
- [3] L. P. EISENHART, *Riemannian Geometry*, 2nd ed., Princeton Univ. Press, Princeton, 1950.
- [4] ———, *Continuous groups of transformations*, Princeton Univ. Press, Princeton, 1933.
- [5] M. IKEDA, S. OGAWA AND Y. OHNOKI, *A possible symmetry in Sakata's model for Bosons-baryons system*, Progr. Theoret. Phys., 22 (1959), pp. 715–724; 23 (1960), pp. 1073–1099.
- [6] M. IKEDA, *Symmetry of dynamical systems*, Butsuri, 25 (1970), pp. 9–13. (In Japanese).
- [7] M. IKEDA AND T. FUJITANI, *On linear first integrals of natural systems in classical mechanics*, Math. Japon., 15 (1971), pp. 143–153.

- [8] M. IKEDA AND Y. NISHINO, *On groups of scalar-preserving isometries in Riemannian spaces, with application to dynamical systems*, Tensor, 27 (1973), pp. 295–305.
- [9] I. ILIEV, *Classification of linear integrals of a holonomic mechanical system with n degrees of freedom*, J. Appl. Math. Mech., 36 (1972), pp. 112–116.
- [10] D. F. LAWDEN, *Optimal Trajectories for Space Navigation*, Butterworths, London, 1963.
- [11] A. I. LURIE, *Thrust Programming in a Central Gravitational Field*, G. Leitmann, ed., Academic Press, New York, 1967.
- [12] H. G. MOYER, *Integrals for impulsive orbit transfer from Noether's theorem*, AIAA J., 7 (1969), pp. 1232–1235.
- [13] E. NOETHER, *Invariante Variationsprobleme*, Nachr. Ges. Göttingen Math.-Phys. Kl. (1918), pp. 235–257.
- [14] L. A. PARS, *A Treatise on Analytical Dynamics*, Wiley, New York, 1968.
- [15] S. PINES, *Constants of the motion for optimum thrust trajectories in a central force field*, AIAA J., 2 (1964), pp. 2010–2014.
- [16] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Wiley, New York, 1962.
- [17] L. I. SCHIFF, *Quantum Mechanics*, 3rd ed., McGraw-Hill, New York, 1968.
- [18] J. L. SYNGE AND B. A. GRIFFITH, *Principles of Mechanics*, 3rd ed., McGraw-Hill, New York, 1959.
- [19] K. YANO, *The Theory of Lie Derivatives and its Applications*, North-Holland, Amsterdam, 1955.

PERIODIC LINEAR DIFFERENTIAL STOCHASTIC PROCESSES*

HUIBERT KWAKERNAAK†

Abstract. Periodic linear differential processes are defined and their properties are analyzed. Equivalent representations are discussed, and the solutions of related optimal estimation problems are given. An extension is presented of Kailath and Geesey's [1] results concerning the innovations representation of stochastic processes with a given covariance function.

1. Introduction. A periodic linear differential stochastic process X_t , $0 \leq t \leq T$, is defined as the solution of the stochastic differential equation

$$(1.1) \quad dX_t = AX_t dt + dW_t, \quad 0 \leq t \leq T,$$

with the periodicity condition

$$(1.2) \quad X_0 = X_T.$$

Here X_t is an n -dimensional real vector stochastic process, T a given real number, A a constant real $n \times n$ matrix, and W_t , $0 \leq t \leq T$, n -dimensional real Brownian motion with $E(dW_t dW_t') = V dt$, where the prime denotes the transpose and V is a given real symmetric nonnegative definite constant matrix. Processes of this kind can be used to model a variety of periodic random phenomena such as occur in communications and physics (t is not necessarily time but may also denote a space variable). In this paper the properties of the solution to (1.1)–(1.2) are studied, equivalent representations of the process X_t , $0 \leq t \leq T$, are given, and filtering, smoothing and prediction problems related to such processes are solved.

2. Properties of periodic differential processes. Solution of the differential equation (1.1) with the use of (1.2) yields the explicit representation

$$(2.1) \quad X_t = e^{At}(I - e^{AT})^{-1} \int_0^T e^{A(T-s)} dW_s + \int_0^t e^{A(t-s)} dW_s, \quad 0 \leq t \leq T,$$

where the integrals are stochastic integrals. The existence of this solution is guaranteed if A has no characteristic values that are an integral multiple of $2\pi i/T$, which henceforth will be assumed. It is easily verified that (2.1) constitutes the unique solution (in a mean square sense) to (1.1) and (1.2).

The expression (2.1) clearly defines a Gaussian process, which, therefore, is completely defined by its mean value function $E(X_t)$, $0 \leq t \leq T$, and its matrix covariance function $\text{cov}(X_t, X_s)$. Obviously,

$$(2.2) \quad E(X_t) = 0, \quad 0 \leq t \leq T.$$

To find the matrix covariance function, (2.1) is rewritten in the form

$$(2.3) \quad X_t = B_1 \int_0^t e^{A(t-s)} dW_s - B_2 \int_t^T e^{A(t-s)} dW_s, \quad 0 \leq t \leq T,$$

*Received by the editors May 15, 1973, and in revised form November 28, 1973.

†Department of Applied Mathematics, Twente University of Technology, Enschede, The Netherlands.

where

$$(2.4) \quad B_1 = (I - e^{AT})^{-1}, \quad B_2 = (I - e^{-AT})^{-1}.$$

Note that $B_1 + B_2 = I$. It follows from (2.3) by direct computation that

$$(2.5) \quad \begin{aligned} R(t, s) = E(X_t X_s') = & B_1 \left[\int_0^s e^{A(t-\theta)} V e^{A'(s-\theta)} d\theta \right] B_1' \\ & - B_1 \left[\int_s^t e^{A(t-\theta)} V e^{A'(s-\theta)} d\theta \right] B_2' \\ & + B_2 \left[\int_t^T e^{A(t-\theta)} V e^{A'(s-\theta)} d\theta \right] B_2' \quad \text{for } s \leq t. \end{aligned}$$

It is seen that the variance matrix is given by

$$(2.6) \quad R(t, t) = B_1 \left[\int_0^t e^{A(t-\theta)} V e^{A'(t-\theta)} d\theta \right] B_1' + B_2 \left[\int_t^T e^{A(t-\theta)} V e^{A'(t-\theta)} d\theta \right] B_2',$$

$$0 \leq t \leq T.$$

By some simple substitutions it follows that

$$(2.7) \quad R(t, t) = B_1 \left[\int_0^T e^{A\theta} V e^{A'\theta} d\theta \right] B_1', \quad 0 \leq t \leq T,$$

which proves that the variance matrix is a constant matrix, which henceforth will be denoted by Q . By premultiplying (2.7) by A , postmultiplying by A' , and integrating by parts it is easily found that Q satisfies the linear matrix equation

$$(2.8) \quad AQ + QA' + B_1 V + V B_1' - V = 0.$$

If $\lambda_j + \lambda_k \neq 0$ for each pair of characteristic values λ_j and λ_k , $j, k = 0, 1, \dots, n$, of A , Q is the unique solution of this matrix equation.

Equation (2.5) shows that for fixed s the matrix covariance function R is a continuous and differentiable function of t . Partial differentiation with respect to t yields

$$(2.9) \quad \frac{\partial R(t, s)}{\partial t} = AR(t, s) - V e^{A'(s-t)} B_2', \quad s \leq t.$$

Integration of this differential equation with the initial condition $R(s, s) = Q$ gives

$$(2.10) \quad \begin{aligned} R(t, s) &= e^{A(t-s)} Q - \left[\int_s^t e^{A(t-\theta)} V e^{A'(s-\theta)} d\theta \right] B_2' \\ &= e^{A(t-s)} Q - e^{A(t-s)} \left[\int_0^{t-s} e^{-A\tau} V e^{-A'\tau} d\tau \right] B_2', \quad s \leq t. \end{aligned}$$

Since $R(t, s) = R'(s, t)$, one immediately obtains

$$(2.11) \quad R(t, s) = Q e^{-A'(t-s)} - B_2 \left[\int_0^{t-s} e^{-A\tau} V e^{-A'\tau} d\tau \right] e^{-A'(t-s)}, \quad t \leq s.$$

THEOREM 2.1 (Covariance function). *Suppose that A has no characteristic values that are integer multiples of $2\pi i/T$. Then the solution of (1.1) and (1.2) is a stationary Gaussian process on $[0, T]$, with zero mean and matrix covariance function*

$$(2.12) \quad R(t, s) = \begin{cases} (I, 0) e^{F(t-s)} \begin{pmatrix} Q \\ B'_2 \end{pmatrix}, & s \leq t, \\ (Q, B_2) e^{-F'(t-s)} \begin{pmatrix} I \\ 0 \end{pmatrix}, & t \leq s, \end{cases}$$

where the variance matrix $Q = \text{var}(X_t)$ is given by (2.7), and where

$$(2.13) \quad F = \begin{pmatrix} A & -V \\ 0 & -A' \end{pmatrix}.$$

Proof. The relations (2.10) and (2.11) show that $R(t, s)$ is a function of $t - s$ alone. Therefore, the process X_t , $0 \leq t \leq T$, is wide-sense stationary, and, hence, strictly stationary. That R may be represented as in (2.12) follows from (2.10) and (2.11) together with the fact that

$$(2.14) \quad e^{F(t-s)} = \begin{pmatrix} e^{A(t-s)} & -e^{A(t-s)} \int_0^{t-s} e^{-A\tau} V e^{-A'\tau} d\tau \\ 0 & e^{-A'(t-s)} \end{pmatrix}. \quad \square$$

It is illuminating to consider the spectral representation of the process X_t , $0 \leq t \leq T$. Let $R(t, s) = \tilde{R}(t - s)$. Then according to Bochner's theorem there exists a matrix spectral distribution function Ψ such that

$$(2.15) \quad \tilde{R}(\theta) = \int_{-\infty}^{\infty} e^{i2\pi\gamma\theta} d\Psi(v), \quad 0 \leq \theta \leq T.$$

Here

$$(2.16) \quad \Psi(v) = \sum_{k=-\infty}^{\infty} \Psi_k H(v - v_k), \quad -\infty \leq v \leq \infty,$$

where H is the Heaviside step function, $v_k = k/T$, and Ψ_k is the coefficient matrix

$$(2.17) \quad \Psi_k = \frac{1}{T} \int_0^T \tilde{R}(\theta) e^{-i2\pi\gamma_k\theta} d\theta, \quad k = 0, \pm 1, \pm 2, \dots$$

It follows from (2.9) that

$$(2.18) \quad \frac{d\tilde{R}(\theta)}{d\theta} = A\tilde{R}(\theta) - V e^{-A'\theta} B'_2, \quad 0 \leq \theta \leq T.$$

Multiplication of both sides of (2.18) by $(1/T) \exp(-i2\pi\gamma_k\theta)$ and integration over $[0, T]$ yields

$$(2.19) \quad \Psi_k = \frac{1}{T} \Phi(i2\pi\gamma_k), \quad k \text{ integer},$$

where

$$(2.20) \quad \Phi(s) = (sI - A)^{-1}V(-sI - A')^{-1}, \quad s \text{ complex}.$$

This result is interesting, for the following reason. Suppose that A is stable, i.e., all its characteristic values have strictly negative real parts. Then the matrix equation

$$(2.21) \quad A\bar{Q} + \bar{Q}A' + V = 0$$

has a unique symmetric nonnegative definite solution \bar{Q} . Furthermore, the stochastic differential equation

$$(2.22) \quad d\bar{X}_t = A\bar{X}_t + dW_t, \quad t \geq 0,$$

with $W_t, t \geq 0$, Brownian motion with $E(dW_t dW_t') = V dt$, and \bar{X}_0 a Gaussian stochastic variable, independent of $W_t, t \geq 0$, with expectation zero and variance matrix \bar{Q} , generates a stationary Gaussian stochastic process $\bar{X}_t, t \geq 0$, with spectral density matrix $\Phi(i2\pi\nu)$. The process $\bar{X}_t, t \geq 0$, may be called the *long term version* of the solution of the stochastic differential equation (1.1), and the solution of (1.1) and (1.2) the *periodic version* of this solution. Thus, the spectral coefficients of the periodic version are directly related to the spectral density matrix of the long term version by (2.19). It furthermore follows from (2.19) that the covariance matrix function \tilde{R} of the periodic version and the covariance matrix function $\bar{R}(t-s) = E(\bar{X}_t \bar{X}_s')$ of the long term version of the process are related by

$$(2.23) \quad \tilde{R}(\theta) = \sum_{n=-\infty}^{\infty} \bar{R}(\theta + nT), \quad 0 \leq \theta \leq T.$$

This suggests that the periodic and long term versions of the process are related by

$$X_t = \text{l.i.m.}_{N \rightarrow \infty} \frac{1}{\sqrt{N+1}} \sum_{n=0}^{\infty} \bar{X}_{t+nT}, \quad 0 \leq t \leq T,$$

which may be verified. These connections between the periodic and long term versions of the process of course make no sense if A is not stable, because in this case the long term version of the process is not defined.

3. Equivalent representations. This section is addressed to the question whether there exists another process $X_t^*, 0 \leq t \leq T$, defined by

$$(3.1) \quad \begin{aligned} dX_t^* &= A^* X_t^* dt + dW_t^*, \quad 0 \leq t \leq T, \\ X_0^* &= X_T^*, \end{aligned}$$

with $E(dW_t^* dW_t^{*'}) = V^* dt$, such that the processes $X_t^*, 0 \leq t \leq T$, and $X_t, 0 \leq t \leq T$, have the same matrix covariance functions. Let

$$(3.2) \quad \Phi^*(s) = (sI - A^*)^{-1}V^*(-sI - A^{*'})^{-1}.$$

Then, since $\Phi(s)$ and $\Phi^*(s)$ agree for all $s = i2\pi k/T, k$ integer,

$$(3.3) \quad \Phi(s) = \Phi^*(s), \quad \text{all complex } s.$$

Assume now that $V = BB'$, and that the pair (A, B) is completely controllable. For the system theoretic terminology employed in this section, reference is made, for example, to Kalman, Falb and Arbib [2]. Then if $\Phi(s)$ is considered as the transfer matrix of a time-invariant linear differential system, (F, G, H) forms a realization of this system, i.e., $\Phi(s) = H(sI - F)^{-1}G$, where

$$(3.4) \quad F = \begin{pmatrix} A & -V \\ 0 & -A' \end{pmatrix}, \quad G = \begin{pmatrix} 0 \\ I \end{pmatrix}, \quad H = (I, 0).$$

From the fact that (A, B) is completely controllable it may be shown by straightforward verification that (F, G) is completely controllable and (F, H) completely observable, so that (F, G, H) is a minimal realization of $\Phi(s)$, i.e., the matrices F , G and H have the smallest possible dimensions. Let

$$(3.5) \quad F^* = \begin{pmatrix} A^* & -V^* \\ 0 & -A^{*'} \end{pmatrix}.$$

Since $\Phi(s) \equiv \Phi^*(s)$, (F^*, G, H) is another realization of $\Phi(s)$. As this realization has the same dimension as (F, G, H) , it is also minimal. Therefore, there exists [2] a nonsingular $2n \times 2n$ transformation matrix M such that $M^{-1}FM = F^*$, $M^{-1}G = G$, and $HM = H$. It is easily found that the latter two equalities are satisfied if and only if

$$(3.6) \quad M = \begin{pmatrix} I & 0 \\ P & 0 \end{pmatrix}, \quad M^{-1} = \begin{pmatrix} I & 0 \\ -P & I \end{pmatrix},$$

where P is an arbitrary $n \times n$ matrix. With this it follows from $M^{-1}FM = F^*$ that

$$(3.7) \quad \begin{pmatrix} A - VP & -V \\ -A'P - PA + PVP & -A' + PV \end{pmatrix} = \begin{pmatrix} A^* & -V^* \\ 0 & -A^{*'} \end{pmatrix},$$

which leads to the following result.

THEOREM 3.1 (Equivalent representations). *Suppose that $V = BB'$, and that (A, B) is completely controllable. Then if the process X_t^* , $0 \leq t \leq T$, has the same matrix covariance function as X_t , $0 \leq t \leq T$,*

$$(3.8) \quad \begin{aligned} A^* &= A - VP, \\ V^* &= V, \end{aligned}$$

where P is a solution of the nonlinear matrix equation

$$(3.9) \quad 0 = A'P + PA - PVP.$$

Equation (3.9) is a special form of the algebraic matrix Riccati equation, about which a great deal is known (see e.g., [3]).

4. Innovations representation and estimation of stochastic processes with a given covariance function. In this section an extension is given of Kailath and Geesey's results [1] concerning the representation and estimation of stochastic processes with given covariance functions. The results given here differ from those of Kailath and Geesey in that vector-valued processes are considered,

nonzero means are accounted for, the existence of the solution of an essential matrix differential equation is proved, and expressions are given for the variance matrices of the reconstruction errors. The results of this section will be applied to periodic differential processes in § 5.

Consider a vector-valued, mean square continuous stochastic process Z_t , $t \in \Theta$, where $\Theta = [t_0, t_1]$, with the mean value function

$$(4.1) \quad E(Z_t) = \int_{t_0}^t m_Z(s) ds, \quad t \in \Theta,$$

and the matrix covariance function

$$(4.2) \quad \text{cov}(Z_t, Z_s) = \int_{t_0}^t dt' \int_{t_0}^s L_Z(t', s') ds' + \int_{t_0}^{\min(t, s)} V(t') dt', \quad t, s \in \Theta.$$

Here $V(t)$, $t \in \Theta$, is a given, continuous, symmetric matrix function, positive definite for every $t \in \Theta$. The function m_Z is given in the form

$$(4.3) \quad m_Z(t) = M(t)\Psi(t, t_0)m_0, \quad t \in \Theta,$$

while L_Z is given in the form

$$(4.4) \quad L_Z(t, s) = \begin{cases} M(t)\Psi(t, s)N(s), & s \leq t, \\ N'(t)\Psi'(s, t)M'(s), & t \leq s. \end{cases}$$

Here M and N are continuous matrix functions on Θ , m_0 is a constant vector, and Ψ is a fundamental matrix satisfying the matrix differential equation

$$(4.5) \quad \begin{aligned} \frac{\partial}{\partial t} \Psi(t, s) &= F(t)\Psi(t, s), & t, s \in \Theta, \\ \Psi(s, s) &= I, & s \in \Theta, \end{aligned}$$

with F a continuous matrix function on Θ . It will finally be assumed that the self-adjoint operator Q defined by

$$(4.6) \quad (Qv)(t) = V(t)v(t) + \int_{t_0}^{t_1} L_Z(t, s)v(s) ds, \quad t_0 \leq t \leq t_1,$$

where $v \in L_2[t_0, t_1]$, is positive definite.

The following result (Kailath and Geesey [1]) is essential.

THEOREM 4.1 (Innovations representation). *Suppose that the matrix differential equation*

$$(4.7) \quad \begin{aligned} \dot{S}(t) &= F(t)S(t) + S(t)F'(t) + K(t)V(t)K'(t), & t \in \Theta, \\ S(t_0) &= 0, \end{aligned}$$

where

$$(4.8) \quad K(t) = [N(t) - S(t)M'(t)]V^{-1}(t), \quad t \in \Theta,$$

has a solution on Θ . Then the process Z_t^* , $t \in \Theta$, defined by

$$(4.9) \quad \begin{aligned} dZ_t^* &= M(t)P_t dt + dI_t, \quad t \in \Theta, \\ Z_{t_0} &= 0, \end{aligned}$$

$$(4.10) \quad \begin{aligned} dP_t &= F(t)P_t dt + K(t) dI_t, \quad t \in \Theta, \\ P_{t_0} &= m_0, \end{aligned}$$

where I_t , $t \in \Theta$, is Brownian motion with $E(dI_t dI_t') = V(t) dt$, has the mean value function defined by (4.1) and (4.3), and the matrix covariance function defined by (4.2) and (4.4).

Proof. This theorem may be proved by direct calculation of the mean value function and matrix covariance function of the process Z_t^* , $t \in \Theta$, as defined by (4.7)–(4.10). This calculation is laborious but straightforward. It helps to note that $S(t) = \text{var}(P_t)$, $t \in \Theta$.

THEOREM 4.2 (Existence of the solution of the matrix differential equation). *The matrix differential equation (4.7)–(4.8) has a unique solution on every finite interval $[t_0, t_1]$.*

Proof. The proof of Kailath and Geesey of the existence of the solution of the matrix differential equation (4.7)–(4.8) is based on the assumption of the existence of some lumped Markov model for the process Z . This assumption is not required in the proof to follow. Define the matrix functions

$$(4.11) \quad \begin{aligned} S^*(t) &= S(t_0 + t_1 - t), \quad F^*(t) = F(t_0 + t_1 - t), \quad N^*(t) = N(t_0 + t_1 - t), \\ M^*(t) &= M(t_0 + t_1 - t), \quad V^*(t) = V(t_0 + t_1 - t), \end{aligned}$$

for all $t \in [t_0, t_1]$. Then it easily follows that (4.7)–(4.8) can be rewritten in the form

$$(4.12) \quad \begin{aligned} -\dot{S}^*(t) &= [F^{*'}(t) - M^{*'}(t)V^{*-1}(t)N^{*'}(t)]'S^*(t) \\ &\quad + S^*(t)[F^{*'}(t) - M^{*'}(t)V^{*-1}(t)N^{*'}(t)] \\ &\quad + S^*(t)M^{*'}(t)V^{*-1}(t)M^*(t)S^*(t) + N^*(t)V^{*-1}(t)N^{*'}(t), \end{aligned}$$

$$t_0 \leq t \leq t_1,$$

$$S^*(t_1) = 0.$$

This matrix differential equation, and hence (4.7)–(4.8), has a unique solution if and only if the following optimal control problem has a unique solution [4]. Maximize

$$(4.13) \quad \int_{t_0}^{t_1} [x'(t)N^*(t)V^{*-1}(t)N^{*'}(t)x(t) - u'(t)V^*(t)u(t)] dt$$

with respect to $u(t)$ and $x(t)$, $t_0 \leq t \leq t_1$ subject to

$$(4.14) \quad \begin{aligned} \dot{x}(t) &= [F^{*'}(t) - M^{*'}(t)V^{*-1}(t)N^{*'}(t)]x(t) + M^{*'}(t)u(t), \quad t_0 \leq t \leq t_1, \\ x(t_0) &= x_0, \end{aligned}$$

with x_0 a given vector. By substituting $u(t) - V^{*-1}(t)N^{*'}(t)x(t) = u^*(t)$, $t_0 \leq t \leq t_1$, it follows that this problem is equivalent to the problem of minimizing

$$(4.15) \quad J = \int_{t_0}^{t_1} [u^{*'}(t)V^*(t)u^*(t) + 2u^{*'}(t)N^{*'}(t)x(t)] dt$$

with respect to $u^*(t)$ and $x(t)$, $t_0 \leq t \leq t_1$, subject to

$$(4.16) \quad \begin{aligned} \dot{x}(t) &= F^{*'}(t)x(t) + M^{*'}(t)u^*(t), \quad t_0 \leq t \leq t_1, \\ x(t_0) &= x_0. \end{aligned}$$

Let $\Psi^*(t, s) = \Psi'(t_0 + t_1 - s, t_0 + t_1 - t)$ denote the fundamental matrix corresponding to $\dot{x}(t) = F^{*'}(t)x(t)$. Then substitution of

$$(4.17) \quad x(t) = \Psi^*(t, t_0)x_0 + \int_{t_0}^t \Psi^*(t, s)M^{*'}(s)u^*(s) ds, \quad t_0 \leq t \leq t_1,$$

into (4.15) yields, with changes of integration variables from s to $t_0 + t_1 - s$ and t to $t_0 + t_1 - t$, and with the notation $u^*(t_0 + t_1 - t) = v(t)$, $t_0 \leq t \leq t_1$,

$$(4.18) \quad \begin{aligned} J &= \int_{t_0}^{t_1} v'(t)V(t)v(t) dt + 2 \int_{t_0}^{t_1} v'(t)N'(t)\Psi'(t_1, t)x_0 dt \\ &\quad + 2 \int_{t_0}^{t_1} v'(t)N'(t) dt \int_t^{t_1} \Psi'(s, t)M'(s)v(s) ds \\ &= \int_{t_0}^{t_1} v'(t)V(t)v(t) dt + 2 \int_{t_0}^{t_1} v'(t)N'(t)\Psi'(t_1, t)x_0 dt \\ &\quad + \int_{t_0}^{t_1} dt \int_t^{t_1} v'(t)N'(t)\Psi'(s, t)M'(s)v(s) ds \\ &\quad + \int_{t_0}^{t_1} ds \int_{t_0}^s v'(s)M(s)\Psi(s, t)N(t)v(t) dt \\ &= \int_{t_0}^{t_1} v'(t)V(t)v(t) dt + 2 \int_{t_0}^{t_1} v'(t)N'(t)\Psi'(t_1, t)x_0 dt \\ &\quad + \int_{t_0}^{t_1} \int_{t_0}^{t_1} v'(t)L_Z(t, s)v(s) dt ds. \end{aligned}$$

In functional analytic form this may be rewritten as

$$(4.19) \quad J = \langle v, Qv \rangle + 2\langle v, f \rangle,$$

where Q is the operator defined by (4.6), f is the function $f(t) = N'(t)\Psi'(t_1, t)x_0$, $t_0 \leq t \leq t_1$ and $\langle \cdot, \cdot \rangle$ denotes the inner product in $L_2[t_0, t_1]$. Since by assumption Q is positive definite, there exists a unique $v \in L_2[t_0, t_1]$ that minimizes J . Hence, the matrix equation (4.7)–(4.8) has a unique solution on any interval $[t_0, t_1]$. \square

Once the innovations representation (4.9)–(4.10) has been obtained, it is easy to obtain the solution to various filtering problems (Kailath and Geesey [1]).

Let R_t , $t \in \Theta$, be another vector-valued, mean square continuous process, defined on the same probability space as the process Z_t , $t \in \Theta$, with the mean value function

$$(4.20) \quad E(R_t) = M_r(t)\Psi(t, t_0)m_0, \quad t \in \Theta.$$

Suppose that the matrix cross covariance function of the processes R and Z is given in the form

$$(4.21) \quad \text{cov}(R_t, Z_s) = \int_{t_0}^s L_{RZ}(t, s') ds', \quad t, s \in \Theta,$$

where

$$(4.22) \quad L_{RZ}(t, s) = \begin{cases} M_r(t)\Psi(t, s)N(s), & s \leq t, \\ N'_r(t)\Psi'(s, t)M'(s), & t \leq s. \end{cases}$$

Here N_r and M_r are given, continuous matrix functions on Θ .

THEOREM 4.3 (Filtering problem). *Let Z_t , $t \in \Theta$, and R_t , $t \in \Theta$, be mean square continuous, vector-valued stochastic processes, with mean value functions and (cross) covariance matrix functions defined by (4.1)–(4.4) and (4.20)–(4.22), respectively. Let $R_{t|t}$ denote the minimum variance unbiased linear estimator of R_t given Z_θ , $t_0 \leq \theta \leq t$. Then*

$$(4.23) \quad R_{t|t} = M_r(t)P_t, \quad t \in \Theta,$$

where P_t , $t \in \Theta$, is the solution of the stochastic differential equation

$$(4.24) \quad \begin{aligned} dP_t &= F(t)P_t dt + K(t)[dZ_t - M(t)P_t dt], \quad t \in \Theta, \\ P_{t_0} &= m_0. \end{aligned}$$

The variance matrix of the estimation error is given by

$$(4.25) \quad \text{var}(R_{t|t} - R_t) = \text{var}(R_t) - M_r(t)S(t)M'_r(t), \quad t \in \Theta.$$

Proof. $R_{t|s}$ is the minimum variance unbiased linear estimator of Z_t given Z_θ , $t_0 \leq \theta \leq s$, if and only if

$$(4.26) \quad E(R_{t|s} - R_t) = 0,$$

and

$$(4.27) \quad \text{cov}(R_{t|s} - R_t, Z_\theta) = 0, \quad t_0 \leq \theta \leq s.$$

It may be verified by straightforward but tedious calculations that the estimator $R_{t|t}$ as given by (4.23) satisfies these conditions for $s = t$. The variance matrix of the estimation error as given by (4.25) follows from the fact that

$$(4.28) \quad \text{var}(R_{t|s} - R_t) = \text{var}(R_t) - \text{var}(R_{t|s}). \quad \square$$

THEOREM 4.4 (Prediction problem). *Let the processes Z_t , $t \in \Theta$, and R_t , $t \in \Theta$, be defined as before. Let $R_{t|s}$, $t \geq s$, denote the minimum variance unbiased linear estimator of R_t , given Z_θ , $t_0 \leq \theta \leq s$. Then*

$$(4.29) \quad R_{t|s} = M_r(t)\Psi(t, s)P_s, \quad t \geq s,$$

where P_t , $t \in \Theta$, is the solution of (4.24). The prediction error variance matrix is given by

$$(4.30) \quad \text{var}(R_{t|s} - R_t) = \text{var}(R_t) - M_r(t)\Psi(t, s)S(s)\Psi'(t, s)M_r'(t), \quad t \geq s.$$

Proof. Again it may be verified by direct computation that the estimator as given by (4.29) satisfies the conditions (4.26) and (4.27). The variance matrix (4.30) follows from (4.28). \square

THEOREM 4.5 (Smoothing problem). *Let the processes Z_t , $t \in \Theta$, and R_t , $t \in \Theta$, be defined as before. Let $R_{t|s}$, $t \leq s$, denote the minimum variance unbiased linear estimator of R_t , given Z_θ , $t_0 \leq \theta \leq s$. Then*

$$(4.31) \quad R_{t|s} = R_{t|t} + [N_r'(t) - M_r(t)S(t)]\Lambda_{t|s}, \quad t \leq s,$$

where $\Lambda_{t|s}$, $t_0 \leq t \leq s$, can be solved from the stochastic differential equation

$$(4.32) \quad \begin{aligned} d\Lambda_{t|s} &= -[F(t) - K(t)M(t)]'\Lambda_{t|s}dt - M'(t)V_2^{-1}(t)[dZ_t - M(t)P_t dt], \quad t \leq s, \\ \Lambda_{s|s} &= 0. \end{aligned}$$

The smoothing error variance matrix is given by

$$(4.33) \quad \begin{aligned} \text{var}(R_{t|s} - R_t) &= \text{var}(R_t) - M_r(t)S(t)M_r'(t) \\ &\quad - [N_r'(t) - M_r(t)S(t)] \left[\int_t^s \Xi'(\theta, s)M'(\theta)V_2^{-1}(\theta)M(\theta)\Xi(\theta, s)d\theta \right] \\ &\quad [N_r(t) - S(t)M_r'(t)], \quad t \leq s, \end{aligned}$$

where the fundamental matrix Ξ is the solution of

$$(4.34) \quad \begin{aligned} \frac{\partial}{\partial t}\Xi(t, s) &= [F(t) - K(t)M(t)]\Xi(t, s), \quad t, s \in \Theta, \\ \Xi(s, s) &= I, \quad s \in \Theta. \end{aligned}$$

Proof. Again it may be verified by direct calculation that (4.31) satisfies (4.26) and (4.27). The variance matrix (4.33) follows from (4.28). \square

5. Filtering, prediction and smoothing for periodic linear differential processes.

In this section the results of the preceding section are applied to the periodic processes of §§ 1 and 2. Consider the periodic process defined by

$$(5.1) \quad \begin{aligned} dX_t &= AX_t dt + dW_{1,t}, \quad 0 \leq t \leq T, \\ X_0 &= X_T, \end{aligned}$$

with $E(dW_{1,t}dW_{1,t}') = V_1 dt$. Suppose furthermore that a process Z_t , $0 \leq t \leq T$, is observed, which is given by

$$(5.2) \quad \begin{aligned} dZ_t &= CX_t dt + dW_{2,t}, \quad 0 \leq t \leq T, \\ Z_0 &= 0, \end{aligned}$$

where C is a constant matrix, and $W_{2,t}$, $0 \leq t \leq T$, Brownian motion, independent of $W_{1,t}$, with $E(dW_{2,t}dW_{2,t}') = V_2 dt$. Using the results of § 2, it is easily seen that the mean value function of the process Z_t , $0 \leq t \leq T$, is identical to zero, while

its covariance matrix function may be expressed in the form

$$(5.3) \quad \text{cov}(Z_t, Z_s) = \int_0^t dt' \int_0^s L_Z(t', s') ds' + \int_0^{\min(t,s)} V_2 dt', \quad 0 \leq t, s \leq T,$$

with

$$(5.4) \quad L_Z(t, s) = \begin{cases} M\Psi(t, s)N, & s \leq t, \\ N'\Psi'(s, t)M', & t \leq s, \end{cases}$$

where

$$(5.5) \quad M = (C, 0), \quad \Psi(t, s) = e^{F(t-s)}, \quad N = \begin{pmatrix} QC' \\ B_2' C' \end{pmatrix}.$$

Here

$$(5.6) \quad F = \begin{pmatrix} A & -V_1 \\ 0 & -A' \end{pmatrix},$$

while Q is the variance matrix of the process X_t . Furthermore, it is easily established that the cross covariance matrix of the processes X_t and Z_t is given by

$$(5.7) \quad \text{cov}(X_t, Z_s) = \int_0^s L_{XZ}(t, s') ds', \quad 0 \leq t, s \leq T,$$

where

$$(5.8) \quad L_{XZ}(t, s) = \begin{cases} M_x \Psi(t, s)N, & s \leq t, \\ N'_x \Psi'(s, t)M', & t \leq s, \end{cases}$$

such that

$$(5.9) \quad M_x = (I, 0), \quad N_x = \begin{pmatrix} Q \\ B_2' \end{pmatrix}.$$

The results (5.3)–(5.4) and (5.7)–(5.8) conform to the assumptions of § 4. The matrix differential equation (4.7), which plays a central role, in the present case takes the form

$$(5.10) \quad \begin{aligned} \dot{S}(t) &= \begin{pmatrix} A & -V_1 \\ 0 & -A' \end{pmatrix} S(t) + S(t) \begin{pmatrix} A' & 0 \\ -V_1 & -A \end{pmatrix} + K(t) V_2 K'(t), \quad 0 \leq t \leq T, \\ S(0) &= 0, \end{aligned}$$

where

$$(5.11) \quad K(t) = \left[\begin{pmatrix} QC' \\ B_2' C' \end{pmatrix} - S(t) \begin{pmatrix} C' \\ 0 \end{pmatrix} \right] V_2^{-1}, \quad 0 \leq t \leq T.$$

It is convenient to make the substitution

$$(5.12) \quad U(t) = \begin{pmatrix} Q & B_2' \\ B_2' & 0 \end{pmatrix} - S(t), \quad 0 \leq t \leq T.$$

This results in the matrix differential equation

$$(5.13) \quad \begin{aligned} \dot{U}(t) &= FU(t) + U(t)F' - U(t)H'V_2^{-1}HU(t) + GV_1G', \quad 0 \leq t \leq T, \\ U(0) &= \begin{pmatrix} Q & B_2 \\ B_2' & 0 \end{pmatrix}, \end{aligned}$$

while

$$(5.14) \quad K(t) = U(t)H'V_2^{-1}, \quad 0 \leq t \leq T.$$

Here, with a change from an earlier notation,

$$(5.15) \quad F = \begin{pmatrix} A & -V_1 \\ 0 & -A' \end{pmatrix}, \quad G = \begin{pmatrix} I \\ 0 \end{pmatrix}, \quad H = (C, 0).$$

The existence and uniqueness of the solutions of (5.10)–(5.11) and (5.13) are guaranteed by Theorem 4.2.

The solutions of the filtering, prediction and smoothing problems for the processes described by (5.1)–(5.2) are now easily solved using the results of § 4. It is advantageous to express these solutions in terms of the solution of the matrix differential equation (5.13).

THEOREM 5.1 (Filtering problem). *The minimum variance unbiased estimator $X_{t|t}$ of X_t , given Z_θ , $0 \leq \theta \leq t$, is given by*

$$(5.16) \quad X_{t|t} = (I, 0)P_t, \quad 0 \leq t \leq T,$$

where the $2n$ -dimensional process P_t , $0 \leq t \leq T$, is the solution of

$$(5.17) \quad \begin{aligned} dP_t &= \begin{pmatrix} A & -V_1 \\ 0 & -A' \end{pmatrix} P_t dt + K(t)[dZ_t - (C, 0)P_t dt], \quad 0 \leq t \leq T, \\ P_0 &= 0. \end{aligned}$$

The filtering error variance matrix is given by

$$(5.18) \quad \text{var}(X_{t|t} - X_t) = U_{11}(t), \quad 0 \leq t \leq T,$$

where U_{11} is obtained by partitioning $U(t)$ into four $n \times n$ matrices as

$$(5.19) \quad U(t) = \begin{pmatrix} U_{11}(t) & U_{12}(t) \\ U'_{12}(t) & U_{22}(t) \end{pmatrix}, \quad 0 \leq t \leq T.$$

THEOREM 5.2 (Prediction problem). *The minimum variance unbiased estimator $X_{t|s}$ of X_t , given Z_θ , $0 \leq \theta \leq s$, is given by*

$$(5.20) \quad X_{t|s} = (I, 0) e^{F(t-s)} P_s, \quad 0 \leq s \leq t \leq T,$$

with F given by (5.15) and where P_t , $0 \leq t \leq T$, is obtained from (5.17). The prediction error variance matrix is

$$(5.21) \quad \begin{aligned} \text{var}(X_{t|s} - X_t) &= (I, 0) \left[e^{F(t-s)} U(s) e^{F'(t-s)} \right. \\ &\quad \left. + \int_s^t e^{F(t-\theta)} G V_1 G' e^{F'(t-\theta)} d\theta \right] \begin{pmatrix} I \\ 0 \end{pmatrix}, \quad 0 \leq s \leq t \leq T. \end{aligned}$$

THEOREM 5.3 (Smoothing problem). *The minimum variance unbiased estimator $X_{t|s}$ of X_t , given Z_θ , $0 \leq \theta \leq s$, is given by*

$$(5.22) \quad X_{t|s} = X_{t|t} + (I, 0)U(t)\Lambda_{t|s}, \quad 0 \leq t \leq s \leq T,$$

where $\Lambda_{t|s}$ is solved from

$$(5.23) \quad \begin{aligned} d\Lambda_{t|s} &= -[F - K(t)H]' \Lambda_{t|s} dt + H' V_2^{-1} [dZ_t - H P_t dt], \quad 0 \leq t \leq s, \\ \Lambda_{s|s} &= 0. \end{aligned}$$

The smoothing error variance matrix may be expressed as

$$(5.24) \quad \begin{aligned} \text{var}(X_{t|s} - X_t) &= U_{11}(t) - (I, 0)U(t) \left[\int_t^s \Xi'(\theta, s) H' V_2^{-1} H \Xi(\theta, s) d\theta \right] U(t) \begin{pmatrix} I \\ 0 \end{pmatrix}, \\ &0 \leq t \leq s \leq T, \end{aligned}$$

where Ξ is the fundamental matrix satisfying

$$(5.25) \quad \begin{aligned} \frac{\partial}{\partial t} \Xi(t, s) &= [F - K(t)H] \Xi(t, s), \quad 0 \leq t, s \leq T, \\ \Xi(s, s) &= I, \quad 0 \leq s \leq T. \end{aligned}$$

It is noted that the filtering, prediction and smoothing solutions as given would have been obtained if the problem had been considered of estimating the first n components of a $2n$ -dimensional process D_t , $0 \leq t \leq T$, which is the solution of

$$(5.26) \quad dD_t = F D_t dt + d\Sigma_{1,t}, \quad 0 \leq t \leq T,$$

with $\Sigma_{1,t}$, $0 \leq t \leq T$, Brownian motion with

$$(5.27) \quad E(d\Sigma_{1,t} d\Sigma'_{1,t}) = \begin{pmatrix} V_1 & 0 \\ 0 & 0 \end{pmatrix} dt,$$

together with the observation equation

$$(5.28) \quad dZ_t = (C, 0)D_t dt + d\Sigma_{2,t}, \quad 0 \leq t \leq T, \quad Z_0 = 0,$$

where $\Sigma_{2,t}$, $0 \leq t \leq T$, is Brownian motion independent of Σ_1 , with

$$E(d\Sigma_{2,t} d\Sigma'_{2,t}) = V_2 dt,$$

and where, finally, D_0 is a stochastic vector, independent of the Brownian motion processes, with expectation zero and variance matrix

$$(5.29) \quad \text{var}(D_0) = \begin{pmatrix} Q & B_2 \\ B_2' & 0 \end{pmatrix}.$$

Such a process D_t , $0 \leq t \leq T$, does not exist, however, since the right-hand side of (5.29) is an indefinite matrix, and hence is not a variance matrix.

A simplification of the solution of the smoothing problem is obtained for $s = T$.

THEOREM 5.4 (Simplification of the smoothing solution). *The minimum variance unbiased estimator $X_{t|T}$ of X_t , given Z_θ , $0 \leq \theta \leq T$, may be solved from*

$$(5.30) \quad dX_{t|T} = AX_{t|T} dt + V_1 \Gamma_t dt, \quad 0 \leq t \leq T,$$

where Γ_t , $0 \leq t \leq T$, satisfies

$$(5.31) \quad d\Gamma_t = -A'\Gamma_t dt - C'V_2^{-1}(dZ_t - CX_{t|T} dt), \quad 0 \leq t \leq T.$$

The boundary conditions are

$$(5.32) \quad X_{0|T} = X_{T|T}, \quad \Gamma_0 = \Gamma_T.$$

The reconstruction error variance matrix $\text{var}(X_{t|T} - X_t)$ is constant on $[0, T]$.

Proof. Equations (5.30)–(5.32) may be proved by differentiating $X_{t|T}$ as derived from (5.22). That $\text{var}(X_{t|T} - X_t)$ is a constant matrix follows by recognizing that $E_t = X_t - X_{t|T}$ and Γ_t satisfy

$$(5.33) \quad \begin{pmatrix} dE_t \\ d\Gamma_t \end{pmatrix} = \begin{pmatrix} A & -V_1 \\ -C'V_2^{-1}C & -A' \end{pmatrix} \begin{pmatrix} E_t \\ \Gamma_t \end{pmatrix} dt + \begin{pmatrix} dW_{1,t} \\ -C'V_2^{-1}dW_{2,t} \end{pmatrix}, \quad 0 \leq t \leq T,$$

$$\begin{pmatrix} E_0 \\ \Gamma_0 \end{pmatrix} = \begin{pmatrix} E_T \\ \Gamma_T \end{pmatrix}.$$

Equation (5.33) defines a periodic differential process, which proves that $\text{var}(E_t, \Gamma_t)$ and hence also $\text{var}(E_t) = \text{var}(X_{t|T} - X_t)$ is constant on $[0, T]$. \square

6. Conclusions. Periodic differential stochastic processes turn out to have several interesting properties. Kailath and Geesey's results provide a convenient method to solve filtering problems for such processes.

REFERENCES

- [1] T. KAILATH AND R. A. GEESEY, *An innovations approach to least squares estimation—Part IV: Recursive estimation given lumped covariance functions*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 720–727.
- [2] R. E. KALMAN, P. L. FALB AND M. A. ARBIB, *Topics in Mathematical Systems Theory*, McGraw-Hill, New York, 1969.
- [3] H. KWAKERNAAK AND R. SIVAN, *Linear Optimal Control Systems*, Wiley-Interscience, New York, 1972.
- [4] R. E. KALMAN, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana, 5 (1960), pp. 102–119.

ON THE NUMBER OF DIRECTIONS NEEDED TO ACHIEVE CONTROLLABILITY*

HÉCTOR J. SUSSMANN†

Abstract. A set S of vector fields on a manifold M is said to be *controllable* if for every pair (m, m') of points of M there is a trajectory of S from m to m' . Going backwards in time is not allowed. If every point has a fundamental system of neighborhoods in which S is controllable, then S is said to be *locally controllable*. We show that on every manifold there exists a locally controllable system of *three* vector fields, and that no system of *two* vector fields can be locally controllable (if $\dim M \geq 2$). It follows that on every connected manifold there exists a controllable system of three vector fields. For some manifolds (such as the special orthogonal groups, spheres and Euclidean spaces) we show how to construct a controllable system of *two* vector fields.

In this note we shall discuss the following question: given a connected C^∞ -manifold M , how many vector fields are needed to define a system which is *controllable* on M ? The precise definitions are given below. However, we emphasize that in this paper a system is said to be *controllable* if, for every point $m \in M$, the set of points reachable from m by trajectories of the system *that go forward in time* is M itself. As an illustration, the vector fields

$$\frac{\partial}{\partial x}, \frac{\partial}{\partial y}$$

in the x, y -plane do *not* define a controllable system. (For instance, the set reachable from the origin is the positive quadrant $\{(x, y): x \geq 0 \text{ and } y \geq 0\}$.) To get a controllable system it is sufficient to add a third vector field. Indeed, the system

$$\left\{ \frac{\partial}{\partial x}, \frac{\partial}{\partial y}, -\left(\frac{\partial}{\partial x} + \frac{\partial}{\partial y} \right) \right\}$$

is obviously controllable.

As we shall see, the situation of the preceding example is typical. We shall show that on *every* connected manifold there exists a controllable system that consists of *three* vector fields. Also, we show how to construct controllable systems of *two* vector fields in some simple situations (such as $M = SO(n)$, or $M = S^n$, or $M = \mathbb{R}^n$). It turns out that controllable systems of *two* vector fields exist on every connected manifold. The proof of this result (due to N. Levitt and the present author) will be given in [8].

We shall also consider the same question for the property of *local controllability*, defined below. Here, our answer is complete: local controllability can always be achieved with *three* vector fields and (except for the trivial case of one-dimensional manifolds) it can never be achieved with two.

We now introduce our notations and definitions. If M is a C^∞ -manifold, we use $V(M)$ to denote the set of all C^∞ -vector fields on M . A *vector field system* is

* Received by the editors November 23, 1973, and in revised form January 24, 1974.

† Department of Mathematics, Rutgers University, University Heights Campus, New Brunswick, New Jersey 08903. This work was supported in part by the National Science Foundation Grant GP-37488.

a subset of $V(M)$. If S is a vector field system, a *trajectory* of S is a continuous mapping γ from a closed interval $[t_i, t_f]$ such that there are reals t_0, \dots, t_k and vector fields X^1, \dots, X^k in S with the property that $t_i = t_0 < \dots < t_k = t_f$ and that, for $j = 1, \dots, k$, the restriction of γ to $[t_{j-1}, t_j]$ is an integral curve of X^j . The points $\gamma(t_i), \gamma(t_f)$ are, respectively, the *initial* and the *final point* of the trajectory. If m and m' are points in M , we say that m' is *S-reachable* from m if there is a trajectory of S whose initial point is m and whose final point is m' . The set of all points that are *S-reachable* from m is called the *positive S-orbit* of m . We shall use $\mathcal{O}_+(m, S)$ to denote this set.

The system S is said to be *controllable* if for every $m \in M$ the positive S -orbit of m is all of M . Clearly, controllable systems can only exist on connected manifolds.

A system S is said to be *locally controllable* if every point has arbitrarily small neighborhoods on which S is controllable. Precisely, this means that for every $m \in M$ and every neighborhood V of m there is an open neighborhood U of m such that $U \subseteq V$ and that the system S_U , whose elements are the restrictions to U of the elements of S , is controllable. If S is locally controllable then the positive orbits are open (trivial) and closed (if m' is in the closure of $\mathcal{O}_+(m, S)$, let U be a neighborhood of m' on which S is controllable. Then U contains a point $m'' \in \mathcal{O}_+(m, S)$. Since m'' is S -reachable from m , and m' is S -reachable from m'' , it follows that $m' \in \mathcal{O}_+(m, S)$). Therefore, a locally controllable system on a connected manifold is controllable. The converse is, of course, not true in general.

Our main result can now be stated.

THEOREM. *Let M be a C^∞ -manifold. Then there exist three C^∞ -vector fields X, Y, Z on M such that the system $\{X, Y, Z\}$ is locally controllable. If the dimension of M is at least two, then no system of two vector fields can be locally controllable.*

The remarks preceding the statement of the theorem enable us to conclude the following.

COROLLARY. *On every connected C^∞ -manifold there exists a controllable system that consists of three vector fields.*

We now turn to the proof of our theorem. To prove the first statement we shall rely on a result of C. Lobry. In [5], Lobry proved that the following property (C) of pairs (X, Y) of vector fields is generic:

(C) Let $B(X, Y)$ denote the smallest set of vector fields which contains X and Y and which is closed under the operation of Lie bracket. Then, for every $m \in M$, the vectors $Z(m), Z \in B(X, Y)$, span the entire tangent space of M at m .

The assertion that property (C) is “generic” means that every pair of C^∞ -vector fields can be approximated, in a suitable topology, by pairs of C^∞ -vector fields for which (C) holds. We shall only need a weaker form of this result.

LEMMA 1. *On every manifold M there exists a pair of vector fields which has property (C).*

In addition to Lobry’s result, we shall use the “positive form of Chow’s theorem.” A very short and elegant proof can be found in Krener [3] (cf. also [1], [4], [6] and [7]).

LEMMA 2. *Let (X, Y) be a pair of vector fields that has property (C). Then every positive orbit of the system $\{X, Y\}$ has a nonempty interior.*

We shall use Lemma 2 to prove the following.

LEMMA 3. *Let (X, Y) be a pair of vector fields that has property (C). Suppose*

the manifold M is connected. Let $Z = -X - Y$. Then the system $\{X, Y, Z\}$ is controllable.

To prove Lemma 3, let $S = \{X, Y, Z\}$. We first show that the positive S -orbits are dense in M . Let $m \in M$ and let N denote the closure of the positive S -orbit of m . For a vector field $W \in V(M)$, let $\{W_t\}$ denote the one-parameter family of local diffeomorphisms induced by W so that, for $p \in M$, the curve $t \rightarrow W_t(p)$ is the integral curve of W which goes through p when $t = 0$. It is easy to see that whenever $n \in N$ and $t \geq 0$, then $X_t(n)$, $Y_t(n)$ and $Z_t(n)$ also belong to N . Next recall that the identity

$$(V + W)_t(n) = \lim_{k \rightarrow \infty} (V_{t/k} W_{t/k})^k(n)$$

is valid for V, W in $V(M)$.

(This can be proved as follows: let $I_{t,k}$ denote the union of the intervals $2it/k, (2i + 1)t/k$, with $i = 0, \dots, k - 1$. Let u_k denote the characteristic function of $I_{t,k}$. For a function $t \rightarrow u(t)$ defined on $[0, 2t]$, use $x_u(s)$ to denote the solution of

$$\dot{x}(s) = (1 - u(s))V(x(s)) + u(s)W(x(s))$$

with initial condition $x(0) = n$. As $k \rightarrow \infty$, it is clear that the functions u_k converge weakly in the interval $[0, 2t]$ to the constant function $u(s) = \frac{1}{2}$. Therefore,

$$\lim_{k \rightarrow \infty} x_{u_k}(2t) = x_u(2t).$$

But $x_{u_k}(2t) = (V_{t/k} W_{t/k})^k(n)$, and $x_u(2t) = (V + W)_t(n)$.

We can apply this identity with $V = Z$, $W = Y$, so that $V + W = -X$. For $t \geq 0$, successive applications of $W_{t/k}$ and of $V_{t/k}$ take points of N into points of N . Since N is closed, we conclude that $(-X)_t(n)$ belongs to N . But $(-X)_t = X_{-t}$. Therefore, we have proved that, if $n \in N$, then $X_t(n) \in N$ for every real t for which it is defined. A similar conclusion is, of course, valid with X replaced by Y . Therefore N is a union of orbits (or leafs) of the system $\{X, Y\}$ (cf. [6], [7]). However, the assumption that (X, Y) has property (C) enables us to apply Chow's theorem (cf. [1], [3], [4], [6] or [7]), and conclude that the $\{X, Y\}$ -orbit of m is all of M . Therefore $N = M$, and this implies that $\mathcal{O}_+(m, S)$ is dense in M .

To conclude the proof of Lemma 3, we apply Lemma 2 to the vector field system whose elements are $-X$ and $-Y$. It follows that the positive $\{-X, -Y\}$ -orbit of any point $m' \in M$ has a nonempty interior. If $m \in M$, the density of $\mathcal{O}_+(m, S)$ implies that there is $m'' \in M$ which is both reachable from m by an S -trajectory and from m' by a $\{-X, -Y\}$ -trajectory. Now, it is clear that a $\{-X, -Y\}$ -trajectory, if "run in reverse," becomes an $\{X, Y\}$ -trajectory and, *a fortiori*, an S -trajectory. Therefore m' is S -reachable from m'' . Since m'' is S -reachable from m , we conclude that $m' \in \mathcal{O}_+(m, S)$. But m and m' were arbitrary. Therefore, S is controllable and Lemma 3 is proved.

Using Lemmas 1 and 3, we can prove the first assertion of our Theorem. Let $\{X, Y\}$ be a pair of vector fields that has property (C) (such a pair exists by Lemma 1). Let $Z = -X - Y$. Since property (C) is obviously local, it follows that the pair $\{X/U, Y/U\}$ has property (C) for every open subset U of M . (Here " $/U$ " means "restricted to U ".) If U is connected, then $\{X/U, Y/U, Z/U\}$ is

controllable by Lemma 3. Since every point of M has arbitrarily small connected neighborhoods, it follows that $\{X, Y, Z\}$ is locally controllable.

To complete the proof of the theorem, we must show that a system of two vector fields X and Y cannot be locally controllable. Assume that a locally controllable system $\{X, Y\}$ is given on a manifold M whose dimension is at least two. Then there must exist a point $m \in M$ such that $X(m)$ and $Y(m)$ are linearly independent. (Otherwise, the orbits of the system $\{X, Y\}$ would be zero- or one-dimensional.) Let w be an element of the cotangent space of M at m such that $\langle w, X(m) \rangle = \langle w, Y(m) \rangle = 1$. Let f be a C^∞ -function defined on a neighborhood V of m , such that the differential of f at m is w . Then $(Xf)(m) = (Yf)(m) = 1$. By taking a smaller V , if necessary, we can assume that the functions Xf and Yf are positive throughout V . Let

$$V_- = \{m' : m' \in V \text{ and } f(m') < f(m)\}.$$

Clearly, f is nondecreasing along every trajectory of $\{X, Y\}$ which is contained in V . Therefore no point of V_- can be reached from m by such a trajectory. But V_- contains points arbitrarily close to m (because the differential of f at m does not vanish). From this it is clear that V contains no neighborhood U of m such that the system $\{X, Y\}$, restricted to M , is controllable. The proof of our theorem is now complete.

The remainder of this paper is devoted to the discussion of some particular examples of manifolds M for which we know how to show the existence of a controllable system of two vector fields by "elementary" means. The proof that this can be done in general will be given in [8].

To begin with, one can construct examples of this situation by letting M be a compact connected Lie group G whose Lie algebra is generated by two elements. We take X and Y to be generators of this Lie algebra, and we assert that $\{X, Y\}$ is controllable on G . Indeed, the way X and Y were defined guarantees that condition (C) holds. It then follows from Theorem 7.1 of [2] that the system is controllable.

An example of a Lie group for which this construction works is $SO(n)$, the special orthogonal group in dimension n (where n is arbitrary). Indeed, $SO(n)$ is compact and connected. Moreover, the Lie algebra of $SO(n)$ can be identified with the Lie algebra of all skew-symmetric $n \times n$ real matrices. It is easy to find a pair of generators (for instance, the matrices A and B given by $A = (a_{ij})$, $1 \leq i, j \leq n$, where $a_{12} = 1$, $a_{21} = -1$ and all other entries equal to zero, and $B = (b_{ij})$, $i \leq i, j \leq n$, with

$$\begin{aligned} b_{12} &= b_{23} = \cdots = b_{n-1,n} = 1, \\ b_{21} &= b_{32} = \cdots = b_{n,n-1} = -1, \end{aligned}$$

and all other entries equal to zero).

In fact, the construction of the preceding paragraphs can be used in a more general situation. Suppose the manifold M is a coset space of a Lie group G for which our construction works. If X belongs to the Lie algebra of G , then to the one-parameter subgroup $t \rightarrow \exp(tX)$ of G there corresponds a one-parameter group of diffeomorphisms of M whose infinitesimal generator is a vector field X .

It is easy to see that, if $\{X, Y\}$ is controllable on G , then $\{X, Y\}$ is controllable on M .

Since the n -dimensional sphere S^n is a coset space of $SO(n+1)$, the preceding remarks show, in particular, how to construct a controllable system $\{X, Y\}$ on S^n .

As a last example, we will prove a lemma which, in particular, shows how to construct a controllable pair of vector fields on n -dimensional Euclidean space \mathbb{R}^n . The idea is simply as follows: \mathbb{R}^n is S^n with a point π removed. Let $\{X, Y\}$ be a controllable pair on S^n . Let X', Y' be the restrictions of X, Y to $S^n - \{\pi\}$ (i.e., \mathbb{R}^n). It seems intuitively obvious that $\{X', Y'\}$ is controllable in \mathbb{R}^n (if $n \geq 2$). Indeed, if p and q are any two points of \mathbb{R}^n , they can be joined by a trajectory of $\{X, Y\}$ in S^n . In fact, there are infinitely many such trajectories, and it seems highly unlikely that all these trajectories will go through π . And, as long as we can find a trajectory γ which does not go through π , then γ will be a trajectory of $\{X', Y'\}$ in \mathbb{R}^n . The following lemma gives a proof that points (and, more generally, closed submanifolds of codimension ≥ 2) can be removed without destroying controllability.

LEMMA 4. *Let S be a vector field system on a manifold M . Assume that S is controllable and that the Lie subalgebra of $V(M)$ generated by S has maximal dimension at each point of M . Let M' be an open submanifold of M which is obtained by removing from M a closed submanifold of codimension ≥ 2 . Let S' denote the set of restrictions to M' of the elements of S . Then S' is controllable.*

Proof. Let p and q be points of M' . We must show that p and q can be joined by a trajectory of S which is entirely contained in M' . Let us call such trajectories "good". It follows easily from the "positive form of Chow's theorem" that there is an open subset U of M' such that every point of U is reachable by a good trajectory from p . Similarly, if we apply the positive form of Chow's theorem to the "reverse system" (i.e., $\{X: X \in V(M') \text{ and } -X \in S'\}$) we conclude that there is an open $V \subseteq M'$ such that q is reachable from every point of V by a good trajectory. Take points r, s in U, V , respectively. Since S is controllable, there is a trajectory γ of S which goes from r to s . (Of course, γ need not be good.) Let $k, t_0, \dots, t_k, X^1, \dots, X^k$ be such that

$$t_i = t_0 < t_2 < \dots < t_k = t_f.$$

$X^j \in S$ for $1 \leq j \leq k$, and the restriction of γ to $[t_{j-1}, t_j]$ is an integral curve of X^j . Clearly, we can assume that $t_0 = 0$. For each t such that $0 \leq t \leq t_f$, let Γ_t denote the local diffeomorphism

$$X_{t-t_{j-1}}^j X_{t_{j-1}-t_{j-2}}^{j-1} \dots X_{t_2-t_1}^2 X_{t_1}^1,$$

where j is such that $t_{j-1} \leq t \leq t_j$. Thus, for each point $m \in M$, the curve $t \rightarrow \Gamma_t(m)$ is the trajectory that corresponds to the same "control" as γ , but whose initial point is m . In particular, the curve $t \rightarrow \Gamma_t(r)$ is precisely γ , and $\Gamma_{t_f}(r) = s$. We can certainly assume that $\Gamma_{t_f}(U) = V$. (Otherwise, let $U' = \Gamma_{t_f}^{-1}(V \cap \Gamma_{t_f}(U))$ and replace U by U' and V by $\Gamma_{t_f}(U')$.) Let A be the set of all points of M which are of the form $\Gamma_t^{-1}(a)$ for some $a \in M - M'$, and some t such that $0 \leq t \leq t_f$. Then A is the image of $[0, t_f] \times (M - M')$ under the map $(t, a) \rightarrow \Gamma_t^{-1}(a)$. The domain of this map is a manifold whose dimension is strictly less than that of M . Moreover, this domain is the union of finitely many pieces in each of which the map is C^∞ .

Therefore, by Sard's theorem, the set A has measure zero in M . In particular, the interior of A is empty. It follows that U contains a point r' which does not belong to A . Therefore, for every t in $[0, t_f]$, the point $\Gamma_t(r')$ is in M' . It follows that the trajectory $t \rightarrow \Gamma_t(r')$ ($0 \leq t \leq t_f$) is good. Let $s' = \Gamma_{t_f}(r')$. Then $s' \in V$, so that there is a good trajectory from s' to q . Since $r' \in U$, there is a good trajectory from p to r' . Finally, we have seen that there is a good trajectory from r' to s' . Therefore, there is a good trajectory from p to q , and the proof of our Lemma is complete.

Acknowledgment. The author is grateful to D. Elliott for his encouragement and helpful discussions, and to an anonymous referee for helpful comments.

REFERENCES

- [1] R. HERMANN, *On the accessibility problem in control theory*, International Symposium on Nonlinear Differential Equations and Nonlinear Mechanics, Academic Press, New York, 1963, pp. 325–332.
- [2] V. JURDJEVIC AND H. SUSSMANN, *Control systems on Lie groups*, J. Differential Equations, 12 (1972), pp. 313–329.
- [3] A. KRENER, *A generalization of Chow's theorem and the bang-bang theorem to nonlinear control problems*, this Journal, 12 (1974), pp. 43–52.
- [4] C. LOBRY, *Contrôlabilité des systèmes non linéaires*, this Journal, 8 (1970), pp. 573–605.
- [5] ———, *Une propriété générique des couples de champs de vecteurs*, Czechoslovak Math. J., 22, 97 (1972), pp. 230–237.
- [6] H. SUSSMANN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95–116.
- [7] H. SUSSMANN, *Orbits of families of vector fields and integrability of distributions*, Trans. Amer. Math. Soc., 180 (1973), pp. 171–188.
- [8] N. LEVITT AND H. SUSSMANN, *On controllability by means of two vector fields*, submitted.

INSTABILITY CRITERIA FOR TIME-VARYING NONLINEAR FUNCTIONAL DIFFERENTIAL SYSTEMS*

E. NOLDUS†

Abstract. Instability criteria are derived for feedback systems containing a time-invariant linear element in the forward branch, and a time-varying nonlinear feedback amplifier. The transfer function of the linear element is not restricted to rational functions. The results are obtained using Lyapunov techniques in function space. They are considerably stronger than those in early papers on the subject, which are confirmed as special cases of the present ones.

1. Introduction. In 1967 Brockett and Lee [3] derived the instability counterpart of the circle criterion for finite-dimensional systems. Since then a number of instability criteria for feedback systems have appeared in the control theory literature. The basic technique in proving these criteria may either involve the use of Lyapunov functions derived from path integrals in state space [10], [11] or Riccati-type algebraic equations, or rely on positive operator methods, requiring the introduction of notions of causality and invertibility [1], [5], [12], [13]. The main advantages of the latter methods are their basic simplicity, the intuitive nature of the arguments, and their independence of the dimensionality of the state space. Lyapunov techniques however have their own merits. Specifically they seem to provide the only means of investigating a system's properties in a subset of the state space. This is the case e.g., when a type of instability is examined which drives the system in a sustained, bounded oscillation. Essentially it must be shown in this case, that the system's state is driven into a bounded subset Ω of the state space, which does not contain the resting state, and afterwards never leaves Ω again. By estimating Ω , such concepts as instability regions or oscillation amplitudes could then be calculated. Results such as these seem to remain out of reach of positive operator methods as presently applied. However they can be obtained, potentially at least, by Lyapunov methods.

As a result some efforts have been made to extend the theory of Lyapunov to include infinite-dimensional systems, and in particular, systems governed by functional differential equations which will be studied here. Basic problems concerning the existence and uniqueness of solutions, and the generalization of LaSalle's [9] Lyapunov theorems to functional differential systems have been covered by Hale [8].

A method for the systematic generation of Lyapunov functionals for such systems has been given by Gruber [6]. It represents an extension to infinite-dimensional systems of a technique originally developed by Brockett for ordinary differential equations [2]. The method relies on a description of functional differential systems in terms of convolution equations involving distributions with compact support. A state space is defined, and quadratic functionals of the state are generated by means of path independent line integrals in state space. Gruber has successfully applied his method for solving certain problems in stability theory and in optimal control. Here it is used for deriving instability criteria.

* Received by the editors December 1, 1972.

† Laboratory De Winne for Control Theory, University of Ghent, B-9710 Zwijnaarde, Belgium.

The results of this paper can be divided into two classes: those pertaining to the existence of unbounded motions, and those dealing with instability in the sense of bounded, self-excited oscillations. The former ones have also been obtained using positive operator techniques [12], the latter ones have not. The systems under investigation are described by an equation

$$(1) \quad p * z + f(q * z, t) = 0$$

where p and q are distributions with compact support such that $p = D^n \delta + p'$, with $D^n \delta$ the n th derivative of the delta functional and p' and q distributions of order less than or equal to $n - 1$, and with support in $[0, T]$, $T > 0$. $f(u, t)$ is the characteristic of a nonlinear, time-varying feedback amplifier. Although the restrictions on p and q are somewhat stronger than necessary, (1) represents a wide variety of dynamical feedback equations, including ordinary differential equations, differential-difference equations and broad classes of functional differential equations. A block diagram for (1) is shown in Fig. 1, where

$$H(s) = \frac{q(s)}{p(s)}$$

is the transfer function of a linear time-invariant operator, $p(s)$ and $q(s)$ denoting the Laplace transforms of p and q . The state space of (1) is $C = C^{n-1}[-T, 0]$, the vector space of all $(n - 1)$ -times continuously derivable scalar functions $\varphi(\theta)$, with domain $[-T, 0]$. $z_t(\theta) = [z(t + \theta), \theta \in [-T, 0]] \in C$ is the state at time t .

The quadratic Lyapunov functionals of the state φ used below are written in the general form

$$(2) \quad V_1(\varphi) = \int_{t(0)}^{t(\varphi)} \left[p * z \cdot q * z - \sum_i (u_i * z)^2 - (\hat{r} * z)^2 \right] dt,$$

where p , q , u_i and \hat{r} are distributions of order smaller than or equal to n , and with support in $[0, T]$. p , q and u_i are given, while \hat{r} is defined by a relation

$$G(s) = \frac{1}{2}[p(s)q(-s) + p(-s)q(s)] - \sum_i u_i(s)u_i(-s) = \hat{r}(s)\hat{r}(-s).$$

From the theory of the spectral factorization of entire functions one knows that such a $\hat{r}(s)$ exists if

$$\operatorname{Re} G(j\omega) \geq 0 \quad \text{for all real } \omega.$$

If $z(t)$ is any function such that $z_t(\theta) = \varphi$ at $t = t(\varphi)$, and $z_t(\theta) = 0$ at $t = t(0)$,

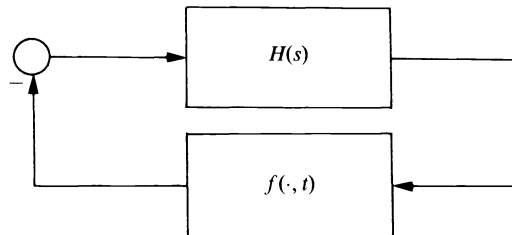


FIG. 1. A time-varying nonlinear feedback system

then the integral (2) is interpreted as the line integral along the path in C , defined by $z(t)$. If the integral exists, i.e., the convolutions in the integrand are square integrable, then it is independent of the path, and hence a functional of the state φ only. The theory of these line integrals and their applications in stability theory have been thoroughly studied by Brockett [2], [3], Willems [4], [7] and Gruber [6]. We refer to these authors for further mathematical background. The paper by Gruber provides an excellent introduction.

2. Objectives and definitions. Since the appearance of Popov's work, remarkable developments in the frequency techniques area have been reported. Applications have appeared in stability theory, optimization, sensitivity analysis and network theory. The following sections contain a contribution on frequency methods in instability theory. The definitions of instability are chosen strong enough to imply a well-specified asymptotic behavior. The results always appear in pairs:

(a) a criterion which guarantees that the system state will leave a given neighborhood of the (unique) equilibrium state, and afterwards will never return to this neighborhood. A dissipative system (i.e., a system whose solutions are ultimately bounded by a ball of sufficiently large radius), which satisfies such a criterion, builds up a sustained oscillatory motion.

(b) a criterion which ensures the existence of unbounded motions. A system which satisfies such a criterion in a bounded subset of its state space must leave this subset after a finite time lapse. One advantage of instability criteria is that they may provide a test for the sharpness of corresponding stability criteria. Necessary conditions for stability being unavailable, one may successively apply a stability criterion and an instability criterion to a given system. If both criteria are sufficiently strong, the asymptotic behavior may be determined for most, if not all, values of the system parameters. For such purposes instability criteria in earlier papers are too poor. Assume e.g., that the nonlinear function $f(u, t)$ satisfies for all u, t ,

$$k_1 u^2 \leq u f(u, t) \leq k_2 u^2.$$

Instability conditions as stated e.g., in [1], [3], [13] then imply that the linearized system, with $f(u, t) = ku$, must be unstable for all $k_1 \leq k \leq k_2$. In reality a feedback loop becomes unstable under much weaker conditions. Suppose that the system is unstable when linearized at the equilibrium state, but becomes stable when linearized at points sufficiently far from equilibrium. Then one may suspect that an unstable behavior of type (a) will occur. In the following sections we shall be looking for criteria that apply to such cases.

We shall restrict ourselves to criteria of the Popov type. By this is meant that in the frequency condition a multiplier of the form $(1 + \alpha s)$ shows up. The results however could easily be further generalized, in exactly the same manner as Popov's theorem has been generalized in stability theory, for example, as in [4]. Some manipulations on the feedback equation will be necessary. Let us consider the case of a time-invariant system

$$(3) \quad p * z + f(q * z) = 0.$$

Assuming that $f(0) = 0$ and that $f(u)$ is differentiable at $u = 0$, let $p_1(s) = p(s) + f^{(1)}(0)q(s)$, and $f_1(u) = f(u) - f^{(1)}(0)u$. The linearization of (3) at the equilibrium state, $z_t = 0$, is $p_1 * z = 0$, and (3) can be written as

$$p_1 * z + f_1(q * z) = 0, \quad f_1^{(1)}(0) = 0.$$

In what follows an important role is played by the zeros in the open right-half plane (RHP) of $p_1(s)$. Assume there are a finite number n_1 such zeros, $\lambda_1, \lambda_2, \dots, \lambda_{n_1}$. Let

$$v(s) = \prod_{i=1}^{n_1} (s - \lambda_i).$$

Then the relation

$$p_1(s) = v(s)w(s)$$

defines $w(s)$ as an entire function. It is easy to see that the inverse Laplace transform $w = \mathcal{L}^{-1}w(s)$ has the order $(n - n_1)$ and support $\in [0, T]$. Furthermore if z is a solution of (3), $\hat{z} = w * z$ is an $(n_1 - 1)$ -times continuously derivable function, and (3) is equivalent to

$$(4) \quad v(D)\hat{z} + f_1(q * z) = 0,$$

$$(5) \quad \hat{z} = w * z.$$

Equation (4) can be written in vector form as

$$(6) \quad \dot{\hat{x}} = M\hat{x} - bf_1(q * z),$$

where

$$(7) \quad \hat{x} = \text{adj}(DI - M) \cdot b\hat{z}$$

and M is such that $\det(sI - M) = v(s)$. Note that all characteristic values of M are confined to the open RHP. In the sequence we shall often use the transformation of variables

$$(8) \quad z = e^{-rt}y$$

with r a real constant. By this transformation (3)–(5) become

$$(3') \quad p_r * y + e^{rt}f(e^{-rt}q_r * y) = 0,$$

$$(4') \quad v_r(D)\hat{y} + e^{rt}f_1(e^{-rt}q_r * y) = 0,$$

$$(5') \quad \hat{y} = w_r * y,$$

where $\hat{z} = e^{-rt}\hat{y}$, and the distributions p_r, q_r, v_r and w_r are defined by their Laplace transforms

$$p_r(s) = p(s - r), \quad q_r(s) = q(s - r), \quad v_r(s) = v(s - r), \quad w_r(s) = w(s - r).$$

Finally let us state our definitions of instability.

DEFINITION 1. Assume that for the system (3) an initial state z_0 at $t = 0$ can be selected arbitrarily close to the origin, such that for some $t_0 > 0$ and some $\varepsilon > 0$, ε independent of z_0 , the following relation holds:

$$(9) \quad \|z_t\| \triangleq \sum_{i=0}^{n-1} \sup_{-T \leq \theta \leq 0} |z_t^{(i)}(\theta)| \geq \varepsilon > 0 \quad \text{for all } t \geq t_0.$$

Then we shall say that (3) is *weakly unstable*.

The left-hand side of (9) is a suitable definition for the norm in the state space C .

DEFINITION 2. Assume that for some $r < 0$, (3') is weakly unstable. Then it follows from (8) that

$$(10) \quad \|z_t\| \rightarrow \infty, \quad t \rightarrow +\infty,$$

where the initial state z_0 at $t = 0$ can be selected arbitrarily close to the origin. We shall say that under these conditions (3) is *strongly unstable*.

Definitions 1 and 2 correspond to types (a) and (b) of unstable asymptotic behavior described before.

3. Instability criteria for time-invariant nonlinear systems. In this section we state and discuss four Popov-type instability criteria for time-invariant nonlinear loops. The proofs are rather lengthy. A detailed derivation of Criteria 1 and 2 is given in Appendix A. The proof of Criteria 3 and 4 is quite similar, and an outline can be found in Appendix B. We have the following.

CRITERION 1. *Suppose*

$$(i) \quad k_1 u^2 \leq uf(u) \leq k_2 u^2 \text{ for all } u;$$

$$(ii) \quad 1 + f^{(1)}(0)H(s) \text{ has } n_1 \geq 1 \text{ zeros in the open RHP and all other zeros in } \operatorname{Re} s \leq -d < 0;$$

$$(iii) \quad \text{scalars } \alpha \geq 0 \text{ and } r, 0 \leq r < d, \text{ can be found such that}$$

$$(11) \quad \operatorname{Re} [1 - \alpha(j\omega - r)] \left[\frac{1 + k_2 H(j\omega - r)}{1 + k_1 H(j\omega - r)} \right]^{\pm 1} - \alpha r \geq \eta^2 > 0 \quad \text{for all real } \omega.$$

Then (3) is weakly unstable. ± 1 stands for either $+1$ or -1 .

CRITERION 2. *Suppose*

$$(i) \quad k_1 u^2 \leq uf(u) \leq k_2 u^2 \text{ for all } u;$$

$$(ii) \quad 1 + f^{(1)}(0)H(s - r) \text{ has } n_1 \geq 1 \text{ zeros in the open RHP and all other zeros in } \operatorname{Re} s \leq -\varepsilon < 0, \text{ for a scalar } r < 0;$$

$$(iii) \quad \text{for a scalar } \alpha \leq 0, (11) \text{ holds.}$$

Then (3) is *strongly unstable*.

In the next two criteria, an additional condition is imposed on the amplifier characteristic $f(u)$. On the other hand, the frequency condition is somewhat weakened. The modifications read as follows.

CRITERION 3. *In Criterion 1, condition (iii) may be replaced by:*

$$(iii') \quad \text{scalars } \alpha \geq 0 \text{ and } r, 0 \leq r < d, \text{ can be found such that either}$$

$$(12) \quad \int_0^u f(\theta) d\theta \geq \frac{1}{2}uf(u) \quad \text{for all } u,$$

$$(13) \quad \operatorname{Re} (1 - \alpha j\omega) \left[\frac{1 + k_2 H(j\omega - r)}{1 + k_1 H(j\omega - r)} \right] \geq \eta^2 > 0 \quad \text{for all real } \omega$$

or

$$(14) \quad \int_0^u f(\theta) d\theta \leq \frac{1}{2}uf(u) \quad \text{for all } u,$$

$$(15) \quad \operatorname{Re} (1 - \alpha j\omega) \left[\frac{1 + k_1 H(j\omega - r)}{1 + k_2 H(j\omega - r)} \right] \geq \eta^2 > 0 \quad \text{for all real } \omega.$$

CRITERION 4. In Criterion 2, condition (iii) may be replaced by:

(iii') for a scalar $\alpha \leq 0$ either (1), (13), or (14), (15) hold.

The analogy between Popov's stability theorem and Criteria 1 and 2 is apparent. The nonlinearity is characterized in the same way. In Criteria 3 and 4 there is an additional constraint on the nonlinearity. The frequency condition however is less severe. To see this, note that (11) can be written as

$$(16) \quad \operatorname{Re}(1 - \beta j\omega) \left[\frac{1 + k_2 H(j\omega - r)}{1 + k_1 H(j\omega - r)} \right]^{\pm 1} - \frac{\beta r}{1 - \beta r} \geq \eta^2 > 0,$$

where $0 \leq \beta r < 1$.

This is stronger than (13) or (15), where αr may be any nonnegative scalar. Criteria 3 and 4 are important for practical applications, since many technical nonlinearities satisfy condition (12) or (14). Nonlinearities caused by saturation, cutoff or a dead zone belong to this class. If the amplifier characteristic is exponential, as, for example, in nuclear reactor models, (12) or (14) are not true, and Criteria 1 and 2 must be applied.

Criteria 3 and 4 have been published previously by the author, for the cases of ordinary [10] and differential-difference equations [11]. They are now extended to functional differential equations. Criteria 1 and 2 are new. For $r = 0$ the latter ones are reduced to the results by Brockett and Lee [3] for ordinary differential equations. As a rule however zero is a poor choice for r , such that the present criteria are substantially better. As a simple example, let $H(s) = 1/(s + 1)^3$. Then $1 + f^{(1)}(0)H(s)$ has two zeros in the open RHP for $f^{(1)}(0) > 8$, while the third zero is in $\operatorname{Re} s < -3$. For $\alpha = 0$, $r = 1$, (11) is certainly satisfied with $k_1 = 0$, $k_2 > 0$ arbitrary. Hence the feedback loop is weakly unstable if

$$f^{(1)}(0) > 8, \quad uf(u) \geq 0 \quad \text{for all } u.$$

This conclusion cannot be reached by selecting $r = 0$. In some cases, such as for systems with time delays, there may be no "best" choice for r : Every value of r yields another class of nonlinearities for which instability is ensured. For $\alpha = r = 0$ we have the counterpart of the circle theorem as stated by Brockett and Lee [3]. For $\alpha = 0$, $r \neq 0$, generalized circle instability theorems are obtained. These remain valid for more general equations, for example, of the form

$$p * z + f(v * z) \cdot q * z = 0$$

with

$$k_1 \leq f(u) \leq k_2 \quad \text{for all } u.$$

4. Time-varying nonlinear systems. Criteria 1–4 can be generalized to systems described by equation (1), with a time-varying nonlinear feedback gain. Intuitively it is clear that if $f(u, t)$ varies very slowly with t , the preceding criteria should remain valid. In this section the speed of variation of $f(u, t)$ with t is measured by a parameter a . Subsequently this parameter appears in the frequency condition, yielding instability criteria that involve a trade-off between restrictions on the amplifier characteristic, on its speed of variation, and on the transfer function of the linear element. We first state the generalizations of Criteria 2 and 4.

CRITERION 5. Suppose

- (i) $k_1 u^2 \leq uf(u, t) \leq k_2 u^2$ for all u, t ;
- (ii) $1 + k_0 H(s - r)$ has $n_1 \geq 1$ zeros in the open RHP and all other zeros in $\operatorname{Re} s \leq -\varepsilon < 0$, for all k_0 in $k_1 \leq k_0 \leq k_2$ and for a scalar $r < 0$;
- (iii) for scalars $\alpha \leq 0$ and $a \leq 0$ either

$$(17) \quad \Delta[f(u, t), a] \leq 0 \quad \text{for all } u, t, \text{ and}$$

$$(18) \quad \operatorname{Re} [1 - \alpha(j\omega - r - a)] \left[\frac{1 + k_2 H(j\omega - r)}{1 + k_1 H(j\omega - r)} \right] - \alpha r \geq \eta^2 > 0 \quad \text{for all real } \omega$$

or

$$(19) \quad \Delta[f(u, t), -a] \geq 0 \quad \text{for all } u, t, \text{ and}$$

$$(20) \quad \operatorname{Re} [1 - \alpha(j\omega - r - a)] \left[\frac{1 + k_1 H(j\omega - r)}{1 + k_2 H(j\omega - r)} \right] - \alpha r \geq \eta^2 > 0 \quad \text{for all real } \omega.$$

Then (1) is strongly unstable. Here

$$\Delta[f(u, t), a] = \int_0^u \frac{\partial f(\theta, t)}{\partial t} d\theta + a \frac{[k_2 u - f(u, t)][f(u, t) - k_1 u]}{(k_2 - k_1)}.$$

CRITERION 6. In Criterion 5 condition (iii) may be replaced by:

(iii') for scalars $\alpha \leq 0$ and $a \leq 0$ either

$$(21) \quad \int_0^u f(\theta, t) d\theta \geq \frac{1}{2} uf(u, t) \quad \text{for all } u, t,$$

condition (17) holds and

$$(22) \quad \operatorname{Re} [1 - \alpha(j\omega - a)] \left[\frac{1 + k_2 H(j\omega - r)}{1 + k_1 H(j\omega - r)} \right] \geq \eta^2 > 0 \quad \text{for all real } \omega$$

or

$$(23) \quad \int_0^u f(\theta, t) d\theta \leq \frac{1}{2} uf(u, t) \quad \text{for all } u, t,$$

condition (19) holds and

$$(24) \quad \operatorname{Re} [1 - \alpha(j\omega - a)] \left[\frac{1 + k_1 H(j\omega - r)}{1 + k_2 H(j\omega - r)} \right] \geq \eta^2 > 0 \quad \text{for all real } \omega.$$

Criteria 1 and 3 cannot be generalized along the same lines unless an additional constraint is imposed on the nonlinearity: The linearized system at the equilibrium state must be time-invariant. Specifically we have the conditions:

$$(25) \left\{ \begin{array}{l} \text{(a) } f_0 \triangleq \partial f(u, t) / \partial u|_{u=0} \text{ is independent of } t; \\ \text{(b) Scalars } \varepsilon \text{ and } K \text{ can be found such that for all } |u| \leq \varepsilon \text{ and for all } t, \\ \left| \int_0^u f_1(\theta, t) d\theta \right| \triangleq \left| \int_0^u [f(\theta, t) - f_0 \theta] d\theta \right| \leq K|u|^3. \end{array} \right.$$

In practical applications $f(u, t)$ is often periodic in t . Then (b) is automatically satisfied. We now have the following.

CRITERION 7. Suppose

- (i) $k_1 u^2 \leq uf(u, t) \leq k_2 u^2$ for all u, t , and $f(u, t)$ satisfies conditions (25);
- (ii) $1 + f_0 H(s)$ has $n_1 \geq 1$ zeros in the open RHP and all other zeros in $\text{Re } s \leq -d < 0$;
- (iii) scalars $\alpha \geq 0$, $a \geq 0$ and r , $0 \leq r < d$, can be found such that either

$$(26) \quad \Delta[f(u, t), a] \geq 0 \quad \text{for all } u, t$$

and (18) holds, or

$$(27) \quad \Delta[f(u, t), -a] \leq 0 \quad \text{for all } u, t$$

and (20) holds.

Then (1) is weakly unstable.

CRITERION 8. In Criterion 7 condition (iii) may be replaced by:

- (iii') scalars $\alpha \geq 0$, $a \geq 0$ and r , $0 \leq r < d$, can be found such that either (21), (22), (26), or (23), (24), (27) hold.

For an outline of the proofs, see Appendix C. If $f(u, t)$ is independent of t , one may choose $a = 0$. Then Criteria 1–4 are obtained again. If $\partial f(u, t)/\partial t$ is unknown, one has to select $|a| = \infty$, the resulting criteria being of the circle type, with a frequency condition

$$\text{Re} \frac{1 + k_2 H(j\omega - r)}{1 + k_1 H(j\omega - r)} \geq \eta^2 > 0 \quad \text{for all real } \omega.$$

Another special case of Criterion 6 occurs when $f(u, t) = k(t)u$, i.e., the case of a time-varying linear system. Then conditions (21), (23) are satisfied, while (17), (19) are reduced to

$$\dot{k}(t) + 2a \frac{[k_2 - k(t)][k(t) - k_1]}{k_2 - k_1} \leq 0 \quad \text{for all } t,$$

resp.,

$$\dot{k}(t) - 2a \frac{[k_2 - k(t)][k(t) - k_1]}{k_2 - k_1} \geq 0 \quad \text{for all } t.$$

Further extensions of the results in this paper may be obtained by imposing stronger conditions on the nonlinearity and correspondingly weakening the restrictions on the linear element. This weakening may take the form of allowing, in the frequency condition, a broader class of multipliers than the Popov multiplier $(1 + \alpha s)$. As a matter of fact, using a technique developed by Willems [14], it can be shown that for time-varying linear systems, the Popov multiplier $(1 - \alpha(s - a))$ in the frequency conditions (22), (24) may be replaced by any rational function $Z(s - a)$, such that $Z(s)$ is positive real. This is the instability counterpart of a generalized circle criterion by Gruber and Willems [7].

5. Conclusion. It has been shown that frequency instability criteria for functional differential systems can systematically be constructed using Lyapunov ideas. Instability definitions have been stated in terms of the system's asymptotic motions. The obtained criteria are comparable to their stability counterparts in

formulation and in sharpness of results, as is demonstrated in specific applications. Much known material has been confirmed as special cases of the new results. At least part of these seem to be out of reach of the actual methods in positive operators research, which lately have been increasingly proposed for instability analysis.

Appendix A. Proof of Criteria 1 and 2. We shall assume that $k_1 = 0$, the general case being reducible to this one by standard manipulations. Then (3) is equivalent to

$$(A.1) \quad p * z + h[(1/k_2)p + q] * z = 0,$$

where the function $h(u)$ is defined by its inverse

$$h^{-1}(u) = f^{-1}(u) - \frac{u}{k_2}$$

and $uh(u) \geq 0$ for all u . (3') becomes

$$(A.1') \quad p_r * y + e^{rt} h[e^{-rt}((1/k_2)p_r + q_r) * y] = 0.$$

The criteria are proved in four steps:

(a) Define the quadratic functional

$$V_1(\varphi) = \int_{t(0)}^{t(\varphi)} [(1/k_2)p_r + q_r - \alpha(D - r)q_r] * y \cdot p_r * y - \Psi \, dt,$$

$$\Psi = \eta_1^2(q_r * y)^2 + \eta_2^2(Dq_r * y)^2 + (\hat{r} * y)^2,$$

where $\eta_i > 0$, $i = 1, 2$, and \hat{r} is determined by a spectral factorization as explained before. This factorization is possible if for all real ω ,

$$1/k_2 + \operatorname{Re} [1 - \alpha(j\omega - r)]H(j\omega - r) \geq (\eta_1^2 + \eta_2^2\omega^2)|H(j\omega - r)|^2,$$

which is satisfied for $|\eta_i|$ sufficiently small if

$$(A.2) \quad \operatorname{Re} [1 - \alpha(j\omega - r)][1 + k_2 H(j\omega - r)] - \alpha r \geq \eta^2 > 0 \quad \text{for all real } \omega$$

since the order of q is less than the order of p . Along the solutions of (3') or (A.1'), we have

$$\dot{V}_1(y_t) = -e^{2rt} \left[\left(\frac{1}{k_2} p + q \right) * e^{-rt} y \right] h \left[\left(\frac{1}{k_2} p + q \right) * e^{-rt} y \right] \\ + \alpha e^{2rt} f(q * e^{-rt} y) Dq * e^{-rt} y - \Psi.$$

Let

$$V_2(\varphi, t) = \alpha e^{2rt} \int_0^{e^{-rt} q_r * y} f(\theta) d\theta|_{y_t = \varphi}$$

and

$$V(\varphi, t) = V_1(\varphi) - V_2(\varphi, t).$$

Then

$$\begin{aligned} \dot{V}(y_t, t) = & -e^{2rt} \left[\left(\frac{1}{k_2} p + q \right) * e^{-rt} y \right] h \left[\left(\frac{1}{k_2} p + q \right) * e^{-rt} y \right] \\ & - \Psi - 2r\alpha e^{2rt} \int_0^{e^{-rt} q_r * y} f(\theta) d\theta \leq 0 \quad \text{for} \\ (A.3) \quad & \alpha r \geq 0. \end{aligned}$$

(b) The region $V(\varphi, 0) < 0$ is not empty, if:

$$(A.4) \quad p_{1r}(s) = p_1(s - r) \text{ has } n_1 \geq 1 \text{ zeros in the open RHP.}$$

Recall that $p(s) + f^{(1)}(0)q(s) = p_1(s)$, and

$$p(s) + h^{(1)}(0) \left[\frac{1}{k_2} p(s) + q(s) \right] = \lambda p_1(s), \quad \lambda = 1 + \frac{h^{(1)}(0)}{k_2}.$$

After some manipulations, $V(\varphi, t)$ can be written as

$$\begin{aligned} (A.5) \quad V(\varphi, t) = & \int_{t(0)}^{t(\varphi)} \left[\lambda \left(\frac{1}{k_2} p_r + q_r \right) - \alpha(D - r)q_r \right] * y \cdot p_{1r} * y dt \\ & - \int_{t(0)}^{t(\varphi)} \left[\Psi + h^{(1)}(0) \left[\left(\frac{1}{k_2} p_r + q_r \right) * y \right]^2 \right. \\ & \quad \left. + \alpha r f^{(1)}(0)(q_r * y)^2 \right] dt \\ & - \alpha e^{2rt} \int_0^{e^{-rt} q_r * y} f_1(\theta) d\theta|_{y_t = \varphi}. \end{aligned}$$

Assume that, for example, $\mu > 0$ is a zero of $p_{1r}(s)$. Choose $\varphi = c e^{\mu\theta}$, $-T \leq \theta \leq 0$, and compute the line integrals in the first two terms of (A.5) along the path defined by the function $y(t) = c e^{\mu t}$, hence with $y_t(\theta) = \varphi e^{\mu t}$, $t(0) = -\infty$, $t(\varphi) = 0$. The first term disappears since $p_{1r} * y \equiv 0$. The second term has the form $-Bc^2$, $B > 0$, since $\hat{r} * y$ is not identically zero along the integration path. The third term (where $t = 0$) is infinitely small of third order as $c \rightarrow 0$, since $f_1^{(1)}(0) = 0$. So $V(\varphi, 0) < 0$ for c sufficiently small. (a) and (b) imply that along the solution of (3') that starts at φ at $t = 0$,

$$(A.6) \quad V(y_t, t) \leq 0 \quad \text{for all } t \geq 0.$$

(c1) Assume that $p_1(s)$ has $n_1 \geq 1$ zeros in the open RHP and all other zeros in $\text{Re } s \leq -d < 0$. Select

$$(A.7) \quad 0 \leq r < d.$$

Then $p_{1r}(s) = v_r(s)w_r(s)$, where all zeros of $w_r(s)$ are confined to $\text{Re } s \leq -\varepsilon = r - d < 0$. Define

$$\hat{p}(s) = w_r(s) \prod_{i=1}^{n_1} (s + g_i), \quad g_i \geq \varepsilon > 0,$$

where \hat{p} has the order n and support $\in [0, T]$. The equation

$$(A.8) \quad \hat{p} * y = 0$$

therefore has a unique solution $y(\varphi)$ that starts at φ at $t = 0$, and approaches zero for $t \rightarrow +\infty$, since all zeros of $\hat{p}(s)$ are in $\text{Re } s \leq -\varepsilon < 0$. Compute $V(\varphi, t)$ along the path defined by $y(\varphi)$. Then $t(\varphi) = 0$, $t(0) = +\infty$. (A.8) can be written in vector form as

$$\dot{\hat{w}} = -G\hat{w}, \quad G = \text{diag}(g_1 \cdots g_{n_1}),$$

where the components of \hat{w} are independent linear combinations of \hat{y} , $D\hat{y}$, \dots , $D^{n_1-1}\hat{y}$. Along the path $y(\varphi)$ we have

$$A^2(\hat{w}_1^2 + \cdots + \hat{w}_{n_1}^2)|_{y_t=\varphi} = -\int_{+\infty}^0 2A^2(g_1\hat{w}_1^2 + \cdots + g_{n_1}\hat{w}_{n_1}^2) dt.$$

Since along the solutions of (A.8), $p_{1r} * y = v_r(D)\hat{y}$ is a linear combination of $\hat{w}_1 \cdots \hat{w}_{n_1}$, it is now clear from (A.5) that

$$V(\varphi, t) = \left[-A^2(\hat{w}_1^2 + \cdots + \hat{w}_{n_1}^2) + \delta_0(q_r * y)^2 - \alpha e^{2rt} \int_0^{e^{-rt}q_r * y} f_1(\theta) d\theta \right]_{y_t=\varphi} + \Phi,$$

where $\Phi \geq 0$ for A^2 sufficiently large, and $\delta_0 > 0$ is sufficiently small. Because of (A.6) this implies that for all $t \geq 0$,

$$-A^2(\hat{w}_1^2 + \cdots + \hat{w}_{n_1}^2) + \delta_0(q_r * y)^2 - \alpha e^{2rt} \int_0^{e^{-rt}q_r * y} f_1(\theta) d\theta < 0$$

or, after multiplying with e^{-2rt} ,

$$(A.9) \quad A^2(\hat{v}_1^2 + \cdots + \hat{v}_{n_1}^2) > \delta_0(q * z)^2 - \alpha \int_0^{q * z} f_1(\theta) d\theta,$$

where $\hat{v}_1 \cdots \hat{v}_{n_1}$ are linear combinations of \hat{z} , $D\hat{z}$, \dots , $D^{n_1-1}\hat{z}$, i.e., of the components \hat{x}_i of the vector \hat{x} .

(d) We shall now construct a neighborhood $N(z_t)$ around the origin, and show that the system state must leave it. Select $\zeta_0 > 0$ sufficiently small such that for all $|u| \leq \zeta_0$,

$$\left| \alpha \int_0^u f_1(\theta) d\theta \right| \leq \frac{1}{2} \delta_0 u^2.$$

This is possible since $f_1^{(1)}(0) = 0$. Because of (A.9) it follows that

$$(A.10) \quad A^2(\hat{v}_1^2 + \cdots + \hat{v}_{n_1}^2) > \frac{1}{2} \delta_0 (q * z)^2$$

if

$$(A.11) \quad (q * z)^2 \leq \zeta_0^2.$$

Define

$$(A.12) \quad \begin{aligned} N(z_t) &= [z_t; (q * z)^2 \leq \zeta_0^2, U(z_t) \leq \varepsilon_0^2], \\ U(z_t) &= \hat{x}' U \hat{x}, \end{aligned}$$

where $U = U' > 0$ is the solution of

$$M'U + UM = I.^1$$

(i) An initial state z_0 can be selected arbitrarily close to zero such that the corresponding trajectory leaves $N(z_t)$ after a finite time interval. Indeed, suppose that $(q * z)^2 \leq \zeta_0^2$ for all $t \geq 0$. Then $t_0 > 0$ must be found such that $U(z_t) > \varepsilon_0^2$. Deriving (A.12) with the aid of (6) yields

$$\dot{U}(z_t) = \hat{x}'\hat{x} - 2b'\hat{x}f_1(q * z).$$

For $U(z_t) = \varepsilon^2$, the first term in the right-hand side is infinitely small of second order and positive. The second term is infinitely small of third order, because of (A.10) and $f_1^{(1)}(0) = 0$. Hence there exist scalars ε_0^2 and $\delta > 0$ such that

$$\dot{U}(z_t) - \delta U(z_t) \geq 0 \quad \text{for } U(z_t) \leq \varepsilon_0^2,$$

which proves the existence of a finite t_0 for which $U(z_{t_0}) > \varepsilon_0^2$.

(ii) This trajectory cannot re-enter $N(z_t)$ for $t > t_0$. Indeed, if a trajectory enters $N(z_t)$ at time t , there are two possibilities:

$$\text{either:} \quad U(z_t) \leq \varepsilon_0^2, \quad (q * z)^2 = \zeta_0^2;$$

This is impossible because of (A.10) if we choose ε_0^2 sufficiently small—

$$\text{or:} \quad U(z_t) = \varepsilon_0^2, \quad (q * z)^2 \leq \zeta_0.$$

Now the trajectory cannot enter $N(z_t)$ as $\dot{U}(z_t) > 0$. We conclude that

$$(A.13) \quad U(z_t) + (q * z)^2 \geq \min(\zeta_0^2, \varepsilon_0^2) > 0 \quad \text{for all } t \geq t_0.$$

Since $\hat{x}_i, i = 1, \dots, n_1$, are linear combinations of $w * z, Dw * z, \dots, D^{n_1-1}w * z$, which all, as well as $q * z$, are convolutions of z with distributions of order $\leq n - 1$ and support in $[0, T]$, (A.13) implies (9). The imposed conditions are (A.2), (A.3), (A.4), (A.7), which can be restated and generalized as in Criterion 1.

Let us now start again at the third step.

(c2) Choose $r < 0$ and assume that $p_{1,r}(s)$ has $n_1 \geq 1$ zeros in the open RHP, and all other zeros in $\text{Re } s \leq -\varepsilon < 0$. $V(\varphi, t)$ is computed as before. One finds

$$A^2(\hat{w}_1^2 + \dots + \hat{w}_{n_1}^2) + \alpha e^{2rt} \int_0^{e^{-rt}q_r * y} f_1(\theta) d\theta \geq -V(\varphi, 0) > 0 \quad \text{for all } t \geq 0$$

or, after multiplying with e^{-2rt} ,

$$(A.14) \quad A^2(\hat{v}_1^2 + \dots + \hat{v}_{n_1}^2) + \alpha \int_0^{q * z} f_1(\theta) d\theta \rightarrow +\infty \quad \text{for } t \rightarrow +\infty,$$

which implies (10). This proves Criterion 2.

Appendix B. Proof of Criteria 3 and 4. The proof is quite similar to the preceding one. The only difference occurs in the definition of $V_1(\varphi)$, for which we now take

$$V_1(\varphi) = \int_{t(0)}^{t(\varphi)} \left[\left(\frac{1}{k_2} p_r + q_r - \alpha D q_r \right) * y \cdot p_r * y - \Psi \right] dt.$$

¹ $U = U' > 0$ denotes a symmetric positive definite matrix.

This requires that

$$\operatorname{Re}(1 - \alpha j\omega)[1 + k_2 H(j\omega - r)] \geq \eta^2 > 0 \quad \text{for all real } \omega.$$

Choosing $V_2(\varphi, t)$ and $V(\varphi, t)$ as before, and differentiating yields

$$\begin{aligned} \dot{V}(y_t, t) &= -e^{2rt} \left(\frac{1}{k_2} p + q \right) * e^{-rt} y \cdot h \left[\left(\frac{1}{k_2} p + q \right) * e^{-rt} y \right] - \Psi \\ &\quad + \alpha r e^{2rt} \left[u f(u) - 2 \int_0^u f(\theta) d\theta \right]_{u=q * e^{-rt} y} \\ &\leq 0 \end{aligned}$$

for

$$\alpha r \left[u f(u) - 2 \int_0^u f(\theta) d\theta \right] \leq 0 \quad \text{for all } u.$$

Instead of (A.5) we have

$$\begin{aligned} V(\varphi, t) &= \int_{t(0)}^{t(\varphi)} \left[\lambda \left(\frac{1}{k_2} p_r + q_r \right) - \alpha D q_r \right] * y \cdot p_{1r} * y \cdot dt \\ &\quad - \int_{t(0)}^{t(\varphi)} \left[\Psi + h^{(1)}(0) \left[\left(\frac{1}{k_2} p_r + q_r \right) * y \right]^2 \right] dt \\ &\quad - \alpha e^{2rt} \int_0^{e^{-rt} q_r * y} f_1(\theta) d\theta|_{y_t = \varphi}. \end{aligned}$$

The proof then proceeds as in Appendix A.

Appendix C. Proof of Criteria 5–8. We start from the system equation (1) and the analogues of (3'), (4'), (A.1), (A.1'), with $f(u, t)$ and $h(u, t)$ replacing $f(u)$ and $h(u)$. To prove Criteria 5 and 7, let

$$V_1(\varphi) = \int_{t(0)}^{t(\varphi)} \left[\left[(1 + \alpha a) \left(\frac{1}{k_2} p_r + q_r \right) - \alpha(D - r) q_r \right] * y \cdot p_r * y - \Psi \right] dt,$$

which introduces the frequency condition

$$\operatorname{Re}[1 - \alpha(j\omega - r - a)][1 + k_2 H(j\omega - r)] - \alpha r \geq \eta^2 > 0 \quad \text{for all real } \omega.$$

$V_2(\varphi, t)$ and $V(\varphi, t)$ are defined as before, replacing $f(u)$ by $f(u, t)$. Differentiation yields

$$\begin{aligned} \dot{V}(y_t, t) &= -e^{2rt} \left[\left(\frac{1}{k_2} p + q \right) * e^{-rt} y \right] h \left[\left(\frac{1}{k_2} p + q \right) * e^{-rt} y, t \right] \\ &\quad - \Psi - 2r\alpha e^{2rt} \int_0^{e^{-rt} q_r * y} f(\theta, t) d\theta \\ &\quad - \alpha e^{2rt} \left[a \left[u - \frac{1}{k_2} f(u, t) \right] f(u, t) + \int_0^u \frac{\partial f(\theta, t)}{\partial t} d\theta \right]_{u=q * e^{-rt} y} \\ &\leq 0 \quad \text{for } \alpha r \geq 0 \end{aligned}$$

and

$$\alpha \left[a \left[u - \frac{1}{k_2} f(u, t) \right] f(u, t) + \int_0^u \frac{\partial f(\theta, t)}{\partial t} d\theta \right] \geq 0 \quad \text{for all } u, t.$$

So the new elements in the proof are:

(a) In the derivative of V_2 an additional term appears, due to the fact that the amplifier is nonstationary.

(b) Since $\dot{V}(y_t, t)$ must be nonpositive, this requires an extra term in the quadratic part of $V(\varphi, t)$, which causes the parameter a to show up in the frequency condition.

Now we proceed as for time-invariant systems. For the proof of Criterion 5, note that φ can be so selected that $V(\varphi, 0) < 0$, if $1 + k_0 H(s - r)$ has at least one zero in the open RHP, with $k_0 = \partial f(u, 0) / \partial u|_{u=0}$, hence, $k_1 \leq k_0 \leq k_2$. This is sufficient to derive a relation such as (A.14). The proof of Criterion 7 can be continued along the lines of Appendix A, if (25) is satisfied. Criteria 6 and 8 are left to the reader.

REFERENCES

- [1] A. R. BERGEN AND S. TAKEDA, *On instability of feedback systems with a single nonlinear time-varying gain*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 462–464.
- [2] R. W. BROCKETT, *On the stability of nonlinear feedback systems*, IEEE Trans. Appl. and Ind., 83 (1964), pp. 443–449.
- [3] R. W. BROCKETT AND H. B. LEE, *Frequency domain instability criteria for time-varying and nonlinear systems*, Proc. IEEE, 55 (1967), pp. 604–619.
- [4] R. W. BROCKETT AND J. L. WILLEMS, *Frequency domain stability criteria*, IEEE Trans. Automatic Control, AC-10 (1965), pp. 255–261 and 401–413.
- [5] J. H. DAVIS, *Mean-square gain criteria for the stability and instability of time-varying systems*, Ibid., AC-17 (1972), pp. 214–219.
- [6] M. GRUBER, *Path integrals and Lyapunov functionals*, Ibid., AC-14 (1969), pp. 465–475.
- [7] M. GRUBER AND J. L. WILLEMS, *On a generalization of the circle criterion*, Proc. 4th Ann. Allerton Conf. Circuit and System Theory, Univ. of Illinois, Urbana, 1966, pp. 827–835.
- [8] J. K. HALE, *Sufficient conditions for stability and instability of autonomous functional-differential equations*, Differential Equations, 1 (1965), pp. 452–482.
- [9] J. P. LASALLE, *An invariance principle in the theory of stability*, Differential Equations and Dynamical Systems, J. K. Hale and J. P. LaSalle, eds., Academic Press, New York, 1967, pp. 277–286.
- [10] E. NOLDUS, *Instability of time-varying and nonlinear feedback systems*, J. Engrg. Math., 4 (1970), pp. 243–259.
- [11] ———, *Asymptotic behaviour of time-varying and nonlinear differential-difference equations*, Internat. J. Control, 14 (1971), pp. 73–81.
- [12] ———, *Criteria for unbounded motion by positive operator methods*, Ibid., 18 (1973), pp. 289–296.
- [13] J. C. WILLEMS, *Stability, instability, invertibility and causality*, this Journal, 7 (1969), pp. 645–671.
- [14] J. L. WILLEMS, *A general stability criterion for non-linear time-varying feedback systems*, Internat. J. Control, 11 (1970), pp. 625–631.

ON THE DIMENSIONS OF CONTROLLABILITY SUBSPACES: A CHARACTERIZATION VIA POLYNOMIAL MATRICES AND KRONECKER INVARIANTS*

MICHAEL E. WARREN AND ADRIAN E. ECKBERG, JR.†

Abstract. The controllability subspaces of a pair (A, B) , instrumental in the formulation of the geometric theory of decoupling, are shown to have a natural analog in terms of the kernel of the singular pencil of matrices $(\lambda I - A; -B)$. In addition the pencil of matrices leads directly to the multivariable canonical form of Brunovsky.

The possible dimensions of controllability subspaces are shown to be completely determined by a set of invariants of the pencil of matrices. The minimum dimension of controllability subspaces which contain arbitrary subspaces of the image of B is ascertained, and a construction for such subspaces is given.

1. Introduction. The theory of the decoupling of constant linear systems by state feedback received a considerable boost with the advent of the geometric theory of Wonham and Morse [10]. Their formulation relied heavily on the concept of a controllability subspace (c.s.), that is, a vector subspace satisfying certain restrictive conditions. Solvability of decoupling problems then became equivalent to finding suitable sets of c.s.

At approximately the same time, the results of Wolovich and Falb [9], Brunovsky [1], Popov [7], Kalman [6] and Rosenbrock [8] led to a definitive canonical form for the input-state dynamics of time invariant linear systems. Kalman [6] and Rosenbrock [8] sensed the relationship between this canonical decomposition and Kronecker's classical theory of pencils of matrices, while Wonham and Morse [11] recognized that the decomposition was in terms of controllability subspaces.

The purpose of this paper is two-fold. First we show there is a strong and natural connection between controllability subspaces and elements in the kernel of a singular pencil of matrices. Furthermore, the pencil of matrices leads easily to the canonical form of Brunovsky. Then exploiting certain invariants of the pencil of matrices, we are able to make definitive statements about the dimensions of possible c.s., including the existence and uniqueness of c.s. of a given dimension.

In what follows, we shall be concerned with linear time invariant systems whose input-state dynamics are described by either the difference equation

$$x_{k+1} = Ax_k + Bu_k,$$

or the differential equation

$$\dot{x}(t) = Ax(t) + Bu(t),$$

* Received by the editors August 31, 1973.

† Electronic Systems Laboratory, Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts. The first author is now at the Electrical Engineering Department, University of Florida, Gainesville, Florida 32603. The second author is now at Bell Laboratories, Holmdel, New Jersey 07733. This research was conducted at the Decision and Control Sciences Group of the M.I.T. Electronic Systems Laboratory, and supported by the National Science Foundation under Grant GK-25781.

where $x \in R^n$, a real n -dimensional vector space, and $u \in R^m$, with A and B real matrices of appropriate dimensions. The algebraic structure of A and B which we will develop is independent of the specific dynamics, and thus will apply equally well to systems governed by either differential or difference equations. Furthermore, although we choose real vector spaces for concreteness, everything that follows will hold for vector spaces over an arbitrary field.

Except where otherwise specified, we shall use upper case italic letters to refer to linear transformations between vector spaces or their associated matrices. Script letters will denote vector spaces or subspaces. Lower case italic letters will indicate vectors while reals (field elements) will be represented by lower case Greek letters. When appropriate, we shall also indicate the image of a map B by \mathcal{B} ; and if $b \in \mathcal{B}$, ℓ stands for the subspace spanned by b . Further, if $A: R^n \rightarrow R^n$ is a linear map, and \mathcal{R} is A -invariant, we write $A|_{\mathcal{R}}$ to indicate the restriction of A to the subdomain \mathcal{R} . The dimension of a subspace \mathcal{R} is given by $\dim \mathcal{R}$.

The space of polynomials with coefficients in R^n is denoted by $R^n[\lambda]$; note that this is an $R[\lambda]$ -module. The degree of a polynomial $u(\lambda)$ is indicated by $\deg u(\lambda)$. We let \underline{m} denote the set of integers $\{1, 2, \dots, m\}$, and for an ordered set of scalars $\{v_1, \dots, v_k\}$ we let

$$v_j^* = \sum_{i \leq j} v_i, \quad j \in \underline{k}.$$

If $\{\cdot\}$ is a set of vectors, we refer to the subspace spanned by the elements as $\text{span } \{\cdot\}$, and we indicate the j th standard unit vector (1 in the j th position, zeros elsewhere) by e_j . Finally the transpose of A is given by A^T .

If $A: \mathcal{X} \rightarrow \mathcal{X}$, $\mathcal{B} \subset \mathcal{X}$, and $\dim \mathcal{X} = n$, then $\{A|_{\mathcal{B}}\} = \mathcal{B} + A\mathcal{B} + \dots + A^{n-1}\mathcal{B}$, i.e., the space reachable under the action of A from inputs to \mathcal{B} . We say that (A, B) is a controllable pair if $\{A|_{\mathcal{B}}\} = \mathcal{X}$. Following [10], a subspace $\mathcal{V} \subset \mathcal{X}$ is (A, B) -invariant if there exists a map $F: R^n \rightarrow R^m$ such that $(A + BF)\mathcal{V} \subset \mathcal{V}$. A subspace \mathcal{R} is a c.s. of (A, B) if \mathcal{R} is (A, B) -invariant and $\mathcal{R} = \{A + BF|_{\mathcal{B}} \cap \mathcal{R}\}$ for some F ; that is, every element in \mathcal{R} is reachable under the action of $A + BF$ from inputs to $\mathcal{B} \cap \mathcal{R}$.

2. Some preliminary results. In this section we show that the concept of a controllability subspace for a pair (A, B) has an analog in terms of the kernel of a particular polynomial matrix. Furthermore, certain invariants of this polynomial matrix are shown to lead quite naturally to the canonical form for controllable pairs developed in [1], [6], [7], [8] and [9].

LEMMA 1. *If \mathcal{R} is a c.s., then for every nonzero $b \in \mathcal{B} \cap \mathcal{R}$, there exists a matrix F such that*

$$(1) \quad \{A + BF|_{\mathcal{R}}\} = \mathcal{R}.$$

Moreover, F can be chosen to satisfy both (1) and

$$(2) \quad (A + BF)^r b = 0,$$

where $r = \dim \mathcal{R}$.

Proof. This is a special case of Theorem 4.2 in [10]. We are choosing F so that \mathcal{R} is cyclic with respect to $A + BF$, with generator b , and so that $(A + BF)|_{\mathcal{R}}$ is nilpotent.

This last result leads to an interesting characterization of controllability subspaces.

LEMMA 2. A subspace $\mathcal{R} \subset R^n$ of dimension r is a c.s. if and only if there exist $x(\lambda) \in R^n[\lambda]$ and $u(\lambda) \in R^m[\lambda]$ such that

- (i) $\deg u(\lambda) = k$ and $\deg x(\lambda) = k - 1$, for some $k \geq r$;
- (ii) $(\lambda I - A)x(\lambda) = Bu(\lambda)$;
- (iii) If $x(\lambda) = \sum_{i \in \underline{k}} \lambda^{i-1} x_{i-1}$, then $\mathcal{R} = \text{span} \{x_{i-1}, i \in \underline{k}\}$.

Proof. Necessity. Suppose \mathcal{R} is a c.s. of dimension r . Let $b \in \mathcal{B} \cap \mathcal{R}$ and F be chosen to satisfy (1) and (2). Define

$$u(\lambda) = \sum_{i=0}^r \lambda^i u_i \in R^m[\lambda] \quad \text{and} \quad x(\lambda) = \sum_{i=0}^{r-1} \lambda^i x_i$$

so that

$$\begin{aligned} Bu_r &= b, \\ u_i &= F(A + BF)^{r-i-1}b, \quad 0 \leq i \leq r-1, \\ x_i &= (A + BF)^{r-i-1}b, \quad 0 \leq i \leq r-1. \end{aligned}$$

Then (i) is trivially satisfied; (ii) follows by comparing coefficients of powers of λ , and by using (2); (iii) follows from (1).

Sufficiency. Let $u(\lambda) \in R^m[\lambda]$ and $x(\lambda) \in R^n[\lambda]$ satisfy (i)–(iii). We shall demonstrate that

$$(3) \quad A\mathcal{R} \subset \mathcal{R} + \mathcal{B}$$

and that

$$(4) \quad \mathcal{R} = \mathcal{W}_k,$$

where $\mathcal{W}_0 = 0$, and $\mathcal{W}_i = (A\mathcal{W}_{i-1} + \mathcal{B}) \cap \mathcal{R}$ for $i \in \underline{k}$.

The result will then follow from Theorem 4.1 in [10].

From (ii) it is easily seen that $Ax_i = x_{i-1} - Bu_i$ for $1 \leq i \leq k-1$, and that $Ax_0 = -Bu_0$; thus (3) follows from (iii).

To demonstrate (4), define subspaces \mathcal{S}_i as

$$\mathcal{S}_i = \text{span} \{x_{k-1}, x_{k-2}, \dots, x_{k-i}\} \quad \text{for } i \in \underline{k}.$$

Clearly, $\mathcal{S}_1 \subset \mathcal{W}_1$; moreover, from (ii) it is easily seen that $\mathcal{S}_i \subset (A\mathcal{S}_{i-1} + \mathcal{B}) \cap \mathcal{R}$ for $2 \leq i \leq k$, whence it follows inductively that $\mathcal{S}_i \subset \mathcal{W}_i$ for all $i \in \underline{k}$. But clearly, $\mathcal{S}_k = \mathcal{R}$, and (4) follows.

Remark 1. If a pair $(x(\lambda), u(\lambda))$ can be found which satisfies conditions (i)–(iii) of Lemma 2, and satisfies the additional condition that the coefficients in $x(\lambda)$ are independent, then one can find a matrix F such that $Fx_{i-1} = u_{i-1}$ for all $i \in \underline{k}$. It then follows that $x_{k-1} \in \mathcal{B} \cap \mathcal{R}$ is a cyclic generator for \mathcal{R} with respect to the matrix $A + BF$.

We have thus established a characterization of controllability subspaces in terms of elements of $\ker(\lambda I - A; -B)$, where the matrix $(\lambda I - A; -B)$ is to be interpreted as representing an $R[\lambda]$ -module morphism: $R^{m+n}[\lambda] \rightarrow R^n[\lambda]$. Elements in $\ker(\lambda I - A; -B)$ may in turn be characterized by the *minimal column*

indices $\{v_i, i \in \underline{m}\}$ and a *fundamental series* $\{z_i(\lambda), i \in \underline{m}\}$ associated with the *singular pencil of matrices* $(\lambda I - A; -B)$. These two sets are determined as follows (see [5, Chap. 12]):

(i) Let v_1 be the least degree of all nonzero elements of $\ker(\lambda I - A; -B)$, and choose $z_1(\lambda) \in \ker(\lambda I - A; -B)$ so that $\deg z_1(\lambda) = v_1$;

(ii) For each i , $1 \leq i \leq m-1$, after having chosen $\{z_j(\lambda), j \in \underline{i}\}$ we define v_{i+1} to be the least degree of all elements $z(\lambda) \in \ker(\lambda I - A; -B)$ such that $z(\lambda)$ is not an element of the submodule generated by the set $\{z_j(\lambda), j \in \underline{i}\}$. Then choose $z_{i+1}(\lambda) \in \ker(\lambda I - A; -B)$ so that $\deg z_{i+1}(\lambda) = v_{i+1}$ and so that $z_{i+1}(\lambda)$ is not an element of the submodule generated by $\{z_j(\lambda), j \in \underline{i}\}$.

We shall call the set $\{v_i, i \in \underline{m}\}$, so obtained, the set of *Kronecker invariants* of the pair (A, B) . Note that by the construction of this set, the v_i 's are ordered as $0 \leq v_1 \leq v_2 \leq \dots \leq v_m$. The sets $\{v_i, i \in \underline{m}\}$ and $\{z_i(\lambda), i \in \underline{m}\}$ enjoy other properties, which we now state.

PROPOSITION 1. *Let $(A, B) \in R^{n \times n} \times R^{n \times m}$ be a controllable pair such that $\text{rank } B = m$. Then the Kronecker invariants and the fundamental series, as determined above, satisfy:*

- (i) *The set $\{v_i, i \in \underline{m}\}$ is well-defined and unique;*
- (ii) *$v_i > 0$, all $i \in \underline{m}$;*
- (iii) *$\sum_{i \in \underline{m}} v_i = n$;*
- (iv) *$\{z_i(\lambda), i \in \underline{m}\}$ is a set of free generators for $\ker(\lambda I - A; -B)$, and any $z(\lambda) \in \ker(\lambda I - A; -B)$ can be uniquely written as*

$$z(\lambda) = \sum_{i: v_i \leq \deg z(\lambda)} z_i(\lambda) \alpha_i(\lambda)$$

for appropriate $\alpha_i(\lambda) \in R[\lambda]$ such that $\deg \alpha_i(\lambda) \leq \deg z(\lambda) - v_i$;

(v) *The fundamental series $\{z_i(\lambda), i \in \underline{m}\}$ is not uniquely determined; however, for each i such that $v_i < v_{i+1}$, the submodule $\mathcal{M}_i \triangleq$ (submodule generated by $\{z_j(\lambda), j \in \underline{i}\}$) is invariant with respect to the choice of fundamental series;*

(vi) *If each $z_i(\lambda)$ is partitioned as $z_i(\lambda) = (s_i^T(\lambda); t_i^T(\lambda))^T$, where $t_i(\lambda) \in R^m[\lambda]$ and $s_i(\lambda) \in R^n[\lambda]$, then $\deg s_i(\lambda) = v_i - 1$ and the collections of coefficients $\{t_{i,v_i}, i \in \underline{m}\}$ and $\{s_{ij}; 0 \leq j \leq v_i - 1, i \in \underline{m}\}$ are bases for R^m and R^n .*

Proof. See references [1], [6], or [7] for (i)–(iii); [2] and [4] for (iv); [5] for (v); and [2, Thm. (4.5–22)] or [9] for (vi).

Now consider the pair $(s_i(\lambda), t_i(\lambda))$, as determined by $z_i(\lambda)$. This pair of polynomial vectors satisfies

$$(\lambda I - A)s_i(\lambda) = Bt_i(\lambda).$$

Thus, from statement (vi) of Proposition 1 and from Remark 1, it follows that $\text{span}\{s_{ij}, 0 \leq j \leq v_i - 1\}$ is a c.s. generated by s_{i,v_i-1} . We call this c.s. \mathcal{R}_i :

$$(5) \quad \mathcal{R}_i \triangleq \text{span}\{s_{ij}, 0 \leq j \leq v_i - 1\} \quad \text{for } i \in \underline{m}.$$

Since $\{s_{ij}; 0 \leq j \leq v_i - 1, i \in \underline{m}\}$ is a basis for R^n , it is clear that

$$(6) \quad R^n = \mathcal{R}_1 \oplus \dots \oplus \mathcal{R}_m.$$

However, because the fundamental series $\{z_i(\lambda), i \in \underline{m}\}$ is not unique, the decomposition of R^n via (6) is not unique. In spite of this fact, there are certain

properties of the decomposition (6) which are invariant with respect to the choice of fundamental series $\{z_i(\lambda), i \in \underline{m}\}$.

PROPOSITION 2. *The subspaces $\mathcal{V}_i = \mathcal{R}_1 \oplus \cdots \oplus \mathcal{R}_i$ for which $v_i < v_{i+1}$ are invariant with respect to the choice of fundamental series $\{z_i(\lambda), i \in \underline{m}\}$.*

Proof. From Proposition 1, (iv)–(v), it is easily seen that when $v_i < v_{i+1}$, $\text{span}\{s_{kj}; 0 \leq j \leq v_k - 1, k \in \underline{i}\}$ is invariant with respect to the choice of fundamental series $\{z_i(\lambda), i \in \underline{m}\}$. This proves the proposition.

PROPOSITION 3. *The subspace $\mathcal{V}_i = \mathcal{R}_1 \oplus \cdots \oplus \mathcal{R}_i$ for which $v_i < v_{i+1}$ is the maximal c.s. contained in the subspace $A^{-v_i}(\mathcal{B} + \cdots + A^{v_i-1}\mathcal{B})$.*

Proof. \mathcal{V}_i is obviously a c.s. and is contained in

$$\mathcal{S}_{v_i} = A^{-v_i}(\mathcal{B} + \cdots + A^{v_i-1}\mathcal{B}).$$

Let \mathcal{V} be a c.s. larger than \mathcal{V}_i . Let P_i denote the projection on \mathcal{R}_i and along $\bigoplus_{j \neq i} \mathcal{R}_j$. Then clearly $P_j \mathcal{V} \neq 0$ for some $j > i$, and since \mathcal{V} is a c.s. we must have $P_j \mathcal{V} = \mathcal{R}_j$. Hence there exists an $x \in \mathcal{V}$ such that $x = \sum_{k,r} \alpha_{k,r} s_{k,r}$ with $\alpha_{j,v_j-1} \neq 0$. But $s_{j,v_j-1} \notin \mathcal{S}_{v_i}$ for $j > i$, and as the $s_{i,j}$ are a basis, clearly $x \notin \mathcal{S}_{v_i}$, proving the assertion.

Finally, from the fundamental series $\{z_i(\lambda) = (s_i^T(\lambda); t_i^T(\lambda))^T, i \in \underline{m}\}$ we can determine a feedback matrix F such that the fundamental series associated with the controllable pair $(A + BF, B)$ is of a particularly simple form; this will then lead to a “canonical” form for (A, B) . We define F as follows. Since $\{s_{ij}\}$ is a basis for R^n , there exists a matrix F such that

$$Fs_{i,j} = t_{i,j}, \quad 0 \leq j \leq v_i - 1, \quad i \in \underline{m}.$$

It now follows easily that

$$(\lambda I - A - BF)s_i(\lambda) = B\lambda^{v_i}t_{i,v_i} \quad \text{for each } i \in \underline{m}.$$

This last relation completely specifies the maps $A + BF: R^n \rightarrow R^n$ and $B: R^m \rightarrow R^n$ with respect to the bases $\{s_{i,j}\}$ (in R^n) and $\{t_{i,v_i}\}$ (in R^m). That is,

$$Bt_{i,v_i} = s_{i,v_i-1}, \quad i \in \underline{m},$$

while

$$(A + BF)s_{i,j} = \begin{cases} s_{i,j-1} & \text{if } 1 \leq j \leq v_i - 1, \quad i \in \underline{m}, \\ 0 & \text{if } j = 0, \quad i \in \underline{m}. \end{cases}$$

Thus, with nonsingular matrices S and G defined as

$$S = (s_{1,0}; s_{1,1}; \cdots; s_{1,v_1-1}; s_{2,0}; \cdots; s_{m,v_m-1})$$

and

$$G = (t_{1,v_1}; \cdots; t_{m,v_m}),$$

it follows that

$$S^{-1}BG = (e_{v_1^*}; e_{v_2^*}; \cdots; e_{v_m^*}),$$

$$S^{-1}(A + BF)S = \text{block diagonal } (H_{v_1}; \cdots; H_{v_m}),$$

where H_k is $k \times k$ with ones on the superdiagonal and zeros elsewhere.

We shall refer to the pair $(S^{-1}(A + BF)S, S^{-1}BG)$ as the Brunovsky canonical form for the pair (A, B) . In the sequel it will be convenient to work with

this canonical form. We note that the fundamental series associated with $(S^{-1}(A + BF)S, S^{-1}BG)$ is

$$\{z_i(\lambda) = (s_i^T(\lambda); t_i^T(\lambda))^T, i \in \underline{m}\}$$

with

$$t_i(\lambda) = \lambda^{v_i} \hat{e}_i, \quad i \in \underline{m},$$

$$s_i(\lambda) = \lambda^{v_i-1} e_{v_i^*} + \cdots + e_{v_i^*-v_i+1}, \quad i \in \underline{m},$$

where \hat{e}_i is the i th standard unit vector in R^m . For this choice of fundamental series the subspaces $\mathcal{R}_i \subset R^n$ are given as

$$\mathcal{R}_i = \text{span} \{e_{v_i^*}, \cdots, e_{v_i^*-v_i+1}\}, \quad i \in \underline{m}.$$

3. Dimensions of controllability subspaces. We consider a controllable pair (A, B) in the Brunovsky canonical form. As indicated in the previous section, this form is compatible with a natural direct sum decomposition of the state space into controllability subspaces $\mathcal{R}_i, i \in \underline{m}$, with

$$\mathcal{R}_j = \text{span} \{e_{v_j^*-v_j+1}, \cdots, e_{v_j^*}\},$$

where e_j is the j th unit vector in the canonical basis for R^n . Denote the projection on \mathcal{R}_i and along $\bigoplus_{j \neq i} \mathcal{R}_j$ by $P_i, i \in \underline{m}$, and the set $\{j \in \underline{m} | P_j \mathcal{R} \neq 0\}$ by $M(\mathcal{R})$ for any subspace \mathcal{R} . Then we have the following bound on the dimension of a c.s.

LEMMA 3. *Let \mathcal{R} be a c.s. of the pair (A, B) . Then $\dim \mathcal{R} \geq \max \{v_j | j \in M(\mathcal{R})\}$.*

Proof. Since \mathcal{R} is a c.s., $\mathcal{R} = \{A + BF | \mathcal{B} \cap \mathcal{R}\}$ for some appropriate F . Further, we know that \mathcal{R} may be singly generated by a $b \in \mathcal{B} \cap \mathcal{R}$. Pick such a $b = \sum_{i \in \underline{m}} \alpha_i e_{v_i^*}$, since the $e_{v_i^*}$ are a basis for \mathcal{B} . By the particular form of (A, B) it is obvious that the c.s. \mathcal{R} is spanned by the columns of the matrix

$$W = \begin{pmatrix} \cdot & \cdot & \cdot & \alpha_1 & \cdot & \cdot & \cdot \\ \cdot & & \alpha_1 & \beta_1 & \cdot & \cdot & \cdot \\ \cdot & \alpha_1 & \beta_1 & \gamma_1 & \cdot & \cdot & \cdot \\ \alpha_1 & \beta_1 & \gamma_1 & \delta_1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \alpha_m & \cdot & \cdot & \cdot \\ & & \alpha_m & \beta_m & \cdot & \cdot & \cdot \\ & \alpha_m & \beta_m & \gamma_m & \cdot & \cdot & \cdot \\ \alpha_m & \beta_m & \gamma_m & \delta_m & \cdot & \cdot & \cdot \end{pmatrix},$$

where only possibly nonzero elements have been shown. It is obvious that $P_i \mathcal{R} = 0$ iff the $v_i^* - v_i + 1$ through v_i^* rows of W are identically zero. Thus if $P_i \mathcal{R} \neq 0$, W must contain a nonsingular lower triangular submatrix of dimension $v_i \times v_i$, hence must have column rank $\geq v_i$, which proves the lemma.

Remark 2. This lemma is the state space analog of Proposition 1 (iv).

The dimension of a c.s. may be similarly bounded from above.

LEMMA 4. Let \mathcal{R} be a c.s. of the pair (A, B) . Then

$$\dim \mathcal{R} \leq \sum_{j \in M(\mathcal{R})} v_j.$$

Proof. Clearly $P_i \mathcal{R} = 0$ for $i \notin M(\mathcal{R})$. Then $(\sum_{i \notin M(\mathcal{R})} P_i) \mathcal{R} = 0$ or equivalently $\mathcal{R} \subset \ker(\sum_{i \notin M(\mathcal{R})} P_i) = \sum_{j \in M(\mathcal{R})} \mathcal{R}_j$. Since the \mathcal{R}_j 's are independent, this latter subspace has dimension $\sum_{j \in M(\mathcal{R})} v_j$ and contains \mathcal{R} , proving the lemma.

THEOREM 1. Let $V = \{v_1, \dots, v_m\}$ be the set of Kronecker invariants of the controllable pair (A, B) . Then there exists a c.s. \mathcal{R} of dimension p iff

$$(7) \quad \max \{v_i | v_i \in U\} \leq p \leq \sum_{v_i \in U} v_i$$

for some subset U of V .

Proof. Necessity. Let $U = \{v_i | i \in M(\mathcal{R})\}$. Then the result is immediate from the preceding two lemmas.

Sufficiency. Given a subset $U \subset V$ and a p satisfying (7) we shall construct a c.s. \mathcal{R} of dimension p by summing the c.s. \mathcal{R}_i possibly with some "overlap". First order the elements of U in decreasing size $(v_{i_1}, \dots, v_{i_k})$ and define s as the smallest integer such that $\eta_s = \sum_{j \leq s} v_{i_j}$ is greater than or equal to p . If $p = \eta_s$ then we may construct a c.s. of dimension p by forming the direct sum of c.s. $\mathcal{R}_{i_1} \oplus \dots \oplus \mathcal{R}_{i_s}$. If η_s exceeds p , then for $s > 2$, we shall construct a c.s. of the form $\mathcal{R}_{i_1} \oplus \dots \oplus \mathcal{R}_{i_{s-2}} \oplus \mathcal{Q}$, where \mathcal{Q} is a c.s. of dimension $p - \eta_{s-2}$ obtained by "overlapping" the c.s. $\mathcal{R}_{i_{s-1}}$ and \mathcal{R}_{i_s} . Finally if $\eta_s > p$ and $s = 2$, then we shall construct a c.s. of the form \mathcal{Q} above. Hence it suffices to consider only the cases $p = \eta_s$ for some s , or $\eta_1 < p < \eta_2$.

To prove the first case we need only show how to form the direct sum of two c.s., say $\mathcal{R}_i \oplus \mathcal{R}_j$. Consider the feedback which changes the $(v_j^*, v_i^* - v_i + 1)$ -element of A from zero to one. With this feedback, $(A + BF)^{v_i} e_{v_i^*} = e_{v_j^*}$; that is the c.s. $\mathcal{R}_i \oplus \mathcal{R}_j$ is generated by the generator of \mathcal{R}_i .

To prove the second case let $p < v_i + v_j$. By choosing feedback such that the $(v_j^*, v_i^* - (p - v_j - 1))$ -element of A is changed from zero to one, it is easily seen that the resulting c.s. generated by $e_{v_i^*}$ is spanned by the set of p independent vectors

$$\{e_{v_i^*}, e_{v_i^*-1}, \dots, e_{v_i^*-p+v_j} + e_{v_j^*}, \dots, e_{v_i^*-v_i+1} + e_{v_j^*-v_i-v_j+p+1}, \\ e_{v_j^*-v_i-v_j+p}, \dots, e_{v_j^*-v_j+1}\}$$

and hence is of dimension p .

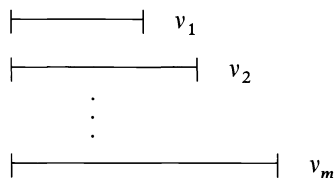
Remark 3. The constructions used to prove sufficiency have an analog in the context of Lemma 2. Let $[s_i^T(\lambda); t_i^T(\lambda)]^T$ be the i th free generator of $\ker[\lambda I - A; -B]$, that is, \mathcal{R}_i is the span of the coefficients of $s_i(\lambda)$. Then for any $k \geq 0$ it is easily seen that

$$(\lambda I - A)(\lambda^k s_i(\lambda) + s_j(\lambda)) = B(\lambda^k t_i(\lambda) + t_j(\lambda)).$$

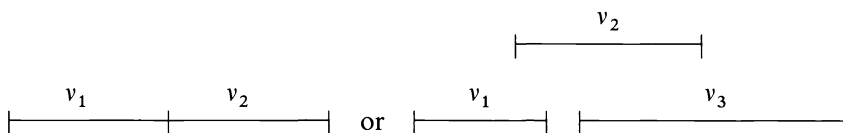
When $k = v_j$, the spans of the coefficients of $\lambda^{v_j} s_i(\lambda) + s_j(\lambda)$ yield the c.s. $\mathcal{R}_i \oplus \mathcal{R}_j$. For $0 < k < v_j \leq v_i$ the span of the coefficients of $\lambda^k s_i(\lambda) + s_j(\lambda)$ is a c.s. of dimension $v_i + k$ of the form \mathcal{Q} in the proof of Theorem 1. It should be clear that these constructions are not unique. We may replace λ^k by $\alpha(\lambda)$, any polynomial

of degree k , and achieve controllability subspaces, albeit possibly different ones, of the appropriate dimensions.

Theorem 1 indicates that the possible dimensions of controllability subspaces of a pair (A, B) are directly determined by the Kronecker invariants of (A, B) . If one considers the Kronecker invariants to be represented by line segments of length v_i



then the theorem states that the corresponding dimensions of possible c.s. are given by the lengths of line segments obtained by joining together some of the above line segments, with the possibility of integral overlap, for example,



etc.

Furthermore we have the following corollary.

COROLLARY 1. *If for some $j \in \underline{m-1}$, $v_{j+1} > v_j^* + 1$, then there exists no c.s. of dimension p , where p is any integer satisfying $v_j^* < p < v_{j+1}$.*

Proof. This follows directly from Theorem 1, as any subset of the set $U = \{v_i | i \leq j\}$ clearly fails the upper bound in (7), while if U includes any elements v_k for $k > j$, it likewise fails the lower bound in (7).

For example, if the Kronecker invariants of (A, B) are $(1, 2, 5)$, then c.s. of dimensions 1, 2, 3, 5, 6, 7, 8 exist, whereas no c.s. of dimension 4 exists.

The constructions of Theorem 1 and Remark 3 are suggestive of the uniqueness of c.s. of certain dimensions. Indeed, we have the following.

COROLLARY 2. *Let \mathcal{R} be a c.s. of dimension p . Then \mathcal{R} is the unique c.s. of dimension p iff $p = v_j^* < v_{j+1}$ for some $j \in \underline{m-1}$. In particular, if $v_2 \neq v_1$, then \mathcal{R}_1 is the unique c.s. of dimension v_1 .*

Proof. Sufficiency. Assume $\dim \mathcal{R} = p = v_j^* < v_{j+1}$. By Lemma 3, $P_i \mathcal{R} = 0$ for $i > j$, so $\mathcal{R} \subset \bigcap_{i > j} \ker P_i = \bigoplus_{k \leq j} \mathcal{R}_k$. Since $\dim \bigoplus_{k \leq j} \mathcal{R}_k = v_j^*$ we must have $\mathcal{R} = \bigoplus_{k \leq j} \mathcal{R}_k$. However, regardless of the nonuniqueness of the set of c.s. \mathcal{R}_k , the subspace $\bigoplus_{k \leq j} \mathcal{R}_k$ is unique by Proposition 2, since $v_{j+1} > v_j$.

Necessity. We will show that if there exists any c.s. of dimension p such that either $p \neq v_j^*$ or $p = v_j^* < v_{j+1}$ for any $j \in \underline{m}$, then there are, in general, many different c.s. of identical dimension.

Consider first the case $p \neq v_j^*$ for any $j \in \underline{m}$. Assume there exists a c.s. of dimension p . Define s to be the largest integer such that $v_s^* < p$, and let $q = p - v_s^*$. Then by Remark 3 following Theorem 1 it is clear that given the polynomial vector

$$u_s(\lambda) = \lambda^{v_s^*-1+q} t_s(\lambda) + \lambda^{v_s^*-2+q} t_{s-1}(\lambda) + \cdots + \lambda^q t_1(\lambda) + \alpha t_{s+1}(\lambda)$$

for any nonzero $\alpha \in R$, the span of the coefficients of the corresponding $x_\alpha(\lambda)$ is a c.s. of dimension p . Further, since $q < v_{s+1}$, it is clear that for $\alpha \neq \beta$, both nonzero, the spans of the coefficients of $x_\alpha(\lambda)$ and $x_\beta(\lambda)$ differ.

Now assume $p = v_j^*$ for some $j \in \underline{m}$, but $v_j^* \geq v_{j+1}$. Clearly the polynomial vector

$$u_0(\lambda) = \lambda^{v_j^*-1}t_j(\lambda) + \lambda^{v_j^*-2}t_{j-1}(\lambda) + \cdots + t_1(\lambda)$$

has an associated $x_0(\lambda)$, the span of whose coefficients is given by $\mathcal{R}_1 \oplus \cdots \oplus \mathcal{R}_j$, a c.s. of dimension p . However, the polynomial vector

$$u_\alpha(\lambda) = \lambda^{v_j^*-1}t_j(\lambda) + \lambda^{v_j^*-2}t_{j-1}(\lambda) + \cdots + t_1(\lambda) + \alpha t_{j+1}(\lambda),$$

where $\alpha \in R$, has an associated $x_\alpha(\lambda)$ whose coefficients span a different c.s. of dimension p . Further, if $\alpha \neq \beta$, then the c.s. associated with $u_\alpha(\lambda)$ differs from that associated with $u_\beta(\lambda)$.

Note that we have shown that for systems defined over the reals (or any other uncountably infinite field), if there exists more than one c.s. of a given dimension, then there exists an uncountable number. In the example with Kronecker invariants, 1, 2 and 5, the c.s. of dimensions 1, 3 and 8 are unique, while there are nondenumerably many c.s. of dimensions 2, 5, 6 and 7.

4. Minimal dimension controllability subspaces. In §2 a characterization of controllability subspaces in terms of the free generators of the kernel of the singular pencil of matrices $[\lambda I - A; -B]$ was developed. Using this representation, requirements on the possible dimensions of c.s. were derived in §3. In this section we wish to explore c.s. constrained to contain, or cover, a given subspace; in particular we will construct minimal dimension c.s. covering subspaces of \mathcal{B} .

Consider an element $b \in \mathcal{B}$. If \mathcal{R} is a c.s. containing b , then by Lemma 1, there exists a feedback map F such that

$$\mathcal{R} = \text{span} \{b, (A + BF)b, \cdots, (A + BF)^{n-1}b\},$$

and $(A + BF)^nb = 0$. Combining this fact with the characterization of c.s. in terms of the pencil of matrices, we may view any c.s. \mathcal{R} containing b as the span of a trajectory generated by driving b to zero. Clearly then, the minimal dimension c.s. containing b are in 1-1 correspondence with the spans of the trajectories arising from driving b to zero in a minimal number of steps, i.e., the spans of trajectories $\{b, (A + BF)b, \cdots\}$ that contain a minimal number of nonzero vectors.

It should be noted that driving a vector x to zero in r steps implies the construction of an input string $\{u_{r-1}, \cdots, u_0\}$ such that if

$$x_{r-1} = x \quad \text{and} \quad x_{r-i-1} = Ax_{r-1} + Bu_{r-i}, \quad 1 \leq i \leq r,$$

then $x_{-1} = 0$. If $x \in \mathcal{B}$, this is of course equivalent to finding $u(\lambda) = \sum_{i=0}^r \lambda^i u_i$ and $x(\lambda) = \sum_{i=0}^{r-1} \lambda^i x_i$ such that $Bu_r = x$ and $(\lambda I - A)x(\lambda) = Bu(\lambda)$. It surely suffices to find a feedback map F such that if

$$x_{r-1} = x \quad \text{and} \quad x_{r-i-1} = (A + BF)x_{r-i}, \quad 1 \leq i \leq r,$$

then $x_{-1} = 0$.

If we wish to drive an element $b \in \mathcal{B}$ to zero in a minimal number of steps, it seems natural that the span of the trajectory excluding b should be independent of \mathcal{B} . This is indeed the case.

LEMMA 5. *Let $b \in \mathcal{B}$. If there exists a feedback map F and a trajectory $\{b, (A + BF)b, \dots, (A + BF)^{n-1}b\}$ such that $\sum_{i=1}^r \alpha_i (A + BF)^i b$, $\alpha_r \neq 0$, is an element of \mathcal{B} for $1 \leq r \leq n - 1$, then b may be driven to zero in r or fewer steps.*

Proof. We write $\hat{b} = \sum_{i=1}^r \alpha_i (A + BF)^i b$ and assume without loss of generality that $\alpha_r = 1$. Since B has full rank there exist unique elements \hat{u} and u such that $\hat{u} = B^{-1}\hat{b}$, $u = B^{-1}b$. Consider the input string u_i , $r - 1 \geq i \geq 0$, defined by

$$\begin{aligned} u_{r-1} &= Fb + \alpha_{r-1}u, \\ u_{r-2} &= F(A + BF)b + \alpha_{r-1}Fb + \alpha_{r-2}u, \\ &\vdots \\ u_0 &= F(A + BF)^{r-1}b + \alpha_{r-1}F(A + BF)^{r-2}b + \dots + \alpha_1 Fb - \hat{u}. \end{aligned}$$

Now let $x_{r-1} = b$, and consider the sequence generated by the recursion

$$x_{r-i-1} = Ax_{r-i} + Bu_{r-i}, \quad 1 \leq i \leq r.$$

Then it follows that

$$x_{-1} = (A + BF)^r b + \alpha_{r-1}(A + BF)^{r-1}b + \dots + \alpha_1(A + BF)b - \hat{b} = 0,$$

and hence b may be driven to zero in r steps.

Note that Lemma 5 implies that for $b \in \mathcal{B}$, if \mathcal{R} is a c.s. of minimum dimension containing b , then $\mathcal{R} \cap \mathcal{B} = \mathcal{B}$. It is now natural to ask: What is the minimum dimension of a c.s. covering an element $x \in \mathcal{X}$? For $x \in \mathcal{B}$, we can easily answer this query.

Recall that $M(\mathcal{R})$ was defined as the set $\{j \in \underline{m} | P_j \mathcal{R} \neq \emptyset\}$ for any subspace \mathcal{R} .

LEMMA 6. *Let $b \in \mathcal{B}$. Then the minimum dimension of a c.s. containing b is given by $\mu = \max \{v_j | j \in M(\mathcal{B})\}$.*

Proof. If \mathcal{R} is a c.s. containing b , then by Lemma 3, $\dim \mathcal{R} \geq \mu$. Now consider the trajectory $(b, Ab, \dots, A^{n-1}b)$, where A, B are assumed in the Brunovsky canonical form. Since $b = \sum_{j \in M(\mathcal{B})} \gamma_j e_{v_j^*}$ and $A^{v_j} e_{v_j^*} = 0$, it follows that $A^\mu b = 0$, yielding a covering c.s. of dimension μ .

We now can turn our attention to the case where we desire to cover an arbitrary subspace of \mathcal{B} . As we shall need a minor construction, we first prove a lemma to motivate that construction.

LEMMA 7. *Let b_1 and b_2 be elements of \mathcal{B} such that $M(\mathcal{L}_1) \cap M(\mathcal{L}_2) = \emptyset$, and denote $\max \{v_j | j \in M(\mathcal{L}_i)\}$ by μ_i , $i \in \underline{2}$. Then if \mathcal{R} is a c.s. covering b_1 and b_2 , $\dim \mathcal{R} \geq \mu_1 + \mu_2$.*

Proof. If \mathcal{R} contains b_1 and b_2 , then \mathcal{R} contains c.s. which may be generated by b_1 and b_2 respectively (by Lemma 1). Then it follows that for some F_1, F_2 ,

$$\text{span} \{b_1, (A + BF_1)b_1, \dots, (A + BF_1)^{n-1}b_1, b_2, \dots, (A + BF_2)^{n-1}b_2\} \subset \mathcal{R}.$$

Recalling that $\{s_{i, v_i-1}, i \in \underline{m}\}$ is a basis for \mathcal{B} , we may write

$$b_1 = \sum_{j \in M(\mathcal{L}_1)} \gamma_{1j} s_{j, v_j-1} \quad \text{and} \quad b_2 = \sum_{j \in M(\mathcal{L}_2)} \gamma_{2j} s_{j, v_j-1}.$$

Since the lemma obviously follows for $\mu_1 = \mu_2 = 1$, we may assume that for some i , $\mu_i > 1$. Now for $\mu_i > 1$ we have

$$(8) \quad (A + BF_i)b_i = \sum_{j \in M(\ell_i)} \gamma_{ij} s_{j, v_j - 2} + \sum_{k \in \underline{m}} \alpha_{ik1} s_{k, v_k - 1} \neq 0$$

for some α_{ik1} , $k \in \underline{m}$. Note that the second term on the right is an element of \mathcal{B} and represents the arbitrary nature of the feedback map F_i . Continuing, we have for $\mu_i > r$,

$$(9) \quad (A + BF_i)^r b_i = \sum_{j \in M(\ell_i)} \gamma_{ij} s_{j, v_j - r - 1} + \sum_{p \in \underline{r}} \sum_{k \in \underline{m}} \alpha_{ikp} s_{k, v_k - (r + 1 - p)} \neq 0$$

for some α_{ikp} , $k \in \underline{m}$, $p \in \underline{r}$, where $s_{i,j} \triangleq 0$ for $j < 0$. Comparing the forms of elements from (8) and (9), it follows from the hypothesis $M(\ell_1) \cap M(\ell_2) = \emptyset$ and the fact that $\{s_{i,j}; 0 \leq j \leq v_i - 1, i \in \underline{m}\}$ is a basis for R^n , that the vectors

$$\{b_1, (A + BF_1)b_1, \dots, (A + BF_1)^{\mu_1 - 1} b_1, b_2, \dots, (A + BF_2)^{\mu_2 - 1} b_2\}$$

are independent, and hence \mathcal{R} is of dimension $\geq \mu_1 + \mu_2$.

COROLLARY 3. Let b_1 and b_2 be elements of \mathcal{B} such that $M(\ell_1) \cap M(\ell_2) = \emptyset$. If \mathcal{V}_1 and \mathcal{V}_2 are minimal dimension covering c.s. for b_1 and b_2 respectively, then $\mathcal{V}_1 \cap \mathcal{V}_2 = 0$ and $\mathcal{V}_1 \oplus \mathcal{V}_2$ is a minimal dimension c.s. covering $\text{span}\{b_1, b_2\}$.

Proof. The proof follows immediately from Lemmas 6 and 7.

THEOREM 2. Let $\mathcal{D} \subset \mathcal{B}$ and $\{b_1, \dots, b_k\}$ be a basis for \mathcal{D} such that $M(\ell_i) \cap M(\ell_j) = \emptyset$ for $i \neq j$, $i, j \in \underline{k}$. Then if \mathcal{R} is a c.s. covering \mathcal{D} , then $\dim \mathcal{R} \geq \sum_{i \in \underline{k}} \mu_i$. Furthermore, if \mathcal{V}_i is a minimal dimension c.s. covering b_i , $i \in \underline{k}$, then $\{\mathcal{V}_i, i \in \underline{k}\}$ is an independent set of subspaces, and $\mathcal{V}_1 \oplus \dots \oplus \mathcal{V}_k$ is a minimal dimension c.s. containing \mathcal{D} .

Proof. First we note that any $\mathcal{D} \subset \mathcal{B}$ has such a basis. Let $\{d_1, \dots, d_k\}$ be any basis for \mathcal{D} and let D be a matrix whose columns are given by the d_i , $i \in \underline{k}$, with respect to the basis for \mathcal{B} , $\{s_{i, v_i - 1}; i \in \underline{m}\}$. Then by applying elementary column operations, it is possible to transform D to a matrix D_0 whose columns are a basis for D and have the desired property (only one nonzero entry per row).

By expanding the argument of Lemma 7 to the case where for appropriate F_1, \dots, F_k ,

$$\text{span}\{b_1, \dots, (A + BF_1)^{n-1} b_1, \dots, b_k, \dots, (A + BF_k)^{n-1} b_k\} \subset \mathcal{R},$$

it is straightforward to show that \mathcal{R} contains a subspace of dimension $\sum_{i \in \underline{k}} \mu_i$. Furthermore, it is clear that $\sum_{i \in \underline{k}} \mathcal{V}_i$ is a c.s. covering \mathcal{D} , hence

$$\dim \left(\sum_{i \in \underline{k}} \mathcal{V}_i \right) \geq \sum_{i \in \underline{k}} \mu_i = \sum_{i \in \underline{k}} \dim \mathcal{V}_i,$$

which implies that the \mathcal{V}_i , $i \in \underline{k}$, are independent subspaces.

Remark 4. It has been noted by one of the reviewers that the results of Theorem 2 are encompassed by Theorem 2.1 of [11]. That is, finding an (A, B) -invariant subspace \mathcal{W} of least dimension such that \mathcal{W} contains \mathcal{D} is equivalent to finding a c.s. of minimal dimension containing \mathcal{D} when $\mathcal{D} \subset \mathcal{B}$.

5. Conclusions. The description of controllability subspaces in terms of the kernel of the singular pencil of matrices $(\lambda I - A; -B)$ extends the notion of a c.s. presented in [10], and provides a basic link between c.s. and structural properties of linear systems. In addition, the pencil of matrices presents an alternative, and algebraically appealing determination of the canonical form of Brunovsky.

In §3 we showed how the Kronecker invariants completely specify the dimensions of possible c.s. It is to be noted that, generically, all the Kronecker invariants will be either $[n/m]$ or $[n/m] + 1$, where $[a]$ is the largest integer not greater than a , in which case there should be considerable freedom in the construction of c.s. (see [3] and/or [7]). However, the results of that section allow one to ascertain that particular systems are structurally not decoupleable by state feedback, which strengthens the results of [10].

The ideas on minimal dimension c.s. developed in §4 are interesting in terms of limiting the effect of a system input, and in the dual sense of observability subspaces, for the determination of limited order observers. The general problem suggested in that section, finding minimal dimension c.s. containing arbitrary subspaces of \mathcal{X} , remains an open issue subject to further investigation.

Acknowledgment. The authors wish to thank Dr. Sanjoy K. Mitter for his helpful comments during the course of this research.

REFERENCES

- [1] P. BRUNOVSKY, *A classification of linear controllable systems*, Kybernetika, 3 (1970), pp. 173–187.
- [2] A. E. ECKBERG, *Algebraic system theory with application to decentralized control*, Ph.D. thesis, Mass. Inst. of Tech., Cambridge, Mass., 1973.
- [3] E. FABIAN AND W. M. WONHAM, *Generic solvability of the decoupling problem*, Control System Rep. 7301, Univ. of Toronto, 1973.
- [4] G. D. FORNEY, *Minimal bases of rational vector spaces with applications to multivariable linear system*, this Journal 13 (1975), pp. 493–520.
- [5] F. R. GANTMACHER, *The Theory of Matrices*, vol. II, Chelsea, New York, 1960.
- [6] R. E. KALMAN, *Kronecker invariants and feedback*, Proc. of Conf. on Ordinary Diff. Eq., NRL Math. Res. Cent., 1971.
- [7] V. M. POPOV, *Invariant description of linear, time-invariant controllable system*, this Journal, 10 (1972), pp. 252–264.
- [8] H. H. ROSENBRACK, *State Space and Multivariable Theory*, Nelson-Wiley, London, 1970.
- [9] W. A. WOLOVICH AND P. L. FALB, *On the structure of multivariable systems*, this Journal, 7 (1969), pp. 437–451.
- [10] W. M. WONHAM AND A. S. MORSE, *Decoupling and pole assignment in linear multivariable systems: A geometric approach*, this Journal, 8 (1970), pp. 93–100.
- [11] ———, *Feedback invariants of linear multivariable systems*, Automatica, 8 (1972), pp. 1–18.

MATRIX OPERATIONS INDUCED BY NETWORK CONNECTIONS*

WILLIAM N. ANDERSON, JR.,[†] RICHARD J. DUFFIN[‡] AND GEORGE E. TRAPP[¶]

Abstract. The interconnection of two electrical networks is used to define an operation, called the router sum, on the corresponding impedance matrices. For the special case of the series connection, the router sum is the ordinary matrix sum; for the parallel connection, the parallel sum is obtained [3]. General connections are considered in this paper. Conditions for commutativity and associativity of the router sum are given. Inequalities related to minimizing power are proved. An explicit formula for the router sum is given in terms of the shorted operator [2].

1. Introduction. In this paper we study a new class of operations defined for pairs of Hermitian semidefinite matrices. These operations arise from consideration of the interconnections of electrical networks; most of the theorems are motivated by the network model. However, the reader may ignore all network references and treat this as a paper about linear algebra; perhaps such a reading will suggest nonelectrical applications of our work [5]. In the linear algebra setting, the interconnection of two networks is defined by a vector space which we term a confluence; the networks themselves are represented by their impedance matrices.

An n -port electrical network may be viewed as a black box with $2n$ external terminals. The terminals are divided into pairs—called ports. By using ideal isolation transformers in the wiring lines to ports, we may insure that the current into one terminal of a port is equal to the current out of the other terminal of the same port. Given the current vector \mathbf{i} into an n -port network with transformers, the voltage vector \mathbf{v} is given by $\mathbf{v} = \mathbf{Z}\mathbf{i}$, where \mathbf{Z} is an $n \times n$ symmetric positive semidefinite matrix. \mathbf{Z} is termed the impedance matrix of the network.

If two n -ports are connected in series, the resulting n -port has impedance matrix $\mathbf{A} + \mathbf{B}$, where \mathbf{A} and \mathbf{B} are the component impedance matrices. The parallel connection of n -ports generates the parallel sum of impedance matrices. The parallel sum is defined by $(\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1}$ and denoted $\mathbf{A}:\mathbf{B}$, if the indicated inverses exist [3]. In the case of singular matrices the definition $\mathbf{A}:\mathbf{B} = \mathbf{A}(\mathbf{A} + \mathbf{B})^+ \mathbf{B}$ is the correct impedance matrix, where \mathbf{C}^+ is the Moore–Penrose generalized inverse of \mathbf{C} . Parallel addition has been studied in the scalar case by Erickson [15] and Lehman [20]. Anderson and Duffin have formulated the theory of parallel addition in the matrix case, see [3].

Most of the classical studies of the interconnection of electrical networks have been restricted to the series and parallel connections; in fact these connections are adequate for most purposes of network synthesis [7], [11]. However, the synthesis problem can be considerably simplified if the hybrid connection (some ports in series and the rest in parallel) is used [12]. The cascade connection has also been studied in the network literature [19].

* Received by the editors May 24, 1973, and in revised form January 3, 1974.

[†] Department of Mathematics, University of Maryland, College Park, Maryland 20742.

[‡] Department of Mathematics, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213.

[¶] Department of Statistics and Computer Science, West Virginia University, Morgantown, West Virginia 26506.

Both the hybrid and cascade connections are amenable to an algebraic treatment [13], [25]. Since these connections have no one-port analogs, the hybrid and cascade additions are not defined for scalars, but only for operators of appropriate dimensionality.

Even if questions of convenience are disregarded, there are fundamental limitations on the series-parallel structure, as Fialkow has shown [16]. Thus it is of interest to try to study the most general kind of network connection; this is our second paper in this direction. In [4] we gave graphical characterizations of those connections which will lead to matrix operations. In this paper we study the resulting operations, but in an axiomatic algebraic setting. Many operations which do not arise from graphs also satisfy our axioms.

Many of the properties of interconnected electrical networks may be obtained by simple physical reasoning; one test on the correctness of our model is the possibility of a mathematical derivation of the same properties [14]. The elementary properties of matrix addition can be viewed this way. For example, physically it makes no difference whether network A is connected in series with network B , or vice-versa. Thus series addition should be commutative; from a similar argument it follows that series addition should be associative. A deeper property is the principle of Maxwell that currents will distribute themselves in such a way as to minimize the power, constrained only by Kirchhoff's current law [23]. We are in fact able to give algebraic formulations for a number of physical principles, including the above, and to prove the corresponding theorems.

In §3 we define the concept of *confluence*. A confluence is a subspace of $3n$ -dimensional space satisfying certain axioms. We put an indefinite inner product on this space, and prove that the orthogonal complement of a confluence is again a confluence, called the *dual confluence*. Physically, a confluence represents the vector space of all currents allowed by a given network interconnection; the space of voltages is then the dual confluence. In this section we also establish certain matrix representations for confluences.

In §4 we define an *associative confluence*, and prove that a confluence is associative if and only if its dual is associative. A confluence arising from a network connection will be associative if given three networks R , S , and T , the result of connecting R to S and then to T is the same as connecting S to T and then R to the connection of S and T . The series, parallel, hybrid and cascade connections all give rise to associative confluences; it is easy however to design a confluence that is not associative.

In §5 we treat the general question of deriving matrix operations from confluences. We do this by setting up the linear equations which correspond to the electrical problem: given a current c applied to a conjoined network, find the currents and voltages in the component networks and the voltage in the connection. We prove that these equations define γ , a unique voltage of the connection; thus there is a matrix C so that $Cc = \gamma$. We call this matrix C the *router sum* of A and B , written $C = A*B$, where A and B are the matrices of the components. We then derive some algebraic properties of $A*B$. If A and B are Hermitian positive semi-definite (HSD), then $A*B$ is HSD; if the confluence is associative then the router operation is associative. Letting $A \geq B$ mean $A - B$ is HSD, we show $A \geq B$ implies that $A*C \geq B*C$, for any HSD C . A fundamental variational principle

due to Maxwell is expressed in three inequalities, known as the *power inequality*, the *series-router inequality*, and the *parallel-router inequality*. Finally we show the relationship between the router operation defined by a confluence and the router operation defined in a similar manner from the dual confluence.

In § 6 we turn to the question of obtaining an explicit form for the router sum. In order to do this it becomes necessary to introduce the concepts of generalized inverse and shorted operators [1], [2]. By using the explicit form we are then able to prove that the router operation is continuous; another result that one would expect from the physical model.

In the last section, we examine possible extensions of this work, and discuss future areas of work.

2. Preliminaries. We denote by C^n an n -dimensional complex vector space. Let $V = C^n \times C^n \times C^n$, the space of all triples of vectors from C^n . A vector x in V is written $x = (a, b, c)$, where a, b and c are in C^n . We define an indefinite inner product on V by the formula:

$$\begin{aligned}\langle x, y \rangle &= \langle (a_1, b_1, c_1), (a_2, b_2, c_2) \rangle \\ &= (a_1, a_2) + (b_1, b_2) - (c_1, c_2),\end{aligned}$$

where (a_1, a_2) , (b_1, b_2) and (c_1, c_2) each are the usual inner product on C^n .

Similarly we will consider $C^n \times C^n \times C^n \times C^n$ with the inner product

$$\langle (a_1, b_1, c_1, d_1), (a_2, b_2, c_2, d_2) \rangle = (a_1, a_2) + (b_1, b_2) + (c_1, c_2) - (d_1, d_2).$$

Not all of the usual Euclidian properties are preserved using an indefinite inner product; however the following results will be needed. We will only sketch the proofs; details may be found in Greub [17, Chap. XII].

LEMMA 1. If T is a subspace of V , we define T^\perp in the usual way:

$$T^\perp = \{y | \langle x, y \rangle = 0, \forall x \in T\}.$$

Then

- (i) T^\perp is a subspace,
- (ii) $T^{\perp\perp} = T$,
- (iii) $\dim T + \dim T^\perp = \dim V$.

Proof. The proof is similar to that for Euclidian spaces; it should be noted, however, that in general $T \cap T^\perp \neq 0$.

LEMMA 2. If $x \in V$ and $\langle x, y \rangle = 0$ for all y then $x = 0$.

Proof. Let $x = (a, b, c)$, and $y = (a, 0, 0)$. Then $\langle x, y \rangle = 0$ implies $(a, a) = 0$ which implies that $a = 0$. A similar treatment shows that b and c must also be 0. Q.E.D.

LEMMA 3. If A is a linear operator from V to V , then the formula $\langle Ax, y \rangle = \langle x, A^*y \rangle$ defines another linear operator A^* , called the adjoint of A ; furthermore,

- (i) $\text{range } (A) = (\text{null space } (A^*))^\perp$, and
- (ii) If

$$A = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix},$$

then

$$A^* = \begin{bmatrix} A_{11}^* & A_{21}^* & -A_{31}^* \\ A_{12}^* & A_{22}^* & -A_{32}^* \\ -A_{13}^* & -A_{23}^* & A_{33}^* \end{bmatrix}.$$

Proof. The method of proof is the same as in the case of a Euclidian inner product. The minus signs in the adjoint of A are caused by the minus sign in the inner product.

Lemma 3 is a form of the Fredholm theorem, which we will need in § 5.

An Hermitian $n \times n$ matrix A is said to be positive semidefinite if $(Ax, x) \geq 0$ for all x ; the abbreviation HSD will be used. For an HSD matrix it is easy to prove that $(Ax, x) = 0$ if and only if $Ax = 0$. If A and B are HSD matrices, then by $A \geq B$ we mean that $A - B$ is HSD.

The HSD matrix A when restricted to its range is one-to-one, and thus a right inverse may be defined there. By A^+ we mean an extension of this partially defined inverse to the whole vector space; the Moore–Penrose generalized inverse is one commonly studied extension [1], [24], but for our purposes any extension will suffice.

3. Confluences and dual confluences. As mentioned in the Introduction, in a previous paper we considered interconnection of sets of terminals [4]. With proper restrictions on the connections we were able to introduce the concept of current flows and to show that the vector space of all current flows possessed certain properties. Here we take these properties as axioms; the resulting vector subspace we call a confluence.

DEFINITION. A *confluence* is a subspace G of $V = C^n \times C^n \times C^n$ such that:

- (i) for each pair (a, b) there is at most one vector c such that (a, b, c) is in G ;
- (ii) for each vector c there is a pair (a, b) such that (a, b, c) is in G .

Note that since G is a subspace, the following condition is equivalent to (i).

- (i') if $(0, 0, c)$ is in G then $c = 0$.

THEOREM 4. If G is a confluence, then G^\perp is a confluence.

Proof. To verify condition (i') suppose that $(0, 0, c_0)$ is in G^\perp . By (ii), for each c there is a triple (a, b, c) such that (a, b, c) is in G . Then $0 = \langle (a, b, c), (0, 0, c_0) \rangle = -(c, c_0)$ for all c , therefore $c_0 = 0$. For (ii), we note that the set of all vectors c for which there is a triple (a, b, c) in G^\perp is a subspace S of C^n ; let c_0 be orthogonal to S . Then $(0, 0, c_0)$ is orthogonal to all vectors in G^\perp . That is $(0, 0, c_0)$ is in $G^{\perp\perp}$. But this is G by Lemma 1. Therefore $c_0 = 0$ and thus c is arbitrary. Q.E.D.

If (a, b, c) is in G , we say that c is the confluence of a and b , written $c = a \# b$; similarly, if (a, b, c) is in G^\perp , we write $c = a \#^\perp b$.

Physically, the reason for studying dual confluences is the fact that if G is the set of all possible current vectors for a network connection, then G^\perp is the set of all possible voltage vectors. The analog of this result for the networks themselves, which is due to Weyl [26], is the basis of the duality theory of Bott and Duffin [8]. Alternatively we may regard both G and G^\perp as sets of current vectors; the corresponding network connections are then called duals. For example, the series connection is dual to the parallel connection.

We may think of the confluence as being a graph of a linear transformation f from $C^n \times C^n$ to C^n . The domain of f , $D(f)$ is the set of all pairs (a, b) for which there is a triple (a, b, c) in G ; the null space of f , $N(f)$, is the subspace of $D(f)$ consisting of all pairs (a, b) such that $(a, b, 0)$ is in G . Let f^\perp be the transformation obtained in a similar manner from G . Then the transformations f and f^\perp have the following relation.

THEOREM 5. $D(f)^\perp = N(f^\perp)$.

Proof. (a_1, b_1) is in $D(f)^\perp$

iff (a_2, b_2) is in $D(f)$ implies that $(a_1, a_2) + (b_1, b_2) = 0$,

iff (a_2, b_2, c_2) is in G implies that $\langle (a_1, b_1, 0), (a_2, b_2, c_2) \rangle = 0$,

iff $(a_1, b_1, 0)$ is in G^\perp ,

iff (a_1, b_1) is in $N(f^\perp)$. Q.E.D.

A third representation of confluences may be given in terms of matrices.

THEOREM 6. *There exist $n \times n$ matrices W, X, Y and Z such that (a, b, c) is in G if and only if*

$$(1) \quad \begin{aligned} Wa + Xb &= c, \\ Ya + Zb &= 0. \end{aligned}$$

Proof. A triple (a, b, c) will be in G if and only if it is orthogonal to G^\perp ; to check this we need only check on a basis for G^\perp . Since G has at least n independent vectors, it follows from Lemma 1 that the dimension of G^\perp is not more than $2n$. By (ii) of the confluence definition, we may then choose the matrix

$$(2) \quad \begin{bmatrix} W^* & Y^* \\ X^* & Z^* \\ I & 0 \end{bmatrix}$$

whose columns span G^\perp . Thus (a, b, c) will be in G if and only if (a, b, c) is orthogonal to all columns of (2); that is (a, b, c) is in G if and only if

$$(3) \quad \begin{bmatrix} W & X & -I \\ Y & Z & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

from which (1) follows (the $-I$ in (3) arises because of the indefinite inner product (see Lemma 3)). Q.E.D.

Similarly, we may choose matrices Φ, Ψ, Ξ , and Ω such that (α, β, γ) is in G^\perp if and only if

$$(4) \quad \begin{aligned} \Phi\alpha + \Psi\beta &= \gamma, \\ \Xi\alpha + \Omega\beta &= 0. \end{aligned}$$

In our previous paper [4] where we studied confluences arising from graphical connections of sets of terminals, the matrices of Theorem 6 were defined by the

incidence matrices of the connection. We then used these matrices to define the confluence. The present treatment is considerably more general, since it may easily be shown that not all confluences can arise from a graph. For example, let $n = 1$, $W = 1$, $X = 2$, and $Y = Z = 0$.

Using Theorem 6, we now derive two important corollaries.

COROLLARY 7. *The vector (a, b, c) is in G if and only if there exists a vector t such that*

$$(5) \quad \begin{bmatrix} \Phi^* & \Xi^* \\ \Psi^* & \Omega^* \end{bmatrix} \begin{bmatrix} c \\ t \end{bmatrix} = \begin{bmatrix} a \\ b \end{bmatrix}.$$

Proof. From the definition of Φ , Ψ , Ξ , and Ω , we have that the null space of

$$\begin{bmatrix} \Phi & \Psi & -I \\ \Xi & \Omega & 0 \end{bmatrix}$$

is G^\perp , therefore by Lemma 3, the range of

$$\begin{bmatrix} \Phi^* & \Xi^* \\ \Psi^* & \Omega^* \\ I & 0 \end{bmatrix}$$

is G . If (a, b, c) is in G , then there exist t and u so that

$$(6) \quad \begin{bmatrix} \Phi^* & \Xi^* \\ \Psi^* & \Omega^* \\ I & 0 \end{bmatrix} \begin{bmatrix} c \\ u \\ t \end{bmatrix} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}.$$

Rewriting (6) we see that $u = c$ and that

$$\begin{bmatrix} \Phi^* & \Xi^* \\ \Psi^* & \Omega^* \end{bmatrix} \begin{bmatrix} c \\ t \end{bmatrix} = \begin{bmatrix} a \\ b \end{bmatrix}. \quad \text{Q.E.D.}$$

The vector t will appear later as a Lagrange multiplier in a variational problem.

COROLLARY 8. *Let W , X , Y , Z , Φ , Ψ , Ξ , and Ω be defined as above. Then*

$$(7) \quad \begin{bmatrix} W & X \\ Y & Z \end{bmatrix} \begin{bmatrix} \Phi^* & \Xi^* \\ \Psi^* & \Omega^* \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}.$$

Proof. Let c and t be arbitrary, and let a and b be defined as follows:

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \Phi^* & \Xi^* \\ \Psi^* & \Omega^* \end{bmatrix} \begin{bmatrix} c \\ t \end{bmatrix}.$$

Now by Theorem 6 we have

$$\begin{bmatrix} W & X \\ Y & Z \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} c \\ 0 \end{bmatrix}.$$

Combining these two equations, we have

$$\begin{bmatrix} W & X \\ Y & Z \end{bmatrix} \begin{bmatrix} \Phi^* & \Xi^* \\ \Psi^* & \Omega^* \end{bmatrix} \begin{bmatrix} c \\ t \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} c \\ t \end{bmatrix},$$

and the result follows. Q.E.D.

We may easily write the matrices for confluences arising from familiar network connections. For example, for the series connection we have

$$\begin{bmatrix} W & X \\ Y & Z \end{bmatrix} = \begin{bmatrix} I & 0 \\ I & -I \end{bmatrix}.$$

A defining set of matrices for the parallel connection is given by

$$\begin{bmatrix} W & X \\ Y & Z \end{bmatrix} = \begin{bmatrix} I & I \\ 0 & 0 \end{bmatrix}.$$

As mentioned above, the series and parallel confluences are duals. Using the above matrices one may easily verify this statement. A combination of the series and parallel confluences is known as the hybrid confluence. The matrices for this confluence are given by

$$\begin{bmatrix} W & X \\ Y & Z \end{bmatrix} = \left[\begin{array}{cc|cc} I & 0 & 0 & 0 \\ 0 & I & 0 & I \\ \hline I & 0 & -I & 0 \\ 0 & 0 & 0 & 0 \end{array} \right].$$

Another common network connection is the cascade connection, for which the matrices are

$$\begin{bmatrix} W & X \\ Y & Z \end{bmatrix} = \left[\begin{array}{cc|cc} I & 0 & 0 & 0 \\ 0 & 0 & 0 & I \\ \hline 0 & I & I & 0 \\ 0 & 0 & 0 & 0 \end{array} \right].$$

4. Associative and commutative confluences. Given a confluence G and vectors a, b and c , it may happen that there is a vector d such that $d = (a \# b) \# c$, or a vector e such that $e = a \# (b \# c)$. In many of the familiar examples, we then have that $d = e$. With this in mind we define the confluence G to be *associative* if whenever either $(a \# b) \# c$ or $a \# (b \# c)$ is defined, then both are defined and they are equal.

THEOREM 9. *The confluence G is associative if and only if the confluence G^\perp is associative.*

To prove the theorem, we first need to establish the following lemmas.

LEMMA 10. $d = (a \# b) \# c$ if and only if

$$(8) \quad \begin{bmatrix} WW & WX & X & -I \\ YW & YX & Z & 0 \\ Y & Z & 0 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Proof. By Corollary 7 we have that $d = (a \# b) \# c$ if and only if there is an e such that

$$(9) \quad \begin{bmatrix} W & X \\ Y & Z \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} e \\ 0 \end{bmatrix},$$

and

$$(10) \quad \begin{bmatrix} W & X \\ Y & Z \end{bmatrix} \begin{bmatrix} e \\ c \end{bmatrix} = \begin{bmatrix} d \\ 0 \end{bmatrix}.$$

If (9) and (10) hold, then (8) is established by eliminating e . Conversely, if (8) holds, then $Ya + Zb = 0$. Thus e is defined by (9), and (10) will hold. Q.E.D.

LEMMA 11. $d = (a \# b) \# c$ if and only if there exist t, u and v such that

$$(11) \quad \begin{bmatrix} \Phi^* \Phi^* & \Phi^* \Xi^* & \Xi^* \\ \Psi^* \Phi^* & \Psi^* \Xi^* & \Omega^* \\ \Psi^* & \Omega^* & 0 \\ I & 0 & 0 \end{bmatrix} \begin{bmatrix} t \\ u \\ v \end{bmatrix} = \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix}.$$

Proof. This is a restatement of Lemma 10 in terms of G^\perp instead of G . Equation (11) states that (a, b, c, d) is in the range of the given matrix, but this is equivalent by Lemma 3 to asserting that (a, b, c, d) is in the null space of the adjoint matrix. Q.E.D.

To complete the proof of Theorem 9, we need to establish the following two lemmas.

LEMMA 12. Let T_1 be the subspace consisting of all quadruples (a, b, c, d) such that $d = (a \# b) \# c$, and let U_1 be the subspace of all quadruples $(\alpha, \beta, \gamma, \delta)$ such that $\delta = (\alpha \# \beta) \# \gamma$. Then $T_1 = U_1^\perp$.

Proof. By Lemma 11, $d = (a \# b) \# c$ if and only if there is a solution to equation (11). But by Lemma 3 there will be a solution to (11) if and only if (a, b, c, d) is orthogonal to the null space of

$$(12) \quad \begin{bmatrix} \Phi\Phi & \Phi\Psi & \Psi & -I \\ \Xi\Phi & \Xi\Psi & \Omega & 0 \\ \Xi & \Omega & 0 & 0 \end{bmatrix}.$$

Lemma 10 assures us that the null space of (12) is U_1 . Therefore (a, b, c, d) is in T_1 if and only if $(a, b, c, d) \perp (\alpha, \beta, \gamma, \delta)$ for all $(\alpha, \beta, \gamma, \delta)$ in U_1 . Q.E.D.

LEMMA 13. Let T_2 be the subspace consisting of all quadruples (a, b, c, d) such that $d = a \# (b \# c)$, and U_2 the subspace of all quadruples $(\alpha, \beta, \gamma, \delta)$ such that $\delta = \alpha \# (\beta \# \gamma)$. Then $T_2 = U_2^\perp$.

Proof. The proof is analogous to the proof of Lemma 12.

Proof of Theorem 9. The confluence G is associative if and only if $T_1 = T_2$, and G^\perp is associative if and only if $U_1 = U_2$; but by Lemmas 12 and 13 these two conditions are equivalent. Q.E.D.

In the next section we consider matrix operations derived from a confluence.

In order to help analyze the operations, we make the following two definitions. A confluence G is *commutative* whenever $a \# b = b \# a$; in terms of subspaces this means that (a, b, c) is in G if and only if (b, a, c) is in G . Similarly, we define a confluence to be *anticommutative* whenever $a \# b = -(b \# a)$.

THEOREM 14. *The confluence G is (anti) commutative if and only if G^\perp is (anti) commutative.*

Proof. Suppose that G is commutative and (α, β, γ) is not in G^\perp . Then there is some triple (a, b, c) such that (a, b, c) is in G and moreover, $\langle (a, b, c), (\alpha, \beta, \gamma) \rangle \neq 0$. But then since G is commutative, we have (b, a, c) is in G and $\langle (b, a, c), (\beta, \alpha, \gamma) \rangle \neq 0$, and thus (β, α, γ) is not in G^\perp . The proof for anticommutative G is similar. Q.E.D.

The series, parallel, hybrid and cascade confluences may easily be shown to be associative. The series, parallel and hybrid confluences are commutative, while the cascade is not. An example of an anticommutative confluence which can be realized by a graphical connection is given by

$$\begin{bmatrix} W & X \\ Y & Z \end{bmatrix} = \begin{bmatrix} I & 0 \\ I & I \end{bmatrix}.$$

5. Matrix operations. Using confluences, in this section we determine matrix operations on the set of Hermitian semidefinite matrices. We abbreviate Hermitian semidefinite as HSD. The necessary properties of HSD matrices are given in § 2.

In previous papers [3], [13], particular electrical connections were used to generate matrix operations. The following theorem shows that any network connection that has a confluence representation gives rise to an HSD matrix operation.

THEOREM 15. *Let G be a confluence, and A and B be HSD matrices. Then for each vector c there is a unique vector γ , and vectors a and b such that (a, b, c) is in G and (Aa, Bb, γ) is in G^\perp .*

Proof. Consider the equations

$$(13) \quad \begin{bmatrix} 0 & 0 & W & X \\ 0 & 0 & Y & Z \\ W^* & Y^* & -A & 0 \\ X^* & Z^* & 0 & -B \end{bmatrix} \begin{bmatrix} \gamma \\ \lambda \\ a \\ b \end{bmatrix} = \begin{bmatrix} c \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

By Theorem 6 the first two equations express the fact that (a, b, c) is in G , and by Corollary 7 the last two express the fact that (Aa, Bb, γ) is in G^\perp . To prove the theorem we need to show that for each c , there is a unique γ satisfying (13). To establish the existence of solutions we consider the homogeneous adjoint system to (13):

$$(14) \quad \begin{bmatrix} 0 & 0 & W & X \\ 0 & 0 & Y & Z \\ W^* & Y^* & -A & 0 \\ X^* & Z^* & 0 & -B \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

If (u_1, u_2, u_3, u_4) satisfies (14) then we have

$$(15) \quad \begin{bmatrix} W^* & Y^* \\ X^* & Z^* \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} Au_3 \\ Bu_4 \end{bmatrix},$$

and therefore (Au_3, Bu_4) is orthogonal to all solutions to the homogeneous adjoint system to (15); but by (14), (u_3, u_4) is such a solution. Therefore we have that

$$(16) \quad (Au_3, u_3) + (Bu_4, u_4) = 0.$$

Since A and B are HSD, (16) implies that $Au_3 = 0$ and $Bu_4 = 0$. Thus the right-hand side of (15) is 0, and (15) then becomes the homogeneous adjoint system for the system:

$$(17) \quad \begin{bmatrix} W & X \\ Y & Z \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} c \\ 0 \end{bmatrix}.$$

Since G is a confluence, Theorem 6 guarantees that (17) has solutions for all c ; therefore (u_1, u_2) is orthogonal to all vectors $(c, 0)$, and thus $u_1 = 0$. Therefore $(c, 0, 0, 0)$ is orthogonal to all solutions of (14), and we then have existence of solutions to (13). Moreover, we have shown that if $c = 0$, then $\gamma = 0$, which establishes uniqueness. Q.E.D.

The correspondence between c and γ is linear; thus by Theorem 15 there is a matrix C such that $Cc = \gamma$. We call C the *router sum* of A and B , written $C = A*B$.

THEOREM 16. *If A and B are HSD matrices, then $A*B$ is HSD.*

Proof. Since (a, b, c) and (Aa, Bb, γ) are in G and G^\perp respectively, we have that $(Cc, c) = (\gamma, c) = (Aa, a) + (Bb, b)$. Since A and B are HSD, the right-hand side of the above is real and nonnegative. Therefore the quadratic form of C is real and nonnegative; thus C is HSD [18]. Q.E.D.

In previous papers we have studied the matrix operations arising from the series, parallel, hybrid and cascade confluences [2], [13], and [25]. The operation arising from the series confluence is ordinary addition; from the parallel confluence we obtain the operation of parallel addition, and we denote the parallel sum of A and B as $A:B$. Many of the theorems previously obtained are special cases of the theorems of this section.

The next theorem expresses the well-known physical principle that the currents in a network will minimize the power dissipated, subject only to Kirchhoff's current law.

THEOREM 17. *Let G be a confluence, and A and B be HSD matrices. Let c and γ be vectors such that $A*Bc = \gamma$. Then for any vector (a, b, c) in G we have*

$$(18) \quad (Aa, a) + (Bb, b) \geq (\gamma, c).$$

Moreover, equality will hold in (18) if and only if a and b are as in Theorem 15.

Proof. Let $(\gamma, \lambda, a_0, b_0)$ be a solution to (13). Then (a_0, b_0, c) is in G and (Aa_0, Bb_0, γ) is in G^\perp ; thus we have

$$(19) \quad (Aa_0, a_0) + (Bb_0, b_0) = (\gamma, c),$$

so that equality holds in (18). Moreover, since G is a confluence, the vector $(a - a_0, b - b_0, 0)$ is in G , and therefore we have the following:

$$(20) \quad (Aa_0, (a - a_0)) + (Bb_0, (b - b_0)) = 0.$$

Expanding the right-hand side of equation (18) yields

$$\begin{aligned}
 (Aa, a) + (Bb, b) &= (A(a_0 + (a - a_0)), (a_0 + (a - a_0))) \\
 &\quad + (B(b_0 + (b - b_0)), (b_0 + (b - b_0))) \\
 &= (Aa_0, a_0) + (Bb_0, b_0) + (A(a - a_0), (a - a_0)) \\
 &\quad + (B(b - b_0), (b - b_0)) + (Aa_0, (a - a_0)) + (Bb_0, (b - b_0)) \\
 &\quad + (A(a - a_0), a_0) + (B(b - b_0), b_0) \\
 &= (\gamma, c) + (A(a - a_0), (a - a_0)) + (B(b - b_0), (b - b_0)) \\
 &\geq (\gamma, c) \quad \text{since } A \text{ and } B \text{ are HSD.}
 \end{aligned}$$

If equality holds in (18), then $(A(a - a_0), a - a_0) + (B(b - b_0), b - b_0) = 0$; since A and B are HSD, it follows that $Aa = Aa_0$ and $Bb = Bb_0$. Therefore (Aa, Bb, γ) is in G . Q.E.D.

COROLLARY 18. *If A and B are HSD matrices, then for any vectors a, b and c such that $c = a + b$,*

$$(21) \quad (Aa, a) + (Bb, b) \geq (A : Bc, c).$$

Moreover, for any vector c , the vectors a and b can be chosen so that equality holds in (21).

Proof. This corollary, which is Lemma 18 of [3], is the special case of Theorem 17 for the parallel confluence.

As mentioned in § 3, our theory is based upon the physical principle that if the currents allowed by a connection form the confluence G , then the voltages will form the dual confluence G^\perp . Alternatively, we could have based our work on the variational principle expressed by Theorem 17; the vector λ of Theorem 15 would then appear as the Lagrange multiplier for the minimization problem.

The following theorem, known as the *series-router inequality*, may also be motivated by a variational principle; the exact argument is given in [25]. Let us recall that for matrices A and B , $A \geq B$ is defined to mean that $A - B$ is HSD.

THEOREM 19. *Let A, B, C and D be HSD matrices. Then*

$$(A + B) * (C + D) \geq (A * C) + (B * D).$$

Proof. For any vector c , let a and b be vectors such that equality holds in (18) with $A + B$ and $C + D$ replacing A and B . Then

$$\begin{aligned}
 ((A + B) * (C + D)c, c) &= ((A + B)a, a) + ((C + D)b, b) \\
 &= (Aa, a) + (Cb, b) + (Ba, a) + (Db, b) \\
 &\geq (A * Cc, c) + (B * Dc, c).
 \end{aligned}$$

The last inequality follows from Theorem 17. Q.E.D.

Because of the duality between series and parallel addition, we are led to consider the dual theorem, known as the *parallel-router inequality*.

THEOREM 20. *Let A, B, C and D be HSD matrices. Then*

$$(A : B) * (C : D) \leq (A * C) : (B * D).$$

Proof. For any vector c , by Corollary 18 there are vectors a and b such that $c = a + b$ and moreover

$$((A*C):(B*D)c, c) = (A*Ca, a) + (B*Db, b).$$

Then by Theorem 17 there are a_1 and a_2 such that (a_1, a_2, a) is in G and

$$(A*Ca, a) = (Aa_1, a_1) + (Ca_2, a_2).$$

Similarly there are b_1 and b_2 such that (b_1, b_2, b) is in G and

$$(B*Db, b) = (Bb_1, b_1) + (Db_2, b_2).$$

Moreover, since G is a subspace, $(a_1 + a_2, b_1 + b_2, a + b) = (a_1 + a_2, b_1 + b_2, c)$ is in G . Therefore we have

$$\begin{aligned} ((A*C):(B*D)c, c) &= (A*Ca, a) + (B*Db, b) \\ &= (Aa_1, a_1) + (Ca_2, a_2) + (Bb_1, b_1) + (Bb_2, b_2) \\ &\geq (A:B(a_1 + b_1), (a_1 + b_1)) + (C:D(a_2 + b_2), (a_2 + b_2)) \\ &\geq ((A:B)*(C:D)c, c). \end{aligned}$$

The penultimate inequality follows from Corollary 18, and the final inequality from Theorem 17. Q.E.D.

Using Theorem 19 (or Theorem 20), we can now show that a router operation preserves the Hermitian semidefinite partial order.

COROLLARY 21. *If A, B and C are HSD, then $A \geq B$ implies that $A*C \geq B*C$.*

Proof. Since $A \geq B$ we have $A = B + D$ where D is HSD. Then

$$A*C = (B + D)*C = (B + D)*(C + 0) \geq B*C + D*0 \geq B*C.$$

Here we have used Theorem 19 and the fact that $D*0$ is HSD. Q.E.D.

A similar argument will verify the following related corollary.

COROLLARY 22. *If A, B, C and D are HSD then $A \geq B$ implies that $C*A \geq C*B$.*

We have already seen that a router operation is closed on the set of HSD matrices and that it preserves the HSD partial ordering. The next two theorems give conditions so that router addition is associative and commutative. An operation with these properties is called a (commutative) semigroup operation.

THEOREM 23. *Let G be an associative confluence. Then the router addition $*$ derived from G is associative.*

Proof. We will use the notation of the proof of Theorem 9. Let d be arbitrary. If $\delta_1 = A*(B*C)d$, then there are vectors a, b and c such that (a, b, c, d) is in T_1 and (Aa, Bb, Cc, δ_1) is in T_1^\perp . Similarly, if $\delta' = (A*B)*Cd$, then there are vectors a', b' and c' such that (a', b', c', d) is in T_2 and (Aa', Bb', Cc', δ') is in T_2^\perp . However, if G is an associative confluence, then $T_1 = T_2$, and thus by the uniqueness part of Theorem 15, we have $\delta = \delta'$. Since d was arbitrary we see that $A*(B*C) = (A*B)*C$. Q.E.D.

COROLLARY 24. *Series, parallel, hybrid and cascade addition are associative.*

Proof. We have already seen that the corresponding confluences are associative. Q.E.D.

In our previous papers we proved that parallel and hybrid additions are associative; the proofs used the explicit matrix form for these additions and were quite difficult. Cascade addition has not previously been proved associative.

THEOREM 25. *Let G be a commutative or an anticommutative confluence. Then the router addition $*$ derived from G is commutative.*

Proof. Suppose that G is commutative, and $\gamma = A*Bc$. Then there are vectors a and b such that (a, b, c) is in G and (Aa, Bb, γ) is in G^\perp . But then by hypothesis (b, a, c) is in G and by Theorem 14, (Bb, Aa, γ) is in G^\perp . Therefore $\gamma = B*Ac$ and thus $A*B = B*A$. If G is anticommutative, it follows that $(-b, -a, c)$ is in G and $(-Bb, -Aa, \gamma)$ is in G^\perp ; again we have $A*B = B*A$. Q.E.D.

If the matrices A and B are positive definite, then there is a familiar form for $A:B$; in fact, $A:B = (A^{-1} + B^{-1})^{-1}$. This formula can be viewed as a relation between series addition and its dual. In the general case we have the following theorem.

THEOREM 26. *Let the router addition $*$ be derived from the confluence G , and let $*^\perp$ be derived from G^\perp . Then if A and B are Hermitian positive definite, we have*

$$A*B = (A^{-1}*^\perp B^{-1})^{-1}.$$

Proof. If $A*Bc = \gamma$, then there exist vectors a and b such that (a, b, c) is in G and (Aa, Bb, γ) is in G^\perp . Let $\alpha = Aa$ and $\beta = Bb$. Then $(A^{-1}\alpha, B^{-1}\beta, c)$ is in G and (α, β, γ) is in G^\perp ; that is, $c = A^{-1}*^\perp B^{-1}\gamma$. Q.E.D.

The relationship between the dual operation and semidefinite matrices is not known. One immediate problem is that in general $A:B \neq (A^+ + B^+)^+$. It may be shown, however, that $A:B(A^+ + B^+)$ is a nonorthogonal projection; perhaps the general result will be of this form.

6. An explicit form for the router sum. In the previous section, we showed that a confluence generates a router operation on the set of HSD matrices. In this section, we determine an explicit representation for the router sum in terms of the Φ , Ψ , Ξ and Ω matrices.

THEOREM 27. *Let G be a confluence and A and B be HSD matrices. Then $A*B$, the router sum, is given by*

$$(22) \quad B = \Phi A \Phi^* + \Psi B \Psi^* - (\Phi A \Xi^* + \Psi B \Omega^*)(\Xi A \Xi^* + \Omega B \Omega^*)^+(\Xi A \Phi^* + \Omega B \Psi^*).$$

Proof. Let c be a vector, and let (γ, λ, a, b) be a solution to (13). Then since (a, b, c) is in G , there is a vector t such that

$$(23) \quad \begin{bmatrix} \Phi^* & \Xi^* \\ \Psi^* & \Omega^* \end{bmatrix} \begin{bmatrix} c \\ t \end{bmatrix} = \begin{bmatrix} a \\ b \end{bmatrix}.$$

Since (Aa, Bb, γ) is in G , we also have that

$$(24) \quad \begin{bmatrix} \Phi & \Psi \\ \Xi & \Omega \end{bmatrix} \begin{bmatrix} Aa \\ Bb \end{bmatrix} = \begin{bmatrix} \gamma \\ 0 \end{bmatrix}.$$

Combining (23) and (24), we have

$$(25) \quad \begin{bmatrix} \gamma \\ 0 \end{bmatrix} = \begin{bmatrix} \Phi & \Psi \\ \Xi & \Omega \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix} \begin{bmatrix} \Phi^* & \Xi^* \\ \Psi^* & \Omega^* \end{bmatrix} \begin{bmatrix} c \\ t \end{bmatrix}.$$

Since for each vector c there is a unique γ , we may rewrite (25) using matrices :

$$(26) \quad \begin{bmatrix} A*B \\ 0 \end{bmatrix} = \begin{bmatrix} \Phi & \Psi \\ \Xi & \Omega \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix} \begin{bmatrix} \Phi^* & \Xi^* \\ \Psi^* & \Omega^* \end{bmatrix} \begin{bmatrix} I \\ T \end{bmatrix}.$$

The second equation of (26) yields $T = -(\Xi A \Xi^* + \Omega B \Omega^*)^+ (\Xi A \Phi^* + \Omega B \Psi^*)$. Solving the first gives $A*B = \Phi A \Phi^* + \Psi B \Psi^* + (\Phi A \Xi^* + \Psi B \Omega^*)T$. Q.E.D.

Formulas similar to (22) have been studied in a number of contexts. Particularly relevant to the present work is the concept of the *shorted operator* [2], [6]. If A is an HSD matrix, partitioned

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

the shorted operator of A , $S(A)$ is defined by

$$(27) \quad S(A) = \begin{bmatrix} a_{11} & -a_{12}a_{22}^+a_{21} & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

The same formula has been given by Carlson, Haynsworth and Markham [9] to define the *generalized Schur complement*, and is used by Albert [1] in connection with regression analysis. In the case where a_{22} is invertible, the formula is familiar in electrical network literature [10]; the noninvertible case is not usually treated.

In terms of the shorted operator, we may rewrite Theorem 27 as follows.

THEOREM 27'.

$$\begin{bmatrix} A*B & 0 \\ 0 & 0 \end{bmatrix} = S \left(\begin{bmatrix} \Phi & \Psi \\ \Xi & \Omega \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix} \begin{bmatrix} \Phi^* & \Xi^* \\ \Psi^* & \Omega^* \end{bmatrix} \right).$$

An alternate approach to the study of router operation may be based directly on Theorem 27'. Let M be any $r \times 2n$ matrix. Then for $n \times n$ HSD matrices A and B , we may define

$$(28) \quad \begin{bmatrix} A*B & 0 \\ 0 & 0 \end{bmatrix} = S \left(M \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix} M^* \right),$$

where $A*B$ is an $s \times s$ matrix, for some $s \leq r$; this new operation is called a *generalized router operation*. Analogues of Theorems 16, 19 and 20 of the present paper follow directly from Theorems 1, 3 and 5 of [2]. We will discuss these matters in more detail in another paper.

The representation of a router operation in terms of the shorted enables us to prove the following continuity result.

THEOREM 28. Let A, B, E, F be HSD matrices and $*$ a router operation. Then $(A + E)*(B + F) \rightarrow 0$ as $E, F \rightarrow 0$.

Proof. In Corollary 2 to Theorem 1 of [6] it is proved that $S(A + E) \rightarrow S(A)$ as $E \rightarrow 0$. The proof follows by using Theorem 27'.

7. Extensions. The equations (13) which we use to define the router sum may be written for any pair of linear operators A and B ; however, without some restrictions on the class of operators considered we cannot in general guarantee

that a unique solution will exist, or that the properties of the router sum will hold. In addition to the case of this paper—Hermitian semidefinite operators on a finite-dimensional space—we have considered a number of other cases. A brief sketch of these extensions is given below; the details will be given in later papers.

(a) As mentioned in the Introduction, the original interest in router operations came from the problem of network synthesis. The only networks whose impedance matrices are Hermitian are those networks composed of resistors and transformers; for interesting synthesis applications we need to consider networks with reactive elements. In this case the impedance matrices will be positive real matrices [11] as functions of the complex variable frequency, and will in general be non-Hermitian. Certain parts of our theory will extend, in particular, if A and B are positive real matrices. Then $A*B$ will be positive real. A synthesis method based on hybrid addition has been given [12]; we are studying applications of the more general router operations defined here.

(b) A linear operator A on a real vector space is said to be almost positive definite if $(Ax, x) \geq 0$ for all x and $(Ax, x) = 0$ if and only if $Ax = 0$ [22]; alternatively if $A = H + S$ with H symmetric and positive semidefinite and $\text{range}(S) \subset \text{range}(H)$. To a certain extent the symmetric and skew parts of an almost positive definite matrix mimic the real and imaginary parts of a positive real matrix. Theorems 15, 16, 23, 24, 25, 26 and 27 hold for almost positive definite matrices; the inequalities, however, do not.

(c) Infinite dimensional Hilbert space is a natural setting for an extension of this work. Theorem 15 is no longer true in this context; but the necessary theorems about the shorted operator are available in [6] to enable the theory of router sums to be based on (30). The special case of parallel addition is analyzed in [6]; hybrid and cascade addition can be similarly treated. Some recent work of Zemanian [27] furnishes a natural network model for these extensions.

(d) Except in our discussion of associativity, we have considered only the router sum of two operators. With a few changes of solutions, we could instead consider the router sum of m operators; the confluence then becomes a subspace of $C^{(m+1)n}$. The corresponding network model has been called an *operator network* by Zemanian [28]. In Zemanian's work the operators are assumed invertible; the method of confluence enables one to study operator networks without such assumptions.

(e) Our treatment of the interconnection of n -port electrical networks is only valid under the assumption that ideal isolation transformers are present to ensure proper port behavior; without these transformers even the formula for series addition fails [19]. The determination of those situations where ideal transformers are not necessary seems to be difficult, only the parallel case has been successfully treated [21].

REFERENCES

- [1] A. ALBERT, *Regression and the Moore–Penrose Pseudoinverse*, Academic Press, New York, 1972.
- [2] W. N. ANDERSON, JR., *Shorted operators*, SIAM J. Appl. Math., 20 (1971), pp. 520–525.
- [3] W. N. ANDERSON, JR. AND R. J. DUFFIN, *Series and parallel addition of matrices*, J. Math. Anal. Appl., 26 (1969), pp. 576–594.

- [4] W. N. ANDERSON, JR., R. J. DUFFIN AND G. E. TRAPP, *Tripartite graphs to analyze the interconnection of networks, graph theory and applications*, Lecture Notes No. 303, Springer-Verlag, Berlin, 1972.
- [5] W. N. ANDERSON, JR., G. D. KLEINDORFER, P. R. KLEINDORFER AND M. B. WOODROOFE, *Consistent estimates of the parameters of a linear system*, Ann. Math. Statist., 40 (1969), pp. 2064–2075.
- [6] W. N. ANDERSON, JR. AND G. E. TRAPP, *Shorted operators II*, SIAM J. Appl. Math., 28 (1975), pp. 60–71.
- [7] R. BOTT AND R. J. DUFFIN, *Impedance synthesis without the use of transformers*, J. Appl. Phys., 20 (1949), p. 816.
- [8] ———, *On the algebra of networks*, Trans. Amer. Math. Soc., 74 (1953), pp. 99–109.
- [9] D. CARLSON, E. HAYNSWORTH AND T. MARKHAM, *A generalization of the Schur complement by means of the Moore–Penrose inverse*, SIAM J. Appl. Math., 26 (1974), pp. 169–175.
- [10] I. CEDERBAUM, *On equivalence of resistive n -port networks*, IEEE Trans. Circuit Theory, CT-12 (1965), pp. 338–344.
- [11] R. J. DUFFIN, *Elementary operations which generate network matrices*, Proc. Amer. Math. Soc., 6 (1955), pp. 335–339.
- [12] R. J. DUFFIN, D. HAZONY AND N. MORRISON, *Network synthesis through hybrid matrices*, SIAM J. Appl. Math., 14 (1966), pp. 390–413.
- [13] R. J. DUFFIN AND G. E. TRAPP, *Hybrid addition of matrices—A network theory concept*, J. Applicable Analysis, 2 (1972), pp. 241–254.
- [14] R. J. DUFFIN, *Network Models, Mathematical Aspects of Electrical Network Analysis*, American Mathematical Society, Providence, R.I., 1971.
- [15] K. E. ERIKSON, *A new operation for analyzing series-parallel networks*, IEEE Trans. Circuit Theory, CT-6 (1959), pp. 124–126.
- [16] A. D. FIALKOW, *A limitation on the series-parallel structure*, IEEE Trans. Circuit Theory, CT-15 (1968), pp. 124–132.
- [17] W. H. GREUB, *Linear Algebra*, Springer-Verlag, Berlin, 1963.
- [18] P. R. HALMOS, *Finite Dimensional Vector Spaces*, Van Nostrand, Princeton, N.J., 1968.
- [19] L. HUELSMAN, *Circuits, Matrices and Linear Vector Spaces*, McGraw-Hill, New York, 1963.
- [20] A. LEHMAN, *Problem 60–5. A resistor network inequality*, SIAM Rev., 4 (1962), pp. 150–155.
- [21] A. LEMPEL AND I. CEDERBAUM, *Parallel interconnection of n -port networks*, IEEE Trans. Circuit Theory, CT-14 (1967), pp. 274–279.
- [22] T. LEWIS AND T. NEWMAN, *Pseudoinverses of positive semidefinite matrices*, SIAM J. Appl. Math., 16 (1968), pp. 701–703.
- [23] J. C. MAXWELL, *A Treatise on Electricity and Magnetism*, 3rd ed., reprinted by Dover, New York, 1954.
- [24] R. PENROSE, *A generalized inverse for matrices*, Proc. Cambridge Philos. Soc., 51 (1955), pp. 406–413.
- [25] G. E. TRAPP, *Algebraic operations derived from electrical networks*, Doctoral dissertation, Carnegie-Mellon University, Pittsburgh, Pa., 1970.
- [26] H. WEYL, *Reparticion de corriente en una red conductora*, Rev. Mat. Hisp.-Amer., 5 (1923), pp. 153–164.
- [27] A. H. ZEMANIAN, *The Hilbert port*, SIAM J. Appl. Math., 18 (1970), pp. 98–130.
- [28] ———, *Passive operator networks*, IEEE Trans. Circuits and Systems, CAS-21 (1974), pp. 184–193.
- [29] ———, *Infinite networks of positive operators*, Circuit Theory and Appl., 2 (1974), pp. 69–78.

CONTROLLABILITY AND OBSERVABILITY IN BANACH SPACE WITH BOUNDED OPERATORS*

ROBERTO TRIGGIANI†

Abstract. The classical theory of (state and output) controllability and observability in finite-dimensional spaces is extended to linear abstract systems defined on infinite-dimensional Banach spaces, under the basic assumption that the operator acting on the state be bounded. Tests for approximate controllability as well as observability, expressed only in terms of the coefficients of the system, are proved via a consequence of the Hahn-Banach theorem, and new phenomena arising in infinite dimensions are studied: for instance, by using Baire category arguments, it is shown that state exact controllability, under large conditions met in cases of physical interest, never arises in infinite-dimensional Banach spaces, even with free final instant. Several examples are presented throughout; in particular, for dynamical systems modeled by integro-differential equations of Volterra type, the present theory leads in turn to explicit, easy-to-check criteria for approximate controllability and observability.

An example shows the applicability of our results to the unbounded operator case.

1. Introduction. The present paper studies the notion of (state and output) controllability and observability for dynamical systems, described by linear differential equations in (infinite-dimensional, separable) Banach spaces, under the basic assumption that the operator acting on the state be bounded.

A more general model for the system was adopted, in the study of state controllability, by Fattorini (see e.g., [12], [13]) in the sense that the operator A acting on the state was assumed to be (closed, linear, with dense domain and) the infinitesimal generator of a strongly continuous semigroup. Then a necessary and sufficient condition for state approximate controllability (in our terminology) was derived in [13], for the case when A is self-adjoint, semibounded above (or normal, with certain further properties) and defined on a Hilbert space, and the dynamical system has only a finite number of scalar controls. The application of Fattorini's criterion requires the knowledge of some ordered representation. Some other technical results were given in [12]. Fattorini then applies his investigation to some differential operators of physical interest.

Our present assumption on the boundedness of the operator acting on the state is motivated by:

(a) the fact that when such operator is instead the infinitesimal generator of a strongly continuous semigroup, the state approximate controllability of the correspondent system can be reduced to, and is equivalent to, that of an associated system with bounded operator. More precisely, Fattorini showed [12, Prop. 2.3] that the closure of the union of the sets of attainability from the origin over all positive times of the two systems

$$\dot{x} = Ax + Bu(t) \quad \text{and} \quad \dot{x} = R(\lambda_0, A)x + Bu(t)$$

* Received by the editors March 5, 1973, and in revised form November 1, 1973.

† Center for Control Sciences, University of Minnesota, Minneapolis. Now at Department of Mathematics, State University of New York at Albany, New York 12222. This paper is based on part of a Ph.D. thesis at the University of Minnesota. This work was supported by the Air Force under Grant AF-AFOSR-72-2243.

is the same and so they are either both approximate controllable in finite time or both not. Such a result has never been exploited. Here $R(\lambda_0, A)$ is the (bounded!) resolvent of A , computed for instance at any point λ_0 in the half-plane $\operatorname{Re} \lambda > \omega_0$, while the spectrum of the infinitesimal generator A is contained in the complementary half-plane $\operatorname{Re} \lambda \leq \omega_0$. We also recall that $R(\cdot, A)$ is a natural bounded operator associated with A [10, pp. 626–7, etc.].

(b) the desire to acquire explicitly verifiable conditions for controllability and observability, expressed solely in terms of the coefficients of the process (suitably smooth if nonautonomous).

The characteristic conditions we shall derive are generalizations of the familiar ones [21], [23], [31] when the state space and the control space are both finite-dimensional and they in turn lead to very explicit, easy-to-check tests for classes of processes described by integro-differential equations of Volterra type. The application of our results to differential operators, via their resolvent $R(\cdot, A)$ as indicated in (a) above is illustrated with one example—(3.2.7).

With reference to infinite-dimensional (Banach) spaces, two types of state and output controllability are considered, approximate and exact, both of which generalize the classical ones of finite-dimensional spaces [36]. In addition the usual notion of observability is introduced. Our analysis then shows:

(i) the classical finite-dimensional theory is extended precisely to the approximate controllability (Theorems 3.1.1, 4.1, etc.) as well as to the observability (Theorems 5.1.1 and 5.3.1) providing the wanted tests in the sense of (b) above. In particular, in a reflexive Banach space, observability is the “dual” of state approximate controllability.

(ii) On the other hand, a new phenomenon arises, namely, under large conditions met in cases of physical interest (e.g., when the control is either finite-dimensional or acts on the system through an integral operator), state exact controllability can never arise in infinite-dimensional Banach spaces, even if the final instant of the time interval is left free (possibly depending on the pair of initial and final points) (Theorem 3.3.3).

However, from an engineering viewpoint, when the analysis of the system is limited over a finite interval, the impossibility of achieving state exact controllability is not crucial and approximate one is a satisfactory substitute.¹ The same conclusion holds *mathematically* in the physically significant case, when the output of the system is a finite-dimensional vector (corresponding to finite data derived from the global distribution of the state) (Corollary 6.2).

At the time this research was initiated, the only paper available in the literature (to the present author's knowledge) dealing with the controllability problem for a class of systems, defined on abstract spaces and with bounded operators acting on the state, was [32]. More precisely, in [32] an autonomous linear system defined on a Hilbert space and with just a scalar controller is con-

¹ We exclusively refer here to the distinction of the two concepts, approximate and exact controllability, as stated by their respective definitions. As pointed out by the referee, the distinction of the two concepts may instead be crucial, as far as their implications (or lack thereof) over other problems, that—at least in finite-dimensional systems—are known to be related to state (exact) controllability (e.g., the maximum principle that is necessary *and* sufficient for time optimal control of autonomous controllable systems).

sidered. Then three theorems are stated: the first gives a necessary condition for exact controllability; the second shows that the necessary condition is not sufficient, when coupled with a certain uniqueness property; and the third gives an asymptotic behavior result. However, [32] is partly incorrect: as mentioned above in (ii), it is in fact shown here that exact controllability for even more general systems than the one considered in [32] never arises²; hence, Theorem 1 in [32] is meaningless, since the assumption under which it is derived never holds. Theorem 3 in [32] seems to contain a flaw in its proof, since, after equation (10), a term-by-term integration, which is correct for every finite interval, is extended to the interval $[0, \infty]$: however, the stated result is correct and indeed we present here (Corollary 3.1.4) an improved version yielding a stronger conclusion under weaker assumptions for more general systems.

In very recent times, a few other papers have appeared in the literature, which consider a few of the problems discussed here: [25], [24], [15]. [25] and [24] appeared in English almost at the same time that a preliminary part from the present research was announced at the Sixth Annual Princeton Conference, overlapping two of our results. More precisely, [25] shows, in an analytic way and without using, as we do here, bounded linear functionals, the special case of our Corollary 3.1.2 for $m = 1$; [24] proves our Theorem 3.3.1 (but not its generalization, Theorem 3.3.3) without making use of the Baire category argument, as we do here, and relying instead, for part (iii), on entire functions of exponential type. Finally, [15] deals with *discrete* autonomous systems and, in addition, shows through an approach, different from Fattorini's, how the unbounded operator case can be reduced (at least in a Hilbert space) to the bounded operator case.

We also remark that the problem of observability in the context discussed here does not appear to have been treated previously.

Several examples are discussed to illustrate the theory.

For known finite-dimensional results appearing in several works, usually only one is selected here as reference, regardless of priority.

Notation and terminology. If X is a complex separable Banach space, X^* will be its dual and x^* an element of X^* . We shall write indifferently $x^*(x)$ or x^*x . The Banach space of all bounded linear operators from a Banach space U to a Banach space X will be denoted by $\mathcal{B}(U, X)$; in particular, $\mathcal{B}(X)$ will stand for $\mathcal{B}(X, X)$. If $B \in \mathcal{B}(U, X)$, then $B^* \in \mathcal{B}(X^*, U^*)$ will be its adjoint operator. If X is a Hilbert space and $B \in \mathcal{B}(X)$, its Hilbert space adjoint in $\mathcal{B}(X)$ will also be indicated by B^* . $\|\cdot\|$ will denote the norm of whatever space under consideration and O will be its origin. Occasionally, we shall write O_X , O_{X^*} , etc. $\mathcal{R}(\cdot)$ and $\mathcal{N}(\cdot)$ will indicate the range and the null space of the operator (\cdot) . The range of, say, B on U will also be indicated by BU . $L_p[[0, T], U]$ will be the Banach space of U -valued functions with norm $\{\int_0^T \|u(t)\|^p dt\}^{1/p}$, $1 \leq p$. For this and related notions like Bochner integrability, strong and weak measurability, see [18, Chap. 3]. If E_n is a sequence of subspaces, $n = 0, 1, \dots$, then $\text{Cl sp } \{E_n, n \geq 0\}$ will indicate the closure of their span (i.e., totality of finite linear combinations of elements e_n , $e_n \in E_n$). Also, an (abstract) Volterra operator on X will be, as in

² The fact that a scalar admissible controller is here a bounded measurable function, while in [32] it is a function of bounded variation, is not essential.

[16], a compact operator on X , whose spectrum is just the origin. Finally, a vector-valued function of time t is called analytic on $[t_0, t_1]$ in case, as usual, it is representable in a neighborhood of any point \bar{t} in $[t_0, t_1]$ as a convergent power series of $(t - \bar{t})$ (the convergence is intended with respect to the norm of the space containing the range of the function).

2. System and its properties.

2.1. Mathematical model and background. The most general state equation considered in this paper is the following abstract linear differential equation

$$(\mathcal{L}) \quad \dot{x} = A(t)x + B(t)u, \quad t \geq t_0,$$

where x and u are vectors in complex³ separable Banach spaces x (state space) and U (control space), respectively; $A(t)$ and $B(t)$ are, at each $t \geq t_0$, in $\mathcal{B}(X)$ and $\mathcal{B}(U, X)$, respectively, and, for the purpose of what follows, (at least) continuous in t . \dot{x} is the time derivative with respect to the norm of X . Admissible controllers for \mathcal{L} on some finite interval $[t_0, T]$ are all U -valued functions $u(t)$ that are Bochner integrable and have bounded norm $\|u(t)\|$ on $[t_0, T]$. Admissible controllers form a linear space in each Banach space $L_p[[t_0, T], U]$.

Then for each admissible controller $u(t)$, there is a unique solution of the Cauchy problem consisting of the equation \mathcal{L} and $x(t_0) = x_0$ in X , satisfying the equation \mathcal{L} a.e. on $[t_0, T]$. Such a solution can be written in a manner formally analogous to the finite-dimensional case

$$x(t, t_0, x_0, u) = \Phi(t, t_0)x_0 + \int_{t_0}^t \Phi(t, \tau)B(\tau)u(\tau) d\tau,$$

where $\Phi(\cdot, \cdot)$ is a one-to-one and onto operator in $\mathcal{B}(X)$, a solution of the homogeneous equation and satisfying the usual properties [6, Chap. 3]. In the autonomous case: with $A(t) \equiv A$, then $\Phi(t, t_0) = e^{A(t-t_0)}$ (where the operator given by the exponential is expressed by the usual series,

$$I + A(t - t_0) + A^2(t - t_0)^2/2! + \dots$$

convergent in the norm of $\mathcal{B}(X)$). Hence the solution (taking $t_0 = 0$ without loss of generality) becomes

$$x(t, x_0, u) = e^{At}x_0 + \int_0^t e^{A(t-\tau)}Bu(\tau) d\tau,$$

and, for later reference, we notice that ($u(t)$ being Bochner integrable)

$$(2.1.1) \quad \int_{t_0}^t e^{A(t-\tau)}Bu(\tau) d\tau = \sum_{n=0}^{\infty} A^n B \left[\int_{t_0}^t \frac{(t-\tau)^n}{n!} u(\tau) d\tau \right]$$

for every finite interval $t_0 \leq t \leq t_1 < \infty$, as can be checked directly. The case deserves special attention when the control space U is finite-dimensional, say of dimension m , and so is (isometrically isomorphic to) R^m . Then \mathcal{L} can be more conveniently written as

$$(\mathcal{L}_m) \quad \dot{x} = A(t)x + \sum b_i(t)u_i(t), \quad i = 1, \dots, m,$$

³ For the application of the results to the case of practical interest when X and U are *real*, see the remark in [13, p. 398] or in [12, p. 694].

with $b_i(t) = B(t)e_i$, $\{e_i\}$ a basis for U , and u_i components of u . In this case $u(t)$ admissible implies that the scalar functions $u_i(t)$ are measurable and bounded. Conversely, any system \mathcal{L}_m with bounded measurable $u_i(t)$ is a special case of \mathcal{L} with $U = R^m$ and u admissible. For $m = 1$, we shall write b instead of b_1 .

We now augment the description of the previous state equation(s) by providing an output (or observation) equation: $y = H(t)x$, where $H(t)$, $t \geq t_0$, is a bounded linear operator from the Banach space X into the Banach space Y (output space). (Completeness of Y when the system is autonomous will not be needed.) The complex of the state and output equations is referred to as an observed system. The output at time t initiating at time t_0 , due to the control $u(t)$ acting from the initial state x_0 , will be indicated by

$$y(t, t_0, x_0, u) = H(t)x(t, t_0, x_0, u),$$

and the specification of $t_0 = 0$ will be dropped in the autonomous case.

Presumably, the case of greater physical interest is when the output space Y is finite-dimensional, say $Y = R^r$, and so $H(t)$ is an r -tuple of bounded linear functions $H(t) = [h_1(t), \dots, h_r(t)]$, $h_j(t) \in X^*$. This means that, although the state of the process is a distributed quantity, (say, a distribution of temperature), external measurements reveal some (finite) data obtained from the global distribution. Therefore, results pertinent to the case $Y = R^r$ will be particularly stressed and explicitly displayed. The following consequence of the Hahn–Banach theorem will be the main tool exploited in the entire paper; it will be referred to as the fundamental proposition.

FUNDAMENTAL PROPOSITION [22, p. 145]. *Let X be a normed linear space and E an arbitrary set in X . Then $\text{Cl sp } \{E\} = X$ if and only if the zero functional is the only bounded linear functional that vanishes on E .*

2.2. Definitions. We collect here the definitions of all the concepts to be investigated in this paper, with reference to the system \mathcal{L} introduced in the previous section.

DEFINITION 1. The nonautonomous system \mathcal{L} is (completely) state

(i) *approximately controllable* on finite interval $[t_0, T]$ in case: given $\varepsilon > 0$ and two arbitrary initial and final points x_0 and x_1 in X , there is an admissible controller $u(t)$ on $[t_0, T]$ steering x_0 , along a solution curve of \mathcal{L} , to an ε -sphere of x_1 , that is such that $\|x(T, t_0, x_0, u) - x_1\| \leq \varepsilon$. In other words: for each initial point x_0 , the set of all points to which x_0 can be steered by admissible controllers on $[t_0, T]$ is dense in X .

(ii) *exactly controllable* on $[t_0, T]$, if $\varepsilon = 0$ in (i).

Without loss of generality, we can take the initial point $x_0 = 0$, since the translation by $\Phi(T, t_0)x_0$ of a dense subspace (the whole space) is still dense in X (the whole space).

If the state space X is finite-dimensional, the two above definitions coincide, and so they both can be thought of as generalizations to infinite-dimensional spaces of the classical concept of controllability, to which they reduce when $X = R^n$ and $U = R^m$. In fact with X finite-dimensional, the subspace of all states reachable from the origin on $[t_0, T]$ is closed and hence it cannot be dense in X unless it is all of X .

The next two definitions refer to the observed process.

DEFINITION 2. The nonautonomous observed process \mathcal{L} is *completely observable* on the finite interval $[t_0, T]$ in case: for each admissible control on $[t_0, T]$, and for every two responses $x(t)$ and $\bar{x}(t)$ with distinct initial states, the outputs $H(t)x(t)$ and $H(t)\bar{x}(t)$ are distinct.

By linearity of \mathcal{L} , the above definition is equivalent to saying that, for the null input $u(t) \equiv 0$ (free system), the null output $H(t)x(t) \equiv 0$ implies

$$x(t) = \Phi(t, t_0)x_0 \equiv 0$$

i.e., $x_0 = 0$, since the operator $\Phi(t, t_0)$ is one-to-one and onto.

DEFINITION 3. The nonautonomous system \mathcal{L} is (completely) *output*

(i) *approximately controllable* on $[t_0, T]$ in case: given $\varepsilon > 0$, x_0 and y_1 arbitrary in X and Y respectively, there is an admissible controller $u(t)$ on $[t_0, T]$ such that $\|y(T, t_0, x_0, u) - y_1\| \leq \varepsilon$.

(ii) *exactly controllable* on $[t_0, T]$, if $\varepsilon = 0$ in (i).

Again, if Y is finite-dimensional, (i) and (ii) in the last definition coincide. Also, without loss of generality, we will always take $x_0 = 0$ and, if the system is autonomous, also $t_0 = 0$.

3. State controllability. Autonomous case.

3.1. **State approximate controllability.** When $X = R^n$ and $U = R^m$, and so A and B are matrices, \mathcal{L} is controllable if and only if $\text{rank } [B, AB, \dots, A^{n-1}B] = n$, [21]. The generalization of this algebraic test to infinite-dimensional spaces is given by the following characterization, expressed solely in terms of the operators A and B and, in particular, independent on the particular finite interval, (whose mention will therefore be henceforth dropped).

THEOREM 3.1.1. *The autonomous system \mathcal{L} is approximately controllable on $[0, T]$ if and only if*

$$\text{Cl sp } \{A^n B U, n \geq 0\} = X$$

or equivalently, if and only if

$$\bigcap_{n=0}^{\infty} \mathcal{N}\{B^*(A^*)^n\} = \{O_{X^*}\}.$$

Also, approximate controllability of \mathcal{L} is retained by using only continuous U -valued controls (via Lusin's theorem that holds also for functions with values in an abstract space [11, p. 474], [10, pp. 1218]).

COROLLARY 3.1.2. *The autonomous system \mathcal{L}_m is approximately controllable on $[0, T]$ if and only if*

$$\text{Cl sp } \{A^n b_i\} = \text{Cl sp } \{b_1, \dots, b_m, Ab_1, \dots, Ab_m, A^2 b_1, \dots\} = X.$$

Also, approximate controllability of \mathcal{L}_m is retained by using only C^∞ -controllers (say, polynomials) via Lusin's and Weistrass' theorem.

Remarks 3.1.1. Notice that while $\text{sp } \{A^n B U, n \geq 0\} = X$ may occur (e.g., when the range of B is all of X), the subspace $\text{sp } \{A^n b_i\}$ is always properly contained in the infinite-dimensional space X ; otherwise, in fact, the sequence in brackets would be a countable Hamel basis for X , which is impossible [10, p. 74]. The above result also justifies why the space X was taken separable to start with.

Finally, approximate controllability is preserved under a linear change of variable $\bar{x} = Px$, where P is a 1-1 onto operator in $\mathcal{B}(X)$.

We also remark that Theorem 3.1.1 can be further extended, in a rather straightforward way, to cover systems with retarded controls, the finite-dimensional theory of which has been treated in § 3 of [2]. Details will be given elsewhere.

As far as the proof of Theorem 3.1.1 is concerned, we remark that, although the necessary part can be proved directly by elementary means in a few different ways [35], we prefer to present here a more abstract method of proof that has the advantage to work equally well for the necessary as well as the sufficient part. Such a method is based on the fundamental proposition, via the introduction of bounded linear functionals.

Proof of Theorem 3.1.1. The proof hinges on the following three steps that are worth pointing out:

(a) \mathcal{L} is approximately controllable on $[0, T]$ if and only if $x^*(e^{A(T-t)}Bu) \equiv 0$ on $[0, T] \times U$ implies $x^* = 0$ (equivalently, if and only if $B^*e^{A^*(T-t)}x^* \equiv 0$ on $[0, T]$ implies $x^* = 0$; in this form see also [13]).

Only if: Otherwise, in fact, there is a nonzero $\bar{x}^* \in X^*$ such that the subspace K_T of all states reachable from the origin on $[0, T]$ is contained in the (closed) hyperplane $\{x: x \in X, \bar{x}^*(x) = 0\}$, which is not dense in X by the fundamental proposition.

If: If the subspace K_T is not dense in X , then $\bar{x}^*(K_T) = 0$ for some nonzero \bar{x}^* in X^* and this easily leads to $\bar{x}^*(e^{A(T-t)}Bu) = 0$ on $[0, T] \times U$.

(b) $x^*(e^{At}Bu) \equiv 0$ on $[0, \infty) \times U$ implies $x^* = 0$ if and only if: $x^*(A^n Bu) = 0$, $n = 0, 1, \dots$, all $u \in U$, implies $x^* = 0$. (Equivalently, $B^*e^{A^*t}x^* = 0$ on $[0, \infty)$ implies $x^* = 0$ if and only if $B^*(A^*)^n x^* = 0$, $n = 0, 1, \dots$, implies $x^* = 0$.)

The proof follows easily, e.g., arguing by contradiction with a nonzero x^* : set $t = 0$, differentiate in t setting $t = 0$ again, etc. in one direction; use the series expansion for the exponential in the other direction.

(c) Theorem 3.1.1 follows now from (a) and (b) via the fundamental proposition. Q.E.D.

Remark 3.1.2. One easily convinces oneself that the condition in Theorem 3.1.1 is necessary (and sufficient) also for a more relaxed definition of approximate controllability, in which the final (finite) instant T is not fixed in advance, but is left free, possibly depending on the pair of initial and final points x_0 and x_1 . In this case, in fact, t in part (a) ranges over $[0, \infty)$, precisely as in part (b).

One also expects that some standard corollaries to the algebraic test for controllability in the finite-dimensional theory, as e.g., Corollaries 1, 2 and 3 in [26, pp. 83-85], can be extended in the "approximate" (i.e., dense) sense to infinite-dimensional spaces via Theorem 3.1.1.

This is indeed the case. We report here only one such result involving the domain of null controllability [26, p. 84], and refer for other results and detailed proofs to the author's dissertation [35].

COROLLARY 3.1.3. *Let the autonomous system \mathcal{L} be approximately controllable. Assume also that $u = 0$ lies in the interior of the control restraint set (in U) and that the spectrum of the operator A lies in the open left-half plane of the complex plane. Then the closure of the domain of null controllability is the entire space X .*

Also, from Theorem 3.1.1, we can easily get an asymptotic result.

COROLLARY 3.1.4. *Let the system \mathcal{L} be approximately controllable and let the real function $\|e^{At}\|$ be bounded for $t \geq 0$ (this implies that the spectrum of A lies in the closed left-half plane in the complex plane [6, § 1.4.2]). Then, given $\varepsilon > 0$ and an arbitrarily small time interval $[0, T]$, each point x_0 in X can be steered to the ε -sphere of the origin over $[0, T]$ and kept herein thereafter by a controller that is admissible in $[0, T]$ and identically zero afterwards.*

Even in the particular case in which the system in question is \mathcal{L}_1 and the state space X is Hilbert, the above corollary improves Theorem 2 in [32], yielding a stronger conclusion under weaker assumptions (see the definition of “basis” in [32]).

The asymptotic behavior of \mathcal{L} is studied in more detail in [35] and will not be reported here.

We close this section by presenting the decomposition problem. It is well known that in the classical finite-dimensional theory, an arbitrary linear system can be uniquely decomposed in a controllable and uncontrollable part [26, p. 110]. An analysis of the argument employed to derive such a decomposition, shows that it relies on three factors: (a) the domain \mathcal{C} of null controllability with no control restraint is a subspace of X [26, p. 97]. (b) existence of a projector P (i.e., linear, with $P^2 = P$) projecting X onto \mathcal{C} ; (c) such a projector is continuous, so that $x_c = Px$, $x'_c = (I - P)x$ imply $\dot{x}_c = P\dot{x}$ and $\dot{x}'_c = (I - P)\dot{x}$. The boundedness of the projector P in turn follows [34, p. 242] from the fact that both the range and the null space of P , being finite-dimensional, are always closed. When the state space X is an infinite-dimensional Banach space, (a) is still valid, as one can convince oneself by examining [26, p. 97]. Consequently (b) also holds true [34, p. 241]. However, now \mathcal{C} need not be closed and, more important than that, even if we replace \mathcal{C} by its closure $\bar{\mathcal{C}}$ the projector operator of X onto $\bar{\mathcal{C}}$ need not be continuous if X is a general Banach space, even if reflexive [34, p. 242]. On the contrary, if X is instead a Hilbert space, the projector operator P of X onto $\bar{\mathcal{C}}$ is not only bounded but also orthogonal ($P = P^*$). Therefore, if one attempts to generalize the decomposition theorem [26, Thm. 10, p. 97] to systems defined on infinite-dimensional spaces, the distinction between a Hilbert or a Banach state space X is fundamental. In the former case, the wanted generalization is given by the following theorem. Its proof follows the classical one as in [26] with appropriate modifications and is therefore omitted (for details see [35]).

THEOREM 3.1.5. *Consider the autonomous system \mathcal{L} defined on a Hilbert space X . Then there exists a subspace \mathcal{C} of X , precisely the domain of null controllability of \mathcal{L} with no control restraint, such that*

(i) *no point of \mathcal{C} can be steered out of \mathcal{C} and no point out of $\bar{\mathcal{C}}$ can be steered into $\bar{\mathcal{C}}$,*

(ii) *\mathcal{L} , restricted on $\bar{\mathcal{C}}$, is approximately controllable. Moreover $\bar{\mathcal{C}}$ is unique and there exists an orthogonal projector P , such that $\bar{\mathcal{C}}$ is just equal to PX , $\bar{\mathcal{C}} = PX$, and \mathcal{L} can be written as*

$$\dot{x}_c = A_1 x_c + A_2 x'_c + PBu,$$

$$\dot{x}'_c = A_3 x'_c$$

with $x_c = Px$ and $x'_c = (I - P)x$; A_1 is the operator PA restricted over $\bar{\mathcal{C}}$, while A_2 and A_3 are, respectively, the operators PA and $(I - P)A$, each restricted over the orthogonal complement of $\bar{\mathcal{C}}$. Also

$$\text{Cl sp } \{A^n BU, n \geq 0\} = \bar{\mathcal{C}}.$$

3.2. Examples of approximate controllable systems.

Approximately controllable systems \mathcal{L}_1 . The condition: $\text{Cl sp } \{A^n b, n \geq 0\} = X$ that characterizes approximate controllability for the system $\mathcal{L}_1: \dot{x} = Ax + bu$ with just one scalar controller is often referred to, in the mathematical literature, as cyclicity of the pair $\langle A, b \rangle$: A is a cyclic operator (on X) having b as cyclic vector.

Example 3.2.1. Consider the following integro-differential equation of Volterra type:

$$(3.2.1) \quad \frac{\partial w(t, \xi)}{\partial t} = \int_0^\xi w(t, s) ds + v(\xi)u(t)$$

for the scalar function $w(t, \xi)$ in two scalar variables t and ξ , say, $0 \leq \xi \leq 1$, and where $v(\xi)$ is a given scalar function, through which the scalar control $u(t)$ affects the solution. Equations similar to (3.2.1) arise in some chemical engineering processes [1]. Let us distinguish two cases.

(i) If the function $v(\xi)$ is continuous, then we take $X = C[0, 1]$ and one checks that (3.2.1) can be written in the abstract form $\dot{x} = Ax + bu$, by taking the state $x(t)$ to be the function $w(t, \cdot)$, the vector b to be the function $v(\cdot)$ and the operator A to be the (Volterra) integral operator $(Af)(\xi) = \int_0^\xi f(s) ds$ defined on $C[0, 1]$.

Then the necessary and sufficient condition for the process (3.2.1) to be approximately controllable (i.e., for $\langle A, b \rangle$ to be a cyclic pair on $C[0, 1]$) is simply that $v(0) \neq 0$ [8].

(ii) More generally, if the given function $v(\xi)$ is p -integrable, then we take $X = L_p[0, 1]$, $1 \leq p < \infty$, while $x(t)$, b and A are defined formally as before (use [30, p. 91] to get $\dot{x}(t) = \partial w(t, \xi)/\partial t$; also [18, pp. 68–70]).

Then the necessary and sufficient condition for (3.2.1) to be approximately controllable (i.e., for $\langle A, b \rangle$ to be a cyclic pair on $L_p[0, 1]$) is that for any $\delta > 0$, $v(\xi) \neq 0$ on a subset of positive measure of $[0, \delta]$ (equivalently, zero belongs to the support of the function $v(\xi)$) [8], [20], [16, p. 37].

As we see, the test for approximate controllability of the process (3.2.1) becomes a trivial matter.

If the interval of integration $[0, \xi]$ in (3.2.1) is replaced by $[\xi, 1]$, then the characterization for approximate controllability becomes: for any $\delta > 0$, $v(\xi) \neq 0$ on a subset of positive measure of $[1 - \delta, 1]$ (equivalently, one belongs to the support of $v(\xi)$) [19, Thm. 3], [16, pp. 37–38].

The previous example offers the opportunity to pause and make the following comments regarding perturbation.

Remark 3.2.1. Recall first that, when X and U are finite-dimensional, controllable systems are open and dense in the totality of linear autonomous systems [26, p. 100]. The same result was recently extended to nonautonomous systems in

the L_p -norm [7]. However, when the state space X is infinite-dimensional, we can easily deduce from the above example, defined, say, on $X = L_2[0, 1]$, that the openness property fails, even if the operator A is left unperturbed and A is (Volterra and) a proper contraction ($\|A\| = 2/\pi < 1$ [17, p. 300]); so that, if $A^n b$ span all of X and $\varepsilon > 0$ is given, we can find b' , with $\|b - b'\| < \varepsilon$ (and hence $\|A^n b - A^n b'\| < \varepsilon$, $n = 0, 1, \dots$) and yet $A^n b'$ will not span X . (In what follows, we write the vectors b and b' in X as functions $v_b(\cdot)$ and $v_{b'}(\cdot)$ in $L_2[0, 1]$.)

In fact, let the process (3.2.1) with $v(\xi) = v_b(\xi)$ be approximately controllable, so that the function $v_b(\xi)$ is different from zero a.e. on some initial interval $[0, \delta]$, $0 < \delta \leq 1$. Next, consider a function $v_{b'}(\xi)$, identical to $v(\xi)$ except on an initial interval, where $v_{b'}(\xi)$ is set identically zero. By selecting the initial interval suitably small, the function $v_{b'}(\xi)$ can be made arbitrarily close to $v_b(\xi)$ in the norm of X and yet the process (3.2.1) with $v(\xi) = v_{b'}(\xi)$ is not approximately controllable.

Using similar considerations, one should notice, however, that in this case the density property does hold, (with A unperturbed); that is, if the process (3.2.1), is not approximately controllable with $v(\xi) = v_{b'}(\xi)$, then it does become such with a function $v(\xi) = v_b(\xi)$ arbitrarily close to $v_{b'}(\xi)$ in the norm of X .

Equally easy tests for the approximate controllability of other more complicated integro-differential equations of Volterra type, with nontrivial kernels under the integration sign, are given next. They are drawn from the scattered literature on cyclic operators.

(An attempt to gather information in cyclic operators with implications on control problems has been made in the Appendix of [35], where references are also given.)

Example 3.2.2. Consider the integro-differential equation of Volterra type

$$(3.2.2) \quad \frac{\partial w(t, \xi)}{\partial t} = \frac{1}{\Gamma(\alpha)} \int_0^\xi (\xi - s)^{\alpha-1} w(t, s) ds + v(\xi)u(t),$$

where the solution $w(t, \xi)$ is sought, say, in $L_2[0, 1]$ in ξ , $0 < \alpha < \infty$, and $v(\xi)$ is a given function in $L_2[0, 1]$. The integral is the Riemann–Liouville integral of fractional order. In this case the operator A is given by the following Volterra operator:

$$(Af)(\xi) = \frac{1}{\Gamma(\alpha)} \int_0^\xi (\xi - s)^{\alpha-1} f(s) ds, \quad f(\xi) \in L_2[0, 1].$$

Introduce the quantity $l_f (\geq 0)$ defined by

$$\int_0^{l_f} |f(\xi)|^2 d\xi = 0, \quad \int_{l_f}^{l_f + \varepsilon} |f(\xi)|^2 d\xi > 0, \quad \varepsilon > 0.$$

Then for the process (3.2.2), the condition of approximate controllability of Corollary 3.1.2 is simply equivalent to $l_v = 0$ [16, p. 399].

Example 3.2.3. A general kernel is now presented. Consider the integro-differential equation of Volterra type

$$(3.2.3) \quad \frac{\partial w(t, \xi)}{\partial t} = \int_\xi^1 F(\xi, s) w(t, s) ds + v(\xi)u(t)$$

with solution $w(t, \xi)$ sought in $L_p[0, 1]$, $1 < p < \infty$, in ξ ; here $v(\xi)$ is a given

function in $L_p[0, 1]$ and the kernel $F(\xi, s)$ satisfies any of the technical conditions of Theorem 3 in [19, p. 492]:

- (i) $F(\xi, s)$ is C^2 and $F(s, s)$ real-valued and $\neq 0$;
- (ii) $F(\xi, s) = (s - \xi)^{\alpha-1}G(\xi, s)$ is analytic, α a positive integer, $G(s, s)$ real-valued and $\neq 0$;
- (iii) $F(\xi, s) = F(s - \xi) = (s - \xi)^{\alpha-1}k(\xi - s) + n(\xi - s)$, α order of F ; $k \in C^2$, $n \in L_1[0, 1]$ and $n = 0$ in a neighborhood of $\xi = s$.

Then, for the process (3.2.3), the condition of approximate controllability given in Corollary 3.1.2 becomes equivalent to the requirement that for any $\delta > 0$, $v(\xi) \neq 0$ on a subset of positive measure of $[1 - \delta, 1]$ (equivalently, one belongs to the support of $v(\xi)$).

Further information, sometimes in the form of weaker sufficient conditions on the kernel $F(\xi, s)$ to make the above process (3.2.3) approximately controllable with a suitable $v(\xi)$ (and $p = 2$), is given in [4, p. 237], [38], [14], [29].

Example 3.2.4. In all the above examples (as well as in the following example 3.2.5), the operator acting on the state of the system \mathcal{L}_1 is unicellular (an operator on X is unicellular in case, of any two distinct invariant subspaces, one is always contained in the other, and so all the invariant subspaces are nested; unicellular operators form a special class of cyclic operators [33, p. 140], [16, p. 36], [14, p. 192]).

In order to enlarge the class of systems \mathcal{L}_1 that may become approximately controllable, with a suitable choice of the vector b (at least when defined on a Hilbert space), the following considerations are in order:

The Volterra integral operator $(Af)(\xi) = \int_0^\xi f(s) ds$ on $L_2[0, 1]$ (or its adjoint [4, p. 113] $(A^*f)(\xi) = \int_\xi^1 f(s) ds$) can be considered to be a "canonical form" within a certain class of operators (simple, dissipative, with one-dimensional imaginary component), in the sense that each operator of this class is unitarily equivalent to a scalar multiple of it [16, p. 33], [33, p. 363].

Since the unicellularity of A (or A^*) is preserved under unitary equivalence [16, p. 36], all the operators in this class are unicellular, hence cyclic. Therefore their corresponding dynamical systems \mathcal{L}_1 can be made approximately controllable. Moreover, as it follows from a general result [33, p. 140], the totality of vectors b , for which one such system is not approximately controllable, form a set of first category. Compare with Remark 3.2.1.

We now wish also to present two other examples of approximately controllable systems \mathcal{L}_1 , describing systems of countably many linear differential equations.

Example 3.2.5. Consider the system $\dot{x} = Ax + bu$ defined on $X = l_p$, $1 \leq p \leq \infty$, $x = [x_0, x_1, \dots]$, $\dot{x} = [\dot{x}_0, \dot{x}_1, \dots]$, $b = [b_0, b_1, \dots]$ while A is the operator, whose matrix representation, with respect to the usual basis, has the following entries: positive nonincreasing numbers λ_j , $j = 1, 2, \dots$, such that $\sum \lambda_j^\alpha < \infty$ for some α ($0 < \alpha < \infty$) in the diagonal immediately below the main diagonal and zero elsewhere. Such an operator A is compact, unicellular [16, p. 404] (hence simple and Volterra with $p = 2$, [16, p. 35]) and hence cyclic. Moreover, for the above system, the condition of approximate controllability of Corollary 3.1.2 holds if and only if [28] $b_0 \neq 0$. See also [8] for $\lambda_j = 2^{-j}$.

Example 3.2.6. Consider the system $\dot{x} = Ax + bu$ defined on $X = l_2$ with x, b, \dot{x} as in the previous example, while A is now the (bounded but noncompact)

right shift operator: $Ae_i = e_{i+1}$, $i = 0, 1, \dots$, $\{e_i\}$ basis for X , whose matrix representation has the entry 1 in the diagonal immediately below the main diagonal and zero elsewhere [34, p. 266].

While the condition $b_0 \neq 0$ is obviously necessary for approximate controllability (on any l_p), to express the necessary and sufficient condition of Corollary 3.1.2 in terms of the coordinates of b is more delicate. To do this conveniently, one introduces the Hardy space $H^2 = H^2(D)$ of all analytic functions

$$f(z) = \sum c_i(f)z^i$$

for which $\sum |c_i|^2 < \infty$. Every such function is automatically holomorphic in the open unit disc D . Then l_2 is mapped unitarily onto H^2 [16, p. 403], by associating with each vector x in l_2 , $x = \sum x_i e_i$, the function $f_x(z)$ in H^2 , $f_x(z) = \sum x_i z^i$. Then the above system is approximately controllable (i.e., the function f_b is cyclic, also "outer" in Beurling terminology) if and only if [3, p. 245]

$$\ln |f_b(0)| = \ln |b_0| = \frac{1}{2\pi} \int_0^{2\pi} \ln |f_b(e^{i\theta})| d\theta.$$

If $f_b(z)$ does not have zeros in D , the above condition (Poisson's formula) always holds [39, p. 205].

Before turning to the case $m > 1$, we remark that very technical necessary and sufficient conditions for some abstract bounded operators in Hilbert space to be unicellular are given in [16, §2 in Appendix], [4, pp. 34, 166, 187, 237], [33, pp. 363, 371].

We next present, as promised in the Introduction, a nontrivial example, showing the application of our results with bounded operators to the general case when the operator acting on the state is unbounded and generates a strongly continuous semigroup. More precisely, as explained in (a) in the Introduction, the characterization for approximate controllability in finite time of the original system with an unbounded operator will be found from the equivalent condition of approximate controllability on $[0, T]$ of the associated system with the resolvent operator acting on the state.

Example 3.2.7.

(a) Let $X = C_0[0, 1]$ be the subspace of $C[0, 1]$ of all continuous functions vanishing at the origin:

$$C_0[0, 1] = \{f: f \in C[0, 1], f(0) = 0\}.$$

Now let A be the (first order) differential operator defined by $Af = -f'$, with domain

$$D(A) = \{f: f \in C_0[0, 1]; f' \text{ exists and belongs to } C_0[0, 1]\}.$$

A acts from $D(A)$ onto $C_0[0, 1]$. Notice that:

- (i) $D(A)$ is dense in $C_0[0, 1]$ with respect to the usual uniform norm: $\|f\| = \max \{|f(\xi)|, 0 \leq \xi \leq 1\}$;
- (ii) A is linear and closed;
- (iii) the spectrum $\sigma(A)$ of A is empty;

(iv) the resolvent $R(\lambda, A) = (\lambda I - A)^{-1}$ is an entire function of λ given by

$$\begin{aligned} [R(\lambda, A)v](\xi) &= \int_0^\xi e^{-\lambda(\xi-s)}v(s)ds = \int_0^\xi e^{-\lambda\tau}v(\xi-\tau)d\tau \\ &= \int_0^\xi e^{-\lambda\tau}S(\tau)v(\xi)d\tau = \int_0^\infty e^{-\lambda\tau}S(\tau)v(\xi)d\tau, \quad \tau \geq 0, \end{aligned}$$

where $S(\tau)$, $\tau \geq 0$, is the strongly continuous semigroup of translations defined on $C_0[0, 1]$ by

$$[S(\tau)v](\xi) = \begin{cases} v(\xi - \tau) & \text{for } 0 \leq \tau \leq \xi, \\ 0 & \text{for } \xi < \tau. \end{cases}$$

The above relationship between $R(\lambda, A)$ and $S(\tau)$ holds also as a Bochner integral on X . Using Corollary 16, p. 627 in [10], we then see that $S(\tau)$ is the semigroup with infinitesimal generator A .

Since the spectrum $\sigma(A)$ is empty, we can take, for example, $R(0, A)$ and hence, as explained in the Introduction, the abstract system

$$(3.2.4) \quad \dot{x} = Ax + bu \quad \text{on } X = C_0[0, 1]$$

is approximately controllable in finite time if and only if the associated system

$$\dot{x} = R(0, A)x + bu \quad \text{also on } X = C_0[0, 1]$$

is. Using Corollary 3.1.2 with $m = 1$, this last condition is equivalent to

$$(3.2.5) \quad \text{Cl sp } \{R^n(0, A)b, n \geq 0\} = C_0[0, 1].$$

But $R(0, A)$ is the Volterra integral operator

$$[R(0, A)v](\xi) = \int_0^\xi v(s)ds$$

defined on $\mathcal{R}(A) = C_0[0, 1]$. According to a result of [8], already used in Example 3.2.1, the totality of closed invariant subspaces for the same integral operator, when defined on the space $C[0, 1]$, is given by the subspaces \mathcal{M}_a , consisting of all continuous functions vanishing on $[0, a]$. It follows that if we indicate by b the vector of $C_0[0, 1]$ corresponding to the function $v(\xi)$, the condition (3.2.5) holds if and only if, for any $\delta > 0$, one has $v(\xi) \neq 0$ in some subinterval of positive length of $(0, \delta]$. We now interpret our results in explicit form. Consider the one-scalar-control system modeled by the following partial differential equation:

$$\frac{\partial w(t, \xi)}{\partial t} = -\frac{\partial w(t, \xi)}{\partial \xi} + v(\xi)u(t)$$

with $v(\xi)$ in $C_0[0, 1]$ and with solution $w(t, \cdot)$ sought in $C_0[0, 1]$, corresponding to the abstract form (3.2.4). Such a system is approximately controllable in finite time if and only if, for any $\delta > 0$, $v(\xi) \neq 0$ in some subinterval of $(0, \delta]$.

(b) Similar considerations can be made on the space $X = C_1[0, 1]$ of all continuous functions $f(\xi)$ vanishing at $\xi = 1$, with respect to the (first order) differential operator A defined by $Af = f'$, with domain $D(A) = \{f: f \in C_1[0, 1];$

f' exists and belongs to $C_1[0, 1]$. See [18, p. 537]: A is a closed linear operator from $D(A)$ onto $C_1[0, 1]$, with dense domain, and is the infinitesimal generator of a strongly continuous semigroup. The spectrum of A is empty and the resolvent $R(0, A) = (-A)^{-1}$ is given by

$$[R(0, A)v](\xi) = \int_0^{1-\xi} v(\xi + \tau) d\tau = \int_{\xi}^1 v(s) ds,$$

which is precisely the same operator (except defined on $C_1[0, 1]$) considered at the end of Example 3.2.1.

Approximately controllable systems \mathcal{L}_m , $m > 1$. A simple, dissipative, Volterra operator on Hilbert space X with nuclear imaginary component of p -dimensional range ($p < \infty$) is [16, p. 353][4, p. 209] [38]

(i) either unicellular, and so there is a vector b in X , such that the pair $\langle A, b \rangle$ is approximately controllable;

(ii) or decomposes (nonuniquely) into a quasi-direct sum $A = A_1 \dot{+} \cdots \dot{+} A_m$ of unicellular operators A_i , $i = 1, \dots, m \leq p$. This means that there are subspaces X_i , invariant under A , satisfying the conditions:

(a) the subspaces X_i are linearly independent (i.e., $f_1 + \cdots + f_m = 0$, $f_i \in X_i$, implies $f_1 = f_2 = \cdots = f_m = 0$); (b) the span of the subspaces X_i is dense in X ; (c) the operator A_i induced by A on X_i is unicellular. Accordingly, one writes

$$X = X_1 \dot{+} \cdots \dot{+} X_m.$$

In this second case, there are vectors b_i in X_i such that $\text{Cl sp } \{A^n b_i, n \geq 0\} = X_i$ and hence

$$\text{Cl sp } \{A^n b_i, 1 \leq i \leq m; n \geq 0\} = X$$

since

$$\begin{aligned} X &= \text{Cl sp } \{X_i, 1 \leq i \leq m\} = \text{Cl sp } \{\text{Cl sp } \{A^n b_i, n \geq 0\}, 1 \leq i \leq m\} \\ &= \text{Cl sp } \{A^n b_i, 1 \leq i \leq m; n \geq 0\}. \end{aligned}$$

Consequently, the correspondent system \mathcal{L}_m is approximately controllable.

Another example of an approximately controllable system \mathcal{L}_m is obtained by defining the bounded but noncompact operator A by: $Ae_i = e_{i+m}$, $i = 1, 2, \dots$, and taking $b_i = e_i$, $i = 1, \dots, m$.

It would be commendable to single out from the class of operators referred to above, those that give rise to physically significant dynamical systems.

Approximately controllable systems \mathcal{L} . Let $X = U$ be a Hilbert space and let A be a simple operator [16, p. 29] (called also completely non-self-adjoint [4, p. 3], that is A and A^* do not have a common invariant subspace on which they coincide) and let $B = A - A^*$. Then $\text{Cl sp } \{A^n BU, n \geq 0\} = X$ [16, p. 30] and so the pair $\langle A, B \rangle$ is approximately controllable.

In particular, the Volterra integral operator defined on

$$X = L_2[0, 1] \quad \text{by } (Af)(\xi) = \int_0^{\xi} f(s) ds$$

is simple and $(A^*f)(\xi) = \int_{\xi}^1 f(s) ds$. Then, the corresponding integro-differential equation of Volterra type

$$(3.2.6) \quad \frac{\partial w(t, \xi)}{\partial t} = \int_0^{\xi} w(t, s) ds + \int_0^1 \mu(t, s) ds - 2 \int_{\xi}^1 \mu(t, s) ds,$$

with scalar functions w and μ such that $x(t) = w(t, \cdot)$ and $u(t) = \mu(t, \cdot)$ are both in $L_2[0, 1]$ for each $t (\geq 0)$ fixed, is approximately controllable. However, as will follow from Theorem 3.3.1, it is not exactly controllable, since the operator B is compact. The same conclusion applies to every pair $\langle A, B \rangle$, with A simple and compact, $B = A - A^*$, defined on a Hilbert space $X = U$.

Of course, the approximate controllability tests for the systems \mathcal{L}_1 given in the previous part of this section can be profitably used to decide the approximate controllability of systems \mathcal{L} having the same operator A .

For instance one can establish the approximate controllability of (3.2.6) on any $X = L_p[0, 1]$, $1 \leq p < \infty$, even if the controllers are restricted to be constant in the space coordinate. In fact, with $\mu(t, \cdot) \equiv \mu(t)$, the control action on the integro-differential equation is merely given by: $v(\xi)\mu(t)$, with $v(\xi) = [-1 + 2\xi]$, so that the test in Example 3.2.1 applies.

Similarly, if in (3.2.6) the control action is replaced by $\int_I \mu(t, s) ds$, with $I = [0, 1]$, or $[0, \xi]$, or $[\xi, 1]$, the process is still approximately controllable on any $X = L_p[0, 1]$ with controls that are constant in the space coordinate.

Remark 3.2.2. While in the classical finite-dimensional theory, controllability of the pair of (real) matrices $\langle A, B \rangle$ is equivalent to the possibility of selecting a (real) matrix D such that the spectrum of $A + BD$ is preassigned [37] compatible with its reality, this property fails in infinite dimension, where now D is an operator in $\mathcal{B}(X, U)$. In fact all the above are examples of approximately controllable systems with both A and B compact operators and yet, for each D in $\mathcal{B}(X, U)$, the spectrum of the compact operator $A + BD$ always contains the origin.

3.3. State exact controllability. A direct study of some of the approximately controllable systems \mathcal{L}_m given in the previous section, shows [35, § 2.6] that these systems are however not exactly controllable on any finite interval: even more, an example is considered in [35, § 2.6] of a system \mathcal{L}_1 , defined on $X = L_2[0, 1]$, where the sequence b, Ab, A^2b, \dots is an orthonormal basis (stronger condition than approximate controllability) and yet the system is not exactly controllable on any finite interval, since all the functions reachable from the origin in finite time by means of admissible controllers are always continuous functions. All this advances the conjecture that the system \mathcal{L}_m , having a finite number of controls, can never be exactly controllable on a finite interval $[0, T]$, when the state space X is infinite-dimensional. This is indeed the case and in fact even more is contained in the following.⁴

THEOREM 3.3.1. *The autonomous system \mathcal{L}_m , defined on an infinite-dimensional Banach space X , can never be exactly controllable on any (fixed) finite interval $[0, T]$. More precisely,*

(i) *the subspace K_T of all points reachable from the origin on $[0, T]$ by $L_1[0, T]$ -control functions does not fill all of X ;*

⁴ See also Remark 10.2, p. 207 in [41].

(ii) given an arbitrary initial point x_0 (final point x_1) there is a point x_1 in X (x_0 in X) that is never reachable from x_0 (steerable to x_1) by any $L_1[0, T]$ -control function.

The negative assertion (i) can be even strengthened as follows:

(iii) the set of all points reachable from the origin in any finite time (possibly depending on the particular point to reach) by locally L_1 -control functions does not fill all of X . In symbols: $\bigcup K_T \neq X$, where the union is taken over all T in $[0, \infty)$.

Remark 3.3.1. $\bigcup K_T$ is the domain of null controllability with no control restrictions for the system $\dot{x} = -Ax - \sum b_i u_i$ that can be shown, as in the case when X is finite-dimensional [26, p. 98], to be a subspace. Compare Theorem 3.3.1 with Corollary 3.1.3.

Proof. (i) We first note that the operator

$$u_i \rightarrow \int_0^T e^{A(T-t)} b_i u_i(t) dt$$

from $L_1[0, T]$ into X is compact [9, p. 369], [10, p. 507] since $e^{A(T-t)} b_i$ describes a compact set in X , for $0 \leq t \leq T$. Hence the operator Q_T defined by

$$Q_T u = x(T, 0, u) = \sum_{i=1}^m \int_0^T e^{A(T-t)} b_i u_i(t) dt$$

from $L_1[[0, T], R^m]$ into X is also compact.

Therefore the image $K_n(T)$ under Q_T of the sphere in $L_1[[0, T], R^m]$ of radius n , centered at the origin, is a precompact set. Let $\bar{K}_n(T)$ be its closure. Since X is infinite-dimensional, $\bar{K}_n(T)$ cannot contain spheres [40, p. 31] and hence is nowhere dense in X . Next observe that exact controllability on $[0, T]$ demands $X = \bigcup_{n=1}^{\infty} K_n(T)$, but this is impossible by the Baire category theorem [30, p. 139]. Consequently the subspace $K_T = \bigcup_{n=1}^{\infty} K_n(T)$, range of the operator Q_T , does not fill all of X . Actually, even $\bigcup_{n=1}^{\infty} \bar{K}_n(T)$ is not all of X .

(ii) Let x_0 be given and let $\bar{x} \notin K_T$. Then the point x_1 in X defined by $x_1 = e^{AT} x_0 + \bar{x}$ cannot be reached from x_0 by $L_1[[0, T], R^m]$ controls. If x_1 is given, then define x_0 by $x_0 = e^{-AT}(x_1 - \bar{x})$ and hence x_0 is not steerable to x_1 over $[0, T]$.

(iii) Take the sequence of time intervals $[0, i]$, $i = 1, 2, \dots$. Then part (i) says that the subspace $K_i = \bigcup_{n=1}^{\infty} K_n(i)$ is a set of first category. Consequently, $K = \bigcup_{i=1}^{\infty} K_i$ is also a set of first category [30, p. 140] and hence K does not fill all of X . But: $i - 1 < T < i$, $i = 1, 2, \dots$, implies

$$K_n(i - 1) \subset K_n(T) \subset K_n(i), \quad n = 1, 2, \dots, \quad i = 1, 2, \dots,$$

(since a point reachable from the origin over $[0, t_1]$ using $u(t)$ is also reachable over a larger interval $[0, t_2]$ by applying first $u_1(t) \equiv 0$, $0 \leq t < t_2 - t_1$, and then $u_2(t) = u(t - (t_2 - t_1))$, $t_2 - t_1 \leq t \leq t_2$). Hence, from the above inclusions, taking first the union over all n , then the union over all T in $[i - 1, i]$ and finally the union over all i , one arrives at

$$K = \bigcup_{i=1}^{\infty} K_{i-1} \subset \bigcup K_T \subset \bigcup_{i=1}^{\infty} K_i = K.$$

Hence $\bigcup K_T = K$ and $\bigcup K_T$ does not fill all of X . Actually, even $\bigcup_T [\bigcup_{n=1}^{\infty} \bar{K}_n(T)]$ is not all of X . Q.E.D.

Remark 3.3.2. The assumption that A be bounded in the first part of the argument in the proof of Theorem 3.3.1 is not crucial and can be replaced by the weaker one that A be (closed, linear, with dense domain, and) the infinitesimal generator of a strongly continuous semigroup $S(t)$, $t \geq 0$. Then $S(T-t)b_i$ still describes a compact set in X , for $0 \leq t \leq T$ and so the operator

$$u_i \rightarrow \int_0^T S(T-t)b_i u_i(t) dt$$

from $L_1[0, T]$ into X is again compact.

Hence the previous Theorem 3.3.1 holds true also for a control process defined by the integral version

$$x(t, x_0, u) = S(t)x_0 + \int_0^t S(t-\tau) \sum b_i u_i(\tau) d\tau.$$

Notice, however, that for A unbounded and generating $S(t)$, the differential equation

$$\dot{x} = Ax + \sum b_i u_i$$

admits indeed the above integral expression as its unique solution, at least when $u(t)$ is C^1 and x_0 is in the domain of A [40, p. 486]. That such a differential equation cannot be exactly controllable in finite time (with C^1 -controllers) is plain, since its solution is, at each t , contained in the domain of A which is only dense in X . Similar comments apply to the more general process of the subsequent Theorem 3.3.3. Next, we have a refinement of the previous theorem.

COROLLARY 3.3.2. *The autonomous system \mathcal{L} , defined on an infinite-dimensional Banach space X can never be exactly controllable on any finite interval by using locally L_1 -controls, if the operator $B: U \rightarrow X$ is of finite-dimensional range.*

Proof. We omit the details, see [35, § 2.13]. Write $Bu(t) = b_1 c_1(u(t)) + \dots + b_m c_m(u(t))$, where m is the dimension of the range of B . Then one shows that all the vectors $u(\cdot)$ in the unit sphere of $L_1[[0, T], U]$ give rise to coefficients $c_i(u(t))$ contained in a finite sphere of $L_1[[0, T], R^m]$. Then proceed as in part (i) of the previous theorem via the Baire category theorem. Q.E.D.

We now combine the above Corollary 3.3.2 with an approximation technique to extend the conclusion of Theorem 3.3.1 to the important case when the operator B is compact.

THEOREM 3.3.3. *The conclusions (i), (ii) and (iii) of Theorem 3.3.1 hold also for the autonomous system \mathcal{L} , defined on an infinite-dimensional Banach space X , under the assumption that the operator $B: U \rightarrow X$ is compact and that X has a Schauder basis.*

Proof. We just need to prove conclusion (i), since (ii) and (iii) will then follow as in Theorem 3.3.1. Now, since B is compact and X has a basis, there is [10, p. 515] a sequence B_k of operators: $U \rightarrow X$, of finite-dimensional range, converging, in the uniform topology, to B . Define the operators Q_k and Q by

$$Q_k u = \int_0^T e^{A(T-t)} B_k u(t) dt, \quad k = 1, 2, \dots,$$

and

$$Qu = \int_0^T e^{A(T-t)} Bu(t) dt,$$

both from $L_1[[0, T], U]$ to X .

By the previous Corollary 3.3.2, Q_k is compact. Moreover, Q_k converges in the uniform topology to Q , and in fact

$$\|Q_k - Q\| \leq e^{\|A\|T} \|B_k - B\|.$$

Therefore Q is compact. The usual Baire category argument as in Theorem 3.3.1 applies and (i) is proved. Q.E.D.

Remark 3.3.3. The above theorem acquires a greater significance since, in light of the example in § 3.3, there are classes of approximately controllable systems that yet are not exactly controllable; e.g., all those given by compact, simple operators A , with $B = A - A^*$, defined on a Hilbert space $X = U$.

It is of interest in some cases [35, Chap. 5] to know a class of easily recognizable points in X , that cannot be steered to the origin (or reachable from the origin) in any finite time by admissible controllers. Information in this direction is provided by the following theorem (and subsequent remark) and should be viewed as a complement to the previous (negative) results on exact controllability. It was suggested to us by a similar result, stated just for systems with one scalar controller, defined on a Hilbert space and with a different class of admissible controllers (functions of bounded variation) [32]. Therefore the proof is only sketched.

THEOREM 3.3.4. *Let the state space X be infinite-dimensional. Assume that the autonomous system \mathcal{L}_m satisfy a property stronger than approximate controllability, namely, let the sequence $A^n b_i: b_1, \dots, b_m, Ab_1, \dots, Ab_m, A^2 b_1, \dots, A^2 b_m, \dots$ be a Schauder basis for X . Write the coordinates of a point x_0 in X as $x_0^{n,i}: x_0^{0,1}, \dots, x_0^{0,m}, x_0^{1,1}, \dots, x_0^{1,m}, x_0^{2,1}, \dots, x_0^{2,m}, \dots$. If $x_0^{n,i} \geq 0$ (or $x_0^{n,i} \leq 0$) for some fixed i and all n greater than some N , then the point x_0 cannot be steered to the origin in any finite time by bounded measurable controls.*

Proof. (Sketch; for details see [35, § 2.6]). The proof makes use of the Hausdorff theorem on the solvability of the moment problem on a finite interval, within the class of bounded measurable functions [27, p. 134]. Consider first the case $m = 1$, and write b instead of b_1 . Let x_0 be a point in X with coordinates x_0^n satisfying $x_0^n \geq 0$ (or $x_0^n \leq 0$) for all $n > N > 0$, $x_0^N \neq 0$, as in the assumption.

We need to show that the response $x(T, x_0, u)$, T arbitrary but fixed, to any bounded measurable control u , cannot be the origin, that is the equality

$$\int_0^T e^{-At} bu(t) dt = \sum_{n=0}^{\infty} \left[\frac{(-1)^n}{n!} \int_0^T t^n u(t) dt \right] A^n b = -x_0 = -\sum_{n=0}^{\infty} x_0^n A^n b$$

cannot hold. This is equivalent (since $A^n b$ is a basis and setting $\tau = t/T$) to showing that the moment problem

$$\int_0^1 \tau^n u(\tau) d\tau = (-1)^{n+1} x_0^n \frac{n!}{T^{n+1}} = \mu_n, \quad n = 0, 1, \dots,$$

is not solvable for $u(\tau)$ bounded measurable; that is, by the theorem in [27, p. 134] the inequality

$$(n+1)C_n^k |\Delta^{n-k}\mu_k| \leq M, \quad k = 0, 1, \dots, n,$$

with M a constant independent of n , cannot hold. And in fact, from [27, p. 131]

$$\Delta^{n-N}\mu_N = \mu_N - C_n^1\mu_{N+1} + C_n^2\mu_{N+2} + \dots + (-1)^n\mu_n, \quad n > N,$$

we deduce that $|\Delta^{n-N}\mu_N| \geq |\mu_N| > 0$, $n > N > 0$, since the μ_n 's corresponding to the point x_0 are alternate in sign (or possibly zero) for all $n > N$. From here one checks that, when $k = N < n$, the left-hand side of the Hausdorff inequality goes to infinity as $n \rightarrow \infty$ and therefore the moment problem is not solvable.

The conclusion for the case $m > 1$ follows from the case $m = 1$, applied to the sequence $A^n b_i$ in question, using

$$\int_0^T e^{-AT} \sum_{i=1}^m b_i u_i(t) dt = \sum_{n=0}^{\infty} \left[\sum_{i=1}^m \frac{(-1)^n}{n!} \int_0^T t^n u_i(t) dt \right] A^n b_i. \quad \text{Q.E.D.}$$

Remark 3.3.4. We argue only for $m = 1$. The modifications needed for $m > 1$ are obvious. Theorem 3.3.4 applies in particular to the points b, Ab, A^2b, \dots (or a finite linear combination of them) that therefore are not steerable to the origin in finite time by bounded measurable controllers. However, for this special class of initial points, the theorem holds under somewhat relaxed assumptions, namely that the sequence $A^n b$ need not be a Schauder basis for X , but it is enough that it satisfies the weaker property of unique representation of the origin, i.e.,

$$\sum c_n A^n b = 0 \rightarrow c_n = 0, \quad n = 0, 1, \dots$$

In fact, when x_0 is one of the vectors of the sequence $A^n b$, this is just what is needed in the proof of Theorem 3.3.4.

An important example where this uniqueness property of the origin in terms of the $A^n b$'s holds, while the sequence $A^n b$ is not a Schauder basis, is given by the case when A is the Volterra integral operator $(Af)(\xi) = \int_0^\xi f(s) ds$ and b is the unit function in $L_p[0, 1]$. In this case, in fact [34, p. 291], $(A^n b)(\xi) = \xi^n/n!$.

Also Theorem 3.3.4 admits a counterpart concerning points reachable from the origin in finite time. For instance, for $m = 1$, one verifies, by paralleling the argument of the previous Theorem 3.3.4, that nonzero points, whose coordinates in the basis b, Ab, A^2b, \dots are alternate in sign (or possibly zero) from a certain place on, are not reachable from the origin in finite time by bounded measurable controls.

So far we have provided neither condition(s) for a system \mathcal{L} (defined on an infinite-dimensional state space X) to be exactly controllable nor any example of such sort and one may very well wonder whether exactly controllable systems exist at all. The answer is of course in the affirmative sense, the simplest example being given by $A = 0$, $B = I$ on $X = U$. Then, the U -valued (admissible) controller $u(t) \equiv x_1/T$ reaches from the origin the arbitrary point x_1 in $[0, T]$.

More generally, if the operator B is onto (and so B is not compact with X infinite-dimensional !) then the autonomous system \mathcal{L} is exactly controllable on any finite interval $[0, T]$. In fact, in this case, for any x_1 in X , we can define (perhaps

nonuniquely) a U -valued function $u(t)$ by $Bu(t) = e^{-A(T-t)}x_1/T$, $0 \leq t \leq T$, whose response at T from the origin is precisely x_1 . Also, $u(t)$ is indeed an admissible controller, since, by the open mapping theorem, it entirely lies in some finite sphere of U .

If in addition, B is either one-to-one (and so B^{-1} is bounded) or else $X = U$ and A and B commute, then an admissible controller steering the origin to the arbitrary point x_1 in $[0, T]$ is given explicitly by $u(t) = B^{-1} e^{-A(T-t)}x_1/T$ and $u(t) = e^{-A(T-t)}u_1/T$, with $Bu_1 = x_1$, respectively.

An obvious necessary condition for exact controllability of \mathcal{L} is (using (2.1.1)) that, for each x in X , there is a sequence $u_n(x)$ in U (depending on x) such that x can be written as $x = \sum_{n=0}^{\infty} A^n B u_n(x)$. However if, in addition, $u_n(x)$ is unique, given x , then the above necessary condition is not sufficient. In fact, take any sequence u_1, u_2, \dots in U and add a point u_0 not contained in their span. Then by the Hahn-Banach theorem, there is $u^* \in U^*$ with $u^*(u_0) = 1$ and $u^*(u_n) = 0$, $n = 1, 2, \dots$. Consider then the point $x_1 = \sum_{n=0}^{\infty} A^n B u_n$. If there were an admissible control $u(t)$ steering the origin to x_1 in finite time T along a solution curve of \mathcal{L} , then, using (2.1.1) and the uniqueness of the representation, the following classical Hausdorff moment problem,

$$\int_0^T (T-t)^n u^*(u(t)) dt = u^*(u_n) \cdot n!, \quad n = 0, 1, \dots,$$

would be solvable for a bounded measurable function $u^*(u(t))$. But this is impossible in our case (see Remark 3.3.4).

4. State controllability. Nonautonomous case. In this section we intend to generalize the theory of approximate controllability to the nonautonomous system \mathcal{L} under at least the standard assumptions as in § 2. The aim is the same as before, that is to characterize the above concept solely in terms of the coefficients appearing in the equation of the process. However, we recall that even with a finite-dimensional state space, this goal is achieved for nonautonomous systems, provided appropriate assumptions of smoothness are placed on the coefficients [31]. We shall see that, with an infinite-dimensional state space, the situation is substantially not any worse.

The proofs will only be sketched, since the underlying idea is the same as in the autonomous case, although new technicalities now appear.

Let the coefficients of the nonautonomous system \mathcal{L} , i.e., the operators $A(t)$ and $B(t)$ be infinitely many times differentiable with respect to t (the limits are computed in the norm of $\mathcal{B}(X)$ and $\mathcal{B}(U, X)$, respectively). Then state approximate controllability will be described by the following sequence of operators $\Gamma_n(t)$ in $\mathcal{B}(U, X)$, obtained from $A(t)$ and $B(t)$ by successive differentiations:

$$\Gamma_0(t) = B(t), \quad \Gamma_n(t) = -A(t)\Gamma_{n-1}(t) + \dot{\Gamma}_{n-1}(t), \quad n = 1, 2, \dots$$

Notice that, since $d\Phi(T, t)/dt = -\Phi(T, t)A(t)$, one has

$$(4.1) \quad \Phi(T, t)\Gamma_n(t)u = \frac{d^n}{dt^n} \Phi(T, t)\Gamma_0(t)u,$$

and hence for any $x^* \in X^*$,

$$(4.2) \quad x^*(\Phi(T, t)\Gamma_n(t)u) = \frac{d^n}{dt^n} x^*(\Phi(T, t)\Gamma_0(t)u).$$

(Recall the definition of an analytic vector-valued function given in § 1.) It follows that, if $\Phi(t_1, t)\Gamma_0(t)u$ is analytic (with convergence in X):

$$\Phi(T, t)\Gamma_0(t)u = \Phi(T, \bar{t})\Gamma_0(\bar{t})u + \sum_{n=1}^{\infty} \Phi(T, \bar{t})\Gamma_n(\bar{t})u(t - \bar{t})^n/n!,$$

t in a neighborhood of \bar{t} , so is $x^*(\Phi(T, t)\Gamma_0(t)u)$ and we have

$$(4.3) \quad x^*(\Phi(T, t)\Gamma_0(t)u) = x^*(\Phi(T, \bar{t})\Gamma_0(\bar{t})u) + \sum_{n=1}^{\infty} x^*(\Phi(T, \bar{t})\Gamma_n(\bar{t})u)(t - \bar{t})^n/n!.$$

THEOREM 4.1. *The nonautonomous system \mathcal{L} is approximately controllable on $[t_0, T]$ if*

$$\text{Cl sp } \{\Gamma_n(t)U, n \geq 0\} = X$$

or, equivalently, if $\bigcap_{n=0}^{\infty} \mathcal{N}\{\Gamma_n^*(t)\} = \{O_{X^*}\}$ for some t in $[t_0, T]$.

Conversely, if \mathcal{L} is approximately controllable and $A(t)$ and $B(t)$ are analytic on $[t_0, T]$, then the above condition holds for (some and hence) all t in $[t_0, T]$.

Remark 4.1. In the autonomous case $A(t) \equiv A$ and $B(t) \equiv B$, we have $\Gamma_n(t) = (-1)^n A^n B^n$ and Theorem 4.1 reduces to Theorem 3.1.1. Also Theorem 4.1 extends results in [31] to the case when both X and U are infinite-dimensional.

In particular, for the nonautonomous system \mathcal{L}_m , with $A(t)$ and $b_i(t)$ infinitely many times differentiable, $\Gamma_n(t)$ is replaced by the following sequence $\gamma_n(t)$ of m -vectors in X :

$$\gamma_0(t) = [b_1(t), \dots, b_m(t)], \quad \gamma_n(t) = -A(t)\gamma_{n-1}(t) + \dot{\gamma}_{n-1}(t), \quad n = 1, 2, \dots$$

COROLLARY 4.2. *The nonautonomous system \mathcal{L}_m is approximately controllable on $[t_0, T]$ if, for some t in $[t_0, T]$,*

$$\text{Cl sp } \{\gamma_n(t), n \geq 0\} = X.$$

Conversely, if \mathcal{L}_m is approximately controllable and $A(t)$ and $b_i(t)$ are analytic on $[t_0, T]$, then the above condition holds for (some and hence) all t in $[t_0, T]$.

Proof of Theorem 4.1. (Sketch; see [35] for details). (a) \mathcal{L} is approximately controllable on $[t_0, T]$ if and only if $x^*(\Phi(T, t)B(t)u) = 0$ on $[t_0, T] \times U$ implies $x^* = 0$ (equivalently, if and only if $B^*(t)\Phi^*(T, t)x^* = 0$ on $[t_0, T]$ implies $x^* = 0$). The proof is as in Theorem 3.1.1.

(b) A sufficient condition for (a) to hold is that: $x^*(\Gamma_n(t)U) = 0$, for some t in $[t_0, T]$, $n = 0, 1, \dots$, implies $x^* = 0$ (or that $\Gamma_n^*(t)x^* = 0$ for some t , $n = 0, 1, \dots$, implies $x^* = 0$).

Otherwise, from $\bar{x}^*(\Phi(T, t)B(t)u) = 0$ on $[t_0, T] \times U$ for some nonzero \bar{x}^* , it follows, via (4.2), that

$$\bar{x}^*(\Phi(T, t)\Gamma_n(t)u) = 0 \quad \text{on } [t_0, T] \times U, \quad n = 0, 1, \dots$$

By the fundamental proposition, the sequence of subspaces $\Phi(T, t)\Gamma_n(t)U$ does not span X at any t in $[t_0, T]$. Since $\Phi(T, t)$ is one-to-one and onto, the same conclusion holds for $\Gamma_n(t)U$, and this, by the fundamental proposition, is a contradiction and proves the sufficiency part of the theorem.

(c) *Necessity.* From the analyticity of $A(t)$ (that implies the analyticity of $\Phi(T, t)$ on $\mathcal{B}(X)^5$) and the analyticity of $B(t)$, it follows that $\Phi(T, t)\Gamma_0(t)u$ is analytic in $[t_0, T]$ and so the expansion (4.3) applies.

Now, if the sequence of subspaces $\Gamma_n(\bar{t})U$ does not span X , \bar{t} in $[t_0, T]$, neither does $\Phi(T, \bar{t})\Gamma_n(\bar{t})U$. Hence, by the fundamental proposition and (4.3), we get that $\bar{x}^*(\Phi(T, \bar{t})B(\bar{t})U) = 0$ (in a neighborhood of \bar{t} and hence) on $[t_0, T]$ for a non-zero \bar{x}^* . So (a) does not hold. Also, differentiating yields $\bar{x}^*(\Phi(T, t)\Gamma_n(t)U) = 0$ on $[t_0, T]$ and hence $\Gamma_n(t)U$ does not span X at any t . Q.E.D.

We next study the observability problem. For the sake of clarity the autonomous and nonautonomous cases will again be handled separately.

5. Observability.

5.1. Autonomous case. For autonomous observed systems, the characterization of observability is expressed solely in terms of the coefficients A and H of the free operating condition and is independent on the particular time interval length.

THEOREM 5.1.1. *The autonomous observed system \mathcal{L} is observable on $[0, T]$ if and only if*

$$(5.1.1) \quad \bigcap_{n=0}^{\infty} \mathcal{N}\{HA^n\} = \{O_X\}.$$

If the (Banach) state space X is reflexive, this characterization is equivalent to

$$(5.1.2) \quad \text{Cl sp } \{(A^*)^n H^* Y^*, n \geq 0\} = X^*.$$

Under the additional assumption that the output space is finite-dimensional, $Y = R^r$, and so $H = [h_1, \dots, h_r]$, $h_j \in X^$, the above condition (5.1.2) becomes*

$$\begin{aligned} \text{Cl sp } \{(A^*)^n h_j\} &= \text{Cl sp } \{h_1, \dots, h_r, A^* h_1, \dots, A^* h_r, (A^*)^2 h_1, \dots, (A^*)^2 h_r, \dots\} \\ &= X^*. \end{aligned}$$

Proof. As in Theorem 3.1.1 (b): $He^{At}x_0 = 0$, $0 \leq t \leq T$, implies $x_0 = 0$ if and only if: $HA^n x_0 = 0$, $n = 0, 1, \dots$, implies $x_0 = 0$, and (5.1.1) is proved. Next, observe that (5.1.2), by the fundamental proposition, is equivalent to: $x^{**}((A^*)^n H^* Y^*) = 0$, $n = 0, 1, \dots$, implies $x^{**} = 0$ which can also be written as $\bigcap_{n=0}^{\infty} \mathcal{N}\{(H^*)^*((A^*)^n)^*\} = \{O_{X^{**}}\}$. From here, with X reflexive, and hence isometrically isomorphic to X^{**} and $(H^*)^* = H$ and $(A^*)^* = A$, we get

$$\bigcap_{n=0}^{\infty} \mathcal{N}\{HA^n\} = \{O_X\}.$$

And this, in view of the first part, proves the second. Q.E.D.

Remark 5.1.1. In view of Theorems 3.1.1 and 5.1.1, it follows that, when the state space X is reflexive, the free observed system

$$\dot{x} = Ax, \quad y = Hx \quad (\dot{x} = Ax, \quad y = [h_1, \dots, h_r]x)$$

⁵ A standard proof of this property for a finite-dimensional X as in [5, p. 90] carries over to an infinite-dimensional Banach space X , without modifications, when $A(t) \in \mathcal{B}(X)$.

is observable if and only if the “dual system” defined on X^* ,

$$\dot{x}^* = A^*x^* + H^*u, \quad U = Y^* \quad (\dot{x}^* = A^*x^* + \sum h_j u_j, 1 \leq j \leq r),$$

is approximately controllable.

If, moreover, the state space X is Hilbert, we can identify X with X^* , via the Riesz representation theorem, and the dual system is also defined on X .

Notice also that, when X is not reflexive, (5.1.1) is equivalent to

$$\text{Cl sp } \{(A^*)^n H^* Y^*, n \geq 0\} = X_1^*,$$

where X_1^* is the largest (closed) subspace of X^* , such that $x^{**} = 0$ is the only vector in the range $\phi[X]$ of the natural isomorphism ϕ of X into X^{**} that vanishes on X_1^* .

5.2. Examples of observable systems. Examples of observable systems can be easily constructed from the examples of approximate controllable systems given in § 3.2, via the duality property, pointed out in the above remark.

Hence, we will here confine ourselves only to two examples.

Example 5.2.1. Let X be the reflexive $L_p[0, 1]$, $1 < p < \infty$, so that the dual X^* is (isometrically isomorphic to) $L_q[0, 1]$, $1 < q < \infty$. By the Riesz theorem, a bounded linear functional h on $X = L_p[0, 1]$ is given by

$$\int_0^1 h(\xi) f(\xi) d\xi, \quad h(\xi) \in L_q[0, 1], \quad f(\xi) \in L_p[0, 1].$$

Let A be the Volterra integral operator on X defined by

$$(Af)(\xi) = \int_\xi^1 f(s) ds, \quad f(\cdot) \in L_p[0, 1],$$

so its dual A^* on X^* is (integrating by parts from the definition $(A^*x^*)x = x^*(Ax)$)

$$(A^*g)(\xi) = \int_0^\xi g(s) ds, \quad g(\cdot) \in L_q[0, 1].$$

Then consider the free observed system

$$\frac{\partial w(t, \xi)}{\partial t} = \int_\xi^1 w(t, s) ds, \quad y(t) = \int_0^1 h(\xi) w(t, \xi) d\xi,$$

whose abstract version is given by

$$\dot{x} = Ax, \quad y = hx \quad (Y = R^1)$$

with $x(t) = w(t, \cdot) \in L_p[0, 1]$. By Theorem 5.1.1, the above system is observable if and only if the pair $\langle A^*, h \rangle$ is cyclic on $L_q[0, 1]$; that is, recalling Example 3.3.1, if and only if, for any $\delta > 0$, $h(\xi) \neq 0$ on a subset of positive measure of $[0, \delta]$ (equivalently, zero belongs to the support of the function $h(\cdot)$).

Example 5.2.2. Let $X = Y$ be a Hilbert space, let A , and hence A^* , be a simple operator and let $H = A - A^*$. Then $H^* = A^* - A$ and $\text{Cl sp } \{(A^*)^n H^* Y, n \geq 0\} = X$ [16, p. 30] and so the pair $\langle A, H \rangle$ is observable.

In particular, recalling § 3.2, the free process

$$\frac{\partial w(t, \xi)}{\partial t} = \int_{\xi}^1 w(t, s) ds, \quad y(t, \xi) = \int_0^1 w(t, s) - 2 \int_0^{\xi} w(t, s) ds$$

is observable.

At this point, it is not difficult to construct examples of observed systems that are both approximately controllable and observable. The following are general results in this direction, valid on a Hilbert space:

(i) if A is unicellular, there are vectors cyclic for both A and A^* [33, p. 140]; if b is one such vector, the observed system $\mathcal{L}_1: x = Ax + bu, y = (b, x)$ is both approximately controllable and observable.

(ii) if A decomposes into a quasi-direct sum, then A^* also does [4, p. 211] and hence, if moreover $(A - A^*)$ is of p -dimensional range, it is possible to construct observed systems $\mathcal{L}_m: \langle A; (b_1, \dots, b_m); (h_1 \dots h_r) \rangle, m \leq p, r \leq p$, that are both approximately controllable and observable.

5.3. Nonautonomous case. When the coefficients of the nonautonomous observed free system \mathcal{L} , i.e., the operators $A(t)$ and $H(t)$, are infinitely many times differentiable, observability will be described by the following sequence $\mathcal{H}_n(t)$ of operators in $\mathcal{B}(X, Y)$ obtained from $A(t)$ and $H(t)$ by successive differentiations:

$$\mathcal{H}_0(t) = H(t), \quad \mathcal{H}_n(t) = \mathcal{H}_{n-1}A(t) + \dot{\mathcal{H}}_{n-1}(t), \quad n = 1, 2, \dots,$$

(or, equivalently, by their adjoint

$$\mathcal{H}^*(t) = A^*(t)\mathcal{H}_{n-1}^* + \dot{\mathcal{H}}_{n-1}^*).$$

(Here the completeness of Y is needed.) Notice that

$$\mathcal{H}_n(t)\Phi(t, t_0) = \frac{d^n}{dt^n} \mathcal{H}_0(t)\Phi(t, t_0),$$

so that, if $\mathcal{H}_0(t)\Phi(t, t_0)$ is analytic at \bar{t} , we can write

$$(5.3.1) \quad \mathcal{H}_0(t)\Phi(t, t_0) = \mathcal{H}_0(\bar{t})\Phi(\bar{t}, t_0) + \sum_{n=1}^{\infty} \mathcal{H}_n(\bar{t})\Phi(\bar{t}, t_0)(t - \bar{t})^n/n!$$

in a neighborhood of \bar{t} .

When the output space is finite-dimensional, say $Y = R^r$, and so

$$H(t) = [h_1(t), \dots, h_r(t)], \quad h_j(t) \in X^*,$$

$H_n^*(t)$ is replaced by the following sequence $\ell_n(t)$ of r -vectors in X^* :

$$\ell_0(t) = [h_1(t), \dots, h_r(t)], \quad \ell_n(t) = A^*(t)\ell_{n-1}(t) + \dot{\ell}_{n-1}(t), \quad n = 1, 2, \dots.$$

We also remark that the sequence $\mathcal{H}_n(\ell_n)$ reduces to $\mathcal{H}_n = HA^n$ ($\ell_n = (A^*)^n h_n$) in the autonomous case: $H(t) \equiv H, A(t) \equiv A$.

Therefore the following theorem generalizes Theorem 5.1.1 to the non-autonomous case.

THEOREM 5.3.1. *The nonautonomous observed system \mathcal{L} is observable on $[t_0, T]$ if, for some t in $[t_0, T]$,*

$$\bigcap_{n=0}^{\infty} \mathcal{N}\{\mathcal{H}_n(t)\} = \{O_X\}.$$

If the (Banach) state space X is reflexive, the above condition is equivalent, as in Theorem 5.1.1, to

$$\text{Cl sp } \{\mathcal{H}_n^*(t)Y^*, n \geq 0\} = X^*$$

for some t in $[t_0, T]$.

Under the additional assumption that the output space is finite-dimensional, $Y = R^r$, and so $H(t) = [h_1(t), \dots, h_r(t)]$, $h_j(t) \in X^$, then the above sufficient condition becomes*

$$\text{Cl sp } \{h_n(t), n \geq 0\} = X^*$$

for some t in $[t_0, T]$.

Conversely, if \mathcal{L} is observable on $[t_0, T]$ and $A(t)$ and $H(t)$ are analytic on $[t_0, T]$, then the above conditions hold for (some and hence) all t in $[t_0, T]$.

Proof. Sufficiency. If \mathcal{L} is not observable on $[t_0, T]$, there is a nonzero vector x_0 in X such that $\mathcal{H}_0(t)\Phi(t, t_0)x_0 = 0$ on $[t_0, T]$. Hence, by successive differentiations, one gets $\mathcal{H}_n(t)\Phi(t, t_0)x_0 = 0$ on $[t_0, T]$, $n = 0, 1, \dots$, which is a contradiction since $\Phi(t, t_0)x_0$ is nonzero.

Necessity. Suppose there is a nonzero vector x_0 in X such that $\mathcal{H}_n(\bar{t})x_0 = 0$ and hence $(\Phi(\bar{t}, t_0)$ being one-to-one and onto) $\mathcal{H}_n(\bar{t})\Phi(\bar{t}, t_0)\bar{x}_0 = 0$, $n = 0, 1, \dots$, for some \bar{t} in $[t_0, T]$. The analyticity of $A(t)$ (hence of $\Phi(t, t_0)$) and of $H(t)$ implies that $H(t)\Phi(t, t_0)$ is analytic. It follows, by (5.3.1), that $H(t)\Phi(t, t_0)\bar{x}_0 = 0$ in a neighborhood of \bar{t} , hence in $[t_0, T]$ and \mathcal{L} is not observable.

For the equivalence when X is reflexive, see the proof of Theorem 5.1.1.

Q.E.D.

Remark 5.3.1. In view of Theorems 4.1 (Corollary 4.2) and 5.3.1, when the state space X is reflexive, the free observed system

$$\dot{x} = A(t)x, \quad y = H(t)x \quad (x = A(t)x, \quad y = [h_1(t), \dots, h_r(t)]x)$$

with $A(t)$ and $H(t)$ ($h_1(t), \dots, h_r(t)$) analytic in $[t_0, T]$ is observable in $[t_0, T]$ if and only if the "dual system" defined on X^* ,

$$\dot{x}^* = -A^*(t)x^* + H^*(t)u, \quad U = Y^* \quad (\dot{x}^* = -A^*(t)x^* + \sum h_j(t)u_j, 1 \leq j \leq r)$$

is approximately controllable on $[t_0, T]$.

6. Relationship between state and output controllability. There are two basic problems regarding output controllability that we shall deal with. The first is to investigate the relationship between state and output controllability; the second is to characterize output controllability solely in terms of the coefficients appearing in the equations of the observed process. The first problem is solved by the following theorem, that reduces to known results [23], when X , U and Y are finite-dimensional; while the second problem is treated in the next section.

THEOREM 6.1. *If the nonautonomous observed process \mathcal{L} is output approximately controllable on $[t_0, T]$, then $\text{Cl } \mathcal{R}(H(T)) = Y$.*

Conversely, if the state equation of \mathcal{L} is approximately controllable on $[t_0, T]$ and $\text{Cl } \mathcal{R}(H(T)) = Y$, then \mathcal{L} is output approximately controllable on $[t_0, T]$.

Proof. The first part is plain and so we prove only the second. Let x_0 and y_1 be arbitrary points in X and Y , respectively, and $\varepsilon > 0$ be given. Then there is x'_1 in X such that $y'_1 = H(T)x'_1$ is within $\varepsilon/2$ from y_1 . But then there is u admissible in $[t_0, T]$, such that its response $x(T, t_0, x_0, u)$, initiating at x_0 at the time t_0 is within $\varepsilon/2\|H(T)\|$ from x'_1 and so $H(T)x(T, t_0, x_0, u)$ is within $(\varepsilon/2)$ from y'_1 and hence within ε from y_1 . Q.E.D.

COROLLARY 6.2. *In the special case when the output space Y is finite-dimensional, say $Y = R^r$, and so $H(t) = [h_1(t), \dots, h_r(t)]$, $h_j(t) \in X^*$, the output exact controllability on $[t_0, T]$ of the nonautonomous system \mathcal{L} implies that $h_1(T), \dots, h_r(T)$ are linearly independent (vectors in X^*).*

Conversely, if the state equation of \mathcal{L} is approximately controllable on $[t_0, T]$ and $h_1(T), \dots, h_r(T)$ are linearly independent, then \mathcal{L} is also output exactly controllable on $[t_0, T]$.

7. Output controllability. In this section we derive characterizations for output controllability expressed solely in terms of the coefficients of the observed system (under suitable conditions of smoothness in the nonautonomous case).

7.1. Autonomous case. For the autonomous observed system \mathcal{L} , output approximate controllability is independent on the particular time interval length and completely characterized by the coefficients A, B, H .

THEOREM 7.1.1. *The autonomous observed system \mathcal{L} is output approximately controllable on $[0, T]$ if and only if*

$$\text{Cl sp } \{HA^nBU, n \geq 0\} = Y,$$

or, equivalently, if and only if

$$\bigcap_{n=0}^{\infty} \mathcal{N}\{B^*(A^*)^nH^*\} = \{O_{X^*}\}.$$

In particular, for the autonomous observed system \mathcal{L}_m , the above characterization becomes

$$\text{Cl sp } \{HA^n b_i\} = \text{Cl sp } \{Hb_1, \dots, Hb_m, HAb_1, \dots, HAb_m, HA^2b_1, \dots\} = Y.$$

Proof. The proof closely parallels that of Theorem 3.1.1 and is therefore omitted. Q.E.D.

Remark 7.1.1. Since H is in $\mathcal{B}(X, Y)$, we have

$$\text{Cl sp } \{HA^nBU, n \geq 0\} = \text{Cl } H\{\text{Cl sp } \{A^nBU, n \geq 0\}\}$$

and this offers an alternative proof to Theorem 6.1, at least in the autonomous case: In particular, by Theorems 3.1.1 and 7.1.1, it follows that, if $\text{Cl } \mathcal{R}(H) = X$ and the state equation is approximately controllable, then output approximate controllability is automatically guaranteed. Notice also that Theorem 3.3.3 can be extended to the output space Y , when H is bounded and so HQ is a compact operator from $L_1[[0, T], U]$ into Y .

is then given explicitly by

$$\frac{\partial w(t, \xi)}{\partial t} = \int_0^\xi w(t, s) ds + v(\xi)u(t),$$

$$y(t) = \left[\int_0^1 h_1(s)w(t, s) ds, \dots, \int_0^1 h_r(s)w(t, s) ds \right].$$

By Theorem 6.1, the above process is output exactly controllable if the state equation is approximately controllable (that is (Example 3.2.1) in case for any $\delta > 0$, $v(\xi) \neq 0$ on a subset of positive measure of $[0, \delta]$) and the functions $h_1(\cdot), \dots, h_r(\cdot)$ are linearly independent a.e. in $[0, 1]$.

Alternatively, we now use Corollary 7.1.2 to test that the above process is output exactly controllable, when, say, $v(\xi)$ is the unit function on $[0, 1]$ and $h_1(\cdot), \dots, h_r(\cdot)$ are linearly independent a.e. in $[0, 1]$. In fact, in this case, one has [34, p. 291]

$$(A^n b)(\xi) = \frac{\xi^n}{n!}, \quad h_j A^n b = \int_0^1 h_j(s) \frac{s^n}{n!} ds, \quad n = 0, 1, \dots.$$

Next observe that [10, p. 627]

$$\int_0^1 [\alpha_1 h_1(s) + \dots + \alpha_r h_r(s)] \frac{s^n}{n!} ds = 0, \quad n = 0, 1, \dots,$$

implies $\alpha_1 h_1(s) + \dots + \alpha_r h_r(s) = 0$ a.e., on $[0, 1]$ and hence, by assumption, $\alpha_1 = \dots = \alpha_r = 0$. This precisely says that the sequences

$$h_1 A^n b, \dots, h_r A^n b, \quad n = 0, 1, \dots,$$

are linearly independent and the result is proved.

7.3. Nonautonomous case. The results of § 7.1 generalize as follows for nonautonomous, smooth systems, via the sequence $\Gamma_n(t)$ of operators in $\mathcal{B}(U, X)$ (or the sequence of r -vectors $\gamma_n(t)$) defined in § 4.

THEOREM 7.3.1. *The nonautonomous observed system \mathcal{L} is output approximately controllable on $[t_0, T]$ if for some t in $[t_0, T]$,*

$$\text{Cl sp } \{H(T)\Gamma_n(t)U, n \geq 0\} = Y,$$

or, equivalently, if

$$\bigcap_{n=0}^{\infty} \mathcal{N}\{\Gamma_n^*(t)H^*(T)\} = \{O_{X^*}\}.$$

In particular, \mathcal{L}_m is output approximately controllable on $[t_0, T]$ if, for some t in $[t_0, T]$,

$$\text{Cl sp } \{H(T)\gamma_n(t), n \geq 0\} = Y.$$

Conversely, if \mathcal{L} or \mathcal{L}_m is output approximately controllable on $[t_0, T]$ and $A(t)$ and $B(t)$, or $b_i(t)$, are analytic in $[t_0, T]$, then the above conditions hold for (some and hence) all t in $[t_0, T]$.

Proof. The proof follows that given for Theorem 4.1. Q.E.D.

- [13] ———, *On complete controllability of linear systems*, J. Differential Equations, 3 (1967), pp. 391–402.
- [14] L. FREEMAN, *Volterra operators similar to $J: f \rightarrow \int_0^x f(t) dt$* , Trans. Amer. Math. Soc., 116 (1965), pp. 181–192.
- [15] P. A. FUHRMANN, *On weak and strong reachability and controllability of infinite dimensional linear systems*, J. Optimization Theory Appl., 9 (1972), pp. 77–89.
- [16] I. C. GOHBERG AND M. G. KREIN, *Theory and applications of Volterra operators in Hilbert space*, Translations Math. Monographs, vol. 24, American Mathematical Society, Providence, R.I., 1970.
- [17] P. R. HALMOS, *A Hilbert Space Problem Book*, Van Nostrand, Princeton, N.J., 1968.
- [18] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semigroups*, Amer. Math. Soc., Providence, R.I., 1957.
- [19] G. K. KALISCH, *On similarity, reducing manifolds, and unitary equivalence of certain Volterra operators*, Ann. of Math., 66 (1967), pp. 481–494.
- [20] ———, *A functional analysis proof of Titchmarsh's theory on convolution*, J. Math. Anal. Appl., 5 (1962), pp. 176–183.
- [21] R. E. KALMAN, Y. C. HO AND K. S. NARENDRA, *Controllability of linear dynamical systems*, Contributions to Differential Equations, 2 (1963), pp. 189–213.
- [22] L. U. KANTOROVICH AND G. P. AKILOV, *Functional Analysis in Normed Spaces*, Pergamon Press, New York, 1964.
- [23] E. KREINDLER AND P. E. SARACHIK, *On the concepts of controllability and observability of linear systems*, IEEE Trans. Automatic Control, AC-9 (1964), pp. 129–136.
- [24] L. M. KUPERMAN AND JU. M. REPIN, *On controllability in infinite-dimensional spaces*, Soviet Math. Dokl., 12 (1971), pp. 1469–1472.
- [25] A. B. KURJANSKII, *On controllability in Banach spaces*, Differencial'nye Uravnenija, 5 (1969), pp. 1269–1271.
- [26] E. B. LEE AND L. MARKUS, *Foundation of Optimal Control Theory*, John Wiley, New York, 1967.
- [27] I. P. NATANSON, *Constructive Function Theory*, vol. II, Frederick Ungar, New York, 1965.
- [28] N. K. NIKOL'SKII, *Nonstandard ideals, unicellularity and algebras associated with a shift operator*, Investigations in Linear Operators and Function Theory, Part I, Consultants Bureau, New York, 1972.
- [29] S. J. OSHER, *Two papers on similarity of certain Volterra integral operators*, Mem. Ann. Math. Soc., 73 (1967).
- [30] H. L. ROYDEN, *Real Analysis*, 2nd ed., Macmillan, New York, 1968.
- [31] L. M. SILVERMAN AND H. E. MEADOWS, *Controllability and observability in time-variable linear systems*, this Journal, 5 (1967), pp. 64–73.
- [32] F. A. SHOLOKHOVICH, *On controllability in Hilbert space*, Differencial'nye Uravnenija, 3 (1967), pp. 479–484.
- [33] SZ-NAGY AND C. FOIAS, *Harmonic analysis of operators in Hilbert Space*, North-Holland, Amsterdam, 1970.
- [34] A. E. TAYLOR, *Introduction to Functional Analysis*, John Wiley, New York, 1958.
- [35] R. TRIGGIANI, *Controllability, observability and stabilizability of dynamical systems in Banach space with bounded operators*, Ph.D. thesis, University of Minnesota, Minneapolis, 1972.
- [36] M. VIDYASAGAR, *On the controllability of infinite-dimensional linear systems*, J. Optimization Theory Appl., 6 (1970), pp. 171–173.
- [37] W. M. WONHAM, *On pole assignment in multi-input controllable linear systems*, IEEE, Trans. Automatic Control, AC-12 (1967), pp. 660–665.
- [38] G. E. KISILEVSKII, *Invariant subspaces of dissipative operators with nuclear imaginary components*, Math. USSR-Izv., 2 (1968), pp. 1–20.
- [39] L. V. AHLFORS, *Complex Analysis*, 2nd ed., McGraw-Hill, New York, 1966.
- [40] T. KATO, *Perturbation Theory of Linear Operators*, Springer-Verlag, Berlin, 1966.
- [41] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, English transl., Springer-Verlag, Berlin, 1971.

MINIMAL BASES OF RATIONAL VECTOR SPACES, WITH APPLICATIONS TO MULTIVARIABLE LINEAR SYSTEMS*

G. DAVID FORNEY, JR.†

Abstract. A minimal basis of a vector space V of n -tuples of rational functions is defined as a polynomial basis such that the sum of the degrees of the basis n -tuples is minimum. Conditions for a matrix G to represent a minimal basis are derived. By imposing additional conditions on G we arrive at a minimal basis for V that is unique. We show how minimal bases can be used to factor a transfer function matrix G in the form $G = ND^{-1}$, where N and D are polynomial matrices that display the controllability indices of G and its controller canonical realization. Transfer function matrices G solving equations of the form $PG = Q$ are also obtained by this method; applications to the problem of finding minimal order inverse systems are given. Previous applications to convolutional coding theory are noted. This range of applications suggests that minimal basis ideas will be useful throughout the theory of multivariable linear systems. A restatement of these ideas in the language of valuation theory is given in an Appendix.

1. Introduction. The major conceptual development in system theory over the past decade or so has been the replacement of the classical transfer function approach by the so-called state space approach. Nowhere have the advantages of the latter method seemed more pronounced than in the analysis of multivariable linear systems.

In the past few years a counterrevolution has begun, with the objective of achieving a synthesis of the two approaches that would combine the advantages of both. Prominent among the counterrevolutionaries has been Rosenbrock [1], who has shown how to derive many state-space results through the analysis of certain polynomial matrices (system matrices). Independently, and about the same time, Popov [2] showed how such seemingly state-space-theoretic problems as realization of systems in controller canonical form and determination of controllability indices could be elegantly solved by polynomial matrix methods, starting from the transfer function matrix. More recently, Wolovich [3], Wang [4] and Wang and Davison [5] have used such methods to solve other problems successfully.

The present paper is offered as a contribution to this counterrevolutionary program. By placing some results derived earlier (1968–70) in studies of convolutional codes [6] in a more general setting, we are able to state some fundamental results concerning polynomial and rational matrices that seem to underlie many of the above results, and through which they can be rederived in a unified way.

Our point of view is rather algebraic. Transfer function matrices are characterized as matrices of rational functions (ratios of polynomials). As the rational functions form a field, it is natural to consider rational matrices as bases for certain rational vector spaces. Among all bases for a given rational vector space, we define as minimal those which are polynomial and of least order in a certain sense. Our main theorem then gives a number of equivalent conditions for a basis to be minimal; we also give an algorithm for reducing a given matrix to a minimal basis. We show that every such vector space has a unique minimal basis satisfying

* Received by the editors June 25, 1973, and in revised form February 11, 1974.

† Codex Corporation, Newton, Massachusetts 02195.

certain additional constraints. Finally, we give a few interesting related results on dual bases and inverses. All of this is pure mathematics. In the Appendix, we recast some of these results in the language of valuation theory, which we believe may be the most natural language for these problems and indeed for algebraic linear system theory generally.

We next apply these ideas to several areas of linear system theory. Following Popov [2], we show how to derive from a given transfer function matrix G a polynomial matrix called a minimal system matrix, which may be viewed as a factorization of G in which the controllability (resp. observability) indices and indeed a controller canonical realization are explicitly exhibited. These results are closely related to certain results of Rosenbrock [1] and largely overlap those of Wolovich [3] and Wang [4]. Then, following Wang and Davison [5], we apply these notions to the class of problems that involve finding a transfer function matrix G that solves the matrix equation $PG = Q$, which includes the problem of finding minimal order inverse systems. Finally, we briefly summarize our own earlier results on convolutional codes [6].

The results we obtain are largely contained in [1]–[6], although most are stated differently. The proofs are all new. Our main objective is to introduce ideas, and we hope that sufficient illustration of their range, power and elegance is given that readers faced with new problems will be inspired to give these tools a try.

2. Definition of a minimal basis. In linear system theory one frequently deals with matrices whose elements are rational functions; i.e., ratios $n(x)/d(x)$ of polynomials $n(x)$ and $d(x) \neq 0$ in some indeterminate x (e.g., $x = s, z, D, \dots$) with coefficients in some field F (e.g., the real or complex numbers, finite fields, etc.). The set of all rational functions in x over F form a field, conventionally denoted $F(x)$.

If G is a $k \times n$ matrix with elements in some field, its row space (the set of all linear combinations of its rows \mathbf{g}_i , $1 \leq i \leq k$) is a vector space V_G over that field, and the dimension of V_G is the rank of G . Conversely, if V is a vector space of n -tuples of field elements of dimension k , it has a basis of k linearly independent n -tuples \mathbf{g}_i , $1 \leq i \leq k$, which may be arranged to form a $k \times n$ matrix G of rank k . We call any $k \times n$ matrix which has rank k a full-rank matrix (implying $k \leq n$), and refer to the matrix as a basis for the corresponding vector space.

From elementary linear algebra we know that two full-rank $k \times n$ matrices G_1 and G_2 are bases for the same vector space if and only if $G_1 = TG_2$ for some full-rank (nonsingular, $\det T \neq 0$) $k \times k$ matrix T . Our objective is to find certain canonical bases for vector spaces of n -tuples over the field of rational functions $F(x)$.

The set of polynomials in x over F is a subset of $F(x)$ and is conventionally denoted as $F[x]$. Certain bases for any vector space over $F(x)$ are entirely polynomial (multiply the elements in any rational basis by their least common denominator).

DEFINITION 1. The *degree* $\deg \mathbf{g}$ of an n -tuple $\mathbf{g} = (g_1, \dots, g_n)$ of polynomials is the greatest degree of its components g_j , $1 \leq j \leq n$.

DEFINITION 2. If G is a $k \times n$ polynomial matrix with rows \mathbf{g}_i , the i th *index* of G is defined as $v_i = \deg \mathbf{g}_i$, $1 \leq i \leq k$, and the *order* of G is defined as $v = \sum_{i=1}^k v_i$.

Now we can define our minimal bases.

DEFINITION 3. If V is a k -dimensional vector space of n -tuples over $F(x)$, a *minimal basis* of V is a $k \times n$ polynomial matrix G such that G is a basis for V and G has least order among all polynomial bases for V .

We now develop a simple lower bound on the order of a minimal basis that will turn out to be tight. The $k \times k$ minors of a $k \times n$ matrix G are the determinants of all $\binom{n}{k}$ square $k \times k$ submatrices obtained by selecting subsets of k columns of G .

The $k \times k$ minors of TG are those of G multiplied by $\det T$. Hence all bases of a given vector space over $F(x)$ have the same set of $k \times k$ minors up to a common constant multiple. Define the *normalized minors* of an $F(x)$ -matrix G as its $k \times k$ minors multiplied by the least common multiple of their denominators and divided by the greatest common divisor of their numerators; then all bases of a given vector space V have the same set of normalized minors. Because of this invariance we may define the *minors of V* to be the normalized minors of any basis of V .

Now if G is polynomial, it has polynomial minors, whose maximum degree cannot exceed the order of G , since a minor is a sum of products of polynomials, one from each row. Further, this maximum degree cannot be less than the maximum degree of the normalized minors of G . Hence the order of a minimal basis for V is lower-bounded by the maximum degree of the minors of V , which can be determined from any basis for V . We see in the next section that this bound is always met with equality.

3. The main theorem. To state our main theorem, we shall need a few more definitions.

If V is a vector space of n -tuples over $F(x)$, we denote by M_V the set of all polynomial n -tuples in V . (M_V is a free $F[x]$ -module.) We denote by V_d the set of all polynomial n -tuples in V with degree (Def. 1) less than d , for all integers $d \geq 0$. V_d is only a vector space over F ; we denote its dimension by $\dim_F V_d$.

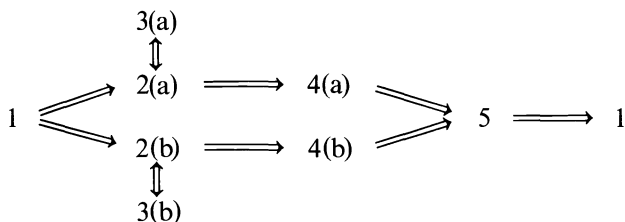
If G is a $k \times n$ polynomial matrix with indices v_i , $1 \leq i \leq k$, its high order coefficient matrix $[G]_h$ is the $k \times n$ F -matrix whose i th row consists of the coefficients of x^{v_i} in the i th row \mathbf{g}_i of G .

MAIN THEOREM. Let V be a k -dimensional vector space of n -tuples over $F(x)$, and let G be a $k \times n$ polynomial basis for V with indices v_i , $1 \leq i \leq k$, and order $v = \sum_i v_i$. Then the following statements are equivalent:

1. G is a minimal basis for V .
2. (a) G is nonsingular modulo $p(x)$ for all irreducible polynomials $p(x) \in F[x]$, and (b) its high order coefficient matrix $[G]_h$ has full rank,
3. (a) The greatest common divisor of the $k \times k$ minors of G is 1, and (b) their greatest degree is v .
4. If $\mathbf{y} = \mathbf{x}G$ is a polynomial n -tuple, then (a) \mathbf{x} must be a polynomial k -tuple and (b) $\deg \mathbf{y} = \max_{1 \leq i \leq k} (\deg x_i + v_i)$.
5. The indices v_i are such that for all $d \geq 0$, $\dim_F V_d = \sum_{i: v_i \leq d} (d - v_i)$.

Proof. A proof of this theorem can be pieced together from the results of [6]. An elementary direct proof that does not rely on the invariant factor theorem follows.

The logical order of the proof is:



(1 \Rightarrow 2) The reduction algorithm of the next section shows that the order of G may be reduced if G does not have full rank mod $p(x)$ or if its high order coefficient matrix does not have full rank.

(2(a) \Leftrightarrow 3(a)) G does not have full rank mod $p(x)$ iff all its $m \times m$ minors are congruent to zero mod $p(x)$, which is to say iff all of them are divisible by $p(x)$, which is to say iff $p(x)$ is a factor of their greatest common divisor.

(2(b) \Leftrightarrow 3(b)). The coefficient of x^v in a given $m \times m$ minor of G is the corresponding $m \times m$ minor of $[G]_h$.

(2(a) \Rightarrow 4(a)) Suppose there exists a nonpolynomial \mathbf{x} such that $\mathbf{y} = \mathbf{x}G$ is polynomial. Let ψ be the monic polynomial of least degree such that $\psi\mathbf{x}$ is polynomial, and let p be any irreducible factor of ψ . Then $\psi\mathbf{x} \not\equiv 0 \pmod{p}$, else $(\psi/p)\mathbf{x}$ would be a polynomial of lower degree than ψ such that $(\psi/p)\mathbf{x}$ was polynomial. But $(\psi\mathbf{x})G = \psi\mathbf{y} \equiv 0 \pmod{p}$, since \mathbf{y} is polynomial. Hence G cannot have full rank mod p , a contradiction.

(2(b) \Rightarrow 4(b)) If $[G]_h$ has full rank, then

$$[\mathbf{y}]_h = \sum [x_i]_h [\mathbf{g}_i]_h \neq 0,$$

where $[\mathbf{y}]_h$ are the coefficients of \mathbf{y} of order $d = \max(\deg x_i + v_i)$ and $[x_i]_h$ are the coefficients of x_i of order $d - v_i$.

(4 \Rightarrow 5) In view of condition 4, $\mathbf{y} = \mathbf{x}G$ is polynomial and has degree less than d iff \mathbf{x} is polynomial and $\max(\deg x_i + v_i) < d$, or $\deg x_i < d - v_i$ for all i . The vector space (over F) of all polynomials x_i with $\deg x_i < d - v_i$ has dimension $d - v_i$ if $v_i \leq d$ and zero otherwise.

(5 \Rightarrow 1) Reorder the rows of G if necessary so that $v_1 \leq v_2 \leq \dots \leq v_k$. The n -tuples $\sum_{i:v_i > d} x_i \mathbf{g}_i$ with $\deg x_i < d - v_i$ generate a subspace of V_d of dimension $\sum_{i:v_i \geq d} (d - v_i)$, and by condition 5 these are all the n -tuples in V_d . Hence there are fewer than i linearly independent (over $F(x)$) polynomial n -tuples in V of degree less than i . By induction on i , any i linearly independent polynomial n -tuples in V must have sum of degrees at least $\sum_{j=1}^i v_j$.

Remark 1. As this theorem merely spells out fundamental properties of vector spaces over $F(x)$, it would be most surprising if it were not to be found somewhere in the mathematical literature; however, we have yet to discover it. The appropriate mathematical framework in which to approach this result (and certain other basic results in system theory) appears to be that of valuation theory. In the Appendix we outline such an approach and state a generalization of the above theorem, in which it is apparent that the parts (a) and (b) of the above conditions are simply two sides of the same coin.

Remark 2. Those parts of conditions 2, 3 and 4 labeled (a) are equivalent conditions for G to be a basis of the free $F[x]$ -module M_V of polynomial n -tuples in V —i.e., all $y \in M_V$ can be expressed as

$$y = \sum_{i=1}^k x_i g_i$$

for some polynomial k -tuple x —as is explicit in part 4(a). These could be replaced by any other necessary and sufficient condition for G to be a basis of M_V . For example, a left divisor of a $k \times n$ polynomial matrix G is a $k \times k$ polynomial matrix T such that $G = TG'$ for some polynomial matrix G' ; G is a basis for M_V if and only if all its left divisors are unimodular—i.e., have constant (degree zero) determinants. Rosenbrock [1] terms a $k \times k$ matrix T and a $k \times (n - k)$ matrix U “relatively left prime” if and only if $G = [T; U]$ is a basis for M_V .

Remark 3. Condition 4(b) was called the “predictable degree property” in [6], for the following reason. In general, for polynomial x_i and g_i one has $\deg(x_i g_i) = \deg x_i + \deg g_i$, $\deg(\sum x_i g_i) \leq \max(\deg x_i + \deg g_i)$, where the inequality may occur due to cancellation of high order coefficients. Condition 4(b) tells us this never happens when G is minimal, so that knowledge of $\deg x_i$ and v_i , $1 \leq i \leq k$, is sufficient to predict the degree of $y = \sum x_i g_i$. The importance of $[G]_h$ having full rank was recognized by Wolovich [3], who called such matrices “row proper”.

Remark 4. Condition 5 tells us how many polynomial n -tuples there are in V of each degree. Furthermore the formula is invertible; from the numbers $\dim_F V_d$, $d \geq 0$, we can determine the indices v_i . (Example: if $k = 2$ and $\dim_F V_d = 0, 0, 1, 2, 4, 6, 8, \dots$ for $d = 0, 1, 2, \dots$, then $v_1 = 1, v_2 = 2$.) Hence the indices v_i (and also the order v of course) are parameters of V , invariant over all minimal bases. We give them a mellifluous appellation intended to distinguish them from the many other indices in the recent literature.

DEFINITION 4. The *invariant dynamical indices* v_i of a vector space V of n -tuples over $F(x)$ are the indices of any minimal basis for V . Its *invariant dynamical order* v is the sum of the v_i .

4. Reduction of a basis to a minimal basis. We suppose that some basis G for a given vector space V can be found. We now show how one can compute a minimal basis from it.

In general, a reduction algorithm will consist of three steps:

Step 1. If G is not polynomial, multiply each row through by its least common denominator to obtain a polynomial basis.

Step 2. Reduce the given polynomial basis to a basis for the module M_V of polynomial n -tuples in V .

Step 3. Reduce the resulting basis to one with a full-rank high order coefficient matrix; i.e., a “row proper” basis.

The hardest part is Step 2. We now give a method of doing Step 2 which is based on condition 2(a) of the main theorem and allows us to complete the proof that $1 \Rightarrow 2(a)$, but that is undoubtedly not the simplest. One alternate approach is to use the invariant factor theorem, as in [1] and [6]. Wolovich (private communication) believes that the simplest approach is to find the greatest left divisor of G —i.e., the T with largest degree determinant $\det T$ such that $G = TG'$, implying that G' is a basis for M_V .

ALGORITHM FOR STEP 2. Compute the $k \times k$ minors of G and determine their greatest common divisor $\delta(x)$. If the greatest common divisor $\delta(x)$ is not 1, let $p(x)$ be any irreducible polynomial factor of $\delta(x)$. Modulo $p(x)$,¹ G does not have full rank, and therefore there exists some linear combination of the rows of G that is congruent to zero mod $p(x)$:

$$\mathbf{g}' = \sum f_i \mathbf{g}_i \equiv 0 \pmod{p(x)},$$

where the f_i may be taken as polynomials of degree strictly less than $\deg p(x)$ (since they represent residues mod $p(x)$). Since $\mathbf{g}' \equiv 0 \pmod{p(x)}$, \mathbf{g}' is divisible by $p(x)$, and \mathbf{g}'/p has degree

$$\begin{aligned} \deg(\mathbf{g}'/p) &= \deg \mathbf{g}' - \deg p \\ &\leq \max (\deg f_i + \deg \mathbf{g}_i) - \deg p \\ &< \max \deg \mathbf{g}_i, \end{aligned}$$

where the maximum may be taken only over those \mathbf{g}_i for which $f_i \neq 0$. Let \mathbf{g}_{i_0} be a row for which the maximum is attained; then \mathbf{g}_{i_0} may be replaced by \mathbf{g}'/p to get another basis of lower order and with $\delta'(x) = \delta(x)/p(x)$. Repeat until $\delta(x) = 1$. Stop.

Step 3 is easier, as it involves only operations over the base field F . The following algorithm completes the proof that $1 \Rightarrow 2(b)$.

ALGORITHM FOR STEP 3. If the order of G is not equal to the maximum degree of its $k \times k$ minors, then its high order coefficient matrix $[G]_h$ does not have full rank over F . Hence there exists some linear combination $\sum f_i [\mathbf{g}_i]_h$ of the rows $[\mathbf{g}_i]_h$ of $[G]_h$ equal to zero, where $f_i \in F$, $1 \leq i \leq k$. Let v_{i_0} be the greatest of the indices v_i for which $f_i \neq 0$. Then

$$\mathbf{g}' = \sum f_i \mathbf{g}_i x^{v_{i_0} - v_i}$$

is a polynomial n -tuple of degree less than v_{i_0} , since the v_{i_0} th order coefficients cancel, and \mathbf{g}_{i_0} may be replaced by \mathbf{g}' to get another basis of lower order. Repeat until $[G]_h$ has full rank, when the order of G will equal the greatest degree of its $k \times k$ minors, and thus G will be minimal. Stop.

Example.

$F = \text{real numbers,}$

minors

$$G = \begin{bmatrix} 1 & x^{-1} & x^{-1} \\ 1 & x^{-1} & x^{-2} \end{bmatrix}.$$

$$0, x^{-2} - x^{-1}, x^{-3} - x^{-2}$$

Step 1.

$$G = \begin{bmatrix} x & 1 & 1 \\ x^2 & x & 1 \end{bmatrix}.$$

$$0, x^2 - x, x - 1$$

Step 2. Compute $\delta(x) = x - 1$.

$$G \bmod x - 1 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix},$$

¹ Recall that the polynomials modulo an irreducible polynomial form a field.

$$\begin{aligned} \mathbf{g}' &= \mathbf{g}_2 - \mathbf{g}_1 = [x^2 - x, x - 1, 0], \\ \mathbf{g}'/p &= [x \quad 1 \quad 0], \\ G &= \begin{bmatrix} x & 1 & 1 \\ x & 1 & 0 \end{bmatrix}. \end{aligned} \quad \begin{array}{l} \text{minors} \\ 0, x, 1 \end{array}$$

Step 3.

$$\begin{aligned} [G]_h &= \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \\ \mathbf{g}' &= \mathbf{g}_1 - \mathbf{g}_2 = [0 \quad 0 \quad 1], \\ G &= \begin{bmatrix} 0 & 0 & 1 \\ x & 1 & 0 \end{bmatrix}. \end{aligned} \quad 0, x, 1$$

End. $v_1 = 0, v_2 = 1, v = 1$.

5. Minimal bases in echelon form. In general a vector space V has many minimal bases. For some purposes it may be desirable to specify one of these bases as canonical, so that the vector space can be uniquely identified with its canonical basis. We shall now specify such a basis, which because of its resemblance to the ordinary echelon canonical form we shall call a *minimal basis in echelon form*. The proof that every V has such a basis contains a constructive algorithm for computing it.²

DEFINITION 5. Let G be a minimal basis with ordered indices $v_1 \leq v_2 \leq \cdots \leq v_k$. The i th *pivot index* γ_i of G is the least integer such that the matrix G_i formed from the intersection of columns $\gamma_1, \dots, \gamma_i$ with the rows of G of index $\leq v_i$ has high order coefficient matrix $[G_i]_h$ of rank i .

Remark 5. Note that there may be more than i rows of G of index $\leq v_i$ if $v_{i+1} = v_i$.

Remark 6. Another way of stating this definition would be as follows. Let there be n_1 rows with index v_1 . Find the first (lowest index) n_1 columns of $[G]_h$ such that the $n_1 \times n_1$ submatrix of $[G]_h$ so defined is nonsingular over F . The indices of these columns, in order, form the first n_1 pivot indices. Delete these n_1 rows and n_1 columns from G and repeat the above procedure to find the next group of pivot indices, corresponding to the rows with the next distinct index value; and so forth.

LEMMA 1. All minimal bases (with ordered indices) for the same vector space V have the same pivot indices.

Proof. Let G and G' be two minimal bases with ordered indices for the same V . Then \mathbf{g}'_i , being a polynomial n -tuple in V of degree v_i , is a linear combination of the rows \mathbf{g}_j of G of degree $v_j \leq v_i$:

$$\mathbf{g}'_i = \sum_{j: v_j \leq v_i} x_j \mathbf{g}_j, \quad \deg x_j \leq v_j - v_i.$$

²This section was written after A. Eckberg of M.I.T. and B. Dickinson and M. Morf of Stanford had noted that a connection to Popov's work could be made by properly defining a canonical basis. While the details have here been independently rederived, credit for priority undoubtedly belongs jointly to these gentlemen.

Further, the i th row $[\mathbf{g}'_i]_h$ of the high order coefficient matrix of G' is a linear (over F) combination of the rows $[\mathbf{g}_j]_h$ of $[G]_h$ such that $v_j \leq v_i$:

$$[\mathbf{g}'_i]_h = \sum_{j: v_j \leq v_i} [x_{ij}]_h [\mathbf{g}_j]_h.$$

In other words, the matrix $[X]_h$ such that $[G']_h = [X]_h [G]_h$ is block lower triangular, where the blocks correspond to rows \mathbf{g}_j of the same index v_i ; e.g., if $k = 5$ and $v_1 = v_2 < v_3 = v_4 = v_5$, then $[X]_h$ has the form

$$[X]_h = \left[\begin{array}{cc|ccc} X & X & 0 & 0 & 0 \\ X & X & 0 & 0 & 0 \\ \hline X & X & X & X & X \\ X & X & X & X & X \\ X & X & X & X & X \end{array} \right].$$

It follows that the matrix $[G'_i]_h$ formed by the intersection of some set of i columns $(\beta_1, \dots, \beta_i)$ with the rows of $[G'_i]_h$ of index $\leq v_i$ can have rank i only if the corresponding matrix $[G_i]_h$ has rank i , and vice versa since the argument can be inverted. Since any minimal basis G must have some set of i columns such that $[G_i]_h$ has rank i , since $[G]_h$ has full rank k , the lemma is proved.

DEFINITION 6. A minimal basis G is said to be in *echelon form* if

- (a) its indices are ordered so that $v_1 \leq v_2 \leq \dots \leq v_k$;
- (b) the polynomials g_{i, γ_i} are monic polynomials of degree v_i , where γ_i is the i th pivot index;
- (c) for any i and i' such that $v_i \leq v_{i'}$, $\deg g_{i', \gamma_{i'}} < v_i$.

Remark 7. These conditions imply that the high order coefficient matrix $[G]_h$ of G has a 1 as its (i, γ_i) th element, and zeros in all positions (i', γ_i) such that $v_{i'} \geq v_i$. For example, if $v_1 = v_2 < v_3 = v_4 = v_5$, a typical $[G]_h$ for G in echelon form would look like

$$\left[\begin{array}{ccccccccc} 0 & 0 & 1 & X & X & 0 & X & X \\ 0 & 0 & 0 & 0 & 0 & 1 & X & X \\ \hline 1 & X & 0 & 0 & 0 & 0 & X & X \\ 0 & 0 & 0 & 1 & 0 & 0 & X & X \\ 0 & 0 & 0 & 0 & 1 & 0 & X & X \end{array} \right],$$

where the pivot indices are (3, 6, 1, 4, 5).

THEOREM 2. Every vector space V of n -tuples over $F(x)$ has a unique minimal basis in echelon form.

Proof. We first show how to go from a given minimal basis G to another minimal basis G' in echelon form and then prove that the resulting basis is unique. First reorder the rows of G if necessary so that $v_1 \leq v_2 \leq \dots \leq v_k$, and temporarily reorder the columns so that the pivot indices are $1, \dots, k$ —i.e., reindex the γ_1 th column as the first column, the γ_2 th as the second, etc. If there are m distinct indices

v_i , let us think of G as being divided into the corresponding blocks; e.g., if $v_1 = v_2 < v_3 = v_4 = v_5$, then $m = 2$ and G is partitioned as follows:

$$G = \begin{bmatrix} X & X & X & X & X & X & X & X & \cdots \\ X & X & X & X & X & X & X & X & \cdots \\ \hline X & X & X & X & X & X & X & X & \cdots \\ X & X & X & X & X & X & X & X & \cdots \\ X & X & X & X & X & X & X & X & \cdots \end{bmatrix}.$$

We call the m square submatrices that appear in this partition the diagonal blocks. Let $[G_{d1}]_h$ be the first diagonal block of $[G]_h$; by the definition of pivotal indices $[G_{d1}]_h$ is nonsingular. Now multiply G on the left by the $m \times m$ block matrix whose first diagonal block is $[G_{d1}]_h^{-1}$, whose other diagonal blocks are identity matrices, and whose off-diagonal blocks are zero; e.g.,

$$T_1 = \begin{bmatrix} [G_{d1}]_h^{-1} & 0 \\ \hline 0 & I \end{bmatrix}.$$

Then the indices of G are unaffected, but the first diagonal block of $[G]_h$ becomes an identity matrix; e.g.,

$$[G]_h = \begin{bmatrix} 1 & 0 & X & X & X & X & X & X \\ 0 & 1 & X & X & X & X & X & X \\ \hline X & X & X & X & X & X & X & X \\ X & X & X & X & X & X & X & X \\ X & X & X & X & X & X & X & X \end{bmatrix}.$$

Let us now consider the set of elements g_{ji} such that $v_i = v_1$ and $v_j > v_1$; i.e., the elements of G lying below the first diagonal block. We call these the elements *dominated* by the first diagonal block. We now show how we can reduce G to a matrix in which all these elements have degree $< v_1$. Let g_{ji} be the element of greatest degree dominated by the first diagonal block, let μ be its degree, and let α be the coefficient of x^μ . Then if $\mu \geq v_i = v_1$, subtraction of $\alpha x^{\mu-v_1} \mathbf{g}_i$ from \mathbf{g}_j gives a row \mathbf{g}'_j in which $\deg g'_{ij} < \mu$; furthermore, no new element of degree μ is introduced by this step since $\deg x^{\mu-v_1} \mathbf{g}_{i'} < \mu$ for $i' \neq i$ such that $v_{i'} = v_1$. Hence repetition of this procedure can only terminate when all elements dominated by the first diagonal block have degree less than v_1 .

Now let $[G_{d2}]_h$ be the second diagonal block of the resulting $[G]_h$. It must be nonsingular since

$$[G]_h = \begin{bmatrix} I & X & \cdots \\ \hline 0 & [G_{d2}]_h & \cdots \\ \hline \vdots & \vdots & \ddots \end{bmatrix}$$

and the definition of pivot indices forces this corner of $[G]_h$ to have full rank. Left multiply G by the constant diagonal block matrix

$$T_2 = \left[\begin{array}{c|c|c} I & 0 & 0 \\ \hline 0 & [G_{d^2}]_h^{-1} & 0 \\ \hline 0 & 0 & I \end{array} \right].$$

This does not affect the first block of rows of G , nor does it increase the maximum degree of the elements dominated by the first diagonal block since T_2 just mixes constant multiples of these rows together. However, the second diagonal block of $[G]_h$ is now an identity matrix. Continuing, we may reduce the degrees of the elements dominated by the second diagonal block to less than the second distinct index of G , then make the third diagonal block of $[G]_h$ the identity, etc.; finally, returning the columns to their original order, we arrive at a minimal basis in echelon form.

To prove uniqueness, let G be some minimal basis in echelon form with indices $\{v_i\}$ and pivot indices $\{\gamma_i\}$. The i th row of any minimal basis in echelon form must have a monic polynomial of degree v_i in column γ_i , polynomials of degree $< v_j$ in columns γ_j such that $v_j \leq v_i$, and polynomials of degree $\leq v_i$ in all other columns. Such an n -tuple must be a linear combination of the rows of G of index $\leq v_i$. But if we add any multiple of any generator \mathbf{g}_j of index $v_j \leq v_i$ to \mathbf{g}_i , we get an n -tuple with coefficient of degree at least v_j in column γ_j . Hence V contains only one polynomial n -tuple satisfying the above conditions for each i . This completes the proof.

Example 1. The minimal basis

$$G = \begin{bmatrix} 0 & 0 & 1 \\ x & 1 & 0 \end{bmatrix}$$

previously found is already in echelon form; its pivot indices are (3, 1). Note that whenever $v_1 = 0$ all elements dominated by the first diagonal block must be zero.

Example 2. For a slightly less trivial example, consider the minimal basis ($F = \text{real numbers}$)

$$G = \begin{bmatrix} 1 & x & x-1 & x-2 \\ 1 & 0 & x+1 & 1 \\ x^2 & x^2-1 & 0 & 0 \end{bmatrix}$$

with indices 1, 1, 2. From the high order coefficient matrix

$$[G]_h = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

we determine that the pivot indices are (2, 3, 1). Premultiplying G by

$$T_1 = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

we obtain

$$G = \begin{bmatrix} 0 & x & -2 & x-3 \\ 1 & 0 & x+1 & 1 \\ x^2 & x^2-1 & 0 & 0 \end{bmatrix}.$$

The last row is successively replaced by $[x^2, -1, 2x, -x^2 + 3x]$ ($\mathbf{g}_3 - x\mathbf{g}_1$) and $[x^2 - 2, -1, -2, -x^2 + 1]$ ($\mathbf{g}_3 - 2\mathbf{g}_2$). Since $\mathbf{g}_{3,1}$ is already monic we now have our basis in echelon form:

$$G = \begin{bmatrix} 0 & x & -2 & x-3 \\ 1 & 0 & x+1 & 1 \\ x^2-2 & -1 & -2 & -x^2+1 \end{bmatrix}.$$

6. Dual spaces and inverses. If G is a full-rank $k \times n$ matrix over $F(x)$, then the set of all column n -tuples \mathbf{z} such that $G\mathbf{z} = 0$ is a vector space over $F(x)$, called the *dual space* V^\perp . Every n -tuple $\mathbf{z} \in V^\perp$ is orthogonal to every n -tuple $\mathbf{y} \in V$ ($\mathbf{y}\mathbf{z} = 0$). The dimension of V^\perp is $n - k$, and it has some $n \times (n - k)$ minimal basis G^\perp .

The following theorem appears in [6, Lemma 8]; it is restated here with a proof more in the spirit of this paper.

THEOREM 3. *Let V^\perp be the dual space to V . Then the minors of V^\perp are the minors of V .*

Proof. Let G be any basis for V . Since G has rank k , it has some nonsingular $k \times k$ submatrix W , which (by relabeling columns if necessary) we may take to comprise the first k columns. Then $W^{-1}B$ is a basis for V of the form $[I_k; H]$, where I_k is the $k \times k$ identity matrix. Now the $n \times (n - k)$ matrix

$$G^\perp \triangleq \begin{bmatrix} H \\ \text{-----} \\ -I_{n-k} \end{bmatrix}$$

clearly has rank $n - k$ and satisfies $W^{-1}GG^\perp = 0$, hence is a basis for V^\perp . Further, the determinant of the submatrix formed from the set of columns $J = \{j_1, j_2, \dots, j_k\}$ of $W^{-1}G$ is the same (up to sign) as the determinant of the submatrix formed from the complementary set of rows $\bar{J} = \{1, 2, \dots, n\} - J$ in G^\perp . (Verification: if J_H is the set of indices in J greater than k and $J_L = J - J_H$, then the determinant in either case is (up to sign) that of the rows \bar{J}_L and columns J_H of H .) Hence $W^{-1}G$ and G^\perp have the same set of minors, hence the same set of normalized minors; hence V and V^\perp have the same minors.

COROLLARY. *The invariant dynamical orders of V and V^\perp are the same.*

Of course the invariant dynamical indices are in general different since $k \neq n - k$ in general.

A related result that will not be proved here is the following.

THEOREM 4. *Let G be a minimal basis. Then there exists an $(n - k) \times n$ polynomial matrix H such that the square matrix*

$$B = \begin{bmatrix} G \\ --- \\ H \end{bmatrix}$$

has unit (constant) determinant (is unimodular). Hence B has a polynomial inverse $B^{-1} = [G^{-1}; H^{-1}]$, and G has a polynomial right inverse G^{-1} . H^{-1} may be taken to be a minimal basis of the dual space.

The simplest proof is by the invariant factor theorem [1], [6] (Smith canonical form). For Theorem 4 to hold it is not actually necessary that G be minimal, only that it be a basis for M_v (conditions 2(a), 3(a) or 4(a) of the main theorem). That G must have a polynomial right inverse is almost obvious from condition 4(a).

7. Application I: Controller canonical realizations from transfer function matrices. A transfer function matrix of an m -input, p -output constant finite-dimensional linear system is normally written as a $p \times m$ -matrix G of rational functions, namely $F(z)$ or $F(D)$ in the case of discrete-time systems (z -transforms, D -transforms), or $F(s)$ in the case of continuous-time systems (Laplace transforms). We shall mostly speak in terms of discrete-time systems and z -transforms, but as is well known the results for continuous-time systems are analogous. The m input sequences

$$u_i = u_{id_i} z^{-d_i} + u_{id_i+1} z^{-d_i-1} + \cdots, \quad 1 \leq i \leq m,$$

are written as a column m -tuple \mathbf{u} , and the p output sequences y_j , $1 \leq j \leq p$, as a column p -tuple \mathbf{y} . A given input \mathbf{u} generates the output $\mathbf{y} = G\mathbf{u}$.

A rational function is called proper if the degree of its numerator is not greater than the degree of its denominator, and strictly proper if strict inequality holds. When speaking about discrete-time systems we shall use as equivalents the terms causal and strictly causal; when expanded as sequences such functions have non-zero coefficients only for nonpositive and negative powers of z , respectively. A proper (strictly proper, causal, strictly causal) matrix is a rational matrix all of whose elements are proper (strictly proper, etc.). We shall consider only proper systems; for an extension to nonproper systems see Rosenbrock and Hayton [18].

We define the *input/output vector* corresponding to an m -tuple \mathbf{u} as the $(m + p)$ -tuple

$$\begin{bmatrix} \mathbf{u} \\ --- \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} I_m \\ --- \\ G \end{bmatrix} \mathbf{u} = \tilde{G}\mathbf{u},$$

where the $(m + p) \times m$ matrix $\tilde{G} \triangleq [I_m; G^T]^T$ is defined as the *extended transfer function matrix* corresponding to G . The space of all input/output vectors is therefore the vector space V_G of column $(m + p)$ -tuples over $F(z)$ generated by \tilde{G} .

The polynomial $(m + p)$ -tuples in this space have a special system-theoretic significance. A polynomial sequence, in the discrete-time case, is one that "stops" at time 0. If an input/output vector is polynomial, then it corresponds to an input that stops no later than time 0 and that generates an output that also stops no

later than time 0—i.e., that has no observable consequences at time 0 or later. Thus in a minimal realization such an input should leave the system in the $\mathbf{0}$ state at time 0^+ , and the set of all inputs that leave the system in state $\mathbf{0}$ at time 0^+ are precisely the set of inputs associated with the module M_G of polynomial $(m + p)$ -tuples in V_G .

It is therefore not surprising that a minimal basis for the space V_G should have especially useful properties.

DEFINITION 7. A *minimal (controllability) system matrix* for a system with $p \times m$ transfer function matrix G is a minimal basis $S = \begin{bmatrix} D \\ - \\ N \end{bmatrix}$ for the space V_G of input/output vectors generated by $\tilde{G} = \begin{bmatrix} I_m \\ - \\ G \end{bmatrix}$, where D is $m \times m$ and N is $p \times m$.

The columns $\begin{bmatrix} \mathbf{d}_i \\ - \\ \mathbf{n}_i \end{bmatrix}$ of a minimal (controllability) system matrix S are of course themselves polynomial input/output vectors, which will turn out to generate a convenient basis for the state space (controller canonical realization). We shall call them *elementary input/output vectors*.

The following theorem appears to be one of those results that develops piecemeal over the years through the loosely related contributions of numerous authors, and we do not have the temerity to attempt properly to apportion the credit. If specialized to minimal system matrices in echelon form, it seems to be essentially due to Popov [2], who did not however publish his proof on grounds of length. Rosenbrock [18] notes his own earlier related work, which does not however identify the v_i as the controllability indices. With regard to the realization algorithms, Wolovich [3] and Wang [4] have presented similar ones.

THEOREM 5. Let $S = \begin{bmatrix} D \\ - \\ N \end{bmatrix}$ be a minimal (controllability) system matrix for a proper matrix G , and let v_i , $1 \leq i \leq m$, and v be the indices and order of S . Then:

1. $\det D = \psi(z)$ is the characteristic polynomial of G , and $v = \deg \psi$;
2. $G = ND^{-1}$;
3. the v_i are the controllability indices of G ;
4. G has a minimal realization of dimension v ;
5. N and D determine a minimal realization of G in controller canonical (Luenberger [7]) form.

Proof. We first note that the minors of \tilde{G} are all causal, and of course its principal minor (the determinant of the first m rows) is $|I_m| = 1$. Hence the normalized minors of \tilde{G} are obtained simply by multiplying through by their least common denominator $\psi(z)$, and all have degree $\leq \deg \psi$, with equality for the principal normalized minor $\psi(z)$. Hence $v = \deg \psi$, $|D| = \psi(z)$, and $[D]_h$ has full rank. We can identify $\psi(z)$ as the characteristic polynomial by noting that the $m \times m$ minors of G are the minors of G of all orders, since it is well known that the characteristic polynomial of G is the least common denominator of its minors of all orders.

The space of input/output vectors can also be characterized as the space of $(m + p)$ -tuples orthogonal to $\tilde{G}^\perp = [G \mid -I_p]$, a $p \times (m + p)$ matrix. Property 2 then follows simply from $\tilde{G}^\perp S = 0$, or $GD = N$, since D has a nonzero determinant and hence is invertible.

To prove conditions 3 and 4, we shall first offer an abstract discussion; subsequently we give a constructive proof by constructing the controller canonical form of condition 5.

Let \mathbf{u} be any polynomial input m -tuple; that is, a set of sequences that are anticausal, or stop at time zero or before. The (I/O) state $\mathbf{s}(\mathbf{u})$ generated by \mathbf{u} is then defined as the sequence of outputs observed at time 1 and later after an input of \mathbf{u} ; symbolically, $\mathbf{s}(\mathbf{u}) = Q(G\mathbf{u})$, where Q is the operator that truncates sequences to time 1 and later ($Q(\cdots + u_0 + u_1 z^{-1} + u_2 z^{-2} + \cdots) = u_1 z^{-1} + u_2 z^{-2} + \cdots$). Two anticausal inputs \mathbf{u}_1 and \mathbf{u}_2 are said to be (Nerode) equivalent if they lead to the same state.

The map $f: F^m[z] \rightarrow \Sigma; \mathbf{u} \mapsto \mathbf{s}(\mathbf{u})$ is a linear map (over F) from the space $F^m[z]$ of all anticausal inputs to the I/O state space Σ . The kernel $\text{Ker } f$ of this map is the set of all \mathbf{u} such that $\mathbf{s}(\mathbf{u}) = \mathbf{0}$, or the set of all \mathbf{u} such that the input/output vector $\tilde{G}\mathbf{u}$ is polynomial; i.e., the module M_G of the initial m -tuples in all polynomial $(m + p)$ -tuples in the space generated by \tilde{G} . By a standard algebraic result the I/O state space Σ is isomorphic to the (Nerode) equivalence classes of $F^m[z]$; i.e., $\Sigma \cong F^m[z]/\text{Ker } f = F^m[z]/M_G$.

From condition 5 of the main theorem, we verify that the dimension of the space of all anticausal inputs of degree less than d up to equivalence mod M_G is

$$md - \sum_{i: d \geq v_i} (d - v_i) = v - \sum_{i: v_i \geq d} (v_i - d).$$

For $d \geq \max(v_i)$, this dimension is v , so that the I/O state space has dimension v . Since every I/O state must correspond to a distinct physical state, the physical state space must be at least as large as the I/O state space; that a realization always exists in which these two spaces are isomorphic is a standard result that we shall shortly reprove.

There are a number of definitions of controllability indices, but an immediate consequence of any of them is that the dimension of the space of anticausal inputs of degree less than d up to Nerode equivalence satisfies one of the formulas above, so that we can identify the v_i as the controllability indices of G (i.e., the controllability indices of $[A, B]$ in any minimal realization of G).

The proof is completed by the algorithm for constructing a controller canonical realization, which shall be our next topic.

REALIZATION ALGORITHM. We are given a $p \times m$ causal transfer function matrix G . For convenience in notation we shall work with its transpose G^T . We first reduce the extended transfer function matrix $\tilde{G}^T = [I_m \mid G^T]$ to a minimal system matrix $S^T = [D^T \mid N^T]$, whose rows are the m elementary input/output vectors, and whose indices and order are v_i and $v = \sum v_i$.

We already know that a minimal realization will have v memory elements. The idea of the construction is to set up these memory elements as m shift registers of lengths v_i , $1 \leq i \leq m$, in such a way that when one of the elementary inputs \mathbf{u}_i occurs, starting at time $-v_i$ and ending at time 0, a single 1 enters the i th shift

register at time $-v_i + 1$ and simply shifts down one stage at each succeeding time unit until it drops off the end after time 0. In other words, set up a v -dimensional state vector $\mathbf{x} = \{x_{ij}, 0 \leq j \leq v_i - 1, 1 \leq i \leq m\}$; choose the $v \times m$ input/state transfer function G_0 so that the minimal system matrix for G_0 has the form $[D^T; U^T]$, where U^T is the $m \times v$ matrix

$$U^T = \begin{bmatrix} 1 & z & \dots & z^{v_1-1} & 0 & 0 & 0 \\ 0 & 1 & \dots & z^{v_2-1} & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & 0 & 0 & 1 & \dots & z^{v_m-1} \end{bmatrix}.$$

(For any i such that $v_i = 0$ the corresponding blocks are missing.)

It is natural to regard \mathbf{x} as a column m -vector of polynomials in z^{-1} , $\{x_i, 1 \leq i \leq m\}$, where

$$x_i = \sum_{j=0}^{v_i-1} x_{ij} z^{-j}.$$

We define $[\mathbf{x}]_h$ as the column m -vector of high order coefficients $\{x_{i,v_i-1}, 1 \leq i \leq m\}$. Finally, we define $[D\mathbf{x}]_0$ as the column m -vector of coefficients of z^0 in the matrix product $D\mathbf{x}$. Letting $[D]_h$ be the matrix of high order coefficients of D and $[D]_h^{-1}$ its inverse, we arrive at the desired input/state equations.

LEMMA 2. *The input/state equations*

$$\begin{aligned} [\mathbf{x}^{t+1}]_h &= [D]_h^{-1}(-[D\mathbf{x}^t]_0 + \mathbf{u}^t), \\ x_{ij}^{t+1} &= x_{i,j+1}^t, \quad 0 \leq j \leq v_i - 2, \quad 1 \leq i \leq m, \end{aligned}$$

realize a system whose input/state minimal system matrix is

$$S^T = [D^T; U^T].$$

Proof. Let the i th column of D be the elementary input vector \mathbf{d}_i . Let the state \mathbf{x}^{-v_i} at time $-v_i$ be zero, and let the input sequence actually be \mathbf{d}_i . Then at time $-v_i$ the input will be $[\mathbf{d}_i]_h$, and since $[D]_h^{-1}[\mathbf{d}_i]_h = \mathbf{e}_i$ (the column vector with a single 1 in the i th position and 0 elsewhere), the state \mathbf{x}^{-v_i+1} will be $z^{-v_i+1}\mathbf{e}_i$. But now $[Dz^{-v_i+1}\mathbf{e}_i]_0 = [D\mathbf{e}_i]_{v_i-1} = [d_i]_{v_i-1}$ where the notation $[\cdot]_l$ means the coefficient of z^l ; hence if the input \mathbf{u}^{-v_i+1} at time $-v_i + 1$ is equal to $[\mathbf{d}_i]_{v_i-1}$, then $\mathbf{u}^t - [D\mathbf{x}^t]_0 = 0$, and $[\mathbf{x}^{-v_i+2}]_h = 0$; thus \mathbf{x}^{-v_i+2} will be $z^{-v_i+2}\mathbf{e}_i$ by the second state equation. Continuing by induction, we prove that when \mathbf{d}_i is the input sequence, the state sequence satisfies

$$\mathbf{x}^t = \begin{cases} z^t \mathbf{e}_i, & -v_i + 1 \leq t \leq 0, \\ 0, & \text{otherwise.} \end{cases}$$

This proves that the columns of S are all valid input/state vectors. But since D is nonsingular, all inputs are linear combinations of these elementary inputs, and the lemma is proved.

Remark 8. These input/state equations are of the so-called controller canonical form, which is useful in considering controllability and state feedback. It is obvious

that such a system can be transformed into pure shift-register (control canonical, Brunovsky [8]) form by state feedback to the input ($+ [D\mathbf{x}^t]$) and change of basis of the input space ($[D]_h^{-1}$). That the controllability indices are invariant parameters under state feedback is one of their most important properties; it is proved by noting that state feedback cannot alter the dimensions of the spaces V_d of input/output vectors of degree less than d .

Remark 9. For a given transfer function matrix G we can obtain a unique D by requiring that the minimal system matrix S be in echelon form. It is not hard to show [B. Dickinson and M. Morf, private communication; also A. Eckberg] that in this case the coefficients of the polynomial d_{ij} in the D -matrix are Popov's parameters α_{ijk} [2]. Popov has also shown [9] that these parameters are a complete set of invariants for $[A, B]$ matrices under state-space transformations; our methods here applied to the D -matrix provide an easy alternate proof.

Remark 10. The foregoing results are closely related to certain results of Rosenbrock [1]. Rosenbrock considers the space of all state/input vectors, defined

as the set of all $(n + m)$ -tuples $\begin{bmatrix} \mathbf{x} \\ - \\ \mathbf{u} \end{bmatrix}$ that are orthogonal to the matrix $[zI_n - A \mid B]$.

His results can be related to those of this paper by statements such as the following: The controllability indices of the matrices $[A, B]$, or the minimal indices of the singular pencil of matrices $[zI_n - A \mid B]$ (in the sense of Kronecker), are the invariant dynamical indices of the vector space dual to the space generated by $[zI_n - A \mid B]$.

We are not quite done in the realization; we still need an output equation. This is easy if G is strictly causal, for then in the minimal system matrix $S^T = [D^T \mid N^T]$ we must have $\deg n_{ij}$ strictly less than v_i , since all outputs are delayed at least one from the corresponding input. The output equation

$$\mathbf{y}^t = [N\mathbf{x}^t]_0$$

is then seen to realize G ; for if \mathbf{d}_i is the input, the state at time t is $\mathbf{x}^t = z^t \mathbf{e}_i$ for $-v_i + 1 \leq t \leq 0$, and the output is $\mathbf{y}^t = [Nz^t \mathbf{e}_i]_0 = [\mathbf{n}_i]_{-t}$ as it should be.

If G is not strictly causal, then we can write it uniquely as the sum of a strictly causal part G_0 and an anticausal remainder term E which will simply be a constant matrix if G is causal. Then if $S_0^T = [D_0^T \mid N_0^T]$ is a minimal system matrix for G_0 , we have the realization

$$\begin{aligned} [\mathbf{x}^{t+1}]_h &= [D_0]_h^{-1}(-[D_0\mathbf{x}^t] + \mathbf{u}^t), \\ x_{ij}^{t+1} &= x_{i,j+1}^t, \quad 0 \leq j \leq v_i - 2, \quad 1 \leq i \leq m, \\ \mathbf{y}^t &= [N_0\mathbf{x}^t]_0 + E\mathbf{u}^t \end{aligned}$$

when E is a constant matrix. An alternate approach, which is more cumbersome if the goal is the (A, B, C, E) matrices but is perhaps more consistent generally with our methods, is to use the minimal system matrix $S^T = [D^T \mid N^T]$ for G directly, from which we get the equations

$$\begin{aligned} [\mathbf{x}^{t+1}]_h &= [D]_h^{-1}(-[D\mathbf{x}^t] + \mathbf{u}^t), \\ x_{ij}^{t+1} &= x_{i,j+1}^t, \quad 0 \leq j \leq v_i - 2, \quad 1 \leq i \leq m, \\ \mathbf{y}^t &= [N\mathbf{x}^t]_0 + [N]_h[\mathbf{x}^{t+1}]_h. \end{aligned}$$

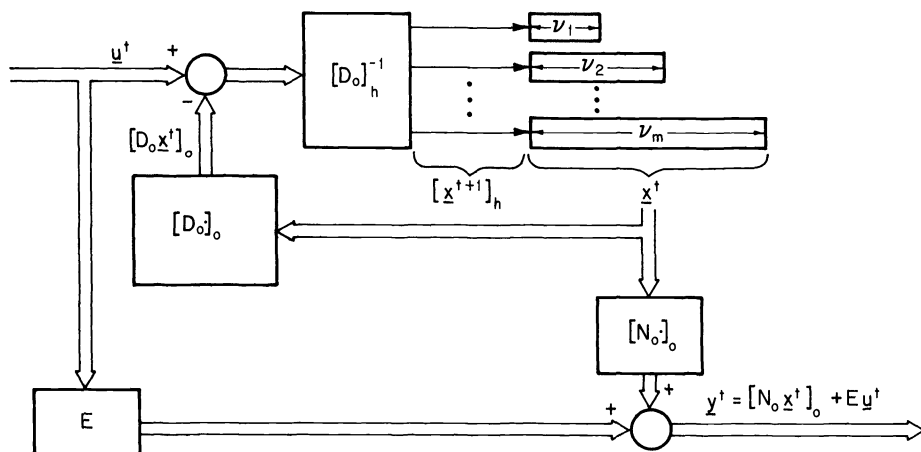


FIG. 1. Controller canonical realization of transfer function matrix $G = G_0 + E$, where G_0 is strictly proper and has minimal system matrix $S_0 = [D_0 \ N_0]$ and controllability indices v_i , $1 \leq i \leq m$

These two structures are illustrated in Figs. 1 and 2.

Example. Let

$$G^T = \frac{1}{z^2 + 3z + 2} \begin{bmatrix} z + 1 & z + 3 & z^2 + 3z \\ z + 2 & z^2 + 2z & 0 \end{bmatrix}.$$

Since G is not strictly causal, we decompose it as follows:

$$\begin{aligned} G^T &= \frac{1}{z^2 + 3z + 2} \begin{bmatrix} z + 1 & z + 3 & -2 \\ z + 2 & -z - 2 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \\ &= G_0^T + E^T. \end{aligned}$$

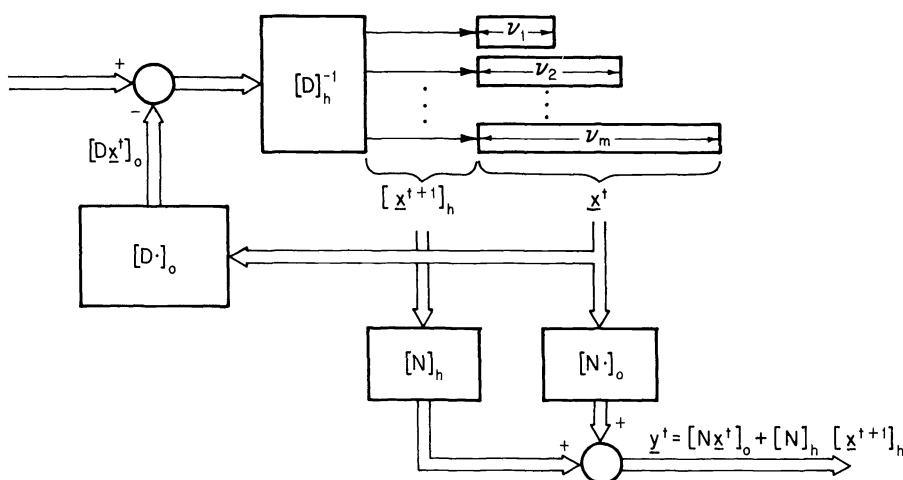


FIG. 2. Alternate form for controller canonical realization, from the minimal system matrix $S = [D \ N]$ of G

Then

$$\tilde{G}_0^T = \frac{1}{z^2 + 3z + 2} \left[\begin{array}{cc|ccc} z^2 + 3z + 2 & 0 & z + 1 & z + 3 & -2 \\ 0 & z^2 + 3z + 2 & z + 2 & -z - 2 & 0 \end{array} \right].$$

We obtain a minimal system matrix by multiplying through by $z^2 + 3z + 2$ and dividing the second row by $z + 2$ to get

$$S_0^T = \left[\begin{array}{cc|ccc} z^2 + 3z + 2 & 0 & z + 1 & z + 3 & -2 \\ 0 & z + 1 & 1 & -1 & 0 \end{array} \right].$$

The controllability indices are thus $v_1 = 2$, $v_2 = 1$, and the state-space form is

$$\begin{aligned} \begin{bmatrix} x_{10}^{t+1} \\ x_{11}^{t+1} \\ x_{20}^{t+1} \end{bmatrix} &= \begin{bmatrix} 0 & 1 & 0 \\ -2 & -3 & 0 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} x_{10}^t \\ x_{11}^t \\ x_{20}^t \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_1^t \\ u_2^t \end{bmatrix}, \\ \begin{bmatrix} y_1^t \\ y_2^t \\ y_3^t \end{bmatrix} &= \begin{bmatrix} 1 & 1 & 1 \\ 3 & 1 & -1 \\ -2 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_{10}^t \\ x_{11}^t \\ x_{20}^t \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} u_1^t \\ u_2^t \end{bmatrix}. \end{aligned}$$

Here the coefficients of D_0 and N_0 appear directly in the A and C matrices since $[D_0]_h$ is the identity matrix. This realization is illustrated in Fig. 3. An alternate realization in the form of Fig. 2 (but with the same state-space equations) can be obtained from the minimal system matrix

$$S^T = [D^T : N^T] = \left[\begin{array}{ccc|ccc} z^2 + 3z + 2 & 0 & z + 1 & z + 3 & z^2 + 3z \\ 0 & z + 1 & 1 & z & 0 \end{array} \right]$$

and is illustrated in Fig. 4.

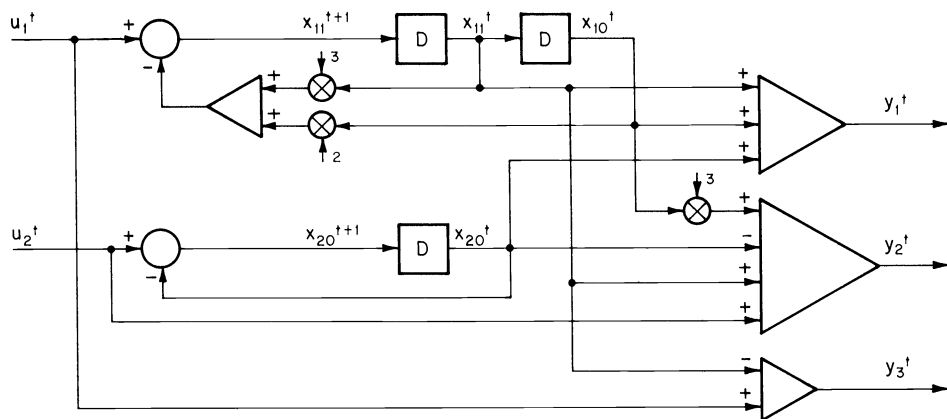
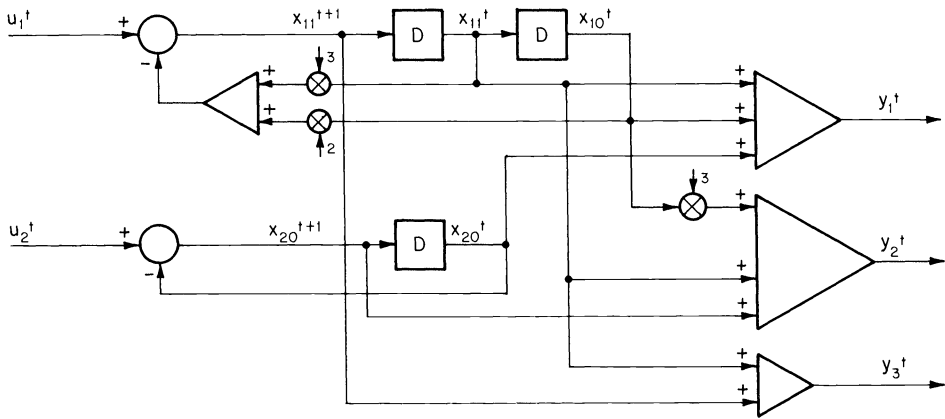


FIG. 3. Realization of the type of Fig. 1 for the G of §7

FIG. 4. Realization of the type of Fig. 2 for the G of §7

The corresponding results for observability indices, observer canonical forms, and so forth can be obtained from the standard duality results. The analogues of Definition 7 and Theorem 5 are stated here for reference.

DEFINITION 7¹. A *minimal (observability) system matrix* for a system with $p \times m$ transfer function matrix G is a minimal basis $S = [D : N]$ for the space $V_{\tilde{G}}$ of $(p + m)$ -tuples generated by $\tilde{G} = [I_p : G]$, where D is $p \times p$ and N is $p \times m$.

THEOREM 5¹. Let $S = [D : N]$ be a minimal (observability) system matrix for G , and let μ_j , $1 \leq j \leq p$, and μ be the indices and order of S . Then

1. $\det D = \psi(z)$ is the characteristic polynomial of G , and $\mu = \deg \psi$;
2. $G = D^{-1}N$;
3. the μ_i are the observability indices of G ;
4. G has a minimal realization of dimension μ ;
5. N and D determine a minimal realization of G in observer canonical form.

A more interesting approach to observability questions will appear in a forthcoming note.

8. Application II: Minimal causal solutions to $PG = Q$. In an interesting paper that was the direct stimulus for this work, Wang and Davison [5] consider the problem of finding a causal $p \times m$ $F(z)$ -matrix (or proper $F(s)$ -matrix) G of minimal order (as a transfer function matrix) that solves the equation $PG = Q$, where P and Q are given $l \times p$ and $l \times m$ matrices, respectively. They point out that such problems frequently arise in multivariable linear system theory; for example, the problem of finding a minimal order inverse H^{-1} to a given system H can be cast in this form, with $P = H^T$, $G = (H^{-1})^T$ and $Q = I_m$.

In this section, using a minimal basis approach and the results of the last section, we reproduce and to some extent generalize Wang and Davison's [5] results on this problem. In our view this approach is much more transparent.

The general idea is as follows. Let G be any causal solution to $PG = Q$ and let $S^T = [D^T : N^T]$ be a minimal (controllability) system matrix for G ; then $G = ND^{-1}$, so $PN = QD$, or $\tilde{P}S = 0$, where \tilde{P} is defined as $[Q : -P]$. Thus the

columns s_i for any minimal system matrix for a G that solves $PG = Q$ lie in the vector space $V_{\tilde{P}}^\perp$ of column $(m + p)$ -tuples orthogonal to \tilde{P} . By finding a minimal basis S^* for $V_{\tilde{P}}^\perp$, we find a basis for all such polynomial $(m + p)$ -tuples. S is thus a minimal system matrix for a G satisfying $PG = Q$ if and only if the columns of S are linear combinations of the columns of S^* and S is a minimal system matrix. We now need the following lemma.

LEMMA 3. *A matrix $S^T = [D^T \ ; \ N^T]$ is a minimal system matrix for a causal transfer function matrix $G = ND^{-1}$ if and only if S is a minimal basis for the space it spans and $[D]_h$ has full rank.*

Remark 11. $[D]_h$ has full rank implies $\deg n_{ij} \leq v_i$ for all i, j .

Proof. Necessity was proven in Theorem 5. For sufficiency, if $[D]_h$ has full rank, then D spans the space of all inputs, so for any \mathbf{u} we can write $\mathbf{u} = \sum \psi_i \mathbf{d}_i$ for some rational coefficients ψ_i , or in fact $\mathbf{u} = \sum \psi_i' \mathbf{d}_i z^{-v_i}$ for some shifted sequences ψ_i' . Since $\mathbf{d}_i z^{-v_i}$ starts at time 0, and $[D]_h = \{\mathbf{d}_i z^{-v_i}\}_0$ has full rank, \mathbf{u} must start at the first time one of the ψ_i' starts, say time μ , since complete cancellation of the $z^{-\mu}$ coefficients in \mathbf{u} cannot occur. But the output \mathbf{y} generated by \mathbf{u} is $\mathbf{y} = \sum \psi_i' \mathbf{n}_i z^{-v_i}$, which starts at time μ or later since all the ψ_i' start at time μ or later and all the $\mathbf{n}_i z^{-v_i}$ start at time 0 or later because $\deg n_{ij} \leq v_i$. Hence $G = ND^{-1}$ is causal.

THEOREM 6. *Let P and Q be given $l \times p$ and $l \times m$ matrices over $F(z)$, let $\tilde{P} = [Q \ ; \ -P]$, and let $r = \text{rank } \tilde{P}$. Let S^* be a minimal basis for the space of $(p + m)$ -tuples orthogonal to \tilde{P} , hence of dimension $(p + m) \times (p + m - r)$, and let $S^{*T} = [D^{*T} \ ; \ N^{*T}]$, where D^* is $m \times (p + m - r)$ and N^* is $p \times (p + m - r)$. Then there exists a causal $p \times m$ matrix G solving $PG = Q$ if and only if $[D^*]_h$ has rank m .*

Proof. If $\text{rank } [D^*]_h = m$, then we may pick columns of S^* corresponding to m linearly independent columns of $[D^*]_h$ to obtain a $(p + m) \times m$ matrix $S = \begin{bmatrix} D \\ - \\ N \end{bmatrix}$

with $[D]_h$ full rank. Since $\mathbf{y} = S^* \mathbf{x}$ is polynomial only if \mathbf{x} is polynomial, $\mathbf{y}' = S \mathbf{x}'$ is polynomial only if \mathbf{x}' is polynomial, and the predictable degree property (condition 4(a) of the main theorem) holds for the columns of S since it holds for the columns of S^* . By the main theorem, condition 4, S is therefore a minimal basis for the space it spans. Consequently S satisfies the conditions of the lemma, and $G = ND^{-1}$ is the required solution.

Conversely, suppose $\text{rank } [D^*]_h < m$. This means that the integers $(1, \dots, m)$ are not all included among the pivot indices of S^* . Thus, since all minimal bases have the same pivot indices, $\text{rank } [D^*]_h < m$ for all minimal bases of $V_{\tilde{P}}^\perp$. Now if S were a set of $m(m + p)$ -tuples forming a minimal basis with $\text{rank } [D]_h = m$, then we could add $p - r$ more columns to S to form a minimal basis for S^* with $\text{rank } [D^*]_h = m$, a contradiction.

THEOREM 7. *Let $P, Q, \tilde{P}, S^*, D^*$ and N^* be as above, and assume $\text{rank } [D^*]_h = m$. Let the columns of S^* be ordered by degree, i.e., $v_1^* \leq v_2^* \leq \dots \leq v_{p+m-r}^*$, and let v_1, v_2, \dots, v_m be the indices of the first m linearly independent columns of $[D^*]_h$. Then*

1. *There exists a causal solution G to $PG = Q$ with these v_i as its controllability indices, and order $v = \sum_{i=1}^m v_i$.*
2. *This solution has least order among all solutions.*
3. *All least order solutions have the same set of controllability indices.*

4. A column of index μ of the minimal system matrix of any solution G is a linear combination of the columns of S^* of indices $v_i^* \leq \mu$.

Proof. 1. Select the corresponding columns as a minimal system matrix S of the desired solution G , as in the proof of Theorem 6. The controllability indices of G are the indices of S by Theorem 5.

4. A column of the minimal system matrix of any solution G is a polynomial $(p + m)$ -tuple orthogonal to \tilde{P} , hence in the space spanned by S^* . By condition 4 of the main theorem, a polynomial $(p + m)$ -tuple linearly dependent on a basis $(p + m)$ -tuple of degree v_i has degree $\geq v_i$.

2 and 3. From the definition of the v_i , the columns of $[D^*]_h$ corresponding to columns of S^* of degree less than v_i have rank less than i for any i . From condition 4 it follows that in any set of i linearly independent polynomial $(p + m)$ -tuples orthogonal to \tilde{P} , there is at least one $(p + m)$ -tuple of degree $\geq v_i$. Since the minimal system matrix of any solution G is a set of m linearly independent polynomial $(p + m)$ -tuples orthogonal to \tilde{P} , the conclusions follow.

COROLLARY. *If a causal solution to $PG = Q$ exists, there exists a solution with order not greater than the invariant dynamical order of the space $V_{\tilde{P}}$ generated by $\tilde{P} = [Q \mid -P]$.*

Proof. The order of S^* is the invariant dynamical order of $V_{\tilde{P}}$ by the corollary to Theorem 3. Any minimal system matrix S (if one exists) constructed from S^* as above can evidently have no greater order.

Example 1. Consider as in [5] the transfer function matrix

$$H = \frac{1}{s^2 + 3s + 2} \begin{bmatrix} s + 1 & s + 2 \\ s + 3 & s^2 + 2s \\ s^2 + 3 & 0 \end{bmatrix}.$$

We wish to find a proper left inverse of least order; i.e., a minimal system G such that $GH = I_2$ or $H^T G^T = I_2$. We thus have $P = [I_2 \mid -H^T]$. A basis for the dual

space is the 5×3 matrix $\begin{bmatrix} H^T \\ - \\ I_3 \end{bmatrix}$.³ Reduction to a minimal basis in echelon form by

the algorithm given earlier (admittedly tedious) yields

$$S^* = \begin{bmatrix} s + 2 & 1/2 & 3/2 \\ 1 & s - 1/2 & 3/2 \\ 1 & -1/2 & s + 3/2 \\ 1 & s + 1/2 & 1/2 \\ s + 2 & -1/2 & 1/2 \end{bmatrix}.$$

Since

$$[D^*]_h = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

³ Which is only to say that the input/output vectors of H^{-1} are the output/input vectors of H .

the first two columns of S^* may be taken as the minimal (controllability) system matrix S of the desired G^T , with controllability indices $v_1 = 1$, $v_2 = 1$. By the standard duality theorems, S^T is the minimal (observability) system matrix of the desired minimal order left inverse $G = H^{-1}$, with observability indices $v_1 = 1$, $v_2 = 1$. All left inverses of order 2 have the same set of observability indices, and in fact the set of all minimal order inverses corresponds to the set of all pairs of linearly independent 5-tuples of degree 1 in the space spanned by S^* , namely, all $S = S^*T$ where T is any 3×2 constant matrix of rank 2.

Example 2. The following example appears in Olson [10] and Moore and Silverman [11]. Let F be the binary field $GF(2)$, and let

$$H = \begin{bmatrix} z^{-1} + z^{-2} & 1 & 1 \\ 0 & z^{-1} & 1 \\ z^{-1} + z^{-2} & 1 & z/(1+z) \\ 1 & 1 + z^{-1} & 0 \end{bmatrix}.$$

Proceeding as above to find the minimal order left inverse of H , we find that a minimal basis S^* for the space spanned by $\begin{bmatrix} H^T \\ - \\ I_4 \end{bmatrix}$ is

$$S^* = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 + z^2 \\ 1 & z & 1 & 0 \\ 0 & 0 & 1 + z & 1 \\ 1 & z & 0 & 1 + z + z^2 \\ 0 & 0 & 1 + z & 1 + z^2 \\ 1 & 0 & 0 & 1 + z \end{bmatrix}.$$

Since

$$[D^*]_h = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

is nonsingular, H does have a causal left inverse; however, even though the indices of S^* are $(0, 1, 1, 2)$, it is not possible to find an inverse of order 2 because the third column is not a realizable input/output vector. There exists a family of inverses of order 3 with observability indices $(0, 1, 2)$. If we require minimal (observability) system matrices to be in echelon form, then all members of this family are characterized in terms of the transposes s_1, s_2, s_3 and s_4 of the columns of S^* as follows:

First row: s_1 ;

Second row: $s_2 + a_1 s_3$;

Third row: $a_2 s_3 + a_3(s_2 + z s_3) + s_4$;

where the coefficients a_i , $1 \leq i \leq 3$, are arbitrary elements of $F = GF(2)$; i.e., 0 or 1. Thus there are $2^3 = 8$ different minimal order inverses. It is straightforward to verify that the only feedforward inverse of order 3 (that is, an inverse with $|D| = z^3$) is obtained from the minimal system matrix $(s_1, s_2, s_2 + zs_3 + s_4)$. From a given minimal (observability) system matrix we can construct a realization in observer canonical form as in the previous section.

We note that, in view of the fact that the invariant dynamical order of the space generated by $[I_m | -H^T]$ (the space of input/output vectors of H , up to sign and transpose) is the order of H , the corollary to Theorem 7 provides yet another proof that if H has a causal inverse, it has such an inverse of order no greater than the order of H . That there exist $p \times m$ matrices H of all sizes and orders meeting this bound with equality is easily shown by examples of the following type:

$$H = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & z^{-n} \end{bmatrix}.$$

9. Application III. Convolutional codes. A rate k/n convolutional encoder G is a k -input, n -output, finite-dimensional, invertible linear system over some finite field, F , such as the binary field $GF(2)$. It generates a vector space V of *code words* over $F(z)$, namely, the set of all outputs $G\mathbf{u}$ as the inputs \mathbf{u} range over all rational k -tuples. The space V is called a *code*. Two encoders G_1 and G_2 that generate the same code V may be regarded as equivalent for coding purposes.

The reader is referred to [6] and its sequel [12] for a more complete discussion. Here we simply summarize the results of [6] in one theorem.

THEOREM 8. *Any convolutional code V can be generated by a "minimal encoder" G which is a minimal basis for V as a vector space over $F(D)$, where $D = z^{-1}$, and which therefore has the following properties:*

1. G is feedback-free and can be realized with v memory elements arranged as k shift registers of lengths v_i , $1 \leq i \leq k$, where v is the invariant dynamical order ("overall constraint length") of V and the v_i are the invariant dynamical indices ("i-th constraint lengths") of V .

2. If F is the finite field $GF(q)$ with q elements, the number of finite causal code words in V that stop at a time less than or equal to d is $q^{D(d)}$, where

$$D(d) = \sum_{i: d \geq v_i} (d - v_i).$$

3. G has a feedback-free inverse G^{-1} and a minimal dual basis G^\perp ("syndrome former") such that

$$\begin{bmatrix} G^{-1} \\ \vdots \\ G^\perp \end{bmatrix} G = \begin{bmatrix} I_k \\ \vdots \\ 0 \end{bmatrix}.$$

4. Any encoder G' that generates G requires at least v memory elements.

10. Conclusion. In this paper we have introduced the notion of minimal bases and developed some of their fundamental properties, which are simply properties of the corresponding rational vector spaces. While the applications we

have presented have not gone far beyond the results previously obtained by Popov, Rosenbrock, Wang, Wang and Davison, Wolovich, Eckberg, Dickinson and Morf, and the author, we believe we have obtained these results in a more unified and direct manner. Since the initial draft of this paper, Dickinson, Morf and Kailath [16] have found these ideas useful in developing efficient recursive solutions to the problem of realization of multivariable systems; Warren and Eckberg [17] have used some of these results in the study of controllability subspaces, which arise in the theory of decoupling; and Rosenbrock [18], [19] has extended Theorem 5 to nonproper systems and used these notions in the development of certain structural properties of multivariable systems. We hope that the tools developed here will be of continuing usefulness in linear system theory.

Appendix. Valuation theory and generalized minimal bases. The similarity in statement and proof of conditions 2(a) and 2(b), 3(a) and 3(b), and 4(a) and 4(b) of the main theorem suggest that a common formulation might exist. Such a formulation arises naturally in the language of valuation theory, which we shall introduce in this Appendix. (For further reference, see [13]–[15].) No valuation-theoretic results are used, but the language does suggest a generalization of our notion of a minimal basis for which a theorem generalizing the main theorem can be proved. More generally, it seems to us that this language is a natural one in which to frame certain discussions in systems theory; for instance, it lays to rest the question of whether discrete-time systems should be represented by z -transforms or D -transforms ($D = z^{-1}$), since it treats both on an equal footing and in the same breath.

A valuation $|\cdot|$ on a field K generalizes the notion of an absolute value on the real or complex field. It is a function from K into the nonnegative real numbers satisfying

- (i) $|k| \geq 0$, with equality if and only if $k = 0$;
- (ii) $|k_1 k_2| = |k_1| |k_2|$;
- (iii) (triangle inequality) $|k_1 + k_2| \leq |k_1| + |k_2|$.

A *non-Archimedean valuation* is one that satisfies

- (iii') (strong triangle inequality) $|k_1 + k_2| \leq \max[|k_1|, |k_2|]$.

An *exponential valuation* $v(k)$ is the negative logarithm of a non-Archimedean valuation. (Any base may be used, since $v(k)$ is essentially unaffected by multiplication by a scalar.) It is therefore a function from k into the real numbers (including ∞) that satisfies

- (i) $v(k) \leq \infty$, with equality if and only if $k = 0$;
- (ii) $v(k_1 k_2) = v(k_1) + v(k_2)$;
- (iii') $v(k_1 + k_2) \geq \min[v(k_1), v(k_2)]$.

The *trivial valuation* is the one such that $|k| = 0$ if $k = 0$ and $|k| = 1$ otherwise; in exponential form,

$$v(k) = \begin{cases} \infty, & k = 0, \\ 0, & k \neq 0. \end{cases}$$

For example, if k is the field of rational numbers, we can easily describe the set of all nontrivial valuations. First, we have the ordinary absolute value. Second, for each prime p , we have the p -adic valuations, defined as follows. Any nonzero

rational number can be written uniquely as $r = p^n a/b$, where a and b are relatively prime integers not divisible by p . (Here n may be negative or zero as well as positive.) Then the p -adic valuation $|r|_p$ is p^{-n} . Of course $|0|_p$ must be defined as 0. The reader may satisfy himself that (i)–(iii') hold, and hence all p -adic valuations are non-Archimedean. Their exponential form is $v_p(p^n a/b) = n$ (where we have taken logarithms to base p).

We are particularly concerned with the field $K = F(x)$ of rational functions in x over F . On this field the set of all interesting (trivial on F) valuations is as follows. First, let p be any monic irreducible polynomial. Any $k \neq 0 \in F(x)$ can be written uniquely as $p^e q_1/q_2$, where e is some integer, called the order of k at p , and q_1 and q_2 are polynomials not divisible by p . If we take $v_p(k) = \alpha e$, where α is any constant and e is the order of k at p , then it is straightforward to verify that $v_p(\cdot)$ satisfies (i)–(iii'). (Of course we take $v_p(0) = \infty$.) Conventionally the constant α is taken as the integer $\deg p$, in order to make the product formula hold (see below).

This gives an essentially different valuation for every $p \in \mathcal{P}$, where \mathcal{P} is the set of all irreducible polynomials in $F[x]$. There is one more. Express any $k \neq 0 \in F(x)$ as a ratio of polynomials q_1/q_2 ; then $v_{x^{-1}}(k)$ is defined as $\deg q_2 - \deg q_1$. Again verification of (i)–(iii') is straightforward. Why do we use the subscript x^{-1} ? Recall that a ratio of polynomials in x can equally well be expressed as a ratio of polynomials in x^{-1} ; i.e., $F(x) = F(x^{-1})$. If $p(x)$ is an irreducible polynomial other than x , then $p(x^{-1}) = p(x)x^{-\deg p}$ is an irreducible polynomial of the same degree in x^{-1} . The order of k at p is the same as the order of k at \tilde{p} . However the order of k at x is the difference between the denominator and numerator degrees of k as a ratio of polynomials in x^{-1} . (Example: $x^2 + x^3 = (1 + x^{-1})/x^{-3}$, $v_x(x^2 + x^3) = 2 = 3 - 1$.) Similarly, $v_{x^{-1}}(k)$ is the order of k at x^{-1} . Hence there is complete symmetry between x and x^{-1} , and we need not distinguish between $F(x)$ and $F(x^{-1})$.

To be more concrete, let $F = \mathbb{C}$, the field of complex numbers. The set \mathcal{P} of all irreducible polynomials is the set of all $p(x) = x - \alpha$, for every $\alpha \in \mathbb{C}$. Let $f(x) = n(x)/d(x)$ be a rational function in $\mathbb{C}(x)$, reduced to lowest terms. We write $v_\alpha(\cdot)$ for $v_{x-\alpha}(\cdot)$. If $v_\alpha[f(x)] > 0$, then $n(x)$ is divisible by $(x - \alpha)^{v_\alpha[f(x)]}$, and we say $f(x)$ has a zero of order $v_\alpha[f(x)]$ at α . Similarly, if $v_\alpha[f(x)] < 0$, we say $f(x)$ has a pole of order $-v_\alpha[f(x)]$ at α . Now $v_{x^{-1}}[f(x)] = \deg d(x) - \deg n(x)$ is often said to be the order of the pole of $f(x)$ at infinity, and we might write $v_\infty(\cdot)$ instead of $v_{x^{-1}}(\cdot)$. If $x = z$ and $f(z)$ is causal (i.e., $\deg d(z) \geq \deg n(z)$), it cannot have zeros at infinity; i.e., $v_{z^{-1}}[f] \geq 0$, or $v_D[f] \geq 0$ if f is expressed as a D -transform instead of a z -transform, where $D \triangleq z^{-1}$.

Let \mathcal{P}^* then be the set $\{x, x^{-1}, 1 + x, \dots\}$; i.e., \mathcal{P} plus x^{-1} . To each $p \in \mathcal{P}^*$ corresponds an essentially different valuation, and in valuation theory it is shown that these comprise all the essentially different valuations on $F(x)$ (that are trivial on F). All are non-Archimedean.

For each $p \in \mathcal{P}^*$, we may express any nonzero element $k \in F(x)$ as a formal Laurent series

$$k = \bar{a}_e p^e + \bar{a}_{e+1} p^{e+1} + \dots,$$

where e is the order of k at p , as follows. Express k as $p^e q_1/q_2$, and let \bar{q}_1 and \bar{q}_2 be the residues of q_1 and q_2 modulo p , considered as elements of the residue class field \bar{F}_p of polynomials modulo p ; then $\bar{a}_e = \bar{q}_1/\bar{q}_2$, considered as a polynomial of degree less than $\deg p$. (For $p = x^{-1}$, operate as though k were in $F(x^{-1})$.) It is easily verified that $k - \bar{a}_e p^e$ has order $e + 1$ or greater at p ; hence the procedure can be repeated indefinitely to give the possibly infinite formal Laurent series. (Example: $k = x^2 - x^3$; $v_x(k) = 2$, $\bar{a}_2 = 1$, $\bar{a}_3 = -1$, $\bar{a}_4 = 0, \dots$; $v_{x^{-1}}(k) = -3$, $\bar{a}_{-3} = -1$, $\bar{a}_{-2} = 1$, $\bar{a}_{-1} = 0, \dots$; $v_{x+1}(k) = 0$, $k = 2 - 5(x+1) + 4(x+1)^2 - (x+1)^3$.) We let $[k]_p$ stand for the residue class representative \bar{a}_e ; i.e.,

$$k = [k]_p p^e [1 + O(p)].$$

Finally, we have the *product formula* (which appears as a sum since we are using exponential valuations):

$$\sum_{p \in \mathcal{P}^*} v_p(k) = 0 \quad \text{if } k \neq 0.$$

When k is a polynomial, this merely says that the degree of k ($= -v_{x^{-1}}(k)$) is equal to the sum of the degrees of the irreducible factors of k . When k is a ratio of polynomials $k = q_1/q_2$, the relation $v_p(k) = v_p(q_1) - v_p(q_2)$, which holds for all $p \in \mathcal{P}^*$, proves the product formula since $\sum v_p(q_1) = 0$, $\sum v_p(q_2) = 0$.

We now proceed to norms. If V is a vector space over a field K with valuation $|\cdot|$, then a norm $\|\cdot\|$ is a function from V to the nonnegative real numbers that satisfies

- (i) $\|v\| \geq 0$, with equality if and only if $v = 0$;
- (ii) $\|kv\| = |k| \|v\|$ for $k \in K, v \in V$;
- (iii) $\|v_1 + v_2\| \leq \|v_1\| + \|v_2\|$;

or, in the non-Archimedean case,

- (iii') $\|v_1 + v_2\| \leq \max[\|v_1\|, \|v_2\|]$.

Similarly, an exponential norm is a function $v(\cdot)$ from V to the real numbers (including ∞) that satisfies

- (i) $v(v) \leq \infty$, equality iff $v = 0$;
- (ii) $v(kv) = v(k) + v(v)$;
- (iii') $v(v_1 + v_2) \geq \min[v(v_1), v(v_2)]$.

The standard norm corresponding to an exponential (non-Archimedean) valuation $v(\cdot)$ is the "box-norm", which is defined in terms of a basis for V as follows. Let e_1, \dots, e_n be a set of vectors that span V , with $n = \dim V$; then every $v \in V$ has the unique representation

$$v = \sum_{i=1}^n v_i e_i$$

for some n -tuple \mathbf{v} of elements of K . The "box-norm", defined by

$$v(v) = \min_i v(v_i),$$

is easily shown to satisfy (i)–(iii') for any valuation $v(\cdot)$ on K .

When $K = F(x)$ and V is the vector space of n -tuples $\mathbf{v} = (v_1, \dots, v_n)$, we take e_1, \dots, e_n as the unit n -tuples; if $p \in \mathcal{P}^*$, the norm $v_p(\mathbf{v})$ is defined as $v_p(\mathbf{v}) = \min_i v_p(v_i)$.

The residue $[v]_p$ of an n -tuple \mathbf{v} is the n -tuple of coefficients of $p^{v_p(\mathbf{v})}$ in the Laurent series expansions of the v_i . Note that if $v_p(v_i) > v_p(\mathbf{v})$, then $[v_i]_p = 0$. The product formula need no longer hold for norms; indeed,

$$\sum_{p \in \mathcal{P}^*} v_p(\mathbf{v}) \leq \sum_{p \in \mathcal{P}^*} v_p(v_1) = 0.$$

This observation inspires the one definition in this Appendix that is nonstandard. We define the *defect* of an n -tuple \mathbf{v} as

$$\text{def } \mathbf{v} \triangleq - \sum_{p \in \mathcal{P}^*} v_p(\mathbf{v}).$$

The defect is thus nonnegative. Further, for any $k \neq 0 \in F(x)$, we have

$$\text{def } k\mathbf{v} = - \sum_{p \in \mathcal{P}^*} v_p(k\mathbf{v}) = - \sum_{p \in \mathcal{P}^*} v_p(k) - \sum_{p \in \mathcal{P}^*} v_p(\mathbf{v}) = \text{def } \mathbf{v}.$$

Finally, if \mathbf{v} is a polynomial n -tuple with g.c.d. $\{v_i\} = 1$, then $\text{def } \mathbf{v} = \deg \mathbf{v}$ since $v_p(\mathbf{v}) = 0$ for $p \neq x^{-1}$, $v_{x^{-1}}(\mathbf{v}) = -\deg \mathbf{v}$; hence, concretely, the defect of an n -tuple is the degree of the polynomial n -tuple we get by multiplying \mathbf{v} by the least common multiple of its denominators and dividing by the g.c.d. of its numerators.

In general, for any exponential (non-Archimedean) norm on a vector space V , if $\mathbf{v} = \sum k_i \mathbf{g}_i$, then

$$(A.1) \quad v(\mathbf{v}) \geq \min_i v(k_i \mathbf{g}_i) = \min_i v(k_i) + v(\mathbf{g}_i).$$

A set of k vectors $\{\mathbf{g}_1, \dots, \mathbf{g}_k\}$ for which equality always holds in (A.1) is called *orthogonal* [13]. We shall call a set of n -tuples \mathbf{g}_i *p-orthogonal* if it is orthogonal for the norm $v_p(\cdot)$, and *globally orthogonal* if *p-orthogonal* for all $p \in \mathcal{P}^*$.

Now we are in a position to recast our results on minimal bases in this language.

DEFINITION 8. If G is a $k \times n$ rational matrix with rows \mathbf{g}_i , the *i-th generalized index* of G is defined as $v_i = \text{def } \mathbf{g}_i$, $1 \leq i \leq k$, and the *generalized order* of G is defined as $v = \sum v_i$.

DEFINITION 9. If V is a k -dimensional vector space of n -tuples over $F(x)$, a *generalized minimal basis* for V is a rational $k \times n$ matrix G such that G is a basis for V and has least generalized order among all bases for V .

DEFINITION 10. If G is a $k \times n$ rational matrix with rows \mathbf{g}_i and $p \in \mathcal{P}^*$, then $[G]_p$ is the $k \times n$ \bar{F}_p -matrix with rows $[\mathbf{g}_i]_p$.

GENERALIZED MAIN THEOREM. Let G be a $k \times n$ basis for a vector space V over $F(x)$ with generalized indices v_i , $1 \leq i \leq k$, and generalized order $v = \sum v_i$. Then the following statements are equivalent:

1. G is a generalized minimal basis for V .
2. $[G]_p$ has full rank over \bar{F}_p for all $p \in \mathcal{P}^*$.
3. Let $v_p(k \times k \text{ minors})$ be the minimum p -valuation among all the $k \times k$ minors of G ; then

$$v_p(k \times k \text{ minors}) = \sum_{i=1}^k v_p(\mathbf{g}_i) \quad \text{for all } p \in \mathcal{P}^*.$$

4. The k rows \mathbf{g}_i are globally orthogonal.

We leave it to the reader to verify that conditions 1–4 of the main theorem are the appropriate special cases of 1–4 here, and to supply a proof along the same lines

as there. We omit a generalization of statement 5 since the space S_d of all n -tuples \mathbf{v} of defect less than d is not particularly interesting; note that even if $\mathbf{s}_1, \mathbf{s}_2 \in S_d$, $\mathbf{s}_1 + \mathbf{s}_2$ may not be in S_d .

Acknowledgment. The stimulus of the interesting paper by S. H. Wang and E. J. Davison has already been acknowledged; for access to this paper and for other help I am indebted to L. M. Silverman. The hospitality and stimulating environment provided by Stanford during a year's visit contributed greatly to this work, particularly conversations with B. Dickinson, T. Kailath and B. F. Wyman. The connection to Popov's work and the importance of a unique minimal basis were explained to me by Dickinson and M. Morf at Stanford, and A. Eckberg at M.I.T. W. A. Wolovich offered several helpful comments on the algorithms. Finally, I owe my exposure to valuation theory to Prof. Wyman.

REFERENCES

- [1] H. H. ROSENBRICK, *State-Space and Multivariable Theory*, John Wiley, New York, 1970.
- [2] V. M. POPOV, *Some properties of control systems with irreducible matrix transfer functions*, Seminar on Differential Equations and Dynamical Systems, Lecture Notes in Mathematics No. 144, Springer, Berlin, 1969, pp. 169–180.
- [3] W. A. WOLOVICH, *The determination of state-space representations for linear multivariable systems*, Automatica, 9 (1973), pp. 97–106.
- [4] S. H. WANG, *Design of linear multivariable systems*, Memo. ERL-M309, Electronics Research Lab., University of California, Berkeley, 1971.
- [5] S. H. WANG AND E. J. DAVISON, *A minimization algorithm for the design of multivariable systems*, IEEE Trans. Automatic Control, AC-18 (1973), pp. 220–225.
- [6] G. D. FORNEY, JR., *Convolutional codes I: Algebraic structure*, IEEE Trans. Information Theory, IT-16 (1970), pp. 720–738. *Correction*, Ibid., IT-17 (1971), p. 360.
- [7] D. G. LUENBERGER, *Canonical forms for linear multivariable systems*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 290–293.
- [8] P. BRUNOVSKY, *A classification of linear controllable systems*, Kybernetika, 3 (1970), pp. 173–187.
- [9] V. M. POPOV, *Invariant description of linear, time-invariant controllable systems*, this Journal, 10 (1972), pp. 252–264.
- [10] R. R. OLSON, *Note on feedforward inverses for linear sequential circuits*, IEEE Trans. Computers, C-19 (1970), pp. 1216–1221.
- [11] B. C. MOORE AND L. M. SILVERMAN, *A new characterization of feedforward delay-free inverses*, IEEE Trans. Information Theory, IT-19 (1973), pp. 126–129.
- [12] G. D. FORNEY, JR., *Structural analysis of convolutional codes via dual codes*, Ibid., IT-19 (1973).
- [13] A. F. MONNA, *Analyse Non-Archimédienne*, Springer, Berlin, 1970.
- [14] G. BACHMAN, *Introduction to p -adic Numbers and Valuation Theory*, Academic Press, New York, 1964.
- [15] L. NARICI, E. BECKENSTEIN AND G. BACHMAN, *Functional Analysis and Valuation Theory*, M. Dekker, New York, 1971.
- [16] B. W. DICKINSON, M. MORF AND T. KAILATH, *A minimal realization algorithm for matrix sequences*, IEEE Trans. Automatic Control, AC-19 (1974), pp. 31–38.
- [17] M. E. WARREN AND A. E. ECKBERG, JR., *On the dimensions of controllability subspaces: A characterization via polynomial matrices and Kronecker invariants*, Rep. ESL-R-512, Electronic Systems Laboratory, M.I.T., Cambridge, Mass., 1973.
- [18] H. H. ROSENBRICK AND G. E. HAYTON, *Dynamical indices of a transfer function matrix*, Rep. 219, Control Systems Centre, U.M.I.S.T., Manchester, England, 1973.
- [19] H. H. ROSENBRICK, *Structural properties of first-order dynamical systems*, Rep. 220, Control Systems Centre, U.M.I.S.T., Manchester, England, 1973.

COMBINED PRIMAL-DUAL AND PENALTY METHODS FOR CONSTRAINED MINIMIZATION*

DIMITRI P. BERTSEKAS†

Abstract. In this paper we consider a class of combined primal-dual and penalty methods often called methods of multipliers. The analysis focuses mainly on the rate of convergence of these methods. It is shown that this rate is considerably more favorable than the corresponding rate for penalty function methods. Some efficient versions of multiplier methods are also considered whereby the intermediate unconstrained minimizations involved are approximate and only asymptotically exact. It is shown that such approximation schemes may lead to a substantial deterioration of the convergence rate, and a special approximation scheme is proposed which exhibits the same rate as the method with exact minimization. Finally, we analyze the properties of the step size rule of the multiplier method in relation to other possible step sizes, and we consider a modified step size rule for the case of the convex programming problem.

1. Introduction. During recent years, penalty function methods (see, e.g., [7]) have gained recognition as one of the most effective class of methods for solving constrained minimization problems. Characteristic of such methods is that they require the solution of a sequence of unconstrained minimizations of the objective function of the problem to which an increasingly high penalty term is added. It is well known that these unconstrained minimization problems have increasingly unfavorable structure due to ill-conditioning [7], [17], a fact which often leads to slow convergence despite the use of efficient unconstrained minimization algorithms. Another important class of methods for constrained minimization is the so-called class of primal-dual methods (see, e.g., [17]). Such methods are, in effect, iterative ascent algorithms for solving the dual problem (defined under suitable local convexity assumptions [17]). Similar to penalty methods, they involve the solution of a sequence of unconstrained minimizations of a Lagrangian function, each of which yields the value and the gradient of the dual functional at the current value of the Lagrange multiplier. At the end of each minimization, the Lagrange multiplier is updated by means of an ascent iteration. Primal-dual methods are known to have serious disadvantages. First, the problem must have a locally convex structure in order for the dual functional to be defined. Second, it is usually necessary to solve a large number of unconstrained minimization problems since the ascent iteration converges only moderately fast. Thus primal-dual methods have found application only in the limited class of problems where the unconstrained minimizations can be carried out very efficiently due to special structure.

In the last few years, a number of researchers have proposed a new class of methods, often called methods of multipliers, in which the penalty idea is merged with the primal-dual philosophy. In these methods, the penalty term is added not to the objective function but rather to the Lagrangian function which is ordinarily

* Received by the editors February 12, 1973, and in final revised form March 8, 1974.

† This work was carried out at the Department of Engineering-Economic Systems, Stanford University, Stanford, California, and supported by the National Science Foundation under Grant GK-29237. Now at Department of Electrical Engineering, University of Illinois, Urbana, Illinois 61801.

minimized in primal-dual methods. Again, a sequence of unconstrained minimization problems is solved; however, each minimization is followed by an ascent iteration on the Lagrange multiplier which is aimed at solving the dual problem. In contrast with penalty methods, the penalty term need not be increased to infinity, thus avoiding the associated extreme ill-conditioning. In addition, the ascent iteration converges fast, thus necessitating only a few unconstrained minimization cycles. By moderating the disadvantages of both penalty and primal-dual methods, multiplier methods have emerged as a very attractive class of algorithms for constrained minimization, a fact substantiated by the limited computational experience presently available. This paper provides an analysis of some aspects of these methods mostly related to their convergence rate and their efficient implementation.

The methods that we consider were initially proposed by Hestenes [12] and Powell [32], and somewhat later by Haarhoff and Buys [11]. Hestenes gave no interpretation or convergence proof of his method of multipliers, and Powell was motivated by a penalty function viewpoint. The primal-dual interpretation was given later by Luenberger [17], who in addition gave an argument indicating the fast convergence of the method, and by Buys [6], who in his recent thesis provided an extensive and well written analysis of multiplier methods. Buys [6] also proved local convergence of the method of multipliers both for the case of exact and approximate unconstrained minimization under the assumption that the penalty parameter is constant but sufficiently large. A similar convergence result for exact minimization was proved by Rupp [41], [42]. Global convergence results for nonconvex problems were proved recently by the author in [3] and [5]. For quadratic problems with linear constraints, global convergence was also proved by Martensson [19] who in addition proposed some variations on the multiplier method. The method of multipliers has been applied to the solution of some infinite-dimensional problems by Rupp [39], [40], [41]. Some variations of the method of multipliers were proposed by Miele, et al. [20], [21], and Tripathi and Narendra [44]. In these particular variations, the Lagrange multiplier is updated at the end of every gradient step or every conjugate gradient cycle in the unconstrained minimization problem. The convergence properties and the precise motivation for such methods is not as yet quite well understood. They seem to be somewhat related to multiplier methods with asymptotically exact unconstrained minimization, as will be explained later on in this paper. They are also related to the Lagrangian algorithms of Arrow, Hurwicz and Uzawa [2] (particularly the chapter by Arrow and Solow) as applied to the "penalized" problem (5) of the next section. Finally we note that multiplier methods as proposed in the above references are mainly applicable to problems with equality constraints. More recently, considerable attention has been directed towards extension of the method to treat inequality constraints. At the same time, the properties of the method when applied to convex programming problems have been analyzed in detail. In this connection, we mention the excellent papers by Rockafellar [34]–[37], the groundwork for which was laid in his early paper [33], and the work of B. Kort and the author [14]–[16], [4]. Generally speaking, methods of multipliers, as adapted to treat inequality constraints, exhibit similar behavior as for the case of equality constraints. However, for convex programming problems, the methods have some

very attractive properties, namely that they converge globally for any positive value of the penalty parameter [14], [35], [15], [4]. We mention also that there is a very interesting duality theory associated with multiplier methods primarily developed by Rockafellar [33], [34], [36] (see also [1], [3], [15], [18], [28]). Aside from its intrinsic value, this theory can form the basis for the development of efficient large-step Lagrangian methods. For one such algorithm based on Newton's method, see Mangasarian [18]. We note that Lagrangian methods utilizing the penalty idea have been proposed by Fletcher [8], [10], Fletcher and Lill [9], and by Miele and his associates [22], [23]. The precise connection of these methods with methods of multipliers is as yet unclear. Finally we mention that some work related to the method of multipliers has been reported recently in [26] and [43].

The present paper is organized as follows. In the next section we describe the basic method of multipliers in a framework which is suitable for analysis of its convergence rate. Subsequently in § 3 we obtain a useful expression for the rate of convergence of the method. It is shown in particular that as the penalty parameter is increased, the rate of convergence of the dual iteration approaches a superlinear rate. Furthermore, it is shown that we can expect multiplier methods to converge considerably faster than penalty methods which are operated sequentially. In § 4 we consider some efficient variants of the method of multipliers whereby the unconstrained minimizations are only asymptotically exact. We show that such approximate minimizations may lead, in general, to a substantial deterioration of the convergence rate, and we propose a particular approximation scheme which exhibits the same asymptotic convergence rate as the method with exact minimization. In § 5 we compare the step size of the multiplier method with other possible step size rules. We show that for certain problems which are not locally convex, the multiplier method step size is nearly optimal. For locally convex problems, we explain that this is not necessarily true and we propose an alternative step size rule which exhibits an improved convergence rate over the ordinary method. Finally in § 6 we present results of numerical experiments which generally support the conclusions of the theoretical analysis.

2. The method of multipliers. Consider the following constrained minimization problem:

$$(1) \quad \text{minimize } f(x) \quad \text{subject to } h(x) = 0,$$

where $f: R^n \rightarrow R$ is a given twice continuously differentiable function and $h: R^n \rightarrow R^m, m \leq n$, is a given twice continuously differentiable mapping.

Let x^* be an optimal solution of problem (1). We shall assume that x^* satisfies the second order sufficiency conditions for an isolated local minimum, i.e., the matrix $\nabla^2 h(x^*)$ has full rank and there exists a unique Lagrange multiplier (row) vector λ^* such that

$$(2) \quad \nabla l(x^*, \lambda^*) = \nabla f(x^*) + \lambda^* \nabla h(x^*) = 0$$

and

$$(3) \quad y' L(x^*, \lambda^*) y > 0$$

for all $y \in R^n$ such that $\nabla h(x^*)y = 0, y \neq 0$. In the above relations, $\nabla l(x^*, \lambda^*)$ and $L(x^*, \lambda^*)$ denote the gradient relative to x and the Hessian matrix relative to x , respectively, of the Lagrangian function

$$(4) \quad l(x, \lambda) = f(x) + \lambda h(x)$$

evaluated at (x^*, λ^*) . The $m \times n$ matrix $\nabla h(x)$ denotes the matrix having as rows the gradients $\nabla h_i(x), i = 1, \dots, m$, and a prime denotes transposition.

It is clear that problem (1) is equivalent to the following problem obtained from problem (1) by adding a penalty term to the objective function:

$$(5) \quad \text{minimize } f(x) + \frac{1}{2}c\|h(x)\|^2 \quad \text{subject to } h(x) = 0,$$

where c is a positive scalar.

Consider now the Lagrangian function corresponding to problem (5):

$$(6) \quad l(x, \lambda, c) = f(x) + \lambda h(x) + \frac{1}{2}c\|h(x)\|^2,$$

and its Hessian evaluated at (x^*, λ^*) :

$$(7) \quad L(x^*, \lambda^*, c) = L(x^*, \lambda^*) + c\nabla h(x^*)'\nabla h(x^*).$$

It follows from (3) that

$$(8) \quad y'L(x^*, \lambda^*, c)y > 0 \quad \forall y \in R^n, \quad y \neq 0$$

if $c \geq c^* > 0$, where c^* is sufficiently large to guarantee that the matrix $L(x^*, \lambda^*, c^*)$ is positive definite. As a result, for every c with $c \geq c^*$, problem (5) has locally convex structure according to the definition of [17], and thus we can define for each $c \geq c^*$ the dual functional

$$g_c(\lambda) = \min_x l(x, \lambda, c).$$

In the above equation, the dual functional $g_c(\lambda)$ is defined in a neighborhood of λ^* and the minimization is understood to be local in a neighborhood of x^* . The implicit function theorem and our assumptions guarantee that such neighborhoods exist for every $c \geq c^*$. Since, however, in the algorithm which we shall describe, the scalar c may vary from one iteration to the next, it is necessary to provide a uniform definition of the dual functional over neighborhoods which do not depend on c . We shall restrict, however, the scalar c to take values in an interval $[c^*, \bar{c}]$, where \bar{c} is an arbitrarily large constant. For practical purposes, this restriction results in no great loss of generality.

For any element z of a finite-dimensional space with the usual Euclidean norm and for any scalar $s > 0$, we denote by $B(z; s)$ the open ball centered at z and having radius s . We denote by $\bar{B}(z; s)$ the corresponding closed ball. We now have the following proposition.

PROPOSITION 1. *There exist positive scalars ε^* and δ^* such that for all $\lambda \in B(\lambda^*; \delta^*)$ and all $c \in [c^*, \bar{c}]$, the problem*

$$\text{minimize } l(x, \lambda, c) = f(x) + \lambda h(x) + \frac{1}{2}c\|h(x)\|^2 \quad \text{subject to } x \in B(x^*; \varepsilon^*)$$

has a unique solution $x(\lambda, c)$. Furthermore, for every ε with $0 < \varepsilon \leq \varepsilon^$, there exists a δ with $0 < \delta \leq \delta^*$ such that*

$$x(\lambda, c) \in B(x^*; \varepsilon) \quad \forall \lambda \in B(\lambda^*; \delta), \quad c \in [c^*, \bar{c}].$$

Proof. The proof is based on a fixed-point argument similar to one used for the proof of the implicit function theorem (see, e.g., [13], [25]).

For each $\lambda \in R^m$ and $c \in [c^*, \bar{c}]$, consider the mapping $Q^{\lambda, c}: R^n \rightarrow R^n$ defined by

$$Q^{\lambda, c}(x) = x - [L(x^*, \lambda^*, c)]^{-1} \nabla l(x, \lambda, c),$$

where L , and ∇l denote the Hessian and gradient of the augmented Lagrangian given by (6) and (7). Taking the gradient of $Q^{\lambda, c}$ with respect to x , we have

$$\nabla Q^{\lambda, c}(x) = [L(x^*, \lambda^*, c)]^{-1} [L(x^*, \lambda^*, c) - L(x, \lambda, c)]$$

and

$$\|\nabla Q^{\lambda, c}(x)\| \leq \| [L(x^*, \lambda^*, c)]^{-1} \| \|L(x^*, \lambda^*, c) - L(x, \lambda, c)\|.$$

Now given any $a \in (0, 1)$, there exist an $\varepsilon > 0$ and $\delta > 0$ such that $\|\nabla Q^{\lambda, c}(x)\| \leq a < 1$ for $x \in \bar{B}(x^*; \varepsilon)$, $\lambda \in \bar{B}(\lambda^*; \delta)$, $c \in [c^*, \bar{c}]$.

On the other hand, we have

$$\begin{aligned} \|Q^{\lambda, c}(x^*) - x^*\| &\leq \| [L(x^*, \lambda^*, c)]^{-1} \| \|\nabla l(x^*, \lambda, c)\| \\ &= \| [L(x^*, \lambda^*, c)]^{-1} \| \|\nabla l(x^*, \lambda, c) - \nabla l(x^*, \lambda^*, c)\|, \end{aligned}$$

and by letting δ be sufficiently small, we can assert that

$$\|Q^{\lambda, c}(x^*) - x^*\| \leq \varepsilon(1 - a) \quad \forall c \in [c^*, \bar{c}].$$

Now we have

$$\begin{aligned} \|Q^{\lambda, c}(x) - x^*\| &\leq \|Q^{\lambda, c}(x^*) - x^*\| + \|Q^{\lambda, c}(x) - Q^{\lambda, c}(x^*)\| \\ &\leq \varepsilon(1 - a) + \sup_{0 < t \leq 1} \|\nabla Q^{\lambda, c}[x^* + t(x - x^*)]\| \|x - x^*\| \leq \varepsilon(1 - a) + a\varepsilon = \varepsilon \end{aligned}$$

for all $x \in \bar{B}(x^*; \varepsilon)$, $\lambda \in \bar{B}(\lambda^*; \delta)$, $c \in [c^*, \bar{c}]$.

Thus we have $Q^{\lambda, c}: \bar{B}(x^*; \varepsilon) \rightarrow \bar{B}(x^*; \varepsilon)$ and

$$\|\nabla Q^{\lambda, c}(x)\| \leq a < 1 \quad \text{for each } \lambda \in \bar{B}(\lambda^*; \delta), \quad c \in [c^*, \bar{c}].$$

Hence $Q^{\lambda, c}$ has a unique fixed point $x(\lambda, c)$, i.e.,

$$x(\lambda, c) = Q^{\lambda, c}[x(\lambda, c)] = x(\lambda, c) - [L(x^*, \lambda^*, c)]^{-1} \nabla l[x(\lambda, c), \lambda, c],$$

from which

$$\nabla l[x(\lambda, c), \lambda, c] = 0.$$

If we take, in addition, ε and δ sufficiently small so that $L(x, \lambda, c)$ is positive definite for all $(x, \lambda) \in \bar{B}(x^*; \varepsilon) \times \bar{B}(\lambda^*; \delta)$ and $c \in [c^*, \bar{c}]$, we have that the corresponding $x(\lambda, c)$ is a unique unconstrained minimum of $l(x, \lambda, c)$ within $B(x^*; \varepsilon^*)$ for ε^* sufficiently small, and the result of the proposition follows easily. Q.E.D.

The following corollary is an easy consequence of Proposition 1.

COROLLARY 1.1. *Let L be such that*

$$\|h(x) - h(y)\| \leq L\|x - y\| \quad \forall x, y \in \bar{B}(x^*; \varepsilon^*),$$

and ε^*, δ^* be as in Proposition 1. Then for every ε with $0 < \varepsilon \leq \varepsilon^*$, there exists a δ with $0 < \delta \leq \delta^*$ such that

$$x(\lambda, c) \in B(x^*; \varepsilon), \quad \lambda + ch[x(\lambda, c)] \in B(\lambda^*; \delta + \bar{c}L\varepsilon)$$

for all $\lambda \in B(\lambda^*; \delta)$, $c \in [c^*, \bar{c}]$.

Proof. If δ corresponds to ε as in Proposition 1, then $x(\lambda, c) \in B(x^*; \varepsilon)$ and

$$\begin{aligned} \|\lambda + ch[x(\lambda, c)] - \lambda^*\| &\leq \|\lambda - \lambda^*\| + \bar{c}\|h[x(\lambda, c)]\| \\ &= \|\lambda - \lambda^*\| + \bar{c}\|h[x(\lambda, c)] - h(x^*)\| < \delta + \bar{c}L\varepsilon. \quad \text{Q.E.D.} \end{aligned}$$

Proposition 1 essentially says that by locally minimizing the augmented Lagrangian, one obtains points which are arbitrarily and uniformly close to x^* provided λ is sufficiently close to λ^* . Furthermore, the proposition provides a means for defining the dual functional over a domain which is common for all $c \in [c^*, \bar{c}]$. We define, for all $\lambda \in B(\lambda^*; \delta^*)$ and all $c \in [c^*, \bar{c}]$, the dual functional as

$$(9) \quad g_c(\lambda) = \min_{x \in B(x^*; \varepsilon^*)} \{f(x) + \lambda h(x) + \tfrac{1}{2}c\|h(x)\|^2\},$$

where the minimum over the open ball $B(x^*; \varepsilon^*)$ is attained by Proposition 1. It can be easily shown (see also [6], [17]) that the scalars ε^* and δ^* in Proposition 1 and Corollary 1.1 can be chosen so that the dual functional $g_c(\lambda)$ is twice continuously differentiable and concave in $B(\lambda^*; \delta^*)$ for all $c \in [c^*, \bar{c}]$. We shall assume that ε^* and δ^* have been so chosen. The gradient ∇g_c and Hessian matrix G_c are given by

$$(10) \quad \nabla g_c(\lambda) = h[x(\lambda, c)]',$$

$$(11) \quad G_c(\lambda) = -\nabla h[x(\lambda, c)]\{L[x(\lambda, c), \lambda, c]\}^{-1}\nabla h[x(\lambda, c)]'.$$

Furthermore, the dual functionals $g_c(\lambda)$, $c \in [c^*, \bar{c}]$ have a common maximizing point, the Lagrange multiplier λ^* , and a common optimal value f^* which is equal to the optimal value $f(x^*)$ of problem (1):

$$g_c(\lambda^*) = \max_{\lambda} g_c(\lambda) = f^* \quad \forall c \in [c^*, \bar{c}].$$

The method of multipliers is simply a gradient method for maximizing the dual functional by means of the iteration

$$(12) \quad \lambda_{k+1} = \lambda_k + c\nabla g_c(\lambda_k).$$

The gradient $\nabla g_c(\lambda_k)$ is given by (10), where $x(\lambda_k, c)$ is an unconstrained minimum (within $B(x^*; \varepsilon^*)$) of the augmented Lagrangian

$$(13) \quad l(x, \lambda_k, c) = f(x) + \lambda_k h(x) + \tfrac{1}{2}c\|h(x)\|^2.$$

The iteration (12) is a fixed step size gradient method for solving the dual problem which can be shown [6], [41] to converge to λ^* provided the constant c is sufficiently large. This fact will also be proved in the next section in a more general setting where c may vary from one iteration to the next. It should be noted that in order for the method to converge, it is not necessary that the initial Lagrange multiplier estimate λ_0 is in $B(\lambda^*; \delta^*)$. Since the method can also be viewed as a penalty function method, it can be shown [3] that if the initial penalty parameter c is sufficiently large and the corresponding minimization problem yields a solution close to x^* , then the next point λ_1 will be arbitrarily close to λ^* . Thus in the initial

iterations, the penalty nature of the method is dominant and provides points sufficiently close to λ^* , and in subsequent iterations, the gradient nature of the algorithm becomes more pronounced.

It is important to realize that it is not necessary to keep the penalty parameter c fixed during the computation. Each constant c defines a dual functional $g_c(\lambda)$ via (9). The collection of all these dual functionals has the same local maximum λ^* . Thus when a different c (say c_k) is used at the k th unconstrained minimization,

$$(14) \quad \text{minimize } f(x) + \lambda_k h(x) + \frac{1}{2} c_k \|h(x)\|^2,$$

the iteration

$$(15) \quad \lambda_{k+1} = \lambda_k + c_k h[x(\lambda_k, c_k)]'$$

can be viewed as a gradient step for maximizing the corresponding dual functional $g_{c_k}(\lambda)$, which attains its maximum at λ^* . Furthermore, it is possible to let the sequence c_k increase to infinity. While the intermediate unconstrained minimization problems become increasingly ill-conditioned, the dual iteration (15) has increasingly faster convergence rate, as will be shown in the next section, and on balance, the method performs well. A reasonable method to update c suggested by Powell [32] and Buys [6], is to multiply c by a constant greater than one (say 5–10) at the end of each unconstrained minimization for which the resulting constraint violation as measured by $\|h(x)\|$ is not decreased by a certain factor. An alternative method for updating c has been suggested by Miele, et al. [20] in a somewhat different setting.

The method of multipliers can be easily extended to handle inequality constraints. As shown by Rockafellar [33]–[36], one may use slack variables to convert inequality constraints into equality constraints. However, the minimization with respect to the slack variables can be carried out explicitly, and as a result, the dimension of the unconstrained minimization problem is not increased. We do not further discuss inequality constraints in this paper, and we refer to [3], [15] and [16] for a discussion of the related rate of convergence aspects. Among other things, one may show that the approximate Lagrange multipliers corresponding to inactive constraints converge to zero in a finite number of steps. As a result, inactive constraints do not enter in any rate of convergence estimates, and the results of this paper under a strict complementarity assumption carry over to the inequality case in a straightforward manner.

We mention finally that the method of multipliers has an economic interpretation similar to the one given by Arrow and Solow [2] for their combined Lagrangian and penalty method. In this interpretation, the iterations of λ_k are viewed as market price adjustments to excess demand or supply, and the iterations of x_k are viewed as production vector changes in response to extrapolated market price changes.

3. Convergence rate of the method of multipliers. As mentioned in the previous section, the method of multipliers can be viewed as a gradient method for solving the dual problem. Thus one can obtain its convergence rate by using a corresponding result on gradient methods (see, e.g., [29], [30]). This result, however, is rather

uninformative, since it involves the eigenvalues of the Hessian $G_c(\lambda)$, which strongly depend on c . The following proposition is obtained by a modification of this result and provides an expression for the convergence rate which is more amenable to proper interpretation.

Let us consider the matrix

$$(16) \quad D(x, \lambda) = -\nabla h(x)[L(x, \lambda)]^{-1}\nabla h(x)',$$

where $L(x, \lambda)$ is the Hessian relative to x of the Lagrangian (4). Notice that $D(x^*, \lambda^*)$ would be the Hessian at λ^* of the ordinary dual functional g_0 if the problem had a locally convex structure [17]. Assume that $D(x, \lambda)$ is defined and is invertible in a set $\bar{B}(x^*; \varepsilon) \times \bar{B}(\lambda^*; \delta + \bar{c}L\varepsilon)$, where ε and δ are positive scalars such that $x(\lambda, c) \in B(x^*; \varepsilon)$, $\lambda + ch[x(\lambda, c)] \in B(\lambda^*; \delta + \bar{c}L\varepsilon)$ for all $\lambda \in B(\lambda^*; \delta)$ and all $c \in [c^*, \bar{c}]$ in accordance with Proposition 1 and Corollary 1.1. Assume also that the algorithm of (14), (15) yields a sequence of vectors (x_k, λ_k) converging to (x^*, λ^*) and that after some index \bar{k} , the vectors (x_k, λ_k) are contained in $B(x^*; \varepsilon) \times B(\lambda^*; \delta)$. Then we have the following proposition.

PROPOSITION 2. *Under the preceding assumptions, we have*

$$(17) \quad \|\lambda_{k+1} - \lambda^*\| \leq r_k \|\lambda_k - \lambda^*\| \quad \forall k \geq \bar{k},$$

with

$$(18) \quad r_k = \max_{\substack{(x, \lambda) \in \bar{B}(x^*; \varepsilon) \times \bar{B}(\lambda^*; \delta + \bar{c}L\varepsilon) \\ i=1, \dots, m}} \left| \frac{1}{1 - c_k e_i[D(x, \lambda)]} \right|,$$

where $e_i[D(x, \lambda)]$ denotes the i -th eigenvalue of $D(x, \lambda)$.

Proof. Consider the Hessian matrix (11). We have

$$G_c(\lambda) = -\nabla h[x(\lambda, c)]\{L[x(\lambda, c), \bar{\lambda}] + c\nabla h[x(\lambda, c)]'\nabla h[x(\lambda, c)]\}^{-1}\nabla h[x(\lambda, c)],$$

where $\bar{\lambda} = \lambda + ch[x(\lambda, c)]'$. From a well-known matrix identity, we have

$$[I - cD[x(\lambda, c), \bar{\lambda}]]^{-1} = I + cG_c(\lambda),$$

and hence for the corresponding eigenvalues of $G_c(\lambda)$ and $D[x(\lambda, c), \bar{\lambda}]$, we have

$$(19) \quad \frac{1}{1 - ce_i[D[x(\lambda, c), \bar{\lambda}]]} = 1 + ce_i[G_c(\lambda)].$$

Now by using the iteration (15), we have

$$\|\lambda_{k+1} - \lambda^*\| = \|\lambda_k - \lambda^* + c_k h(x_k)'\| = \left\| \lambda_k - \lambda^* + c_k \int_0^1 (\lambda_k - \lambda^*) G_{c_k}(\lambda) dt \right\|,$$

where $\lambda = \lambda^* + t(\lambda_k - \lambda^*)$. Hence

$$\begin{aligned} \|\lambda_{k+1} - \lambda^*\| &\leq \|\lambda_k - \lambda^*\| \left\| \int_0^1 [I + c_k G_{c_k}(\lambda)] dt \right\| \\ &\leq \|\lambda_k - \lambda^*\| \max_{\substack{\lambda = \lambda^* + t(\lambda_k - \lambda^*) \\ t \in [0, 1] \\ i=1, \dots, m}} |1 + c_k e_i[G_{c_k}(\lambda)]|. \end{aligned}$$

By using (19) and Corollary 1.1, it follows that

$$\begin{aligned} \|\lambda_{k+1} - \lambda^*\| &\leq \|\lambda_k - \lambda^*\| \max_{\substack{(x, \lambda) \in \bar{B}(x^*; \varepsilon) \times \bar{B}(\lambda^*; \delta + \bar{c}L\varepsilon) \\ i=1, \dots, m}} \left| \frac{1}{1 - c_k e_i[D(x, \lambda)]} \right| \\ &= r_k \|\lambda_k - \lambda^*\|. \end{aligned} \quad \text{Q.E.D.}$$

Some important observations can be made from the result of Proposition 2. First of all, a trivial modification of its proof yields the following local convergence result.

COROLLARY 2.1. *Let ε and δ be positive scalars such that $x(\lambda, c) \in B(x^*; \varepsilon)$ and $\lambda + ch[x(\lambda, c)]' \in B(\lambda^*; \delta + \bar{c}L\varepsilon)$ for all $\lambda \in B(\lambda^*; \delta)$ and $c \in [c^*, \bar{c}]$ in accordance with Proposition 1 and Corollary 1.1. Assume that ε and δ are sufficiently small and c_k is sufficiently large so that for some constant μ ,*

$$(20) \quad c_k \geq \mu > \max \left\{ 0, \frac{2}{e_i[D(x, \lambda)]} \right\} \quad \forall k > 0$$

for all eigenvalues $e_i[D(x, \lambda)]$ of $D(x, \lambda)$ over $\bar{B}(x^; \varepsilon) \times \bar{B}(\lambda^*; \delta + \bar{c}L\varepsilon)$. Assume also that $\lambda_0 \in B(\lambda^*; \delta)$. Then the sequence $\{\lambda_k\}$ generated by the iteration (15) remains in $B(\lambda^*; \delta)$ and converges to λ^* .*

Proof. By exact repetition of the argument in the proof of Proposition 1, we have $\|\lambda_1 - \lambda^*\| \leq r_0 \|\lambda_0 - \lambda^*\|$, where r_0 is given by (18). By our assumption (20),

$$r_0 < \max \left| \frac{1}{1 - \mu e_i[D(x, \lambda)]} \right| = \rho < 1.$$

Hence $\|\lambda_1 - \lambda^*\| \leq \rho \|\lambda_0 - \lambda^*\|$ and $\lambda_1 \in B(\lambda^*; \delta)$. Proceeding similarly, we prove for all k that $\|\lambda_k - \lambda^*\| \leq \rho^k \|\lambda_0 - \lambda^*\|$, and the result follows immediately. Q.E.D.

The most important observation from Proposition 2 is that the sequence $\|\lambda_k - \lambda^*\|$ converges linearly with stepwise convergence ratio r_k . Furthermore, r_k decreases to zero as c_k is increased. Thus a superlinear rate is approached as c_k tends to infinity. This is consistent with the argument of Luenberger [17], who observed that as c_k increases, the gradient iteration (15) approaches a Newton step for solving the dual. If the sequence c_k converges to a finite value c , then we have

$$(21) \quad \limsup_{k \rightarrow \infty} \frac{\|\lambda_{k+1} - \lambda^*\|}{\|\lambda_k - \lambda^*\|} \leq \max_i \left| \frac{1}{1 - c e_i[D(x^*, \lambda^*)]} \right| = \bar{r}.$$

It can easily be shown that at least for quadratic problems with linear constraints and a constant sequence c_k , \bar{r} is a sharp bound in the sense that there exist starting points λ_0 for which (21) holds with equality.

It is of interest to compare the convergence rate of the multiplier method with the convergence rate of penalty function methods which are based on sequential unconstrained minimization of the function

$$(22) \quad f(x) + \frac{1}{2} c_k \|h(x)\|^2$$

for a sequence $c_k \rightarrow \infty$. If the sequence $\{x_k\}$ of minimizing points of (22) converges to the point x^* , then the sequence $\{\lambda_k\}$, where $\lambda_k = c_k h(x_k)'$, converges to λ^* . It has

been shown [31], [24] that such penalty function methods generally exhibit a convergence rate governed by the relation

$$(23) \quad \|\lambda_k - \lambda^*\| \leq q/c_k \quad \forall k > \bar{k},$$

where \bar{k} is some index and q is a constant depending on the problem. By comparing (17), (18) and (23), it can be seen that the sequence $\{\lambda_k\}$ can be expected to converge considerably faster in the multiplier method than in the quadratic penalty function method. This fact has been substantiated by numerical experiments. Given that the two methods involve a comparable amount of computation at each unconstrained minimization and share the advantage of simplicity, it appears that multiplier methods should be generally considered preferable to penalty function methods. For further elaboration on the comparison between penalty methods and multiplier methods we refer to [3] and [5].

4. Efficient implementations of the multiplier method. The multiplier method described in the previous section has the drawback that the unconstrained minimization of the augmented Lagrangian must be carried out exactly in order to update the Lagrange multiplier via the gradient iteration (15). This requires an unreasonably high amount of computation for the unconstrained minimizations. It appears that a more efficient scheme results if only moderate accuracy is demanded in the initial minimizations, and the accuracy is increased at later iterations. Such a procedure has been suggested by Buys [6] in a similar vein as in corresponding penalty function methods [24], [27], [31]. In this procedure, the minimization process in the problem

$$(24) \quad \text{minimize } l(x, \lambda_k, c_k) = f(x) + \lambda_k h(x) + \frac{1}{2} c_k \|h(x)\|^2$$

is terminated at a point x_k such that

$$(25) \quad \|\nabla l(x_k, \lambda_k, c_k)\| \leq \varepsilon_k,$$

where $\{\varepsilon_k\}$ is a preselected decreasing sequence tending to zero. The corresponding dual iteration can take several alternate forms. One possibility is to use the iteration of the previous section

$$(26) \quad \lambda_{k+1} = \lambda_k + c_k h(x_k)'.$$

Other possible methods of updating include the iteration

$$(27) \quad \lambda_{k+1} = \lambda_k + \beta_k h(x_k)',$$

where

$$(28) \quad \beta_k = c_k - \frac{h(x_k)' \nabla h(x_k) \nabla l(x_k, \lambda_k, c_k)}{h(x_k)' \nabla h(x_k) \nabla h(x_k)' h(x_k)},$$

proposed by Miele, et al. [18] in a somewhat different setting, and the iteration

$$(29) \quad \lambda_{k+1} = -\nabla f(x_k) \nabla h(x_k)' [\nabla h(x_k) \nabla h(x_k)']^{-1}$$

suggested by Haarhoff and Buys [11], Buys [6], and Miele, et al. [22].

One way of justifying the iteration (27) is by observing that β_k as given by (28) minimizes the quantity $\|\nabla l(x_k, \lambda_k, \beta)\|$ over β [22]. Hence, lacking further

information, the vector $h(x_k)'$ can be considered as a more accurate approximation to the gradient $\nabla g_{\beta_k}(\lambda_k)$ of the dual functional $g_{\beta_k}(\lambda)$ than to the gradient $\nabla g_{c_k}(\lambda_k)$. A similar interpretation can be given for the iteration (29). It should be mentioned that both iterations (27) and (29) reduce to the basic iteration (26) if the unconstrained minimization (24) is carried out exactly.

First let us consider the algorithm with the termination criterion (25) and the updating rule (26) (call it Algorithm A1). Let us consider again the matrix

$$(30) \quad D(x, \lambda) = -\nabla h(x)[L(x, \lambda)]^{-1}\nabla h(x)'$$

defined over $\bar{B}(x^*; \varepsilon) \times \bar{B}(\lambda^*; \delta + \bar{c}L\varepsilon)$, where ε and δ are as in Proposition 1, and assume that the algorithm generates a sequence (x_k, λ_k) converging to (x^*, λ^*) and that after some index \bar{k} , the vectors (x_k, λ_k) are contained in $B(x^*; \varepsilon) \times B(\lambda^*; \delta)$. By Proposition 1, the exact minimizing point $x(\lambda_k, c_k)$ of $l(x, \lambda_k, c_k)$ belongs to $B(x^*; \varepsilon)$.

Let $L > 0$ be as in Corollary 1.1, i.e.,

$$(31) \quad \|h(x) - h(y)\| \leq L\|x - y\| \quad \forall x, y \in B(x^*; \varepsilon),$$

and let M denote the minimum of the eigenvalues of the Hessian $L(x, \lambda, c)$ for $(x, \lambda) \in \bar{B}(x^*; \varepsilon) \times \bar{B}(\lambda^*; \delta)$, $c \in [c^*, \bar{c}]$, i.e.,

$$(32) \quad M = \min_{\substack{(x, \lambda) \in \bar{B}(x^*; \varepsilon) \times \bar{B}(\lambda^*; \delta) \\ c \in [c^*, \bar{c}] \\ i=1, \dots, n}} e_i[L(x, \lambda, c)].$$

We assume that ε and δ are sufficiently small to guarantee that $M > 0$. We have the following proposition.

PROPOSITION 3. *Under the preceding assumptions, we have, for Algorithm A1,*

$$(33) \quad \|\lambda_{k+1} - \lambda^*\| \leq r_k \|\lambda_k - \lambda^*\| + \varepsilon_k c_k (L/M) \quad \forall k \geq \bar{k},$$

where

$$(34) \quad r_k = \max_{(x, \lambda) \in \bar{B}(x^*; \varepsilon) \times \bar{B}(\lambda^*; \delta + \bar{c}L\varepsilon)} \left| \frac{1}{1 - c_k e_i[D(x, \lambda)]} \right|.$$

Proof. We have, by using (26),

$$\begin{aligned} \|\lambda_{k+1} - \lambda^*\| &= \|\lambda_k - \lambda^* + c_k h(x_k)'\| \\ &\leq \|\lambda_k - \lambda^* + c_k h[x(\lambda_k, c_k)]'\| + c_k \|h(x_k) - h[x(\lambda_k, c_k)]\|. \end{aligned}$$

Now, similarly as in Proposition 1, we have

$$\|\lambda_k - \lambda^* + c_k h[x(\lambda_k, c_k)]'\| \leq r_k \|\lambda_k - \lambda^*\|.$$

On the other hand, we have

$$\begin{aligned} \|h(x_k) - h[x(\lambda_k, c_k)]\| &\leq L\|x_k - x(\lambda_k, c_k)\| \\ &\leq \frac{L}{M} \|\nabla l(x_k, \lambda_k, c_k) - \nabla l[x(\lambda_k, c_k), \lambda_k, c_k]\| = \frac{L}{M} \|\nabla l(x_k, \lambda_k, c_k)\| \leq \frac{L\varepsilon_k}{M} \end{aligned}$$

and the result follows. Q.E.D.

The result of Proposition 3 indicates that the convergence rate of Algorithm A1 may be different from the convergence rate of the multiplier method of the previous section. Indeed, if the sequence ε_k does not converge as fast as $\|\lambda_k - \lambda^*\|$, the convergence of the sequence $\|\lambda_k - \lambda^*\|$ may not even be linear. To illustrate this fact consider the following example.

Example. Consider the problem

$$\text{minimize } \frac{1}{2}x^2 \quad \text{subject to } x = 0.$$

Take $c = 1$, and let the accuracy of the unconstrained minimization be determined by

$$\|\nabla l(x, \lambda_k, 1)\| \leq \varepsilon_k = \frac{k-1}{k(k+1)}, \quad k \geq 2,$$

where the augmented Lagrangian $l(x, \lambda, 1)$ is given by

$$l(x, \lambda, 1) = \frac{1}{2}x^2 + \lambda x + \frac{1}{2}x^2.$$

Then by direct computation, it can be seen that a possible sequence $\{\lambda_k\}$ generated by the algorithm is the sequence

$$\lambda_k = 1/k, \quad k \geq 2,$$

if the starting point is $\lambda_2 = \frac{1}{2}$. Since $\lambda^* = 0$ for this problem, we have

$$|\lambda_{k+1} - \lambda^*|/|\lambda_k - \lambda^*| = k/k + 1, \quad k \geq 2,$$

showing that the convergence of the sequence $\{\lambda_k\}$ is not linear.

In order to preserve the convergence rate of the multiplier method, it is necessary to use an approximation scheme which guarantees that the minimization is sufficiently accurate at least when we are close to the solution. Such a scheme is obtained by using, instead of the termination criterion (25), the following termination criterion:

$$(35) \quad \|\nabla l(x_k, \lambda_k, c)\| \leq \eta_k \|h(x_k)\|,$$

where $\{\eta_k\}$ is a decreasing sequence converging to zero. We shall call the algorithm resulting from use of the criterion (35) and the dual iteration (26) Algorithm A2. We can now prove the following proposition, under the assumptions of Proposition 3 and the additional assumption that $M - \eta_k L > 0$ for all $k \geq \bar{k}$.

PROPOSITION 4. *Under the preceding assumptions, we have, for Algorithm A2,*

$$(36) \quad \|\lambda_{k+1} - \lambda^*\| \leq (r_k + \frac{\eta_k L}{M - \eta_k L} p_k) \|\lambda_k - \lambda^*\| \quad \forall k \geq \bar{k},$$

where

$$(37) \quad r_k = \max_{\substack{(x, \lambda) \in \bar{B}(x^*; \varepsilon) \times \bar{B}(\lambda^*; \delta + \bar{c}Le) \\ i=1, \dots, m}} \left| \frac{1}{1 - c_k e_i[D(x, \lambda)]} \right|,$$

$$(38) \quad p_k = \max_{\substack{(x, \lambda) \in \bar{B}(x^*; \varepsilon) \times \bar{B}(\lambda^*; \delta + \bar{c}Le) \\ i=1, \dots, m}} \left| \frac{c_k e_i[D(x, \lambda)]}{1 - c_k e_i[D(x, \lambda)]} \right|.$$

Proof. We have

$$(39) \quad \begin{aligned} \|\lambda_{k+1} - \lambda^*\| &= \|\lambda_k - \lambda^* + c_k h(x_k)'\| \\ &\leq \|\lambda_k - \lambda^* + c_k h[x(\lambda_k, c_k)]'\| + c_k \|h(x_k) - h[x(\lambda_k, c_k)]\|. \end{aligned}$$

Similarly as in Propositions 2 and 3, we have

$$(40) \quad \|\lambda_k - \lambda^* + c_k h[x(\lambda_k, c_k)]'\| \leq r_k \|\lambda_k - \lambda^*\|.$$

Also we have

$$\begin{aligned} \|h(x_k) - h[x(\lambda_k, c_k)]\| &\leq L \|x_k - x(\lambda_k, c_k)\| \\ &\leq \frac{L}{M} \|\nabla l(x_k, \lambda_k, c_k)\| \leq \frac{\eta_k L}{M} \|h(x_k)\| \\ &\leq \frac{\eta_k L}{M} (\|h(x_k) - h[x(\lambda_k, c_k)]\| + \|h[x(\lambda_k, c_k)]\|), \end{aligned}$$

from which

$$(41) \quad c_k \|h(x_k) - h[x(\lambda_k, c_k)]\| \leq \frac{\eta_k L c_k}{M - \eta_k L} \|h[x(\lambda_k, c_k)]\|.$$

Since $h[x(\lambda_k, c_k)]$ is the gradient of the dual functional $g_{c_k}(\lambda)$ at λ_k and $\nabla g_{c_k}(\lambda^*) = 0$, we have

$$c_k \|h[x(\lambda_k, c_k)]\| \leq c_k \max_{\substack{\lambda \in \bar{B}(\lambda^*; \delta) \\ i=1, \dots, m}} |e_i[G_{c_k}(\lambda)]| \|\lambda_k - \lambda^*\|,$$

and by using (19) and (38),

$$(42) \quad c_k \|h[x(\lambda_k, c_k)]\| \leq p_k \|\lambda_k - \lambda^*\|.$$

By combining now (39), (40), (41) and (42), the result follows. Q.E.D.

It is to be noted that p_k is bounded and tends to unity as c_k increases, so that for large c_k and small η_k , (36) becomes approximately

$$\|\lambda_{k+1} - \lambda^*\| \leq (r_k + \eta_k L/M) \|\lambda_k - \lambda^*\|.$$

A comparison of Propositions 2 and 4 reveals now that Algorithm A2 has identical asymptotic convergence ratio with the method of multipliers.

It is easy to see that when the updating rule (27) is used instead of (26), the estimate of Proposition 4 becomes

$$(43) \quad \|\lambda_{k+1} - \lambda^*\| \leq \left(\tilde{r}_k + \frac{\eta_k L}{M - \eta_k L} \tilde{p}_k \right) \|\lambda_k - \lambda^*\|,$$

where

$$\begin{aligned} \tilde{r}_k &= \max_{\substack{(x, \lambda) \in \bar{B}(x^*; \varepsilon) \times \bar{B}(\lambda^*; \delta + \bar{c}L\varepsilon) \\ i=1, \dots, m}} \left| \frac{1}{1 - \beta_k e_i[D(x, \lambda)]} \right|, \\ \tilde{p}_k &= \max_{\substack{(x, \lambda) \in \bar{B}(x^*; \varepsilon) \times \bar{B}(\lambda^*; \delta + \bar{c}L\varepsilon) \\ i=1, \dots, m}} \left| \frac{\beta_k e_i[D(x, \lambda)]}{1 - \beta_k e_i[D(x, \lambda)]} \right|. \end{aligned}$$

When the sequence c_k converges to a finite value c , then in view of (28) and (35), we have $\beta_k \rightarrow c$, and the relation (42) yields

$$\limsup_{k \rightarrow \infty} \frac{\|\lambda_{k+1} - \lambda^*\|}{\|\lambda_k - \lambda^*\|} \leq \bar{r},$$

where

$$\bar{r} = \max_{i=1, \dots, m} \left| \frac{1}{1 - ce_i[D(x^*, \lambda^*)]} \right|.$$

In other words, Algorithm A2 with the iteration (27), (28) instead of (26) has the same asymptotic convergence ratio as the multiplier method of the previous section, which requires exact unconstrained minimization.

It should be noted that all the results of this and the previous section can be generalized for the more general algorithm which involves the (exact or approximate) unconstrained minimization of

$$l(x, \lambda_k, M_k) = f(x) + \lambda_k h(x) + \frac{1}{2} h(x)' M_k h(x),$$

where M_k are positive definite symmetric matrices. The dual iteration becomes, in this case, $\lambda_{k+1} = \lambda_k + h(x_k)' M_k$. When $M_k = c_k I$, we obtain the method of multipliers discussed earlier. The use of the matrices M_k has the effect of scaling the constraint equation. If $M_k \rightarrow M$, the convergence rate (21) becomes

$$\limsup_{k \rightarrow \infty} \frac{\|M_k^{-1/2}(\lambda_{k+1} - \lambda^*)\|}{\|M_k^{-1/2}(\lambda_k - \lambda^*)\|} \leq \max_i \left| \frac{1}{1 - e_i[M^{1/2} D(x^*, \lambda^*) M^{1/2}]} \right|,$$

and similar results as those of Propositions 2, 3 and 4 can be obtained. The updating rule (29) can be justified in the context of this more general algorithm in that λ_{k+1} as given by (29) can be written as $\lambda_{k+1} = \lambda_k + h(x_k)' M_k$ for some diagonal matrix M_k , and furthermore, M_k minimizes $\|\nabla l(x_k, \lambda_k, M)\|$ over all diagonal matrices M [22].

We finally mention that it is easy to establish local convergence results, similar to Corollary 2.1, for Algorithms A1 and A2 by making use of the arguments in the proofs of Propositions 3 and 4. The additional assumption required is that the sequences $\{\varepsilon_k\}$ and $\{\eta_k\}$ are bounded above by sufficiently small positive numbers. This assumption is necessary in order to guarantee that the generated sequence $\{\lambda_k\}$ satisfies $\|\lambda_k - \lambda^*\| \leq \|\lambda_0 - \lambda^*\| < \delta$ and hence the sequence $\{\lambda_k\}$ remains in the neighborhood $B(\lambda^*; \delta)$. Stronger global convergence results for algorithms similar to A1 and A2 have been obtained recently in [3] and [5].

5. Alternative step size choices for the method of multipliers. As mentioned in § 2, the method of multipliers can be interpreted as a fixed step size gradient method of maximizing the “penalized” dual functional $g_c(\lambda)$. It is well known [29] that in such gradient methods, the choice of step size parameter is crucial both in terms of the convergence and in terms of the rate of convergence of the method. It is a rather remarkable fact that the particular step size parameter c which is used in the multiplier method works so well from the point of view of both convergence and rate of convergence. Nonetheless, it is of interest to try to compare the step

size c with other possible step sizes and in particular with the optimal step size. This is what we attempt to do in this section. As it turns out for certain problems which are not locally convex, the step size c is close to the optimal and can hardly be improved upon. However, for the locally convex case (e.g., a convex programming problem), the analysis indicates the possibility of a significant improvement by modification of the step size. In what follows, we suggest a modified step size rule which has worked well in numerical experiments.

In order to simplify the analysis, we initially restrict attention to the case in which the objective function f is quadratic (with not necessarily positive definite or even semidefinite Hessian matrix) and the constraint functions h_i are linear. Since we shall be using results which have been proved in generality [29], it is a routine matter to extend our analysis to the general case.

Consider a multiplier method where c is held fixed for the purpose of unconstrained minimization. The step size now, however, is taken to be α rather than c , i.e., the iteration

$$(44) \quad \lambda_{k+1} = \lambda_k + \alpha \nabla g_c(\lambda_k) \quad \forall k$$

is used. Then by [29, Thm. 6], the iteration above converges for

$$(45) \quad 0 < \alpha < 2/E_c,$$

where E_c is the largest eigenvalue of the negative Hessian $-G_c(\lambda^*)$, provided that $G_c(\lambda^*)$ is a negative definite matrix. Furthermore, the rate of convergence is linear and governed by

$$(46) \quad \frac{\|\lambda_{k+1} - \lambda^*\|}{\|\lambda_k - \lambda^*\|} \leq r(\alpha) \quad \forall k,$$

where

$$(47) \quad r(\alpha) = \max \{ |1 - \alpha E_c|, |1 - \alpha e_c| \},$$

with E_c and e_c denoting the largest and smallest eigenvalues of $-G_c(\lambda^*)$. The optimal convergence ratio is attained for the step size α^* minimizing $r(\alpha)$ over α

$$(48) \quad \alpha^* = 2/(E_c + e_c)$$

and is given by

$$(49) \quad r(\alpha^*) = \frac{E_c - e_c}{E_c + e_c}.$$

In general, it is quite difficult to find the optimal step size, since this requires knowledge of the eigenvalues E_c and e_c .

Now by equation (19), we have

$$e_i[-G_c(\lambda^*)] = \frac{1}{\frac{1}{e_i[-D(x^*, \lambda^*)] + c}}$$

for all eigenvalues of the matrices $-G_c(\lambda^*)$ and $-D(x^*, \lambda^*)$. Let E_0 and e_0 denote

the eigenvalues of $-D(x^*, \lambda^*)$ corresponding to E_c and e_c in accordance with the above relation :

$$(50) \quad E_c = \frac{1}{(1/E_0) + c}, \quad e_c = \frac{1}{(1/e_0) + c}.$$

If $-D(x^*, \lambda^*)$ is positive definite, then E_0 and e_0 are its largest and smallest eigenvalues. If, however, $-D(x^*, \lambda^*)$ is neither positive nor negative definite, then E_0 is its largest negative eigenvalue, e_0 is its smallest positive eigenvalue, and $E_0 < 0 < e_0$. In view of (45) and (50), we have that convergence occurs for all step sizes α satisfying

$$(51) \quad 0 < \alpha < (2/E_0) + 2c.$$

It follows that for c much larger than $1/E_0$, the step size $\alpha = c$ of the method of multipliers is approximately in the middle of the interval of convergence, a fact which explains, to some extent, the excellent numerical stability of the method. It may also be observed that as $c \rightarrow \infty$, we have $E_c \rightarrow 1/c$, $e_c \rightarrow 1/c$, $G_c(x^*) \rightarrow -(1/c)I$, and the multiplier method iteration approaches a Newton step as noted by Luenberger [17].

The convergence ratio corresponding to the step size c is given by (cf. (47), (50))

$$(52) \quad r(c) = \max \left\{ \left| \frac{1}{1 + cE_0} \right|, \left| \frac{1}{1 + ce_0} \right| \right\}.$$

By (48), (49) and (50), the convergence ratio corresponding to the optimal step size

$$(53) \quad \alpha^* = \frac{2(1 + cE_0)(1 + ce_0)}{E_0 + e_0 + 2cE_0e_0}$$

is given by

$$(54) \quad r(\alpha^*) = \frac{E_0 - e_0}{E_0 + e_0 + 2cE_0e_0}.$$

We now distinguish two cases of interest.

Case (a) ($E_0 < 0 < e_0$). Here we assume that the matrix $D(x^*, \lambda^*)$ is neither positive semidefinite nor negative semidefinite. In this case, by (51) we must have $(-1/E_0) < c$ in order to guarantee local convexity ($0 < e_c$), in which case there exist some step sizes α which achieve convergence ($r(\alpha) < 1$). However, the particular step size $\alpha = c$ guarantees convergence only if $(-2/E_0) < c$, in which case $r(c) < 1$ (cf. Corollary 2.1). For values of c close to $-2/E_0$, equation (52) shows that the convergence ratio $r(c)$ is poor (close to one). However, as c increases, not only does the convergence ratio $r(c)$ improve, but also the ratio $r(c)/(r(\alpha^*))$ decreases, and in fact from (52) and (54),

$$\lim_{c \rightarrow \infty} \frac{r(c)}{r(\alpha^*)} = \max \left\{ \frac{2e_0}{e_0 - E_0}, \frac{2E_0}{E_0 - e_0} \right\} < 2.$$

Furthermore, it may be shown by direct calculation from (52) and (54) that if $c > (e_0 - 3E_0)/2E_0^2$, then $r(c)/(r(\alpha^*)) < 2$.

Thus for the case $E_0 < 0 < e_0$, not only is the convergence ratio $r(c)$ small for large c , but also $r(c)$ is close to being optimal and can be improved only by a factor of at most 2 by optimal step size choice. Given that $r(c)$ is already low for large c , it appears that for $E_0 < 0 < e_0$, there is rather little room for improvement of the performance of the multiplier method by alternative step size choice. This is particularly so since there are no simple ways for finding or approximating the optimal step size without explicit knowledge of the eigenvalues E_c, e_c of $-G_c$.

Case (b) ($0 < e_0 \leq E_0$). This is the locally convex case, which includes convex programming problems. For this case, the ordinary dual functional

$$g_0(\lambda) = \min_x \{f(x) + \lambda h(x)\}$$

is well-defined as a concave quadratic function. For any given $c > 0$, any step size α with $0 < \alpha < 2c$ satisfies $r(\alpha) < 1$ by (47) and (51), and hence achieves convergence. However, by direct calculation from (52) and (53), we have (assuming $e_0 \neq E_0$)

$$\frac{r(c)}{r(\alpha^*)} = \frac{1 + (e_0/E_0) + 2ce_0}{1 + ce_0} \cdot \frac{1}{1 - e_0/E_0}$$

and

$$\lim_{c \rightarrow \infty} \frac{r(c)}{r(\alpha^*)} = \frac{2}{1 - e_0/E_0} > 2.$$

The relations above show that, contrary to the previous case, there may be a substantial improvement of the convergence ratio if the optimal step size α^* can be found or approximated. The potential gain is increased as e_0 is close to E_0 , i.e., the ordinary dual problem is well-conditioned.

While the exact optimal step size α^* cannot be found except by a complete eigenvalue analysis of the matrix $D(x, \lambda)$, one may devise simple means for improving the convergence ratio by alternative step size choice. For example, if an upper bound E is known for E_0 , then the step size $\alpha = c + 1/E$ can readily be shown to yield a better convergence ratio.

In what follows, we describe a step size rule which is based on approximation of the minimum of the ordinary dual functional $g_0[\lambda_k + \alpha \nabla g_c(\lambda_k)]$ over α by means of a quadratic or cubic fit. The approximation is used every second iteration. We present the algorithm for a variable value of penalty parameter c_k .

Given λ_{2k} , and c_{2k} , $k = 0, 1, \dots$, we obtain x_{2k} and $h(x_{2k})$ by unconstrained minimization of the augmented Lagrangian, and we set

$$(55) \quad \lambda_{2k+1} = \lambda_{2k} + c_{2k} h(x_{2k})'.$$

Similarly we obtain x_{2k+1} , $h(x_{2k+1})$ by means of unconstrained minimization of the augmented Lagrangian. However, now we set

$$(56) \quad \lambda_{2k+2} = \lambda_{2k+1} + \alpha_{2k+1} h(x_{2k+1})',$$

where

$$(57) \quad \alpha_{2k+1} = c_{2k+1} \frac{h(x_{2k+1})' h(x_{2k})}{h(x_{2k+1})' h(x_{2k}) - \|h(x_{2k+1})\|^2}$$

This step size rule is obtained by observing that $h(x_{2k})$ is equal to the gradient $\nabla g_0(\lambda_{2k+1})$ and $h(x_{2k+1})$ is equal to the gradient $\nabla g_0[\lambda_{2k+1} + c_{2k+1}h(x_{2k+1})]$. Thus a quadratic approximation of $g_0[\lambda_{2k+1} + \alpha h(x_{2k+1})]$ can be made based on the two gradients and the difference $c_{2k+1}h(x_{2k+1})$ between the two points. The step size α_{2k+1} of (57) maximizes the quadratic approximation over α .

Another possibility is to determine the step size α_{2k+1} by means of a cubic fit based on the gradients $\nabla g_0(\lambda_{2k+1})$, $\nabla g_0[\lambda_{2k+1} + c_{2k+1}h(x_{2k+1})]$ and the values of the dual functional g_0 :

$$(58) \quad g_0(\lambda_{2k+1}) = f(x_{2k}) + \lambda_{2k+1}h(x_{2k}),$$

$$(59) \quad g_0[\lambda_{2k+1} + c_{2k+1}h(x_{2k+1})] = f(x_{2k+1}) + \lambda_{2k+1}h(x_{2k+1}) + c_{2k+1}\|h(x_{2k+1})\|^2.$$

The corresponding formulas for α_{2k+1} are somewhat more complicated (see [17]), but the cubic fit is more accurate than the quadratic and can be expected to yield better results for nonquadratic problems. Also, alternate quadratic fits are possible by using the values (58), (59) and one of the two gradients.

It may be shown that the sequence $\{\lambda_{2k}\}$ generated by the modified multiplier method described above satisfies, for the case of a quadratic problem,

$$\frac{\|\lambda_{2k+2} - \lambda^*\|}{\|\lambda_{2k} - \lambda^*\|} \leq \frac{1 - e_0/E_0}{(1 + c_{2k}e_0)(1 + c_{2k+1}e_0)}.$$

The bound above, though not sharp, compares favorably with the corresponding result

$$\frac{\|\lambda_{2k+2} - \lambda^*\|}{\|\lambda_{2k} - \lambda^*\|} \leq \frac{1}{(1 + c_{2k}e_0)(1 + c_{2k+1}e_0)}$$

associated with the ordinary method.

Consider now a general locally convex problem with nonquadratic objective function or nonlinear constraints. In this case it is necessary to restrict the step size α_{2k+1} of (56) to the interval $[c_{2k+1}, 2c_{2k+1}]$ in order to prove local convergence. This choice of interval is guided by (51) and by the fact that (47) and (50) yield $r(\alpha) \geq r(c)$ for all $0 < \alpha < c$ when $E_0 \geq e_0 > 0$. Thus (56) is modified to take the form

$$(60) \quad \lambda_{2k+2} = \lambda_{2k+1} + \bar{\alpha}_{2k+1}h(x_{2k+1})',$$

where

$$(61) \quad \bar{\alpha}_{2k+1} = \begin{cases} 2c_{2k+1} & \text{if } 2c_{2k+1} < \alpha_{2k+1}, \\ \alpha_{2k+1} & \text{if } c_{2k+1} \leq \alpha_{2k+1} \leq 2c_{2k+1}, \\ c_{2k+1} & \text{if } \alpha_{2k+1} < c_{2k+1}, \end{cases}$$

where α_{2k+1} is given by (57) or is obtained by means of the cubic fit mentioned earlier. We shall prove local convergence of the dual iteration (55), (60), (61) by viewing it as a special case of a more general algorithm which will be shown to be locally convergent both for the case of exact and approximate minimization of the augmented Lagrangian.

Referring to the problem of § 2, we consider the special case in which the eigenvalues of the matrix $D(x, \lambda)$ of (16) satisfy

$$(62) \quad e_i[D(x, \lambda)] < 0, \quad i = 1, \dots, m,$$

for all (x, λ) in a set $\bar{B}(x^*, \varepsilon) \times \bar{B}(\lambda^*; \delta + \bar{c}L\varepsilon)$. The positive scalars ε and δ are as in Proposition 1 and Corollary 1.1. Let the sequence $\{c_k\}$ satisfy $0 < c^* \leq c_k \leq (\bar{c}/2)$ (it is assumed that $2c^* \leq \bar{c}$) and consider the iteration

$$(63) \quad \lambda_{k+1} = \lambda_k + s_k h[x(\lambda_k, c_k)]',$$

where s_k satisfies, for all k ,

$$(64) \quad c_k \leq s_k \leq 2c_k.$$

Then we have the following local convergence result, which parallels Proposition 2 and Corollary 2.1.

PROPOSITION 5. *Assume that the initial point λ_0 belongs to $B(\lambda^*; \delta)$. Then the sequence $\{\lambda_k\}$ generated by any iteration of the form (63), (64) remains in $B(\lambda^*; \delta)$ and converges to λ^* . Furthermore, we have*

$$(65) \quad \|\lambda_{k+1} - \lambda^*\| \leq \bar{r}_k \|\lambda_k - \lambda^*\| \quad \forall k,$$

where

$$(66) \quad \bar{r}_k = \max_{\substack{(x, \lambda) \in \bar{B}(x^*; \varepsilon) \times \bar{B}(\lambda^*; \delta + \bar{c}L\varepsilon) \\ 0 < c^* \leq c_k \leq \bar{c}/2 \\ c_k \leq s_k \leq 2c_k \\ i=1, \dots, m}} \left| \frac{1 + (s_k - c_k)e_i[D(x, \lambda)]}{1 - c_k e_i[D(x, \lambda)]} \right|.$$

Proof. First, by using the facts $c_k \leq s_k \leq 2c_k$ and $e_i[D(x, \lambda)] < 0$, we have for every k ,

$$\bar{r}_k \leq \max_{\substack{(x, \lambda) \in \bar{B}(x^*; \varepsilon) \times \bar{B}(\lambda^*; \delta + \bar{c}L\varepsilon) \\ 0 < c^* \leq c_k \leq \bar{c}/2 \\ c_k \leq s_k \leq 2c_k \\ i=1, \dots, m}} \left| \frac{1 + (s_k - c_k)e_i[D(x, \lambda)]}{1 - c_k e_i[D(x, \lambda)]} \right| = p < 1.$$

Now by substituting the step size s_0 in place of c in the proofs of Propositions 1 and 2, and by using (19) and (63), we have

$$\|\lambda_1 - \lambda^*\| \leq \bar{r}_0 \|\lambda_0 - \lambda^*\| \leq p \|\lambda_0 - \lambda^*\|,$$

showing that $\lambda_1 \in B(\lambda^*; \delta)$. Proceeding similarly, we have for all k , $\lambda_k \in B(\lambda^*; \delta)$ and $\|\lambda_k - \lambda^*\| \leq p^k \|\lambda_0 - \lambda^*\|$. Hence $\lambda_k \rightarrow \lambda^*$. Q.E.D.

One may also prove propositions similar to Propositions 3 and 4 for the algorithm (63), (64) for the case of inexact minimization with either one of the termination criteria (25) or (35). For the criterion (25), we have the estimate (cf. (33))

$$\|\lambda_{k+1} - \lambda^*\| \leq \bar{r}_k \|\lambda_k - \lambda^*\| + \varepsilon_k s_k L/M,$$

and for the criterion (35) the estimate (cf. (36))

$$\|\lambda_{k+1} - \lambda^*\| \leq \left(\bar{r}_k + \frac{\eta_k L}{M - \eta_k L} \bar{p}_k \right) \|\lambda_k - \lambda^*\|,$$

where

$$\bar{p}_k = \max_{\substack{(x, \lambda) \in \bar{B}(x^*; \varepsilon) \times \bar{B}(\lambda^*; \delta + cLe) \\ i=1, \dots, m}} \left| \frac{s_k e_i[D(x, \lambda)]}{1 - c_k e_i[D(x, \lambda)]} \right|$$

and \bar{r}_k is given by (66). Local convergence results similar to Proposition 5 may also be proved assuming the sequences $\{\varepsilon_k\}$ and $\{\eta_k\}$ are bounded above by sufficiently small positive numbers.

Now the local convergence results obtained clearly apply (cf. (55), (61), (64)) to the iteration given by (55), (60), (61). As shown in the next section, this iteration worked very well in numerical experiments. The iteration (63), (64) (and hence also the iterations (55), (60), (61)) can be easily extended to the case of inequality constraints by using slack variables and therefore is fully applicable to the solution of convex programming problems. In fact, for such problems, the iteration can be shown to converge globally, i.e., for an arbitrary starting point λ_0 [4].

6. Computational experience. A limited number of numerical experiments were performed to test the analysis of this paper. As a general rule, the method of multipliers performed considerably better than the corresponding quadratic penalty function method ($\lambda_k = 0$ for all k). This was true for both exact and approximate unconstrained minimization. The schemes based on approximate minimization performed considerably better than the schemes based on exact minimization both for the penalty method and the multiplier method. The modified step size rule of the previous section performed better than the regular step size rule of the multiplier method in all runs except one. It was generally found that it is better to increase the penalty parameter c at each iteration rather than to keep it at a fixed value. It is interesting to note that for the approximate minimization schemes, the unconstrained minimizations typically required one cycle of the variable metric method after the first dual iteration. Thus the approximate minimization schemes were, in effect, similar to the conjugate gradient scheme proposed by Miele, et al. [20]. We present below some detailed results for the Rosen–Suzuki problem [38]:

$$\text{Minimize } f(x) = x_1^2 + x_2^2 + 2x_3^2 + x_4^2 - 5x_1 - 5x_2 - 21x_3 + 7x_4$$

subject to

$$h_1(x) = 2x_1^2 + x_2^2 + x_3^2 + 2x_1 - x_2 - x_4 - 5 \leq 0,$$

$$h_2(x) = x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_1 - x_2 + x_3 - x_4 - 8 \leq 0,$$

$$h_3(x) = x_1^2 + 2x_2^2 + x_3^2 + 2x_4^2 - x_1 - x_4 - 10 \leq 0.$$

The optimal solution is $x^* = (0, 1, 2, -1)'$, and the Lagrange multiplier is $\lambda^* = (2, 1, 0)$. The optimal value of the objective is $f^* = -44.0$. The eigenvalues of the negative Hessian of the ordinary dual at λ^* are $e_0 \cong .29$ and $E_0 \cong 6.95$. Here only the active constraints $h_1(x) \leq 0$, $h_2(x) \leq 0$ are considered.

The inequality constraints were converted to equality constraints by using a vector of slack variables $z = (z_1, z_2, z_3)'$:

$$\bar{h}_i(x, z) = h_i(x) + z_i^2 = 0, \quad i = 1, 2, 3.$$

The resulting augmented Lagrangian is given by

$$l(x, z, \lambda, c) = f(x) + \sum_{i=1}^3 \lambda^i [h_i(x) + z_i^2] + \frac{1}{2}c \sum_{i=1}^3 [h_i(x) + z_i^2]^2.$$

However, rather than minimizing $l(x, z, \lambda, c)$ jointly with respect to x and z , the minimization was first done explicitly with respect to z to yield

$$(67) \quad \bar{l}(x, \lambda, c) = \min_z l(x, z, \lambda, c) = f(x) + \frac{1}{2c} \sum_{i=1}^3 \{\max[0, \lambda^i + ch_i(x)]\}^2 - (\lambda^i)^2.$$

Subsequently, $\bar{l}(x, \lambda, c)$ was minimized with respect to x by using the Fletcher–Powell method (available on the IBM-360 as the FMFP Scientific Subroutine). The iteration for λ_k in the method of multipliers takes the form

$$\lambda_{k+1}^i = \max[0, \lambda_k^i + ch_i(x_k)], \quad i = 1, 2, 3,$$

where x_k is the minimizing point. This updating formula is obtained from the ordinary iteration of the method of multipliers

$$\lambda_{k+1}^i = \lambda_k^i + c\bar{h}_i(x_k, z_k), \quad i = 1, 2, 3,$$

after substitution of the minimizing value z_k obtained from (67). Table 1 shows the number of function evaluations required by the multiplier method with and without quadratic fit, and by the pure penalty method. Each function evaluation corresponds to a calculation of the values of the objective and constraint functions and their gradients.

In runs 1–5 in Table 1, accuracy to 7 significant digits of the optimal value of the objective function was attained. In runs 6–8, the accuracy was to 4 significant digits. For the runs with approximate minimization, the termination criterion (25) was used.

TABLE 1

run no.	c_k	ϵ_k	λ_0	number of function evaluations		
				multiplier	multiplier with quadratic fit	penalty
1	10^k	1×10^{-k}	(1, 1, 1)	110	107	221
2	5^k	1×5^{-k}	(0, 0, 0)	96	92	260
3	4^k	$.1 \times 4^{-k}$	(1, 1, 1)	112	119	282
4	2^k	10^{-5}	(0, 0, 0)	174	126	555
5	8^k	$.25 \times 8^{-k}$	(0, 0, 0)	93	92	192
6	1	$.1 \times 10^{-k}$	(1, 1, 1)	201	118	
7	1	$.1 \times 10^{-k}$	(0, 0, 0)	216	119	
8	1	10^{-5}	(1, 1, 1)	279	186	

7. Conclusions. This paper provided an analysis of the convergence rate of multiplier methods with exact and approximate unconstrained minimization. The results show that such methods can be expected to converge considerably faster than conventional penalty function methods. Furthermore, it appears that the approximate minimization schemes result in more efficient computation than schemes with exact minimization. The modified step size rule considered in § 5 appears to be promising for convex programming problems. While both theoretical and experimental evidence strongly indicate the faster convergence property of multiplier methods over penalty methods, it does not seem appropriate to predict that penalty methods will be totally replaced in the future by multiplier methods. In many problems where solution accuracy is not of paramount importance, penalty methods are not operated sequentially, but rather a single unconstrained minimization problem is solved with what is considered to be a sufficiently high value of penalty parameter. The solution of this problem is then taken as the final answer. When such a philosophy is adopted, multiplier methods can offer no advantage over penalty methods.

Acknowledgment. The author would like to acknowledge his interactions with Daniel Gabay, Barry Kort and David Luenberger, which were helpful in shaping the results of this paper. Many thanks also are due to one of the referees for his careful review and numerous helpful comments.

REFERENCES

- [1] K. J. ARROW, F. J. GOULD AND S. M. HOWE, *A general saddle point result for constrained optimization*. Inst. of Statistics Mimeo Ser. no. 774, Univ. of North Carolina, Chapel Hill, N.C., 1971.
- [2] K. J. ARROW, L. HURWICZ AND H. UZAWA, *Studies in Linear and Nonlinear Programming*, Stanford University Press, Calif., 1958.
- [3] D. P. BERTSEKAS, *On penalty and multiplier methods*, Dept. of Engrg.-Economic Syst. Working Paper, Stanford Univ., Stanford, Calif., 1973; this Journal, to appear.
- [4] ———, *On the method of multipliers for convex programming*. Dept. of Engrg.-Economic Syst. Working Paper, Stanford Univ., Stanford, Calif., 1973.
- [5] ———, *Convergence rate of penalty and multiplier methods*, Proc. 1973 IEEE Conf. on Decision and Control, San Diego, Calif., pp. 260–264.
- [6] J. D. BUYS, *Dual algorithms for constrained optimization*, Ph. D. thesis, Rijksuniversiteit de Leiden, the Netherlands, 1972.
- [7] A. V. FIACCO AND G. P. MCCORMICK, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley, New York, 1968.
- [8] R. FLETCHER, *A class of methods for nonlinear programming with termination and convergence properties*, Integer and Nonlinear Programming, J. Abadie (ed.), North-Holland, Amsterdam, 1970.
- [9] R. FLETCHER AND S. LILL, *A class of methods for nonlinear programming. II: Computational experience*, Nonlinear Programming, J. B. Rosen, O. L. Mangasarian and K. Ritter, eds., Academic Press, New York, 1971.
- [10] R. FLETCHER, *A class of methods for nonlinear programming. III: Rates of convergence*, Numerical Methods for Nonlinear Optimization, F. A. Lootsma, ed., Academic Press, New York, 1973.
- [11] P. C. HAARHOFF AND J. D. BUYS, *A new method for the optimization of a nonlinear function subject to nonlinear constraints*, Comput. J., 13 (1970), pp. 178–184.
- [12] M. R. HESTENES, *Multiplier and Gradient methods*, J. Optimization Theory Appl., 4 (1969), pp. 303–320.
- [13] L. V. KANTOROVICH AND G. P. AKILOV, *Functional Analysis in Normed Spaces*, Pergamon Press, New York, 1964.

- [14] B. W. KORT AND D. P. BERTSEKAS, *A new penalty function method for constrained minimization*, Proc. 1972 IEEE Conf. on Decision and Control, New Orleans, La., pp. 162–166.
- [15] ———, *Combined primal dual and penalty methods for convex programming*, Dept. of Engrg.-Economic Syst. Working Paper, Stanford Univ., Stanford, Calif., 1973.
- [16] ———, *Multiplier methods for convex programming*, Proc. 1973 IEEE Conference on Decision and Control, San Diego, Calif., pp. 428–432.
- [17] D. G. LUENBERGER, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, Mass., 1973.
- [18] O. L. MANGASARIAN, *Unconstrained Lagrangians in nonlinear programming*, Computer Sciences Tech. Rep. 174, Univ. of Wisconsin, Madison, 1973; this Journal, 13 (1975), pp. 772–791.
- [19] K. MARTSENSON, *New approaches to the numerical solution of optimal control problems*, Rep. 7206, Lund Inst. of Tech., Division of Automatic Control, Lund, Sweden, 1972.
- [20] A. MIELE, P. E. MOSELEY, A. V. LEVY AND G. M. COGGINS, *On the method of multipliers for mathematical programming problems*, J. Optimization Theory Appl., 10 (1972), pp. 1–33.
- [21] A. MIELE, P. E. MOSELEY AND E. E. CRAGG, *A modification of the method of multipliers for mathematical programming problems*, Techniques of Optimization, A. V. Balakrishnan, ed., Academic Press, New York, 1972.
- [22] A. MIELE, E. E. CRAGG, R. R. IYER AND A. V. LEVY, *Use of the augmented penalty function in mathematical programming problems, Part I*, J. Optimization Theory Appl., 8 (1971), pp. 115–130.
- [23] A. MIELE, E. E. CRAGG AND A. V. LEVY, *Use of the augmented penalty function in mathematical programming problems, Part II*, Ibid., 8 (1971), pp. 131–153.
- [24] R. MIFFLIN, *Convergence bounds for nonlinear programming algorithms*, Tech. Rep. 57, Dept. of Administrative Sciences, Yale Univ., New Haven, Conn., 1972.
- [25] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1969.
- [26] D. A. PIERRE, *Multiplier algorithms for nonlinear programming*, 7th Internat. Symp. on Math. Programming, Stanford Univ., Stanford, Calif., 1973.
- [27] E. POLAK, *Computational Methods in Optimization*, Academic Press, New York, 1971.
- [28] M. A. POLLATSCHEK, *Generalized duality theory in nonlinear programming*, Operations Res., Statistics and Economics, Mimeo Ser. 122, Technion, Haifa, Israel, 1973.
- [29] B. T. POLYAK, *Gradient methods for the minimization of functionals*, Ž. Vyčisl. Mat. i Mat. Fiz., 3 (1963), pp. 643–653.
- [30] ———, *Iterative methods using Lagrange multipliers for solving extremal problems with constraints of the equation type*, Ibid., 10 (1970), pp. 1098–1106.
- [31] ———, *The convergence rate of the penalty function method*, Ibid., 11 (1971), pp. 3–11.
- [32] M. J. D. POWELL, *A method for nonlinear constraints in minimization problems*, Optimization, R. Fletcher, ed., Academic Press, New York, 1969, pp. 283–298.
- [33] R. T. ROCKAFELLAR, *New applications of duality in convex programming*, 7th Internat. Symp. on Math. Programming, the Hague, 1970; published in Proc. 4th Conf. on Probability, Brasov, Romania, 1971.
- [34] ———, *A dual approach to solving nonlinear programming problems by unconstrained optimization*, Math. Programming, 5 (1973), pp. 354–373.
- [35] ———, *The multiplier method of Hestenes and Powell applied to convex programming*, J. Optimization Theory Appl., 12 (1973), pp. 555–562.
- [36] ———, *Augmented Lagrange multiplier functions and duality in nonconvex programming*, this Journal, 12 (1974), pp. 268–285.
- [37] ———, *Penalty methods and augmented Lagrangians in nonlinear programming*, Proc. 5th IFIP Conf. on Optimization Techniques, Rome, 1973.
- [38] J. B. ROSEN AND S. SUZUKI, *Construction on nonlinear programming test problems*, Comm. ACM, 8 (1965), p. 113.
- [39] R. D. RUPP, *Approximation of the classical isoperimetric problem*, J. Optimization Theory Appl., 9 (1972), pp. 251–264.
- [40] ———, *A method for solving a quadratic optimal control problem*, Ibid., 9 (1972), pp. 238–250.
- [41] ———, *A nonlinear optimal control minimization technique*, Trans. Amer. Math. Soc., 178 (1973), pp. 357–381.

- [42] ———, *On the combination of the multiplier method of Hestenes and Powell with Newton's method*, S.U.N.Y. Rep., Albany, N.Y., 1973.
- [43] H. SAYAMA, Y. KAMEYAMA, H. NAKAYAMA AND Y. SAWARAGI, *The generalized Lagrangian functions for mathematical programming*, 8th Internat. Symp. on Math. Programming, Stanford Univ., Stanford, Calif., 1973.
- [44] S. S. TRIPATHI AND K. S. NARENDRA, *Constrained optimization problems using multiplier methods*, J. Optimization Theory Appl., 9 (1972), pp. 59–70.

GAUSSIAN OPEN LOOP CONTROL PROBLEMS*

CHARLES J. HOLLAND†

Abstract. Expansions in powers of ε of the optimal open loop cost and control are derived for a special class of fixed stopping time small noise control problems.

Introduction. In this paper we derive expansions in powers of ε of the optimal open loop cost and control for a special class of fixed stopping time small noise open loop control problems. These problems arise by adding an additive white noise term with a small coefficient $(2\varepsilon)^{1/2}\sigma(t)$ to the system equations in the deterministic control problem. Two crucial properties are that each open loop control generates a nondegenerate Gaussian process and that the control set $K = \mathbb{R}^k$. The first property allows the conversion of the stochastic control problem into an equivalent deterministic control problem. The expansions are then established using the second property and other assumptions which guarantee that the solutions to the two-point boundary value problems determined from Pontryagin's maximum principle depend smoothly on the parameter ε .

1. The problem. Suppose that the state $\eta(t)$ evolves according to the stochastic differential equations

$$(1) \quad d\eta(t) = A(t, U(t))\eta(t) + B(t)U(t) dt + (2\varepsilon)^{1/2}\sigma(t) dw$$

with initial condition $\eta(s_0) = x_0$, x_0 a constant in \mathbb{R}^n . In (1), w is an n -dimensional Brownian motion with $w(s_0) = 0$, $U(t)$ is a control with values in the control set K , and

$$A(t, u) = A_0(t) + \sum_{j=1}^k u_j A_j(t),$$

where the $A_j(t)$, $j = 0, 1, \dots, k$, are $n \times n$ matrices and u_j is the j th component of the vector u . For each $\varepsilon \geq 0$ we seek to minimize

$$(2) \quad J_1(U, \varepsilon) = E \left\{ \int_{s_0}^T M(t, \eta(t), U(t)) dt \mid \eta(s_0) = x_0 \right\}$$

over the class of open loop controls \mathcal{U} . An *open loop control* is a bounded Borel measurable function on $[s_0, T]$ with values in K .

Throughout we assume the following:

(i) The initial point (s_0, x_0) is a fixed constant in \mathbb{R}^{n+1} , and is known to the controller. There exists a unique optimal open loop control U^0 for the deterministic control problem (1), (2) with $\varepsilon = 0$.

(ii) $K = \mathbb{R}^k$.

(iii) A is a C^∞ -function on $[s_0, T] \times \mathbb{R}^k$, and B and σ are C^∞ -functions on $[s_0, T]$.

* Received by the editors June 19, 1973, and in revised form February 1, 1974.

† Department of Mathematics, Purdue University, West Lafayette, Indiana 47907.

(iv) There exist $C_0 > 0$ and a function $d(u)$ with $d(u)/|u| \rightarrow \infty$ as $|u| \rightarrow \infty$ such that both $L(s, x, u) \geq d(u)$ and $v'L_{uu}(s, x, u)v \geq C_0 v'v$ for all $v \in \mathbb{R}^k$ and $(s, x, u) \in [s_0, T] \times \mathbb{R}^n \times \mathbb{R}^k$.

(v) Let α be an n -vector of nonnegative integers, $|\alpha|^* = \sum_{j=1}^n \alpha_j$, and $x^\alpha = x_1^{\alpha_1} \cdots x_n^{\alpha_n}$. Then

$$M(t, x, u) = \sum_{|\alpha|^* \leq m} g_\alpha(t, u) x^\alpha$$

for some positive integer m .

(vi) For each $U \in \mathcal{U}$ and $\varepsilon > 0$, the covariance matrix of the process (1) is nondegenerate Gaussian for $t > s_0$.

(vii) Let η^0 denote the trajectory corresponding to using U^0 in (1), and let λ^0 satisfy

$$(3) \quad \frac{d\lambda^{0'}}{dt}(t) = -\lambda^{0'}(t)A(t, U^0(t)) - M_x(t, \eta^0(t), U^0(t))$$

with $\lambda^0(T) = 0$. Define $H(t, x, \lambda, u) = M(t, x, u) + \lambda'(A(t, u)x + B(t)u)$ and $r(t) = (t, \eta^0(t), \lambda^0(t), U^0(t))$. Then the matrix

$$\begin{pmatrix} H_{xx}(r(t)) & H_{xu}(r(t)) \\ H_{ux}(r(t)) & H_{uu}(r(t)) \end{pmatrix}$$

is positive definite for each $t \in [s_0, T]$.

For each $\varepsilon \geq 0$ let $\mathcal{X}^\varepsilon = \inf_{U \in \mathcal{U}} J_1(U, \varepsilon)$ and let U^ε be an optimal control. The primary result of this paper is the following theorem.

THEOREM 1. *There exists a unique U^ε for sufficiently small ε . Further there exist constants \mathcal{X}_j , $j = 1, 2, \dots$, and functions U_j on $[s_0, T]$, $j = 1, 2, \dots$, such that for any positive integer l ,*

$$(4) \quad \mathcal{X}^\varepsilon = \mathcal{X}^0 + \sum_{j=1}^l \varepsilon^j \mathcal{X}_j + o(\varepsilon^l)$$

and

$$(5) \quad U^\varepsilon(t) = U^0(t) + \sum_{j=1}^l \varepsilon^j U_j(t) + o(\varepsilon^l)$$

for each $t \in [s_0, T]$.

2. Proof of the theorem. We first convert the stochastic control problem into an equivalent deterministic control problem. For each $U \in \mathcal{U}$, let the n -vector $\xi(t)$ and the $n \times n$ matrix $Q(t) = (q_{ij}(t))$ denote the mean and covariance of the nondegenerate Gaussian process (1). Let $\Gamma = (q_{11}, \dots, q_{1n}, \dots, q_{n1}, \dots, q_{nn})'$ and

$$(6) \quad L(t, x, \Gamma, u) = \sum_{|\alpha|^* \leq m} g_\alpha(t, u) \frac{\partial^{|\alpha|^*} \phi(0)}{\partial v^\alpha},$$

where ϕ , the characteristic function of $\eta(t)$, is given by

$$\phi(v) = \exp \left\{ i \sum_{j=1}^n v_j \xi_j(t) - \frac{1}{2} v' Q(t) v \right\}, \quad i = \sqrt{-1}.$$

Then

$$E \left\{ \int_{s_0}^T M(t, \eta(t), U(t)) dt | \eta(s_0) = x_0 \right\} = \int_{s_0}^T L(t, \xi(t), \Gamma(t), U(t)) dt.$$

Thus the original problem is equivalent to minimizing

$$(7) \quad J_2(U, \varepsilon) = \int_{s_0}^T L(t, \xi(t), \Gamma(t), U(t)) dt$$

over the class of open loop controls \mathcal{U} with state equations

$$(8) \quad \frac{d\xi}{dt}(t) = A(t, U(t))\xi(t) + B(t)U(t)$$

and

$$(9) \quad \frac{d\Gamma}{dt}(t) = R(t, U(t))\Gamma(t) + S(\varepsilon)$$

and initial conditions $\xi(s_0) = x_0$, $\Gamma(s_0) = 0$. The matrices R and S are determined from the equation satisfied by the covariance $Q(t)$:

$$(10) \quad \frac{dQ}{dt} = A(t, U(t))Q(t) + Q(t)A(t, U(t)) + (2\varepsilon)\sigma(t)\sigma'(t), \quad Q(s_0) = 0.$$

The existence of an optimal control U^ε for the J_2 -problem follows from Cesari [1].

We now formulate Pontryagin's maximum principle for the J_2 -problem. Let p be an n -vector and Λ an n^2 -vector. For each t , x , p , Γ , Λ , let

$$L(t, x, \Gamma, u) + p'(A(t, u)x + B(t)u) + \Lambda'R(t, u)\Gamma$$

attain its minimum on K at $u = Y(t, x, p, \Gamma, \Lambda)$. Then $U^\varepsilon(t) = Y(t, \xi^\varepsilon(t), p^\varepsilon(t), \Gamma^\varepsilon(t), \Lambda^\varepsilon(t))$, where ξ^ε , p^ε , Γ^ε and Λ^ε are the solutions with $U(t) = U^\varepsilon(t)$ to the two-point boundary value problems (8), (9), and

$$(11) \quad \begin{aligned} \frac{dp'}{dt} &= -p'A(t, U(t)) - L_x(t, \xi(t), \Gamma(t), U(t)), \\ \frac{d\Lambda'}{dt} &= -\Lambda'R(t, U(t)) - L_\Gamma(t, \xi(t), \Gamma(t), U(t)) \end{aligned}$$

with boundary conditions $p(T) = 0$, $\Lambda(T) = 0$, $\xi(s_0) = x_0$, $\Gamma(s_0) = 0$.

From (6) it follows that L is a polynomial in x and Γ with coefficients depending on t and u such that L is a C^∞ -function of the variables t , x , Γ and u . Using the smoothness assumptions on A , C and R , and the fact that $K = \mathbb{R}^k$, we can conclude that Y is a C^∞ -function of the variables t , x , p , Γ and Λ . Therefore, to establish the expansion (5), it suffices to show that $\xi^\varepsilon(t)$, $p^\varepsilon(t)$, $\Gamma^\varepsilon(t)$ and $\Lambda^\varepsilon(t)$ are C^∞ -functions of ε for each t in $[s_0, T]$.

Let $\omega = (\beta, \gamma)$, β an n -vector, γ an n^2 -vector, and let $\xi(t, \varepsilon, \omega)$, $p(t, \varepsilon, \omega)$, $\Gamma(t, \varepsilon, \omega)$ and $\Lambda(t, \varepsilon, \omega)$ be the solutions to the final value problem (8), (9) and (11) with

$$U(t) = Y(t, \xi(t, \varepsilon, \omega), p(t, \varepsilon, \omega), \Gamma(t, \varepsilon, \omega), \Lambda(t, \varepsilon, \omega))$$

and final conditions $\xi(T, \varepsilon, \omega) = \beta$, $p(T, \varepsilon, \omega) = 0$, $\Gamma(T, \varepsilon, \omega) = \gamma$ and $\Lambda(T, \varepsilon, \omega) = 0$. Define $\xi^0(T) = \beta^0$, $\omega^0 = (\beta^0, 0)$ and $\Omega(\varepsilon, \omega) = (\xi(s_0, \varepsilon, \omega) - x_0, \Gamma(s_0, \varepsilon, \omega))$. If the matrix of partial derivatives $\Omega_\omega(0, \omega^0)$ is nonsingular, there exist constants δ^* , ε^* , and a unique C^∞ -function of ε , $h(\varepsilon)$, satisfying $\Omega(\varepsilon, h(\varepsilon)) = 0$ with $|h(\varepsilon) - \omega^0| \leq \delta^*$ for $0 \leq \varepsilon \leq \varepsilon^*$.

We shall show that

$$\Omega_\omega(0, \omega^0) = \begin{pmatrix} \frac{\partial \xi}{\partial \beta}(s_0, 0, \omega^0) & \frac{\partial \xi}{\partial \gamma}(s_0, 0, \omega^0) \\ \frac{\partial \Gamma}{\partial \beta}(s_0, 0, \omega^0) & \frac{\partial \Gamma}{\partial \gamma}(s_0, 0, \omega^0) \end{pmatrix}$$

is nonsingular. When $\varepsilon = 0$, $\Gamma(t, 0, (\beta, 0)) = 0$ for all $t \in [s_0, T]$ and all β ; therefore $(\partial \Gamma / \partial \beta)(s_0, 0, \omega^0) = 0$. Next $(\partial \Gamma / \partial \gamma)(t, 0, \omega^0)$ satisfies

$$\frac{d}{dt} \left(\frac{\partial \Gamma}{\partial \gamma}(t, 0, \omega^0) \right) = R(t, U^0(t)) \frac{\partial \Gamma}{\partial \gamma}(t, 0, \omega^0)$$

with final condition $(\partial \Gamma / \partial \gamma)(T, 0, \omega^0) = I$ and hence is nonsingular on $[s_0, T]$. Thus to establish the nonsingularity of $\Omega_\omega(0, \omega^0)$, it suffices to show that $(\partial \xi / \partial \beta)(s_0, 0, \omega^0)$ is nonsingular.

We show that the singularity of $(\partial \xi / \partial \beta)(s_0, 0, \omega^0)$ leads to a contradiction. Let $Z(t) = (\partial \xi / \partial \beta)(t, 0, \omega^0)$, $W(t) = (\partial p / \partial \beta)(t, 0, \omega^0)$. Since $H_u(r(t)) \equiv 0$ (recall the definition of $r(t)$ in (vii)), then

$$(12) \quad \frac{d}{dt} \begin{bmatrix} Z \\ W \end{bmatrix} = \begin{bmatrix} D(t) - E(t)H_{uu}^{-1}(r(t))H_{ux}(r(t)) & E(t)H_{uu}^{-1}(r(t))E'(t) \\ -H_{xx}(r(t)) & -D'(t) \\ +H_{xu}(r(t))H_{uu}^{-1}(r(t))H_{ux}(r(t)) & +H_{xu}(r(t))H_{uu}^{-1}(r(t))E'(t) \end{bmatrix} \begin{bmatrix} Z \\ W \end{bmatrix}$$

with final condition $Z(T) = I$, $W(T) = 0$, where we have defined

$$D(t) = A(t, U^0(t)), \quad E(t) = B(t) + A_u(t, U^0(t)).$$

If $Z(s_0) = (\partial \xi / \partial \beta)(t, 0, \omega^0)$ is singular, then there exists a nontrivial combination of the columns of $Z(s_0)$ such that $c_1 Z_{i1}(s_0) + \cdots + c_n Z_{in}(s_0) = 0$, $i = 1, \dots, n$. Let $c = (c_1, \dots, c_n)'$, $z(t) = Z(t)c$, $w(t) = W(t)c$. Then $z(t)$, $w(t)$ is a solution to the differential equation (12) with $z(s_0) = w(T) = 0$, but $z(T) \neq 0$.

Consider the control problem with state equations

$$dz = D(t)z + E(t)y, \quad z(s_0) = 0,$$

$y \in R^k$ a control variable, and cost function

$$\int_{s_0}^T z' H_{xx}(r(t))z + z' H_{xu}(r(t))y + y' H_{ux}(r(t))z + y' H_{uu}^{-1}(r(t))y dt$$

which we seek to minimize. Clearly, the unique optimal control is $y(t) \equiv 0$ since (vii) holds. For this special problem Pontryagin's maximum principle is both necessary and sufficient. But $z(t)$, $w(t)$ satisfy the state and costate equation and therefore imply that $y(t) = -H_{uu}^{-1}(r(t))\{H_{xu}(r(t))z(t) + E(t)w(t)\}$ is a nonzero

optimal control since $z(t)$ is not identically zero. This is a contradiction and hence $\Omega_\omega(0, \omega^0)$ is nonsingular.

Hence $\xi(t, \varepsilon, h(\varepsilon))$, $p(t, \varepsilon, h(\varepsilon))$, $\Lambda(t, \varepsilon, h(\varepsilon))$ and $\Gamma(t, \varepsilon, h(\varepsilon))$ are the unique solutions to the previously defined two-point boundary value problems with $0 \leq \varepsilon \leq \varepsilon^*$ and $|\beta - \beta^0| \leq \delta^*$. A modification of Lemma 3.3 in [2] shows that $|\xi^\varepsilon(T) - \xi^0(T)| \rightarrow 0$ as $\varepsilon \rightarrow 0$. (Uniqueness of U^0 is used here.) Therefore, for sufficiently small ε , $\xi^\varepsilon(t) = \xi(t, \varepsilon, h(\varepsilon))$, $p^\varepsilon(t) = p(t, \varepsilon, h(\varepsilon))$, $\Lambda^\varepsilon(t) = \Lambda(t, \varepsilon, h(\varepsilon))$ and $\Gamma^\varepsilon(t) = \Gamma(t, \varepsilon, h(\varepsilon))$. Therefore both $\xi^\varepsilon(t)$ and $U^\varepsilon(t) = Y(t, \xi^\varepsilon(t), p^\varepsilon(t), \Lambda^\varepsilon(t), \Gamma^\varepsilon(t))$ are C^∞ -functions of ε , and U^ε is unique for ε sufficiently small.

We now establish (5). Let l be given and define the multi-index $\beta = (x, u, \varepsilon)$ and $c^\varepsilon(t) = (t, \xi^\varepsilon(t), U^\varepsilon(t), \varepsilon)$. Then

$$\begin{aligned} \mathcal{X}^\varepsilon - \mathcal{X}^0 &= \int_{s_0}^T \sum_{|\alpha|^* \leq l} \frac{\partial^{|\alpha|^*} L(c^0(t))}{\partial \beta^\alpha} (c^\varepsilon(t) - c^0(t))^\alpha \\ &\quad + \sum_{|\alpha|^* = l+1} \frac{\partial^{l+1} L(S^{\varepsilon, \alpha})}{\partial \beta^\alpha} (c^\varepsilon(t) - c^0(t))^\alpha dt, \end{aligned}$$

where $S^{\varepsilon, \alpha}$ lies on the line segment between $c^0(t)$ and $c^\varepsilon(t)$. Performing the integration on the first term on the right, one obtains a polynomial in ε^j , $j \leq l$, plus terms which are $o(\varepsilon^l)$. Since $\xi^\varepsilon(t)$, $U^\varepsilon(t)$ are uniformly bounded for $0 \leq \varepsilon \leq \varepsilon_0$, then the last term is bounded by $C\varepsilon^{l+1}$ for some constant C . Dividing by ε^l and using the Lebesgue dominated convergence theorem, one obtains the conclusion.

3. Conclusions. The expansions (4) and (5) suggest using $U^0 + \varepsilon U_1$ as a suboptimal control for sufficiently small ε . The determination of U_1 is illustrated below in two simple examples.

Example 1. Consider the scalar open loop control problem

$$d\eta(t) = U(t)\eta(t) dt + (2\varepsilon)^{1/2} dw, \quad \eta(0) = 0$$

with cost function $E \int_0^1 \eta(t)^2 + U(t)^2 dt$. The corresponding deterministic control problem is

$$\begin{aligned} \frac{d\xi}{dt}(t) &= U(t)\xi(t), & \xi(0) &= 0, \\ \frac{d\Gamma}{dt}(t) &= 2U(t)\Gamma(t) + 2\varepsilon, & \Gamma(0) &= 0 \end{aligned} \quad (13)$$

with cost function $\int_0^1 \xi(t)^2 + \Gamma(t) + U(t)^2 dt$. Since $\xi(t) \equiv 0$ for all $U \in \mathcal{U}$, we obtain that $U^\varepsilon(t) = -\Lambda^\varepsilon(t)\Gamma^\varepsilon(t)$, where Γ^ε satisfies (13) with $U = U^\varepsilon$ and

$$\frac{d\Lambda^\varepsilon}{dt}(t) = -2U^\varepsilon(t)\Lambda^\varepsilon(t) - 1, \quad \Lambda^\varepsilon(1) = 0.$$

For $\varepsilon = 0$, $U^0 \equiv 0$ which achieves zero cost. Use of U^0 in the ε -problem results in a cost of ε . Let us use $U^0 + \varepsilon U_1$ as a suboptimal control, where U_1 is defined by (5). Now $U_1 = (\partial U^0 / \partial \varepsilon)(t) = (\partial U^\varepsilon / \partial \varepsilon)(t)|_{\varepsilon=0}$ satisfies

$$U_1(t) = -\frac{\partial \Lambda^0}{\partial \varepsilon}(t) \cdot \Gamma^0(t) - \Lambda^0(t) \frac{\partial \Gamma^0}{\partial \varepsilon}(t),$$

where

$$\begin{aligned} \frac{d}{dt}\left(\frac{\partial \Gamma^0}{\partial \varepsilon}(t)\right) &= -2\frac{\partial \Lambda^0}{\partial \varepsilon}(t) \cdot \Gamma^0(t)^2 - 4\Lambda^0(t)\Gamma^0(t)\frac{\partial \Gamma^0}{\partial \varepsilon}(t) + 2, \\ \frac{\partial \Gamma^0}{\partial \varepsilon}(0) &= 0, \end{aligned}$$

and

$$\frac{d}{dt}\left(\frac{\partial \Lambda^0}{\partial \varepsilon}(t)\right) = 2\frac{\partial \Gamma^0}{\partial \varepsilon}(t) \cdot \Lambda^0(t)^2 + 4\Lambda^0(t)\Gamma^0(t)\frac{\partial \Lambda^0}{\partial \varepsilon}(t), \quad \frac{\partial \Lambda^0}{\partial \varepsilon}(1) = 0.$$

Since $\Gamma^0(t) \equiv 0$, then $U_1(t) = -(1 - t)(2t)$. Numerical calculations given in Table 1 show that $U^0 + \varepsilon U_1 = 2\varepsilon t(t - 1)$ in the ε -problem results in a lower cost than using U^0 for $\varepsilon < 4.3$.

TABLE 1

ε = Cost of Using U^0	Cost of Using $U^0 + \varepsilon U_1$
0	0
0.1	0.09873
0.2	0.19513
0.3	0.28948
0.4	0.38207
0.5	0.47315
1	0.91389
2	1.79306
4	3.91618
4.3	4.29956
4.35	4.36534
5	5.27137

Example 2. Consider $d\eta(t) = U(t) dt + (2\varepsilon)^{1/2} dw$, $\eta(0) = 0$, with cost function

$$E \int_0^1 (\eta(t)^2 + \eta'(t))^2 + \eta(t)^2 + \tfrac{1}{2}U(t)^2 dt.$$

Since $A_i(t) \equiv 0, i \geq 1$, the covariance $Q(t)$ is independent of the control and $\Gamma(t) = 2\varepsilon t$ for all $U \in \mathcal{U}$. The deterministic control problem becomes

$$\frac{d\xi}{dt}(t) = U(t), \quad \xi(0) = 0$$

with cost function

$$\begin{aligned} \int_0^1 [\xi(t)^4 + 2\xi(t)^3 + (2 + 12\varepsilon t)\xi(t)^2 + 12\varepsilon t\xi(t) + \tfrac{1}{2}U(t)^2 \\ + 4\varepsilon t + 12\varepsilon^2 t^2] dt. \end{aligned}$$

Note that this new problem is nonautonomous. For $\varepsilon = 0$, $U^0(t) \equiv 0$ achieves zero cost, while for $\varepsilon > 0$, use of U^0 results in a cost of $2\varepsilon + 4\varepsilon^2$. Since $U^\varepsilon(t) = -p^\varepsilon(t)$, then $U_1(t) = -(\partial p^0 / \partial \varepsilon)(t)$. Now

$$\frac{d}{dt} \left(\frac{\partial \xi^0}{\partial \varepsilon}(t) \right) = -\frac{\partial p^0}{\partial \varepsilon}(t), \quad \frac{\partial \xi^0}{\partial \varepsilon}(0) = 0,$$

and

$$\frac{d}{dt} \left(\frac{\partial p^0}{\partial \varepsilon}(t) \right) = -4 \frac{\partial \xi^0}{\partial \varepsilon}(t) - 12t, \quad \frac{\partial p^0}{\partial \varepsilon}(1) = 0.$$

One finds that $U_1(t) = -3(1 - (\sinh 2) \cosh 2t)$. Numerical calculations, summarized in Table 2, show that using $U_0 + \varepsilon U_1$ in the ε -problem results in a lower cost than using U^0 for $\varepsilon < 0.2$. For $\varepsilon = 1$, using $U^0 + \varepsilon U_1$ results in more than twice the cost of using U^0 . Unfortunately the range of ε for which $U^0 + \varepsilon U_1$ is better than U^0 is not known a priori.

TABLE 2

ε	Cost of Using U^0	Cost of Using $U^0 + \varepsilon U_1$ - Cost of Using U^0
0	0	0
0.04	0.0864	-0.00232
0.08	0.1856	-0.00714
0.12	0.2976	-0.011946
0.16	0.4224	-0.01375
0.20	0.56	-0.00952
0.24	0.7104	0.00388
0.40	1.44	0.21644
0.80	4.16	3.20844
1.00	6	7.08288

Remarks. Using weaker assumptions on the state equations than assumed in Theorem 1 here, Fleming, in [2, Thm. 7.1], established that both the optimal cost and feedback control for the corresponding completely observable problem are C^∞ -functions of ε . If the control set K is assumed compact, then the expansions (4) and (5) are not to be expected. In [3] this author established a one-term expansion of the open loop optimal cost under assumptions including the compactness of K . Other approaches to the open loop control problem include the work of Mortensen [4] and VanSlyke and Wets [5].

REFERENCES

- [1] L. CESARI, *Existence theorems for optimal solutions in Pontryagin and Lagrange problems*, this Journal, 3 (1965), pp. 475-498.
- [2] W. FLEMING, *Stochastic control for small noise intensities*, this Journal, 9 (1971), pp. 483-517.
- [3] C. HOLLAND, *Small noise open loop control*, this Journal, 12 (1974), pp. 380-388.
- [4] R. MORTENSEN, *Stochastic control with noisy observations*, Internat. J. Control, 4 (1966), pp. 455-464.
- [5] R. VANSLYKE AND R. WETS, *Programming under uncertainty and stochastic optimal control*, this Journal, 4 (1966), pp. 179-193.

THE SEQUENTIAL CONSTRUCTION OF MINIMAL PARTIAL REALIZATIONS FROM FINITE INPUT-OUTPUT DATA*

B. M. ANDERSON,[†] F. M. BRASCH, JR.,[‡] AND P. V. LOPRESTI[¶]

Abstract. Any strictly proper transfer function matrix of a continuous or discrete, linear, constant, multivariable system can be written as the product of a numerator polynomial matrix with the inverse of another polynomial matrix, the denominator. Since a realization is easily constructed from the polynomial matrix representation, the minimal partial realization problem is translated to that of extracting a minimal order partial denominator polynomial matrix from a finite length matrix sequence. It is shown that minimal partial denominator matrices evolve recursively; that is, a minimal partial denominator matrix for any finite length sequence is a combination of the minimal partial denominator matrices of its proper subsequences. A computationally efficient algorithm that sequentially constructs a minimal partial denominator matrix for a given finite length sequence is presented. A theorem by Anderson and Brasch leads to a definition of uniqueness for the resulting denominator matrix based upon its invariant factors. Parameters used during execution of the algorithm are shown to be sufficient for enumerating all invariant factor sets in the equivalence class of minimal partial realizations. The results apply to continuous and discrete linear systems including finite state machines.

1. Introduction. Consider the following discrete, linear, constant dynamical system:

$$(1) \quad \begin{aligned} x(k+1) &= Ax(k) + Bu(k), \\ y(k) &= Cx(k), \end{aligned} \quad k = 0, 1, 2, \dots$$

The vectors and matrices have real-valued elements or, if (1) represents a finite state machine, the vectors and matrices may be defined over a finite field. The state is denoted by the n -vector x ; u , an m -vector, and y , an r -vector, are the external input and output, respectively. Thus A , B and C are constant matrices of dimension $n \times n$, $n \times m$ and $r \times n$, respectively, over some appropriate but fixed field \mathcal{F} . For a continuous, linear, constant system,

$$(2) \quad \begin{aligned} \frac{dx(t)}{dt} &= Ax(t) + Bu(t), \\ y(t) &= Cx(t). \end{aligned}$$

Here the vectors and matrices have real-valued elements.

Both systems are characterized externally by a strictly proper rational matrix $M(z)$ called a transfer function and given by

$$(3) \quad M(z) = C(zI - A)^{-1}B.$$

* Received by the editors July 5, 1973, and in revised form February 11, 1974. This research was supported in part by the National Aeronautics and Space Administration under the Graduate Fellowship Program.

[†] MGD Graphics Systems Division, Rockwell International, Downers Grove, Illinois.

[‡] Departments of Electrical Engineering and Computer Sciences, Northwestern University, Evanston, Illinois 60201.

[¶] Western Electric Engineering Research Center, Princeton, New Jersey 08540.

When dealing with discrete-time systems, the polynomial indeterminate z can be thought of as the z -transform variable. For continuous systems z can be thought of as the Laplace transform variable.

For either the discrete or continuous system $M(z)$ is said to have a *realization* [6] given by the matrix triple $\Sigma = (A, B, C)$ if A, B and C satisfy (3). Let (3) be expanded in a Laurent series,

$$(4) \quad M(z) = \sum_{i=1}^{\infty} M_i z^{-i},$$

where the $r \times m$ matrices M_i are called Markov parameters [6] and

$$(5) \quad M_i = CA^{i-1}B, \quad i = 1, 2, \dots$$

Alternatively Σ is a realization if A, B and C satisfy (5) for all i . If the dimension of A is minimized over all matrix triples satisfying (3), then Σ is called a *minimal realization*.

The transfer function may also be expressed as the product of a numerator polynomial matrix, $Q(z) \in \mathcal{F}^{r \times m}(z)$, with the inverse of a denominator polynomial matrix, $P(z) \in \mathcal{F}_m^{m \times m}(z)$, i.e.,

$$(6) \quad M(z) = Q(z)P^{-1}(z).$$

The columns of $P(z)$ are called *column annihilating polynomials* (CAP) and are written

$$(7) \quad p_i(z) = \sum_{j=0}^{n_i} p_{ji} z^{n_i-j}, \quad 1 \leq i \leq m.$$

The matrix $P_0 = [p_{01} \ p_{02} \ \cdots \ p_{0m}]$ is called the *leading coefficient matrix*, provided $|P_0| \neq 0$. The nonnegative integers n_i , $1 \leq i \leq m$, will be called *column degrees*; their sum is the *composite degree* of $P(z)$. Then from (6),

$$(8) \quad M(z)p_i(z) = q_i(z) = \begin{cases} \sum_{j=1}^{n_i} q_{ji} z^{n_i-j}, & n_i > 0, \\ 0, & n_i = 0, \end{cases}$$

where the r -vector coefficients of column i in $Q(z)$ are given by

$$(9) \quad q_{ji} = \sum_{l=0}^{j-1} M_{j-l} p_{li}.$$

Since $|P_0| \neq 0$, let

$$(10) \quad S(z) = P_0^{-1}P(z).$$

For any $k \geq 1$,

$$(11) \quad \sum_{j=0}^{n_i} M_{k+n_i-j} p_{ji} = 0.$$

The vector π_i of length $(n_i + 1)m$ constructed from the coefficients of $p_i(z)$ as

$$(12) \quad \pi_i = \begin{bmatrix} p_{n_i,i} \\ p_{n_i-1,i} \\ \vdots \\ p_{1,i} \\ p_{0,i} \end{bmatrix}$$

will be called a *column annihilating vector* (CAV).

It has been established [4] that a realization for $M(z)$ having dimension equal to the composite degree of $P(z)$ is easily constructed from a representation $[Q(z), P(z)]$ satisfying (6). Moreover, a minimal representation [13] exists for any proper transfer function $M(z)$; that is, there exists a $P(z)$ with composite degree equal to the dimension of a minimal realization and satisfying (6). A property of $P(z)$, originally proved in [4], is given by the following theorem.

THEOREM 1. *Let $\Sigma = (A, B, C)$ be a minimal realization and let $[Q(z), P(z)]$ be a minimal representation of $M(z)$. Then the m highest degree invariant factors of $zI - A$ are identical to the invariant factors of $P(z)$.*

The invariant factors [5] of $P(z)$ are monic polynomials denoted $\gamma_{P_i}(z)$, with the property that $\gamma_{P_i}(z)$ divides $\gamma_{P_{i+1}}(z)$, $1 \leq i \leq m-1$. Thus the $m \times m$ matrix $P(z)$ contains the same information about the system dynamics that is contained in the system matrix A .

Let $\{M_i\}$ denote any infinite sequence of $r \times m$ matrices, and consider the finite Markov parameter sequence of length N , $\{M_i\}_N$. This finite sequence is said to have a *partial realization* Σ if (5) is satisfied for $i = 1, 2, \dots, N$. That is, a partial realization of $\{M_i\}_N$ has a transfer function with a Laurent expansion whose first N terms correspond identically with $\{M_i\}_N$; remaining coefficients of the Laurent expansion are called the *extension sequence*. It has been shown [12] that every finite sequence has a *minimal partial realization* that is computed from the elements M_1, M_2, \dots, M_N . It will be convenient to arrange the Markov parameters in an array called a Hankel matrix given by

$$(13) \quad \sigma^k H(i, j) = \begin{bmatrix} M_{k+1} & M_{k+2} & \cdots & M_{k+j} \\ M_{k+2} & M_{k+3} & \cdots & M_{k+j+1} \\ \vdots & \vdots & \ddots & \vdots \\ M_{k+i} & M_{k+i+1} & \cdots & M_{k+i+j-1} \end{bmatrix},$$

where the shift operator σ^k , $k \geq 0$, effectively adds k to the subscript of each block element. If σ^k is omitted, k is understood to be zero.

Since $\{M_i\}_N$ has a minimal partial realization, it also has a pair of minimal partial numerator and denominator matrices, $Q_N(z)$ and $P_N(z)$, that satisfy (6) for the transfer function of any minimal partial realization.

THEOREM 2. *Let $P_N(z)$ be a minimal partial denominator matrix for $\{M_i\}_N$. Then the column degrees of $P_N(z)$ are bounded by*

$$(14) \quad n_{Ni} \leq N, \quad 1 \leq i \leq m.$$

Proof. Since the extension sequence may be arbitrarily chosen, the degree of any column of $P_N(z)$ need not exceed N . In fact, if for any i , $n_{Ni} = N$, coefficients for $p_{Ni}(z)$, excepting the leading coefficient, may be specified arbitrarily. \square

Since (i) a minimal (partial) realization is easily constructed from a minimal (partial) representation [4], [13], and (ii) given a minimal $P(z)$ and the Markov parameter sequence, $Q(z)$ is easily obtained from (8)–(9), the minimal (partial) realization problem [12], [8] translates to the problem of extracting a minimal (partial) denominator matrix from a given finite sequence of Markov parameters.

For an arbitrary infinite sequence of $r \times m$ matrices this paper will establish that minimal partial denominator polynomial matrices for the finite length subsequences evolve recursively. This is a straightforward and useful approach to solving the multivariable minimal partial realization problem on the digital computer. These results are motivated by the work of Massey [9] where Massey's minimal length shift register synthesis algorithm is seen as a recursive means of constructing a minimal partial realization from a scalar, i.e., single-input, single-output sequence.

The main result of § 2 forms the theoretical basis for the sequential realization algorithm presented in § 3. In § 4 uniqueness is defined for sequentially generated denominator polynomial matrices. The sequential realization method is evaluated and compared with some existing realization techniques in § 5.

Notation. All scalar elements and polynomial coefficients belong to an arbitrary but fixed field \mathcal{F} ; 0 (zero) denotes the additive and 1 the multiplicative identity elements of \mathcal{F} . An $m \times n$ matrix X having rank r over \mathcal{F} is written $X \in \mathcal{F}_r^{m \times n}$. I_n represents the $n \times n$ identity matrix, 0 is any matrix of zeros, and the transpose of X is written X' . The range and null space of a matrix X are denoted $\mathcal{R}(X)$ and $\mathcal{N}(X)$, respectively. Elements of the integral domain $\mathcal{F}(z)$ are polynomials of degree l , $0 \leq l < \infty$, with coefficients in \mathcal{F} . If $X(z)$ is an $m \times n$ matrix of rank r over $\mathcal{F}(z)$, $X(z) \in \mathcal{F}_r^{m \times n}(z)$. The units of $\mathcal{F}(z)$ are the nonzero elements of \mathcal{F} ; an element is monic if its leading, i.e., highest degree, coefficient is 1. Additional notation will be presented as needed.

2. The sequential realization theorem. This section establishes the theoretical basis for the sequential realization algorithm. A lower bound on the dimension of a minimal partial realization is now given.

THEOREM 3. *The dimension of a minimal partial realization of the sequence $\{M_i\}_N$, $1 \leq N < \infty$, is zero if and only if*

$$(15) \quad M_i = 0, \quad i = 1, 2, \dots, N.$$

Proof. Sufficiency. Assume (15) holds. Then $P_N(z)$ is any nonsingular matrix, and the composite degree is zero. Hence $n = 0$.

Necessity. Assume that $M_j \neq 0$ for any $j = 1, 2, \dots, N$. Then from (11) at least one column of $P_N(z)$ has degree $n_i > 0$, implying $n > 0$. \square

This theorem may obviously be extended to include minimal realizations of the infinite sequence of zero matrices. Any sequence for which the dimension of a minimal (partial) realization is zero will be called the *zero sequence*. Any nonsingular $m \times m$ matrix is a minimal denominator polynomial matrix for the zero sequence.

DEFINITION 1. The *zero length sequence*, denoted $\{M_i\}_0$, is the sequence having no elements. Since every infinite sequence, including the zero sequence, is an extension of the zero length sequence, a minimal partial realization of $\{M_i\}_0$ has dimension zero.

DEFINITION 2. The sequence $\{\bar{M}_i\}_j$ is said to be a *subsequence* of $\{M_i\}_k$ if $j \leq k$ and $\bar{M}_i = M_i$, $i = 1, 2, \dots, j$; $\{\bar{M}_i\}_j$ is a *proper subsequence* if $j < k$. Thus for any N , $1 \leq N < \infty$, $\{M_i\}_N$ has N distinct proper subsequences, $\{M_i\}_0, \{M_i\}_1, \dots, \{M_i\}_{N-1}$.

The main result of this section is the following theorem.

THEOREM 4 (The sequential realization theorem). Let $\{M_i\}_0, \{M_i\}_1, \dots, \{M_i\}_j, \dots$ be the distinct proper subsequences of an arbitrary infinite sequence $\{M_i\}$. Then for any $N \geq 0$, the information contained in the minimal partial denominator matrices for $\{M_i\}_0, \{M_i\}_1, \dots, \{M_i\}_N$, denoted by $P_0(z), P_1(z), \dots, P_N(z)$, respectively, is sufficient to calculate a minimal partial denominator matrix $P_{N+1}(z)$ for $\{M_i\}_{N+1}$.

Proof. This theorem is proved in a vector space rather than a polynomial matrix formulation. Some notational preliminaries and lemmas are required before proceeding with the main proof.

To column i , $1 \leq i \leq m$, of $P_j(z)$, $0 \leq j \leq N$, there corresponds a CAV π_{ji} with degree n_{ji} ; by Theorem 2, $0 \leq n_{ji} \leq j$. Applying π_{ji} to the shifted Hankel matrix of $\{M_i\}_{j+1}$ below yields a vector d_{ji} , where

$$(16) \quad d_{ji} = \sigma^{j-n_{ji}} H(1, n_{ji} + 1) \pi_{ji}.$$

DEFINITION 3. The $r \times m$ matrix D_j given by

$$(17) \quad D_j = [d_{j1} \quad d_{j2} \quad \dots \quad d_{jm}]$$

is called the *j-th discrepancy matrix*. If D_j is zero, it is clear from (11) that $P_j(z)$ is also a minimal partial denominator matrix for $\{M_i\}_{j+1}$. More likely, $D_j \neq 0$, so an alternative, less obvious method for finding $P_{j+1}(z)$ is required. This is the topic of the remainder of this section.

The next two definitions are made primarily for notational convenience.

DEFINITION 4. For the vector π_{ji} of degree n_{ji} define the *augmented column annihilating vector* $\pi_{ji}(l, n_{ji} + 1, k)$ as

$$(18) \quad \pi_{ji}(l, n_{ji} + 1, k) = \begin{bmatrix} 0 \\ \vdots \\ \pi_{ji} \\ \vdots \\ 0 \end{bmatrix}.$$

That is, the vector $\pi_{ji}(l, n_{ji} + 1, k)$ consists of π_{ji} embedded between two zero vectors of lengths lm and km , respectively. The length of the augmented vector is $m(l + m_{ji} + 1 + k) \geq m(n_{ji} + 1)$. Thus $\pi_{ji}(0, n_{ji} + 1, 0) = \pi_{ji}$. The degree associated with an augmented CAV is the same as the degree of the CAV being augmented.

DEFINITION 5. Associated with π_{ji} and $p_{ji}(z)$ is an integer k_{ji} called the *accumulation index relative to N* and given by

$$(19) \quad k_{ji} = (N - j) + n_{ji}, \quad 0 \leq j \leq N, \quad 1 \leq i \leq m.$$

It will be shown that $p_{N+1,i}(z)$ is a polynomial combination of columns in $P_j(z)$, $0 \leq j \leq N$; the degree of $p_{N+1,i}(z)$ is determined from the largest accumulation index of elements comprising the combination. Equation (16) becomes

$$(20) \quad d_{ji} = \sigma^{j-n_{ji}} H(1, k_{ji} + 1) \pi_{ji}(0, n_{ji} + 1, N - j).$$

A slight rearrangement of (19) substituted into (20) yields

$$(21) \quad d_{ji} = H(1, N + 1) \pi_{ji}(N - k_{ji}, n_{ji} + 1, k_{ji} - n_{ji}).$$

LEMMA 1. For any i and j , $1 \leq i \leq m$, $0 \leq j \leq N$, to each $p_{ji}(z)$ there corresponds a set of $N - k_{ji}$ linearly independent augmented CAV's in $\mathcal{N}[H(1, N + 1)]$.

The proof follows from observing that for every $n_{ji} < j$ the structure of the Hankel matrices for $\{M_i\}_{N+1}$ implies

$$(22) \quad \sigma^{j-n_{ji}-l} H(1, k_{ji} + 1) \pi_{ji}(0, n_{ji} + 1, N - j) = 0, \quad 1 \leq l \leq j - n_{ji}.$$

Since $N - k_{ji} = j - n_{ji}$, (22) becomes

$$(23) \quad H(1, N + 1) \pi_{ji}(N - k_{ji} - l, n_{ji} + 1, N - j + l) = 0, \quad 1 \leq l \leq N - k_{ji},$$

which proves the lemma.

Now let augmented CAV's from $P_j(z)$ form the columns of $\theta_j(N + 1) \in \mathcal{F}_m^{(N+1)m \times n}$, i.e.,

$$(24) \quad \theta_j(N + 1) = [\pi_{j1}(N - k_{j1}, n_{j1} + 1, N - j) \cdots \pi_{jm}(N - k_{jm}, n_{jm} + 1, N - j)]$$

so that from (21),

$$(25) \quad H(1, N + 1) \theta_j(N + 1) = D_j, \quad 0 \leq j \leq N.$$

By letting

$$(26) \quad \Theta(N + 1) = [\theta_0(N + 1) \quad \theta_1(N + 1) \quad \cdots \quad \theta_N(N + 1)]$$

and

$$(27) \quad \Delta(N + 1) = [D_0 \quad D_1 \quad \cdots \quad D_N] = [\Delta(N) : D_N],$$

equation (25) yields

$$(28) \quad H(1, N + 1) \Theta(N + 1) = \Delta(N + 1).$$

The matrix $\Theta(N + 1)$ is square and block upper triangular of size $(N + 1)m$; by construction the $m \times m$ diagonal blocks are the nonsingular leading coefficient matrices of $P_0(z), P_1(z), \dots, P_N(z)$. Thus $\Theta(N + 1)$ is nonsingular.

LEMMA 2. $\mathcal{R}[H(1, N + 1)] = \mathcal{R}[\Delta(N + 1)]$.

The proof of this lemma is obvious from (28) and the nonsingularity of $\Theta(N + 1)$. In accordance with Definition 1 define $\Delta(0)$ and $H(1, 0)$ to be the zero vector so $\mathcal{R}[H(1, 0)] = \mathcal{R}[\Delta(0)]$, the space spanned by the zero vector.

The number of columns of $P_{N+1}(z)$ having degree equal to $N + 1$ may be determined from the N th discrepancy matrix D_N and $\Delta(N)$. Let $\hat{\theta}_N(N + 1)$ and \hat{D}_N with columns \hat{d}_{Ni} be given by

$$(29) \quad \begin{aligned} \hat{D}_N &= D_N \hat{U}_N, \\ \hat{\theta}_N(N + 1) &= \theta_N(N + 1) \hat{U}_N, \end{aligned}$$

where \hat{U}_N is obtained as follows. Define D_N^* with columns

$$(30) \quad d_{Ni}^* = \begin{cases} d_{Ni} & \text{if } d_{Ni} \notin \mathcal{R}[\Delta(N)], \\ 0 & \text{otherwise,} \end{cases} \quad 1 \leq i \leq m.$$

Then form \hat{U}_N unit upper triangular so that the nonzero columns of

$$(31) \quad \bar{D}_N = D_N^* \hat{U}_N$$

are linearly independent.

LEMMA 3. *There is a one-to-one correspondence between the nonzero columns of \bar{D}_N and the columns of $P_{N+1}(z)$ having degree $N + 1$.*

By Theorem 2 each column of $P_{N+1}(z)$ may be placed in one of two categories: those with degree equal to $N + 1$, and those with degree less than $N + 1$. By Lemma 2 and (29), any linear combination of the columns of $H(1, N + 1)$ is also a linear combination of the columns of $[\Delta(N) : \hat{D}_N]$. No combination of the columns of $[\Delta(N) : \hat{D}_N]$ that includes a nonzero column of \bar{D}_N can equal zero because the nonzero columns of \bar{D}_N are linearly independent and are independent of the columns of $[\Delta(N) : (\hat{D}_N - \bar{D}_N)]$. Since the leading coefficient matrix of $P_{N+1}(z)$ is nonsingular and every column is a CAP, to every nonzero column of \bar{D}_N there must correspond a column of $P_{N+1}(z)$ with degree $N + 1$. It remains to show that to every zero column of \bar{D}_N there corresponds a column of $P_{N+1}(z)$ with degree less than $N + 1$. For any i , $1 \leq i \leq m$, suppose $\bar{d}_{Ni} = 0$. Then there exists a vector t such that

$$\Delta(N)t + \hat{d}_{Ni} = 0.$$

By Lemma 2 there exists a vector x such that

$$H(1, N)x + \hat{d}_{Ni} = 0.$$

Substituting for \hat{d}_{Ni} ,

$$(32) \quad H(1, N + 1) \left[\begin{bmatrix} x \\ -\frac{x}{0} \end{bmatrix} + \hat{\pi}_{Ni}(N - \hat{n}_{Ni}, \hat{n}_{Ni} + 1, 0) \right] = 0,$$

where $\hat{\pi}_{Ni}$ is the i th column of $\hat{\theta}_N(N + 1)$. The vector postmultiplying $H(1, N + 1)$ in (32) is clearly an augmented CAV for $\{M_i\}_{N+1}$. Thus an augmented CAV satisfying (32) exists for every zero column of \bar{D}_N . Since $P_{N+1}(z)$ is minimal, Lemma 3 is proved.

If there are r linearly independent columns in $\Delta(N + 1)$, it is of full row rank. By Lemma 3 under the restrictions imposed by (29)–(31), there can be no more than r columns in all of the $P_j(z)$, $0 \leq j \leq N + 1$, for which $n_{ji} = j$. Aside from these at most r columns, the degree associated with each column of $P_j(z)$ must be strictly less than j , $0 \leq j \leq N + 1$.

Let the $(N + 1)m$ -vector ϕ represent any augmented CAV of degree $n_\phi < N + 1$ associated with the finite sequence $\{M_i\}_{N+1}$. Moreover, assume ϕ is partitioned in m -vector segments as

$$\phi = \begin{bmatrix} \phi_N \\ \phi_{N-1} \\ \vdots \\ \phi_1 \\ \phi_0 \end{bmatrix}, \quad \phi_0 \neq 0.$$

Then

$$(33) \quad H(1, N + 1)\phi = \sum_{j=0}^{n_\phi} M_{N+1-j}\phi_j = 0.$$

Since $\Theta(N + 1)$ is nonsingular, there exists a vector x , partitioned like ϕ and satisfying,

$$(34) \quad \phi = \Theta(N + 1)x = \sum_{j=0}^N \theta_{N-j}(N + 1)x_j.$$

LEMMA 4. Let ϕ and x be as given above and let x_{ji} , $1 \leq i \leq m$, denote the elements of x_j . Then ϕ is a linear combination of augmented CAV's from $P_0(z)$, $P_1(z)$, \dots , $P_N(z)$, and its degree is

$$(35) \quad n_\phi = \max_{0 \leq j \leq N} \left\{ \max_{1 \leq i \leq m} \{k_{N-j,i} | x_{ji} \neq 0\} \right\}.$$

The first part of the lemma follows from (34); it remains to prove (35). First assume that $x_0 \neq 0$ and that

$$n_\phi < \max_{1 \leq i \leq m} \{k_{Ni} | x_{0i} \neq 0\}.$$

Then replace the column of $\theta_N(N + 1)$ with nonzero x_{0i} and the largest accumulation index with ϕ . Note that $k_{Ni} = n_{Ni}$. The result is a new matrix, $\tilde{\theta}_N(N + 1)$, of augmented CAV's with a lower composite degree than $\theta_N(N + 1)$. This would contradict the minimality of $P_N(z)$ so

$$(36) \quad n_\phi \geq \max_{1 \leq i \leq m} \{k_{Ni} | x_{0i} \neq 0\}.$$

The proof is continued for $j = 1, 2, \dots, N$ on the vectors ϕ^j given by

$$(37) \quad \phi^j = \phi - \sum_{l=0}^{j-1} \theta_{N-l}(N + 1)x_l.$$

Assume $x_j \neq 0$ and that

$$n_\phi < \max_{1 \leq i \leq m} \{k_{N-j,i} | x_{ji} \neq 0\}.$$

Then ϕ^j replaces the column of $\theta_{N-j}(N+1)$ having the highest accumulation index and $x_{ji} \neq 0$ to yield a matrix of augmented CAV's with a lower composite degree than $\theta_{N-j}(N+1)$. This contradicts the assumed minimality of $P_{N-j}(z)$. Hence combining with (36),

$$n_\phi \geq \max_{0 \leq j \leq N} \left\{ \max_{1 \leq i \leq m} \{k_{N-j,i} | x_{ji} \neq 0\} \right\}.$$

But by Lemma 1 every column of $\Theta(N+1)$ for which $x_{ji} \neq 0$ can, by appropriate internal shifts, generate $N - k_{N-j,i}$ linearly independent vectors in $\mathcal{N}[H(1, N+1)]$. By the same argument ϕ can be shifted internally to produce a total of $N+1 - n_\phi$ augmented CAV's in $\mathcal{N}[H(1, N+1)]$. Hence Lemma 4 is proved.

Now consider the matrix $\theta_{N+1}(N+2) \in \mathcal{F}_m^{(N+2)m \times m}$ with columns

$$\pi_{N+1,i}(N+1 - n_{N+1,i}, n_{N+1,i} + 1, 0), \quad 1 \leq i \leq m,$$

constructed from $P_{N+1}(z)$. The first m rows in any column of $\theta_{N+1}(N+2)$ having degree less than $N+1$ are zero; columns with degree $N+1$ can have the first m row elements set arbitrarily to zero. Thus $\theta_{N+1}(N+2)$ may be written

$$(38) \quad \theta_{N+1}(N+2) = \begin{bmatrix} 0 \\ \bar{F}_N \end{bmatrix},$$

where $F_N \in \mathcal{F}_m^{(N+1)m \times m}$

Certain nonsingular matrices specified in the following lemma postmultiply $\theta_{N+1}(N+2)$ to yield a matrix of augmented CAV's having the same composite degree.

LEMMA 5. *The composite degree of $\theta_{N+1}(N+2)$ remains invariant under postmultiplication by a nonsingular matrix R if (i) R is a diagonal matrix, (ii) R is a permutation matrix or (iii) R is lower triangular and the columns of $\theta_{N+1}(N+2)$ are ordered from the left by decreasing degree, i.e., any column of $\theta_{N+1}(N+2)$ has degree less than or equal to the degree of every column to its left.*

Now to proceed with the proof of Theorem 4. From Lemmas 3 and 4 it is clear that

$$(39) \quad H(1, N+1)F_N = E_N,$$

where the nonzero columns of E_N are linearly independent and are not elements of $\mathcal{R}[H(1, N)]$. Moreover, for every nonzero column of E_N there is a column of $P_{N+1}(z)$ with degree equal to $N+1$. Since $\Theta(N+1)$ is nonsingular, there exist matrices $W_N \in \mathcal{F}_m^{m \times m}$ and $V_N \in \mathcal{F}^{Nm \times m}$ satisfying

$$(40) \quad F_N = \Theta(N+1) \begin{bmatrix} V_N \\ \bar{W}_N \end{bmatrix}.$$

From Lemma 2 and (27)–(28),

$$(41) \quad H(1, N+1)F_N = [\Delta(N) : D_N] \begin{bmatrix} V_N \\ \bar{W}_N \end{bmatrix}.$$

It is now claimed, without loss of generality, that W_N is unit upper triangular. In its most general form the nonsingular W_N may be factored [3] as follows:

$$(42) \quad W_N = U_N R_N L_N,$$

where U_N is unit upper triangular, L_N is lower triangular and R_N is a permutation matrix. Should W_N have the form (42) with the columns of $\theta_{N+1}(N+2)$ ordered from left to right by decreasing degree, $\theta_{N+1}(N+2)$ can be postmultiplied first by L_N^{-1} and then by R_N . According to Lemma 5, the resulting matrix of augmented CAV's has the same composite degree as the original and is thus minimal. Hence $W_N = U_N$. Note also that U_N contains \hat{U}_N of (29)–(31) as a factor, i.e., $U_N, \hat{U}_N, \hat{U}_N^{-1}U_N$ are all unit upper triangular and

$$(43) \quad H(1, N+1)F_N = [\Delta(N) : \hat{D}_N] \begin{bmatrix} V_N \\ \hat{U}_N^{-1} \hat{U}_N \end{bmatrix}.$$

Combining (38)–(40) expresses in augmented column annihilating vector form $P_{N+1}(z)$ as a combination of the columns of $P_0(z), P_1(z), \dots, P_N(z)$, i.e.,

$$(44) \quad \theta_{N+1}(N+2) = \begin{bmatrix} 0 \\ \vdots \\ \Theta(N+1) \end{bmatrix} \begin{bmatrix} V_N \\ \hat{U}_N \end{bmatrix}.$$

This proves Theorem 4. \square

The form of (44), i.e., U_N upper triangular with unity diagonal elements, indicates that, regardless of the associated degree, column i in $\theta_{N+1}(N+2)$ is a shifted linear combination of column i in $\theta_N(N+1)$ with columns to its left in $\Theta(N+1)$.

3. The sequential realization algorithm. The results of the previous section will be used in this section to develop an algorithm for the recursive construction of the minimal partial denominator matrices of the finite length subsequences of a given infinite sequence. The proof of Theorem 4 is constructive because it demonstrates that column i of $\theta_{N+1}(N+2)$ equals column i of $\theta_N(N+1)$ shifted and added to a linear combination of shifted columns to the left of i in $\Theta(N+1)$. The exact numerical form of this linear combination is dependent upon column i of the N th discrepancy matrix, and it is selected to minimize $n_{N+1,i}$. Let

$$(45) \quad \Delta_i(N) = \begin{cases} \Delta(N), & i = 1, \\ [\Delta(N) : d_{N1} \quad d_{N2} \quad \cdots \quad d_{N,i-1}], & 1 < i \leq m. \end{cases}$$

By Lemma 3 and construction, $n_{N+1,i} = N+1$ if and only if $d_{Ni} \notin \mathcal{R}[\Delta_i(N)]$. If $d_{Ni} \notin \mathcal{R}[\Delta_i(N)]$, every row element in column i of $\theta_{N+1}(N+2)$ is arbitrary except for the last $m+1-i$; hence the linear combination is also arbitrary.

If $d_{Ni} \in \mathcal{R}[\Delta_i(N)]$, $n_{N+1,i} < N+1$ and the selected linear combination must satisfy two conditions: when applied to the columns of $\Delta_i(N)$, it must yield $-d_{Ni}$; and when applied to the columns of $\Theta(N+1)$, $n_{N+1,i}$, given by (35) as a function of the accumulation indices of columns included in the combination, must be minimized. If $n_{N+1,i}$ is minimized in this manner column by column, (44) clearly indicates that the composite degree of $\theta_{N+1}(N+2)$ is also minimized.

Let $\rho_{Ni} = \text{rank} [\Delta_i(N)]$. Since a column basis for $\Delta_i(N)$ is formed from ρ_{Ni} linearly independent columns, $0 \leq \rho_{Ni} \leq r$, usually only a subset of the columns of $\Delta_i(N)$ need be considered for any minimal linear combination that produces $-d_{Ni} \in \mathcal{R}[\Delta_i(N)]$; viz. a subset of ρ_{Ni} linearly independent columns generated by a corresponding subset of columns in $\Theta(N+1)$ whose largest accumulation index is minimized. More precisely, let K_{Ni} be a spanning set of columns of $\Delta_i(N)$ generated by

$$(46) \quad H(1, N+1)\Phi_{Ni} = K_{Ni},$$

where

$$(47) \quad \Phi_{Ni} = \begin{cases} [\phi_{Ni1} & \phi_{Ni2} & \cdots & \phi_{Ni\rho_{Ni}}], & \rho_{Ni} \geq 1, \\ 0, & \rho_{Ni} = 0. \end{cases}$$

The columns ϕ_{Nij} are columns of $\Theta(N+1)$ with degree $n_{\phi_{Nij}}$ and accumulation index $k_{\phi_{Nij}}$ selected so as to minimize $\max_{1 \leq j \leq \rho_{Ni}} \{k_{\phi_{Nij}}\}$. By the previous discussion, if $n_{N+1,i} < N+1$, either $n_{N+1,i} = n_{Ni}$ or $n_{N+1,i}$ equals the largest accumulation index of the columns of Φ_{Ni} included in the linear combination. In either case $n_{N+1,i}$ is minimized. Finally, for notational consistency, let $b_{Nij}(z)$ denote the column of $P_0(z), P_1(z), \dots, P_N(z)$ from which ϕ_{Nij} was constructed, and let

$$(48) \quad B_{Ni}(z) = \begin{cases} [b_{Ni1}(z) & b_{Ni2}(z) & \cdots & b_{Ni\rho_{Ni}}(z)], & \rho_{Ni} \geq 1, \\ 0, & \rho_{Ni} = 0. \end{cases}$$

The following algorithm is a procedure for computing a minimal partial denominator polynomial matrix for a finite matrix sequence of length $N_0 \geq 0$. Initially, I_m is assumed as the minimal denominator for the zero length sequence; the procedure halts with the minimal denominator for $\{M_i\}_{N_0}$. The notation is consistent with that presented previously except that the subscript Ni has been dropped for convenience. Also the algorithm is presented in polynomial rather than vector space notation.

THE SEQUENTIAL REALIZATION ALGORITHM.

Step 1. Initialization. Set $N = \rho = 0$, $B(z) = 0$, $K = 0$, $P(z) = I_m$, $n_i = 0$, $1 \leq i \leq m$.

Step 2. If $N = N_0$, stop.

Step 3. Otherwise perform Steps 4–14 for $i = 1, 2, \dots, m$.

Step 4. From column i of $P(z)$ compute

$$d = \sum_{j=0}^{n_i} M_{N+1-j} p_{ji}.$$

Step 5. If $d = 0$, go to Step 14.

Step 6. If $d \neq 0$ and $\rho = 0$, set $\rho = 1$, $K = d$, $B(z) = b_1(z) = p_i(z)$, $k_{\phi_1} = n_{\phi_1} = n_i$, $p_i(z) = p_i(z)z^{N+1-n_i}$, $n_i = N+1$ and go to Step 14.

Step 7. If $d \neq 0$ and $\rho > 0$, compute $x = K^\#d$,

$$\bar{d} = d - Kx = (I - KK^\#)d,$$

where $K^\#$ is a pseudoinverse of K . That is, $K = KK^\#K$. (See [3] for a discussion of pseudoinverses of finite field matrices.)

Step 8. Set

$$n_t = \begin{cases} N + 1 & \text{if } \bar{d} \neq 0, \\ \max \left\{ n_i, \max_{1 \leq j \leq \rho} \{k_{\phi j} | x_j \neq 0\} \right\} & \text{if } \bar{d} = 0 \end{cases}$$

and

$$t(z) = p_i(z)z^{(n_t - n_i)} - \sum_{j=1}^{\rho} b_j(z)x_j z^{(n_t - k_{\phi j})}.$$

Step 9. If $n_t = n_i$, go to Step 13.

Step 10. If $n_t > n_i$, set $K = [K : d]$, $b_{\rho+1}(z) = p_i(z)$, $B(z) = [B(z) : p_i(z)]$, $k_{\phi, \rho+1} = n_{\phi, \rho+1} = n_i$.

Step 11. Order the columns of $B(z)$ from left to right by increasing accumulation index; place the corresponding columns of K in the same order.

Step 12. Set $\rho = \text{rank}(K)$. Discard the linearly dependent column of K , if any, with the largest accumulation index, i.e., the right-most linearly dependent column. Remove from $B(z)$ the column corresponding to any discarded column of K . If necessary, renumber the columns of K and $B(z)$ from 1 through ρ .

Step 13. Set $p_i(z) = t(z)$ and $n_i = n_t$.

Step 14. If $i < m$, increment i by 1 and go to Step 4.

Step 15. If $i = m$, set $N = N + 1$. If $\rho > 0$, set $k_{\phi j} = k_{\phi j} + 1$, $1 \leq j \leq \rho$, and go to Step 2.

Before demonstrating the algorithm on some examples it should be noted that the following values hold at Step 2 for any $N \leq N_0$:

$$P(z) = P_N(z) \quad \text{with column degrees} \quad n_i = n_{N_i}, \quad 1 \leq i \leq m, \quad n = n_N;$$

$$K = K_{N_1} \quad \text{of rank} \quad \rho_{N_1} = \text{rank}[\Delta(N)] = \rho;$$

$$B(z) = B_{N_1}(z) \quad \text{with accumulation indices} \quad k_{\phi j} = k_{\phi N_1 j}, \quad 1 \leq j \leq \rho_{N_1}.$$

The algorithm is now applied to an example from [12] where the sequence is defined over the real number field. Parameter values are given at Step 2 of the algorithm for $N = 0, 1, 2, 3, 4$.

$$\text{Example 1. Let } N_0 = 4 \text{ and } \{M_i\}_4 = \left\{ \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 4 & 3 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 10 & 7 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 22 & 15 \\ 3 & 3 \end{pmatrix} \right\}.$$

$$N = 0: \quad P(z) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad n_1 = n_2 = n = 0,$$

$$K = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \rho = 0,$$

$$B(z) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

$$N = 1: \quad P(z) = \begin{bmatrix} z & -1 \\ 0 & 1 \end{bmatrix}, \quad n_1 = 1, \quad n_2 = 0, \quad n = 1,$$

$$K = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \rho = 1,$$

$$B(z) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad k_{\phi 1} = 1.$$

$$N = 2: \quad P(z) = \begin{bmatrix} z - 4 & -z + 1 \\ 0 & z \end{bmatrix}, \quad n_1 = n_2 = 1, \quad n = 2,$$

$$K = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \quad \rho = 1,$$

$$B(z) = \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \quad k_{\phi 1} = 1.$$

$$N = 3: \quad P(z) = \begin{bmatrix} z^3 + 2z^2 & -z \\ -6z^2 & z + 1 \end{bmatrix}, \quad n_1 = 3, \quad n_2 = 1, \quad n = 4,$$

$$K = \begin{bmatrix} -1 & -6 \\ 0 & 1 \end{bmatrix}, \quad \rho = 2,$$

$$B(z) = \begin{bmatrix} -1 & z - 4 \\ 1 & 0 \end{bmatrix}, \quad k_{\phi 1} = k_{\phi 2} = 2.$$

$$N = 4: \quad P(z) = \begin{bmatrix} z^3 + 3z^2 + 2z & -z^2 - z - 2 \\ -6z^2 - 6z & z^2 + z + 6 \end{bmatrix}, \quad n_1 = 3, \quad n_2 = 2, \quad n = 5,$$

$$K = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \quad \rho = 2,$$

$$B(z) = \begin{bmatrix} -z & -1 \\ z + 1 & 1 \end{bmatrix}, \quad k_{\phi 1} = 2, \quad k_{\phi 2} = 3.$$

4. Uniqueness of minimal partial denominator matrices. In this section uniqueness of a minimal partial denominator polynomial matrix is defined, and a criterion for determining the uniqueness of a sequentially computed denominator matrix is presented. The information carried along while applying the sequential realization algorithm, viz. N , ρ , $P(z)$, $B(z)$, K , $\{n_i, i = 1, \dots, m\}$ and $\{n_{\phi j}, k_{\phi j}, j = 1, \dots, \rho\}$, is shown to be sufficient for constructing the equivalence class of minimal partial denominator matrices after the algorithm terminates.

First consider the set of all minimal partial denominator matrices for $\{M_i\}_{N+1}$, i.e., let

$$(49) \quad \Omega_{N+1} = \{\tilde{P}_{N+1}(z) | \tilde{P}_{N+1}(z) \text{ is a minimal partial denominator matrix for } \{M_i\}_{N+1}\}.$$

It is useful to define uniqueness for a minimal partial denominator matrix in terms of its invariant factors.

DEFINITION 6. The minimal partial denominator matrix $P_{N+1}(z)$ for $\{M_i\}_{N+1}$ is *unique* if it and every other element $\tilde{P}_{N+1}(z) \in \Omega_{N+1}$ has the same set of invariant factors.

The above is consistent with the definition given in [8] for uniqueness of minimal realizations modulo a choice of basis for the state. That is, a minimal realization is said to be unique modulo an equivalence class of similar system matrices. This definition has now been extended to partial realizations via the polynomial matrix formulation.

THEOREM 5. Let $P_{N+1}(z)$ and $B_{N+1,1}(z)$ be computed as above and let $\tilde{P}_{N+1}(z) \in \Omega_{N+1}$. Then there exist polynomial matrices $X_{N+1}(z)$ and $Y_{N+1}(z)$ such that

$$(50) \quad \tilde{P}_{N+1}(z) = P_{N+1}(z)X_{N+1}(z) + B_{N+1,1}(z)Y_{N+1}(z),$$

where $X_{N+1}(z)$ is an elementary matrix.

Proof. Let $\tilde{\theta}_{N+1}(N+2)$ be a set of augmented CAV's constructed from $\tilde{P}_{N+1}(z)$ according to (24). Then $\tilde{\theta}_{N+1}(N+2)$ generates a discrepancy matrix when postmultiplying $H(1, N+2)$. Since $\Theta(N+2)$ is nonsingular and upper block triangular,

$$(51) \quad \begin{aligned} \tilde{\theta}_{N+1}(N+2) &= \Theta(N+2) \begin{bmatrix} T_{N+1} \\ \hline R_{N+1} \end{bmatrix} \\ &= \begin{bmatrix} \Theta(N+1) \\ \hline 0 \end{bmatrix} T_{N+1} + \theta_{N+1}(N+2) R_{N+1}, \end{aligned}$$

where $\theta_{N+1}(N+2)$ is given by (38). Note that elements of R_{N+1} and T_{N+1} are chosen so that the composite degree of $\tilde{\theta}_{N+1}(N+2)$ equals that of $\theta_{N+1}(N+1)$. Lemma 5 gives the restrictions on the nonsingular matrix R_{N+1} . Without loss of generality, column permutations will be ignored so that R_{N+1} must have non-zero diagonal elements, and there is a column degree equivalence between corresponding columns of $\tilde{\theta}_{N+1}(N+2)$ and $\theta_{N+1}(N+2)$. Moreover, for off-diagonal elements

$$(52) \quad \|R_{N+1}\|_{ji} \neq 0 \quad \text{only if } n_{N+1,j} \leq n_{N+1,i}, \quad 1 \leq i \neq j \leq m.$$

That is, only those columns of $\theta_{N+1}(N+2)$ for which the degree is not greater than $n_{N+1,i}$ can be combined with column i of $\theta_{N+1}(N+2)$ to form column i of $\theta_{N+1}(N+2)$. The matrix T_{N+1} has elements restricted for $1 \leq i, l \leq m, 0 \leq j \leq N$, by

$$(53) \quad \|T_{N+1}\|_{jm+l,i} \neq 0 \quad \text{only if } k_{jl} \leq n_{N+1,i},$$

where k_{jl} is the accumulation index relative to $N+1$. The linear combinations of (51) allow every variation of $\tilde{\theta}_{N+1}(N+2)$ that leaves the composite degree unchanged. Thus (51) under the restrictions of (52)–(53) specifies the equivalence class of minimal sets of augmented CAV's for $\{M_i\}_{N+1}$.

Now for the matrix $\theta_{N+1}(N+2)$ define the shift matrix $\sigma^{-1}[\theta_{N+1}(N+2)]$ whose columns consist of every distinct augmented CAV of length $(N+2)m$

associated with $P_{N+1}(z)$ except those in $\theta_{N+1}(N+2)$. Thus by Lemma 1 there are $N+1-n_{N+1,i}$ columns in $\sigma^{-1}[\theta_{N+1}(N+2)]$ for every column $p_{N+1,i}(z)$, $1 \leq i \leq m$. Let $\sigma^{-1}[\theta_{N+1}(N+2)]$ be the zero vector if every column of $P_{N+1}(z)$ has degree $N+1$.

Let $\Phi_{N+1,1}$, given by (46)–(47), be the subset of columns of $\Theta(N+2)$ that generates discrepancies forming a column basis for $\Delta_1(N+1) = \Delta(N+1)$ with minimum largest accumulation index relative to $N+1$. Define the shift matrix $\sigma^{-1}(\Phi_{N+1,1})$ whose columns are every distinct augmented CAV of length $(N+2)m$ associated with $B_{N+1,1}(z)$ with the last m rows zero and not in $\Phi_{N+1,1}$. Thus for every column $b_{N+1,1j}(z)$, $1 \leq j \leq \rho_{N+1,1}$, there are $N+1-k_{\phi,N+1,1j}$ linearly independent columns in $\sigma^{-1}(\Phi_{N+1,1})$; the columns of $\Phi_{N+1,1}$ and $\sigma^{-1}(\Phi_{N+1,1})$ are all linearly independent by construction. If the accumulation index is $N+1$ for every column of $\Phi_{N+1,1}$ or if $\Phi_{N+1,1} = 0$, let $\sigma^{-1}(\Phi_{N+1,1})$ be the zero vector.

LEMMA 6. *The columns of $\theta_{N+1}(N+2)$, $\sigma^{-1}[\theta_{N+1}(N+2)]$, $\Phi_{N+1,1}$ and $\sigma^{-1}(\Phi_{N+1,1})$ are a basis for the column space of $\Theta(N+2)$.*

First from the algorithm presented in § 3 it is readily verified that

$$\theta_N(N+2), \quad \sigma^{-1}[\theta_N(N+2)], \quad \begin{bmatrix} \Phi_{N,1} \\ -\Phi_{N,1} \\ 0 \end{bmatrix} \quad \text{and} \quad \sigma^{-1} \begin{bmatrix} \Phi_{N,1} \\ -\Phi_{N,1} \\ 0 \end{bmatrix}$$

are generated as linear combinations of

$$\sigma^{-1}[\theta_{N+1}(N+2)], \Phi_{N+1,1} \quad \text{and} \quad \sigma^{-1}(\Phi_{N+1,1}).$$

Reversing the algorithm it may be shown by successive induction that the column space of $\Theta(N+2)$ is spanned by columns of $\theta_{N+1}(N+2)$, $\sigma^{-1}[\theta_{N+1}(N+2)]$, $\Phi_{N+1,1}$ and $\sigma^{-1}(\Phi_{N+1,1})$. It remains to show that a mutual linear independence exists between the columns of these four matrices.

Since the last m rows of $\theta_{N+1}(N+2)$ form a nonsingular matrix, its columns are obviously independent of $[\Phi_{N+1,1} : \sigma^{-1}[\theta_{N+1}(N+2)] : \sigma^{-1}(\Phi_{N+1,1})]$. Let x and y be arbitrary with $x \in \mathcal{R}[\sigma^{-1}[\theta_{N+1}(N+2)] : \sigma^{-1}(\Phi_{N+1,1})]$ and $y \in \mathcal{R}[\Phi_{N+1,1}]$. Then $H(1, N+2)x = 0$, but $H(1, N+2)y = d \neq 0$ implying $x \neq y$. Thus the columns of $\Phi_{N+1,1}$ are linearly independent of $[\sigma^{-1}[\theta_{N+1}(N+2)] : \sigma^{-1}(\Phi_{N+1,1})]$. To show that columns of $\sigma^{-1}(\Phi_{N+1,1})$ and $\sigma^{-1}[\theta_{N+1}(N+2)]$ are mutually independent, assume that some element $t \in \mathcal{R}[\sigma^{-1}(\Phi_{N+1,1})]$ is a linear combination of columns in $\sigma^{-1}[\theta_{N+1}(N+2)]$. By the structure of the shift matrix $\sigma^{-1}(\Phi_{N+1,1})$, t can be shifted internally. This yields the contradictory implication that an element in $\mathcal{R}(\Phi_{N+1,1})$ is a linear combination of columns in $\sigma^{-1}[\theta_{N+1}(N+2)]$ and $\sigma^{-1}(\Phi_{N+1,1})$. The proof of Lemma 6 is now complete.

From Lemma 6 it is possible to rewrite (51) as

$$(54) \quad \begin{aligned} \tilde{\theta}_{N+1}(N+2) &= \theta_{N+1}(N+2)R_{N+1,1} + \sigma^{-1}[\theta_{N+1}(N+2)]R_{N+1,2} \\ &\quad + \Phi_{N+1,1}T_{N+1,1} + \sigma^{-1}(\Phi_{N+1,1})T_{N+1,2} \end{aligned}$$

with $R_{N+1,1} = R_{N+1}$ and appropriate restrictions on the elements of $R_{N+1,2}$, $T_{N+1,1}$ and $T_{N+1,2}$. But since an internally shifted CAV has an alternative representation as a CAP multiplied by the polynomial indeterminate to a non-negative power, (54) has the more compact polynomial representation of (50). If

column permutations are again ignored, $X_{N+1}(z)$ has nonzero diagonal elements; the elements of $X_{N+1}(z)$ are given for $1 \leq i, j \leq m$ by

$$(55) \quad x_{N+1,ji}(z) = \begin{cases} \sum_{l=0}^{n_{N+1,i} - n_{N+1,j}} x_{N+1,jil} z^{n_{N+1,i} - n_{N+1,j} - l}, & n_{N+1,i} \geq n_{N+1,j}, \\ 0, & n_{N+1,i} < n_{N+1,j}. \end{cases}$$

Elements of $Y_{N+1}(z)$ are given for $1 \leq j \leq \rho_{N+1,1}$, $1 \leq i \leq m$, by

$$(56) \quad y_{N+1,ji}(z) = \begin{cases} \sum_{l=0}^{n_{N+1,i} - k_{\phi N+1,1j}} y_{N+1,jil} z^{n_{N+1,i} - k_{\phi N+1,1j} - l}, & n_{N+1,i} \geq k_{\phi N+1,1j}, \\ 0, & n_{N+1,i} < k_{\phi N+1,1j}. \end{cases}$$

Suppose that the columns of $P_{N+1}(z)$ are ordered from the left by decreasing degree. It is then clear from (55) that $X_{N+1}(z)$ is block lower triangular with diagonal blocks that are nonsingular matrices of elements in \mathcal{F} . Hence $X_{N+1}(z)$ is an elementary matrix, and Theorem 5 is proved. \square

THEOREM 6. $P_{N+1}(z)$ is a unique (by Definition 6) minimal partial denominator matrix for $\{M_i\}_{N+1}$ if and only if

$$(57) \quad \min_{0 \leq j \leq N} \left\{ \min_{1 \leq i \leq m} \{k_{ji} | d_{ji} \neq 0\} \right\} > \max_{1 \leq i \leq m} \{n_{N+1,i}\},$$

where the accumulation indices k_{ji} are relative to $N+1$.

Proof. By Theorem 5, $X_{N+1}(z)$ is an elementary matrix, and by Lemma 6, no column of $B_{N+1,1}(z)$ is a polynomial combination of columns in $P_{N+1}(z)$. Hence $P_{N+1}(z)$ is unique by Definition 6 if and only if

$$(58) \quad Y_{N+1}(z) \equiv 0.$$

But inspection of (56) shows that (58) holds if and only if

$$(59) \quad \min_{1 \leq j \leq \rho_{N+1,1}} \{k_{\phi N+1,1j}\} > \max_{1 \leq i \leq m} \{n_{N+1,i}\}.$$

By the process of selection the columns of $\Phi_{N+1,1}$, (59) and (57) are equivalent conditions. This completes the proof of Theorem 6. \square

Now (50) with (55)–(56) enumerate the entire equivalence class Ω_{N+1} given the parameters required by the sequential realization algorithm. Also (58) provides a criterion for uniqueness. If (58) does not hold, $P_{N+1}(z)$ is not unique, and it may be desirable to examine the range of variation in the set of invariant factors over the elements in Ω_{N+1} . This task can be simplified considerably by considering

$$(60) \quad \bar{P}_{N+1}(z) = P_{N+1,0}^{-1} \tilde{P}_{N+1}(z) X_{N+1}^{-1}(z).$$

Since $P_{N+1,0}$ and $X_{N+1}(z)$ are elementary matrices, $\bar{P}_{N+1}(z)$ and $\tilde{P}_{N+1}(z)$ are equivalent and thus have the same set of invariant factors. Moreover, $\bar{P}_{N+1}(z)$ can be written

$$(61) \quad \bar{P}_{N+1}(z) = S_{N+1}(z) + \bar{B}_{N+1,1}(z) \bar{Y}_{N+1}(z),$$

where $S_{N+1}(z)$ is obtained from $P_{N+1}(z)$ by (10), $\bar{B}_{N+1,1}(z) = P_{N+1,0}^{-1} B_{N+1,1}(z)$ and $\bar{Y}_{N+1}(z) = Y_{N+1}(z) X_{N+1}^{-1}(z)$. Corresponding columns of $\bar{B}_{N+1,1}(z)$ and $B_{N+1,1}(z)$ have the same degree so that elements of $\bar{Y}_{N+1}(z)$ must obey the same constraints, viz. (56), as $Y_{N+1}(z)$. Every column of $B_{N+1,1}(z) Y_{N+1}(z)$ has degree strictly less than the corresponding column of $P_{N+1}(z) X_{N+1}(z)$ by construction; a similar relation holds between columns of $\bar{B}_{N+1,1}(z) \bar{Y}_{N+1}(z)$ and $S_{N+1}(z)$. Although (61) and (56) simplify the task of enumerating the invariant factor sets of a non-unique minimal partial denominator matrix, in general complicated, nonlinear, algebraic manipulations may be required as the following example serves to illustrate.

Example 2. Continuing with the sequence of Example 1, $P_4(z)$ is clearly not unique by Definition 6. It is the purpose of this example to enumerate all invariant factor sets for $P_4(z)$.

$$P_4(z) = \begin{bmatrix} z^3 + 3z^2 + 2z & -z^2 - z - 2 \\ -6z^2 - 6z & z^2 + z + 6 \end{bmatrix},$$

$$B_{4,1}(z) = \begin{bmatrix} -z & -1 \\ z + 1 & 1 \end{bmatrix}, \quad k_{\phi 1} = 2, \quad k_{\phi 2} = 3,$$

$$P_{4,0} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}, \quad P_{4,0}^{-1} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

$$S_4(z) = \begin{bmatrix} z^3 - 3z^2 - 4z & 4 \\ -6z^2 - 6z & z^2 + z + 6 \end{bmatrix}, \quad \bar{B}_{4,1}(z) = \begin{bmatrix} 1 & 0 \\ z + 1 & 1 \end{bmatrix}.$$

Then $\bar{Y}_4(z)$ has the form

$$\bar{Y}_4(z) = \begin{bmatrix} uz + v & w \\ t & 0 \end{bmatrix}$$

with parameters t, u, v and w . In general,

$$\bar{P}_4(z) = \begin{bmatrix} z^3 - 3z^2 + (u - 4)z + v & w + 4 \\ (u - 6)z^2 + (u + v - 6)z + (v + t) & z^2 + (w + 1)z + (w + 6) \end{bmatrix}.$$

Case 1. $w \neq -4$. Clearly $\gamma_{\bar{P}1} = 1$ and $\gamma_{\bar{P}2}$ is the characteristic equation. Thus

$$\gamma_{\bar{P}2} = z^5 + (w - 2)z^4 + (u - 2w - 1)z^3 + (v - 3u - w + 2)z^2 + (2u - 3v + 2w)z + 2v - t(w + 4).$$

If the desired characteristic equation is

$$\gamma_{\bar{P}2} = z^5 + \beta_1 z^4 + \beta_2 z^3 + \beta_3 z^2 + \beta_4 z + \beta_5,$$

then all but one of the coefficients may be independently specified;

$$\beta_4 = -(15\beta_1 + 7\beta_2 + 3\beta_3 + 31)$$

and

$$w = \beta_1 + 2, \quad t = \frac{1}{\beta_1 + 6}(14\beta_1 + 6\beta_2 + 2\beta_3 - \beta_5 + 30),$$

$$u = 2\beta_1 + \beta_2 + 5, \quad v = 7\beta_1 + 3\beta_2 + \beta_3 + 15.$$

Case 2. $w = -4$.

$$\bar{P}_4(z) = \begin{bmatrix} z^3 - 3z^2 + (u-4)z + v & 0 \\ (u-6)z^2 + (u+v-6)z + (v+t) & z^2 - 3z + 2 \end{bmatrix}.$$

(a) $t \neq 0, \gamma_{\bar{P}_1} = 1, \gamma_{\bar{P}_2} = [z^3 - 3z^2 + (u-4)z + v](z^2 - 3z + 2).$

(b) $t = 0, v = 2(6-u).$

$$\bar{P}_4(z) = \begin{bmatrix} [z^2 - z + (u-6)](z-2) & 0 \\ [(u-6)z + (u-6)](z-2) & (z-1)(z-2) \end{bmatrix},$$

$$\gamma_{\bar{P}_1} = z-2, \quad \gamma_{\bar{P}_2} = [z^2 - z + (u-6)](z^2 - 3z + 2).$$

(c) $t = 0, v = 6-u.$

$$\bar{P}_4(z) = \begin{bmatrix} [z^2 - 2z + (u-6)](z-1) & 0 \\ [(u-6)z + (u-6)](z-1) & (z-2)(z-1) \end{bmatrix},$$

$$\gamma_{\bar{P}_1} = z-1, \quad \gamma_{\bar{P}_2} = [z^2 - 2z + (u-6)](z^2 - 3z + 2).$$

(d) $t = v = 0, u = 6.$

$$\bar{P}_4(z) = \begin{bmatrix} z^3 - 3z^2 + 2z & 0 \\ 0 & z^2 - 3z + 2 \end{bmatrix},$$

$$\gamma_{\bar{P}_1} = z^2 - 3z + 2, \quad \gamma_{\bar{P}_2} = (z^2 - 3z + 2)z.$$

5. Discussion of the sequential realization method. The method of realizing linear constant dynamical systems derived from the sequential realization algorithm has much to recommend it over previous techniques. It may be used to construct linear realizations from transfer function matrices whose elements are ratios of not necessarily co-prime polynomials that need not be in factored form. The method is also useful in constructing a minimal linear model from a set of empirical measurements represented by a finite output data sequence. Hence it is a practical solution to the black box problem.

The sequential realization method is recursive so that the algorithm may be halted at any point in a matrix sequence with a minimal partial denominator matrix from which a minimal partial realization can be easily obtained. It is an improvement over Massey's algorithm because it is applicable to multivariable systems.

Rissanen's recursive realization procedure is essentially a variation of the Ho algorithm [6], [8] where in [10] and [11] the Hankel matrix of a finite sequence is factored as the product of a unit lower triangular matrix with a matrix whose bottom row is zero. The block symmetric structure of the Hankel matrix is exploited to yield a recursive factorization. In order to maintain this recursive

property, it is necessary that the row dimension of the Hankel matrix not exceed the column dimension and that the bottom row be linearly dependent. Thus Rissanen's approach will not always produce a minimal partial realization for an arbitrary finite sequence. For example, no minimal partial realization can be obtained using the method in [10] and [11] for the scalar sequence of length $N \geq 2$ given by $M_1 = M_2 = \cdots = M_{N-1} = 0$, $M_N = 1$. In contrast it has been shown that the sequential realization algorithm yields a minimal partial realization of any finite sequence. Another comparison to be made is based upon the amount of storage required if implementing the method on the computer. Since it is necessary to store and manipulate a Hankel matrix and its factors in order to execute Rissanen's algorithm, the required storage grows rapidly with the length of the sequence, particularly for matrix sequences. To use the sequential realization algorithm it is only necessary to store the given sequence, N , $P(z)$, $B(z)$, $\{n_i, 1 \leq i \leq m\}$, ρ , $\{k_{\beta j}$ and $n_{\beta j}, 1 \leq j \leq \rho\}$ and K .

The sequential realization method applies to the broad class of systems of (1)–(2) including finite state machines. It is a solution to the partial realization problem for arbitrary finite sequences including the zero sequence and the zero length sequence. Within this framework the initialization of $B(z)$ in the sequential realization algorithm is handled in a more natural way than by Massey [9].

Finally the sequential realization method provides a complete answer to an open question in Kalman [7]. He points out that the whole question of invariant factors, cyclicity and so forth of minimal partial realizations is entirely open. In the preceding section the uniqueness of a minimal partial realization was based upon the invariant factors of a minimal partial denominator matrix. Upon termination the parameters generated by the sequential realization algorithm yield the entire range of variation on the invariant factors of the minimal partial realizations. This certainly provides answers to questions like: Do there exist any minimal partial realizations for a given finite sequence having a cyclic system matrix? The sequence of Examples 1 and 2 has been used by Ackerman [1] to point out the superiority of the realization method in [2] over that of [12] because a coupling parameter results that is not available using the method in [12]. However, a complete enumeration of the available minimal partial realizations given in Example 2 reveals that the realization obtained by Ackerman in [1], although an improvement over Tether's, is still more restricted than it needs to be. In particular, Ackerman's realization corresponds to Case 1 in Example 2 with $t = 0$.

REFERENCES

- [1] J. E. ACKERMAN, *On partial realizations*, IEEE Trans. Automatic Control, AC-17 (1972), p. 381.
- [2] J. E. ACKERMAN AND R. S. BUCY, *Canonical minimal realization of a matrix of impulse response sequences*, Information and Control, 19 (1971), pp. 224–231.
- [3] B. M. ANDERSON, *Polynomial matrices—Applications to synthesis and analysis of linear multi-variable systems*, Doctoral dissertation, Northwestern University, Evanston, Ill., 1973.
- [4] B. M. ANDERSON AND F. M. BRASCH, JR., *The feedback synthesis of multivariable systems: A transfer function approach*, Proc. Sixteenth Midwest Symposium on Circuit Theory, Waterloo, Ontario, Canada, 1973.
- [5] F. R. GANTMACHER, *The Theory of Matrices*, vol. I, Chelsea, New York, 1959.

- [6] B. L. HO AND R. E. KALMAN, *Effective construction of linear state-variable models from input/output data*, Proc. Third Allerton Conf. on Circuit and System Theory, Monticello, Ill., 1965, pp. 449–459.
- [7] R. E. KALMAN, *On minimal partial realizations of a linear input/output map*, Aspects of Network and System Theory, R. E. Kalman and N. De Claris, eds., Holt, Rinehart and Winston, New York, 1971, pp. 385–407.
- [8] R. E. KALMAN, P. L. FALB AND M. A. ARBIB, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1969.
- [9] J. L. MASSEY, *Shift-register synthesis and BCH decoding*, IEEE Trans. Information Theory, IT-15 (1969), pp. 122–127.
- [10] J. RISSANEN, *Recursive identification of linear systems*, this Journal, 9 (1971), pp. 420–430.
- [11] J. RISSANEN AND T. KAILATH, *Partial realization of random systems*, Automatica, 8 (1972), pp. 389–396.
- [12] A. J. TETHER, *Construction of minimal linear state-variable models from finite input–output data*, IEEE Trans. Automatic Control, AC-15 (1970), pp. 427–436.
- [13] W. A. WOLOVICH, *The determination of state-space representations for linear multivariable systems*, Second IFAC Symposium on multivariable Technical Control Systems, Duesseldorf, Germany, 1971.

ASYMPTOTIC EXPANSIONS OF SINGULARLY PERTURBED QUASI-LINEAR OPTIMAL SYSTEMS*

PEDDAPULLAIAH SANNUTI†

Abstract. A class of two-point boundary value problems (TPBVP's) which arise in fixed final time free endpoint optimal control problems is considered. An asymptotic power series solution of the TPBVP is constructed with respect to a parameter whose perturbation changes the differential order of the problem. Based on a stability hypothesis, the proof of asymptotic correctness is accomplished through a successive approximation scheme.

1. Problem statement. Let us consider a quasi-linear system

$$(1.1a) \quad \dot{x} = g_1(x, t) + B_1(t)z + C_1(t)u, \quad x(0) = x_0,$$

$$(1.1b) \quad \lambda \dot{z} = G_1(x, t) + B_2(t)z + C_2(t)u, \quad z(0) = z_0,$$

where x and z are \mathbf{n} - and \mathbf{m} -dimensional state vectors respectively, u is an \mathbf{r} -dimensional control vector, and λ is a nonnegative scalar parameter. Note that for $\lambda = 0$, z ceases to be a state vector and the order of system (1.1) reduces to \mathbf{n} . The objective of the design is to find a control $u(t, \lambda)$ on the interval $[0, T]$, where T is a given number, such that the application of this control to the system (1.1) results in a trajectory $(x(t, \lambda), z(t, \lambda))$ which minimizes the cost functional

$$(1.2) \quad J = \int_0^T (V(x, t) + \frac{1}{2}u'R(t)u) dt.$$

Here the prime denotes transpose. The functions $g_1, G_1, B_1, B_2, C_1, C_2, V$ and R might have asymptotic series expansions in λ , but for simplicity we will take them independent of λ . It is assumed that these functions are infinitely differentiable in all their arguments in an appropriately defined domain. Further, for each t , $R(t)$ is a symmetric positive definite matrix and $V(x, t)$ is a positive semidefinite scalar function of x .

To obtain necessary conditions for an optimal control, we introduce the Hamiltonian

$$(1.3) \quad \begin{aligned} H(x, z, p, q, u, t) = & -V - \frac{1}{2}u'Ru + p'(g_1 + B_1z + C_1u) \\ & + q'(G_1 + B_2z + C_2u), \end{aligned}$$

where p and q are the costate variables corresponding to x and z respectively. Note that it is conventional to use λq as the costate variable; however, our use of q here is convenient for singular perturbation analysis. The Pontryagin maximum principle implies [17] that along an optimal trajectory

$$(1.4) \quad \nabla_u H = -Ru + C_1'p + C_2'q = 0,$$

* Received by the editors December 28, 1971, and in revised form March 7, 1974.

† Department of Electrical Engineering, Rutgers University, New Brunswick, New Jersey 08903. The research for this paper was supported in part by the National Science Foundation under Grant GK 4881.

while the costate variables p and q satisfy the equations

$$(1.5) \quad \begin{aligned} \dot{p} &= -\nabla_x H = \nabla_x V - g'_{1x}p - G'_{1x}q, \\ \lambda \dot{q} &= -\nabla_z H = -B'_1p - B'_2q, \end{aligned}$$

along with the final conditions

$$(1.6) \quad p(T) = 0, \quad q(T) = 0.$$

From (1.4),

$$(1.7) \quad u = R^{-1}(C'_1p + C'_2q),$$

so we can rewrite the state and costate equations (1.1) and (1.5) as

$$(1.8) \quad \begin{aligned} \dot{x} &= g_1 + S_1p + B_1z + Sq \equiv g(x, p, z, q, t), \\ \dot{p} &= \nabla_x V - g'_{1x}p - G'_{1x}q \equiv f(x, p, q, t), \\ \lambda \dot{z} &= G_1 + S'p + B_2z + S_2q \equiv G(x, p, z, q, t), \\ \lambda \dot{q} &= -B'_1p - B'_2q \equiv F(p, q, t), \end{aligned}$$

where S , S_1 and S_2 are the matrices

$$S(t) = C_1R^{-1}C'_2, \quad S_i(t) = C_iR^{-1}C'_i, \quad i = 1, 2.$$

The boundary conditions for (1.8) are

$$(1.9) \quad x(0) = x_0, \quad p(T) = 0, \quad z(0) = z_0, \quad q(T) = 0.$$

Equations (1.7)–(1.9) are necessary conditions for an optimal control. Under appropriate hypotheses these are also sufficient conditions [14]. We note that (1.8) and (1.9) define a two-point boundary value problem (TPBVP). For $\lambda = 0$ the differential order of the problem drops from $2(\mathbf{n} + \mathbf{m})$ to $2\mathbf{n}$. Thus we have a singularly perturbed TPBVP. The reduced problem obtained when $\lambda = 0$ is given by

$$(1.10a) \quad \dot{X}_0 = g_1(X_0, t) + S_1(t)P_0 + B_1(t)Z_0 + S(t)Q_0,$$

$$(1.10b) \quad \dot{P}_0 = \nabla_x V(X_0, t) - g'_{1x}(X_0, t)P_0 - G'_{1x}(X_0, t)Q_0,$$

$$(1.10c) \quad 0 = G_1(X_0, t) + S'(t)P_0 + B_2(t)Z_0 + S_2(t)Q_0,$$

$$(1.10d) \quad 0 = -B'_1(t)P_0 - B'_2(t)Q_0.$$

With the matrix B_2 invertible, we can solve for Z_0 and Q_0 from (1.10c) and (1.10d) and substitute them into (1.10a) and (1.10b), thus getting a $2\mathbf{n}$ th order problem for X_0 and P_0 with boundary conditions

$$(1.11) \quad X_0(0) = x_0, \quad P_0(T) = 0.$$

The reduced problem (1.10) and (1.11) can also be obtained by an alternative method. First one neglects λ in the state equations (1.1) directly and then obtains necessary conditions for an optimal control. This leads to a TPBVP which is identical to (1.10) and (1.11).

It is well known that the computational difficulties in obtaining an iterative solution of a TPBVP increase exponentially as the order of the problem increases [2], [18]. Another difficulty with (1.8) is that the standard methods of stepwise integration require the use of step sizes which are generally smaller than λ . Because of this, direct solution of (1.8) and (1.9) for small λ may be computationally prohibitive. Based on intuition, one could try to neglect λ and obtain the control from the solution of the reduced problem (1.10) and (1.11). However, such a control may result in unsatisfactory performance when applied to the full problem (1.1), since in general, the initial value $Z_0(0)$ of the reduced problem is completely different from the prescribed initial condition z_0 (for examples, see [19], [20]). Our interest in this paper is to determine asymptotic expansions for the optimal control by determining an asymptotic solution of the full problem (1.8) in terms of the positive parameter λ as λ tends to zero. We will calculate all the terms in the expansions using only lower order systems.

Kokotović and Sannuti [11] showed at first how singularly perturbed TPBVP's of the type (1.8) and (1.9) arise in control theory. Further, for a certain class of such problems, they gave a design procedure to calculate the first two terms of the expansions [11], [19], [20]. This procedure was based on the work of Vasil'eva [22] and Tupčiev [21]. Hadlock [5] gave then a formal expansion procedure generalizing in a natural way the earlier work. Recently, O'Malley [16] has considered asymptotic expansions for a linear state regulator problem. In this paper we give an explicit method of constructing the expansions and thus simplify the earlier methods. Furthermore, we prove the asymptotic validity of the expansions. The proof is basically an iterative scheme similar to those in [6], [8] and [12]. We follow closely the method of Hadlock [6] who considered a class of TPBVP's with a focus on sufficient conditions for the existence of a solution and for convergence of this solution to a given, known solution of the reduced problem. The analysis was in two main parts: first, a sequence of lemmas about changes of variables and convergence of successive approximations; second, a contraction mapping argument showing that the boundary conditions could be satisfied. We follow, in our proof of asymptotic validity, basically the same steps. Detailed definitions and discussions of asymptotic solutions can be found in Wasow [24].

2. Constructing an asymptotic solution. We shall be able to obtain an asymptotic representation of the solution to the problem (1.8) and (1.9) (in a natural form) under the following three hypotheses:

(H1) The reduced two-point boundary value problem (1.10) and (1.11) has a unique solution, (X_0, P_0, Z_0, Q_0) .

(H2) The matrix $B_2(t)$ is a stable matrix; i.e., all its m eigenvalues have negative real parts for each fixed t , $0 \leq t \leq T$.

(H3) The matrix $H_{xx}(t)$ is negative semidefinite for each fixed t , $0 \leq t \leq T$.

Here and elsewhere when functions are evaluated along the reduced solution (X_0, P_0, Z_0, Q_0) , we use the notation

$$g(t) = g(X_0(t), P_0(t), Z_0(t), Q_0(t), t),$$

$$g_x(t) = \frac{\partial g}{\partial x}(X_0(t), P_0(t), Z_0(t), Q_0(t), t), \quad \text{etc.}$$

Hypothesis (H2) plays several roles. It guarantees the solvability of algebraic equations of the type (1.10c) and (1.10d) for Z_0 and Q_0 . It rules out turning-point behavior [24]. Also, as we shall see, it defines the behavior in the boundary layers. Hypothesis (H3) implies that the solution of the reduced problem satisfies the Jacobi sufficient conditions for a local minimum. In particular, there are no conjugate points to T in the interval $[0, T)$ (see [1]). Since R is positive definite, the Legendre–Clebsch condition is always satisfied [2].

Based on previous experience [5], [7], [16], [24], we seek a solution to (1.8) and (1.9) of the form

$$(2.1) \quad \begin{aligned} x(t) &= X(t, \lambda) + \lambda m_1(\tau, \lambda) + \lambda n_1(\sigma, \lambda), \\ p(t) &= P(t, \lambda) + \lambda m_2(\tau, \lambda) + \lambda n_2(\sigma, \lambda), \\ z(t) &= Z(t, \lambda) + m_3(\tau, \lambda) + n_3(\sigma, \lambda), \\ q(t) &= Q(t, \lambda) + m_4(\tau, \lambda) + n_4(\sigma, \lambda), \end{aligned}$$

where τ and σ are stretched time coordinates

$$\tau = \frac{t}{\lambda}, \quad \sigma = \frac{T-t}{\lambda}.$$

Here the “outer solution” (X, P, Z, Q) will have an asymptotic power series which formally satisfies the system (1.8) throughout $0 \leq t \leq T$. Specifically, let

$$(2.2) \quad (X, P, Z, Q) \sim \sum_{j=0}^{\infty} (X_j(t), P_j(t), Z_j(t), Q_j(t)) \lambda^j \quad \text{as } \lambda \rightarrow 0,$$

where $(X_0(t), P_0(t), Z_0(t), Q_0(t))$ clearly must be the unique solution of the reduced problem. The outer solution alone will not, in general, asymptotically satisfy the boundary conditions (1.9). Thus, the terms $(\lambda m_1, \lambda m_2, m_3, m_4)$ in (2.1) represent the “boundary layer correction” at the initial point $t = 0$, and the terms $(\lambda n_1, \lambda n_2, n_3, n_4)$ represent the “boundary layer correction” at the final point $t = T$. They will have asymptotic power series expansions as $\lambda \rightarrow 0$,

$$(2.3) \quad m_i(\tau, \lambda) \sim \sum_{j=0}^{\infty} m_{ij}(\tau) \lambda^j, \quad n_i(\sigma, \lambda) \sim \sum_{j=0}^{\infty} n_{ij}(\sigma) \lambda^j, \quad i = 1 \text{ to } 4,$$

on the semiinfinite intervals $\tau \geq 0$ and $\sigma \geq 0$ respectively. The coefficients $m_{ij}(\tau)$ and $n_{ij}(\sigma)$ will be so chosen that (2.1) will formally (i.e., as a power series) satisfy the boundary conditions (1.9) asymptotically, and $m_{ij}(\tau)$ will tend to zero as the left stretched coordinate τ tends to infinity, and $n_{ij}(\sigma)$ will tend to zero as the right stretched coordinate σ tends to infinity,

$$(2.4) \quad m_{ij}(\tau) \rightarrow 0 \quad \text{as } \tau \rightarrow \infty, \quad n_{ij}(\sigma) \rightarrow 0 \quad \text{as } \sigma \rightarrow \infty, \quad i = 1 \text{ to } 4.$$

(We shall actually find that the m_{ij} and n_{ij} decay exponentially; i.e., they are of “boundary layer type” in the terminology of Višik and Lyusternik [23]). We remark here that (2.4) replaces the familiar matching or patching conditions used in the method of inner and outer expansions [10], [15].

Since x_0 and z_0 are independent of λ and since $m_i(\tau)$ at $t = T$ and $n_i(\sigma)$ at $t = 0$ are asymptotically negligible, the boundary conditions (1.9) are asymptotically equivalent to

$$(2.5a) \quad X_0(0) = x_0, \quad P_0(T) = 0,$$

$$(2.5b) \quad m_{30}(0) = z_0 - Z_0(0), \quad n_{40}(0) = -Q_0(T) = 0,$$

$$(2.5c) \quad X_j(0) = -m_{1j-1}(0), \quad P_j(T) = -n_{2j-1}(0) \quad \text{for all } j \geq 1,$$

$$(2.5d) \quad m_{3j}(0) = -Z_j(0), \quad n_{4j}(0) = -Q_j(T) \quad \text{for all } j \geq 1.$$

Note that whenever $m_{ij}(\tau)$ and $n_{ij}(\sigma)$ decay exponentially the solution (2.1) will be asymptotically equal to the outer solution in the open interval $0 < t < T$, i.e., for each integer $N \geq 0$,

$$\begin{aligned} x(t) &= \sum_{j=0}^N X_j(t)\lambda^j + O(\lambda^{N+1}), & p(t) &= \sum_{j=0}^N P_j(t)\lambda^j + O(\lambda^{N+1}), \\ z(t) &= \sum_{j=0}^N Z_j(t)\lambda^j + O(\lambda^{N+1}), & q(t) &= \sum_{j=0}^N Q_j(t)\lambda^j + O(\lambda^{N+1}), \end{aligned}$$

uniformly on each interval $\delta \leq t \leq T - \delta$ for any $\delta > 0$. As (2.5) shows, however, the initial values $m_{1j}(0)$ and $n_{2j}(0)$ must be calculated in order to determine the boundary values of the terms in the outer expansion (i.e., the asymptotic expansion of the outer solution).

The outer expansion is formally obtained by substituting (2.2) into (1.8) and equating coefficients of each λ^j . We have equality at $\lambda = 0$ by taking X_0, P_0, Z_0 and Q_0 to be the solution of the reduced problem. Equating first order coefficients, X_1, P_1, Z_1 and Q_1 must satisfy the linear system

$$(2.6a) \quad \dot{X}_1 = g_{1x}(t)X_1 + S_1(t)P_1 + B_1(t)Z_1 + S(t)Q_1,$$

$$(2.6b) \quad \dot{P}_1 = f_x(t)X_1 - g'_{1x}(t)P_1 - G'_{1x}(t)Q_1,$$

$$(2.6c) \quad 0 = G_{1x}(t)X_1 + S'(t)P_1 + B_2(t)Z_1 + S_2(t)Q_1 - \dot{Z}_0,$$

$$(2.6d) \quad 0 = -B'_1(t)P_1 - B'_2(t)Q_1 - \dot{Q}_0.$$

Since B_2 is nonsingular, (2.6c) and (2.6d) can be solved for Z_1 and Q_1 so that (2.6) can be rewritten in the form

$$(2.7a) \quad \dot{X}_1 = A_1(t)X_1 + A_2(t)P_1 + L_{10}(t),$$

$$(2.7b) \quad \dot{P}_1 = A_3(t)X_1 - A'_1(t)P_1 + L_{20}(t),$$

where

$$A_1(t) = g_{1x} - B_1B_2^{-1}G_{1x},$$

$$A_2(t) = (B_1B_2^{-1}C_2 - C_1)R^{-1}(B_1B_2^{-1}C_2 - C_1)',$$

$$A_3(t) = -H_{xx}(t) = f_x(t),$$

and the expressions for L_{10} and L_{20} follow from (2.6). The boundary conditions

for (2.7) are given by (2.5c),

$$(2.8) \quad X_1(0) = -m_{10}(0), \quad P_1(T) = -n_{20}(0).$$

Equations (2.7) and (2.8) represent a well-known TPBVP of a nonhomogeneous linear Hamiltonian system. We note that A_2 is positive semidefinite. Hypothesis (H3) assures that A_3 is also positive semidefinite. Then (2.7) and (2.8) has a unique solution [3], [9]. With

$$(2.9) \quad P_1(t) = -K(t)X_1(t) + \Pi_1(t),$$

this problem is reducible to a final value problem for an $n \times n$ dimensional symmetric matrix Riccati equation

$$(2.10) \quad \dot{K} = -KA_1 - A_1'K + KA_2K - A_3, \quad K(T) = 0,$$

and an n -dimensional linear vector equation

$$(2.11) \quad \dot{\Pi}_1 = (KA_2 - A_1')\Pi_1 + KL_{10} + L_{20}, \quad \Pi_1(T) = -n_{20}(0).$$

Since A_2 and A_3 are positive semidefinite matrices, $K(t)$ exists on $0 \leq t \leq T$ and is unique [3], [9]. Once the matrix K is known, the vector Π_1 is uniquely obtained from the final value problem (2.11). Then, the substitution of (2.9) in (2.7a) allows X_1 to be uniquely obtained as the solution of the initial value problem,

$$(2.12) \quad \dot{X}_1 = (A_1 - A_2K)X_1 + A_2\Pi_1 + L_{10}, \quad X_1(0) = -m_{10}(0).$$

Thus the variables X_1 , P_1 , Z_1 and Q_1 are uniquely obtained. In general, we find X_j , P_j , Z_j and Q_j by solving

$$(2.13) \quad \begin{aligned} \dot{X}_j &= g_{1x}(t)X_j + S_1(t)P_j + B_1(t)Z_j + S(t)Q_j + R_{1j-1}(t), \\ \dot{P}_j &= f_x(t)X_j - g'_{1x}(t)P_j - G'_{1x}(t)Q_j + R_{2j-1}(t), \\ 0 &= G_{1x}(t)X_j + S'(t)P_j + B_2(t)Z_j + S_2(t)Q_j + R_{3j-1}(t), \\ 0 &= -B'_1(t)P_j - B'_2(t)Q_j + R_{4j-1}(t), \end{aligned}$$

with boundary values given by (2.5c), where $R_{ij-1}(t)$, $i = 1$ to 4, are successively known in terms of X_r , P_r , Z_r and Q_r for all $r \leq j-1$. Note that (2.13) differs from (2.6) only in nonhomogeneous terms. Hence the solution can be easily obtained and requires solving only linear equations of the form (2.11) and (2.12). Thus the terms of the outer expansion can be generated recursively with only the boundary values $X_j(0)$ and $P_j(T)$ for $j \geq 1$ to be specified, (2.5c).

The following simple lemma will be repeatedly used to obtain the terms of the expansions for the boundary layer corrections.

LEMMA 1. Consider a linear time invariant system

$$(2.14) \quad \dot{\alpha} = -B\alpha + b \exp[-vt],$$

where b is an m -vector, B is an $m \times m$ matrix whose eigenvalues all have real parts less than $-\mu$ for some $\mu > 0$, and v is a positive parameter. Then there exists a unique α_0 such that the solution $\alpha(t)$ of (2.14) with the auxiliary condition $\alpha(0) = \alpha_0$ satisfies the inequality

$$|\alpha(t)| \leq k \exp[-vt] \quad \text{for all } t \geq 0,$$

where k depends only on v .¹

Proof. Integration of (2.14) implies

$$\alpha(t) = \exp[-Bt] \left(\alpha_0 + \int_0^\infty \exp[(B - Iv)s] b \, ds + \int_\infty^t \exp[(B - Iv)s] b \, ds \right).$$

Since B is a stable matrix, the first integral exists. Denote it by M_0 and choose $\alpha_0 = -M_0$. Then

$$\alpha(t) = \int_\infty^t \exp[B(s - t) - Ivs] b \, ds.$$

Observing the inequality [13]

$$|\exp[B(s - t)]| \leq k_0 \exp[-\mu(s - t)] \quad \text{for all } s \geq t,$$

we have

$$|\alpha(t)| \leq \int_t^\infty k_1 \exp[-\mu(s - t) - vs] \, ds \leq k \exp[-vt].$$

Now we proceed to calculate the boundary layer correction terms. Since the outer expansion (2.2) formally satisfies the system (1.8) and since the right boundary layer correction $n_i(\sigma)$ is asymptotically negligible near $t = 0$, (2.1) and (1.8) imply that the left boundary layer correction $m_i(\tau)$ must satisfy the system of equations

$$\begin{aligned} \frac{dm_1}{d\tau} &= \frac{dx}{dt} - \frac{dX}{dt} = \hat{g}(m_1, m_2, m_3, m_4, \lambda, \tau), \\ \frac{dm_2}{d\tau} &= \frac{dp}{dt} - \frac{dP}{dt} = \hat{f}(m_1, m_2, m_4, \lambda, \tau), \\ \frac{dm_3}{d\tau} &= \lambda \left(\frac{dz}{dt} - \frac{dZ}{dt} \right) = \hat{G}(m_1, m_2, m_3, m_4, \lambda, \tau), \\ \frac{dm_4}{d\tau} &= \lambda \left(\frac{dq}{dt} - \frac{dQ}{dt} \right) = -\lambda B'_1(\lambda\tau)m_2 - B'_2(\lambda\tau)m_4, \end{aligned} \tag{2.15}$$

where

$$\begin{aligned} \hat{g} &= g(X(\lambda\tau) + \lambda m_1(\tau), P(\lambda\tau) + \lambda m_2(\tau), Z(\lambda\tau) + m_3(\tau), Q(\lambda\tau) + m_4(\tau), \lambda\tau) \\ &\quad - g(X(\lambda\tau), P(\lambda\tau), Z(\lambda\tau), Q(\lambda\tau), \lambda\tau) \end{aligned}$$

and \hat{f} and \hat{G} are similarly defined.

Substituting the asymptotic power series for $m_i(\tau)$ into (2.15) and equating coefficients, we obtain a system of equations for the left boundary layer terms.

¹ We shall let the norm of a matrix or a vector A be denoted by $|A|$ and be equal to the sum of the magnitudes of all the elements of A .

In particular, the lowest order terms will satisfy the linear system

$$\begin{aligned}
 \frac{dm_{10}}{d\tau} &= B_1(0)m_{30} + S(0)m_{40}, \\
 \frac{dm_{20}}{d\tau} &= -G'_{1x}(0)m_{40}, \\
 \frac{dm_{30}}{d\tau} &= B_2(0)m_{30} + S_2(0)m_{40}, \\
 \frac{dm_{40}}{d\tau} &= -B'_2(0)m_{40}.
 \end{aligned}
 \tag{2.16}$$

Note from (2.5b), $m_{30}(0) = z_0 - Z_0(0)$. Using Lemma 1, we choose $m_{40}(0) = 0$, and obtain

$$\begin{aligned}
 m_{40}(\tau) &\equiv 0, & m_{30}(\tau) &= \exp[B_2(0)\tau](z_0 - Z_0(0)), \\
 m_{20}(\tau) &\equiv 0, & m_{10}(\tau) &= B_1(0)B_2^{-1}(0) \exp[B_2(0)\tau](z_0 - Z_0(0)).
 \end{aligned}
 \tag{2.17}$$

Thus the $m_{i0}(\tau)$ are uniquely defined as decaying exponentials. Further, (2.5c) gives

$$X_1(0) = -m_{10}(0) = -B_1(0)B_2^{-1}(0)(z_0 - Z_0(0)),$$

as was derived earlier [20] by another method.

In general, $m_{ij}(\tau)$ are obtained from (2.15) by successively equating like coefficients of λ^j ,

$$\begin{aligned}
 \frac{dm_{1j}}{d\tau} &= B_1(0)m_{3j} + S(0)m_{4j} + M_{1j-1}(\tau), \\
 \frac{dm_{2j}}{d\tau} &= -G'_{1x}(0)m_{4j} + M_{2j-1}(\tau), \\
 \frac{dm_{3j}}{d\tau} &= B_2(0)m_{3j} + S_2(0)m_{4j} + M_{3j-1}(\tau), \\
 \frac{dm_{4j}}{d\tau} &= -B'_2(0)m_{4j} + M_{4j-1}(\tau),
 \end{aligned}
 \tag{2.18}$$

where the terms $M_{ij-1}(\tau)$, $i = 1$ to 4, are known successively and satisfy (by induction)

$$M_{ij-1}(\tau) = O(\exp[-v\tau]) \quad \text{for some } v > 0.$$

Using Lemma 1, $m_{4j}(0)$ can be selected such that $m_{4j}(\tau)$ decays exponentially. $m_{3j}(0)$ must assume the initial condition $-Z_j(0)$. Since $B_2(0)$ is a stable matrix, it is evident that $m_{ij}(\tau)$, $i = 1$ to 4, can be uniquely defined by the condition that they decay exponentially as τ tends to infinity. Then (2.5c) gives $X_{j+1}(0) = -m_{1j}(0)$.

The right boundary layer correction terms $n_i(\sigma)$ are determined in an analogous manner. Since the $m_i(\tau)$ are asymptotically negligible at the final point

$t = T$, (2.1) and (1.8) imply that

$$\begin{aligned}
 \frac{dn_1}{d\sigma} &= \tilde{g}(n_1, n_2, n_3, n_4, \lambda, \sigma), \\
 \frac{dn_2}{d\sigma} &= \tilde{f}(n_1, n_2, n_4, \lambda, \sigma), \\
 \frac{dn_3}{d\sigma} &= \tilde{G}(n_1, n_2, n_3, n_4, \lambda, \sigma), \\
 \frac{dn_4}{d\sigma} &= \lambda B'_1(T - \lambda\sigma)n_2 + B'_2(T - \lambda\sigma)n_4,
 \end{aligned}
 \tag{2.19}$$

where

$$\begin{aligned}
 \tilde{g} &= g(X(t), P(t), Z(t), Q(t), t) \\
 &\quad - g(X(t) + \lambda n_1(\sigma), P(t) + \lambda n_2(\sigma), Z(t) + n_3(\sigma), Q(t) + n_4(\sigma), t)
 \end{aligned}$$

and t is replaced by $T - \lambda\sigma$. \tilde{f} and \tilde{G} are similarly defined. Substituting asymptotic power series and equating like coefficients of λ^j , we obtain a system of equations for $n_{ij}(\sigma)$. The lowest-order terms will satisfy

$$\begin{aligned}
 \frac{dn_{10}}{d\sigma} &= -B_1(T)n_{30} - S(T)n_{40}, \\
 \frac{dn_{20}}{d\sigma} &= G'_{1x}(T)n_{40}, \\
 \frac{dn_{30}}{d\sigma} &= -B_2(T)n_{30} - S_2(T)n_{40}, \\
 \frac{dn_{40}}{d\sigma} &= B'_2(T)n_{40}.
 \end{aligned}
 \tag{2.20}$$

The solution $n_{i0}(\sigma)$ should vanish as σ tends to infinity, and since (2.5b) dictates $n_{40}(0) = 0$, we have

$$n_{i0}(\sigma) \equiv 0, \quad i = 1 \text{ to } 4.
 \tag{2.21}$$

Thus (2.5c) implies $P_1(T) = 0$ as was derived earlier [20]. In general, the higher order terms $n_{ij}(\sigma)$ satisfy

$$\begin{aligned}
 \frac{dn_{1j}}{d\sigma} &= -B_1(T)n_{3j} - S(T)n_{4j} + N_{1j-1}(\sigma), \\
 \frac{dn_{2j}}{d\sigma} &= G'_{1x}(T)n_{4j} + N_{2j-1}(\sigma), \\
 \frac{dn_{3j}}{d\sigma} &= -B_2(T)n_{3j} - S_2(T)n_{4j} + N_{3j-1}(\sigma), \\
 \frac{dn_{4j}}{d\sigma} &= B'_2(T)n_{4j} + N_{4j-1}(\sigma),
 \end{aligned}
 \tag{2.22}$$

where, by induction, the exponentially decaying terms $N_{ij-1}(\sigma)$ are known successively. Equation (2.5d) dictates $n_{4j}(0) = -Q_j(T)$. Using Lemma 1, $n_{3j}(0)$ can be selected such that $n_{3j}(\sigma)$ decays exponentially. It is evident, then, that the terms $n_{ij}(\sigma)$ can be uniquely defined and that they decay exponentially as $\sigma \rightarrow \infty$. Equation (2.5c) then gives $P_{j+1}(T) = -n_{2j}(0)$.

This completes the construction of asymptotic expansions for the state and costate variables. To summarize, the reduced problem first determines (X_0, P_0, Z_0, Q_0) . Using (H2) and the reduced solution, the zero order boundary layer terms $(m_{i0}(\tau), n_{i0}(\sigma))$ are then obtained from (2.17) and (2.21), so the boundary conditions $X_1(0)$ and $P_1(T)$ are known from (2.5). Next the solution of (2.7) gives us (X_1, P_1, Z_1, Q_1) . Here (H3) is used. Thus proceeding recursively, the outer expansion terms (X_r, P_r, Z_r, Q_r) for all $r \leq j$ allow us to calculate the j th boundary layer terms $(m_{ij}(\tau), n_{ij}(\sigma))$. These boundary layer terms, in turn, allow us to calculate the $(j+1)$ th outer expansion terms $(X_{j+1}, P_{j+1}, Z_{j+1}, Q_{j+1})$. Finally, knowing the asymptotic expansions for the costate variables $p(t)$ and $q(t)$, the expansions for the optimal control $u(t)$ can easily be determined from (1.7). Since $m_{40}(\tau)$ and $n_{40}(\sigma)$ are identically zero, it is seen explicitly that the optimal control does not have any zero order boundary layers either at the initial point $t = 0$ or at the final point $t = T$ (i.e., the optimal control for the higher order problem will converge uniformly throughout $0 \leq t \leq T$ to that of the reduced problem as $\lambda \rightarrow 0$).

3. Asymptotic validity. Let us define the partial sums

$$\begin{aligned} ((X)_N, (P)_N, (Z)_N, (Q)_N) &= \sum_{j=0}^N (X_j(t), P_j(t), Z_j(t), Q_j(t))\lambda^j, \\ (m_i)_N &= \sum_{j=0}^{N-1} m_{ij}(\tau)\lambda^j, & (n_i)_N &= \sum_{j=0}^{N-1} n_{ij}(\sigma)\lambda^j, & i &= 1, 2, \\ (m_i)_N &= \sum_{j=0}^N m_{ij}(\tau)\lambda^j, & (n_i)_N &= \sum_{j=0}^N n_{ij}(\sigma)\lambda^j, & i &= 3, 4, \\ (x)_N &= (X)_N + \lambda(m_1)_N + \lambda(n_1)_N, & (p)_N &= (P)_N + \lambda(m_2)_N + \lambda(n_2)_N, \\ (z)_N &= (Z)_N + (m_3)_N + (n_3)_N, & (q)_N &= (Q)_N + (m_4)_N + (n_4)_N. \end{aligned}$$

THEOREM. For each integer $N \geq 0$, the two-point boundary value problem (1.8) and (1.9) under the hypotheses (H1)–(H3) has a solution which is of the form

$$(3.1) \quad \begin{aligned} x &= (x)_N + \lambda^N x^*, & p &= (p)_N + \lambda^N p^*, \\ z &= (z)_N + \lambda^N z^*, & q &= (q)_N + \lambda^N q^*, \end{aligned}$$

where x^*, p^*, z^* and q^* are all $O(\lambda)$ uniformly throughout $0 \leq t \leq T$ for λ sufficiently small.

Proof. Equations (1.8) and (3.1) imply

$$\frac{d(X)_N}{dt} + \frac{d(m_1)_N}{d\tau} - \frac{d(n_1)_N}{d\sigma} + \lambda^N \frac{dx^*}{dt} = g(x, p, z, q, t),$$

$$\begin{aligned}\frac{d(P)_N}{dt} + \frac{d(m_2)_N}{d\tau} - \frac{d(n_2)_N}{d\sigma} + \lambda^N \frac{dp^*}{dt} &= f(x, p, q, t), \\ \lambda \frac{d(Z)_N}{dt} + \frac{d(m_3)_N}{d\tau} - \frac{d(n_3)_N}{d\sigma} + \lambda^{N+1} \frac{dz^*}{dt} &= G(x, p, z, q, t), \\ \lambda \frac{d(Q)_N}{dt} + \frac{d(m_4)_N}{d\tau} - \frac{d(n_4)_N}{d\sigma} + \lambda^{N+1} \frac{dq^*}{dt} &= F(p, q, t).\end{aligned}$$

Moreover, by the construction of the outer solution (see (1.8), (1.10), (2.6) and (2.13)), we have

$$\begin{aligned}\frac{d(X)_N}{dt} &= g_1((X)_N, t) + S_1(t)(P)_N + B_1(t)(Z)_N + S(t)(Q)_N + O(\lambda^{N+1}), \\ \lambda \frac{d(Z)_N}{dt} &= G_1((X)_N, t) + S'(t)(P)_N + B_2(t)(Z)_N + S_2(t)(Q)_N + O(\lambda^{N+1}), \\ \frac{d(P)_N}{dt} &= \nabla_x V((X)_N, t) - g'_{1x}((X)_N, t)(P)_N - G'_{1x}((X)_N, t)(Q)_N + O(\lambda^{N+1}), \\ \lambda \frac{d(Q)_N}{dt} &= -B'_1(t)(P)_N - B'_2(t)(Q)_N + O(\lambda^{N+1}).\end{aligned}$$

Similarly, by the construction of the left boundary layer corrections (see (2.15), (2.16) and (2.18)), we have

$$\begin{aligned}\frac{d(m_1)_N}{d\tau} &= g_1((X)_N + \lambda(m_1)_N, \lambda\tau) + \lambda S_1(\lambda\tau)(m_2)_N + B_1(\lambda\tau)(m_3)_N + S(\lambda\tau)(m_4)_N \\ &\quad - g_1((X)_N, \lambda\tau) + O(\lambda^N \exp[-v\tau]), \\ \frac{d(m_2)_N}{d\tau} &= \nabla_x V((X)_N + \lambda(m_1)_N, \lambda\tau) - g'_{1x}((X)_N + \lambda(m_1)_N, \lambda\tau)((P)_N + \lambda(m_2)_N) \\ &\quad - G'_{1x}((X)_N + \lambda(m_1)_N, \lambda\tau)((Q)_N + (m_4)_N) - \nabla_x V((X)_N, \lambda\tau) \\ &\quad + g'_{1x}((X)_N, \lambda\tau)(P)_N + G'_{1x}((X)_N, \lambda\tau)(Q)_N + O(\lambda^N \exp[-v\tau]), \\ \frac{d(m_3)_N}{d\tau} &= G_1((X)_N + \lambda(m_1)_N, \lambda\tau) + \lambda S'(\lambda\tau)(m_2)_N + B_2(\lambda\tau)(m_3)_N + S_2(\lambda\tau)(m_4)_N \\ &\quad - G_1((X)_N, \lambda\tau) + O(\lambda^{N+1} \exp[-v\tau]), \\ \frac{d(m_4)_N}{d\tau} &= -\lambda B'_1(\lambda\tau)(m_2)_N - B'_2(\lambda\tau)(m_4)_N + O(\lambda^{N+1} \exp[-v\tau]).\end{aligned}$$

Finally, by the construction of the right boundary layer corrections (see (2.19), (2.20) and (2.22)), we have

$$\begin{aligned}\frac{d(n_1)_N}{d\sigma} &= -g_1((X)_N + \lambda(n_1)_N, T - \lambda\sigma) - \lambda S_1(T - \lambda\sigma)(n_2)_N - B_1(T - \lambda\sigma)(n_3)_N \\ &\quad - S(T - \lambda\sigma)(n_4)_N + g_1((X)_N, T - \lambda\sigma) + O(\lambda^N \exp[-v\sigma]),\end{aligned}$$

$$\begin{aligned}
\frac{d(n_2)_N}{d\sigma} &= -\nabla_x V((X)_N + \lambda(n_1)_N, T - \lambda\sigma) + g'_{1x}((X)_N + \lambda(n_1)_N, T - \lambda\sigma)((P)_N \\
&\quad + \lambda(n_2)_N) + G'_{1x}((X)_N + \lambda(n_1)_N, T - \lambda\sigma)((Q)_N + (n_4)_N) \\
&\quad + \nabla_x V((X)_N, T - \lambda\sigma) - g'_{1x}((X)_N, T - \lambda\sigma)(P)_N \\
&\quad - G'_{1x}((X)_N, T - \lambda\sigma)(Q)_N + O(\lambda^N \exp[-v\sigma]), \\
\frac{d(n_3)_N}{d\sigma} &= -G_1((X)_N + \lambda(n_1)_N, T - \lambda\sigma) - \lambda S'(T - \lambda\sigma)(n_2)_N - B_2(T - \lambda\sigma)(n_3)_N \\
&\quad - S_2(T - \lambda\sigma)(n_4)_N + G_1((X)_N, T - \lambda\sigma) + O(\lambda^{N+1} \exp[-v\sigma]), \\
\frac{d(n_4)_N}{d\sigma} &= \lambda B'_1(T - \lambda\sigma)(n_2)_N + B'_2(T - \lambda\sigma)(n_4)_N + O(\lambda^{N+1} \exp[-v\sigma]).
\end{aligned}$$

In all of the above equations, the O symbols hold for all t , $0 \leq t \leq T$. Further, the above equations imply that the remainders (x^*, p^*, z^*, q^*) will satisfy the system of equations

$$\begin{aligned}
\dot{x}^* &= g_{1x}(t)x^* + S_1(t)p^* + B_1(t)z^* + S(t)q^* + g^*(x^*, \lambda, t), \\
\dot{p}^* &= f_x(t)x^* - g'_{1x}(t)p^* - G'_{1x}(t)q^* + f^*(x^*, p^*, q^*, \lambda, t), \\
\lambda \dot{z}^* &= G_{1x}(t)x^* + S'(t)p^* + B_2(t)z^* + S_2(t)q^* + G^*(x^*, \lambda, t), \\
\lambda \dot{q}^* &= -B'_1(t)p^* - B'_2(t)q^* + F^*(\lambda, t),
\end{aligned} \tag{3.2}$$

where

$$\begin{aligned}
\lambda^N g^* &= g_1((X)_N + \lambda(m_1)_N + \lambda(n_1)_N + \lambda^N x^*, t) - g_1((X)_N + \lambda(m_1)_N, t) \\
&\quad - g_1((X)_N + \lambda(n_1)_N, t) + g_1((X)_N, t) - \lambda^N g_{1x}(t)x^* \\
&\quad + O(\lambda^{N+1} + \lambda^N \exp[-v\tau] + \lambda^N \exp[-v\sigma]), \\
\lambda^N G^* &= G_1((X)_N + \lambda(m_1)_N + \lambda(n_1)_N + \lambda^N x^*, t) - G_1((X)_N + \lambda(m_1)_N, t) \\
&\quad - G_1((X)_N + \lambda(n_1)_N, t) + G_1((X)_N, t) - \lambda^N G_{1x}(t)x^* + O(\lambda^{N+1}),
\end{aligned} \tag{3.3}$$

and f^* and F^* are similarly defined. We note that g^* , f^* , G^* and F^* satisfy the following two properties.

PROPERTY 1.

$$\begin{aligned}
|g^*(0, \lambda, t)| &\leq k_0(\exp[-vt/\lambda] + \exp[-v(T-t)/\lambda] + \lambda), \\
|f^*(0, 0, 0, \lambda, t)| &\leq k_0(\exp[-vt/\lambda] + \exp[-v(T-t)/\lambda] + \lambda), \\
|G^*(0, \lambda, t)| &\leq k_0\lambda, \\
|F^*(\lambda, t)| &\leq k_0\lambda,
\end{aligned}$$

where k_0 is a positive constant.

PROPERTY 2. For each $\delta > 0$, there exists an $\varepsilon(\delta) > 0$ such that for $|\hat{x}^*|, |\tilde{x}^*|$, etc., and $\lambda < \varepsilon$,

$$\begin{aligned}
|g^*(\hat{x}^*, \lambda, t) - g^*(\tilde{x}^*, \lambda, t)| &\leq \delta|\hat{x}^* - \tilde{x}^*|, \\
|f^*(\hat{x}^*, \hat{p}^*, \hat{q}^*, \lambda, t) - f^*(\tilde{x}^*, \tilde{p}^*, \tilde{q}^*, \lambda, t)| &\leq \delta(|\hat{x}^* - \tilde{x}^*| + |\hat{p}^* - \tilde{p}^*| + |\hat{q}^* - \tilde{q}^*|), \\
|G^*(\hat{x}^*, \lambda, t) - G^*(\tilde{x}^*, \lambda, t)| &\leq \delta|\hat{x}^* - \tilde{x}^*|.
\end{aligned}$$

Proof. After linearizing the system (3.3) around the solution of the reduced problem, observe that the product $(m_1)_N(n_1)_N$ is $O(\lambda^r)$ throughout $0 \leq t \leq T$ for r arbitrarily large, and then use the mean value theorem.

Boundary conditions (1.9) and equations (2.5) and (3.1) imply

$$(3.4) \quad \begin{aligned} x^*(0) &= -\lambda^{-(N-1)}(n_1)_N|_{t=0}, & z^*(0) &= -\lambda^{-N}(n_3)_N|_{t=0}, \\ p^*(T) &= -\lambda^{-(n-1)}(m_2)_N|_{t=T}, & q^*(T) &= -\lambda^{-N}(m_4)_N|_{t=T}. \end{aligned}$$

Since $m_i(\tau)$ and $n_i(\sigma)$ decay exponentially to zero away from $t = 0$ and $t = T$ respectively, we have

$$x^*(0) = O(\lambda^r), \quad p^*(T) = O(\lambda^r), \quad z^*(0) = O(\lambda^r), \quad q^*(T) = O(\lambda^r)$$

for r arbitrarily large.

Lemmas 2 and 3 provide some transformation matrices. Let us first make the abbreviations

$$\chi = \begin{bmatrix} x \\ p \end{bmatrix}, \quad y = \begin{bmatrix} z \\ q \end{bmatrix}, \quad \psi = \begin{bmatrix} g \\ f \end{bmatrix}, \quad \Phi = \begin{bmatrix} G \\ F \end{bmatrix}.$$

LEMMA 2. *There exists a nonsingular, continuously differentiable matrix $\Theta(t)$,*

$$\Theta(t) = \begin{bmatrix} I & E(t) \\ 0 & I \end{bmatrix},$$

such that

$$(3.5) \quad \Theta^{-1}(t)\Phi_y(t)\Theta(t) = \begin{bmatrix} B_2(t) & 0 \\ 0 & -B'_2(t) \end{bmatrix}.$$

Proof. It is shown in [20] that there exists a matrix $E(t)$ which satisfies (3.5).

LEMMA 3. *Let $A(t) = \psi_\chi(t) - \psi_y(t)\Phi_y^{-1}(t)\Phi_\chi(t)$. Then the linear matrix differential equation*

$$(3.6) \quad \dot{W} = A(t)W$$

has a $2n \times 2n$ nonsingular solution $W(t)$ on $[0, T]$ which when partitioned into $n \times n$ block matrices

$$W(t) = \begin{bmatrix} W_1(t) & W_2(t) \\ W_3(t) & W_4(t) \end{bmatrix}$$

satisfies $W_2(0) = 0$ and $W_3(T) = 0$. Further, $W_1(t)$ and $W_4(t)$ are nonsingular.

Proof. Partition $A(t)$ into $n \times n$ blocks

$$A(t) = \begin{bmatrix} A_1(t) & A_2(t) \\ A_3(t) & -A'_1(t) \end{bmatrix}$$

and note that A_1 , A_2 and A_3 are the matrices as defined in (2.7). A_2 and A_3 are positive semidefinite. By the proof of Theorem 1 in [3], H6 of [6] holds. Lemma 1 of [6] now yields the conclusion.

Next we modify (3.2) through a change of variables. Let

$$(3.7) \quad \begin{bmatrix} x^* \\ p^* \end{bmatrix} = W \begin{bmatrix} \omega \\ \pi \end{bmatrix}, \quad \begin{bmatrix} z^* \\ q^* \end{bmatrix} = \Theta \begin{bmatrix} \eta \\ \zeta \end{bmatrix} - \Gamma \begin{bmatrix} \omega \\ \pi \end{bmatrix},$$

where $\Gamma = \Phi_y^{-1} \Phi_x W$. Equation (3.2) transforms into

$$(3.8) \quad \begin{aligned} \dot{\omega} &= D_1(t)\eta + D_2(t)\zeta + h_1(\omega, \pi, \eta, \zeta, \lambda, t), \\ \dot{\pi} &= D_3(t)\eta + D_4(t)\zeta + h_2(\omega, \pi, \eta, \zeta, \lambda, t), \\ \lambda \dot{\eta} &= B_2(t)\eta + h_3(\omega, \pi, \eta, \zeta, \lambda, t), \\ \lambda \dot{\zeta} &= -B'_2(t)\zeta + h_4(\omega, \pi, \eta, \zeta, \lambda, t), \end{aligned}$$

where

$$\begin{bmatrix} D_1(t) & D_2(t) \\ D_3(t) & D_4(t) \end{bmatrix} = W^{-1}(t) \psi_y(t) \Theta(t),$$

$$\begin{bmatrix} h_1 \\ h_2 \end{bmatrix} = W^{-1} \begin{bmatrix} g^* \\ f^* \end{bmatrix},$$

and

$$(3.9) \quad \begin{bmatrix} h_3 \\ h_4 \end{bmatrix} = \Theta^{-1} \begin{bmatrix} G^* \\ F^* \end{bmatrix} + \lambda \Theta^{-1} \left[\dot{\Gamma} \begin{bmatrix} \omega \\ \pi \end{bmatrix} - \dot{\Theta} \begin{bmatrix} \eta \\ \zeta \end{bmatrix} + \Phi_y^{-1} \Phi_x \left(\psi_y \Theta \begin{bmatrix} \eta \\ \zeta \end{bmatrix} + \begin{bmatrix} g^* \\ f^* \end{bmatrix} \right) \right].$$

In view of the Properties 1 and 2, we see the h_i satisfy analogous Properties A and B.

PROPERTY A.

$$|h_i(0, 0, 0, 0, \lambda, t)| \leq k_1(\exp[-vt/\lambda] + \exp[-v(T-t)/\lambda] + \lambda), \quad i = 1, 2,$$

$$|h_j(0, 0, 0, 0, \lambda, t)| \leq k_1\lambda, \quad j = 3, 4.$$

PROPERTY B. For each $\delta > 0$, there exists an $\varepsilon > 0$ such that for $|\hat{\omega}|, |\tilde{\omega}|$, etc., and $\lambda < \varepsilon$,

$$\begin{aligned} & |h_i(\hat{\omega}, \hat{\pi}, \hat{\eta}, \hat{\zeta}, \lambda, t) - h_i(\tilde{\omega}, \tilde{\pi}, \tilde{\eta}, \tilde{\zeta}, \lambda, t)| \\ & \leq \delta(|\hat{\omega} - \tilde{\omega}| + |\hat{\pi} - \tilde{\pi}| + |\hat{\eta} - \tilde{\eta}| + |\hat{\zeta} - \tilde{\zeta}|), \quad i = 1, 2, \\ & |h_j(\hat{\omega}, \hat{\pi}, \hat{\eta}, \hat{\zeta}, \lambda, t) - h_j(\tilde{\omega}, \tilde{\pi}, \tilde{\eta}, \tilde{\zeta}, \lambda, t)| \\ & \leq \delta|\hat{\omega} - \tilde{\omega}| + k_1\lambda(|\hat{\pi} - \tilde{\pi}| + |\hat{\eta} - \tilde{\eta}| + |\hat{\zeta} - \tilde{\zeta}|), \quad j = 3, 4, \end{aligned}$$

and k_1 is a positive constant.

Let $\theta_1(t, s)$ and $\theta_2(t, s)$ be the fundamental matrices

$$\dot{\theta}_1(t, s) = \frac{1}{\lambda} B_2(t) \theta_1(t, s), \quad \theta_1(t, t) = I,$$

$$\dot{\theta}_2(t, s) = -\frac{1}{\lambda} B'_2(t) \theta_2(t, s), \quad \theta_2(t, t) = I.$$

Since $B_2(t)$ is a stable matrix, the following bounds can be established [4], [12]:

$$(3.10) \quad \begin{aligned} |\theta_1(t, s)| &\leq k_2 \exp[-v(t-s)/\lambda], & 0 \leq s \leq t \leq T, \\ |\theta_2(t, s)| &\leq k_2 \exp[-v(s-t)/\lambda], & 0 \leq t \leq s \leq T, \end{aligned}$$

where k_2 and v are positive constants.

Now consider (3.8) with auxiliary conditions

$$(3.11) \quad \omega(0) = a, \quad \pi(T) = b, \quad \eta(0) = c, \quad \zeta(T) = d.$$

Lemmas 4 and 5 are taken from Hadlock [6]. Necessary modifications are made so as to conform with our problem. We remark here that the λ factor on the right side of the inequality (for h_3 and h_4) in Property B is the key factor which leads us to the required modifications. (Note that $(\omega, \pi, \eta, \zeta)$ here stand for (π, ρ, η, ϕ) in [6]).

LEMMA 4. *There exist positive constants k and ε_1 such that if $|a|, |b|, |c|, |d|$ and λ are each $\leq \varepsilon_1$, then (3.8) and (3.11) has a solution for all t , $0 \leq t \leq T$, satisfying the bounds*

$$(3.12) \quad \begin{aligned} |\omega(t)| &\leq k(|a| + |b| + \lambda|c| + \lambda|d| + \lambda), \\ |\pi(t)| &\leq k(|a| + |b| + \lambda|c| + \lambda|d| + \lambda), \\ |\eta(t)| &\leq k(|a| + |b| + \lambda|c| + |c| \exp[-vt/2\lambda] + \lambda|d| + \lambda), \\ |\zeta(t)| &\leq k(|a| + |b| + \lambda|c| + \lambda|d| + |d| \exp[-v(T-t)/2\lambda] + \lambda). \end{aligned}$$

Proof. A successive approximation method as in Lemma 4 of [6] can be used. Successive iterates are defined by $\omega_0(t) = 0$, $\pi_0(t) = 0$, $\eta_0(t) = 0$, $\zeta_0(t) = 0$, and, for each integer $i \geq 0$,

$$\begin{aligned} \omega_{i+1}(t) &= a + \int_0^t [D_1(s)\eta_i(s) + D_2(s)\zeta_i(s) + h_1(\beta_i(s))] ds, \\ \pi_{i+1}(t) &= b + \int_T^t [D_3(s)\eta_i(s) + D_4(s)\zeta_i(s) + h_2(\beta_i(s))] ds, \\ \eta_{i+1}(t) &= \theta_1(t, 0)c + \frac{1}{\lambda} \int_0^t \theta_1(t, s)h_3(\beta_i(s)) ds, \\ \zeta_{i+1}(t) &= \theta_2(t, T)d + \frac{1}{\lambda} \int_T^t \theta_2(t, s)h_4(\beta_i(s)) ds, \end{aligned}$$

where to shorten notation, we have set

$$\beta_i(s) = (\omega_i(s), \pi_i(s), \eta_i(s), \zeta_i(s), \lambda, s).$$

Knowing the Properties A and B and the bounds on θ_1 and θ_2 , (3.10), and using the standard arguments, we find that the successive iterates are well-defined and that there exists a $\delta_1 > 0$ such that if $|a|, |b|, |c|, |d|$ and λ are each $\leq \varepsilon_1(\delta_1)$, the sequence converges uniformly to a solution of (3.8) and (3.11). Also, the estimates (3.12) follow from the iteration.

LEMMA 5. Let $(\tilde{\omega}(t), \tilde{\pi}(t), \tilde{\eta}(t), \tilde{\zeta}(t))$ and $(\hat{\omega}(t), \hat{\pi}(t), \hat{\eta}(t), \hat{\zeta}(t))$ be two different solutions of (3.8) along with the auxiliary conditions (3.13) and (3.14) respectively,

$$(3.13) \quad \tilde{\omega}(0) = a, \quad \tilde{\pi}(T) = b, \quad \tilde{\eta}(0) = \tilde{c}, \quad \tilde{\zeta}(T) = \tilde{d},$$

$$(3.14) \quad \hat{\omega}(0) = a, \quad \hat{\pi}(T) = b, \quad \hat{\eta}(0) = \hat{c}, \quad \hat{\zeta}(T) = \hat{d}.$$

Then, there exist positive constants K_0 and ε_0 such that whenever $|a|, |b|, |\tilde{c}|, |\hat{c}|, |\tilde{d}|, |\hat{d}|$ and λ are each $\leq \varepsilon_0$, the following estimates are satisfied:

$$(3.15) \quad \begin{aligned} |\tilde{\omega}(t) - \hat{\omega}(t)| &\leq K_0 \lambda (|\tilde{c} - \hat{c}| + |\tilde{d} - \hat{d}|), \\ |\tilde{\pi}(t) - \hat{\pi}(t)| &\leq K_0 \lambda (|\tilde{c} - \hat{c}| + |\tilde{d} - \hat{d}|), \\ |\tilde{\eta}(t) - \hat{\eta}(t)| &\leq K_0 \lambda (|\tilde{c} - \hat{c}| + |\tilde{d} - \hat{d}|) + K_0 \exp[-\nu t/2\lambda] |\tilde{c} - \hat{c}|, \\ |\tilde{\zeta}(t) - \hat{\zeta}(t)| &\leq K_0 \lambda (|\tilde{c} - \hat{c}| + |\tilde{d} - \hat{d}|) + K_0 \exp[-\nu(T-t)/2\lambda] |\tilde{d} - \hat{d}|. \end{aligned}$$

Proof. Again a successive approximation method can be used. Following [6], we calculate the bounds on the i th iterates $|\tilde{\omega}_i(t) - \hat{\omega}_i(t)|$, etc., and then show, by choosing K_0 and δ_0 properly, that the limits of these bounds, as $i \rightarrow \infty$, yield the estimates (3.15).

We will now calculate the parameters a, b, c and d such that the solution $(\omega, \pi, \eta, \zeta)$ of (3.8) (in terms of the original variables (x^*, p^*, z^*, q^*)) satisfies (3.2) along with the boundary conditions (3.4). Noting $W_2(0) = 0$ and $W_3(T) = 0$, equations (3.7) and (3.11) imply

$$x^*(0) = W_1(0)a, \quad p^*(T) = W_4(T)b.$$

Thus,

$$(3.16) \quad a = W_1^{-1}(0)x^*(0) = O(\lambda^r), \quad b = W_4^{-1}(T)p^*(T) = O(\lambda^r).$$

Writing $\omega(t) = \omega(c, d, \lambda, t)$, etc., equations (3.7) and (3.11) imply

$$z^*(0) = c + E(0)\zeta(c, d, \lambda, 0) - \Gamma_1(0)a(\lambda) - \Gamma_2(0)\pi(c, d, \lambda, 0)$$

and

$$q^*(T) = d - \Gamma_3(T)\omega(c, d, \lambda, T) - \Gamma_4(T)b(\lambda),$$

where

$$\Gamma(t) = \begin{bmatrix} \Gamma_1(t) & \Gamma_2(t) \\ \Gamma_3(t) & \Gamma_4(t) \end{bmatrix} = \Phi_y^{-1}(t)\Phi_x(t)W(t).$$

Thus c and d are given by

$$(3.17) \quad \begin{aligned} c &= z^*(0) - E(0)\zeta(c, d, \lambda, 0) + \Gamma_1(0)a(\lambda) + \Gamma_2(0)\pi(c, d, \lambda, 0) \equiv \Omega_1(c, d, \lambda), \\ d &= q^*(T) + \Gamma_3(T)\omega(c, d, \lambda, T) + \Gamma_4(T)b(\lambda) \equiv \Omega_2(c, d, \lambda). \end{aligned}$$

We need to show that c and d satisfying (3.17) are of $O(\lambda)$. Let us make the abbreviations

$$\mathbf{c} = \begin{bmatrix} c \\ d \end{bmatrix}, \quad \Omega(\mathbf{c}, \lambda) = \begin{bmatrix} \Omega_1(c, d, \lambda) \\ \Omega_2(c, d, \lambda) \end{bmatrix}.$$

Lemma 5 and (3.17) imply

$$|\Omega(\tilde{\mathbf{c}}, \lambda) - \Omega(\hat{\mathbf{c}}, \lambda)| \leq \lambda K_0(|E(0)| + |\Gamma_2(0)| + |\Gamma_3(T)|)(|\tilde{\mathbf{c}} - \hat{\mathbf{c}}| + |\tilde{\mathbf{d}} - \hat{\mathbf{d}}|) \\ + K_0|E(0)| \exp[-\nu T/2\lambda]|\tilde{\mathbf{d}} - \hat{\mathbf{d}}|,$$

whenever $|\tilde{\mathbf{c}}|, |\hat{\mathbf{c}}|$ and λ are all sufficiently small. This shows that Ω is a contraction mapping for λ sufficiently small. From (3.16) and Lemma 4, it follows that

$$(3.18) \quad |\Omega(0, \lambda)| = O(\lambda).$$

Using (3.18) and the triangle inequality, it is easy to show that $\Omega(\mathbf{c}, \lambda)$ maps a domain $\mathcal{D} = \{\mathbf{c} | |\mathbf{c}| \leq \varepsilon_2\}$ into itself for λ sufficiently small. This shows that Ω has a unique fixed point in \mathcal{D} . Let the fixed point be \mathbf{c}^* . Then we have

$$(3.19) \quad |\mathbf{c}^* - \Omega(0, \lambda)| = |\Omega(\mathbf{c}^*, \lambda) - \Omega(0, \lambda)| \leq \alpha|\mathbf{c}^*|,$$

where α is the contraction mapping constant. Now (3.18) and (3.19) imply

$$|\mathbf{c}^*| \leq \frac{1}{1-\alpha}|\Omega(0, \lambda)| = O(\lambda).$$

Thus when a, b, c and d are all of $O(\lambda)$, Lemma 4 implies that $(\omega, \pi, \eta, \zeta)$ and hence (x^*, p^*, z^*, q^*) are all of $O(\lambda)$. This completes the proof of the theorem.

4. Example. Let

$$\dot{x} = z + \lambda x^2, \quad x(0) = x_0, \\ \lambda \dot{z} = -x - z + u, \quad z(0) = z_0,$$

and let the cost functional to be minimized be

$$J = \frac{1}{2} \int_0^1 (x^2 + u^2) dt.$$

Defining the Hamiltonian,

$$H = -\frac{1}{2}(x^2 + u^2) + p(z + \lambda x^2) + q(-x - z + u),$$

we obtain the necessary conditions for an optimal control,

$$(4.1) \quad \begin{aligned} \dot{x} &= z + \lambda x^2, & x(0) &= x_0, \\ \dot{p} &= x - 2\lambda xp + q, & p(1) &= 0, \\ \lambda \dot{z} &= -x - z + q, & z(0) &= z_0, \\ \lambda \dot{q} &= -p + q, & q(1) &= 0, \\ u &= q. \end{aligned}$$

The reduced problem,

$$\begin{aligned} \dot{X}_0 &= Z_0, & X_0(0) &= x_0, \\ \dot{P}_0 &= X_0 + Q_0, & P_0(1) &= 0, \\ 0 &= -X_0 - Z_0 + Q_0, \\ 0 &= -P_0 + Q_0, \\ U_0 &= Q_0, \end{aligned}$$

has the solution

$$\begin{aligned}X_0(t) &= x_0(a_1 \exp[\sqrt{2}t] + a_2 \exp[-\sqrt{2}t]), \\P_0(t) &= x_0(a_3 \exp[\sqrt{2}t] - a_4 \exp[-\sqrt{2}t]), \\Z_0(t) &= P_0(t) - X_0(t), \\Q_0(t) &= U_0(t) = P_0(t),\end{aligned}$$

where

$$\begin{aligned}a_1 &= \frac{1}{2\sqrt{2}}(\sqrt{2} - 1 - a_5), & a_2 &= \frac{1}{2\sqrt{2}}(\sqrt{2} + 1 + a_5), \\a_3 &= \frac{1}{2\sqrt{2}}(1 - (\sqrt{2} + 1)a_5), & a_4 &= \frac{1}{2\sqrt{2}}(1 + (\sqrt{2} - 1)a_5),\end{aligned}$$

and

$$a_5 = (1 + \sqrt{2} \coth \sqrt{2})^{-1}.$$

The left boundary layer equations are

$$\begin{aligned}\frac{dm_1}{d\tau} &= m_3 + 2\lambda^2 X m_1 + \lambda^3 m_1^2, \\ \frac{dm_2}{d\tau} &= \lambda m_1 + m_4 - 2\lambda^2 (P m_1 + X m_2 + \lambda m_1 m_2), \\ \frac{dm_3}{d\tau} &= -\lambda m_1 - m_3 + m_4, \\ \frac{dm_4}{d\tau} &= -\lambda m_2 + m_4.\end{aligned}$$

The leading terms then satisfy

$$\begin{aligned}\frac{dm_{10}}{d\tau} &= m_{30}, & \frac{dm_{20}}{d\tau} &= m_{40}, \\ \frac{dm_{30}}{d\tau} &= -m_{30} + m_{40}, & \frac{dm_{40}}{d\tau} &= m_{40}.\end{aligned}$$

The auxiliary conditions are that $m_{30}(0) = z_0 - Z_0(0)$ and that the solution decays exponentially. Thus

$$\begin{aligned}m_{10}(\tau) &= -(z_0 - Z_0(0)) \exp[-\tau], & m_{20}(\tau) &\equiv 0, \\ m_{30}(\tau) &= (z_0 - Z_0(0)) \exp[-\tau], & m_{40}(\tau) &\equiv 0.\end{aligned}$$

The initial condition for $X_1(0)$ is

$$X_1(0) = -m_{10}(0) = z_0 - Z_0(0) = (1 + a_5)x_0 + z_0.$$

Analogously, the right boundary layer equations are

$$\begin{aligned}\frac{dn_1}{d\sigma} &= -n_3 - 2\lambda^2 X n_1 - \lambda^3 n_1^2, \\ \frac{dn_2}{d\sigma} &= -\lambda n_1 - n_4 + 2\lambda^2 (P n_1 + X n_2 + \lambda n_1 n_2), \\ \frac{dn_3}{d\sigma} &= \lambda n_1 + n_3 - n_4, \\ \frac{dn_4}{d\sigma} &= \lambda n_2 - n_4.\end{aligned}$$

These equations imply

$$n_{10}(\sigma) = n_{20}(\sigma) = n_{30}(\sigma) = n_{40}(\sigma) \equiv 0,$$

so that $P_1(1) = 0$.

Applying the theorem, we conclude that the true solution of (4.1) within $O(\lambda)$ for all t , $0 \leq t \leq 1$, is approximated by

$$(x)_0 = X_0(t), \quad (p)_0 = P_0(t), \quad (z)_0 = Z_0(t) + m_{30}(t/\lambda), \quad (u)_0 = (q)_0 = Q_0(t).$$

We will now determine the next approximation. The first order terms of the outer expansion are defined by

$$\begin{aligned}\dot{X}_1 &= Z_1 + X_0^2, & \dot{P}_1 &= X_1 + Q_1 - 2X_0 P_0, \\ 0 &= -X_1 - Z_1 + Q_1 - \dot{Z}_0, & 0 &= -P_1 + Q_1 - \dot{Q}_0,\end{aligned}$$

and

$$X_1(0) = (1 + a_5)x_0 + z_0, \quad P_1(1) = 0.$$

Thus

$$\begin{aligned}X_1(t) &= c_1 + (c_2 + \sqrt{2} x_0 a_1 t) \exp[\sqrt{2} t] + (c_3 - \sqrt{2} x_0 a_2 t) \exp[-\sqrt{2} t] \\ &\quad + c_4 \exp[2\sqrt{2} t] + c_5 \exp[-2\sqrt{2} t], \\ P_1(t) &= c_6 + (c_7 + \sqrt{2} x_0 a_3 t) \exp[\sqrt{2} t] + (c_8 + \sqrt{2} x_0 a_4 t) \exp[-\sqrt{2} t] \\ &\quad + c_9 \exp[2\sqrt{2} t] + c_{10} \exp[-2\sqrt{2} t], \\ Z_1(t) &= P_0(t) + P_1(t) - (X_0(t) + X_1(t)), \\ Q_1(t) &= P_0(t) + P_1(t) + X_0(t),\end{aligned}$$

where the c_i are constants which depend on x_0 and z_0 . Further, the first order terms of the left boundary layer are given by

$$\begin{aligned}\frac{dm_{11}}{d\tau} &= m_{31}, & \frac{dm_{21}}{d\tau} &= m_{41} + m_{10}, \\ \frac{dm_{31}}{d\tau} &= -m_{31} + m_{41} - m_{10}, & \frac{dm_{41}}{d\tau} &= m_{41} - m_{20},\end{aligned}$$

with the auxiliary condition $m_{31}(0) = -Z_1(0)$. Thus,

$$m_{11}(\tau) = (Z_1(0) + Z_0(0) - z_0 - (z_0 - Z_0(0))\tau) \exp[-\tau],$$

$$m_{21}(\tau) = (z_0 - Z_0(0)) \exp[-\tau],$$

$$m_{31}(\tau) = ((z_0 - Z_0(0))\tau - Z_1(0)) \exp[-\tau],$$

$$m_{41}(\tau) \equiv 0.$$

This gives $X_2(0) = -m_{11}(0) = -(Z_1(0) + Z_0(0) - z_0)$. Analogously, we find

$$n_{11}(\sigma) = n_{31}(\sigma) = -\frac{1}{2}Q_1(1) \exp[-\sigma],$$

$$n_{21}(\sigma) = n_{41}(\sigma) = -Q_1(1) \exp[-\sigma].$$

This gives $P_2(1) = -n_{21}(0) = Q_1(1)$.

Now applying the theorem, we conclude that the true solution of (4.1) within $O(\lambda^2)$ for all t , $0 \leq t \leq 1$, is approximated by

$$(x)_1 = X_0(t) + \lambda X_1(t) + \lambda m_{10}(t/\lambda) + \lambda n_{10}((1-t)/\lambda),$$

$$(p)_1 = P_0(t) + \lambda P_1(t) + \lambda m_{20}(t/\lambda) + \lambda n_{20}((1-t)/\lambda),$$

$$(z)_1 = Z_0(t) + \lambda Z_1(t) + m_{30}(t/\lambda) + \lambda m_{31}(t/\lambda) + n_{30}((1-t)/\lambda) + \lambda n_{31}((1-t)/\lambda),$$

$$(q)_1 = Q_0(t) + \lambda Q_1(t) + m_{40}(t/\lambda) + \lambda m_{41}(t/\lambda) + n_{40}((1-t)/\lambda) + \lambda n_{41}((1-t)/\lambda),$$

$$(u)_1 = (q)_1.$$

The higher order approximations can be determined in an analogous manner.

Acknowledgment. The author is thankful to Professor R. E. O'Malley, Jr. for helpful discussions. He is also thankful to an unknown referee whose comments helped him to improve this paper.

REFERENCES

- [1] J. V. BREAKWELL AND Y. C. HO, *On the conjugate point condition for the control problem*, Internat. J. Engrg. Sci., 2 (1965), pp. 565-579.
- [2] A. E. BRYSON, JR. AND Y. C. HO, *Applied Optimal Control*, Blaisdell, Waltham, Mass., 1969.
- [3] R. S. BUCY, *Two-point boundary value problem of linear Hamiltonian systems*, SIAM J. Appl. Math., 15 (1967), pp. 1385-1389.
- [4] L. FLATTO AND N. LEVINSON, *Periodic solutions of singularly perturbed systems*, J. Rat. Mech. Anal., 4 (1955), pp. 943-950.
- [5] C. R. HADLOCK, *Singular perturbations of a class of two-point boundary value problems arising in optimal control*, Rep. R-481, Coordinated Science Laboratory, Univ. of Illinois, Urbana, 1970.
- [6] ———, *On a class of singularly perturbed two point boundary value problems*, Proc. Eighth Annual Allerton Conference on Circuit and System Theory, 1970, pp. 331-339. See also, *Existence and dependence on a parameter of solutions of a nonlinear two-point boundary value problem*, J. Differential Equations, 14 (1973), pp. 498-517.
- [7] W. A. HARRIS, JR., *Singular perturbations of a boundary value problem for a nonlinear system of differential equations*, Duke Math. J., 29 (1962), pp. 429-445.
- [8] F. C. HOPPENSTEADT, *Properties of solutions of ordinary differential equations with small parameters*, Comm. Pure Appl. Math., 24 (1971), pp. 807-840.
- [9] R. E. KALMAN, P. L. FALB AND M. A. ARBIB, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1969.

- [10] S. KAPLUN, *Fluid Mechanics and Singular Perturbations*, P. A. LAGERSTROM, L. N. HOWARD AND C. S. LIU, eds., Academic Press, New York, 1967.
- [11] P. V. KOKOTOVIĆ AND P. SANNUTI, *Singular perturbation method for reducing the model order in optimal control design*, IEEE Trans. Automatic Control, 13 (1968), pp. 377–384.
- [12] J. J. LEVIN, *Singular perturbations of nonlinear systems of differential equations related to conditional stability*, Duke Math. J., 23 (1956), pp. 609–620.
- [13] J. J. LEVIN AND N. LEVINSON, *Singular perturbations of nonlinear systems of differential equations and associated boundary layer equation*, J. Rat. Mech. Anal., 3 (1954), pp. 247–270.
- [14] O. L. MANGASARIAN, *Sufficient conditions for the optimal control of nonlinear systems*, this Journal, 4 (1966), pp. 139–152.
- [15] R. E. O'MALLEY, JR., *Topics in singular perturbations*, Advances in Math., 2 (1968), pp. 365–470. Reprinted in Lectures on Ordinary Differential Equations, R. W. MCKELVEY, ed., Academic Press, New York, 1970, pp. 155–260.
- [16] ———, *The singularly perturbed linear state regulator problem*, this Journal, 19 (1972), pp. 399–413.
- [17] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.
- [18] A. P. SAGE, *Optimum Systems Control*, Prentice-Hall, Englewood Cliffs, N.J., 1968.
- [19] P. SANNUTI, *Singular perturbation method in the theory of optimal control*, Rep. R-379, Co-ordinated Science Laboratory, Univ. of Illinois, Urbana, 1968.
- [20] P. SANNUTI AND P. V. KOKOTOVIĆ, *Singular perturbation method for near optimum design of high-order nonlinear systems*, Automatica, 5 (1969), pp. 773–779.
- [21] V. A. TUPČIEV, *Asymptotic behavior of the solution of a boundary problem for systems of differential equations of first order with a small parameter in the derivative*, Soviet Math. Dokl., 3 (1962), pp. 612–616.
- [22] A. B. VASIL'EVA, *Asymptotic behavior of solutions to certain problems involving nonlinear differential equations containing a small parameter multiplying the highest derivatives*, Russian Math. Surveys, 18 (1963), no. 3, pp. 13–81.
- [23] M. I. VIŠIK AND L. A. LYUSTERNIK, *Regular degeneration and boundary layer for linear differential equations with a small parameter*, Uspehi Mat. Nauk, 12 (1957), pp. 3–122; English transl., Amer. Math. Soc. Transl., Ser. 2, 20 (1961), pp. 239–364.
- [24] W. WASOW, *Asymptotic Expansions for Ordinary Differential Equations*, Interscience, New York, 1965.

CIRCLE CRITERIA FOR STABILITY IN HILBERT SPACE*

P. A. COOK†

Abstract. Sufficient conditions are obtained for the stability, in the L^2 sense, of dynamical systems whose inputs and outputs are in Hilbert space. The conditions generalize the circle criterion for the stability of feedback systems.

1. Introduction. This paper is concerned with the generalization of certain already established stability criteria for feedback systems to a wider context. The criteria in question are of the form of the "circle criterion" [1] for a closed-loop system consisting of two components, one being linear and time-invariant while the other can be nonlinear, time-dependent and even history-dependent, but has its response confined within a certain sector. The stability criterion is there expressed in terms of the behavior of the Nyquist plot of the linear, time-invariant part, with respect to a disc in the complex plane corresponding to the limitations on the response of the other component. Generalizations of this criterion have already been made, by Rosenbrock [2] and the present author [3], to the case of multi-input, multi-output systems in which the transfer-function matrix of the linear, time-invariant part has rational elements; we are concerned here with a further generalization to the case where the transfer-functions can be nonrational and the input and output spaces can be infinite-dimensional (more specifically, Hilbert space). There is thus a connection with the Hilbert space stability theory of Freedman, Falb and Zames [4], but the present work is in one sense more, and in another sense less, general than theirs; no assumption is made of normality or approximability for the transfer-function operator of the linear, time-invariant part, and more information about the nonlinear, time-dependent part can be utilized, but the discussion is confined to continuous-time systems, though this is probably not essential. Since there is, in general, no finite-dimensional state space, we work entirely in terms of the input and output variables.

The main results of the paper are embodied in Theorems 3–5, which generalize the author's previous work [3] on finite-dimensional systems.

2. Definitions and preliminaries. We begin by listing various definitions used in the paper, and commenting on them where appropriate.

2.1. Linear spaces. H is a *real separable Hilbert space*, with *inner product*

$$\langle a, b \rangle = \langle b, a \rangle$$

and *norm*

$$\|a\| = \sqrt{\langle a, a \rangle}$$

* Received by the editors July 24, 1973, and in revised form January 28, 1974.

† Department of Mathematics, Loughborough University of Technology, Loughborough, Leicestershire, United Kingdom. Now at Control Systems Centre, University of Manchester Institute of Science and Technology, Manchester, United Kingdom. This work was supported by the Science Research Council under Grant B/SR/8561.

for $a, b \in H$. H^c is its complexification, with inner product

$$\langle a, b \rangle = \langle b, a \rangle^*$$

where $*$ means complex conjugate, and norm

$$\|a\| = \sqrt{\langle a, a \rangle}$$

for $a, b \in H^c$. With respect to an orthonormal basis $\{e_j\}$ of H , where j ranges over an appropriate index set, an element $a \in H$ (H^c) is represented by a real (complex) column vector

$$a = [a_j] = [\langle e_j, a \rangle].$$

$L^2(H)$ is the space of H -valued functions of time t which are Lebesgue square-integrable on $(-\infty, \infty)$. With inner product

$$\langle u, v \rangle = \int_{-\infty}^{\infty} \langle u(t), v(t) \rangle dt$$

and norm

$$\|u\| = \sqrt{\langle u, u \rangle}$$

for $u(\cdot), v(\cdot) \in L^2(H)$, it also is a real separable Hilbert space. For any $u(\cdot) \in L^2(H)$ and any real ω , the Fourier transform $\tilde{u}(\omega)$, satisfying

$$\langle a, \tilde{u}(\omega) \rangle = \frac{1}{2\pi} \text{l.i.m.} \int_{-T}^T \langle a, u(t) \rangle \exp(-i\omega t) dt \quad \forall a \in H,$$

exists and is in H^c , where "l.i.m." signifies "limit in the mean of order 2". Since $\langle a, u(t) \rangle$ is real,

$$\langle a, \tilde{u}(\omega) \rangle^* = \langle a, \tilde{u}(-\omega) \rangle \quad \forall a \in H, \forall \text{ real } \omega.$$

By Parseval's and Plancherel's Theorems [5], we have

$$\langle a, u(t) \rangle = \text{l.i.m.} \int_{-\Omega}^{\Omega} \langle a, \tilde{u}(\omega) \rangle \exp(i\omega t) d\omega \quad \forall a \in H,$$

and

$$\langle u, v \rangle = 2\pi \int_{-\infty}^{\infty} \langle \tilde{u}(\omega), \tilde{v}(\omega) \rangle d\omega$$

for $u(\cdot), v(\cdot) \in L^2(H)$.

For any real T , $L_T^2(H)$ denotes the subspace of $L^2(H)$ consisting of those elements $u(\cdot)$ which satisfy

$$u(t) = 0, \quad t \leq T.$$

By a theorem of Paley and Wiener [5], $u(\cdot)$ belongs to $L_T^2(H)$ if and only if its Fourier transform $\tilde{u}(\omega)$ is the boundary value of an H^c -valued function $\bar{u}(s)$ of a complex variable s in the sense that

$$\langle a, \tilde{u}(\omega) \rangle = \text{l.i.m.} \langle a, \bar{u}(\sigma + i\omega) \rangle \quad \forall a \in H, \quad \sigma \rightarrow 0+$$

where $\bar{u}(s)$ has the properties: $\langle a, \bar{u}(s) \rangle$ is holomorphic in $\operatorname{Re} s > 0$, for all $a \in H$;

$$\langle a, \bar{u}(s) \rangle^* = \langle a, \bar{u}(s^*) \rangle \quad \forall a \in H, \quad \forall \operatorname{Re} s > 0;$$

there exists real M such that

$$\int_{-\infty}^{\infty} \|\bar{u}(\sigma + i\omega)\|^2 d\omega \leq M \exp(-2\sigma T) \quad \forall \sigma > 0.$$

A further consequence of these properties, also established by Paley and Wiener [5], is that $\|\bar{u}(s)\|$ is bounded by

$$\|\bar{u}(s)\| \leq \frac{\|u\|}{\sqrt{4\pi\sigma}} \exp(-\sigma T) \quad \forall \operatorname{Re} s \geq \sigma > 0.$$

$L_r^2(H)$ is the union, taken over all real T , of the spaces $L_T^2(H)$. It is thus incomplete, but dense in $L^2(H)$.

$L_e^2(H)$ is the space of H -valued functions of t , defined on $(-\infty, \infty)$ and Lebesgue square-integrable on $(-\infty, T)$, for all real T . Thus for any $u(\cdot) \in L_e^2(H)$, the *truncation* $u_T(\cdot)$, defined by

$$u_T(t) = \begin{cases} u(t), & t \leq T, \\ 0, & t > T, \end{cases}$$

is an element of $L^2(H)$, for all real T . For $u(\cdot), v(\cdot) \in L_e^2(H)$ and any real T , we define the *truncated inner product*

$$\langle u, v \rangle_T = \langle u_T, v_T \rangle.$$

2.2. Operators. An operator on a linear space V is a mapping of V into itself. We shall use the following notations for particular operators.

The unit operator on any space is denoted by I .

The operators P_T and S_T on $L_e^2(H)$ are defined, for real T , by

$$(P_T u)(t) = u_T(t), \quad (S_T u)(t) = u(t - T),$$

where $u(\cdot) \in L_e^2(H)$.

Any bounded linear operator Q on H induces a linear operator on $L_e^2(H)$, denoted by the same symbol and defined by

$$(Qu)(t) = Qu(t),$$

where $u(\cdot) \in L_e^2(H)$.

An operator on any space will be called *unbiased* if and only if it maps the zero element into itself.

An operator X on $L_e^2(H)$ will be called *time-invariant* if and only if it satisfies

$$XS_T = S_TX \quad \forall \text{ real } T.$$

An operator X on $L^2(H)$ will be called *causal* if and only if it satisfies

$$P_TX = P_TXP_T \quad \forall \text{ real } T.$$

Such an operator always has a unique extension to an operator on $L_e^2(H)$, denoted by the same symbol and defined by

$$(Xu)(t) = (Xu_t)(t),$$

where $u(\cdot) \in L^2_c(H)$, and the extension is still causal in the same sense.

An operator X on $L^2(H)$ will be called *anticausal* if and only if it satisfies

$$XP_T = P_T X P_T \quad \forall \text{ real } T.$$

The class of bounded linear operators on a normed linear space V will be called $B(V)$. If V is a separable Hilbert space with inner product

$$\langle a, b \rangle = \langle b, a \rangle^*$$

the *adjoint* Q^* of an operator Q in $B(V)$, satisfying

$$\langle a, Q^*b \rangle = \langle b, Qa \rangle^* \quad \forall a, b \in V$$

exists in $B(V)$ and is unique. The *norm* of Q is

$$\|Q\| = \inf \{M : \|Qa\| \leq M\|a\| \quad \forall a \in V\} = \|Q^*\|.$$

With respect to an orthonormal basis $\{e_j\}$ of V , Q is represented by a matrix

$$Q = [q_{jk}] = [\langle e_j, Qe_k \rangle].$$

A $B(H^c)$ -valued operator function $Z(s)$ of a complex variable s will be called *bounded holomorphic* in a region (i.e., an open connected set) D of the complex plane if and only if there exists real M such that

$$\|Z(s)\| \leq M \quad \forall s \in D,$$

and $\langle a, Z(s)b \rangle$ is holomorphic in D , for all $a, b \in H^c$.

The class of *bounded, linear, time-invariant* operators on $L^2(H)$ will be called W .

A class W_+ of operators on $L^2(H)$ is defined as follows. To an operator Z in W_+ , there corresponds a $B(H^c)$ -valued function $Z(s)$, bounded holomorphic in $\text{Re } s > 0$ and satisfying

$$\langle a, Z(s)b \rangle^* = \langle a, Z(s^*)b \rangle \quad \forall a, b \in H, \forall \text{Re } s > 0.$$

Then, for $u(\cdot) \in L^2_T(H)$ and $a \in H$, $\langle a, Z(s)\bar{u}(s) \rangle$ is holomorphic in $\text{Re } s > 0$, by Vitali's and Weierstrass' theorems [6], since, by introducing an orthonormal basis, we can construct a sequence of approximants, holomorphic and uniformly bounded in $\text{Re } s \geq \sigma$ for any $\sigma > 0$, converging to $\langle a, Z(s)\bar{u}(s) \rangle \exp(sT)$ for all $\text{Re } s > 0$. Moreover,

$$\langle a, Z(s)\bar{u}(s) \rangle^* = \langle a, Z(s^*)\bar{u}(s^*) \rangle \quad \forall a \in H, \forall \text{Re } s > 0,$$

and there exists real M such that

$$\int_{-\infty}^{\infty} \|Z(\sigma + i\omega)\bar{u}(\sigma + i\omega)\|^2 d\omega \leq M \exp(-2\sigma T) \quad \forall \sigma > 0,$$

so that $Z(s)\bar{u}(s)$ has the property that its boundary value as $\text{Re } s \rightarrow 0$ is the Fourier transform of an element of $L^2_T(H)$. This element is defined to be $Zu(\cdot)$, and so Z is defined on $L^2_r(H)$ and may be extended by continuity to operate on $L^2(H)$. By construction, such an operator is bounded, linear, time-invariant and causal, so that W_+ is a subclass of W and each member of W_+ has a unique causal extension to an operator on $L^2_c(H)$. Also, by a similar construction of approximants, we can

show that the product of two $B(H^c)$ -valued functions, which are bounded holomorphic in a given region, is bounded holomorphic in the same region, and hence that the product of two operators in W_+ is in W_+ .

The subclass of operators in W whose adjoints are in W_+ is called W_- . Operators in W_- are anticausal, since, for Z in W_+ ,

$$Z^*P_T = (P_T Z)^* = (P_T Z P_T)^* = P_T Z^* P_T \quad \forall \text{ real } T.$$

An operator Z in W_+ is called *invertible in W_+* if and only if it has a unique inverse \hat{Z} in W_+ , i.e., $Z(s)$ has a $B(H^c)$ -valued inverse $\hat{Z}(s)$, bounded holomorphic in $\text{Re } s > 0$. The extension of such a Z to an operator on $L_e^2(H)$ has an inverse (the extension of \hat{Z}) which is unique since, otherwise, there exists $u(\cdot) \in L_e^2(H)$ such that

$$Zu \equiv 0,$$

and then, for any $v(\cdot) \in L^2(H)$ and any real T ,

$$\begin{aligned} \langle u, v \rangle_T &= \langle u_T, v_T \rangle \\ &= \langle Zu_T, \hat{Z}^* v_T \rangle \\ &= \langle Zu_T, \hat{Z}^* v_T \rangle_T \\ &= \langle Zu, \hat{Z}^* v_T \rangle_T \\ &= 0, \end{aligned}$$

whence

$$u \equiv 0.$$

2.3. Relations. A *relation* on a linear space V is a subset of the product space $V \times V$. If R is a relation and a, b are elements of V , we shall use the notations

$$\begin{aligned} (a, b) &\in R, \\ b &= Ra, \\ a &= R^{-1}b, \\ (b, a) &\in R^{-1} \end{aligned}$$

to mean the same thing. Thus the inverse R^{-1} is always defined. An operator on V automatically defines a relation on V , denoted by the same symbol.

The *gain* of a relation R on $L_e^2(H)$ is defined by

$$\|R\| = \inf \{M : \|v_T\| \leq M \|u_T\|, \forall (u, v) \in R, \forall \text{ real } T\}.$$

Such a relation is called *finite-gain* if and only if

$$\|R\| < \infty.$$

A relation R on $L_e^2(H)$ is called *positive* if and only if

$$\langle u_T, v_T \rangle \geq 0 \quad \forall (u, v) \in R, \quad \forall \text{ real } T,$$

and *strongly positive* if and only if there exists $\delta > 0$ such that

$$\langle u_T, v_T \rangle \geq \delta \|u_T\|^2 \quad \forall (u, v) \in R, \quad \forall \text{ real } T.$$

3. Stability criteria. The feedback system to be considered is described by the equations

$$(S) \quad \begin{aligned} u &= v - Fy, \\ y &= z + Gu, \end{aligned}$$

where v, z are elements of $L^2(H)$, u, y are elements of $L_e^2(H)$, and F, G are unbiased causal operators on $L^2(H)$ (and consequently also on $L_e^2(H)$). The system (S) will be called *stable* if and only if any solution of (S) has $u, y \in L^2(H)$ and there exists real M , independent of v and z , such that

$$\max(\|u\|, \|y\|) \leq M \max(\|v\|, \|z\|)$$

for every solution of (S).

We now have the following stability theorem, which is an adaptation to our purpose of standard results [1],[4].

THEOREM 1. *Let K, L, Q be operators in $B(H)$, such that $I - KL, I - LK$ and Q have inverses in $B(H)$.*

Then, if the relations $Q(F - K)(I - LF)^{-1}$ and $Q^(G + L)(I + KG)^{-1}$ are both strongly positive, or if one of them is strongly positive and finite-gain while the other is positive, the system (S) is stable.*

Proof. Define $w, x \in L_e^2(H)$ by

$$\begin{aligned} w &= u + K(y - z), \\ x &= y + L(u - v), \end{aligned}$$

so that (S) gives

$$\begin{aligned} w &= (I + KG)u = v - Kz - (F - K)y, \\ x &= (I - LF)y = z - Lv + (G + L)u, \end{aligned}$$

whence

$$\begin{aligned} w &= v - Kz - (F - K)(I - LF)^{-1}x, \\ x &= z - Lv + (G + L)(I + KG)^{-1}w, \end{aligned}$$

and so

$$\begin{aligned} \langle x, Qw \rangle_T &= \langle x, Q(v - Kz) \rangle_T - \langle x, Q(F - K)(I - LF)^{-1}x \rangle_T \\ &= \langle w, Q^*x \rangle_T = \langle w, Q^*(z - Lv) \rangle_T + \langle w, Q^*(G + L)(I + KG)^{-1}w \rangle_T \end{aligned}$$

for all real T . Hence the conditions of the theorem imply that there exist

$$\delta_F \geq 0, \quad \delta_G \geq 0, \quad \text{with } \delta_F + \delta_G > 0,$$

such that

$$\delta_F \|x_T\|^2 + \delta_G \|w_T\|^2 \leq \|Q^*x_T\| \|(v - Kz)_T\| + \|Qw_T\| \|(z - Lv)_T\|,$$

while, if $\delta_F = 0$, there exists $\mu_G \geq 0$ such that

$$\|Q^*x_T\| \leq \|Q^*(z - Lv)_T\| + \mu_G \|w_T\|,$$

and, similarly, if $\delta_G = 0$, there exists $\mu_F \geq 0$ such that

$$\|Qw_T\| \leq \|Q(v - Kz)_T\| + \mu_F \|x_T\|.$$

In any of these cases, it then follows that there exists real μ such that

$$\max(\|w_T\|, \|x_T\|) \leq \mu \max(\|v_T\|, \|z_T\|) \quad \forall \text{ real } T,$$

and hence, since

$$\begin{aligned}u - v &= (I - KL)^{-1}[w - v - K(x - z)], \\y - z &= (I - LK)^{-1}[x - z - L(w - v)],\end{aligned}$$

there exists real M such that

$$\max(\|u_T\|, \|y_T\|) \leq M \max(\|v_T\|, \|z_T\|) \quad \forall \text{ real } T.$$

Thus the conditions for stability of (S) are fulfilled. Q.E.D.

Subsequently, we shall take K, L to be certain symmetric operators, and set

$$Q = (I - KL)^{-1}$$

so that

$$Q^* = (I - LK)^{-1}.$$

Also, we shall take G to be in W_+ and make use of the following lemma, which is a straightforward generalization of a standard type of result [2].

LEMMA 2. Let Z be an operator in W_+ , with

$$\sup_{\operatorname{Re} s > 0} \|Z(s)\| < 1.$$

Then, the operator

$$(I + Z)(I - Z)^{-1}$$

is in W_+ and is strongly positive.

Proof. Since the operator function $Z(s)$ is bounded holomorphic in $\operatorname{Re} s > 0$, with

$$\sup_{\operatorname{Re} s > 0} \|Z(s)\| < 1,$$

we can construct the sequence of approximants

$$Z_n(s) = I + \sum_{r=1}^n Z^r(s),$$

which converges in the norm and hence defines the operator $[I - Z(s)]^{-1}$ in $B(H^c)$, for all $\operatorname{Re} s > 0$, since $B(H^c)$ is complete in its norm topology. Further, for any $a, b \in H^c$, the sequence of approximants $\langle a, Z_n(s)b \rangle$ is holomorphic and uniformly bounded in $\operatorname{Re} s > 0$, whence, by Vitali's and Weierstrass' theorems [6], $\langle a, [I - Z(s)]^{-1}b \rangle$ is holomorphic in $\operatorname{Re} s > 0$, and also

$$\sup_{\operatorname{Re} s > 0} \|[I - Z(s)]^{-1}\| \leq [1 - \sup_{\operatorname{Re} s > 0} \|Z(s)\|]^{-1} < \infty,$$

so that $[I - Z(s)]^{-1}$ is bounded holomorphic in $\operatorname{Re} s > 0$ and hence $(I - Z)^{-1}$ is in W_+ .

Now, for any $u(\cdot) \in L^2(H)$,

$$\begin{aligned}&\langle u, (I + Z)(I - Z)^{-1}u \rangle \\&= \frac{1}{2}[\langle u, (I + Z)(I - Z)^{-1}u \rangle + \langle u, (I - Z^*)^{-1}(I + Z^*)u \rangle] \\&= \frac{1}{2}\langle (I - Z)^{-1}u, [(I - Z^*)(I + Z) + (I + Z^*)(I - Z)](I - Z)^{-1}u \rangle \\&= \langle (I - Z)^{-1}u, (I - Z^*Z)(I - Z)^{-1}u \rangle \\&\geq \|(I - Z)^{-1}u\|^2(1 - \|Z\|^2) \\&\geq (1 - \|Z\|)\|u\|^2/(1 + \|Z\|),\end{aligned}$$

and so $(I + Z)(I - Z)^{-1}$ is strongly positive since

$$\|Z\| \leq \sup_{\operatorname{Re} s > 0} \|Z(s)\| < 1. \quad \text{Q.E.D.}$$

From now on, we work with a fixed representation, taken with respect to an orthonormal basis $\{e_j\}$ of H , where the index j is an integer, ranging from 1 to m if H has finite dimension m and over the positive integers if H has infinite dimension. Summations over j will be assumed to be unrestricted unless otherwise indicated.

Now, suppose $Z(s)$ is a $B(H^c)$ -valued function, defined and uniformly bounded in norm in a region D of the complex s -plane. Then, if $Z(s)$ is bounded holomorphic in D , all its matrix elements $z_{jk}(s)$ are holomorphic in D , by definition. Conversely, if all the $z_{jk}(s)$ are holomorphic in D , then, for any $a, b \in H^c$, we can construct the sequence of approximants

$$\sum_{j=1}^n \sum_{k=1}^n a_j^* z_{jk}(s) b_k, \quad n = \text{positive integer},$$

which are holomorphic and uniformly bounded in D and converge to $\langle a, Z(s)b \rangle$, which is thus holomorphic in D , by Vitali's and Weierstrass' theorems [6], and hence $Z(s)$ is bounded holomorphic in D . Thus, an operator in W_+ is now represented by a matrix function $Z(s)$ whose elements are holomorphic in $\operatorname{Re} s > 0$ and satisfy

$$z_{jk}(s)^* = z_{jk}(s^*) \quad \forall j, k, \quad \forall \operatorname{Re} s > 0,$$

$$\sup_{\operatorname{Re} s > 0} \|Z(s)\| < \infty.$$

We shall take G to be in W_+ , and set

$$K = A,$$

$$L = B^{-1},$$

where A, B are operators in $B(H)$ which are diagonal in the representation we are using, i.e., represented by matrices

$$A = \operatorname{diag}(\alpha_j),$$

$$B = \operatorname{diag}(\beta_j),$$

with

$$\inf_j |\beta_j| > 0$$

so that B^{-1} is in $B(H)$. For real $\sigma, \Omega > 0$, we define the contour $C(\sigma, \Omega)$ in the s -plane as the boundary of the rectangle with vertices $[\sigma \pm i\Omega, \Omega \pm i\Omega]$. We then derive stability criteria involving the behavior of the elements $g_{jk}(s)$ of the matrix function $G(s)$, which represents G , as s goes round the contours $C(\sigma, \Omega)$.

THEOREM 3. *Suppose*

$$\inf_j \alpha_j > 0, \quad \inf_j \beta_j > 0, \quad \inf_j (\alpha_j^{-1} - \beta_j^{-1}) > 0,$$

and that there exist real $\varepsilon, \sigma_0, \Omega_0 > 0$ such that, for all $\sigma \in (0, \sigma_0)$ and $\Omega > \Omega_0$ and for each value of j , $g_{jj}(s)$ does not encircle $-\frac{1}{2}(\alpha_j^{-1} + \beta_j^{-1})$ as s goes round $C(\sigma, \Omega)$ and also

$$|g_{jj}(s) + \frac{1}{2}(\alpha_j^{-1} + \beta_j^{-1})| - \sum_{k(\neq j)} \frac{|g_{jk}(s)| + |g_{kj}(s)|}{2} \geq \frac{1}{2}(\alpha_j^{-1} - \beta_j^{-1})(1 + \varepsilon),$$

for all s on $C(\sigma, \Omega)$.

Then, $(I - B^{-1}A)^{-1}(G + B^{-1})(I + AG)^{-1}$ is strongly positive and finite-gain.

Proof. Define

$$\Theta = \frac{A^{-1} + B^{-1}}{2} = \text{diag}(\theta_j),$$

$$\Phi = \frac{A^{-1} - B^{-1}}{2} = \text{diag}(\phi_j),$$

$$\Phi^{1/2} = \text{diag}(\sqrt{\phi_j}),$$

all of which, together with their inverses, are in $B(H)$. By Theorem A4 of the Appendix, identifying $Z(s) = \Phi^{-1/2}[G(s) + \Theta]\Phi^{-1/2}$, $\lambda_j = \sqrt{\phi_j}$, the operator Z is invertible in W_+ and

$$\sup_{\text{Re } s > 0} \|\hat{Z}(s)\| \leq (1 + \varepsilon)^{-1} < 1,$$

whence, by Lemma 2, there exists in W_+ the strongly positive operator

$$\begin{aligned} A^{-1}\Phi^{-1/2}[I - \Phi^{1/2}(G + \Theta)^{-1}\Phi^{1/2}][I + \Phi^{1/2}(G + \Theta)^{-1}\Phi^{1/2}]^{-1}\Phi^{-1/2}A^{-1} \\ = A^{-1}[I - \Phi(G + \Theta)^{-1}][I + \Phi(G + \Theta)^{-1}]^{-1}A^{-1} \\ = A^{-1}\Phi^{-1}(G + B^{-1})(G + A^{-1})^{-1}A^{-1} \\ = 2(I - B^{-1}A)^{-1}(G + B^{-1})(I + AG)^{-1}. \end{aligned} \quad \text{Q.E.D.}$$

THEOREM 4. Suppose

$$\inf_j (\beta_j - \alpha_j) > 0,$$

and that $G(s)$ has an inverse $\hat{G}(s)$ in $B(H^c)$ for all $\text{Re } s > 0$, with the properties: there exist real $\varepsilon, \sigma_0, \Omega_0 > 0$ such that for all $\sigma \in (0, \sigma_0)$ and $\Omega > \Omega_0$ and for each value of j , $\hat{g}_{jj}(s)$ does not encircle $-\frac{1}{2}(\alpha_j + \beta_j)$ as s goes round $C(\sigma, \Omega)$,

$$|\hat{g}_{jj}(s) + \frac{1}{2}(\alpha_j + \beta_j)| - \sum_{k(\neq j)} \frac{|\hat{g}_{jk}(s)| + |\hat{g}_{kj}(s)|}{2} \geq \frac{1}{2}(\beta_j - \alpha_j)(1 + \varepsilon)$$

for all s on $C(\sigma, \Omega)$, and also there exists real $M(\sigma, \Omega)$ such that $\|\hat{G}(s)\| \leq M(\sigma, \Omega)$ for all s on $C(\sigma, \Omega)$.

Then, $(I - B^{-1}A)^{-1}(G + B^{-1})(I + AG)^{-1}$ is strongly positive and finite-gain.

Proof. Define

$$\Gamma = \frac{A + B}{2} = \text{diag}(\gamma_j),$$

$$\Delta = \frac{B - A}{2} = \text{diag}(\delta_j),$$

$$\Delta^{1/2} = \text{diag}(\sqrt{\delta_j}),$$

so that $\Delta, \Delta^{1/2}$ are in $B(H)$, together with their inverses. By Theorem A.4, identifying $Z(s) = \Delta^{-1/2}[\hat{G}(s) + \Gamma]\Delta^{-1/2}$, $\lambda_j = \sqrt{\delta_j}$, there exists an operator \hat{Z} in W_+ with the properties

$$\Delta^{-1/2}\hat{Z}\Delta^{-1/2}(I + \Gamma G) = G$$

and

$$\sup_{\text{Re } s > 0} \|\hat{Z}(s)\| \leq (1 + \varepsilon)^{-1} < 1,$$

whence, by Lemma 2, there exists in W_+ the strongly positive operator

$$\begin{aligned} \Delta^{-1/2}(I + \hat{Z})(I - Z)^{-1}\Delta^{-1/2} \\ &= \Delta^{-1}(I + \Delta^{1/2}\hat{Z}\Delta^{-1/2})(I - \Delta^{1/2}\hat{Z}\Delta^{-1/2})^{-1} \\ &= \Delta^{-1}(I + \Gamma G + \Delta G)(I + \Gamma G - \Delta G)^{-1} \\ &= \Delta^{-1}(I + BG)(I + AG)^{-1} \\ &= 2(I - B^{-1}A)^{-1}(G + B^{-1})(I + AG)^{-1}. \end{aligned} \quad \text{Q.E.D.}$$

THEOREM 4'. *Suppose*

$$\inf_j (\alpha_j + \beta_j) \geq 0, \quad \sup_j (\alpha_j) < 0,$$

and that there exist real $\varepsilon, \sigma_0, \Omega_0 > 0$ such that for all $\sigma \in (0, \sigma_0)$ and $\Omega > \Omega_0$, $G(s)$ has an inverse $\hat{G}(s)$ in $B(H^c)$ for all s on $C(\sigma, \Omega)$, and also, for each value of j ,

$$|\hat{g}_{jj}(s) + \frac{1}{2}(\alpha_j + \beta_j)| - \sum_{k(\neq j)} \frac{|\hat{g}_{jk}(s)| + |\hat{g}_{kj}(s)|}{2} \geq \frac{1}{2}(\beta_j - \alpha_j)(1 + \varepsilon),$$

for all s on $C(\sigma, \Omega)$.

Then, $(I - B^{-1}A)^{-1}(G + B^{-1})(I + AG)^{-1}$ is strongly positive and finite-gain.

Proof. Since the conditions of the theorem remain valid with β_j replaced by $-\alpha_j$ for each j , consider first the case $B = -A$. Then for all $\sigma \in (0, \sigma_0)$ and $\Omega > \Omega_0$,

$$\|(-A)^{1/2}G(s)(-A)^{1/2}\| \leq (1 + \varepsilon)^{-1} < 1 \quad \forall s \text{ on } C(\sigma, \Omega),$$

by Theorem A.2 of the Appendix, where

$$(-A)^{1/2} = \text{diag}(\sqrt{-\alpha_j}),$$

and hence, by Lemma A.3 of the Appendix,

$$\sup_{\text{Re } s > 0} \|(-A)^{1/2}G(s)(-A)^{1/2}\| < 1,$$

so that, for any operator in $B(H)$ of the form

$$\Psi = \text{diag}(\psi_j),$$

where, for each j ,

$$0 \leq \psi_j \leq -\alpha_j,$$

the operator $(I + \Psi G)^{-1}$ exists in W_+ .

By a finite number of successive steps of this kind, we can show that the operator $[I + (A + B)G/2]^{-1}$ exists in W_+ , and the proof then proceeds as in Theorem 4. Q.E.D.

We can now combine the above results to give more explicit stability criteria. We take the operator F to be of the following form: for some decomposition of H into the direct sum of real Hilbert spaces H_r , each spanned by a subset S_r of $\{e_j\}$, where r ranges over some index set, F is the direct sum of unbiased casual operators F_r on $L^2(H_r)$, and, for each r , there exist real a_r, b_r , with

$$b_r > a_r,$$

such that

$$\langle b_r y - F_r y, F_r y - a_r y \rangle \geq 0 \quad \forall y(\cdot) \in L^2(H_r).$$

We then have the following theorem.

THEOREM 5. *Suppose that, with the identifications*

$$\left. \begin{array}{l} \alpha_j = a_r \\ \beta_j = b_r \end{array} \right\} \quad \text{for each } j \text{ with } e_j \in S_r,$$

the conditions of one of Theorems 3, 4, 4' hold.

Then, the system (S) is stable.

Proof. For any $x \in L^2(H_r)$,

$$\begin{aligned} & \langle x, (F_r - a_r I)(b_r I - F_r)^{-1} x \rangle \\ &= \langle (b_r I - F_r)y, (F_r - a_r I)y \rangle \\ &\geq 0, \end{aligned}$$

identifying

$$y = (b_r I - F_r)^{-1} x,$$

and hence the relation

$$(I - AB^{-1})^{-1}(F - A)(I - B^{-1}F)^{-1} = B(B - A)^{-1}(F - A)(B - F)^{-1}B$$

is positive, so that the conditions of Theorem 1 are satisfied. Q.E.D.

4. Comments and examples. The stability criteria derived above depend on the behavior of $G(s)$ or $\hat{G}(s)$ on $C(\sigma, \Omega)$ for sufficiently small σ and large Ω . In practice, continuity properties as $\sigma \rightarrow 0$ will usually justify restricting our attention to the imaginary axis together with a large return contour in the right-half-plane. Further, in the case of $G(s)$ (though probably not $\hat{G}(s)$), the elements will usually vanish fast enough as $|s| \rightarrow \infty$ with $\text{Re } s > 0$ to enable this part of the contour to be ignored.

Theorems 3, 4 and 4' generalize results already obtained by the author [3] for the case that H is finite-dimensional and the matrix elements of $G(s)$ are rational functions of s . For the criterion based on the behavior of the elements of $G(s)$, Theorem 3 gives an essentially complete generalization. For the criterion based on the elements of $\hat{G}(s)$, Theorems 4 and 4' provide partial generalizations, the difficulty in obtaining a complete one being due to the possible infinite dimension of H , not to the irrationality of the transfer functions; the proofs in the finite-dimensional case involve determinants, which infinite matrices do not generally have. The criteria given here by no means exhaust the list of those which can be derived along the same lines. For example, there are analogues of Theorem 3, for other values of the α 's and β 's, in which the elements of $G(s)$ have to satisfy different constraints, which can be derived from known bounds for the norm of a $B(H^c)$ -valued operator; they have been omitted because of their more complicated nature than those given. More interesting would be still further generalizations allowing the consideration of unbounded operators, but this has not yet been achieved.

Finally, we illustrate the possible use of our stability criteria with the following examples.

Example 1. Let H be the space of L^2 -functions of a real variable θ on $[0, \pi]$. Suppressing the input variables v, z , let the system (S) be given by

$$\frac{\partial^2 y}{\partial t^2} + \lambda \frac{\partial y}{\partial t} - \mu \frac{\partial^2 y}{\partial \theta^2} = (1 + 2\delta \cos \theta)u,$$

$$u = -N(y),$$

where λ, μ, δ are real positive constants, $N(\cdot)$ is a function satisfying

$$[N(y) - ay][by - N(y)] \geq 0$$

for some real a, b , with

$$0 < a < b,$$

and $y(\theta, t)$ is subject to the constraint

$$y(0, t) = y(\pi, t) = 0 \quad \forall t.$$

Then, taking the orthonormal basis $\{\sqrt{2/\pi} \sin j\theta\}$ and Laplace transforming, we have

$$(s^2 + \lambda s + \mu j^2) \bar{y}_j(s) = \bar{u}_j(s) + \delta[\bar{u}_{j-1}(s) + \bar{u}_{j+1}(s)]$$

for all positive integers j , with the understanding $\bar{u}_0 \equiv 0$. Thus, $G(s)$ has the matrix elements

$$g_{jk}(s) = \begin{cases} 1/(s^2 + \lambda s + \mu j^2), & k = j, \\ \delta/(s^2 + \lambda s + \mu j^2), & k = j \pm 1, \\ 0, & \text{otherwise,} \end{cases}$$

and we can determine the stability of the system by drawing, for each j , the Nyquist plot of $1/(\mu j^2 - \omega^2 + i\lambda\omega)$, surrounding each point by a circle of radius

$$\delta[1/|\mu - \omega^2 + i\lambda\omega| + 1/4\mu - \omega^2 + i\lambda\omega]/2, \quad j = 1,$$

$$\delta[1/|\mu j^2 - \omega^2 + i\lambda\omega| + 1/2|\mu(j+1)^2 - \omega^2 + i\lambda\omega| + 1/2|\mu(j-1)^2 - \omega^2 + i\lambda\omega|],$$

$$j > 1,$$

and examining whether or not the bands swept out by these circles, as ω goes from $-\infty$ to ∞ , encircle or intersect the critical disc on $[-1/a, -1/b]$ as diameter. In view of the rapid decrease for large j , only the first few bands will need to be drawn in practice. If the bands do not encircle the critical disc, and remain outside it by a margin $\varepsilon > 0$, the system is stable.

Example 2. Let θ be a real variable defined modulo 2π , and H be the space of L^2 -functions of θ on $[-\pi, \pi]$. Suppressing the input variables, let the system (S) be given by

$$\left(\frac{\partial^2}{\partial t^2} + \lambda \frac{\partial}{\partial t} + v\right)y(\theta, t) + \frac{\gamma}{4} \int_{-\pi}^{\pi} \operatorname{sgn}(\theta - \theta' \bmod 2\pi) y(\theta', t) d\theta' = u(\theta, t),$$

$$u = -N(y),$$

where λ, v, γ are real positive constants and $N(\cdot)$ is a function satisfying

$$|N(y)| \leq b|y|$$

for some real b .

Taking the orthonormal basis $\{e_j\}$ defined by

$$e_j = \begin{cases} 1/\sqrt{2\pi}, & j = 1, \\ (1/\sqrt{\pi}) \cos(j-1)\theta/2, & j \text{ odd} > 1, \\ (1/\sqrt{\pi}) \sin j\theta/2, & j \text{ even}, \end{cases}$$

and Laplace transforming, we find

$$\bar{u}(s) = \hat{G}(s)\bar{y}(s),$$

where $\hat{G}(s)$ has the matrix elements

$$\hat{g}_{jk}(s) = \begin{cases} s^2 + \lambda s + v, & k = j, \\ \gamma/n, & j = 2n, \quad k = 2n+1, \quad n \text{ odd}, \\ -\gamma/n, & j = 2n+1, \quad k = 2n, \quad n \text{ odd}, \\ 0, & \text{otherwise.} \end{cases}$$

We now determine stability by drawing the Nyquist plot of $v^2 - \omega^2 + i\lambda\omega$ and surrounding each point by circles of radii γ/n , for all odd positive integers n . Clearly only the circle for $n = 1$ needs to be drawn, in fact. Then, if the band thus swept out as ω goes from $-\infty$ to ∞ avoids the disc, radius b , centered on the origin, by a margin $\varepsilon > 0$, the system is stable.

Appendix. Here we collect some theorems about matrices, used in the text. The matrices may be finite or infinite; in the latter case, they are bounded operators on l^2 , the space of complex column vectors whose elements form absolutely square-summable sequences. We use the symbol \hat{Z} to denote inverse, the norm

$$\|Q\| = \sup_{\|x\|=1} \|Qx\| = \sup_{\|x\|=\|y\|=1} |y^*Qx|,$$

where $*$ means Hermitian adjoint, and the *numerical radius*

$$w(Q) \equiv \sup_{\|x\|=1} |x^*Qx|.$$

Of the results given here, Theorem A.1 is a simple generalization of known results [7]; Theorem A.2 generalizes a result previously given by the author [3], for finite matrices; Lemma A.3 is an elementary consequence of the maximum-modulus theorem; Theorem A.4 is believed to be new. We remark that Theorems A.1 and A.2, at least, generalize further to the case of partitioned matrices, with appropriate replacement of moduli of elements by spectral norms of blocks, and would lead to similar generalizations of the stability criteria in the text if the required analyticity properties could be established; this, however, seems difficult in general, unless the matrices are finite, and so, as the corresponding generalizations seem rather weak in practice, we have omitted them.

THEOREM A.1. Let $Q = [q_{jk}]$, and let there be real $\mu_j > 0$ such that

$$\sum_k \frac{\mu_k}{\mu_j} \left(\frac{|q_{jk}| + |q_{kj}|}{2} \right) \leq 1 \quad \forall j,$$

Then

$$w(Q) \leq 1.$$

Proof. For any $x = [x_j] \in l^2$,

$$\begin{aligned} \left| \sum_j \sum_k x_j^* q_{jk} x_k \right| &\leq \sum_j \sum_k |q_{jk}| |x_j| |x_k| \\ &\leq \sum_j \sum_k \frac{|q_{jk}|}{2} \left[\frac{\mu_k}{\mu_j} |x_j|^2 + \frac{\mu_j}{\mu_k} |x_k|^2 \right] \\ &= \sum_j \left[|x_j|^2 \sum_k \frac{\mu_k}{\mu_j} \left(\frac{|q_{jk}| + |q_{kj}|}{2} \right) \right] \\ &\leq \|x\|^2, \end{aligned}$$

whence

$$w(Q) = \sup_{\|x\|=1} |x^*Qx| \leq 1.$$

Q.E.D.

THEOREM A.2. Let $Z = [z_{jk}]$, with

$$\|Z\| < \infty$$

and let there be real $\lambda_j > 0$ such that

$$|z_{jj}| - \sum_{k(\neq j)} \frac{\lambda_k}{\lambda_j} \left(\frac{|z_{jk}| + |z_{kj}|}{2} \right) \geq 1 \quad \forall j.$$

Then \hat{Z} exists, with

$$\|\hat{Z}\| \leq 1.$$

Proof. Define the diagonal unitary matrix

$$U = \text{diag} (z_{jj}/|z_{jj}|)$$

and let

$$V = [v_{jk}] = \frac{1}{2}(U^*Z + Z^*U)$$

so that

$$V^* = V$$

and

$$v_{jj} - \sum_{k(\neq j)} \frac{\lambda_k}{\lambda_j} |v_{jk}| \geq 1 \quad \forall j.$$

Then for any $x = [x_j] \in l^2$,

$$\begin{aligned} x^*Vx &= \sum_j \left[v_{jj}|x_j|^2 + \sum_{k(\neq j)} v_{jk}x_j^*x_k \right] \\ &\geq \sum_j \left[v_{jj}|x_j|^2 - \sum_{k(\neq j)} |v_{jk}||x_j||x_k| \right] \\ &\geq \sum_j \left[v_{jj}|x_j|^2 - \frac{1}{2} \sum_{k(\neq j)} |v_{jk}| \left(\frac{\lambda_k}{\lambda_j} |x_j|^2 + \frac{\lambda_j}{\lambda_k} |x_k|^2 \right) \right] \\ &= \sum_j |x_j|^2 \left[v_{jj} - \sum_{k(\neq j)} \frac{\lambda_k}{\lambda_j} |v_{jk}| \right] \\ &\geq \|x\|^2 \end{aligned}$$

and so

$$\begin{aligned} \|Zx\| \|x\| &\geq |x^*U^*Zx| \\ &\geq \text{Re}(x^*U^*Zx) \\ &= x^*Vx \\ &\geq \|x\|^2. \end{aligned}$$

Thus,

$$\inf_{\|x\|=1} \|Zx\| \geq 1$$

and similarly,

$$\inf_{\|x\|=1} \|Z^*x\| \geq 1,$$

whence \hat{Z} exists [8], and

$$\begin{aligned} \|\hat{Z}\| &= 1 / \inf_{\|x\|=1} \|Zx\| \\ &\leq 1. \end{aligned}$$

Q.E.D.

LEMMA A.3. Let $Q(s) = [q_{jk}(s)]$, where s is a complex variable. Let C denote a simple closed Jordan curve, bounding a region R in the complex plane, and suppose that $q_{jk}(s)$ is holomorphic in R and continuous in $\bar{R} = R \cup C$, for all j, k , and that $\|Q(s)\| < \infty$, for all $s \in \bar{R}$.

Then,

$$\sup_{s \in \bar{R}} w(Q(s)) = \sup_{s \in C} w(Q(s))$$

and

$$\sup_{s \in \bar{R}} \|Q(s)\| = \sup_{s \in C} \|Q(s)\|.$$

Proof. Let E denote the set of all nonvanishing $x = [x_j]$ with only a finite number of nonzero elements. Then, for $x, y \in E$, $y^*Q(s)x$ is holomorphic in R and continuous in \bar{R} , so that, by the maximum-modulus theorem, $|y^*Q(s)x|$ attains its supremum over \bar{R} on C .

But, E is dense in l^2 , so we have

$$w(Q(s)) = \sup_{x \in E} \frac{|x^*Q(s)x|}{\|x\|^2},$$

$$\|Q(s)\| = \sup_{x, y \in E} \frac{|y^*Q(s)x|}{\|y\| \|x\|},$$

and the results follow. Q.E.D.

THEOREM A.4. Let $Z(s) = [z_{jk}(s)]$, where s is a complex variable. For real $\sigma, \Omega > 0$, let $R(\sigma, \Omega)$ denote the interior of the rectangle with vertices $[\sigma \pm i\Omega, (1 \pm i)\Omega]$, $\bar{R}(\sigma, \Omega)$ its closure, $C(\sigma, \Omega)$ its boundary. Suppose $z_{jk}(s)$ is holomorphic in $\text{Re } s > 0$, and there exist $\sigma_0, \Omega_0 > 0$ such that, for all $\Omega > \Omega_0$ and all $\sigma \in (0, \sigma_0)$: there exists real $M(\sigma, \Omega) > 0$ such that $\|Z(s)\| \leq M(\sigma, \Omega)$ for all s on $C(\sigma, \Omega)$; there exists real $\lambda_j > 0$ such that

$$|z_{jj}(s)| - \sum_{k(\neq j)} \frac{\lambda_k}{\lambda_j} \left[\frac{|z_{jk}(s)| + |z_{kj}(s)|}{2} \right] \geq 1 \quad \forall s \text{ on } C(\sigma, \Omega), \forall j;$$

$z_{jj}(s)$ does not encircle the origin as s goes round $C(\sigma, \Omega)$ for all j .

Then, $\hat{Z}(s)$ exists, all its elements being holomorphic in $\text{Re } s > 0$, and

$$\|\hat{Z}(s)\| \leq 1 \quad \forall \text{Re } s > 0.$$

Proof. Since $z_{jj}(s)$ is holomorphic in $\text{Re } s > 0$ and fails to encircle the origin as s goes round $C(\sigma, \Omega)$, for sufficiently small σ and large Ω , it has no zeros in $\text{Re } s > 0$ and hence we can define a square root

$$\xi_j(s) = \sqrt{z_{jj}(s)}$$

which is holomorphic in $\text{Re } s > 0$ and satisfies

$$1 \leq |\xi_j(s)| \leq \sqrt{M(\sigma, \Omega)} \quad \forall s \in \bar{R}(\sigma, \Omega), \forall j,$$

for each $\Omega > \Omega_0$ and $\sigma \in (0, \sigma_0)$, by the maximum-modulus theorem.

Next, we define $Q(s) = [q_{jk}(s)]$ by

$$\begin{aligned} q_{jk}(s) &= \xi_j^{-1}(s) z_{jk}(s) \xi_k^{-1}(s), \quad k \neq j, \\ q_{jj}(s) &= 0, \end{aligned}$$

so that

$$Z(s) = \Xi(s)[I + Q(s)]\Xi(s),$$

where

$$\Xi(s) = \text{diag}(\xi_j(s)).$$

Then, setting

$$\mu_j(s) = \lambda_j |\xi_j(s)|,$$

we have for each $\Omega > \Omega_0$ and $\sigma \in (0, \sigma_0)$,

$$\sum_k \frac{\mu_k(s)}{\mu_j(s)} \left[\frac{|q_{jk}(s)| + |q_{kj}(s)|}{2} \right] \leq 1 - \frac{1}{M(\sigma, \Omega)} \quad \forall s \text{ on } C(\sigma, \Omega), \quad \forall j,$$

so that, by Theorem A.1,

$$w(Q(s)) \leq 1 - \frac{1}{M(\sigma, \Omega)} \quad \forall s \text{ on } C(\sigma, \Omega),$$

and hence, by Lemma A.3,

$$w(Q(s)) \leq 1 - \frac{1}{M(\sigma, \Omega)} \quad \forall s \in \bar{R}(\sigma, \Omega).$$

Since the numerical radius is an upper bound for the spectral radius [9], $\rho(Q(s))$, it follows that

$$\rho(Q(s)) < 1 \quad \forall \text{Re } s > 0,$$

and hence, for each s in $\text{Re } s > 0$, $[I + Q(s)]^{-1}$ exists and is bounded. Thus, there also exists

$$\hat{Z}(s) = \Xi^{-1}(s)[I + Q(s)]^{-1}\Xi^{-1}(s),$$

with

$$\|\hat{Z}(s)\| < \infty \quad \forall \text{Re } s > 0.$$

Moreover, if $Z(s)$ has finite dimension $m \times m$,

$$|\det [I + Q(s)]| \geq |1 - \rho(Q(s))|^m > 0 \quad \forall \text{Re } s > 0,$$

so that all the elements of $Z(s)$ are holomorphic in $\text{Re } s > 0$.

If $Z(s)$ has infinite dimension, we construct the sequence of approximants $Z^{(n)}(s) = [z_{jk}^{(n)}(s)]$, defined, for integer $n > 0$, by

$$z_{jk}^{(n)}(s) = \begin{cases} z_{jk}(s), & k \leq n, \\ z_{jj}(s), & j = k > n, \\ 0, & j \neq k > n, \end{cases}$$

so that, by the same argument as for $\hat{Z}(s)$, $\hat{Z}^{(n)}(s)$ exists, with

$$\|\hat{Z}^{(n)}(s)\| < \infty, \quad \forall \operatorname{Re} s > 0, \quad \forall n,$$

Moreover, the elements of $\hat{Z}^{(n)}(s)$ are finite sums of finite products of elements of $Z(s)$, inverses of diagonal elements of $Z(s)$, and elements of the inverse of the leading principal $n \times n$ submatrix of $Z(s)$, and are consequently holomorphic in $\operatorname{Re} s > 0$ for all n . Hence, for each $\Omega > \Omega_0$ and $\sigma \in (0, \sigma_0)$,

$$\sup_{s \in \bar{R}(\sigma, \Omega)} \|\hat{Z}^{(n)}(s)\| = \sup_{s \in C(\sigma, \Omega)} \|\hat{Z}^{(n)}(s)\| \leq 1, \quad \forall n,$$

by Lemma A.3 and Theorem A.2, so that

$$\|\hat{Z}^{(n)}(s)\| \leq 1 \quad \forall \operatorname{Re} s > 0, \quad \forall n.$$

Further, for any $x = [x_j] \in l^2$,

$$\|[Z(s) - Z^{(n)}(s)]x\| \rightarrow 0 \quad \text{as } n \rightarrow \infty \quad \forall \operatorname{Re} s > 0,$$

and so, for any $x = [x_j]$ and $y = [y_j]$ with $\|x\| = \|y\| = 1$,

$$\begin{aligned} |y^*[\hat{Z}^{(n)}(s) - \hat{Z}(s)]x| &= |y^*\hat{Z}^{(n)}(s)[Z(s) - Z^{(n)}(s)]\hat{Z}(s)x| \\ &\leq \|[Z(s) - Z^{(n)}(s)]\hat{Z}(s)x\| \\ &\rightarrow 0 \quad \text{as } n \rightarrow \infty \quad \forall \operatorname{Re} s > 0. \end{aligned}$$

Thus, for each element of $\hat{Z}(s)$, we have a sequence of approximants, holomorphic and uniformly bounded in $\operatorname{Re} s > 0$, convergent for each s in $\operatorname{Re} s > 0$, whence, by Vitali's and Weierstrass' theorems [6], the elements of $\hat{Z}(s)$ are holomorphic in $\operatorname{Re} s > 0$.

Hence, by the same argument as for $\|\hat{Z}^{(n)}(s)\|$,

$$\|\hat{Z}(s)\| \leq 1 \quad \forall \operatorname{Re} s > 0. \quad \text{Q.E.D.}$$

Acknowledgments. The author has benefited greatly from conversations and correspondence with many friends and colleagues, notably Professors H. H. Rosenbrock, C. Storey, J. C. Willems, C. A. Desoer, and Mr. A. I. Mees.

REFERENCES

- [1] G. ZAMES, *On the input-output stability of time-varying nonlinear feedback systems*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 228–238, 465–476.
- [2] H. H. ROSENBRICK, *Multivariable circle theorems*, Recent Mathematical Developments in Control, D. J. Bell, ed., Academic Press, New York, 1973, pp. 345–365.
- [3] P. A. COOK, *Modified multivariable circle theorems*, Recent Mathematical Developments in Control, D. J. Bell, ed., Academic Press, New York, 1973, pp. 367–372.
- [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] R. E. A. C. PALEY AND N. WIENER, *Fourier Transforms in the Complex Domain*, American Mathematical Society Colloquium Publications, vol. 19, Providence, R.I., 1934, pp. 1–9.
- [6] G. SANSONE AND J. GERRETSEN, *Lectures on the Theory of Functions of a Complex Variable*, vol. I, P. Noordhof, Groningen, Netherlands, 1960, pp. 93–94, 100–103.
- [7] W. V. PARKER, *Characteristic roots and field of values of a matrix*, Bull. Amer. Math. Soc., 57 (1951), pp. 103–108.
- [8] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators*, vol. II, Interscience, New York, 1963, p. 924.
- [9] G. BACHMAN AND L. NARICI, *Functional Analysis*, Academic Press, New York, 1966, p. 386.

CHARACTERIZATION OF THE CONTROLLED STATES IN $W_2^{(1)}$ OF LINEAR HEREDITARY SYSTEMS*

H. T. BANKS,[†] MARC Q. JACOBS[‡] AND C. E. LANGENHOP[¶]

Abstract. The question of controlling a linear retarded functional differential equation from an initial function to a terminal function, both functions belonging to the Sobolev space $W_2^{(1)}$, is considered. Necessary and sufficient conditions for full controllability in this function space are derived. These conditions result in computable algebraic criteria. Null controllability in function space is also investigated and conditions that are necessary and sufficient are obtained. In the absence of full controllability, methods for characterizing the attainable set in $W_2^{(1)}$ are discussed in detail along with a number of illustrative examples.

1. Introduction. The main purpose of this paper is to examine some fundamental questions concerning the controllability of linear functional differential equations of retarded type. These equations take the form

$$(1.1) \quad \dot{x} = L(t, x_t) + B(t)u(t)$$

in Hale's notation [12]. Weiss [30] and Gabasov and Kirillova were among the early investigators of the controllability of systems of the above form. An account of Gabasov and Kirillova's work is now available in a research monograph [9]. We will not go into a detailed review of previous work on controllability of functional differential equations (FDEs) since Banks' survey paper [2] is still reasonably current.

Our point of view throughout is that of controlling the true states, x_t , of (1.1). Since the states x_t are functions, the questions must be framed in compatible state and control function spaces. We have elected to use L_2 as the class of admissible controllers and the Sobolev space $W_2^{(1)}$ as the state space. This has proved successful in [4], [17]. Loosely speaking, the compatibility of the state and control function spaces means that if the set of admissible controllers is the space L_p , then a natural state space is $W_p^{(1)}$, $1 \leq p \leq \infty$. Failure to acknowledge this explains why some others have only obtained density results (e.g., see [31]) or approximate controllability.

Theorem 3.1 completely characterizes systems (1.1) which are controllable on an interval $[t_0, t_1]$ (see the definition in § 2). When the theorem is specialized to the case where B is a constant $n \times m$ matrix, we obtain that (1.1) is controllable on $[t_0, t_1]$, if and only if B has rank n . Thus the controllability is purely algebraic and within the limitations of the definition (see § 2, $t_1 > t_0 + h$) is independent

* Received by the editors August 2, 1973, and in final revised form March 12, 1974.

[†] Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. This research was supported in part by the U.S. Army under Grant DA-ARO-D-31-124-71-G12, and in part by the U.S. Air Force under Grant AF-AFOSR-71-2078.

[‡] Department of Mathematics, University of Missouri, Columbia, Missouri 65201. This research was supported by the National Science Foundation under Grant GP-33882.

[¶] Division of Applied Mathematics, Brown University, Providence, Rhode Island, and Department of Mathematics, Southern Illinois University, Carbondale, Illinois 62901. This research was supported in part by the National Science Foundation under Grant GP-28931, and in part by the U.S. Air Force under Grant AF-AFOSR-71-2078.

of the interval $[t_0, t_1]$. This result is somewhat negative in that it shows that controllability imposes limitations on (1.1) which are too severe to cover many of the applications. That rank $B = n$ implies system (1.1) is controllable is well known [4], [9], [17], whereas the converse extends Lemma 3.1 of [17]. In fact, the integrability condition in that lemma (for the case when B depends on t) is inadequate to assure the stated conclusion. Rather, hypothesis (H3) in § 2 of the present paper is what is needed.

Our interest in controllability stems primarily from the fundamental role it plays in the resolution of a number of difficulties encountered in solving optimum settling problems [2], [3], [4], [16], [17]. As noted in [4], the negative aspects of Theorem 3.1 can be ameliorated by a more precise identification of the state space of (1.1). In that paper it was shown in the closure lemma that a certain closed subspace of $W_2^{(1)}$ was the appropriate state space in which to apply the various optimization tools of functional analysis. Thus the main object of our study will be the attainable set of system (1.1) with unlimited control. This is important even if the underlying optimization problem has control constraints. The chief thrust of the present paper is to obtain effective computable criteria for characterizing the attainable set of the controlled process (1.1).

In Theorem 3.3, where it is shown that (H2) and (H3) imply closedness of the attainable sets, we sharpen the closure lemma obtained in [4]. Example 3.3 shows that assumption (H2) cannot be eliminated. Theorem 3.2 establishes the connection between the closure of the attainable sets and null controllability, and Corollary 3.2 effectively identifies the null controllable processes.

In § 4 it is shown how the attainable set can be determined by using the Fredholm alternative. When these results are applied to systems satisfying (H1), (H2) and (H3) of § 2, Theorem 4.1 shows that the attainable set for (1.1) coincides with the trajectory set of an ordinary linear differential equation. This theorem and the Fredholm alternative are then used to give a simple means for computing the attainable set.

In §§ 5 and 6 we generalize results recently obtained by Minjuk [25]. The analysis is given for neutral systems of the form

$$(1.2) \quad \dot{x}(t) = A_{-1}\dot{x}(t-h) + A_0x(t) + A_1x(t-h) + Bu(t),$$

which are not covered by the model (1.1). The conclusions for neutral processes are obtained at no extra expense and so are included here for completeness. In contrast, a number of results in §§ 3 and 4 differ substantially in the neutral case; see, for example, Theorem 3.1 and Example 6.1. A paper will soon be forthcoming treating those aspects of the controllability of neutral systems which differ substantially from the retarded case. In Theorem 5.2 the attainable set $\mathcal{A}(0, t_1, 0)$ is implicitly defined as the set of solutions of a simplified boundary value problem involving only point data. In § 6 methods are given for using this reduction to explicitly determine the attainable set for (1.2). Here we also apply Theorem 5.2 to show that if $A_{-1} = 0$, A_1 is nonsingular and B is $n \times 1$, then any initial state for (1.2) at $t = 0$ which can be controlled to the zero state by some $t_1 > nh$ can be controlled to the zero state by an arbitrary $\tilde{t}_1 > nh$.

A number of examples are presented throughout the paper.

Some authors [5], [6] have suggested that it might be profitable to transform systems such as (1.1) into an abstract evolution equation prior to studying these systems. There is, of course, the question of computational effectiveness. But more important from a theoretical viewpoint with regard to optimal control of systems (1.1) to function space targets [3], [4], [17] is that such an approach can at most yield approximate controllability or density results (cf. [19], [29]). In [5], [6] the state space is $R^n \times L_2$ and the solution operator for (1.1) is compact. It is a well-known fact [27, p. 104] that if K is a compact linear operator from one B -space to another with range K closed, then the range of K is finite-dimensional.

2. Basic notation and definitions. The symbol $L_2([\alpha, \beta], R^p)$ denotes the usual Hilbert space of "square integrable" functions from $[\alpha, \beta]$ into R^p [8]. In all statements where measures are needed, Lebesgue measure is understood unless we explicitly state otherwise. We use $W_2^{(1)}([\alpha, \beta], R^p)$ to denote the Sobolev space consisting of all absolutely continuous functions $x: [\alpha, \beta] \rightarrow R^p$ with the property that $t \mapsto \dot{x}(t) = (dx/dt)(t)$ belongs to $L_2([\alpha, \beta], R^p)$. The space $W_2^{(1)}([\alpha, \beta], R^p)$ is a Hilbert space with inner product defined by

$$(2.1) \quad \langle x, y \rangle \equiv \langle x(\alpha), y(\alpha) \rangle + \int_{\alpha}^{\beta} \langle \dot{x}(t), \dot{y}(t) \rangle dt,$$

for $x, y \in W_2^{(1)}([\alpha, \beta], R^p)$, where the angle brackets on the right-hand side of (2.1) represent the usual inner product in R^p (vectors in R^p are written as $p \times 1$ matrices). If T is a linear mapping, then $\text{Im } T$ and $\ker T$ will be used respectively for the image (range) and kernel (null space) of the transformation. The adjoint of the transformation T is represented by T^* ; in case T is a matrix T^* is the transposed matrix. If S is a subset of a Hilbert space H , then S^{\perp} denotes the orthogonal complement of S [8, p. 249].

The symbol \mathcal{L}_{pq} will be used for the collection of all real $p \times q$ matrices with a suitable matrix norm.

Let $h > 0$ be given. For functions $x: [t_0 - h, t_1] \rightarrow R^n$ and $t \in [t_0, t_1]$, we use x_t to denote the function on $[-h, 0]$ defined by

$$(2.2) \quad x_t(\theta) = x(t + \theta), \quad -h \leq \theta \leq 0.$$

We shall consider systems of linear functional differential equations of retarded type having the form

$$(2.3) \quad \dot{x}(t) = L(t, x_t) + B(t)u(t),$$

where $B(t) \in \mathcal{L}_{nm}$, $t \mapsto B(t)$ is continuous, and for each fixed $t \in R$ the linear operator $\phi \mapsto L(t, \phi)$, $\phi \in W_2^{(1)}([-h, 0], R^n)$ has the form

$$(2.4) \quad L(t, \phi) = \int_{-h}^0 d_{\theta} \eta(t, \theta) \phi(\theta).$$

Here the integral is in the Lebesgue–Stieltjes sense and $(t, \theta) \mapsto \eta(t, \theta)$, $(t, \theta) \in R \times R$, is a mapping with values in \mathcal{L}_{nm} . We further assume that $t \mapsto \eta(t, \theta)$, $t \in R$, is continuous for each fixed $\theta \in [-h, 0]$ and $\theta \mapsto \eta(t, \theta)$ is of bounded variation on $[-h, 0]$ for each fixed $t \in R$. Moreover,

$$\eta(t, \theta) = 0, \quad \theta \geq 0, \quad \eta(t, \theta) = \eta(t, -h), \quad \theta \leq -h,$$

and $\theta \mapsto \eta(t, \theta)$ is left continuous on $(-h, 0)$. The variation function satisfies the inequality

$$\text{var}_{\theta \in R} \eta(t, \theta) \leq \rho(t), \quad t \in R,$$

where $\rho(t)$ is locally integrable on R .

The most important special case of (2.3) known to be useful in the applications is when

$$(2.5) \quad L(t, \phi) = \sum_{i=0}^N A_i(t) \phi(-h_i) + \int_{-h}^0 A(t, \theta) \phi(\theta) d\theta$$

for $\phi \in W_2^{(1)}([-h, 0], R^n)$ and

$$(2.6) \quad h_0 = 0 < h_1 < h_2 < \cdots < h_N \leq h.$$

In (2.5) we shall always assume that $A_i, A(\cdot, \theta)$ for fixed $\theta \in [-h, 0], i = 0, 1, \dots, N$, are continuous mappings of R into \mathcal{L}_m , and for fixed $t \in R$, $A(t, \cdot)$ is integrable on $[-h, 0]$. Such systems do, of course, fulfill the requirements in [1] and [12, pp. 80–81].

If t_0, t_1 are given real numbers, and $\phi \in W_2^{(1)}([-h, 0], R^n), u \in L_2([t_0, t_1], R^m)$, then there is a unique absolutely continuous function $x(\cdot, t_0, \phi, u) = x: [t_0 - h, t_1] \rightarrow R^n$ satisfying (2.3) a.e. on $[t_0, t_1]$ and the initial condition

$$(2.7) \quad x_{t_0} = \phi.$$

Moreover, $x_t(\cdot, t_0, \phi, u) \in W_2^{(1)}([-h, 0], R^n)$ for $t \in [t_0, t_1]$. When the initial time t_0 is understood it will be suppressed in the above notation. The attainable set for system (2.3) is defined by the following equation:

$$\mathcal{A}(t_0, t_1, \phi) = \{x_{t_1}(\cdot, t_0, \phi, u) | u \in L_2([t_0, t_1], R^m)\},$$

where $t_1 > t_0$ and $\phi \in W_2^{(1)}([-h, 0], R^n)$.

System (2.3) will be called *controllable* (respectively, *Euclidean controllable*) on the interval $[t_0, t_1]$ if for each $\phi, \psi \in W_2^{(1)}([-h, 0], R^n)$ (resp., $\phi \in W_2^{(1)}([-h, 0], R^n), \xi \in R^n$) there is a controller $u \in L_2([t_0, t_1], R^m)$ such that $x_{t_1}(\cdot, t_0, \phi, u) = \psi$ (respectively, $x(t_1, t_0, \phi, u) = \xi$). System (2.3) is called *null controllable* (respectively, *Euclidean null controllable*) on $[t_0, t_1]$ if the conditions of the preceding definitions can be met with $\psi \equiv 0$ (respectively, $\xi = 0$). The qualifying phrase “on the interval $[t_0, t_1]$ ” will be dropped in the two definitions of types of controllability to functions when the kind of controllability obtains on *every interval* $[t_0, t_1]$ with $t_1 > t_0 + h$. Similarly, dropping the qualifying phrase “on the interval $[t_0, t_1]$ ” when referring to the two types of Euclidean controllability signifies that the condition holds for every interval $[t_0, t_1], t_1 > t_0$.

Let $X(t, s)$ be the fundamental matrix for the homogeneous equation $\dot{x}(t) = L(t, x_t)$. That is, $X(t, s)$ is the $n \times n$ matrix solution [1] to

$$(2.8) \quad X(t, s) + \int_s^t X(t, \alpha) \eta(\alpha, s - \alpha) d\alpha = I$$

for $s \leq t$, and

$$(2.9) \quad X(t, s) = 0, \quad s > t \quad \text{and} \quad X(t, t) = I,$$

where I is the $n \times n$ identity matrix. Then we have

$$(2.10) \quad \begin{aligned} x(t, t_0, \phi, u) &= x(t, t_0, \phi, 0) + \int_{t_0}^t X(t, s)B(s)u(s) ds, \\ &= x(t, t_0, \phi, 0) + x(t, t_0, 0, u) \end{aligned}$$

for $t_0 \leq t \leq t_1$.

We now list some hypotheses to which we shall refer in the sequel.

(H1) The matrix $G(t_0, t_1 - h) = \int_{t_0}^{t_1 - h} X(t_1 - h, s)B(s)B^*(s)X^*(t_1 - h, s) ds$ has rank n .

Remark 2.1. We note that under the stated assumptions for η in (2.4) we can write

$$(2.11) \quad L(t, \phi) = A_0(t)\phi(0) + \int_{-h}^0 d_\theta \tilde{\eta}(t, \theta)\phi(\theta),$$

where $A_0: R \rightarrow \mathcal{L}_{nn}$ is continuous, $\eta(t, 0^-) = -A_0(t)$, and $\tilde{\eta}$ satisfies for $t \in R$

$$(2.12) \quad \tilde{\eta}(t, \theta) = \begin{cases} 0, & \theta \geq 0, \\ \eta(t, \theta) + A_0(t), & \theta < 0. \end{cases}$$

(H2) Let $B^\dagger(t)$ denote the Moore–Penrose generalized (or pseudo-) inverse of $B(t)$, $t \in R$ (see [22]). For almost every $t \in [t_1 - h, t_1]$,

$$B(t)B^\dagger(t)\tilde{\eta}(t, \theta) = \tilde{\eta}(t, \theta)$$

if $-h \leq \theta \leq 0$.

Remark 2.2. The condition (H2) is the same as requiring $\text{Im } \tilde{\eta}(t, \alpha - t) \subset \text{Im } B(t)$ for the stated values of t, α .

(H3) The map $t \mapsto B^\dagger(t)$, $t \in R$, is essentially bounded on $[t_1 - h, t_1]$.

Remark 2.3. The mapping $t \mapsto B^\dagger(t)$, $t \in R$, is measurable, since B is continuous. This is an immediate consequence of Showalter's approximation theorem [28], or it can be shown by elementary methods.

We shall denote by (H1'), (H2') and (H3') the respective strengthenings of hypotheses (H1), (H2) and (H3) to require the stated conditions to be satisfied for every choice of t_0, t_1 with $t_1 > t_0 + h$. Furthermore, the statement “(H3) holds on the interval $[\alpha, \beta]$ ” will be used in the sequel to mean that $t \mapsto B^\dagger(t)$ is essentially bounded on $[\alpha, \beta]$.

The assumptions (H1), (H2) and (H3) cover a wide class of systems (2.3). In fact (see Theorem 3.2 below), (H1') and (H2') are properties enjoyed by every null controllable system.

3. Controllability and properties of the attainable set. The first theorem and its corollaries completely characterize systems (2.3) which are controllable, and show that the class of controllable systems does not include a large number of those arising in applications. The proofs involve extensions of ideas previously discussed in [17, Lemma 3.1] and [4, Remark 3.2].

THEOREM 3.1. *Suppose (H3) holds. The system (2.3) is controllable on an interval $[t_0, t_1]$ with $t_1 > t_0 + h$ if and only if*

$$(i) \quad \text{rank } B(t) = n \quad \text{on } [t_1 - h, t_1].$$

Proof. The controllability assumption implies that for arbitrary

$$\psi \in W_2^{(1)}([-h, 0], R^n)$$

there is a $u \in L_2([t_0, t_1], R^m)$ such that $x_{t_1}(\cdot, t_0, 0, u) = \psi$. For brevity let $x = x(\cdot, t_0, 0, u)$; then

$$(3.1) \quad \dot{\psi}(t - t_1) = L(t, x_t) + B(t)u(t)$$

a.e. on $[t_1 - h, t_1]$, where $x(\tau) = \psi(\tau - t_1)$, $t_1 - h \leq \tau \leq t_1$, on the right-hand side of (3.1).

Suppose that there is a measurable $E \subset [t_1 - h, t_1]$ such that $m(E) > 0$ (m is Lebesgue measure), and

$$(3.2) \quad \text{rank } B(t) < n, \quad t \in E.$$

Since $\text{rank } B(t) = \text{rank } B(t)B^*(t)$, we have that 0 is an eigenvalue of $B(t)B^*(t)$, $t \in E$. Consequently an eigenvector $e(t)$, $t \in E$, corresponding to the zero eigenvalue can be chosen such that $|e(t)|^2 = \langle e(t), e(t) \rangle = 1$, $t \in E$ and $t \mapsto e(t)$, $t \in E$, is measurable [26]. The functions $t \rightarrow L(t, \phi)$ are measurable (in fact continuous) for each fixed $\phi \in W_2^{(1)}([-h, 0], R^n)$ and the functions $\phi \rightarrow L(t, \phi)$ are uniformly continuous for each fixed $t \in [t_1 - h, t_1]$. Thus the generalized Scorza–Dragoni theorem [15, Thm. 2.1] implies that there is a measurable $F' \subset E$, $m(F') > 0$ such that $L: F' \times W_2^{(1)}([-h, 0], R^n) \rightarrow R^n$ is continuous. By Lusin's theorem there is a measurable $F \subset F'$, $m(F) > 0$ such that $e|_F$ is continuous. We note that

$$\langle e(t), B(t)B^*(t)e(t) \rangle = 0 = \langle B^*(t)e(t), B^*(t)e(t) \rangle,$$

so that $e^*(t)B(t) = 0$, $t \in F$. Now let α be an arbitrary function in $L_2(F, R)$. Define

$$f(t) = \begin{cases} \alpha(t)e(t), & t \in F, \\ 0, & t \notin F, \end{cases}$$

and

$$(3.3) \quad \psi(t - t_1) = \int_{t_1-h}^t f(s) ds, \quad t \in [t_1 - h, t_1],$$

and apply the controllability assumption to get (3.1) for some choice of $u \in L_2([t_0, t_1], R^m)$ depending on ψ in (3.3). If we multiply both sides of (3.1) by $e^*(t)$, $t \in F$, we get

$$(3.4) \quad \alpha(t) = e^*(t)L(t, x_t) \quad \text{a.e. on } F.$$

Since the right-hand side of (3.4) is continuous, we have proved that any $\alpha \in L_2(F, R)$ is equal almost everywhere to a continuous function. This is a contradiction, so that (3.2) is false, i.e. $\text{rank } B(t) = n$ a.e. on $[t_1 - h, t_1]$.

Hypothesis (H3) and the fact that $\text{rank } B(t) = n$ a.e. on $[t_1 - h, t_1]$ imply that condition (i) of Theorem 3.1 holds. In order to show this it suffices to prove that $\det B(t)B^*(t) \neq 0$ for every t in $[t_1 - h, t_1]$. From the above rank condition we have the expression

$$B^\dagger(t) = B^*(t) \text{adj}(B(t)B^*(t))/\det(B(t)B^*(t)) \quad \text{a.e. on } [t_1 - h, t_1],$$

where "adj" denotes the algebraic adjoint. Since B^\dagger is essentially bounded and given by this expression for all t such that $\det(B(t)B^*(t)) \neq 0$, we find that there is a constant $\tilde{M} > 0$ such that

$$|A(t)| \leq \tilde{M} \quad \text{a.e. on } [t_1 - h, t_1],$$

where $A(t) \equiv B^\dagger(t)B^\dagger(t) = \text{adj}(B(t)B^*(t))/\det(B(t)B^*(t))$ for a.e. t . Hence there exists $M > 0$ such that $|\det A(t)| \leq M$ a.e. on $[t_1 - h, t_1]$. However,

$$\begin{aligned} \det A(t) &= [\det(B(t)B^*(t))]^{-n} \det[\text{adj}(B(t)B^*(t))] \\ &= [\det(B(t)B^*(t))]^{-1} \quad \text{a.e. } t \in [t_1 - h, t_1]. \end{aligned}$$

Thus from the continuity of B and the inequality

$$|[\det(B(t)B^*(t))]^{-1}| \leq M \quad \text{a.e. } t \in [t_1 - h, t_1],$$

we conclude that, in fact, $\det(B(t)B^*(t))$ does not vanish for any $t \in [t_1 - h, t_1]$.

Before proving the converse statement of Theorem 3.1, we state a lemma that will be useful here and in the sequel.

LEMMA 3.1. *System (2.3) is Euclidean controllable on $[t_0, \tau]$ if and only if $\text{rank } G(t_0, \tau) = n$.*

A number of authors have given algebraic criteria for $G(t_0, \tau)$ to be nonsingular (e.g., [9], [23], [31] or the survey paper [2]). The proof of Lemma 3.1 is an obvious analogue of the corresponding situation for ordinary differential equations [14] and will be omitted. We thus continue with the proof of Theorem 3.1.

Suppose that (i) and (H3) hold. Then $\text{rank } B(t_1 - h) = n$ and it is not difficult to show that this implies (H1). Let $t_1 > t_0 + h$, $\phi, \psi \in W_2^{(1)}([-h, 0], R^n)$. By Lemma 3.1 there is a $u \in L_2([t_0, t_1 - h], R^m)$ such that

$$(3.5) \quad x(t_1 - h, t_0, \phi, u) = \psi(-h).$$

We will extend u and $x = x(\cdot, t_0, \phi, u)$ to the interval $[t_0, t_1]$ so that (3.1) and $\psi(t - t_1) = x(t)$, $t_1 - h \leq t \leq t_1$, are satisfied and the proof of Theorem 3.1 will be completed. To do this we observe that for any absolutely continuous $y: [t_0 - h, t_1] \rightarrow R^n$, we have from (2.4)

$$\begin{aligned} L(t, y_t) &= \int_{-h}^0 d_\theta \eta(t, \theta) y(t + \theta) \\ &= \eta(t, 0) y(t) - \eta(t, -h) y(t - h) - \int_{t-h}^t \eta(t, \alpha - t) \dot{y}(\alpha) d\alpha \\ (3.6) \quad &= -\eta(t, -h) y(t - h) - \int_{t-h}^{t_1-h} \eta(t, \alpha - t) \dot{y}(\alpha) d\alpha \\ &\quad - \int_{t_1-h}^t \eta(t, \alpha - t) \dot{y}(\alpha) d\alpha, \end{aligned}$$

for $t_1 - h \leq t \leq t_1$. In view of (i) and (3.6), we can define

$$(3.7) \quad u(t) = B^{\dagger}(t) \left\{ \dot{\psi}(t - t_1) + \eta(t, -h)x(t - h) + \int_{t-h}^{t_1-h} \eta(t, \alpha - t)\dot{x}(\alpha) d\alpha \right. \\ \left. + \int_{t_1-h}^t \eta(t, \alpha - t)\dot{\psi}(\alpha - t_1) d\alpha \right\},$$

for $t_1 - h \leq t \leq t_1$, and we get that (3.1) is satisfied. That this u is square-integrable on $[t_1 - h, t_1]$ follows from (H3) and the smoothness properties of $x, \dot{\psi}$.

Remark 3.1. The difficulties that one may encounter in the absence of (H3) can be illustrated by a simple example. Consider the scalar differential equation

$$\dot{x}(t) = B(t)u(t), \quad 0 \leq t \leq 1,$$

with $B(t) = t^{1/2}$, $0 \leq t \leq 1$. Define

$$(3.8) \quad u_v(t) = t^{(2-v)/2v}, \quad 0 \leq t \leq 1.$$

Then $u_v \in L_2([0, 1], R)$, and the corresponding responses with zero initial data are

$$(3.9) \quad x_v(t) = \frac{v}{v+1} t^{(v+1)/v}, \quad 0 \leq t \leq 1, \quad v = 1, 2, 3, \dots$$

Evidently $x_v \rightarrow x$ in $W_2^{(1)}([0, 1], R)$, where $x(t) \equiv t$, $0 \leq t \leq 1$. This limit trajectory x is clearly not attainable by the given differential equation.

Remark 3.2. Note that in (3.1) the term $L(t, x_i)$ depends on the choice of u on the interval $[t_0, t_1 - h]$, so that the controllability assumption does not say that independent choices of u can be made on the intervals $[t_0, t_1 - h]$ and $[t_1 - h, t_1]$. If this were possible we could conclude immediately from (3.1) that $\text{rank } B(t) = n$ a.e. on $[t_1 - h, t_1]$. Thus the proof of Theorem 3.1 is somewhat more involved than one might at first expect.

In the event that the mapping $t \rightarrow B(t)$ is constant, the above results yield very simple conditions which we state as follows.

COROLLARY 3.1. *Suppose the matrix function B in (2.3) is constant. A necessary and sufficient condition that (2.3) be controllable on $[t_0, t_1]$ is that $\text{rank } B = n$.*

We point out that for autonomous systems the above shows that controllability on $[t_0, t_1]$ is independent of the length of the interval (as long as $t_1 > t_0 + h$, of course). As we shall see below, this differs markedly from the situation for null controllability.

The developments above show that controllability is too narrow a concept for systems (2.3). For example, it rules out n th order scalar retarded equations with a single control (cf. [4, Example 3.1, Remark 3.4]). Some type of controllability is going to have to be present if we are to get very far in our study of optimal control problems with systems guided by (2.3). Since systems (2.3) will generally not be controllable, we shall direct our efforts mainly toward identifying as completely as possible the set $\mathcal{A}(t_0, t_1, \phi)$ and some of its properties, and thereby

salvage much of what a "controllability assumption" accomplishes in an optimization problem (cf. [4]).

The example in Remark 3.1 leads one to suspect that for $\mathcal{A}(t_0, t_1, \phi)$ to be closed in $W_2^{(1)}([-h, 0], R^n)$, condition (H3) must be satisfied. (We conjecture that such a theorem is true but have not succeeded in establishing such a result.) Under this therefore not unreasonable assumption, Theorem 3.2 given below shows that most null-controllable processes must satisfy (H1') and (H2').

We shall call (2.3) a *system with strict retardations* (or say simply that (2.3) is *strictly retarded*) if the mapping $\tilde{\eta}$ in (2.12) satisfies the additional requirement that there is a δ , $0 < \delta < h$, such that $\tilde{\eta}(t, \theta) = 0$ for $-\delta \leq \theta \leq 0$, $t \in R$.

The following lemma, in addition to being of interest in itself, will be needed in the sequel.

LEMMA 3.2. *Suppose (2.3) is strictly retarded. Then system (2.3) is Euclidean null controllable if and only if (H1') is satisfied.*

Proof. Hypothesis (H1') clearly implies Euclidean null controllability (see Lemma 3.1).

Conversely, suppose (2.3) is strictly retarded and Euclidean null controllable. Then there exists $\varepsilon > 0$ such that, for any t_1 , $G(t_1 - \varepsilon, t_1)$ has rank n . This follows from the fact that if $0 < \varepsilon < \delta$, Euclidean null controllability of (2.3) implies Euclidean null controllability on $[t_1 - \varepsilon, t_1]$ of the ordinary differential equation system

$$(3.10) \quad \dot{z}(t) = A_0(t)z(t) + B(t)u(t).$$

Furthermore, it is well known [14, p. 92] that a necessary and sufficient condition for this to be true is that

$$(3.11) \quad M(t_1 - \varepsilon, t_1) \equiv \int_{t_1 - \varepsilon}^{t_1} \Phi(t_1, s)B(s)B^*(s)\Phi^*(t_1, s) ds$$

have rank n , where $s \mapsto \Phi(t_1, s)$ satisfies

$$(3.12) \quad \begin{aligned} \dot{\Phi}(t_1, s) &= -\Phi(t_1, s)A_0(s), & s \in [t_1 - \varepsilon, t_1], \\ \Phi(t_1, t_1) &= I. \end{aligned}$$

But under the assumption that (2.3) is strictly retarded and $0 < \varepsilon < \delta$, the equation (2.8) can be equivalently written

$$(3.13) \quad \dot{X}(t_1, s) = -X(t_1, s)A_0(s), \quad s \in [t_1 - \varepsilon, t_1],$$

so that we find $G(t_1 - \varepsilon, t_1) = M(t_1 - \varepsilon, t_1)$ has rank n .

If $t_0, t_1 \in R$ are now chosen subject only to $t_1 > t_0$, then we get for $0 < \varepsilon < \delta$,

$$(3.14) \quad G(t_0, t_1) = G(t_1 - \varepsilon, t_1) + \int_{t_0}^{t_1 - \varepsilon} X(t_1, s)B(s)B^*(s)X^*(t_1, s) ds.$$

Both terms on the right-hand side of (3.14) are positive semidefinite, with $G(t_1 - \varepsilon, t_1)$ actually positive definite. Consequently (H1') must be satisfied.

We comment that the above lemma is true in the case where (2.3) is autonomous even without the restriction that (2.3) be strictly retarded. Since the proof involves slightly more technical arguments and since strictly retarded systems

include most systems of interest (in particular, all nonautonomous differential-difference equations are strictly retarded), we shall not pursue the development of this modification of the lemma here.

PROPOSITION 3.1. *Hypotheses (H1'), (H2') and (H3') imply that (2.3) is null controllable.*

Proof. Suppose $t_1 > t_0 + h$ is given and (H1') and (H2') are satisfied (in fact only (H1) and (H2) for this choice of t_0, t_1 need be assumed). According to Lemma 3.1, if $\phi \in W_2^{(1)}([-h, 0], R^n)$ there is a $u \in L_2([t_0, t_1 - h], R^m)$ such that

$$(3.15) \quad x(t_1 - h, t_0, \phi, u) = 0.$$

Upon applying integration by parts to (2.4) and using (2.12), equation (2.3) takes the form

$$(3.16) \quad \dot{x}(t) = A_0(t)x(t) - \tilde{\eta}(t, -h)x(t-h) - \int_{t-h}^t \tilde{\eta}(t, \alpha - t)\dot{x}(\alpha) d\alpha + B(t)\bar{u}(t),$$

for $\bar{u} \in L_2([t_0, t_1], R^m)$ with $\bar{u}|_{[t_0, t_1-h]} = u$, where u is as given in (3.15). We may use (H2) to write (3.16) in the form

$$(3.17) \quad \begin{aligned} \dot{x}(t) = & A_0(t)x(t) + B(t) \left\{ \bar{u}(t) - B^\dagger(t)\tilde{\eta}(t, -h)x(t-h) \right. \\ & \left. - B^\dagger(t) \int_{t-h}^{t_1-h} \tilde{\eta}(t, \alpha - t)\dot{x}(\alpha) d\alpha \right\} - \int_{t_1-h}^t \tilde{\eta}(t, \alpha - t)\dot{x}(\alpha) d\alpha \end{aligned}$$

a.e. on $[t_0, t_1]$. Thus if we define

$$(3.18) \quad \bar{u}(t) = \begin{cases} u(t), & t_0 \leq t \leq t_1 - h, \\ B^\dagger(t) \left\{ \tilde{\eta}(t, -h)x(t-h) + \int_{t-h}^{t_1-h} \tilde{\eta}(t, \alpha - t)\dot{x}(\alpha) d\alpha \right\}, & t_1 - h < t \leq t_1, \end{cases}$$

then (H3) implies $\bar{u} \in L_2([t_0, t_1], R^m)$. We see at once from (3.15), (3.17), (3.18) that

$$x_{t_1}(\cdot, t_0, \phi, \bar{u}) = 0,$$

so that (2.3) is null controllable.

PROPOSITION 3.2. *Let B satisfy (H3'), and let (2.3) be a system with strict retardations. Then (H2') is satisfied whenever (2.3) is null controllable.*

Proof. Take $0 < \delta < h$ as in the definition of a system with strict retardations. Pick $t_0 \in R$ and set $t_1 = t_0 + h + \delta/2$. Then for any $\phi \in W_2^{(1)}([-h, 0], R^n)$ there is a $u \in W_2^{(1)}([-h, 0], R^n)$ such that

$$(3.19) \quad x_{t_1}(\cdot, t_0, \phi, u) = 0.$$

Using (3.19) and (2.3) (use the form of (2.3) given in (3.16)), we obtain

$$\begin{aligned} B(t)u(t) = & \tilde{\eta}(t, -h)\phi(t - t_0 - h) + \int_{t-h}^{t_0} \tilde{\eta}(t, \alpha - t)\dot{\phi}(\alpha - t_0) d\alpha \\ & + \int_{t_0}^{t_0+\delta/2} \tilde{\eta}(t, \alpha - t)\dot{x}(\alpha) d\alpha \end{aligned}$$

a.e. $t \in [t_0 + \delta/2, t_0 + h]$. If we restrict t to the interval $[t_0 + \delta/2, t_0 + \delta]$, then

this becomes

$$(3.20) \quad B(t)u(t) = \tilde{\eta}(t, -h)\phi(t - t_0 - h) + \int_{t-h}^{t_0} \tilde{\eta}(t, \alpha - t)\dot{\phi}(\alpha - t_0) d\alpha.$$

Let ϕ be any constant function on $[-h, 0]$ with values in R^n . Then (3.20) implies

$$(3.21) \quad \text{Im } \tilde{\eta}(t, -h) \subset \text{Im } B(t)$$

a.e. $t \in [t_0 + \delta/2, t_0 + \delta]$. Since $t_0 \in R$ was arbitrary, (3.21) holds for almost all $t \in R$. Now let $\{\xi_\mu\}$ be a sequence which is dense in $[-h, 0]$ and let $e \in R^n$ be arbitrary. Define

$$\phi_{v\mu}(\theta) \equiv v \int_{-h}^{\theta} e \chi_{[\xi_\mu - 1/v, \xi_\mu]}(\sigma) d\sigma, \quad -h \leq \theta \leq 0,$$

where χ_E denotes the characteristic function of a set E . Then $\phi_{v\mu} \in W_2^{(1)}([-h, 0], R^n)$ and there is a sequence of controllers $u_{v\mu}$ such that (3.19) holds with $\phi = \phi_{v\mu}$ and $u = u_{v\mu}$, $v, \mu = 1, 2, 3, \dots$. Moreover, there is a set E_0 of measure zero such that $x(t, t_0, \phi_{v\mu}, u_{v\mu})$ satisfies the differential equation (2.3) for all $t \in [t_0, t_1] \setminus E_0$ and $v, \mu = 1, 2, 3, \dots$. Applying (3.20), we get

$$(3.22) \quad B(t)u_{v\mu}(t) = \tilde{\eta}(t, -h)\phi_{v\mu}(t - t_0 - h) + v \int_{E_{v\mu}(t)} \tilde{\eta}(t, \alpha - t)e d\alpha,$$

for $t \in [t_0 + \delta/2, t_0 + \delta] \setminus E_0$, where

$$E_{v\mu}(t) \equiv [t - h, t_0] \cap [t_0 + \xi_\mu - 1/v, t_0 + \xi_\mu].$$

Since $\theta \mapsto \tilde{\eta}(t, \theta)$ is left continuous on $(-h, 0)$, (3.22) implies

$$(3.23) \quad \lim_{v \rightarrow \infty} B(t)u_{v\mu}(t) = \tilde{\eta}(t, \xi_\mu + t_0 - t)e$$

for $t \in \{[t_0 + \delta/2, t_0 + \delta] \setminus E_0\} \cap [t_0, \xi_\mu + t_0 + h)$. Restrict ξ_μ to the interval $(-h + \delta, -\delta/2)$ and recall that $e \in R^n$ was arbitrary so that Corollary 3.1 together with (3.23) imply

$$\text{Im } \tilde{\eta}(t, \alpha_\mu - t) \subset \text{Im } B(t),$$

$t \in [t_0 + \delta/2, t_0 + \delta] \setminus E_0$, where $\alpha_\mu \equiv \xi_\mu + t_0$. Since $\{\xi_\mu\}$ is dense in $[-h, 0]$ and $\theta \mapsto \tilde{\eta}(t, \theta)$ is left continuous on $(-h, 0)$, we get

$$(3.24) \quad \text{Im } \tilde{\eta}(t, \alpha - t) \subset \text{Im } B(t),$$

$t \in [t_0 + \delta/2, t_0 + \delta] \setminus E_0$ and $t - h \leq \alpha \leq t - \delta$. The above $t_0 \in R$ was arbitrary so we infer that (3.24) holds for almost every $t \in R$ and each α in the region $t - h \leq \alpha \leq t - \delta$. However, $\tilde{\eta}(t, \theta) = 0$ for $-\delta < \theta \leq 0$ so that (3.24) actually holds for $t - h \leq \alpha \leq t$. Combining this with (3.21), we get that (H2') is satisfied.

THEOREM 3.2. *Let (H3') be satisfied and suppose (2.3) has strict retardations. Then (H1') and (H2') are necessary and sufficient conditions for (2.3) to be null controllable.*

Proof. Null controllability implies Euclidean null controllability so that (H1') follows from Lemma 3.2. The property (H2') results from Proposition 3.2. The converse follows from Proposition 3.1.

Next we proceed to give some easily computed algebraic conditions for (2.3) to be null controllable.

COROLLARY 3.2. *Let (H3') be satisfied, and let (2.3) be strictly retarded. The following statements are equivalent:*

- (i) *System (2.3) is null controllable;*
- (ii) *Hypotheses (H2') and (H1') are satisfied;*
- (iii) *(H2') is satisfied and system*

$$(3.25) \quad \dot{z}(t) = A_0(t)z(t) + B(t)f(t)$$

is Euclidean null controllable,

- (iv) *(H2') is satisfied and the matrix*

$$M(t_0, t_1) \equiv \int_{t_0}^{t_1} \Phi(t_1, s)B(s)B^*(s)\Phi^*(t_1, s) ds$$

has rank n whenever $t_1 > t_0$, where $\Phi(t, s)$ is the fundamental matrix for the homogeneous system

$$(3.26) \quad \dot{z}(t) = A_0(t)z(t).$$

Proof. The equivalence of (i) and (ii) is a restatement of Theorem 3.2. Suppose (i) is satisfied. Let $t_1 > t_0$ be given. Then for $\phi \in W_2^{(1)}([-h, 0], R^n)$ there is a controller $u \in L_2([t_0, t_1 + h], R^m)$ such that

$$x_{t_1+h}(\cdot, t_0, \phi, u) = 0.$$

Using (2.3) in the form of equation (3.17), we have that if $x(t) \equiv x(t, t_0, \phi, u)$, then

$$\dot{x}(t) = A_0(t)x(t) + B(t)\left\{u(t) - B^\dagger(t)\tilde{\eta}(t, -h)x(t-h) - B^\dagger(t)\int_{t-h}^{t_1}\tilde{\eta}(t, \alpha-t)\dot{x}(\alpha)d\alpha\right\}$$

for $t_1 \leq t \leq t_1 + h$ and $x(t_1) = 0$. Hence (3.25) is Euclidean null controllable on $[t_0, t_1]$. That (3.25) being null controllable (Euclidean) is equivalent to rank $M(t_0, t_1) = n, t_1 > t_0$, has already been pointed out (see Lemma 3.2). As the proof of Lemma 3.2 shows, this implies that $G(t_1 - h - \varepsilon, t_1 - h)$ has rank n , when $0 < \varepsilon < \delta$. It follows from

$$G(t_0, t_1 - h) = G(t_1 - h - \varepsilon, t_1 - h) + \int_{t_0}^{t_1 - h - \varepsilon} X(t_1 - h, s)B(s)B^*(s)X^*(t_1 - h, s) ds,$$

where $t_1 > t_0 + h, 0 < \varepsilon < \delta$, that $G(t_0, t_1 - h)$ has rank n and hence, by Lemma 3.2, (2.3) is Euclidean null controllable. This completes the proof of the corollary.

For autonomous systems (i.e., L, B in (2.3) independent of t), we get the following obvious corollary.

COROLLARY 3.3. *Let (2.3) be strictly retarded. System (2.3) is null controllable if and only if (H2') is satisfied and*

$$(3.27) \quad \text{rank } [B, A_0B, \dots, A_0^{n-1}B] = n.$$

The proof is omitted. It suffices to note the equivalence of the two rank conditions in Corollary 3.2 (iv) and (3.27) [14, pp. 93, 94].

Example 3.1. We note that all systems of the form

$$(3.28) \quad \dot{x}(t) = \sum_{i=0}^N A_i x(t - h_i) + Bu(t),$$

where $0 = h_0 < h_1 < \dots < h_N$, A_i , $i = 0, 1, \dots, N$ are $n \times n$ constant matrices, and B is an $n \times m$ constant matrix, are strictly retarded. Hence, if one seeks purely algebraic criteria for system (3.28) to be null controllable, one expects to get null controllability on any interval $[t_0, t_1]$ with $t_1 > t_0 + h$. Corollary 3.3 answers completely the question of when systems (3.28) are null controllable. For example, all n th order scalar differential difference equations of retarded type are null controllable (cf. [4, Example 3.1]).

One might anticipate that for certain systems (3.28) one could obtain null controllability on $[0, t_2]$, but not on $[0, t_1]$, $0 < t_1 < t_2$. This can, indeed, happen, as the next example shows.

Example 3.2. For any state space dimension n one can construct systems of the form (3.28) (with one lag h_1 which we call h) which are null controllable on $[0, t_1]$ if $t_1 > nh$ but not null controllable on $[0, t_1]$ if $t_1 \leq nh$. To see this consider

$$(3.29) \quad \dot{x} = A_1 x(t - h) + Bu(t),$$

where A_1 is the matrix with all of its entries equal to zero except those on the first diagonal above the main diagonal and these are all equal to one, and B is an $n \times 1$ matrix whose first $n - 1$ entries are zero and n th entry is one. System (3.29) is null controllable on $[0, t_1]$, $t_1 > nh$, and is not null controllable on $[0, t_1]$ if $t_1 \leq nh$. This can either be verified directly or one can refer ahead to Examples 4.1 and 6.3 where arguments are given.

The next theorem gives an improvement of the closure lemma in [4]. The result is obtained here for general linear retarded systems without (H1).

THEOREM 3.3. *Let (H2) be satisfied. The attainable set $\mathcal{A}(t_0, t_1, \phi)$ is closed in $W_2^{(1)}([-h, 0], R^n)$ if (H3) is satisfied.*

Proof. Let us abbreviate $\mathcal{A}(t_0, t_1, 0)$ by \mathcal{A}_0 . It suffices for us to prove \mathcal{A}_0 is closed if (H3) is satisfied. Just as in the proof of Lemma 3.1 of [4], the sets

$$\mathcal{A}_0^p \equiv \{x_{t_1}(\cdot, t_0, 0, u) | u \in L_2([t_0, t_1], R^m), \|u\|_2 \leq p\},$$

$p = 1, 2, 3, \dots$, where $\|u\|_2$ denotes the norm of $u \in L_2([t_0, t_1], R^m)$, are all closed in $W_2^{(1)}([-h, 0], R^n)$. The theorem now follows from the next lemma by using essentially the same proof as in the proof of Lemma 3.1 of [4].

We use S_α^W, S_α^L to denote closed balls of radius α , center 0 in $W_2^{(1)}([-h, 0], R^n)$ and $L_2([t_0, t_1], R^m)$ respectively.

LEMMA 3.3. *Let (H2), (H3) be satisfied. Let \mathcal{S} be a subset of \mathcal{A}_0 with $\mathcal{S} \subset S_\alpha^W$ for some $\alpha > 0$. Then there is a set $\mathcal{L}(\mathcal{S})$ and a $K_\alpha > 0$ such that $\mathcal{L}(\mathcal{S}) \subset S_{K_\alpha}^L$ and*

$$\{x_{t_1}(\cdot, t_0, 0, u) | u \in \mathcal{L}(\mathcal{S})\} = \mathcal{S}.$$

Proof. Since $\mathcal{S} \subset \mathcal{A}_0$ there is a set $\mathcal{C}(\mathcal{S}) \subset L_2([t_0, t_1], R^m)$ such that

$$(3.30) \quad \mathcal{S} = \{x_{t_1}(\cdot, t_0, 0, u) | u \in \mathcal{C}(\mathcal{S})\}.$$

We need to verify that there is a bounded set $\mathcal{L}(\mathcal{S})$ such that (3.30) is true with $\mathcal{C}(\mathcal{S})$ replaced by $\mathcal{L}(\mathcal{S})$. Let G denote the mapping $u \mapsto x(t_1 - h, t_0, 0, u)$ of $L_2([t_0, t_1 - h], R^m)$ into R^n . The operator G is a bounded linear operator with closed range. As such there is a bounded generalized inverse

$$G^\dagger : R^n \rightarrow L_2([t_0, t_1 - h], R^m).$$

Since \mathcal{S} is bounded, the set

$$\{v|v = G^{\dagger}Gu, u \in \mathcal{C}(\mathcal{S})\}$$

is therefore a bounded subset of $L_2([t_0, t_1 - h], R^m)$. Using (H2), we write (2.3) on $[t_1 - h, t_1]$ as

$$(3.31) \quad \begin{aligned} \dot{x}(t) = & A_0(t)x(t) + \int_{t_1-h}^t d_{\alpha}\tilde{\eta}(t, \alpha - t)x(\alpha) \\ & + B(t)\left\{u(t) + B^{\dagger}(t) \int_{t-h}^{t_1-h} d_{\alpha}\tilde{\eta}(t, \alpha - t)x(\alpha)\right\}. \end{aligned}$$

Let $x_{t_1}(\cdot, t_0, \phi, u) = \psi$ and define

$$(3.32) \quad \tilde{u}(t) = \begin{cases} (G^{\dagger}Gu)(t), & \text{if } t_0 \leq t \leq t_1 - h, \\ B^{\dagger}(t)\left\{\dot{\psi}(t - t_1) - A_0(t)\psi(t - t_1) - \int_{t_1-h}^t d_{\alpha}\tilde{\eta}(t, \alpha - t)\psi(\alpha - t_1) \right. \\ \quad \left. - \int_{t-h}^{t_1-h} d_{\alpha}\tilde{\eta}(t, \alpha - t)x(\alpha)\right\}, & \text{if } t_1 - h \leq t \leq t_1. \end{cases}$$

In (3.32), $x(\alpha) = x(\alpha, t_0, \phi, G^{\dagger}Gu)$, $t_0 \leq \alpha \leq t_1 - h$. From (3.31) we get

$$\dot{\psi}(t - t_1) - A_0(t)\psi(t - t_1) - \int_{t_1-h}^t d_{\alpha}\tilde{\eta}(t, \alpha - t)\psi(\alpha - t_1)$$

is in the range of $B(t)$ for almost every $t \in [t_1 - h, t_1]$. Hence by (H2) and the fact that $BB^{\dagger}B = B$ it follows that

$$x_{t_1}(\cdot, t_0, \phi, \tilde{u}) = \psi.$$

The desired set is $\mathcal{L}(\mathcal{S}) = \{\tilde{u}|u \in \mathcal{C}(\mathcal{S})\}$, which is evidently a bounded subset of $L_2([t_0, t_1], R^m)$.

COROLLARY 3.4. *The mapping $\tilde{B}: L_2([\alpha, \beta], R^m) \rightarrow L_2([\alpha, \beta], R^n)$ defined by*

$$(\tilde{B}u)(t) = B(t)u(t), \quad \alpha \leq t \leq \beta,$$

has closed range if and only if the mapping $t \rightarrow B^{\dagger}(t)$ is essentially bounded on $[\alpha, \beta]$.

Proof. Theorem 3.3 does not require the hypothesis $t_1 > t_0 + h$. Also note that, (H2) is satisfied for the system $\dot{x}(t) = B(t)u(t)$. Hence an application of Theorem 3.3 to this system with $h = \beta - \alpha$, $t_0 = \alpha$, $t_1 = \beta$ shows that (H3) on $[\alpha, \beta]$ implies \tilde{B} has closed range. Conversely, if \tilde{B} has closed range, then \tilde{B}^{\dagger} is a bounded multiplication operator on L_2 . Hence $t \rightarrow B^{\dagger}(t)$ is essentially bounded [13, pp. 31, 212].

It is noted that if the underlying hypothesis (H2) is not satisfied, then Theorem 3.3 is no longer true. This is demonstrated in the following example.

Example 3.3. Suppose the system equations are

$$\dot{x}_1(t) = u(t), \quad \dot{x}_2(t) = x_1(t - 1).$$

Then $\mathcal{A}(0, 2, 0)$ is not closed in $W_2^{(1)}([-1, 0], R^2)$. Define

$$u_v(t) = \begin{cases} 0, & 0 \leq t \leq \frac{1}{2} - \frac{1}{v}, \\ \frac{v}{2}, & \frac{1}{2} - \frac{1}{v} < t \leq \frac{1}{2} + \frac{1}{v}, \\ 0, & \frac{1}{2} + \frac{1}{v} < t \leq 2. \end{cases}$$

Then $\psi_v = (\psi_{1v}, \psi_{2v})^* \in \mathcal{A}(0, 2, 0)$, $v = 2, 3, \dots$, where ψ_{1v}, ψ_{2v} are given by

$$\psi_{1v}(\theta) = 1, \quad -1 \leq \theta \leq 0,$$

$$\psi_{2v}(\theta) = \int_1^{2+\theta} \int_0^{s-1} u_v(\xi) d\xi ds, \quad -1 \leq \theta \leq 0.$$

Clearly $\psi_v \rightarrow \psi = (\psi_1, \psi_2)^*$ in $W_2^{(1)}([-1, 0], R^2)$, where

$$\psi_1(\theta) = 1, \quad -1 \leq \theta \leq 0,$$

$$\psi_2(\theta) = \int_1^{2+\theta} \chi_E(\xi) d\xi, \quad -1 \leq \theta \leq 0,$$

and $E = [3/2, 2]$. We have that

$$\begin{aligned} \mathcal{A}(0, 2, 0) = \left\{ (\zeta_1, \zeta_2)^* \in W_2^{(1)}([-1, 0], R^2) \mid \zeta_1(\theta) = \int_0^{2+\theta} u(\xi) d\xi, \right. \\ \left. \zeta_2(\theta) = \int_1^{2+\theta} \int_0^{s-1} u(\xi) d\xi ds, -1 \leq \theta \leq 0, \text{ for some } u \in L_2([0, 2], R) \right\}, \end{aligned}$$

and consequently $\psi \notin \mathcal{A}(0, 2, 0)$. Therefore $\mathcal{A}(0, 2, 0)$ is not closed.

Remark 3.3. The intimate relationship between closedness of the attainable set and nontriviality of multipliers in the first order necessary conditions for control problems with function space target sets has been alluded to in [4], [17]. In a recent paper [19a] Kurcyusz has pointed out that there are simple quadratic minimization problems involving the system in Example 3.3 and boundary conditions of the form $x_0 = \phi$, $x_3 = \psi$ for which the Lagrange multipliers must be identically zero if the state space is $W_2^{(1)}([-h, 0], R^2)$. However, when Kurcyusz chooses the state space to be $W_2^{(1)}([-h, 0], R^1) \times W_2^{(2)}([-h, 0], R^1)$ he does obtain nontrivial (albeit more complicated) Lagrange multipliers for these problems.

4. Fredholm alternative methods for determining the attainable set. Assume that (H3) is satisfied on the interval $[t_0, t_1]$, and define an operator

$$\mathcal{K} : W_2^{(1)}[t_0 - h, t_1], R^n \rightarrow L_2([t_0, t_1], R^n)$$

by the equation

$$(4.1) \quad (\mathcal{K}x)(t) \equiv \dot{x}(t) - L(t, x_t), \quad t_0 \leq t \leq t_1.$$

Let $\tilde{B} : L_2([t_0, t_1], R^m) \rightarrow L_2([t_0, t_1], R^n)$ be the map determined by

$$(4.2) \quad (\tilde{B}u)(t) = B(t)u(t), \quad t_0 \leq t \leq t_1.$$

Using this notation, we see that $\psi \in \mathcal{A}(t_0, t_1, \phi)$ is equivalent to saying there is an $x \in W_2^{(1)}([t_0 - h, t_1], R^n)$ such that $x_{t_0} = \phi$, $x_{t_1} = \psi$ and

$$(4.3) \quad \mathcal{K}x \in \text{Im } \tilde{B}.$$

Thus it is of some interest to study the inclusion (4.3). Noting that \tilde{B} is a bounded operator with closed range (see Corollary 3.4), we see that the alternative theorem [8, p. 487] implies that (4.3) is equivalent to

$$(4.4) \quad \mathcal{K}x \in (\ker \tilde{B}^*)^\perp.$$

The inclusion (4.4) is equivalent to

$$(4.5) \quad \int_{t_0}^{t_1} \langle \dot{x}(t) - L(t, x_t), \mu(t) \rangle dt = 0$$

for every $\mu \in \ker \tilde{B}^*$.

For simplicity we now assume that B is a constant matrix function. Thus (H3') is automatically satisfied. Suppose the nullity of the matrix B is v . Choose an orthonormal basis e_1, e_2, \dots, e_v for $\ker B^*$ (note that $\ker B^*$ is the nullspace of the transposed matrix, whereas $\ker \tilde{B}^*$ is an infinite-dimensional space if $v > 0$). Let π be the projection of R^n onto $\ker B^*$ defined by

$$\pi y = \langle e_1, y \rangle e_1 + \dots + \langle e_v, y \rangle e_v, \quad y \in R^n.$$

It can be shown that $\tilde{\pi}$, defined by

$$(4.6) \quad (\tilde{\pi}\mu)(t) = \pi\mu(t), \quad t_0 \leq t \leq t_1,$$

is the projection of $L_2([t_0, t_1], R^n)$ onto $\ker \tilde{B}^*$. Hence (4.5) takes the form

$$\int_{t_0}^{t_1} \langle \dot{x}(t) - L(t, x_t), (\tilde{\pi}\mu)(t) \rangle dt = \int_{t_0}^{t_1} \langle (\tilde{\pi}\mathcal{K}x)(t), \mu(t) \rangle dt = 0$$

for every $\mu \in L_2([t_0, t_1], R^n)$. Consequently (4.5) is equivalent to

$$(4.7) \quad \sum_{i=1}^v \frac{d}{dt} \langle e_i, x(t) \rangle e_i = \sum_{i=1}^v \langle e_i, L(t, x_t) \rangle e_i$$

a.e. on $[t_0, t_1]$. We summarize these results in the next proposition and corollary.

PROPOSITION 4.1. *Let B be a constant matrix. Then $x \in W_2^{(1)}([t_0 - h, t_1], R^n)$ satisfies (4.3) if and only if (4.7) is satisfied.*

COROLLARY 4.1. *Let B be a constant matrix. A function $\psi \in W_2^{(1)}([-h, 0], R^n)$ belongs to $\mathcal{A}(t_0, t_1, \phi)$ if and only if there is an $x \in W_2^{(1)}([t_0 - h, t_1], R^n)$ satisfying (4.7) and*

$$x_{t_0} = \phi, \quad x_{t_1} = \psi.$$

Example 4.1. Consider the system (3.28) in Example 3.1. Let $(A_k)_i$ denote the i th row of the matrix A_k . Then for system (3.28) with $B \equiv (0, 0, \dots, 0, 1)^*$ we get that $e_i = (\delta_{i1}, \delta_{i2}, \dots, \delta_{in})^*$, $i = 1, 2, \dots, n-1$, is an orthonormal basis for $\ker B^*$, where $\delta_{ij} = 0$ if $i \neq j$ and $\delta_{ij} = 1$ if $i = j$. Thus (4.7) takes the form

$$(4.8) \quad \dot{x}_i(t) = \sum_{k=0}^N (A_k)_i x(t - h_k), \quad i = 1, 2, \dots, n-1.$$

If we specialize this example to the situation discussed in Example 3.2, then equations (4.8) take the form

$$(4.8') \quad \dot{x}_i(t) = x_{i+1}(t - h), \quad i = 1, 2, \dots, n-1.$$

In order for $0 \in \mathcal{A}(0, t_1, \phi)$ there must be a solution $x \in W_2^{(1)}([-h, t_1], R^n)$ of the above $n-1$ differential difference equations such that $x_0 = \phi$, $x_{t_1} = 0$. Thus we see that if $h < t_1 \leq nh$, then ϕ cannot be an arbitrary element of $W_2^{(1)}([-h, 0], R^n)$. In fact, equations (4.8') require $\phi_n(\theta) = 0$, $t_1 - nh < \theta \leq 0$, where

$$\phi = (\phi_1, \phi_2, \dots, \phi_n)^*.$$

On the other hand, if $t_1 > nh$ the equations (4.8') do not restrict the initial function ϕ at all. Hence the differential difference equation in Example 3.2 is null controllable on $[0, t_1]$ if $t_1 > nh$, but *not* null controllable if $t_1 \leq nh$.

The above corollary and proposition are somewhat difficult to use. There is, however, a special class of problems covering many of the known applications for which we can give a particularly simple construction of the attainable set. These are the systems which satisfy (H1), (H2) and (H3) (see [4] for a number of examples). These added conditions enable us to reduce the question to one involving only ordinary differential equations.

Along with (2.3) consider the ordinary differential equation

$$(4.9) \quad \dot{z}(t) = A_0(t)z(t) + B(t)f(t).$$

Define

$$\mathcal{R}(t_1) = \{\omega \in W_2^{(1)}([-h, 0], R^n) | \omega = z_{t_1}(\cdot, t_1 - h, \xi, f), \\ \xi \in R^n, f \in L_2([t_1 - h, t_1], R^m)\},$$

where $z(t) = z(t, t_1 - h, \xi, f)$ is the solution to (4.9) satisfying

$$(4.10) \quad z(t_1 - h) = \xi.$$

THEOREM 4.1. *Hypotheses (H1), (H2) and (H3) imply that $\mathcal{R}(t_1) = \mathcal{A}(t_0, t_1, \phi)$, for $t_1 > t_0 + h$ and $\phi \in W_2^{(1)}([-h, 0], R^n)$.*

Proof. If $\omega \in \mathcal{R}(t_1)$, then there is an f in $L_2([t_1 - h, t_1], R^m)$ and $\xi \in R^n$ such that

$$z_{t_1}(\cdot, t_1 - h, \xi, f) = \omega.$$

In view of (H1) and Lemma 3.1, given $\phi \in W_2^{(1)}([-h, 0], R^n)$, there is a $v \in L_2([t_0, t_1 - h], R^m)$ such that

$$x(t_1 - h, t_0, \phi, v) = \xi.$$

Now define

$$\tilde{v}(t) = f(t) + B^t(t) \left\{ \tilde{\eta}(t, -h)x(t - h, t_0, \phi, v) + \int_{t-h}^{t_1-h} \tilde{\eta}(t, \alpha - t)\dot{x}(\alpha, t_0, \phi, v) d\alpha \right. \\ \left. + \int_{t_1-h}^t \tilde{\eta}(t, \alpha - t)\dot{\omega}(\alpha - t_1) d\alpha \right\}$$

for $t_1 - h \leq t \leq t_1$, and let

$$u(t) = \begin{cases} v(t), & t_0 \leq t \leq t_1 - h, \\ \tilde{v}(t), & t_1 - h < t \leq t_1. \end{cases}$$

Then (H3) implies $u \in L_2([t_0, t_1], R^m)$, whereas (H2) and the fact that

$$\omega = z_{t_1}(\cdot, t_1 - h, \xi, f)$$

imply that $x_{t_1}(\cdot, t_0, \phi, u) = \omega$. We conclude that $\mathcal{R}(t_1) \subset \mathcal{A}(t_0, t_1, \phi)$. On the other hand, if $\psi \in \mathcal{A}(t_0, t_1, \phi)$ and $x_{t_1}(\cdot, t_0, \phi, u) = \psi$, $u \in L_2([t_0, t_1], R^m)$, then (H2) clearly implies $\psi = z_{t_1}(\cdot, t_1 - h, \xi, f)$, where $\xi = x(t_1 - h, t_0, \phi, u)$ and

$$f(t) = u(t) - B^t(t) \left\{ \tilde{\eta}(t, -h)x(t - h, t_0, \phi, u) + \int_{t-h}^t \tilde{\eta}(t, \alpha - t)\dot{x}(\alpha, t_0, \phi, u) d\alpha \right\},$$

$t_1 - h \leq t \leq t_1$. Thus $\mathcal{A}(t_0, t_1, \phi) \subset \mathcal{R}(t_1)$, and this completes the proof.

PROPOSITION 4.2. *Suppose (H3') holds. If $\mathcal{A}(t_0, t_1, \phi) = \mathcal{R}(t_1)$ for each t_0, t_1 with $t_1 > t_0 + h$ and $\phi \in W_2^{(1)}([-h, 0], R^n)$, and if (2.3) is a system with strict retardations, then (H1') and (H2') are satisfied.*

Proof. The zero function always belongs to $\mathcal{R}(t_1)$. Hence $\mathcal{A}(t_0, t_1, \phi) = \mathcal{R}(t_1)$ implies there is a $u \in L_2([t_0, t_1], R_m)$ such that $x_{t_1}(\cdot, t_0, \phi, u) = 0$. Since $\phi \in W_2^{(1)}([-h, 0], R^n)$, $t_1, t_0, t_1 > t_0 + h$, are arbitrary, the conclusion follows from Theorem 3.2.

In view of Theorem 4.1 it becomes a question of some interest to devise an effective means of determining $\mathcal{R}(t_1)$. For brevity we will abbreviate $[t_1 - h, t_1]$ by $[\alpha, \beta]$, and define two operators S and T where $S: R^n \rightarrow W_2^{(1)}([\alpha, \beta], R^n)$ and $T: L_2([\alpha, \beta], R^m) \rightarrow W_2^{(1)}([\alpha, \beta], R^n)$ are given by

$$\begin{aligned} (S\xi)(t) &= z(t, \alpha, \xi, 0), & \alpha \leq t \leq \beta, \\ (Tf)(t) &= z(t, \alpha, 0, f), & \alpha \leq t \leq \beta. \end{aligned}$$

Moreover, instead of $\mathcal{R}(t_1)$ we merely write \mathcal{R} . Thus we see that

$$(4.11) \quad \mathcal{R} = \text{Im } S + \text{Im } T.$$

The space $\text{Im } S$ is finite-dimensional, and we will not discuss its construction. We focus our attention on $\text{Im } T$.

Let $\tilde{A}_0: L_2([\alpha, \beta], R^n) \rightarrow L_2([\alpha, \beta], R^n)$ be the multiplication map corresponding to matrix A_0 as in equation (4.2). Let $\langle \cdot, \cdot \rangle_2$ be the inner product in $L_2([\alpha, \beta], R^n)$.

COROLLARY 4.2. *Let (H3) be satisfied. Then $z \in \text{Im } T$ if and only if $z \in W_2^{(1)}([\alpha, \beta], R^n)$, $z(\alpha) = 0$ and $\langle \dot{z} - \tilde{A}_0 z, \mu \rangle_2 = 0$ for every $\mu \in \ker \tilde{B}^*$.*

This corollary is merely a specialization of the result in relation (4.4) to the present ordinary differential equation (4.9).

COROLLARY 4.3. *Let (H3) be satisfied. Then $z \in \text{Im } T$ if and only if $z \in W_2^{(1)}([\alpha, \beta], R^n)$, $z(\alpha) = 0$ and $\langle \dot{z}, \omega \rangle_2 = 0$ for every ω of the form $\omega(t) = \mu(t) - \int_t^\beta A_0^*(s)\mu(s) ds$, a.e. $t \in [\alpha, \beta]$ with $\mu \in \ker \tilde{B}^*$.*

Proof. We note that for $\mu \in \ker \tilde{B}^*$ we have

$$\langle \dot{z} - \tilde{A}_0 z, \mu \rangle_2 = \langle \dot{z}, \mu \rangle_2 - \langle z, \tilde{A}_0^* \mu \rangle_2.$$

Use integration by parts on the second term on the right-hand side of the last equation and the desired conclusion is evident.

Example 4.2. Take the system

$$\dot{z}(t) = A_0(t)z(t) + B(t)u(t),$$

where $B^*(t) = (0, C^*(t))$ and $C^*(t)$ is an $m \times m$ nonsingular matrix (the 0 denotes an $m \times (n - m)$ zero matrix filling out the rest of the matrix $B^*(t)$). Thus

$$\ker \tilde{B}^* = \left\{ \begin{pmatrix} \bar{\mu} \\ 0 \end{pmatrix} \mid \bar{\mu} \in L_2([\alpha, \beta], R^{n-m}) \right\}.$$

In the notation of Example 4.1, we get $z \in \text{Im } T$ if and only if $z \in W_2^{(1)}([\alpha, \beta], R^n)$,

and

$$(4.12) \quad \begin{aligned} \dot{z}_i(t) &= (A_0)_i z(t), & i = 1, 2, \dots, n-m, \\ z(\alpha) &= 0. \end{aligned}$$

Hence if B is as given above, and if system (2.3) satisfies (H1), (H2), (H3), then equations (4.12) can be used to determine $\mathcal{A}(t_0, t_1, \phi)$. At this time we merely observe that this is done with the aid of Theorem 4.1 and the fact that

$$\mathcal{A}(t_0, t_1, \phi) = \text{Im } S + \text{Im } T$$

if we take $\alpha = t_1 - h, \beta = t_1$.

Example 4.3. Consider now the n th order scalar equation

$$(4.13) \quad y^{(n)}(t) = \sum_{i=0}^{n-1} b_i(t)y^{(i)}(t-h) + \sum_{i=0}^{n-1} a_i(t)y^{(i)}(t) + u(t).$$

Taking $x = (x_1, x_2, \dots, x_n)^* = (y, y^{(1)}, \dots, y^{(n-1)})^*$, we write (4.13) as a first order n -dimensional differential-difference equation,

$$(4.14) \quad \dot{x}(t) = A_0(t)x(t) + A_1(t)x(t-h) + Bu(t),$$

where $B^* = (0, 0, \dots, 0, 1)$,

$$A_0(t) = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ a_0(t) & a_1(t) & a_2(t) & \cdots & a_{n-1}(t) \end{bmatrix}$$

and

$$A_1(t) = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ b_0(t) & b_1(t) & \cdots & b_{n-1}(t) \end{bmatrix}.$$

If we assume that the functions a_j have j continuous derivatives on $(-\infty, \infty)$, $j = 0, 1, \dots, n-1$, then the system

$$(4.15) \quad \dot{z}(t) = A_0(t)z(t) + Bu(t)$$

is Euclidean controllable [14, p. 100]. Then, since (4.14) satisfies (H2') and (H3'), it also satisfies (H1') by Corollary 3.3. Hence Theorem 4.1 applies to (4.14) and (4.15) on any interval $[t_0, t_1]$ with $t_1 > t_0 + h$. Specializing the results of Example 4.2 to the present situation, one can readily check that

$$(4.16) \quad \begin{aligned} \mathcal{A}(t_0, t_1, \phi) &= \{\psi = (\psi_1, \dots, \psi_n)^* \in W_2^{(1)}([-h, 0], R^n) | \dot{\psi}_i(\theta) \\ &= \psi_{i+1}(\theta), -h \leq \theta \leq 0, i = 1, \dots, n-1\}. \end{aligned}$$

This simplifies the treatment of Example 3.1 in [4] (cf. [2], [9], [11], [16]).

Example 4.4. Consider the autonomous system

$$(4.17) \quad \dot{z} = A_0 z + bu,$$

with scalar control u (b is an $n \times 1$ matrix). Let

$$\det(A_0 - \lambda I) = (-1)^n \left[\lambda^n - \sum_{i=0}^{n-1} a_i \lambda^i \right].$$

According to the Cayley–Hamilton theorem,

$$(4.18) \quad A_0^n = \sum_{i=0}^{n-1} a_i A_0^i.$$

Let G be the matrix

$$G = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & & 1 \\ a_0 & a_1 & a_2 & \cdots & a_{n-1} \end{bmatrix}.$$

Suppose

$$(4.19) \quad \begin{aligned} P &= [b, A_0 b, \dots, A_0^{n-1} b], \\ Q &= [b_0, Gb_0, \dots, G^{n-1} b_0], \end{aligned}$$

where $b_0 = (0, 0, \dots, 0, 1)^*$. If we assume that (4.17) is Euclidean controllable, then both P and Q are invertible. Moreover, if we make the change of variable $z = PQ^{-1}y$, then (4.17) becomes

$$(4.20) \quad \dot{y} = Gy + b_0 u$$

(cf. [21, p. 91]). The results of Example 4.3 can now be applied to (4.20). If we let $\mathcal{R}_z(t_1)$, $\mathcal{R}_y(t_1)$ denote the sets $\mathcal{R}(t_1)$ corresponding to (4.17) and (4.20) respectively, then $\mathcal{R}_y(t_1)$ is the set described in equation (4.16) and

$$(4.21) \quad \mathcal{R}_z(t_1) = PQ^{-1}\mathcal{R}_y(t_1).$$

Hence $\mathcal{R}_z(t_1)$ is completely determined by the two controllability matrices in (4.19) and the coefficients of the characteristic polynomial for A_0 . It is noted that by (4.21) and the fact that $\mathcal{R}_y(t_1)$ is the same as the set in (4.16), the set $\mathcal{R}_z(t_1)$ is independent of t_1 . This has worthwhile implications. If (2.3) satisfies (H1'), (H2') and (H3'), then the corresponding attainable set $\mathcal{A}(t_0, t_1, \phi) = \mathcal{R}_z(t_1)$ is as given in (4.21) and does not depend on t_0, t_1 or ϕ . This has a number of consequences for the optimization problems studied in [4], [16], [17] and [21]. These will be explored in a forthcoming paper.

Example 4.5. Let us determine $\mathcal{R}(t_1)$ explicitly for two-dimensional systems with scalar control u . Suppose the constant matrices

$$A_0 = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

are given, and that the corresponding controlled system (4.17) is Euclidean controllable. One verifies that

$$A_0^2 = (\text{tr } A_0)A_0 - (\det A_0)I$$

($\text{tr } A_0$ is an abbreviation for the trace of the matrix). Matrix Q in (4.19) is

$$Q = \begin{bmatrix} 0 & 1 \\ 1 & \text{tr } A_0 \end{bmatrix},$$

and

$$Q^{-1} = \begin{bmatrix} -\text{tr } A_0 & 1 \\ 1 & 0 \end{bmatrix}.$$

The matrix P in (4.19) is $[b, A_0b]$, and thus $PQ^{-1} = [-(\text{adj } A_0)b, b]$. If we let $W_2^{(2)}[-h, 0]$ be all functions $\eta \in W_2^{(1)}([-h, 0], R)$ such that $\dot{\eta} \in W_2^{(1)}([-h, 0], R)$, then (4.21) becomes

$$\mathcal{R}_2(t_1) = \{(\dot{\omega}I - \omega \text{adj } A_0)b | \omega \in W_2^{(2)}[-h, 0]\}.$$

5. Implicit determination of the reachable set. As Example 6.1 below indicates, Theorem 3.1 does not extend directly to provide necessary and sufficient conditions for controllability of neutral functional differential equations. Indeed, our preliminary efforts on neutral systems indicate in fact that they differ substantially from systems of retarded type with regard to controllability characteristics. We have therefore to this point restricted our discussions in this paper to retarded systems even though several of the results obtained so far will extend (with proper modifications) to provide results for neutral systems. We shall present our findings for neutral systems in a later paper. However, the techniques to be described herein and in § 6 are such that the extension to neutral differential-difference equations is immediate. We have, therefore, to avoid lengthy and undue repetition at a later date, included certain neutral systems in our discussions below.

We consider then the equation

$$(5.1) \quad \dot{x}(t) = A_{-1}\dot{x}(t-h) + A_0x(t) + A_1x(t-h) + Bu(t),$$

where the A_i are constant $n \times n$ matrices and B is a constant $n \times m$ matrix. When $A_{-1} = 0$, this is a special case of (2.3). The techniques to be described here apply to the case where $A_{-1} \neq 0$ and generalize some results in [25]. Since (5.1) is autonomous, we shall take $t_0 = 0$ in §§ 5 and 6 without loss of generality.

Further notation will be convenient here. By $W_{2,0}^{(1)}(t_1, R^n)$ we denote the collection of functions which are continuous on $(-\infty, t_1]$, vanish on $(-\infty, 0]$ and have restrictions on $[0, t_1]$ in $W_2^{(1)}([0, t_1], R^n)$. Inductively, we define $W_{2,0}^{(k+1)}(t_1, R^n)$ for $k \geq 1$ as the set of integrals $\int_0^t \xi(\tau) d\tau$ of functions $\xi \in W_{2,0}^{(k)}(t_1, R^n)$. By $W_{2,0}^{(0)}(t_1, R^m)$ we shall mean the set of functions which vanish on $(-\infty, 0]$ and have restrictions on $[0, t_1]$ in $L_2([0, t_1], R^m)$.

For each $u \in W_{2,0}^{(0)}(t_1, R^m)$ there exists a unique $x \in W_{2,0}^{(1)}(t_1, R^n)$ satisfying (5.1) a.e. on $[0, t_1]$ and in fact on $(-\infty, t_1]$ since both are zero on $(-\infty, 0]$.

For functions $x \in W_{2,0}^{(k)}(t_1, R^n)$, $k \geq 1$, $v \geq 1$, we define the operators S and

D by

$$(Sx)(t) = x(t - h), \quad (Dx)(t) = \dot{x}(t).$$

The operators S and D commute and each commutes with multiplication by a constant scalar. Moreover, each of these three commutes with the coordinate projections. That is, if $(x)_j = x_j$ denotes the j th coordinate of the v -vector x , then $(Sx)_j = Sx_j$, $j = 1, \dots, v$, with the analogous relations holding also for D and for multiplication by a real number c . Thus D and S may be treated formally as scalars in the usual operations involving constant matrices and vector functions.

With this in mind we may write (5.1) in the form

$$(5.2) \quad (ID - A_{-1}SD - A_0 - A_1S)x = Bu,$$

where I denotes the $n \times n$ identity matrix. We now introduce the operator matrices

$$(5.3) \quad Q(D, S) = ID - A_{-1}SD - A_0 - A_1S$$

and

$$(5.4) \quad P(D, S) = \text{adj } Q(D, S).$$

Here "adj" means the formal algebraic adjoint, the transpose of the corresponding matrix of cofactors. As a relation between matrix operators we have

$$(5.5) \quad Q(D, S)P(D, S) = P(D, S)Q(D, S) = I \det Q(D, S)$$

since these are algebraic identities in the scalar variables D and S . Note that

$$(5.6) \quad P(D, S) = \sum_{j=0}^{n-1} P_j(D)S^j = \sum_{j=0}^{n-1} \hat{P}_j(S)D^j,$$

where the $n \times n$ matrix polynomials $P_j(D)$ and $\hat{P}_j(S)$ are at most of degree $n - 1$ in their arguments. When $A_{-1} = 0$ the polynomial $P_j(D)$ is at most of degree $n - 1 - j$.

LEMMA 5.1. *Let $t_1 > 0$. Then for each $x \in W_{2,0}^{(1)}(t_1, R^n)$ there is a unique $\omega \in W_{2,0}^{(n)}(t_1, R^n)$ such that*

$$(5.7) \quad P(D, S)\omega = x \quad \text{on } (-\infty, t_1].$$

Proof. Observe that $\hat{P}_{n-1}(S) = \text{adj } (I - A_{-1}S)$. Hence (5.7) is equivalent to

$$(5.8) \quad D^{n-1}\omega + \sum_{k=1}^{n-1} C_k S^k D^{n-1}\omega = - \sum_{j=0}^{n-2} \hat{P}_j(S) D^j \omega + x$$

for some constant $n \times n$ matrices C_k . Introducing the auxiliary variables $v^j = D^{j-1}\omega$, $j = 1, \dots, n-1$, we may view (5.8) as a first order equation in the $n(n-1) \times 1$ column \tilde{v} comprising the v^j 's. As such there is on $(-\infty, t_1]$ a unique solution $\tilde{v} \in W_{2,0}^{(2)}(t_1, R^v)$ where $v = n(n-1)$ for each $x \in W_{2,0}^{(1)}(t_1, R^n)$. In particular, $\omega = v^1$ and $D^{n-2}\omega = v^{n-1}$ are each in $W_{2,0}^{(2)}(t_1, R^n)$ whence $\omega \in W_{2,0}^{(n)}(t_1, R^n)$.

Now let b^k denote the k th column of B and define

$$(5.9) \quad K^k(D) = [P_0(D)b^k, P_1(D)b^k, \dots, P_{n-1}(D)b^k].$$

Here the products $P_j(D)b^k$ are to be interpreted as products of operators obtained by formal algebraic matrix multiplication with D as a scalar variable. That is, b^k is here a constant vector multiplicative operator acting on scalar functions and *not* a constant vector function on which the operator $P_j(D)$ acts. The order of the factors is, of course, not immaterial. Each $K^k(D)$ is $n \times n$ with elements which are polynomials in D at most of degree $n - 1$.

THEOREM 5.1. Let B be $n \times m$, $t_1 > 0$, and $u \in W_{2,0}^{(0)}(t_1, R^m)$. Then

$$x \in W_{2,0}^{(1)}(t_1, R^n)$$

is the corresponding solution of (5.1) on $(-\infty, t_1]$ if and only if there exist $\omega^1, \dots, \omega^m \in W_{2,0}^{(m)}(t_1, R^n)$ such that

$$(5.10) \quad \sum_{k=1}^m K^k(D)\omega^k = x \quad \text{on } (-\infty, t_1],$$

$$(5.11) \quad S\omega_j^k = \omega_{j+1}^k, \quad j = 1, \dots, n-1, \quad k = 1, \dots, m,$$

$$(5.12) \quad \det Q(D, S)\omega_1^k = u_k, \quad k = 1, \dots, m.$$

Proof. Let x be the solution of (5.1) on $(-\infty, t_1]$ corresponding to the given u and let ω be the solution of (5.7) for this x as assured by Lemma 5.1. Using (5.5), we then have

$$[I \det Q(D, S)]\omega = Q(D, S)[P(D, S)\omega] = Q(D, S)x = Bu,$$

whence

$$(5.13) \quad \det Q(D, S)\omega_i = \sum_{k=1}^m b_i^k u_k, \quad i = 1, \dots, n, \text{ on } (-\infty, t_1].$$

Now let $\tilde{\omega}^k \in W_{2,0}^{(n)}(t_1, R)$ be the unique solution on $(-\infty, t_1]$ of

$$(5.14) \quad \det Q(D, S)\tilde{\omega}^k = u_k, \quad k = 1, \dots, m.$$

It follows that $\omega_i = \sum_{k=1}^m b_i^k \tilde{\omega}^k$, so

$$(5.15) \quad \omega = \sum_{k=1}^m b^k \tilde{\omega}^k.$$

From (5.7) and (5.6) we then have

$$(5.16) \quad x = \sum_{k=1}^m \sum_{j=0}^{n-1} P_j(D)b^k S^j \tilde{\omega}^k.$$

If we define

$$(5.17) \quad \omega^k = \begin{bmatrix} \tilde{\omega}^k \\ S\tilde{\omega}^k \\ \vdots \\ S^{n-1}\tilde{\omega}^k \end{bmatrix},$$

then (5.16) takes the form (5.10), each $\omega^k \in W_{2,0}^{(n)}(t_1, R^n)$, (5.11) follows from the form specified in (5.17) and (5.12) is (5.14).

Conversely, suppose $\omega^k \in W_{2,0}^{(n)}(t_1, R^n)$, $k = 1, \dots, m$, satisfy (5.11) and (5.12) and that x is defined by (5.10). If we let $\tilde{\omega}^k = \omega_1^k$, then ω^k is given by (5.17) and since each $K^k(D)$ is a polynomial in D at most of degree $n - 1$, then $x \in W_{2,0}^{(1)}(t_1, R^n)$. Equation (5.10) may now be written in the form (5.16), whence

$$x = P(D, S) \sum_{k=1}^m b^k \omega_1^k.$$

Finally, by (5.5),

$$Q(D, S)x = [I \det Q(D, S)] \sum_{k=1}^m b^k \omega_1^k = \sum_{k=1}^m b^k u_k = Bu$$

in view of (5.12). Thus x is a solution of (5.1) on $(-\infty, t_1]$ for the given u , and since this solution is unique the theorem is proved.

Theorem 5.1 generalizes and makes more transparent some of the relations behind the analysis in Minjuk's work [25]. Besides treating only the retarded case with just a single control, Minjuk assumed that $\det K^1(D) \neq 0$ so that his equation corresponding to (5.10) in Laplace transforms could be solved for the transform of ω^1 . Our proof shows this sort of assumption is superfluous. The following theorem is a generalization of Minjuk's result and involves no such hypothesis.

THEOREM 5.2. *Let $t_1 > 0$ and let ρ be the nonnegative integer such that $\rho h < t_1 \leq (\rho + 1)h$. If $\psi \in W_2^{(1)}([t_1 - h, t_1], R^n)$, then there is a control*

$$u \in W_{2,0}^{(0)}(t_1, R^m)$$

such that the corresponding solution $x \in W_{2,0}^{(1)}(t_1, R^n)$ of (5.1) satisfies

$$(5.18) \quad x(t) = \psi(t), \quad t \in [t_1 - h, t_1],$$

if and only if there exist functions $\omega^k \in W_2^{(n)}([t_1 - h, t_1], R^n)$ such that

$$(5.19) \quad \sum_{k=1}^m K^k(D) \omega^k(t) = \psi(t), \quad t \in [t_1 - h, t_1],$$

and for $k = 1, \dots, m$,

$$(5.20) \quad D^i \omega_j^k(t_1 - h) = D^i \omega_{j+1}^k(t_1), \quad i = 0, \dots, n - 1, \quad j = 1, \dots, n - 1,$$

$$(5.21a) \quad D^i \omega_{\rho+1}^k(t_1 - h) = 0, \quad i = 0, \dots, n - 1,$$

$$(5.21b) \quad \omega_{\rho+1}^k(t) = 0, \quad t \in [t_1 - h, \rho h],$$

$$(5.22) \quad \omega_j^k(t) = 0, \quad t \in [t_1 - h, t_1] \text{ if } \rho + 2 \leq j \leq n.$$

Remark 5.1. We note that (5.21a) is implied by (5.21b) if $(n - 1)h < t_1 < nh$, (5.21b) is implied by (5.21a) if $t_1 = nh$, and both are vacuous if $\rho \geq n$; moreover, (5.22) is vacuous if $\rho \geq n - 1$.

Proof of Theorem 5.2. Suppose $u \in W_{2,0}^{(0)}(t_1, R^m)$ exists such that the corresponding solution $x \in W_{2,0}^{(1)}(t_1, R^n)$ of (5.1) on $(-\infty, t_1]$ satisfies (5.18). By Theorem 5.1 we then have (5.10) for some functions $\omega^k \in W_{2,0}^{(n)}(t_1, R^n)$ so (5.19) holds for these ω^k . The ω^k satisfy $\omega_j^k(t - h) = \omega_{j+1}^k(t)$ for $t \leq t_1$ by (5.11) so $D^i \omega_j^k(t_1 - h) = D^i \omega_{j+1}^k(t_1)$ for $i = 0, 1, \dots, n - 1$, so (5.20) must hold. If $\rho \leq n - 1$,

then $\omega_{\rho+1}^k(t) = \omega_1^k(t - \rho h) = 0$ for $t \leq \rho h$ and if $\rho + 2 \leq j \leq n$, then $\omega_j^k(t) = \omega_1^k(t - (j-1)h) = 0$ for $t \leq t_1$ since $\omega_1^k(\tau) = 0$ for $\tau \leq 0$. Hence also (5.21) and (5.22) must hold.

Now suppose the functions $\omega^k \in W_2^{(n)}([t_1 - h, t_1], R^n)$, $k = 1, \dots, m$, satisfy (5.19) through (5.22). We shall define extensions $\hat{\omega}^k \in W_{2,0}^{(n)}(t_1, R^n)$ of the ω^k such that $S\hat{\omega}_j^k = \hat{\omega}_{j+1}^k$, $j = 1, \dots, n-1$. With $u_k = \det Q(D, S)\hat{\omega}_1^k$ then, by Theorem 5.1, $x = \sum_{k=1}^m K^k(D)\hat{\omega}^k$ is the solution in $W_{2,0}^{(1)}(t_1, R^n)$ of (5.1) on $(-\infty, t_1]$ corresponding to the resulting control vector $u \in W_{2,0}^{(0)}(t_1, R^m)$. For $t \in [t_1 - h, t_1]$ we then get (5.18) from (5.19) since $\omega^k = \hat{\omega}^k$ on this interval. The extension is accomplished as follows. Beginning with $\hat{\omega}_1^k(t) = \omega_1^k(t)$ for $t \in [t_1 - h, t_1]$, we define

$$(5.23) \quad \hat{\omega}_1^k(t - jh) = \omega_{j+1}^k(t), \quad j = 1, \dots, n-1, \quad t \in [t_1 - h, t_1).$$

If $t_1 > nh$, then for $t < t_1 - nh$ we take $\hat{\omega}_1^k$ to be any function in $W_{2,0}^{(n)}(t_1 - nh, R)$ such that $D^i \hat{\omega}_1^k$ is continuous at $t_1 - nh$ for $i = 0, 1, \dots, n-1$. If $t_1 \leq nh$ we take $\hat{\omega}_1^k(t) = 0$ for $t < t_1 - nh$. By (5.20) through (5.22) the resulting $\hat{\omega}_1^k$ is in $W_{2,0}^{(n)}(t_1, R)$ regardless of the value of the integer $\rho \geq 0$. We now define $\hat{\omega}^k$ by

$$\hat{\omega}^k(t) = \begin{bmatrix} \hat{\omega}_1^k(t) \\ S\hat{\omega}_1^k(t) \\ \vdots \\ S^{n-1}\hat{\omega}_1^k(t) \end{bmatrix}, \quad t \leq t_1.$$

Clearly $\hat{\omega}^k(t) = \omega^k(t)$ for $t \in [t_1 - h, t_1]$ by (5.23) and also at $t = t_1$ by the continuity of $\hat{\omega}_1^k$. Since $S\hat{\omega}_j^k = \hat{\omega}_{j+1}^k$, $j = 1, \dots, n-1$, by construction, the theorem is proved.

COROLLARY 5.1. *For (5.1) the attainable sets $\mathcal{A}(t_0, t_1, 0)$ from the zero initial function satisfy the relations*

$$(5.24) \quad \mathcal{A}(t_0, t_1, 0) \subseteq \mathcal{A}(t'_0, t'_1, 0)$$

if $t_1 - t_0 \leq t'_1 - t'_0$, with equality holding if $t_1 - t_0 > nh$.

Proof. For (5.1) it is clear by translation of the t variable that $\mathcal{A}(t_0, t_1, \phi) = \mathcal{A}(0, t_1 - t_0, \phi)$ for any initial function $\phi \in W_2^{(1)}([-h, 0], R^n)$. It is also evident that $\mathcal{A}(0, t_1, 0) \subseteq \mathcal{A}(0, t'_1, 0)$ if $t_1 \leq t'_1$. Indeed, if $\psi \in \mathcal{A}(0, t_1, 0)$, then

$$\psi = x_{t_1}(\cdot, 0, 0, u)$$

for some control $u \in L_2([0, t_1], R^m)$. For $t'_1 > t_1$ define $\tilde{u} \in L_2([0, t'_1], R^m)$ by

$$\tilde{u}(t) = \begin{cases} 0, & 0 \leq t < t'_1 - t_1, \\ u(t + t_1 - t'_1), & t'_1 - t_1 \leq t \leq t'_1; \end{cases}$$

clearly $x_{t'_1}(\cdot, 0, 0, \tilde{u}) = x_{t_1}(\cdot, 0, 0, u) = \psi$ and (5.24) follows. To complete the proof we may put Theorem 5.2 in a form which provides a more direct characterization of $\mathcal{A}(0, t_1, 0)$. Define $\tilde{x}(\theta) = x(t_1 + \theta)$, $\tilde{\psi}(\theta) = \psi(t_1 + \theta)$ and $\tilde{\omega}^k(\theta) = \omega^k(t_1 + \theta)$. For $t_1 > nh$, then, since the coefficients in $K^k(D)$ are constant, Theorem 5.2 states, in effect, that $\mathcal{A}(0, t_1, 0)$ is the set of functions $\tilde{\psi} \in W_2^{(1)}([-h, 0], R^n)$ for which

there exist functions $\tilde{\omega}^k \in W_2^{(n)}([-h, 0], R^n)$, $k = 1, 2, \dots, m$, satisfying

$$(5.25) \quad \sum_{k=1}^m K^k(D) \tilde{\omega}^k(\theta) = \tilde{\psi}(\theta), \quad \theta \in [-h, 0],$$

and for $k = 1, \dots, m$,

$$(5.26) \quad D^i \tilde{\omega}_j^k(-h) = D^i \tilde{\omega}_{j+1}^k(0), \quad i = 0, 1, \dots, n-1, \quad j = 1, \dots, n-1.$$

Clearly this set does not depend on t_1 .

Remark 5.2. We observe that the above theorem and corollary indicate that the set reachable from $\phi = 0$ on $[0, t]$ does not increase with t for $t > nh$, while it may be increasing in t for $t < nh$. It is thus not surprising that we find examples of systems that are null controllable on $[0, t_1]$ for $t_1 > nh$ which are not null controllable on $[0, t_1]$ for $t_1 \leq nh$ (see Examples 6.3, 6.4 below).

6. Explicit determination of the reachable set and some applications. For equation (5.1) the reachable set $\mathcal{A}(0, t_1, 0)$ is implicitly defined through Theorem 5.2 as the class of functions $\theta \mapsto \tilde{\psi}(\theta) = \psi(t_1 + \theta)$ in $W_2^{(1)}([-h, 0], R^n)$ for which (5.19) has a solution ω^k , $k = 1, \dots, m$, satisfying conditions (5.20) and also (5.21), (5.22) depending on the size of $t_1 > 0$. The restrictions on ψ in order that this boundary value problem have a solution fall into two classes: those that are imposed by the requirement that (5.19) be consistent (have a solution $\omega^1, \dots, \omega^m$ in the appropriate space $W_2^{(n)}$), and those, assuming (5.19) is consistent, that are imposed by the relevant boundary conditions (5.20)–(5.22). In this section we discuss procedures for ascertaining the restrictions, if any, of the first class. Little of a general nature can be said about those of the second class. Both are illustrated below in Example 6.2.

Let us introduce the $n \times nm$ operator matrix

$$(6.1) \quad K(D) = [K^1(D), \dots, K^m(D)]$$

and the nm -vector function

$$(6.2) \quad \Omega(t) = \begin{bmatrix} \omega^1(t) \\ \vdots \\ \omega^m(t) \end{bmatrix}.$$

In this notation (5.19) assumes the form

$$(6.3) \quad K(D)\Omega(t) = \psi(t), \quad t \in J,$$

where J is some real interval. The coordinates of Ω are those ω_j^k of the vectors ω^k . For each j , k some highest ordered derivative of ω_j^k is present on the left in (6.3). By introducing subsidiary variables for the lower order derivatives, we may convert (6.3) to a first order system of differential equations. In this form the system may be analyzed as described in [10, Chap. 12, pp. 45–49]. One finds generally that in order for (6.3) to be consistent certain well-defined constant-coefficient linear dependence relations must hold among the coordinates of ψ and their derivatives. The character of the consistency conditions and the number of arbitrary coordinate functions and constants of integration are determined by certain algebraic invariants intrinsic to the polynomial matrix $K(D)$.

We propose to outline here an alternative program for studying the consistency of (6.3). Accordingly, we note that since each $K^k(D)$ is at most of degree $n - 1$ in D , then we may write

$$(6.4) \quad K(D) = K_0 D^{n-1} + \cdots + K_{n-1},$$

where each coefficient K_k is a constant real $n \times nm$ matrix. Since $P_0(D) = \text{adj}(ID - A_0) = ID^{n-1} + \text{lower order terms in } D$, we see from (5.9) that the $(n(j-1) + 1)$ th column in K_0 is b^j . Hence if $K_0 = 0$, then $B = 0$. We assume, of course, that $B \neq 0$, whence $K_0 \neq 0$.

Below we show that if rank $K_0 = n$, then (6.3) has a family of solutions $\Omega \in W_2^{(n)}(J, R^{nm})$ for any $\psi \in W_2^{(1)}(J, R^n)$. However, if rank $K_0 < n$, then ψ must satisfy certain conditions in order that (6.3) have such solutions. These conditions can be determined by operating on (6.3) with first order $n \times n$ matrix operators of a special type which reduce $K(D)$ to a canonical form analogous to the row-echelon form in solving a system of linear equations.

If rank $K_0 = n$ in (6.4), then there is a constant nonsingular $nm \times nm$ matrix C such that

$$K_0 C = [\tilde{K}_0, \hat{K}_0],$$

where \tilde{K}_0 has n columns and $\det \tilde{K}_0 \neq 0$. (When $m = 1$, the submatrix \hat{K}_0 does not appear.) We write

$$C^{-1}\Omega = \begin{bmatrix} \tilde{\Omega} \\ \hat{\Omega} \end{bmatrix}, \quad K(D)C = [\tilde{K}(D), \hat{K}(D)],$$

where $\tilde{\Omega}$ is $n \times 1$ and $\tilde{K}(D)$ is $n \times n$. Equation (6.3) can then be written

$$(6.5) \quad \tilde{K}(D)\tilde{\Omega}(t) = \psi(t) - \hat{K}(D)\hat{\Omega}(t).$$

Now let $\Delta(D) = \det \tilde{K}(D)$. Since $\det \tilde{K}_0 \neq 0$, then $\Delta(D)$ is a polynomial in D of degree $\delta = n(n-1)$. For any scalar function ξ continuous on J let $\Delta^{-1}\xi$ denote the unique solution y of the initial value problem

$$\Delta(D)y(t) = \xi(t), \quad t \in J,$$

$$D^i y(\sigma) = 0, \quad i = 0, 1, \dots, \delta - 1,$$

where σ is some fixed number in J . We extend the operator Δ^{-1} coordinate-wise to n -vector functions; that is, $(\Delta^{-1}\xi)_j = \Delta^{-1}\xi_j$ for $j = 1, \dots, n$.

THEOREM 6.1. Suppose $\tilde{K}(D)$ in (6.5) is $n \times n$, of degree $n-1$ in D , $\det \tilde{K}_0 \neq 0$, and $\hat{K}(D)$ is at most of degree $n-1$ in D . If $\psi \in W_2^{(p)}(J, R^n)$ and $\hat{\Omega} \in W_2^{(q)}(J, R^{nm-n})$ with $q - n + 1 \geq p \geq 0$, then (6.5) has a solution $\tilde{\Omega} \in W_2^{(n+p-1)}(J, R^n)$; moreover, each solution $\tilde{\Omega} \in W_2^{(n)}$ is actually in $W_2^{(n+p-1)}$ and is given by

$$(6.6) \quad \tilde{\Omega} = H(D) \left[\Delta^{-1}(\psi - \hat{K}(D)\hat{\Omega}) + \sum_{k=1}^{\delta} \gamma^k y_k \right]$$

for some constant n -vector γ^k where

$$H(D) = \text{adj } \tilde{K}(D)$$

and y_1, \dots, y_{δ} are a basis for the solutions y of

$$(6.7) \quad \Delta(D)y = 0.$$

(Note that $p = 1, q = n$ gives $\Omega \in W_2^{(n)}(J, R^{nm})$.)

Proof. Under the given hypotheses we see that $\psi - \hat{K}(D)\hat{\Omega} \in W_2^{(p)}(J, R^n)$ so $\Delta^{-1}(\psi - \hat{K}(D)\hat{\Omega}) \in W_2^{(p+\delta)}(J, R^n)$. Since the degree of $H(D)$ is at most $(n-1)(n-1) = \delta + 1 - n$ and the y_k are infinitely differentiable, any $\tilde{\Omega}$ given by (6.6) is in $W_2^{(n+p-1)}(J, R^n)$. Moreover, since $I\Delta(D) = \tilde{K}(D)H(D)$ and $\tilde{K}(D)$ is of degree $n-1$, any such $\tilde{\Omega}$ satisfies (6.5) on J by virtue of (6.7) and the definition of Δ^{-1} .

Conversely, suppose $\tilde{\Omega} \in W_2^{(n)}(J, R^n)$ is a solution of (6.5), where $\psi \in W_2^{(p)}(J, R^n)$ and $\hat{\Omega} \in W_2^{(q)}(J, R^n)$ with $q - n + 1 \geq p \geq 0$. Then $\Delta^{-1}\tilde{\Omega} \in W_2^{(n+\delta)}(J, R^n)$ so that $\zeta \equiv \tilde{K}(D)\Delta^{-1}\tilde{\Omega} \in W_2^{(\delta+1)}(J, R^n)$. Now $H(D)$ is of degree at most $(n-1)(n-1) = \delta + 1 - n$ so that we may apply $H(D)$ to ζ obtaining

$$(6.8) \quad H(D)\zeta = \tilde{\Omega}.$$

Since $\zeta \in W_2^{(\delta+1)}(J, R^n)$ and $I\Delta(D) = \tilde{K}(D)H(D)$, we have

$$\tilde{K}(D)[H(D)\zeta] = [\tilde{K}(D)H(D)]\zeta,$$

so that applying $\tilde{K}(D)$ to (6.8) yields

$$I\Delta(D)\zeta = \tilde{K}(D)\tilde{\Omega} = \psi - \hat{K}(D)\hat{\Omega},$$

or

$$\zeta = \Delta^{-1}(\psi - \hat{K}(D)\hat{\Omega}) + \sum_{k=1}^{\delta} \gamma^k y_k.$$

It follows immediately from (6.8) that $\tilde{\Omega}$ has the form (6.6). As we have shown above, any such $\tilde{\Omega}$ is in $W_2^{(n+p-1)}(J, R^n)$.

Remark 6.1. The consistency conclusion provided by Theorem 6.1 under the key hypothesis $\det \tilde{K}_0 \neq 0$ also follows easily from a reduction of (6.5) to a first order system. The representation (6.6) is a slight generalization of that appearing in Frazer [8a, p. 183]. The homogeneous equation $\tilde{K}(D)\tilde{\Omega} = 0$ is equivalent to a first order system for a vector with $n(n-1) = \delta$ coordinates. Hence the set of solutions of $\tilde{K}(D)\tilde{\Omega} = 0$ is a vector space of dimension δ ; thus there are not as many independent arbitrary constants of integration actually in (6.6) as would appear from the totality of coordinates of the γ^k . However, since there are δ arbitrary constants one may expect that the δ conditions (5.20) can be satisfied. This is indeed so in Example 6.1 below.

We now turn to the case when in (6.4) $\text{rank } K_0 = \rho < n$. Below we let $\mu = \text{rank } K(D)$ and observe that $\rho \leq \mu \leq n$. Since $K_0 \neq 0$, $\rho \geq 1$. There is then a nonsingular constant matrix T_0 such that the last $n - \rho$ rows of $T_0 K_0$ are zero. Accordingly, (6.3) is equivalent to

$$(6.9) \quad T_0 K(D)\Omega(t) = T_0 \psi(t), \quad t \in J.$$

We define $M(D)$ and $U(D)$ by

$$(6.10) \quad \begin{bmatrix} M(D) \\ U(D) \end{bmatrix} = T_0 K(D),$$

where $M(D)$ is $\rho \times nm$ and, by (6.4),

$$M(D) = M_0 D^{n-1} + \cdots + M_{n-1}$$

with rank $M_0 = \rho$. The submatrix $U(D)$ is $(n - \rho) \times nm$ and has the form

$$U(D) = U_1 D^{n-2} + \cdots + U_{n-1}.$$

If τ_j , $1 \leq j \leq n - \rho$, denotes the $(\rho + j)$ th row of T_0 , then we see that

$$(6.11) \quad \tau_j \psi \in W_2^{(2)}(J, R)$$

is a necessary condition that (6.9) have a solution $\Omega \in W_2^{(n)}(J, R^{nm})$; even more strongly, we must have

$$(6.12) \quad \tau_j \psi = 0$$

in case the j th row of $U(D)$ is zero.

Now consider $n \times n$ matrix operators of the form

$$(6.13) \quad T(D) = \begin{bmatrix} I & 0 & 0 \\ v & D + a & 0 \\ 0 & 0 & E \end{bmatrix},$$

where I is the $v \times v$ identity, $1 \leq v \leq n - 1$, a is a complex number, v is $1 \times v$ with complex elements, and E is an identity matrix which, along with the corresponding rows and columns, is absent from (6.13) if $v = n - 1$. If $v = \rho - 1 + j$, $1 \leq j \leq n - \rho$, then (6.9) for $\Omega \in W_2^{(n)}(J, R^{nm})$ implies

$$(6.14) \quad T(D)T_0K(D)\Omega(t) = T(D)T_0\psi(t), \quad t \in J,$$

in view of (6.11). Let $u_j(D)$ denote the j th row of $U(D)$ in (6.10) and let $\sigma \in J$. Then one may easily show that

$$(6.15) \quad u_j(D)\Omega(\sigma) = \tau_j\psi(\sigma)$$

and (6.14) imply (6.9).

It can be shown [20] that premultiplication of $T_0K(D)$ by a finite number of matrix operators of the type (6.13), interspersed perhaps with some nonsingular constant matrices which permute only the last $n - \rho$ rows, gives

$$T_q \cdots T_1 T_0 K(D) = \begin{bmatrix} \tilde{M}(D) \\ 0 \end{bmatrix},$$

where

$$\tilde{M}(D) = \tilde{M}_0 D^{n-1} + \cdots + \tilde{M}_{n-1}$$

is $\mu \times nm$ with $\mu = \text{rank } K(D) = \text{rank } \tilde{M}_0$. Here T_1, \dots, T_q are the required operators and permutation matrices. Moreover, after multiplication by each T_k the product to that stage remains a polynomial of degree $n - 1$ in D . If $\tilde{\psi}$ consists of the first μ rows of $T_q \cdots T_1 T_0 \psi$, then (6.9) is equivalent to

$$(6.16) \quad \tilde{M}(D)\Omega(t) = \tilde{\psi}(t), \quad t \in J,$$

along with the corresponding conditions of type (6.11) or (6.12) and (6.15) which appear in connection with operators T_k of the type (6.13). These conditions represent the necessary and sufficient conditions for the consistency of (6.3) in the context of ψ being in $W_2^{(1)}(J, R^n)$ and Ω in $W_2^{(n)}(J, R^{nm})$. When these are satis-

fied the solution of (6.3) may then be obtained by applying Theorem 6.1 to (6.16).

The following two examples illustrate the theory and technique in §§ 5 and 6 to this point.

Example 6.1.

$$\begin{aligned}\dot{x}_1(t) &= \dot{x}_2(t-h) + x_1(t-h), \\ \dot{x}_2(t) &= u(t).\end{aligned}$$

Here $B = b^1$, a single column, so $K(D) = K^1(D)$, $\Omega = \omega^1$ and equation (6.3) in this case is

$$(6.17) \quad \begin{bmatrix} 0 & D \\ D & -1 \end{bmatrix} \Omega(t) = \psi(t).$$

Note that $\tilde{K}(D) = K(D)$ satisfies the hypotheses of Theorem 6.1. Since $\Delta(D) = \det \tilde{K}(D) = -D^2$, the operator Δ^{-1} with $\sigma = t_1 - h$ is given by

$$\Delta^{-1} \xi(t) = \int_{t_1-h}^t (s-t) \xi(s) ds.$$

A basis for the solutions of $\Delta(D)y = 0$ is $\{1, t\}$ so, using (6.6), we find the general solution of (6.17) is

$$(6.18) \quad \Omega(t) = \begin{bmatrix} \int_{t_1-h}^t [(t-s)\psi_1(s) + \psi_2(s)] ds + c_1 + c_2 t \\ \int_{t_1-h}^t \psi_1(s) ds + c_2 \end{bmatrix},$$

where $c_1 = -\gamma_1^1 - \gamma_2^2$ and $c_2 = -\gamma_1^2$ in terms of the coordinates of γ^1 and γ^2 .

Assuming $t_1 > 2h$, we now impose conditions (5.20); viz., $\Omega_1(t_1 - h) = \Omega_2(t_1)$ and $D\Omega_1(t_1 - h) = D\Omega_2(t_1)$. From (6.18) these are, respectively,

$$(6.19) \quad c_1 + c_2(t_1 - h) = \int_{t_1-h}^{t_1} \psi_1(s) ds + c_2,$$

$$(6.20) \quad \psi_2(t_1 - h) + c_2 = \psi_1(t_1).$$

Regardless of $\psi \in W_2^{(1)}([t_1 - h, t_1], R^2)$ we can find c_2 to satisfy (6.20) and then c_1 to satisfy (6.19) for this c_2 . Thus the system is completely controllable from zero; for $t_1 > 2h$ we have $\mathcal{A}(0, t_1, 0) = W_2^{(1)}([-h, 0], R^2)$. Using $\omega_1^1 = \Omega_1$ extended backward as in the proof of Theorem 5.2, we can use (5.12) to calculate a control u which would transfer x from zero to a prescribed ψ on $[t_1 - h, t_1]$.

Finally, we point out that for any $\phi \in W_2^{(1)}([-h, 0], R^n)$, the set $\mathcal{A}(0, t_1, \phi)$ is a translate of $\mathcal{A}(0, t_1, 0)$. Hence the system under consideration in this example is controllable on $[0, t_1]$ if $t_1 > 2h$. Yet we have $B = (0 \ 1)^*$ does not have rank 2, so that Theorem 3.1 does not extend to systems of neutral type.

Example 6.2.

$$\begin{aligned}\dot{x}_1(t) &= x_1(t) + u(t), \\ \dot{x}_2(t) &= x_1(t-h) + au(t),\end{aligned}$$

where a is constant. In this case (6.3) becomes

$$\begin{bmatrix} D & 0 \\ a(D-1) & 1 \end{bmatrix} \Omega(t) = \psi(t)$$

for which $\text{rank } K(D) = 2$. We may take

$$T_0 = \begin{bmatrix} 1 & 0 \\ -a & 1 \end{bmatrix}, \quad T_1 = \begin{bmatrix} 1 & 0 \\ a & D \end{bmatrix}$$

and then,

$$(6.21) \quad T_1 T_0 K(D) = \begin{bmatrix} D & 0 \\ 0 & D \end{bmatrix} = ID,$$

the degrees of $K(D)$, $T_0 K(D)$ and $T_1 T_0 K(D)$ in D all being one. After applying T_0 , we recognize the condition (see (6.11))

$$\psi_2 - a\psi_1 \in W_2^{(2)}([t_1 - h, t_1], R)$$

and, after applying T_1 , we must add to this the condition (see (6.15))

$$(6.22) \quad \Omega_2(t_1) - a\Omega_1(t_1) = \psi_2(t_1) - a\psi_1(t_1).$$

The transformed system assumes the form (see (6.21))

$$(6.23) \quad \begin{aligned} \dot{\Omega}_1(t) &= \psi_1(t), \\ \dot{\Omega}_2(t) &= a\psi_1(t) + \dot{\psi}_2(t) - a\dot{\psi}_1(t) \end{aligned}$$

from which we may write, using (6.22),

$$(6.24) \quad \begin{aligned} \Omega_1(t) &= \Omega_1(t_1 - h) + \int_{t_1 - h}^t \psi_1(s) ds, \\ \Omega_2(t) &= a\Omega_1(t_1) + a \int_{t_1}^t \psi_1(s) ds + \psi_2(t) - a\psi_1(t). \end{aligned}$$

We now impose the conditions

$$(6.25) \quad \Omega_1(t_1 - h) - \Omega_2(t_1) = 0$$

and $\dot{\Omega}_1(t_1 - h) = \dot{\Omega}_2(t_1)$. From (6.23) this becomes

$$a\psi_1(t_1) + \dot{\psi}_2(t_1) - a\dot{\psi}_1(t_1) - \psi_1(t_1 - h) = 0.$$

From (6.24) we get

$$\Omega_1(t_1) = \Omega_1(t_1 - h) + \int_{t_1 - h}^{t_1} \psi_1(s) ds$$

so (6.22) becomes

$$(6.26) \quad -a\Omega_1(t_1 - h) + \Omega_2(t_1) = \psi_2(t_1) - a\psi_1(t_1) + a \int_{t_1 - h}^{t_1} \psi_1(s) ds.$$

If $a \neq 1$, then (6.25) and (6.26) can be solved for $\Omega_1(t_1 - h)$ and $\Omega_2(t_1)$ in any

case; if $a = 1$, then a solution exists if and only if

$$\psi_2(t_1) - a\psi_1(t_1) + a \int_{t_1-h}^{t_1} \psi_1(s) ds = 0.$$

By translation of the t -axis we now have from Theorem 5.2 that for $t_1 > 2h$ the set $\mathcal{A}(0, t_1, 0)$ consists of all functions $\tilde{\psi} \in W_2^{(1)}([-h, 0], R^2)$ such that

$$\tilde{\psi}_2 - a\tilde{\psi}_1 \in W_2^{(2)}([-h, 0], R)$$

and

$$a\tilde{\psi}_1(0) + \dot{\tilde{\psi}}_2(0) - a\dot{\tilde{\psi}}_1(0) - \tilde{\psi}_1(-h) = 0,$$

and, in case $a = 1$, also

$$\tilde{\psi}_2(0) - a\tilde{\psi}_1(0) + a \int_{-h}^0 \tilde{\psi}_1(s) ds = 0.$$

Theorem 5.2 can, under some circumstances, give information regarding null controllability of (5.1). Here we let $x(t, \phi, u)$ denote the solution of (5.1) on $[0, t_1]$ which is continuous on $[-h, t_1]$ and satisfies $x_0(\cdot, \phi, u) = \phi$, where $\phi \in W_2^{(1)}([-h, 0], R^n)$. Since

$$x(t, \phi, u) = x(t, 0, u) + x(t, \phi, 0), \quad t \geq -h,$$

and $x(t, \phi, u)$ is linear in u , we see that $x(t, \phi, u) = 0$ on some interval $J \subset [-h, t_1]$ if and only if for some $u \in W_{2,0}^{(0)}(t_1, R^m)$ we have

$$(6.27) \quad x(t, 0, u) = x(t, \phi, 0), \quad t \in J.$$

THEOREM 6.2. Suppose $A_{-1} = 0$, A_1 is nonsingular, $B = b$ is $n \times 1$ and the degree of $\det K(D)$ in D is $\frac{1}{2}n(n-1)$. If $\phi \in W_2^{(1)}([-h, 0], R^n)$ and (6.27) holds for $t \in [t_1 - h, t_1]$ for some $t_1 > nh$ and $u \in W_{2,0}^{(0)}(t_1, R)$, then for any $\varepsilon > 0$ such that $t_2 = t_1 + \varepsilon - h > nh$, there is a control $\tilde{u} \in W_{2,0}^{(0)}(t_2, R)$ such that

$$(6.28) \quad x(t, 0, \tilde{u}) = x(t, \phi, 0), \quad t \in [t_2 - h, t_2].$$

Proof. We extend u so that

$$(6.29) \quad u(t) = 0, \quad t > t_1.$$

We may then assume that (6.27) holds for $t \in J = [t_1 - h, t_1 + \varepsilon]$ for any $\varepsilon > 0$. If $\varepsilon \geq h$ we may choose $\tilde{u} = u$ to get (6.28) so we assume for the arguments below that $\varepsilon < h$.

Now by Theorem 5.1 and (6.27) there exists $\omega \in W_{2,0}^{(n)}(t_1 + \varepsilon, R^n)$ such that

$$(6.30) \quad K(D)\omega(t) = x(t, \phi, 0), \quad t \in J,$$

with (5.11) and (5.12) holding; that is,

$$(6.31) \quad S\omega_j = \omega_{j+1}, \quad j = 1, \dots, n-1,$$

$$(6.32) \quad \det Q(D, S)\omega_1 = u.$$

We now define the n -vector function Ω to be any particular solution of the system

$$D^{j-1}\Omega_j(t) = \omega_j(t), \quad j = 1, \dots, n, \quad t \in J.$$

Then $\Omega \in W_2^{(n)}(J, R^n)$ and (6.30) implies

$$(6.33) \quad [P_0(D)b, DP_1(D)b, \dots, D^{n-1}P_{n-1}(D)b]\Omega(t) = x(t, \phi, 0), \quad t \in J.$$

If $\tilde{K}(D)$ denotes the operator matrix on the left in (6.33), then

$$\det \tilde{K}(D) = D^{(1/2)n(n-1)} \det K(D)$$

so the degree of $\det \tilde{K}(D)$ in D is $n(n-1)$ by our hypothesis. But since $A_{-1} = 0$, the degree of $P_j(D)$ is at most $n-j-1$, so the degree of $\tilde{K}(D)$ is at most $n-1$ and it follows that $\det \tilde{K}_0 \neq 0$; that is we may apply Theorem 6.1 to (6.33). In doing this we note that there is no $\hat{K}(D)$ or $\hat{\Omega}$ present in this instance. Moreover, $x(\cdot, \phi, 0)$ restricted to J is in $W_2^{(n+1)}(J, R^n)$ since $t_1 > nh$ and $\phi \in W_2^{(1)}([-h, 0], R^n)$. (It is easy to see that solutions of (5.1) with $A_{-1} = 0$ and $u = 0$ get progressively smoother as t is increased by h .) By Theorem 6.1 we conclude that $\Omega \in W_2^{(2n)}(J, R^n)$, whence

$$(6.34) \quad \omega_j \in W_2^{(2n-j+1)}(J, R), \quad j = 1, \dots, n.$$

When $A_{-1} = 0$ we may write

$$(6.35) \quad \det Q(D, S) = \sum_{j=0}^n m_j(D)S^j,$$

where $m_j(D)$ is a scalar polynomial in D of degree $n-j$. By (6.31) and (6.32) the control u is then given by

$$(6.36) \quad u(t) = \sum_{j=1}^n m_{j-1}(D)\omega_j(t) + m_n(D)\omega_1(t-nh).$$

But $m_n(D) = m_n$ is a constant and $\omega_1 \in W_{2,0}^{(n)}(t_1 + \varepsilon, R)$ so it follows by (6.34) that u restricted to J is in $W_2^{(n)}(J, R)$. Hence the mapping $t \rightarrow u(t+h)$ satisfies

$$(6.37) \quad u(\cdot + h) \in W_2^{(n)}([t_2 - h, t_2], R),$$

since $[t_2, t_2 + h] \subset J = [t_1 - h, t_1 + \varepsilon]$.

Substituting (5.3), (5.6) and (6.35) into (5.5) and equating coefficients of like powers of S , we derive the following relations:

$$(6.38) \quad \begin{aligned} I m_0(D) &= (DI - A_0)P_0(D), \\ I m_j(D) &= (DI - A_0)P_j(D) - A_1P_{j-1}(D), \quad j = 1, \dots, n-1, \\ I m_n(D) &= -A_1P_{n-1}(D). \end{aligned}$$

Now $P_{n-1}(D) = \text{adj}(-A_1)$ so $m_n = m_n(D) = \det(-A_1)$ is a nonzero constant. By (6.30) and (5.1) with $u = 0$ it follows that

$$(6.39) \quad x(t, \phi, 0) = A_1^{-1}(DI - A_0)K(D)\omega(t+h), \quad t \in [t_2 - h, t_2].$$

Using (6.38), we get that

$$A_1^{-1}(DI - A_0)K(D) = A_1^{-1}[bm_0(D), \dots, bm_{n-1}(D)] + K(D)N,$$

where

$$N = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}.$$

Hence, by (6.36) and (6.39) we see that for $t \in [t_2 - h, t_2]$,

$$x(t, \phi, 0) = A_1^{-1}b[u(t+h) - m_n\omega_1(t+h-nh)] + K(D)N\omega(t+h).$$

From the last relation in (6.38) it now follows that

$$(6.40) \quad K(D)\tilde{\omega}(t) = x(t, \phi, 0), \quad t \in [t_2 - h, t_2],$$

where for $t \in [t_2 - h, t_2]$ we define

$$(6.41) \quad \tilde{\omega}_j(t) = \omega_{j+1}(t+h) = \omega_1(t+h-jh), \quad j = 1, \dots, n-1,$$

$$(6.42) \quad \tilde{\omega}_n(t) = \omega_1(t+h-nh) - \frac{1}{m_n}u(t+h).$$

Observe that $\tilde{\omega} \in W_2^{(n)}([t_2 - h, t_2], R^n)$ in view of (6.37) and the fact that $\omega_1 \in W_{2,0}^{(n)}(t_1 + \varepsilon, R)$. Assuming $t_2 > nh$, we may now verify the analogues of (5.20), viz.,

$$(6.43) \quad D^i\tilde{\omega}_j(t_2 - h) = D^i\tilde{\omega}_{j+1}(t_2), \quad i = 0, 1, \dots, n-1,$$

for $j = 1, \dots, n-1$. These clearly hold for $j = 1, \dots, n-2$ by (6.41) and the smoothness of ω_1 . They hold for $j = n-1$ for the same reason by (6.42) and (6.29) since $t_2 + h = t_1 + \varepsilon$. Now (6.40) and (6.43) imply the conclusion (6.28), by virtue of Theorem 5.2.

COROLLARY 6.1. Suppose $A_{-1} = 0$, A_1 is nonsingular, $B = b$ is $n \times 1$ and the degree of $\det K(D)$ in D is $\frac{1}{2}n(n-1)$. Let $x(t, \phi, u)$ be the solution of (5.1) with $x_0(\cdot, \phi, u) = \phi$. If $t_1 > nh$ and $x_{t_1}(\cdot, \phi, u) = 0$, then for any $t_2 > nh$ there is a control $\tilde{u} \in W_{2,0}^{(0)}(t_2, R)$ such that $x_{t_2}(\cdot, \phi, \tilde{u}) = 0$.

Proof. The conclusion is clearly true for $t_2 > t_1$; we need merely to set $u(t) = 0$ for $t > t_1$. For $nh < t_2 < t_1$, let $k \geq 1$ be an integer such that $\varepsilon = h - (t_1 - t_2)/k > 0$. Since $x_{t_1}(\cdot, 0, -u) = x_{t_1}(\cdot, \phi, 0)$ and $t_1 + \varepsilon - h \geq t_2 > nh$, we may apply Theorem 6.2 to conclude the existence of a control \tilde{u} such that $x_{t_1+\varepsilon-h}(\cdot, 0, -\tilde{u}) = x_{t_1+\varepsilon-h}(\cdot, \phi, 0)$. But $t_2 = t_1 + k(\varepsilon - h)$ so we may proceed inductively if $k > 1$, and finally conclude the existence of a control $\tilde{u} \in W_{2,0}^{(0)}(t_2, R)$ such that $x_{t_2}(\cdot, 0, -\tilde{u}) = x_{t_2}(\cdot, \phi, 0)$. This, in effect, completes the proof.

We consider once again the example discussed previously (see Example 4.1). In addition to providing an example which is null controllable on $[0, t_1]$ for $t_1 > nh$ but not null controllable for $t_1 \leq nh$, this example illustrates nicely the use of the ideas in §§ 5 and 6 to ascertain null controllability.

Example 6.3.

$$\begin{aligned} \dot{x}_j(t) &= x_{j+1}(t-h), & j &= 1, \dots, n-1, \\ \dot{x}_n(t) &= u(t). \end{aligned}$$

We show that this system is null controllable on $[0, t_1]$ for $t_1 > nh$. From the above discussions, it suffices to show that any ψ of the form $\psi(t) = x(t, \phi, 0)$, $t \in [t_1 - h, t_1]$, where $\phi \in W_2^{(1)}([-h, 0], R^n)$, is attainable from the zero function.

An easy calculation yields

$$(6.44) \quad K(D) = \begin{bmatrix} 0 & \cdot & \cdot & \cdot & \cdot & 0 & 1 \\ 0 & \cdot & \cdot & \cdot & 0 & D & 0 \\ 0 & \cdot & \cdot & 0 & D^2 & 0 & 0 \\ \vdots & & & & & \vdots & \\ \vdots & & & & & \vdots & \\ 0 & D^{n-2} & 0 & \cdot & \cdot & \cdot & 0 \\ D^{n-1} & 0 & \cdot & \cdot & \cdot & \cdot & 0 \end{bmatrix},$$

and we observe that the degree of $\det K(D)$ is $\frac{1}{2}n(n-1)$ (see Theorem 6.2 above).

If ψ is given as above, we see that

$$(6.45) \quad \begin{aligned} \dot{\psi}_j(t) &= \psi_{j+1}(t-h), & j &= 1, 2, \dots, n-1, \\ \dot{\psi}_n(t) &= 0. \end{aligned}$$

It follows that for $t_1 > nh$, $\psi_j \in C^\infty([t_1 - h, t_1], R)$ for $j = 1, 2, \dots, n$.

We apply Theorem 5.2 to ensure that ψ is attainable. For the given ψ on $[t_1 - h, t_1]$ we define

$$(6.46) \quad \begin{aligned} \omega_n(t) &= \psi_1(t), \\ \omega_{n-j}(t) &= \sum_{k=0}^{j-1} \frac{1}{k!} D^k \omega_{n-j+1}(t_1) \{t - (t_1 - h)\}^k \\ &\quad + \int_{t_1-h}^t ds_j \int_{t_1-h}^{s_j} \cdots ds_2 \int_{t_1-h}^{s_2} \psi_{j+1}(s_1) ds_1 \\ &\quad \text{for } j = 1, 2, \dots, n-1. \end{aligned}$$

Here the multiple integrals in defining ω_{n-j} are j -fold integrals.

Defining the n -vector ω as in (6.46), we obtain $\omega \in C^\infty([t_1 - h, t_1], R^n)$ and furthermore we have on $[t_1 - h, t_1]$,

$$(6.47) \quad D^j \omega_{n-j}(t) = \psi_{j+1}(t), \quad j = 0, 1, \dots, n-1,$$

or, in light of (6.44),

$$(6.48) \quad K(D)\omega(t) = \psi(t), \quad t \in [t_1 - h, t_1].$$

Furthermore, for $j = 1, 2, \dots, n-1$, from (6.46) we have for $0 \leq i \leq j-1$,

$$D^i \omega_{n-j}(t_1 - h) = D^i \omega_{n-j+1}(t_1),$$

while for $i = j$ we find

$$\begin{aligned} D^j \omega_{n-j}(t_1 - h) &= \psi_{j+1}(t_1 - h) = \dot{\psi}_j(t_1) = D(D^{j-1} \omega_{n-(j-1)})(t_1) \\ &= D^j \omega_{n-j+1}(t_1). \end{aligned}$$

For $i > j$, (6.47) along with (6.45) imply

$$\begin{aligned} D^i \omega_{n-j}(t_1 - h) &= D^{i-j} D^j \omega_{n-j}(t_1 - h) \\ &= D^{i-j} D^j \omega_{n-j+1}(t_1) = D^i \omega_{n-j+1}(t_1). \end{aligned}$$

We therefore have

$$(6.49) \quad D^i \omega_{n-j}(t_1 - h) = D^i \omega_{n-j+1}(t_1), \quad i = 0, 1, \dots, n-1, j = 1, \dots, n-1.$$

In light of (6.48) and (6.49), Theorem 5.2 yields the existence of $u \in W_{2,0}^{(0)}(t_1, R)$ such that

$$x(t, 0, u) = \psi(t), \quad t \in [t_1 - h, t_1],$$

which implies null controllability. We remark that using the backward extension of ω_1 as in (5.23) one can, using (5.12), easily obtain a control u which drives $\phi \in W_2^{(1)}([-h, 0], R^n)$ to zero at time $t_1 > nh$ (i.e., $x_{t_1}(\phi, u) = 0$). Note, however, that one must know $\psi(t) = x(t, \phi, 0)$, $t \in [t_1 - h, t_1]$, in order to compute ω using (6.46).

The next example shows that neither of the restrictive assumptions of Theorem 6.2 ($\deg \det K(D) = \frac{1}{2}n(n-1)$ and A_1 nonsingular) is needed in using Theorem 5.2 to determine null controllability.

Example 6.4.

$$\begin{aligned} \dot{x}_1(t) &= x_2(t - h) + u(t), \\ \dot{x}_2(t) &= x_1(t) + x_2(t - h). \end{aligned}$$

For this example, it is also easy to argue that for null controllability on $[0, t_1]$ one must have $t_1 > nh = 2h$. As in the previous example, we show that the system is null controllable for $t_1 > 2h$ by using Theorem 5.2 to verify that any ψ satisfying on $[t_1 - h, t_1]$

$$(6.50) \quad \begin{aligned} \dot{\psi}_1(t) &= \psi_2(t - h), \\ \dot{\psi}_2(t) &= \psi_1(t) + \psi_2(t - h) \end{aligned}$$

lies in the set of functions reachable from the zero function, if $\psi(t) = x(t, \phi, 0)$, $\phi \in W_2^{(1)}([-h, 0], R^2)$.

We find that

$$K(D) = \begin{bmatrix} D & -1 \\ 1 & 0 \end{bmatrix}$$

so that $\deg \det K(D) = 0 \neq 1 = \frac{1}{2}n(n-1)$. Also note that

$$A_1 = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$$

is singular. The equation $K(D)\omega = \psi$ then becomes

$$(6.51) \quad \begin{aligned} D\omega_1(t) - \omega_2(t) &= \psi_1(t), \\ \omega_1(t) &= \psi_2(t). \end{aligned}$$

We observe that for $\phi \in W_2^{(1)}$ and $t_1 > 2h$, $\psi \in W_2^{(3)}([t_1 - h, t_1], R^2)$. We then define

$$(6.52) \quad \begin{aligned} \omega_1(t) &= \psi_2(t), \\ \omega_2(t) &= \dot{\psi}_2(t) - \psi_1(t) \end{aligned}$$

and obtain immediately that $\omega \in W_2^{(2)}([t_1 - h, t_1], R^2)$ and satisfies (6.51). Finally, since (6.50) holds, we have

$$\begin{aligned} \psi_2(t_1 - h) &= \dot{\psi}_2(t_1) - \psi_1(t_1), \\ \dot{\psi}_2(t_1 - h) &= \ddot{\psi}_2(t_1) - \dot{\psi}_1(t_1), \end{aligned}$$

which can be written

$$(6.53) \quad D^i \omega_1(t_1 - h) = D^i \omega_2(t_1), \quad i = 0, 1.$$

Application of Theorem 5.2 then yields null controllability on $[0, t_1]$ if $t_1 > 2h$.

Finally, we present an example illustrating the use of Theorem 5.2 in determining that a system is not null controllable.

Example 6.5.

$$\begin{aligned} \dot{x}_1(t) &= x_1(t - h) + u(t), \\ \dot{x}_2(t) &= x_1(t) + u(t). \end{aligned}$$

We show that this system is not null controllable on $[0, t_1]$ for any $t_1 > 0$. It suffices to let t_1 be arbitrary, $t_1 > 2h$. In view of the necessity of the conditions in Theorem 5.2, if $\phi \in W_2^{(1)}([-h, 0], R^n)$ is controllable to the zero function, i.e., $x_{t_1}(\phi, u) = 0$, there must exist ω satisfying

$$K(D)\omega(t) = \psi(t), \quad t \in [t_1 - h, t_1]$$

and

$$D^i \omega_1(t_1 - h) = D^i \omega_2(t_1), \quad i = 0, 1,$$

where $\psi(t) = x(t, \phi, 0)$. Here we find

$$K(D) = \begin{bmatrix} D & 0 \\ D + 1 & -1 \end{bmatrix}$$

so that the first set of equations becomes

$$(6.54) \quad \begin{aligned} D\omega_1(t) &= \psi_1(t), & t \in [t_1 - h, t_1], \\ D\omega_1(t) + \omega_1(t) - \omega_2(t) &= \psi_2(t), & t \in [t_1 - h, t_1], \end{aligned}$$

while the latter equations can be written

$$(6.55) \quad \begin{aligned} \omega_1(t_1 - h) &= \omega_2(t_1), \\ D\omega_1(t_1 - h) &= D\omega_2(t_1). \end{aligned}$$

We are thus forced to define ω_1 by

$$(6.56) \quad \omega_1(t) = \omega_1(t_1 - h) + \int_{t_1 - h}^t \psi_1(s) ds$$

and ω_2 by

$$(6.57) \quad \omega_2(t) = \omega_1(t) + \psi_1(t) - \psi_2(t).$$

Since ψ must satisfy

$$(6.58) \quad \begin{aligned} \dot{\psi}_1(t) &= \psi_1(t-h), \\ \dot{\psi}_2(t) &= \psi_1(t), \end{aligned}$$

(6.56) becomes

$$\omega_1(t) = \omega_1(t_1 - h) + \psi_2(t) - \psi_2(t_1 - h),$$

which upon substitution into (6.57) yields

$$(6.59) \quad \omega_2(t) = \omega_1(t_1 - h) + \psi_1(t) - \psi_2(t_1 - h).$$

The requirements (6.55) then become

$$\begin{aligned} \psi_1(t_1) &= \psi_2(t_1 - h), \\ \psi_1(t_1 - h) &= \dot{\psi}_1(t_1). \end{aligned}$$

The latter requirement is automatically satisfied, while the first may be written

$$(6.60) \quad \psi_1(t_1) = \phi_2(0) + \int_0^{t_1-h} \psi_1(s) ds,$$

where $\phi_2(0)$ can be arbitrarily chosen and ψ_1 must be a solution of

$$\dot{\psi}_1(t) = \psi_1(t-h), \quad t > 0,$$

with $\psi_1(\theta) = \phi_1(\theta)$, $\theta \in [-h, 0]$. For the solution of this equation we find

$$\psi_1(t_1) = \psi_1(0) + \int_{-h}^0 \psi_1(\theta) d\theta + \int_0^{t_1-h} \psi_1(s) ds,$$

which, taken with (6.60), requires

$$(6.61) \quad \phi_2(0) = \phi_1(0) + \int_{-h}^0 \phi_1(\theta) d\theta.$$

Since (6.61) will not hold for arbitrary $\phi \in W_2^{(1)}([-h, 0], R^2)$, we do not have null controllability on $[0, t_1]$.

Acknowledgment. The second coauthor is happy to express his gratitude to D. C. Taylor for several fruitful discussions during the preparation of this manuscript, and to K. J. Bechmann for pointing out an improvement in Theorem 3.1.

REFERENCES

- [1] H. T. BANKS, *Representation for solutions of linear functional differential equations*, *Differential Equations*, 5 (1969), pp. 399–409.
- [2] ———, *Control of functional differential equations with function space boundary conditions*, *Delay and Functional Differential Equations and their Applications*, Klaus Schmitt, ed., Academic Press, New York, 1972, pp. 1–16.
- [3] H. T. BANKS AND G. A. KENT, *Control of functional differential equations of retarded and neutral type to target sets in function space*, this Journal, 10 (1972), pp. 567–594.

- [4] H. T. BANKS AND M. O. JACOBS, *An attainable sets approach to optimal control of functional differential equations with function space side conditions*, J. Differential Equations, 13 (1973), pp. 127–149.
- [5] JU. G. BORISOVIC AND A. S. TURBABIN, *On the Cauchy problem for linear nonhomogeneous differential equations with retarded argument*, Soviet Math. Dokl., 10 (1969), pp. 401–405.
- [6] M. C. DELFOUR AND S. K. MITTER, *Controllability, observability and optimal feedback control of affine hereditary differential systems*, this Journal, 10 (1972), pp. 298–328.
- [7] ———, *Controllability and observability for infinite-dimensional systems*, this Journal, 10 (1972), pp. 329–333.
- [8] N. DUNFORD AND J. SCHWARTZ, *Linear Operators, Part I*, Interscience, New York, 1958.
- [8a] R. A. FRAZER, W. J. DUNCAN AND A. R. COLLAR, *Elementary Matrices*, Cambridge University Press, London, 1963.
- [9] R. GABASOV AND F. KIRILLOVA, *Qualitative Theory of Optimal Processes*, Nauka Publ., Moscow, 1971.
- [10] F. R. GANTMACHER, *The Theory of Matrices*, vol. 2, Chelsea, New York, 1960.
- [11] A. HALANAY, *On the controllability of linear differential-difference equations*, Math. Systems Theory and Economics, II, H. Kuhn and G. Szegö, eds., Lecture Notes in Operations Research and Economics, No. 12, Springer-Verlag, Berlin, 1969, pp. 329–336.
- [12] J. K. HALE, *Functional Differential Equations*, Applied Mathematical Sciences, vol. 3, Springer-Verlag, New York, 1971.
- [13] P. R. HALMOS, *A Hilbert Space Problem Book*, Van Nostrand, Princeton, N.J., 1967.
- [14] H. HERMES AND J. P. LASALLE, *Functional Analysis and Time Optimal Control*, Academic Press, New York, 1969.
- [15] M. Q. JACOBS, *Remarks on some recent extensions of Filippov's implicit functions lemma*, this Journal, 5 (1967), pp. 622–627.
- [16] ———, *An optimization problem for an nth-order scalar neutral functional differential equation with functional side conditions*, Delay and Functional Differential Equations and their Applications, Klaus Schmitt, ed., Academic Press, New York, 1972, pp. 345–352.
- [17] M. Q. JACOBS AND TI-JEUN KAO, *An optimum settling problem for time-lag systems*, J. Math. Anal. Appl., 40 (1972), pp. 687–707.
- [18] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.
- [19] L. M. KUPERMAN AND JU. M. REPIN, *On the question of controllability in infinite-dimensional space*, Dokl. Akad. Nauk SSSR, 200 (1971), pp. 767–769.
- [19a] S. KURCYUSZ, *A local maximum principle for operator constraints and its application to systems with time lags*, Control and Cybernetics, 2 (1973), no. 1.
- [20] C. E. LANGENHOP, *A reduction of λ -matrices*, Linear Algebra and Appl., to appear.
- [21] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
- [22] D. G. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.
- [23] A. MANITIUS AND A. OLBROT, *Controllability conditions for linear systems with delayed state and control*, Arch. Automat. i. Telemekh., 17 (1972), pp. 119–131.
- [24] R. K. MILLER, *Nonlinear Volterra Integral Equations*, Benjamin, Menlo Park, California, 1971.
- [25] S. A. MINJUK, *On complete controllability of linear systems with delay*, Differencial'nye Uravnenija, 8 (1972), pp. 254–259.
- [26] W. T. REID, *Some elementary properties of proper values and proper vectors of matrix functions*, SIAM J. Appl. Math., 18 (1970), pp. 259–266.
- [27] M. SCHECHTER, *Principles of Functional Analysis*, Academic Press, New York, 1971.
- [28] D. SHOWALTER, *Representation and computation of the pseudo-inverse*, Proc. Amer. Math. Soc., 18 (1967), pp. 584–586.
- [29] R. TRIGGIANI, *Controllability, observability and stabilizability of dynamical systems in Banach space with bounded operators*, Doctoral Dissertation, University of Minnesota, Minneapolis, 1972.
- [30] L. WEISS, *On the controllability of delay-differential systems*, this Journal, 5 (1967), pp. 575–587.
- [31] R. B. ZMOOD, *On Euclidean space and function space controllability of control systems with delay*, Doctoral Dissertation, University of Michigan, Ann Arbor, 1971.
- [32] R. B. ZMOOD AND N. H. MCCLAMROCH, *On the pointwise completeness of differential-difference equations*, J. Differential Equations, 12 (1972), pp. 474–486.

AN EFFICIENT ALGORITHM FOR THE SOLUTION OF THE WEBER PROBLEM WITH MIXED NORMS*

ALEJO PLANCHART† AND ARTHUR P. HURTER, JR.‡

Abstract. An algorithm is developed for the solution of the Weber problem when the distances are measured as arbitrary linear combinations of l_p -norms. Euclidean and rectilinear norms are used to illustrate the development. The algorithm applies the Dantzig-Wolfe decomposition method to the geometric programming dual of the Weber problem. The algorithm can be extended to the solution of the multifacilities location problem with constraints.

1. The problem. The problem can be formulated as

$$(1) \quad \varphi = \min \sum_{j=1}^n [w_j \|x - a_j\|_2 + w'_j \|x - a_j\|_1],$$

where w_j and w'_j are nonnegative scalars, x and a_j are k -dimensional vectors in the Euclidean space E^k , $\|\cdot\|_2$ is the Euclidean norm and $\|\cdot\|_1$ is the rectilinear norm.

In most economic applications, the location problem is restricted to E^2 . However, there are economically based problems where $k \geq 3$. For example, the location of equipment within a multistory building is a location problem in E^3 . The location or Weber problem format has potential application to decision problems in other fields as illustrated in [40]. Consequently, it is appropriate to develop techniques applicable to problems in E^k rather than restricting our attention to E^2 .

An efficient algorithm for the solution of (1) is developed. In addition to converging to the optimal solution, the algorithm facilitates sensitivity analysis on the parameters (w_j , w'_j and a_j).

Many of the results of this paper, including the algorithm, are valid for the following generalization of (1):

$$(1') \quad \varphi_1 = \min \sum_{p \in P_L} \sum_{j=1}^n w_{pj} \|x - a_j\|_p,$$

where

$$P_L = \{(p_1, \dots, p_L) | p_i \in P, i = 1, \dots, L, L < \infty\},$$

$$P = \{p | p = 2a/(2b + 1), a, b \text{ positive integers}\} \cup \{1\},$$

$w_{pj} > 0$ are known weights, and

$$\|\cdot\|_p \text{ is a general } l_p\text{-norm, i.e., } \|x\|_p = [|x_1|^p + \dots + |x_k|^p]^{1/p}.$$

* Received by the editors October 23, 1973, and in revised form July 26, 1974. This work and its extensions were supported in part by the National Science Foundation under Contract GK-38336.

† Instituto de Estudios Superiores de Administracion, Caracas, Venezuela.

‡ Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois 60201.

In order to keep the notation as simple as possible, problem (1) instead of (1') will be considered hereafter. The generalizations of the results to problem (1') are usually straightforward.

Problem (1) for the special case when $w'_j = 0$ for all j is the well-known generalized Weber problem of great importance in location problems. Similar heuristic solutions of this case were developed independently by Kuhn and Kuenne [24], Cooper [3], [4], Miehe [29] and Palermo [30]. A dynamic programming solution was developed by Bellman [1] and nonlinear programming algorithms by Vergin and Rogers [36] and Love [25], [26]. The last two papers refer to the more general multifacilities location problem in which more than one facility is to be located.

Problem (1) with $w_j = 0$ for all j was studied by Francis [12] and the multifacilities location problem for this case by Francis [13], Wesolowsky and Love [41], [42], [43] and by Cabot, Francis and Sary [2].

Besides being important on their own, the "single source algorithms" are used as one of the steps in the solution of more complex problems. For example, the "multisource Weber problem" can be solved by the "location-allocation heuristic" of Cooper [5], [7] and Kuenne and Soland [21] which requires the iterative solution of several independent "single source problems" and a linear program. Another recent approach to the solution of problem (1) is that of Eyster, White and Wierwille [11]. Some particular application algorithms have been developed that require the solution of this type of problem in one of the steps, as in Hurter, Schaefer and Wendell [19] and Schaefer and Hurter [32].

Problem (1) is an important problem which seems to have been ignored in the literature. It is important because it is not unusual for two kinds of transportation to be considered in the same problem. In this case, the rectilinear and Euclidean norms would be most likely to be applicable. For example, in a plant layout problem, it would not be unusual to find a machine being fed some of its materials through conveyors or pipes where the Euclidean distance is appropriate; while others are supplied or removed by fork lift trucks operating in aisles where the rectilinear distance is important.

An algorithm is developed which uses the geometric programming dual of problem (1) and then applies the Dantzig-Wolfe decomposition method. First some properties of problem (1) must be established. The following notation is used:

$$\begin{aligned} f_j^1(x) &\triangleq w'_j \|x - a_j\|_1; & f_j^2(x) &\triangleq w_j \|x - a_j\|_2; \\ f_j(x) &\triangleq f_j^1(x) + f_j^2(x); & F(x) &\triangleq \sum_{j=1}^n f_j(x); \\ J_1 &\triangleq \{j | w'_j > 0\}; & J_2 &\triangleq \{j | w_j > 0\}. \end{aligned}$$

THEOREM 1.1. *$F(x)$ is a convex function of x . Furthermore, if the points a_j , $j \in J_2$, are not collinear, and J_2 is not empty, then $F(x)$ is a strictly convex function of x .*

Proof. Both $f_j^1(x)$ and $f_j^2(x)$ are convex functions (since they are norms) and hence $f_j(x)$ and $F(x)$ are convex (since the sum of convex functions is convex). It

has been proved (Kuhn [22, p. 45]) that $\sum_{j \in J_2} f_j^2(x)$ is strictly convex when the points $a_j, j \in J_2$, are not collinear; therefore $\sum_{j \in J_2} f_j^2(x) + \sum_{j \in J_1} f_j^1(x)$ is strictly convex since the sum of convex functions, one of which is strictly convex, is strictly convex.

COROLLARY. *If the points $a_j, j \in J_2$, are not collinear and J_2 is not empty, then $F(x)$ has a unique minimum.*

Proof. It follows from strict convexity.

2. The geometric programming dual. Consider a convex program:

$$\varphi_1 = \inf_{x \in X \cap C} F(x),$$

where X is a vector subspace of E^k , C is the domain of F . Peterson defines the symmetric geometric programming dual and develops theorems that characterize the relationship between the primal and the dual in [31]. The dual is

$$\psi_1 = \inf_{y \in Y \cap D} H(y),$$

where Y is the orthogonal complement of X , D is the domain of $H(\cdot)$ and $H(\cdot)$ is the conjugate transform of $F(\cdot)$ defined as

$$H(y) = \sup_x \{ \langle y, x \rangle - F(x) \}.$$

Problem (1) can be formulated as

$$\varphi = \min \sum_{j=1}^n [f_j^2(x_j) + f_j^1(x_j^1)]$$

such that

$$x_j = x_j^1 = x \quad \text{for all } j.$$

Here X is characterized by the condition $x_j = x_j^1 = x$ for all j and C is E^{2nk} Euclidean space.

Wendell and Peterson [39] obtained the geometric programming dual of the multifacility location problem. Their results are applied to problem (1) and are the basis of the algorithm developed in § 3. We are preparing an algorithm for the multifacility location problem also based on the Wendell and Peterson results.

The geometric programming dual of problem (1):

$$(2) \quad \psi = \min \sum_{j=1}^n [\langle a_j, y_j \rangle + \langle a_j, y'_j \rangle]$$

such that

$$\|y_j\|_2 \leq w_j, \quad \|y'_j\|_\infty \leq w'_j, \\ \sum (y_j + y'_j) = 0,$$

where y_j and y'_j are k -dimensional vectors, $\|y'_j\|_\infty = \max \{|y'_{1j}|, \dots, |y'_{kj}|\}$, and $\|\cdot\|_2, \|\cdot\|_\infty$ are the polars of $\|\cdot\|_2$ and $\|\cdot\|_1$ of problem (1).

The dual for the special case of $w'_j = 0$ for all j was obtained by Kuhn [23] as a generalization of Fasbender's method of solving the three-point Weber

problem. Witzgall [44] generalized that dual for any arbitrary norm. Wendell [37] and Wendell and Peterson [39] generalized it further to include the multi-facility location problem and nonlinear functions of distance (nondecreasing and convex). Francis and Cabot [14] also obtained this dual for the Euclidean case via Sinha's duality [34], [35].

Problem (2) is a convex, well-behaved, and differentiable programming problem as shown in the following more convenient equivalent formulation:

$$\psi = \min H(y) = \min \sum_{j=1}^n [\langle a_j, y_j \rangle + \langle a_j, y'_j \rangle]$$

such that

$$(3.1) \quad (y_{1j})^2 + \cdots + (y_{kj})^2 \leq w_j^2 \quad \text{for all } j = 1, \dots, n,$$

$$(3.2) \quad -w'_j \leq y'_{ij} \leq w'_j \quad \text{for all } i = 1, \dots, k, \quad j = 1, \dots, n$$

$$(3.3) \quad \sum_{j=1}^n (y_j + y'_j) = 0 \quad \text{for all } j = 1, \dots, n.$$

The following properties, developed by Wendell and Peterson [39], will be used in § 3.

Property 1. If x and y are feasible solutions to dual problems (1) and (3), then $F(x) + H(y) \geq 0$ with equality iff $x \in \partial H(y)$ or $y \in \partial F(x)$, in which case x and y are optimal solutions to problems (1) and (3). The notation $x \in \partial H(y)$ means x is contained in the subgradient set of $H(y)$.

Property 2. Both problem (1) and problem (3) have nonempty optimal solution sets and there is no duality gap, so $\varphi + \psi = 0$.

These properties will be used in § 3 where an algorithm for the solution of problem (3) is developed.

3. The algorithm. The Dantzig-Wolfe decomposition method is employed to solve problem (3). Then duality theory is used to obtain the optimal solution of problem (1).

The first step in the algorithm is to "inner-linearize" (Geoffrion [15]) the sets described by (3.1). Define C_j as:

$$C_j \triangleq \{y_j | (y_{1j})^2 + \cdots + (y_{kj})^2 \leq w_j^2\}.$$

Let y_j^τ for $\tau = 1, \dots, t_j$ be known elements on the boundary of C_j . t_j is an arbitrary integer which indicates the number of elements known on C_j . Consider the set D_j as an approximation to the set C_j where

$$D_j \triangleq \left\{ y_j \left| y_j = \sum_{\tau=1}^{t_j} \lambda_j^\tau y_j^\tau; \sum_{\tau=1}^{t_j} \lambda_j^\tau = 1, \lambda_j^\tau \geq 0 \right. \right\}.$$

In other words, the set of points D_j consists of convex linear combinations of some of the boundary points of C_j . Notice that there are as many sets C_j as there are points a_j in (1), since there is a constraint (3.1) for each point ($j = 1, \dots, n$). Therefore, we can state the following lemma.

LEMMA 3.1.

$$D_j \subseteq C_j.$$

Proof. D_j is the set of all the convex combinations of some points of C_j , and hence it is contained in it.

The algorithm involves the iterative solution of a master problem which is equivalent to problem (3); but with the sets D_j instead of the sets C_j and of n nonlinear programming subproblems which will be described later. At step "s", the master problem (which is linear) and its linear programming dual may be written:

MP(s):

$$(4) \quad \chi(s) = \min \sum_{j=1}^n \sum_{\tau=1}^{t_j(s)} \lambda_j^\tau(s) \langle a_j, y_j^\tau \rangle + \sum_{\tau=1}^{t_j(s)} \langle a_j, y_j^\tau(s) \rangle$$

such that

$$(4.1) \quad -w'_j \leq y'_j(s) \leq w'_j, \quad j = 1, \dots, n,$$

$$(4.2) \quad \sum_{j=1}^n \sum_{\tau=1}^{t_j(s)} \lambda_j^\tau(s) y_j^\tau + \sum_{j=1}^n y'_j(s) = 0,$$

$$(4.3) \quad \sum_{\tau=1}^{t_j(s)} \lambda_j^\tau(s) = 1, \quad j = 1, \dots, n,$$

$$\lambda_j^\tau(s) \geq 0.$$

DP(s):

$$(5) \quad D(s) = \max \sum_{j=1}^n \delta_j(s) - \sum_{j=1}^n w'_j \left[\sum_{i=1}^k \theta_j^i(s) + \eta_j^i(s) \right]$$

such that

$$(5.1) \quad \delta_j(s) + \langle \alpha(s) - a_j, y_j^\tau \rangle \leq 0, \quad j = 1, \dots, n, \quad \tau = 1, \dots, t_j(s)$$

$$(5.2) \quad -\eta_j(s) + \theta_j(s) + \alpha(s) = a_j, \quad j = 1, \dots, n,$$

$$\eta_j(s), \theta_j(s) \geq 0.$$

Note that problem (4) has one column for each point taken on the relative boundary of the sets C_j , so that, if all the points on the relative boundary of the sets C_j 's are taken, problem (4) is a semi-infinite program (infinite columns).

The idea of the algorithm is to solve MP(s) determining values for the $y'_j(s)$ and $\lambda_j^\tau(s)$, for a finite number of columns. Then, the solution of the n subproblems shows whether or not there exist some points on C_j such that the columns associated with them would have had a negative reduced cost if they had been considered available when solving MP(s). In that case, they would have been pivoted into the basis, yielding a reduction in the value of the objective of MP(s). If such points exist, a new master problem MP(s + 1) is formed by attaching to MP(s) the column which has the most negative reduced costs on each of the C_j 's (and hence will reduce the objective function the most); and the procedure is repeated until no further reduction is possible or some stopping criterion is satisfied.

The reduced cost of the columns corresponding to any of the boundary points of C_j not included in MP(s) is given by $\langle a_j - \alpha(s), y_j \rangle - \delta_j(s)$. Thus, to find the most negative in each set, it suffices to find the y_j 's which solve the following subproblems:

SP_j:

$$(6) \quad \min_{y_j \in C_j} \langle a_j - \alpha(s), y_j \rangle.$$

The optimizing vector for (6) can be expressed explicitly as

$$y_j^* = -w_j(a_j - \alpha(s))/\|a_j - \alpha(s)\|_2$$

and the optimal value of the objective function as

$$\langle a_j - \alpha(s), y_j^* \rangle = -w_j\|a_j - \alpha(s)\|_2 = -f_j^2(\alpha(s)),$$

so that the column associated with y_j^* will be attached to MP(s) to form MP(s + 1) only if $-f_j^2(\alpha(s)) < \delta_j(s)$.

Notice that the expression for y_j^* is not valid if $\alpha(s) = a_j$. However, in that case, $\langle \alpha(s) - a_j, y_j^* \rangle = 0$ for all $y_j^* \in C_j$. Therefore, the cutting plane generated is $\delta_j(s) \leq 0$. When this occurs, the structure of the problem guarantees that no other y_j^* will be generated from the set C_j during successive iterations while $\alpha(s) = a_j$. The test problem presented below demonstrates the algorithm's ability to deal successfully with this case.

The solution of MP(s) is efficiently obtained using the Dantzig and Van Slyke [9] generalized upper bounding (G.U.B.) technique. An inverse of dimension $k \times k$ (2×2 for E^2 problems) is all that is needed to solve (4).

LEMMA 3.2.

$$\chi(s) \geq \psi.$$

Proof. Lemma 3.1 shows that the feasible set of problem (4) is no larger than the feasible set for problem (3). In all other respects, problems (3) and (4) are equivalent. Thus, the optimal value of the objective of (4), χ must be at least as large as that of (3), ψ .

LEMMA 3.3.

$$\chi(s + 1) < \chi(s).$$

Proof. The assertion follows since MP(s + 1) has as the starting basis the optimal basis for MP(s) and it has some columns with strictly negative reduced cost any of which, once pivoted into the basis, will never increase the value of the objective.

LEMMA 3.4. *If master program (4) is feasible, then it is bounded.*

Proof. Program (4) is equivalent to the minimization of a convex program (linear) on a closed and convex and hence compact set, namely the intersection of a finite number of closed and convex sets, the D_j 's. The assertion of the lemma follows immediately.

THEOREM 3.1.

$$\chi(s) = \psi \quad \text{if } f_j^2(\alpha(s)) \leq -\delta_j(s) \quad \text{for all } j.$$

Proof. Assume the opposite, that is, $\chi(s) > \psi$ (Lemma 3.2) and $f_j^2(\alpha(s)) \leq -\delta_j(s)$, and prove contradiction. Let y_j^* be the solution to problem (3). Obviously $y_j^* \notin D_j$ since otherwise it would have been the solution to (4). Now $f_j^2(\alpha(s)) \leq -\delta_j(s)$ implies that $\langle a_j - \alpha(s), y_j \rangle - \delta_j \geq 0$ for all $y_j \in C_j$ since $-f_j^2(\alpha(s))$ is the minimum value over $\langle a_j - \alpha(s), y_j \rangle$. But attaching to problem MP(s), the column corresponding to y_j^* to form problem MP(s + 1), y_j^* must be

the solution of the latter, which implies that $\chi(s+1) = \psi < \chi(s)$, but this implies in turn that the reduced cost of the column corresponding to y_j^* must have been negative; that is, $\langle a_j - \alpha(s), y_j^* \rangle - \delta_j(s) < 0$ and in particular $f_j^2(\alpha(s)) > -\delta_j(s)$, which contradicts the assertion.

THEOREM 3.2. *The optimal solutions to (4) and its dual provide the following bounds to the optimal solution φ of problem (1):*

$$-\sum_{j=1}^n \delta_j(s) + \sum_{j=1}^n w'_j \left[\sum_{i=1}^n \eta_j^i(s) + \theta_j^i(s) \right] \leq \varphi \leq \sum_{j=1}^n f_j^2(\alpha(s)) + f_j^1(\alpha(s)).$$

Proof. $\varphi \leq \sum_{j=1}^n f_j^2(\alpha(s)) + f_j^1(\alpha(s))$ follows directly since $\alpha(s)$ is a feasible point for problem (1). The rest of the inequality follows from the linear programming dual theorem and Lemma 3.2. Then

$$-\sum_{j=1}^n \delta_j(s) + \sum_{j=1}^n w'_j \left[\sum_{i=1}^k \eta_j^i(s) + \theta_j^i(s) \right] = -D(s) = -\chi(s) \leq -\psi.$$

The geometric dual theorem insures that $\psi + \varphi = 0$ from which Theorem 3.2 follows.

THEOREM 3.3. *If an optimal solution to (4) and its dual (5) is such that*

$$f_j^2(\alpha(s)) \leq -\delta_j(s)$$

for all j 's, then:

$$-\sum_{j=1}^n \delta_j(s) + \sum_{j=1}^n w'_j \left[\sum_{i=1}^k \eta_j^i(s) + \theta_j^i(s) \right] = \varphi = \sum_{j=1}^n f_j^2(\alpha(s)) + f_j^1(\alpha(s)).$$

Proof. Theorem 3.1 and geometric programming duality guarantee that

$$-\sum_{j=1}^n \delta_j(s) + \sum_{j=1}^n w'_j \left[\sum_{i=1}^k \eta_j^i(s) + \theta_j^i(s) \right] = -\chi(s) = -\psi = \varphi,$$

which establishes the left-hand equality. From the slackness conditions, both $\eta_j^i(s)$ and $\theta_j^i(s)$ cannot be simultaneously strictly greater than zero, so that one of the next three alternatives must hold.

(a) $\eta_j^i(s) = \theta_j^i(s) = 0$, but from (5.2) this implies that $\alpha_i(s) = a_{ij}$.

(b) $\eta_j^i(s) > 0$, $\theta_j^i(s) = 0$, which implies $\eta_j^i(s) = \alpha_i(s) - a_{ij}$.

(c) $\eta_j^i(s) = 0$, $\theta_j^i(s) > 0$, which implies $\theta_j^i(s) = a_{ij} - \alpha_i(s)$.

(a), (b) and (c) can be summarized by:

$$|\eta_j^i(s) + \theta_j^i(s)| = |\alpha_i(s) - a_{ij}|$$

and by the triangle inequality and positiveness of η_j^1 and θ_j^1 :

$$|\eta_j^i(s) + \theta_j^i(s)| \leq |\eta_j^i(s)| + |\theta_j^i(s)| = \eta_j^i(s) + \theta_j^i(s),$$

so that

$$\eta_j^i(s) + \theta_j^i(s) \geq |\alpha_i(s) - a_{ij}|.$$

By summing over i , multiplying by w'_j and then adding over j we obtain

$$\sum_{j=1}^n w'_j \left[\sum_{i=1}^k \eta_j^i(s) + \theta_j^i(s) \right] \geq \sum_{j=1}^n f_j^1(\alpha(s)).$$

Finally by summing the conditions $-\delta_j \geq f_j^2(\alpha(s))$ over all j 's and adding to the previous inequality we obtain

$$-\sum_{j=1}^n \delta_j(s) + \sum_{j=1}^n w'_j \left[\sum_{i=1}^k \eta_j^i(s) + \theta_j^i(s) \right] \geq \sum_{j=1}^n f_j^2(\alpha(s)) + f_j^1(\alpha(s)).$$

But Theorem 3.2 guarantees that

$$-\sum_{j=1}^n \delta_j(s) + \sum_{j=1}^n w'_j \left[\sum_{i=1}^k \eta_j^i(s) + \theta_j^i(s) \right] \leq \sum_{j=1}^n f_j^2(\alpha(s)) + f_j^1(\alpha(s)),$$

from which it follows that only an equality can hold.

COROLLARY. At an optimum $f_j^2(\alpha(s)) = -\delta_j(s)$ and

$$f_j^1(\alpha(s)) = w'_j \left[\sum_{i=1}^k \eta_j^i(s) + \theta_j^i(s) \right].$$

Proof. It is obvious from the proof of the theorem.

Theorem 3.3 proves the very important result that the dual variables associated with the master problem constraints (4.2) provide the optimal solution for the original problem (1) when the optimality conditions of Theorem 3.1 are satisfied. Notice that by taking the dual of the "linearized and restricted" (Geoffrion [15]) version of the geometric programming dual of problem (1), the optimal solution of the latter was obtained directly in the linear programming tableau. Furthermore, the dual variables to problem (1), i.e., the y 's are also obtained almost directly from the final tableau. Notice that y_j is the gradient of $f_j^2(x)$, that is, $y_{ij} = \partial f_j^2(x)/\partial x_i$ evaluated at the optimizing point (Francis and Cabot [14]), so that $\sum_{j=1}^n y_{ij} = \partial F(x)/\partial x_i$ and condition (4.2) corresponds to the Kuhn-Tucker first order conditions. The y 's and the δ 's at optimality provide an interesting sensitivity analysis which will be explored in a forthcoming paper.

4. Convergence properties. The algorithm proposed in this paper is equivalent to the Dantzig-Wolfe decomposition method (dual cutting plane algorithm), the convergence of which has been proved by Dantzig [8] and Zangwill [45] for the case in which only inequality constraints are present.

Greenberg and Robbins [18] extended the proofs for the following case.

$$P^0: \quad \max f(y): \quad y \in Y, \quad g(y) \leq 0, \quad h(y) = 0;$$

where $f: Y \rightarrow E^1$, $Y \subseteq E^n$, $g: Y \rightarrow E^m$, $h: Y \rightarrow E^k$, under the following assumptions:

A1: P^0 is feasible;

A2: Y is compact;

A3: f , g and h are continuous on Y ;

A4: There exist $m + 2k$ points $y_0^1, \dots, y_0^m, y_+^1, \dots, y_+^k, y_-^1, \dots, y_-^k$ in Y (not necessarily distinct) such that the following are true:

- (i) $g(y_0^i) \leq 0$, $h(y_0^i) = 0$ and $g_i(y_0^i) < 0$ for $i = 1, \dots, m$,
- (ii) $h_r(y_-^i) = h_r(y_+^i) = 0$ for $i \neq r$,
- (iii) $h_r(y_-^r) < 0$ and $h_r(y_+^r) > 0$.

Assumptions A1–A4 are trivially satisfied by problem (3) when the following notation is used:

$$(a) \quad y \triangleq (y'_1, \dots, y'_n, y_1, \dots, y_n) \in E^{2kn}.$$

$$(b) \quad Y = \{y | -w'_j \leq y'_j \leq w'_j, y_{1j}^2 + \dots + y_{kj}^2 \leq w_j^2\}.$$

$$(c) \quad f(y) = - \sum_{j=1}^n \langle a_j, y_j + y'_j \rangle.$$

$$(d) \quad g_i(y) = (-1)^{i+1} y'_{rq} - w'_q \quad \text{for } i = 1, \dots, 2nk,$$

where $q = \{i/2k\}$ and $r = \{(i - qk)/2\}$ and $\{x\}$ represents the smallest integer greater than or equal to x .

$$(e) \quad h_r(y) = \sum_{j=1}^n y'_{rj} + y_{rj} \quad \text{for } r = 1, \dots, k.$$

$$(f) \quad y_0^r = 0 \quad \text{for } r = 1, \dots, 2nk.$$

$$(g) \quad y_+^r = ({}_+y'_1, \dots, {}_+y'_n, {}_+y_1, \dots, {}_+y_n) \quad \text{such that } {}_+y_j^r = 0 \quad \text{for all } r = 1, \dots, k.$$

$$\text{and} \quad {}_+y_{ij}^r = \begin{cases} y_{ij}^* & \text{for } i = r, \\ 0 & \text{otherwise;} \end{cases}$$

$$y_{ij}^* = +w_j |a_{ij} - \alpha_i(-1)| / \|a_j - \alpha(-1)\|,$$

where $\alpha(-1)$ is an arbitrary k -dimensional vector chosen so that $a_j \neq \alpha(-1)$ for all j .

$$(h) \quad y_-^r = ({}_-y'_1, \dots, {}_-y'_n, {}_-y_1, \dots, {}_-y_n)$$

such that

$${}_-y_j^r = 0 \quad \text{for all } r = 1, \dots, k,$$

$${}_-y = \begin{cases} -y_{ij}^* & \text{for } i = r, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore the algorithm proposed below will converge to the optimal solution of problem (3).

5. Algorithm and computational experience.

Step 0. Take $s = 0$.

Step 1. Choose $\alpha(-1)$ and $x(-1)$ to be the center of gravity of the points $a_j, j \in J_2$, or any arbitrary point such that $\alpha(-1) \neq a_j$ for all $j = 1, \dots, n$.

Step 2. Form master problem (4) MP(0) by taking $2kn + 2k$ points such that A4 is satisfied.

Step 3. Solve MP(s) and get $\delta_j(s), \eta_j^i(s), \theta_j^i(s), y_j'(s)y_j(s)$ and $\alpha(s)$.

Step 4. If $F(\alpha(s)) < F(\alpha(s-1))$, take $x(s) = \alpha(s)$. Otherwise $x(s) = \alpha(s-1)$.

Step 5. Compute $\text{DIF} = F(x(s)) + \chi(s)$ and $\text{ACC} = \text{DIF}/\chi(s)$.

Step 6. If $\text{DIF} = 0$ or ACC less than a specified small number, stop. Otherwise move to Step 7.

Step 7. If $f_j^2(\alpha(s)) \leq -\delta_j(s)$ for all j , stop; otherwise proceed to Step 8.

Step 8. For those j such that $f_j^2(\alpha(s)) > -\delta_j(s)$, add to MP(s) the column defined by

$$y_j^r = -w_j(a_j - \alpha(s))/\|a_j - \alpha(s)\|$$

and make $t_j(s+1) = t_j(s) + 1$.

Step 9. Make $s = s + 1$ and go back to Step 3.

The algorithm was programmed in FORTRAN IV for a CDC 6400. The linear programming code used in Step 5 is the subroutine for experimental optimization (SEXOP) developed by Professor Roy E. Marsten based on the Graves' primal-dual approach to linear programming [17]. It was observed in all the test problems that in no case did a column, once it became nonbasic, re-enter the basis. This suggested that elimination of nonbasic columns would not curtail convergence. The algorithm was then reformulated so that nonbasic variables, except those on the initial set (to satisfy A4), were dropped. Using the work of Eaves and Zangwill [10, p. 534], it was proved that the conditions for dropping cutting planes are satisfied except in the case that in a given iteration (s) the current basis is degenerate and the objective function value does not improve after the optimal solution to MP(s) is obtained. A modification of the algorithm to take care of degeneracy is being developed. The results reported here were obtained using a routine which drops nonbasic columns.

Three sets of problems of different sizes were created in the following way: the first set consists of 5 problems with from 10 to 50 cities (in intervals of 10) in which only the Euclidean distance is considered. The starting point was taken at the origin ($\alpha(-1) = 0$). The second set consists of 8 problems, with from 10 to 80 cities (in intervals of 10) in which only the Euclidean distance is considered and the starting point is taken at the center of gravity $\alpha(-1) = \sum_{j=1}^n w_j a_j / \sum_{j=1}^n w_j$. The first 5 problems of this set are identical to the problems of the previous set. The third set consist of 5 problems with from 10 to 50 cities (in intervals of 10) with both Euclidean and rectilinear distances considered. The starting point is at the origin for problems with 10, 20 and 40 cities and at the center of gravity $\alpha(-1) = \sum_{j \in J_2} w_j a_j / \sum_{j \in J_2} w_j$ for problems with 30 and 50 cities.

All the problems attempted in the three sets were solved to the desired degree of accuracy (ACC), the speed of convergence being very satisfactory. No consistent difference in the convergence rates between sets (1) and (2) was observed, which suggests only weak dependence of the convergence rate on the starting point. Tables 1 through 3 summarize the results corresponding to the three sets of problems solved. These tables show the number of iterations (I) and total computer processing time (T) needed to achieve certain degrees of accuracy (ACC).

Accuracy is defined as in [19], that is,

$$ACC = \left| \frac{A - B}{B} \right| \times 100\%.$$

where A is the algorithm result and B is the lower bound. Table 4 presents some of the results required in the standardized reporting of algorithm performance proposed by Ignizio [20].

TABLE 1
Euclidean distances
 $\alpha(-1) = 0$

ACC	>5%			1%-5%			0.1%-1%			<0.1%			Final accuracy				Comments
	I	T	ACC	I	T	ACC	I	T	ACC	I	T	ACC	I	T	ACC	N	
No. of cities																	
10	3	0.65	6.60	4	0.80	2.48	5	1.06	0.30	7	1.44	0.02	9	1.59	0.005	100	
20	3	1.08	12.17										4	1.60	0.000	74	Exact
30	2	1.70	7.46	3	2.36	1.40	4	3.53	0.33	5	4.31	0.09	6	6.07	0.001	201	
40	2	2.28	35.56	3	4.16	2.77	5	5.93	0.26	6	6.99	0.09	9	8.96	0.003	271	
50	3	2.35	8.73	4	4.57	2.41	5	6.20	0.71	7	11.63	0.02	8	13.00	0.008	276	

I = No. of iterations. T = Time in sec. (CP). ACC = Accuracy. N = No. of pivots.

TABLE 2
Euclidean distances
 $\alpha(-1) = \sum_{j=1}^n w_j a_j / \sum_{j=1}^n w_j$

ACC	>5%			1%-5%			0.1%-1%			<0.1%			Final accuracy			Comments
	I	T	ACC	I	T	ACC	I	T	ACC	I	T	ACC	I	T	ACC	
No. of cities																
10	2	0.29	7.12	3	0.43	2.86	4	0.57	0.89	6	0.97	0.03	7	1.22	0.001	70
20	2	0.58	5.98	3	1.12	4.69	4	1.26	0.11	5	1.54	0.04	6	1.93	0.000	77 exact
30	2	2.02	7.10	3	3.14	1.07	4	3.79	0.21	5	4.95	0.02	6	5.43	0.009	172
40	2	2.62	16.50				3	3.84	0.53	4	6.04	0.06	7	8.39	0.004	208
50	2	4.20	14.35	3	6.22	2.51	4	8.84	0.56	7	11.29	0.07	9	15.56	0.004	316
60	1	4.80	21.07	2	8.29	2.40	3	12.82	0.57	5	19.62	0.01	6	22.42	0.007	290
70	2	5.83	20.70	3	9.52	4.10	4	12.51	0.96	7	23.40	0.02	9	29.15	0.002	468
80	1	2.10	42.43	2	6.28	4.20	3	13.08	0.52	5	25.85	0.04	7	33.26	0.002	421

I = No. of iterations. T = Time in sec. (CP). ACC = Accuracy. N = No. of pivots.

TABLE 3
Euclidean and rectilinear distances

ACC	>5%			1%-5%			0.1%-1%			<0.1%			Final accuracy			Comments
	I	T	ACC	I	T	ACC	I	T	ACC	I	T	ACC	I	T	ACC	
No. of cities																
10	1	0.43	20.14				2	0.78	0.80	4	1.16	0.05	5	1.30	0.006	50 †
20	1	1.41	21.92										2	2.31	0.000	exact†
30	1	2.89	12.39				2	5.14	0.86	3	6.63	0.08	3	8.60	0.000	exact‡
40	1	4.90	29.68	2	8.76	1.08	3	11.72	0.17	4	13.14	0.04	6	16.45	0.002	†
50	1	5.05	15.79	2	12.06	2.55				3	18.34	0.08	6	27.21	0.003	‡

I = iterations. T = Time in sec. (CP). ACC = Accuracy. N = No. of pivots.

† $a(-1) = 0$. ‡ $a(-1) = \sum_{j \in J_2} w_j a_j / \sum_{j \in J_2} w_j$.

TABLE 4

Computer Utilized: CDC 6400.
 Programming language: FORTRAN IV.
 Amount of Internal Storage Used: Words 43,300_g.
 Amount of External Storage Used = none.
 Expression of memory requirements as a function of problem size:

$$9250 + \text{NW} [8 + 8k + 2k^2] + \text{NWW} [1 + 13k + 4k^2] \\ + \text{NDP} [3 + k] + 19k + k^2,$$

where

NDP = n° of demand points,

NW = demand points $a_j, j \in J_2$,

NWW = demand points $a_j, j \in J_1$,

$k = n^\circ$ of coordinates.

The computer code was intended only as a first attempt and is, therefore, a very unsophisticated and imperfect representation of the algorithm. This makes the results obtained even more encouraging. Perhaps one could improve the speed of convergence if the starting point were selected at the center of gravity to the power n as suggested in [16] and [33]. We wish to thank a referee for bringing these references to our attention.

6. Conclusions and extensions. The speed of convergence suggests that this algorithm can be used to solve multisource Weber problems with unknown weights by location-allocation procedures [5], [7], [20] when it is known that a given percentage of the demand must be shipped through a "rectilinear distance path". The algorithm is currently being extended to include constraints and also to solve the multifacility location problem with different norms and linear constraints. Most of the theory developed here will be applicable in these cases in a straightforward way.

Again consider DP(s). This suggests that the primal problem (1) could be re-expressed as:

$$\begin{aligned} \text{minimize } \sum_j \left[q_j + w'_j \sum_i (\theta_j^i + \eta_j^i) \right] \quad \text{over all } q, \theta, \eta \text{ and } x \text{ in } E^k \text{ satisfying} \\ q_j \geq w_j \|x - a_j\|_2, \quad j = 1, \dots, n, \\ x - \eta_j + \theta_j = a_j, \quad \eta_j \geq 0, \quad \theta_j \geq 0, \quad j = 1, \dots, n, \\ q_j \leq w_j \|x^0 - a_j\|_2, \quad j = 1, \dots, n, \end{aligned}$$

where x^0 is an arbitrary point, so that the set of feasible q_j is compact. Then the re-expressed primal problem can be solved directly by the cutting plane algorithm. We are grateful to the editor, Professor Rockafellar, for pointing out this equivalence.

Since work on this paper was completed and during its reviewing process, papers using a similar approach have appeared and the interested reader is

advised to compare our results with those of Love and Kraemer [27] and Love [28]. We are grateful to an unknown referee for bringing these papers to our attention.

REFERENCES

- [1] R. BELLMAN, *An application of dynamic programming to location allocation problems*, SIAM Rev., 7 (1965), pp. 126–128.
- [2] V. CABOT, R. FRANCIS AND M. STARY, *A network flow solution to a rectilinear distance facility location problem*, AIIE Trans., 2 (1970), pp. 132–141.
- [3] LEON COOPER, *Location-allocation problems*, Operations Res., 11 (1963), pp. 331–344.
- [4] ———, *Heuristic methods for location-allocation problems*, SIAM Rev., 6 (1964), pp. 37–53.
- [5] ———, *Solutions of generalized locational equilibrium models*, J. Regional Sci., 7 (1967), pp. 1–18.
- [6] ———, *An extension of the generalized Weber problem*, Ibid., 8 (1968), pp. 181–197.
- [7] ———, *The transportation location problem*, Operations Res., 20 (1972), pp. 94–109.
- [8] G. B. DANTZIG, *Linear Programming and Extensions*, Princeton University Press, Princeton, N.J., 1963.
- [9] G. DANTZIG AND R. M. VAN SLYKE, *Generalized upper bounding techniques*, J. Comput. System Sci., 1 (1967), pp. 213–226.
- [10] B. C. EAVES AND W. I. ZANGWILL, *Generalized cutting plane algorithms*, this Journal, 9 (1971), pp. 529–542.
- [11] J. EYSTER, WHITE AND WIERWILLE, *On solving multifacility location problems using a hyperboloid Approximation procedure*, AIIE Trans., 5 (1973), pp. 1–6.
- [12] R. L. FRANCIS, *A note on the optimum location of new machines in existing plant layouts*, J. Industrial Engrg., 14 (1963), pp. 57–59.
- [13] ———, *On the location of multiple new facilities with respect to existing facilities*, Ibid., 15 (1964).
- [14] R. L. FRANCIS AND V. CABOT, *Properties of a multifacility location problem involving Euclidean distances*, Presented at ORSA meeting, New Orleans, 1972; Naval Res. Logist. Quart., 19 (1972), p. 335.
- [15] A. M. GEOFFRION, *Elements of large scale mathematical programming*, Parts I & II, Management Sci., 16 (1970), pp. 652–675.
- [16] L. A. GOLDSTONE, *A further note on warehouse location*, Ibid., 14 (1968), pp. 132–133.
- [17] G. N. GRAVES, *A complete constructive algorithm for the general mixed linear programming problem*, Naval Res. Logist. Quart., 12 (1965), pp. 1–35.
- [18] H. J. GREENBERG AND T. C. ROBBINS, *Finding Everett's Lagrange multipliers by generalized linear programming*, Tech. Rep. CP-70008, Computer Science/Operations Research Center, Southern Methodist Univ., Dallas, Tex., 1970. Revised 1972.
- [19] A. P. HURTER, M. SCHAEFER AND R. WENDELL, *Solutions of constrained location problems*, Management Sci., to appear.
- [20] J. P. IGNIZIO, *On the establishment of standards for comparing algorithm performances*, Interfaces, 2 (1971), pp. 8–12.
- [21] R. E. KUENNE AND R. M. SOLAND, *The multisource Weber problem*, Economic papers, Institute of Defense Analysis, 1971.
- [22] H. W. KUHN, *Location problems and mathematical programming*, Separatum Colloquium on the Application of Mathematics to Economics (Budapest, 1963). Publishing house of the Hungarian Academy of Sciences, Budapest, 1965, pp. 235–242.
- [23] ———, *On a pair of dual non-linear programs*, Non-linear Programming, J. Abadie, ed., John Wiley, New York, 1967, Chap. 3, pp. 37–54.
- [24] H. W. KUHN AND R. KUENNE, *An efficient algorithm for the numerical solution of the generalized Weber problem in spatial economics*, J. Regional Sci., 4 (1962), pp. 21–33.
- [25] R. F. LOVE, *A non-linear programming approach to facilities location problems*, J. Canad. Operations. Res. Soc., 5 (1967), pp. 136–143.
- [26] ———, *Locating facilities in three-dimensional space by convex programming*, Naval Res. Logist. Quart., 16 (1969), pp. 503–516.
- [27] R. F. LOVE AND SVEND KRAEMER, *A dual decomposition method for minimizing transportation costs in multifacility location problems*, Transportation Sci., 7 (1973).

- [28] R. F. LOVE, *The dual of a hyperbolic approximation to the generalized constrained multi-facility location problem with l_p distances*, Management Sci., 21 (1974), pp. 22–34.
- [29] W. MIEHLE, *Link-length minimization in networks*, Operations Res., 6 (1958), pp. 232–243.
- [30] F. P. PALERMO, *A network minimization problem*, IBM J. Res. Develop., 5 (1961), pp. 335–337.
- [31] E. PETERSON, *Symmetric duality for generalized unconstrained geometric programming*, SIAM J. Appl. Math., 19 (1970), pp. 487–526.
- [32] M. SCHAEFER AND A. HURTER, *The regional allocation of fire resources: A damage minimizing approach*, Urban Analysis, to appear.
- [33] A. SCHARLIG, *An application of operational research: The optimal location of an enterprise*, Swiss Rev. of Political Economy and Statistics, 3 (1971), pp. 599–611.
- [34] S. M. SINHA, *An extension of a theorem on supports of a convex function*, Management Sci., 12 (1966), pp. 380–385.
- [35] ———, *A duality theorem for nonlinear programming*, Management Sci., 12 (1966), pp. 385–391.
- [36] R. C. VERGIN AND J. D. ROGERS, *An algorithm and computational procedure for locating economic facilities*, Management Sci., 13 (1967), pp. B240–B255.
- [37] RICHARD WENDELL, *Some aspects on the theory of location*, Doctoral dissertation, Northwestern Univ., Evanston, Ill., 1971.
- [38] RICHARD WENDELL AND ARTHUR HURTER, *Location theory-dominance and convexity*, Operations Res., 21 (1973), pp. 314–321.
- [39] RICHARD WENDELL AND E. PETERSON, *Duality in generalized location problems*, Working paper, Dept. of Industrial Engineering and Management Sciences, Northwestern Univ., Evanston, Ill., 1971.
- [40] RICHARD E. WENDELL AND STUART J. THORSON, *Some generalizations of social decisions under majority rule*, Econometrica, to appear.
- [41] G. O. WESOLOWSKY AND R. F. LOVE, *The optimal location of new facilities using rectangular distances*, Operations Res., 19 (1971), pp. 124–130.
- [42] ———, *Location of facilities with rectangular distances among point and area destinations*, Naval Res. Logist. Quart., 18 (1971), pp. 83–91.
- [43] ———, *A nonlinear approximation method for solving a generalized rectangular distance Weber problem*, Management Sci., 18 (1972), pp. 656–664.
- [44] C. WITZGALL, *Optimal location of a central facility: Mathematical models and concepts*, Rep. 8388, National Bureau of Standards, Washington, D.C., 1965.
- [45] W. I. ZANGWILL, *Nonlinear Programming: A Unified Approach*, Prentice-Hall, Englewood Cliffs, N.J., 1969.

CONVEXITY OF THE RANGE OF CERTAIN INTEGRALS*

LAMBERTO CESARI†

Abstract. In the line of Lyapunov's convexity theorem, the author considers the subsets H and K obtained from any number h of given integrable n -vector functions by all possible set decompositions and by all possible convex combinations respectively and integration. Both sets are closed and convex, and the author gives a simple proof that $H = K$.

We prove here briefly and by a simple elementary approach a number of statements related to Lyapunov's theorem: the usual dichotomy process (§1), the Lyapunov theorem (Theorem 1), and the statements (Theorems 3, 4, 5 in §2) concerning the sets $H, K \subset E^n$ obtained from any number h of given integrable n -vector functions by all possible "set decompositions" and all possible "convex combinations" respectively and integration. Both sets H and K are closed and convex, and we give here a simple proof that $H = K$. We have made use of the identity $H = K$ in a related paper on existence theorems without convexity conditions [6]. The same identity has been used also by M. B. Suryanarayana [22, 23] and T. S. Angell [1]. Acknowledgment is made in the reference list to earlier works in the past decades, which have influenced the present writing.

1. Range of integrals of vector-valued functions. If $[a, b]$ is any given interval of length $l = b - a$, and $\alpha, 0 \leq \alpha \leq 1$, any number, then the point $t = a + \alpha(b - a) = a + \alpha l$ divides $[a, b]$ into two parts of measures αl and $(1 - \alpha)l$. If we divide $[a, b]$ into two equal parts, and we divide each part as above, the corresponding set

$$D_\alpha^2 = \left[a \leq t < a + \frac{\alpha l}{2} \right] \cup \left[a + \frac{l}{2} \leq t < a + \frac{l}{2} + \frac{\alpha l}{2} \right]$$

has measure αl and is the union of 2^k disjoint intervals. Also, for $0 \leq \alpha < \alpha' \leq 1$ $[a, b]$ into 2^k equal parts, and in each part we take corresponding subintervals, then the set

$$(1.1) \quad D_\alpha^k = \bigcup_{i=1}^{2^k} [a + 2^{-k}(i-1)l, a + 2^{-k}(i-1+\alpha)l]$$

has measure αl and is the union of 2^k disjoint intervals. Also, for $0 \leq \alpha < \alpha' \leq 1$ and the same k we have $D_\alpha^k \subset D_{\alpha'}^k$ and $\text{meas}(D_{\alpha'}^k - D_\alpha^k) = (\alpha' - \alpha)l$.

LEMMA 1. *Given any vector function $f(t) = (f_1, \dots, f_n)$, $a \leq t \leq b$, whose components are L -integrable in $[a, b]$, and any $\varepsilon > 0$, there is an integer K such that for all $k \geq K$ and $\alpha, 0 \leq \alpha \leq 1$, we have*

$$\left| \int_{D_\alpha^k} f(t) dt - \alpha \int_a^b f(t) dt \right| \leq \varepsilon.$$

In other words, if $a_0 = (a_1, \dots, a_n)$ denotes the integral of $f(t)$ on $[a, b]$, then the integral on D_α^k , thought of as a function of $\alpha, 0 \leq \alpha \leq 1$, is uniformly approximated by the linear function $a_0\alpha, 0 \leq \alpha \leq 1$.

* Received by the editors September 27, 1972.

† Department of Mathematics, University of Michigan, Ann Arbor, Michigan 48104. This research was supported in part by US-AFOSR Research Project 71-2122 at the University of Michigan.

Proof. It is not restrictive to assume $a = 0, b = 1$. We know that there is a continuous vector function $g(t), 0 \leq t \leq 1$, such that

$$\int_0^1 |f(t) - g(t)| dt \leq \varepsilon/4.$$

Then $g(t)$ is uniformly continuous in $[0, 1]$, and hence there is $\delta > 0$ such that $t, t' \in [0, 1], |t - t'| \leq \delta$ implies $|g(t) - g(t')| \leq \varepsilon/4$. Let K be the smallest integer with $1/2^K < \delta$. For any $k \geq K$ let $g_k(t), 0 \leq t \leq 1$, be the step function defined by $g_k(t) = g(t_{i-1})$ for all $t_{i-1} \leq t \leq t_i, i = 1, \dots, 2^k$, where $t_i = i/2^k$. Then $|g(t) - g_k(t)| \leq \varepsilon/4$ for all $0 \leq t \leq 1$. Thus

$$\int_0^1 |f(t) - g_k(t)| dt \leq \int_0^1 |f - g| dt + \int_0^1 |g - g_k| dt \leq \frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \frac{\varepsilon}{2},$$

and

$$\begin{aligned} \Delta &= \left| \int_{D_\alpha^k} f(t) dt - \alpha \int_0^1 f(t) dt \right| \leq \left| \int_{D_\alpha^k} f dt - \int_{D_\alpha^k} g_k dt \right| \\ &+ \left| \int_{D_\alpha^k} g_k dt - \alpha \int_0^1 g_k dt \right| + \left| \alpha \int_0^1 g_k dt - \alpha \int_0^1 f dt \right| = s_1 + s_2 + s_3. \end{aligned}$$

Here

$$\begin{aligned} s_2 &= \left| \sum_i g_k(t_i) \left(\frac{\alpha}{2^k} \right) - \alpha \sum_i g_k(t_i) \left(\frac{1}{2^k} \right) \right| = 0, \\ s_1 &\leq \int_{D_\alpha^k} |f - g_k| dt \leq \int_0^1 |f - g_k| dt \leq \frac{\varepsilon}{2}, \\ s_3 &\leq \int_0^1 |g_k - f| dt \leq \frac{\varepsilon}{2}, \end{aligned}$$

and finally $\Delta \leq \varepsilon/2 + 0 + \varepsilon/2 = \varepsilon$. Lemma 1 is thereby proved.

Lemma 1 has a stronger form which of course is less easy to prove.

LEMMA 2. *Given any vector function $f(t) = (f_1, \dots, f_n)$, $a \leq t \leq b$, whose components are L -integrable in $[a, b]$, then for every $\alpha, 0 \leq \alpha \leq 1$, there is a measurable subset E_α of $[a, b]$ such that*

$$(1.2) \quad \int_{E_\alpha} f(t) dt = \alpha \int_a^b f(t) dt, \quad 0 \leq \alpha \leq 1.$$

In other words, if $a_0 = (a_1, \dots, a_n)$ is the integral of f on $[a, b]$, then the first member of (1.2), thought of as a function of α , is a linear function of α , namely, $a_0 \alpha, 0 \leq \alpha \leq 1$.

This statement is a particular case of the following one which we shall prove below.

LEMMA 3. *Given any vector function $f(t) = (f_1, \dots, f_n)$, $a \leq t \leq b$, whose components are L -integrable functions in $[a, b]$, and any measurable subset A of $[a, b]$, then for every $\alpha, 0 \leq \alpha \leq 1$, there is a measurable subset $B_\alpha, B_\alpha \subset A \subset [a, b]$, with*

$$(1.3) \quad \int_{B_\alpha} f(t) dt = \alpha \int_A f(t) dt, \quad 0 \leq \alpha \leq 1.$$

Proof. The proof of Lemma 3 is made up of parts. (a) Let us prove Lemma 3 for $n = 1$ and f a nonnegative scalar function. If $\varphi(t)$ denotes the characteristic function of A , say $\varphi = 1$ on A and $\varphi = 0$ otherwise, then $f(t)\varphi(t)$ is L-integrable in $[a, b]$, and hence $F(t) = \int_a^t f(\tau)\varphi(\tau) d\tau$, $a \leq t \leq b$, is a continuous function taking all values from $F(a) = 0$ to $F(b)$. Thus, there is some c , $a \leq c \leq b$, with $F(c) = \alpha F(b)$, and, if $B_\alpha = [a, c] \cap A$, also

$$\int_{B_\alpha} f dt = \int_a^c f \varphi dt = F(c) = \alpha F(b) = \alpha \int_a^b f \varphi dt = \alpha \int_A f dt.$$

Thus, Lemma 3 is proved for $n = 1$ and f scalar nonnegative.

(b) Let us assume that we know how to determine $B_{1/2}$ for every A and a given vector function $f = (f_1, \dots, f_n)$ whose components are nonnegative L-integrable, and let us prove that we can determine all sets B_α , $0 \leq \alpha \leq 1$, and that we can determine them in such a way that $\alpha < \alpha'$ implies $B_\alpha \subset B_{\alpha'}$.

For the sake of simplicity we shall use the notation

$$(1.4) \quad \mu(E) = \int_E f dt, \quad \mu_i(E) = \int_E f_i dt, \quad i = 1, \dots, n.$$

First, for $B'_{1/2} = A - B_{1/2}$, we have

$$\mu(B'_{1/2}) = \mu(A) - \mu(B_{1/2}) = \mu(A) - (1/2)\mu(A) = (1/2)\mu(A).$$

Then let us determine sets $B_{1/4} \subset B_{1/2}$, $B'_{3/4} \subset B'_{1/2}$ such that

$$\mu(B_{1/4}) = (1/2)\mu(B_{1/2}), \quad \mu(B'_{3/4}) = (1/2)\mu(B'_{1/2})$$

and then, for $B_{3/4} = B_{1/2} \cup B'_{3/4}$, also

$$\mu(B_{3/4}) = (3/4)\mu(A), \quad \mu(B_{1/4}) = (1/4)\mu(A),$$

and if $B_0 = \emptyset$, $B_1 = A$, we have $B_0 \subset B_{1/4} \subset B_{1/2} \subset B_{3/4} \subset B_1$. By repeating this process we obtain sets $B_{i/2^r}$, $i = 0, 1, \dots, 2^r$, $r = 1, 2, \dots$, so that $\mu(B_{i/2^r}) = (i/2^r)\mu(A)$, and for $i < j$, $\lambda = i/2^r$, $\lambda' = j/2^r$, also $B_\lambda \subset B_{\lambda'}$. Now let α be any number $0 < \alpha < 1$, and let $[\lambda_s]$, $[\lambda'_s]$ be sequences of numbers $\lambda_s = i/2^s$, $\lambda'_s = j/2^s$, such that $\lambda_s \leq \lambda_{s+1} < \alpha < \lambda'_{s+1} \leq \lambda'_s$, $\lambda_s \rightarrow \alpha$, $\lambda'_s \rightarrow \alpha$. For

$$B_\alpha = \bigcup B_{\lambda_s} \quad B'_\alpha = \bigcap B_{\lambda'_s},$$

we have $B_\alpha \subset B'_\alpha$ and

$$\lambda_s \int_A f dt = \int_{B_{\lambda_s}} f dt \leq \int_{B_\alpha} f dt \leq \int_{B'_\alpha} f dt \leq \int_{B_{\lambda'_s}} f dt = \lambda'_s \int_A f dt,$$

where \leq means that such a relation holds for each component. As $s \rightarrow \infty$ we obtain

$$\alpha \int_A f dt = \int_{B_\alpha} f dt = \int_{B'_\alpha} f dt, \quad 0 \leq \alpha \leq 1.$$

Finally, $\alpha < \alpha'$ implies $B_\alpha \subset B_{\alpha'}$ as it is clear from the definition of these sets. This proves (b).

Given any two sets E, F we shall write as usual $E \Delta F = (E - F) \cup (F - E)$.

(c) Assume that Lemma 3 has been proved for some vector function $f = (f_1, \dots, f_n)$ whose components are nonnegative L -integrable. Let E, F be any two measurable subsets of $A \subset [a, b]$. Then for every $\alpha, 0 \leq \alpha \leq 1$, there is some subset $C(\alpha)$ of $E \cup F$ with $C(0) = E, C(1) = F$, such that

$$(1.5) \quad \begin{aligned} \int_{C(\alpha)} f \, dt &= (1 - \alpha) \int_E f \, dt + \alpha \int_F f \, dt, \quad 0 \leq \alpha \leq 1, \\ \int_{C(\alpha) \Delta C(\alpha')} f_i \, dt &\leq |\alpha - \alpha'| \int_{E \Delta F} f_i \, dt, \\ 0 &\leq \alpha, \quad \alpha' \leq 1, \quad i = 1, \dots, n. \end{aligned}$$

Indeed, let us apply Lemma 3 to the sets $E - F$ and $F - E$ and the number α . Let $B_\alpha \subset E - F, B'_\alpha \subset F - E$ be the corresponding sets and take

$$C(\alpha) = (E \cap F) \cup (E - F - B_\alpha) \cup B'_\alpha.$$

Then

$$\begin{aligned} \mu(C(\alpha)) &= \mu(E \cap F) + \mu(E - F) - \mu(B_\alpha) + \mu(B'_\alpha) \\ &= \mu(E \cap F) + \mu(E - F) - \alpha\mu(E - F) + \alpha\mu(F - E) \\ &= \mu(E \cap F) + (1 - \alpha)\mu(E - F) + \alpha\mu(F - E) \\ &= (1 - \alpha)\mu[(E \cap F) \cup (E - F)] + \alpha\mu[(E \cap F) \cup (F - E)] \\ &= (1 - \alpha)\mu(E) + \alpha\mu(F). \end{aligned}$$

In addition, for each component f_i and $0 \leq \alpha \leq \alpha' \leq 1$, we have

$$\begin{aligned} C(\alpha) &= (E \cap F) \cup (E - F - B_\alpha) \cup B'_\alpha, \\ C(\alpha') &= (E \cap F) \cup (E - F - B_{\alpha'}) \cup B'_{\alpha'}, \\ C(\alpha) \Delta C(\alpha') &= [C(\alpha) - C(\alpha')] \cup [C(\alpha') - C(\alpha)] \\ &= (B_{\alpha'} - B_\alpha) \cup (B'_{\alpha'} - B'_\alpha), \end{aligned}$$

and finally

$$\begin{aligned} \mu_i[C(\alpha) \Delta C(\alpha')] &= \mu_i(B_{\alpha'} - B_\alpha) + \mu_i(B'_{\alpha'} - B'_\alpha) \\ &= \alpha' \mu_i(E - F) - \alpha \mu_i(E - F) + \alpha' \mu_i(F - E) - \alpha \mu_i(F - E) \\ &\leq |\alpha - \alpha'| \mu_i(E \Delta F). \end{aligned}$$

Thus (c) is proved.

(d) Lemma 3 has been proved for $n = 1$ and f scalar nonnegative. Assume that Lemma 3 has been proved for $n - 1$ and vectors f with nonnegative components, and let us prove it for n . Let \tilde{f} be the $(n - 1)$ -vector $\tilde{f} = (f_1, \dots, f_{n-2}, \tilde{f}_{n-1})$ with $\tilde{f}_{n-1} = f_{n-1} + f_n$, and let $\bar{\mu}, \mu_i, \bar{\mu}_i$ be the set functions defined by (1.4) with f replaced by $\tilde{f}, f_i, \tilde{f}_i$. First by Lemma 3 with $\alpha = 1/2$ applied to \tilde{f} there is a subdivision of A into two parts E, F , with $E \cap F = \emptyset, E \cup F = A$, and

$$(1.6) \quad \bar{\mu}(E) = \bar{\mu}(F) = (1/2)\bar{\mu}(A).$$

Also, by force of (c), there are sets $C(\alpha) \subset E \cup F = A$, $0 \leq \alpha \leq 1$, with $C(0) = E$, $C(1) = F$, and

$$(1.7) \quad \begin{aligned} \bar{\mu}(C(\alpha)) &= (1 - \alpha)\bar{\mu}(E) + \alpha\bar{\mu}(F) = (1 - \alpha)(1/2)\bar{\mu}(A) + \alpha(1/2)\bar{\mu}(A) \\ &= (1/2)\bar{\mu}(A), \quad 0 \leq \alpha \leq 1. \end{aligned}$$

Let us prove that $\mu_{n-1}(C(\alpha))$ is a continuous function of α in $[0, 1]$. Indeed, $\mu_{n-1}(C(\alpha))$ is a scalar, namely the integral of $f_{n-1} \geq 0$ over $C(\alpha)$, and

$$\begin{aligned} &|\mu_{n-1}(C(\alpha)) - \mu_{n-1}(C(\alpha'))| \\ &\leq \mu_{n-1}[C(\alpha) - C(\alpha')] + \mu_{n-1}[C(\alpha') - C(\alpha)] \\ &\leq \bar{\mu}_{n-1}[C(\alpha) - C(\alpha')] + \bar{\mu}_{n-1}[C(\alpha') - C(\alpha)] \\ &\leq |\alpha - \alpha'|(\bar{\mu}_{n-1}(E - F) + \bar{\mu}_{n-1}(F - E)) = |\alpha - \alpha'| \int_A (f_{n-1} + f_n) dt. \end{aligned}$$

This proves that $\mu_{n-1}(C(\alpha))$ is a continuous function of α for $0 \leq \alpha \leq 1$. On the other hand, $\mu_{n-1}(C(0)) = \mu_{n-1}(E)$, $\mu_{n-1}(C(1)) = \mu_{n-1}(F)$. Since E and F are complementary in A then $\mu_{n-1}(E) \leq (1/2)\mu_{n-1}(A)$ according as $\mu_{n-1}(F) \geq (1/2)\mu_{n-1}(A)$. Thus, as α describes $[0, 1]$, $\mu_{n-1}(C(\alpha))$ describes an interval which contains $(1/2)\mu_{n-1}(A)$. We conclude that there is some α , $0 \leq \alpha \leq 1$, such that $\mu_{n-1}(C(\alpha)) = (1/2)\mu_{n-1}(A)$. For this particular value of α , we have from (1.7),

$$\begin{aligned} \int_{C(\alpha)} f_i dt &= \frac{1}{2} \int_A f_i dt, & i = 1, \dots, n-2, \\ \int_{C(\alpha)} (f_{n-1} + f_n) dt &= \frac{1}{2} \int_A (f_{n-1} + f_n) dt, \\ \int_{C(\alpha)} f_{n-1} dt &= \frac{1}{2} \int_A f_{n-1} dt, \end{aligned}$$

and hence, by difference, also

$$\int_{C(\alpha)} f_n dt = \frac{1}{2} \int_A f_n dt,$$

or

$$(1.8) \quad \int_{C(\alpha)} f dt = \frac{1}{2} \int_A f dt.$$

We have proved that for the n -vector $f = (f_1, \dots, f_n)$ we can determine a subset $B_{1/2} = C(\alpha) \subset A$ satisfying (1.8), where A is any measurable subset of $[a, b]$. Thus, by (b), we can determine analogous sets B_α for all α , $0 \leq \alpha \leq 1$, and Lemma 3 is proved for vector-valued functions with nonnegative components.

(e) We have now to prove Lemma 3 for vector functions $f = (f_1, \dots, f_n)$ with L -integrable components of arbitrary signs. For every $i = 1, \dots, n$, and $j = 1, 2$, we consider the sets $A_{i1} \subset A$ where $f_i \geq 0$ and $A_{i2} \subset A$ where $f_i < 0$. We divide A into 2^n disjoint measurable sets $A_r = A_{1j_1} \cap A_{2j_2} \cap \dots \cap A_{nj_n}$, where r denotes any one of the 2^n systems (j_1, j_2, \dots, j_n) of indices 1 and 2. On each set A_r the

components f_i have constant signs, and there are, therefore, sets $B_{r\alpha} \subset A_r$ with $\int_{B_{r\alpha}} f \, dt = \alpha \int_{A_r} f \, dt$, $0 \leq \alpha \leq 1$. The sets $B_\alpha = \bigcup_r B_{r\alpha}$ then satisfy the requirements of Lemma 3. Lemma 3 is thereby proved.

COROLLARY. *Given any vector function $f(t) = (f_1, \dots, f_n)$, $a \leq t \leq b$, whose components are L -integrable in $[a, b]$, and any two fixed measurable sets $E, F \subset [a, b]$, then for every α , $0 \leq \alpha \leq 1$, there is some set $C(\alpha) \subset E \cup F$, with $C(0) = E$, $C(1) = F$, and*

$$\int_{C(\alpha)} f \, dt = (1 - \alpha) \int_E f \, dt + \alpha \int_F f \, dt.$$

This statement is a consequence of part (c) of the proof of Lemma 3.

THEOREM 1. (Lyapunov). *Given any vector function $f(t) = (f_1, \dots, f_n)$, $a \leq t \leq b$, whose components are L -integrable, and any measurable subset A of $[a, b]$, then*

$$(1.9) \quad \mu(E) = \int_E f(t) \, dt$$

describes a convex set H as E describes all possible measurable subsets E of A (in other words, the range of $\mu(E)$ is convex).

Proof. If $\mu_1, \mu_2 \in H$, then there are measurable sets E_1, E_2 in A such that $\mu_i = \mu(E_i) = \int_{E_i} f \, dt$, $i = 1, 2$. Among all measurable subsets of A there certainly are the sets $C(\alpha)$, $0 \leq \alpha \leq 1$, defined in the Corollary above. Then

$$\mu(C(\alpha)) = (1 - \alpha) \int_{E_1} f \, dt + \alpha \int_{E_2} f \, dt = (1 - \alpha)\mu_1 + \alpha\mu_2,$$

that is, all points of the segment $(1 - \alpha)\mu_1 + \alpha\mu_2$, $0 \leq \alpha \leq 1$, belong to H , and H is proved to be convex.

2. Convex combinations. Closure of the range.

THEOREM 2. *Given any two vector functions $f(t) = (f_1, \dots, f_n)$, $g(t) = (g_1, \dots, g_n)$, $a \leq t \leq b$, whose components are L -integrable, let A be any measurable subset of $[a, b]$, let E denote any measurable subset of A , and $h_E(t)$, $t \in A$, the function $h_E(t) = f(t)$ for $t \in E$, $h_E(t) = g(t)$ for $t \in F = A - E$. Then*

$$(2.1) \quad \mu(E) = \int_A h_E(t) \, dt = \int_E f(t) \, dt + \int_F g(t) \, dt$$

describes a convex subset H of the space E_n as E describes all measurable subsets of A .

Proof. For every E as above and $F = A - E$, we have

$$\mu(E) = \int_A h_E \, dt = \int_E f \, dt + \int_F g \, dt = \int_E (f - g) \, dt + \int_A g \, dt.$$

If μ_0 is the fixed value of the last integral, and we apply Theorem 1 to the function $f - g$, we see that the set H of Theorem 2 is simply a translation of the convex set H of Theorem 1 relative to $f - g$.

THEOREM 3. *Given any number of vector functions $f^{(j)}(t) = (f_1^{(j)}, \dots, f_n^{(j)})$, $a \leq t \leq b$, $j = 1, \dots, h$, whose components are all real-valued and L -integrable, and any measurable subset A of $[a, b]$, then*

$$(2.2) \quad \mu(E_1, \dots, E_h) = \int_{E_1} f^{(1)} dt + \dots + \int_{E_h} f^{(h)} dt$$

describes a convex set H of E_n when E_1, \dots, E_h describe all possible decompositions of A into disjoint measurable subsets E_j or A , $j = 1, \dots, h$.

Proof. The statement has just been proved for $h = 2$. Let us prove it for $h > 2$. Let μ_1, μ_2 be any two points of the set H described by (2.2). Then there are two decompositions of A such that $\mu(E_1, \dots, E_h) = \mu_1$, $\mu(F_1, \dots, F_h) = \mu_2$ with $E_i \cup E_j = \emptyset$, $F_i \cup F_j = \emptyset$, $i \neq j$, $i, j = 1, \dots, h$, and $\bigcup_i E_i = A$, $\bigcup_i F_i = A$. Let α be any number $0 \leq \alpha \leq 1$, and let us prove that there is a decomposition of A such that $\mu(G_{1\alpha}, \dots, G_{h\alpha}) = (1 - \alpha)\mu_1 + \alpha\mu_2$. First, let us consider the decomposition of A into the h^2 measurable disjoint sets $A_{ij} = E_i \cap F_j$, $i, j = 1, \dots, h$. For every pair i, j with $i \neq j$ we shall now apply Lemma 3 to the $2n$ -vector $(f^{(i)}, f^{(j)})$ to obtain a decomposition of each set A_{ij} into sets H'_{ij}, H''_{ij} , such that $A_{ij} = H'_{ij} \cup H''_{ij}$, $H'_{ij} \cap H''_{ij} = \emptyset$, and

$$\begin{aligned} \int_{H'_{ij}} f^{(i)} dt &= \alpha \int_{A_{ij}} f^{(i)} dt, & \int_{H'_{ij}} f^{(j)} dt &= \alpha \int_{A_{ij}} f^{(j)} dt, \\ \int_{H''_{ij}} f^{(i)} dt + \int_{H''_{ij}} f^{(j)} dt &= (1 - \alpha) \int_{A_{ij}} f^{(i)} dt + \alpha \int_{A_{ij}} f^{(j)} dt. \end{aligned}$$

If we now take

$$G_{i\alpha} = A_{ii} \cup \left(\bigcup_{s \neq i} H'_{is} \right) \cup \left(\bigcup_{s \neq i} H''_{si} \right), \quad i = 1, \dots, h,$$

then the h sets $G_{i\alpha}$ form a decomposition of A into h measurable disjoint sets, and the identities hold

$$\begin{aligned} E_i &= A_{ii} \cup \left(\bigcup_{s \neq i} A_{is} \right), & i &= 1, \dots, h, \\ F_j &= A_{jj} \cup \left(\bigcup_{s \neq j} A_{sj} \right), & j &= 1, \dots, h. \end{aligned}$$

We now have

$$\begin{aligned} \sum_i \int_{G_{i\alpha}} f^{(i)} dt &= \sum_i \int_{A_{ii}} f^{(i)} dt + \sum_{i \neq j} \sum \int_{H'_{ij}} f^{(i)} dt + \sum_{i \neq j} \sum \int_{H''_{ij}} f^{(j)} dt \\ &= (1 - \alpha) \left[\sum_i \int_{A_{ii}} f^{(i)} dt + \sum_{i \neq j} \sum \int_{A_{ij}} f^{(i)} dt \right] \\ &\quad + \alpha \left[\sum_j \int_{A_{jj}} f^{(j)} dt + \sum_{i \neq j} \sum \int_{A_{ij}} f^{(j)} dt \right] \\ &= (1 - \alpha) \sum_i \int_{E_i} f^{(i)} dt + \alpha \sum_j \int_{F_j} f^{(j)} dt. \end{aligned}$$

Theorem 3 is thereby proved.

THEOREM 4. *Given any set of vector functions $f^{(j)}(t) = (f_1^{(j)}, \dots, f_n^{(j)})$, $a \leq t \leq b$, $j = 1, \dots, h$, whose components are all real-valued and L -integrable, and*

any fixed measurable set A of $[a, b]$, then

$$v(p_1, \dots, p_h) = \sum_{j=1}^h \int_A p_j(t) f^{(j)}(t) dt$$

describes a convex set K of E_n when $p_1(t), \dots, p_h(t)$, $t \in A$, describe all possible sets of scalar measurable functions $p_j(t) \geq 0$, $j = 1, \dots, h$, with $p_1(t) + \dots + p_h(t) = 1$, $t \in A$.

Proof. Let v_1, v_2 be any two points of K . Then there are two systems of weight functions $p_j(t) \geq 0$, $q_j(t) \geq 0$, $\sum_j p_j = 1$, $\sum_j q_j = 1$, such that

$$v_1 = \sum_{j=1}^h \int_A p_j(t) f^{(j)}(t) dt, \quad v_2 = \sum_{j=1}^h \int_A q_j(t) f^{(j)}(t) dt.$$

If α is any number $0 \leq \alpha \leq 1$, and we take

$$r_j(t) = (1 - \alpha)p_j(t) + \alpha q_j(t), \quad t \in A, \quad j = 1, \dots, h,$$

then $r_j(t) \geq 0$, $j = 1, \dots, h$, $\sum_j r_j(t) = 1$, $t \in A$, and

$$v = \sum_{j=1}^h \int_A r_j(t) f^{(j)}(t) dt = (1 - \alpha)v_1 + \alpha v_2.$$

This proves that K is convex.

THEOREM 5. *Under the same hypotheses of Theorems 3 and 4, the sets H and K are identical and $H = K$ is a compact convex set.*

Proof. First, let us prove that $H \subset K$. Indeed, every point $\mu \in H$, or

$$\mu = \sum_{j=1}^h \int_{E_j} f^{(j)}(t) dt, \quad E_i \cap E_j = \emptyset, \quad i \neq j, \quad \bigcup_{j=1}^h E_j = A,$$

can be written in the form

$$\mu = \sum_{j=1}^h \int_A p_j(t) f^{(j)}(t) dt,$$

by taking $p_j(t) = 1$ for $t \in E_j$, $p_s(t) = 0$ for $t \in E_j$, $s \neq j$, and where $j = 1, \dots, h$. Thus, $\mu \in K$ and we have proved that $H \subset K$.

Let us remark that the set K is certainly bounded since for every point $v \in K$ we have

$$|v| = \left| \sum_{j=1}^h \int_A p_j f^{(j)} dt \right| \leq \sum_{j=1}^h \int_A |f^{(j)}(t)| dt.$$

Also, the set K is compact. It is enough to prove that K is closed. Let $\bar{v} \in \text{cl } K$. Then there is a sequence v_r , $r = 1, 2, \dots$, of points $v_r \in K$ with $v_r \rightarrow \bar{v}$ as $r \rightarrow \infty$ and

$$(2.3) \quad v_r = \sum_{j=1}^h \int_A p_{rj}(t) f^{(j)}(t) dt, \quad r = 1, 2, \dots,$$

with $0 \leq p_{rj}(t) \leq 1$, $\sum_j p_{rj}(t) = 1$, $r = 1, 2, \dots$. Therefore, the functions $p_{rj}(t) \in L_\infty(A)$, and are equibounded; in particular, the elements of each of the h sequences $[p_{rj}(t), r = 1, 2, \dots]$, $j = 1, \dots, h$, are equibounded in the bounded

measurable subset A of $[a, b]$. By known properties of weak compactness of $L_\infty(A)$ we conclude that there is a subsequence, say still $[r]$ for the sake of simplicity, such that $p_{rj}(t) \rightarrow p_j(t)$ as $r \rightarrow \tilde{M}$ weakly in $L_\infty(A)$ toward measurable functions $p_j(t)$, and then $0 \leq p_j(t) \leq 1$, $\sum_j p_j(t) = 1$ (a.e. in A). From (2.3) we deduce

$$\bar{v} = \sum_{j=1}^h \int_A p_j(t) f^{(j)}(t) dt,$$

since the functions $f^{(j)}$ are fixed and L_1 -integrable in A . Thus, $\bar{v} \in K$, and K is closed and therefore compact.

Let us denote now by $x = (x^1, \dots, x^n)$ the points of the space E^n , and let γ be the infimum of the values of the x^1 -coordinate of the points $x = (x^1, \dots, x^n) \in K$. Then, γ is finite, and we shall denote by K_γ the set of all points of K (if any) with $x^1 = \gamma$. Let us prove that $H \cap K_\gamma$ is not empty, and then K_γ is certainly not empty. To this purpose, let us note that for any $x = (x^1, \dots, x^n) \in K$ we have

$$x^i = \sum_{j=1}^h \int_A p_j(t) f_i^{(j)}(t) dt, \quad i = 1, \dots, n,$$

where all $f_i^{(j)}$ are scalar. Let us denote by E_j the set of all points $t \in A$ with $f_1^{(j)}(t) \leq f_1^{(s)}(t)$ for all $s \neq j$, $s = 1, \dots, h$. The sets E_j may not be disjoint, but the sets $E'_1 = E_1, E'_2 = E_2 - E'_1, \dots, E'_h = E_h - (E'_1 \cup \dots \cup E'_{h-1})$ are disjoint and their union is still A . Now we have obviously

$$(2.4) \quad \gamma = (x^1)_{\min} = \sum_{j=1}^h \int_{E_j} f_1^{(j)}(t) dt, \quad x^s = \sum_{j=1}^h \int_{E_j} f_s^{(j)}(t) dt, \quad s = 2, \dots, n.$$

If we take $p_j(t) = 1$ for $t \in E_j$, $p_s(t) = 0$ for $t \in E_j$, $s \neq j$, where $j = 1, \dots, h$, then we have also

$$(2.5) \quad \gamma = (x^1)_{\min} = \sum_{j=1}^h \int_A p_j(t) f_1^{(j)}(t) dt,$$

$$x^s = \sum_{j=1}^h \int_A p_j(t) f_s^{(j)}(t) dt, \quad s = 1, \dots, n.$$

This proves that there is at least one point $x \in H \cap K_\gamma$.

The argument above shows also that, for $n = 1$, that is, all $f^{(j)}$ scalar, then H and K are convex sets on the x^1 -axis, $H \subset K$, and K has a minimum element γ which is also an element of H . In the same way we can prove that K has a maximum element γ' which is also an element of H , and then $H = K = [\gamma, \gamma']$ is the segment from γ to γ' on the x^1 -axis. We have proved Theorem 5 for $n = 1$.

Let us assume we have proved Theorem 5 for $1, 2, \dots, n-1$, and let us prove it for n .

Above we have proved that $H \subset K$ and that $H \cap K_\gamma$ is not empty, $H \cap K_\gamma \subset K_\gamma$. By using the induction hypothesis we shall prove first that $H \cap K_\gamma = K_\gamma$. Note that if the sets E_j defined above are disjoint, and even if they are not disjoint but all intersections $E_i \cap E_j$, $1 \leq i < j \leq h$, have measure zero, then the representations (2.4) and (2.5) of the points $x = (\gamma, x^2, \dots, x^n) \in K_\gamma$ are unique (up to

sets of measure zero). In other words, the set K_γ is made up of a single point x_0 , and then certainly $H \cap K_\gamma = K_\gamma = \{x_0\}$.

If some of the sets $E_i \cap E_j$ have positive measure, then we consider all possible systems $\omega = (j_1, \dots, j_k)$ of distinct integers $1 \leq j_1 < j_2 < \dots < j_k \leq h$, $2 \leq k \leq h$. Let $\omega' = (1, 2, \dots, h) - \omega$ denote the complementary set, say $\omega' = (j'_1, \dots, j'_{h-k})$. For $\omega = (j_1, \dots, j_k)$ let E_ω be the set of all $t \in A$ where

$$f_1^{(j_1)}(t) = \dots = f_1^{(j_k)}(t) < f_1^{(s)}(t) \quad \text{for all } s \in \omega'.$$

If E denotes the union of all sets $E_i \cap E_j$, $1 \leq i < j \leq h$, then E has now a decomposition into the finitely many disjoint measurable sets E_ω . Note that, for every $t \in E_0 = A - E$, the point t belongs to one and only one set E_j ; hence $f_1^{(j)}(t) < f_1^{(s)}(t)$ for $t \in E_0 \cap E_j$ and all $s \neq j$, $s = 1, \dots, h$. In other words, $E_0 = A - E$ has the decomposition $[E_0 \cap E_j, j = 1, \dots, h]$ into h disjoint measurable subsets. Thus, the sets $E_0 \cap E_j$, $j = 1, \dots, h$, and all sets E_ω form a decomposition of A into disjoint measurable subsets. We shall denote by $p_j(t)$, $t \in A$, $j = 1, \dots, h$, any system of weight functions $p_j(t) \geq 0$, $j = 1, \dots, h$, $\sum_{j=1}^h p_j(t) = 1$, $t \in A$, defined as follows: (a) for $t \in E_0 \cap E_j$ we take $p_j(t) = 1$ and $p_s(t) = 0$ for $s \neq j$, and here j ranges over all $j = 1, \dots, h$; (b) for $t \in E_\omega$, $\omega = (j_1, \dots, j_k)$, and hence $\omega' = (j'_1, \dots, j'_{h-k})$; we require $\sum_{j \in \omega} p_j(t) = 1$ and hence $p_s(t) = 0$ for all $s \in \omega'$. Note that for every $\omega = (j_1, \dots, j_k)$ we have $f_1^{(j_1)}(t) = \dots = f_1^{(j_k)}(t)$, $t \in E_\omega$, and we denote by $f_1^{(\omega)}(t)$ the common scalar value of these functions.

Now any point $x = (\gamma, x^2, \dots, x^n) \in K_\gamma$ has the following representation:

$$(2.6) \quad \begin{aligned} \gamma &= (x^1)_{\min} = \sum_{j=1}^h \int_{E_0 \cap E_j} f_1^{(j)}(t) dt + \sum_{\omega} \int_{E_\omega} f_1^{(\omega)}(t) dt, \\ x^s &= \sum_{j=1}^h \int_{E_0 \cap E_j} f_s^{(j)}(t) dt + \sum_{\omega} \sum_{\sigma=1}^k \int_{E_\omega} p_{j_\sigma}(t) f_s^{(j_\sigma)}(t) dt, \end{aligned}$$

where $s = 2, \dots, h$, and where the last summation ranges over all elements j_1, \dots, j_k of the system ω . For each system ω and set E_ω , the values

$$(2.7) \quad \sum_{\sigma=1}^k \int_{E_\omega} p_{j_\sigma}(t) f_s^{(j_\sigma)}(t) dt, \quad s = 2, \dots, n,$$

represent the coordinates $y = (x^2, \dots, x^n)$ of a point of a set K' in E^{n-1} . By the induction hypothesis this set coincides with the corresponding set, say H' , each point of which can be written in the form

$$\sum_{\sigma=1}^k \int_{F_\sigma^\omega} f_s^{(j_\sigma)}(t) dt, \quad s = 2, \dots, n,$$

for a suitable decomposition of E_ω into disjoint measurable sets $F_1^\omega, \dots, F_k^\omega$. Thus, from (2.6), we conclude that for every $x \in K_\gamma$ we have

$$x = \sum_{j=1}^h \int_{E_0 \cap E_j} f^{(j)}(t) dt + \sum_{\omega} \sum_{\sigma=1}^k \int_{F_\sigma^\omega} f^{(j_\sigma)}(t) dt.$$

This proves that $K_\gamma = H \cap K_\gamma$ as stated.

Now let \bar{x} be any point of the boundary of the convex compact set K , hence $\bar{x} \in \text{bd}(\text{cl } K)$, where $\text{bd}(\text{cl } K) = \text{bd } K \subset K$. There is some hyperplane $bx - c = 0$, such that $b\bar{x} - c = 0$, and $bx - c \geq 0$ for all $x \in K$. If (ξ^1, \dots, ξ^n) is a new system of coordinates in E_n such that $\xi^1 = bx$, then $\xi^1 = \gamma = c$ is the infimum of the values of the coordinate ξ^1 of the points $\xi = (\xi^1, \dots, \xi^n) \in K$. The same argument above shows that $H \cap K_\gamma = K_\gamma$. In particular $\bar{x} \in K_\gamma = H \cap K_\gamma$ and $\bar{x} \in H$.

We have proved that all boundary points of K are in H . Since any compact convex set is the convex hull of its boundary points, we conclude that $H = K$ and that this set is convex and compact.

REFERENCES

- [1] T. S. ANGELL, *Existence of optimal control without convexity and a bang-bang theorem for Volterra equations*, J. Optimization Theory Appl., to appear.
- [2] D. BLACKWELL, *The range of certain vector integrals*, Proc. Amer. Math. Soc., 2 (1951), pp. 390–395.
- [3] C. CASTAING, *Quelques problèmes de mesurabilité liés à la théorie de la commande*, C. R. Acad. Sci. Paris Ser. A, 262 (1966), pp. 409–411.
- [4] ———, *Sur les équations différentielles multivoques*, Ibid., 263 (1966), pp. 63–66.
- [5] ———, *Sur une nouvelle extension du théorème de Lyapunov*, Ibid., 264 (1967), pp. 333–336.
- [6] ———, *Sur un théorème de représentation intégrale lié à la comparaison des mesures*, Ibid., 264 (1967), pp. 1059–1062.
- [7] L. CESARI, *An existence theorem without convexity conditions*, this Journal, 12 (1974), pp. 319–331.
- [8] A. DVORETZKY, A. WALD AND J. WOLFOWITZ, *Relations among certain ranges of vector measures*, Pacific J. Math., 1 (1951), pp. 59–74.
- [9] H. HALKIN, *Lyapunov's theorem on the range of a vector measure and Pontryagin's maximum principle*, Arch. Rational Mech. Anal., 10 (1962), pp. 296–304.
- [10] ———, *On the necessary condition for optimal control of nonlinear systems*, J. Analyse Math., 12 (1964), pp. 1–82.
- [11] ———, *Some further generalizations of a theorem of Lyapunov*, Arch. Rational Mech. Anal., 17 (1964), pp. 272–277.
- [12] ———, *A generalization of LaSalle's bang-bang principle*, this Journal, 3 (1965), pp. 199–203.
- [13] ———, *On a generalization of a theorem of Lyapunov*, J. Math. Anal. Appl., 10 (1965), pp. 325–329.
- [14] ———, *A property of nonseparated convex sets*, Proc. Amer. Math. Soc., 17 (1966), pp. 1389–1395.
- [15] ———, *Convexity and control theory*, Functional Analysis and Optimization, E. R. Caianello, ed., Academic Press, New York, 1966, pp. 85–97.
- [16] ———, *Mathematical foundation of system optimization*, Topics in Optimization, G. Leitmann, ed., Academic Press, New York, 1967, pp. 197–262.
- [17] P. R. HALMOS, *On the set of values of a finite measure*, Bull. Amer. Math. Soc., 53 (1947), pp. 138–141.
- [18] ———, *The range of a vector measure*, Ibid., 54 (1948), pp. 416–421.
- [19] A. A. LYAPUNOV, *Sur les fonctions vecteurs complètement additives*, Izv. Akad. Nauk SSSR Ser. Mat., 8 (1940), pp. 465–478.
- [20] L. W. NEUSTADT, *The existence of optimal controls in the absence of convexity conditions*, J. Math. Anal. Appl., 7 (1963), pp. 110–117.
- [21] C. OLECH, *Lexicographical order, range of integrals, and bang-bang principle*, Mathematical Theory of Controls, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, New York, 1967, pp. 35–53.
- [22] M. B. SURYANARAYANA, *Linear control problems with total differential equations without convexity*, Trans. Amer. Math. Soc., 200 (1974).
- [23] ———, *Existence theorems for optimization problems concerning linear hyperbolic partial differential equations without convexity conditions*, J. Optimization Theory Appl., to appear.

RELAXED CONTROL PROBLEMS WITH STATE EQUALITY CONSTRAINTS*

A. B. SCHWARZKOPF†

Abstract. This paper derives a set of necessary conditions for a general optimal control problem with bilateral state and endpoint constraints with no assumptions of normality or regularity involving the constraints. The basic result is a separation theorem in a linear topological space and leads to the Pontryagin maximum principle. Although this separation theorem involves functionals from the dual of L^∞ , we show that, under weaker hypotheses than generally assumed, this result leads to the classical maximum principle. The extension of these results to include unilateral constraints is indicated.

1. Introduction. In this paper we will obtain necessary conditions for a fixed time optimal control problem with bounded relaxed controls and continuously differentiable (in the state) state equations subject to both unilateral and bilateral state and endpoint constraints. Such a problem can be adapted to include variable time problems although we do not consider them specifically. More precisely we wish to find necessary conditions for minimizing a functional of the form

$$J(C) = \int_a^b b^0(s, x(s), u(s)) ds$$

over a class C of (generalized) curves $C = (x, u)$ with state equation

$$(1.1) \quad x(t) = x(a) + \int_a^t f(s, x(s), u(s)) ds$$

and constraints

$$(1.2) \quad g^1(x(a)) + \int_a^t b^1(s, x(s), u(s)) ds \leq 0, \quad a \leq t \leq b,$$

$$(1.3) \quad b^2(t, x(t), u(t)) = 0, \quad \text{a.e. for } a \leq t \leq b,$$

and end conditions

$$(1.4) \quad h^1(x(a), x(b)) \leq 0, \quad h^2(x(a), x(b)) = 0.$$

This problem has a long history extending back at least to the work of Young and McShane in the calculus of variations around 1940. Young [25], [26] showed that the introduction of relaxed controls assured the existence of optimal curves for unconstrained problems in the calculus of variations. McShane extended these results to the problem of Bolza, and at the same time showed that the basic necessary conditions for optimal curves—then in the form of Lagrange multipliers and the Wierstrass condition—held with no assumptions of normality of the optimum [6], [7], [8]. More recently Warga showed that for constraints of the type (1.2) and (1.4) (i.e., unilateral), the same necessary conditions hold—in the form of the Pontryagin maximum principle—with no assumption of independence among the constraints (1.4) as previous authors had needed [20], [21]. Neustadt

* Received by the editors October 12, 1972, and in final revised form March 8, 1974.

† Department of Mathematics, University of Oklahoma, Norman, Oklahoma 73069.

has obtained similar results for the case where \mathcal{C} is replaced by the set of ordinary curves [12], [13], [14].

None of these results apply to bilateral constraints (1.3) without imposing the classical assumptions of independence among the constraints. It is the purpose of this paper to extend the results of Warga and Neustadt to bilateral constraints by proving a separation theorem of the type often used to derive the Pontryagin maximum principle without requiring any independence assumptions on the constraints. Our proof is similar to classical arguments, in that it makes use of an implicit function result to obtain a contradiction to optimality when the abstract maximum principle fails. Indeed, the proof of Theorem 1 is essentially the derivation of an implicit function theorem for functions of the form (1.3) with range in L^∞ .

The main result we obtain (Theorem 1) involves a linear functional from the dual of L^∞ and is not very useful in application, so we also present conditions under which the multipliers involved are bounded measurable functions. The additional hypothesis that we require (first proposed in [20], see also [17], and [21]) is that the convex hull of the set $b^2(t, x_0(t), U)$ —where x_0 is the optimal trajectory and U is the control set—contains an open set of fixed radius δ about the origin for all t in $[a, b]$. This requirement is more natural in the context of relaxed controls than the classical hypothesis [4], [7], [15], [19] which requires the introduction of functions q which describe the boundary of U and assumes that the Jacobian matrix of the functions (q, b^1, b^2) with respect to the control variables is non-singular for those components of b^1 for which (1.2) is an equality. We have not given a derivation of the maximum principle implied by the separation theorems (Theorem 1 and Theorem 2) in §2 since similar derivations for special cases of this problem are readily available in the literature [13], [14].

It is convenient to make a few observations about our choice of notation in the rest of the paper. In the first place we have used boldface type to indicate the identity function on a set. Thus $f(\mathbf{t}, \mathbf{x}, \mathbf{u})$ stands for $(t, x, u) \rightarrow f(t, x, u)$ if $f: T \times G \times U \rightarrow R^m$. This convention is an attempt to satisfy the need for mathematical rigor as well as concise notation in dealing with partial derivatives. If G is an open subset of R^n we let $\partial_x f(\mathbf{t}, \mathbf{x}, \mathbf{u})$ denote the $m \times n$ Jacobian matrix function whose entries are partial derivatives of the coordinate functions of f with respect to the variables labeled x , or equivalently the partial Fréchet derivative of f with respect to x .

Finally we should comment on our choice of notation for relaxed controls. We have adopted the older symbol

$$\mathfrak{M}[\phi(t, u); t]$$

of McShane [6], [7], [8], [10] rather than the more recent

$$\phi(t, v(t))$$

used in the books of Young [27] and Warga [24] to denote the value of the function $\phi(\mathbf{t}, \mathbf{u}): T \times U \rightarrow R$ at time t and under the relaxed control \mathfrak{M} or v . We make this choice, in spite of the natural appearance of the latter notation, in order to emphasize the fact that relaxed controls are mean value operators from the space $C[T \times U \rightarrow R]$ of bounded continuous real-valued functions on $T \times U$ to the

space $L^\infty[T \rightarrow R]$. (If $\phi(t, u): T \times U \rightarrow R^k$, then $\mathfrak{M}[\phi(t, u), t] \equiv (\mathfrak{M}[\phi^1(t, u); t], \mathfrak{M}[\phi^2(t, u); t], \dots, \mathfrak{M}[\phi^k(t, u); t])$.) In particular, we make fundamental use of the fact that, if \mathfrak{M}_1 and \mathfrak{M}_2 are both relaxed controls and $\gamma(t): T \rightarrow [0, 1]$ is a measurable function, then

$$\gamma(t)\mathfrak{M}_1[\cdot; t] + (1 - \gamma(t))\mathfrak{M}_2[\cdot; t]$$

is a relaxed control and

$$(\gamma(t)\mathfrak{M}_1 + (1 - \gamma(t))\mathfrak{M}_2)[\phi(t, u); t] = \gamma(t)\mathfrak{M}_1[\phi(t, u); t] + (1 - \gamma(t))\mathfrak{M}_2[\phi(t, u); t]$$

for any $\phi \in C[T \times U \rightarrow R]$. The reader unfamiliar with the concepts of “relaxed control” and “generalized curve” is referred to the discussions in [10]. Additional material is available in papers of Gambill [2], Ghouila Houry [3], and Warga [22] as well as the original papers of Young [25], [26] and McShane [6], [7], [8] already mentioned. A more exhaustive treatment appears in the books of Young [27, see “chattering controls”] and Warga [24].

2. The fundamental separation theorem. In order to simplify the arguments of this section we will amend the problem posed in the previous section to consider only equality (bilateral) state space and endpoint constraints. This is not really a specialization. In the next section we will indicate how Valentine’s method [18] allows us to impose unilateral constraints in this form.

Let us pose our problem more precisely. Suppose we are given an interval $T = [a, b]$ in R , a bounded open set G in R^n , and a compact set U in R^p . Further suppose that we have functions

$$f(t, \mathbf{x}, \mathbf{u}): T \times G \times U \rightarrow R^n, \\ b^0(t, \mathbf{x}, \mathbf{u}): T \times G \times U \rightarrow R \quad \text{and} \quad b(t, \mathbf{x}, \mathbf{u}): T \times G \times U \rightarrow R^m,$$

which are defined and continuous in a compact neighborhood of $T \times G \times U$ along with all first partial derivatives with respect to x . Let there also be given a function $h(\mathbf{x}_1, \mathbf{x}_2): G \times G \rightarrow R^l$ which is bounded and continuous along with all first partial derivatives on a neighborhood of its domain. Let \mathcal{C} be the collection of all generalized curves $C = (x, \mathfrak{M})$, where \mathfrak{M} is a relaxed control function on T with values over U (see [10] or [27]) and x satisfies

$$(2.1) \quad x(t) = x(a) + \int_a^t \mathfrak{M}[f(s, x(s), u); s] ds.$$

We wish to establish necessary conditions for a curve $C_0 = (x_0, \mathfrak{M}_0)$ from \mathcal{C} to minimize the functional

$$(2.2) \quad J(C) = \int_a^b \mathfrak{M}[b^0(s, x(s), u); s] ds$$

subject to the conditions

$$(2.3) \quad \mathfrak{M}[b(t, x(t), u); t] = 0 \quad \text{a.e. in } T,$$

and

$$(2.4) \quad h(x(a), x(b)) = 0.$$

Our main result takes the form of a separation of two sets in a linear topological space and leads to a maximum principle in the usual form. In order to state the theorem we need some further notation. Suppose that $C_0 = (x_0, \mathfrak{M}_0)$ is an optimal curve for the problem above and let \mathscr{Y} denote the collection of all pairs (η, Δ) , where $\eta \in R^n$ and Δ is an operator of the form

$$\Delta = \sum_{i=1}^k \alpha_i (\mathfrak{M}_i - \mathfrak{M}_0), \quad k = 1, 2, \dots,$$

where the α_i are real numbers and the \mathfrak{M}_i are relaxed controls over U . The collection \mathscr{Y} is a linear space under the natural definitions of sum and multiplication by a scalar. For each $y = (\eta, \Delta)$ in \mathscr{Y} (with $\sum |\alpha_i|$ and $|\eta|$ sufficiently small) we can define

$$(2.5) \quad x(t, y) = (x_0(a) + \eta) + \int_a^t (\mathfrak{M}_0 + \Delta)[f(s, x(s, y), u); s] ds,$$

$$(2.6) \quad J(y) = \int_a^b (\mathfrak{M}_0 + \Delta)[b^0(s, x(s, y), u); s] ds,$$

$$(2.7) \quad b(t, y) = (\mathfrak{M}_0 + \Delta)[b(t, x(t, y), u); t]$$

and

$$(2.8) \quad h(y) = h(x(a, y), x(b, y)).$$

Following McShane [10], we can verify that the functions $x(t, \alpha y)$, $J(\alpha y)$, $b(t, \alpha y)$ and $h(\alpha y)$ defined on R for fixed (t, y) in $T \times \mathscr{Y}$ are differentiable at $\alpha = 0$ with derivatives given by the solutions to the following equations:

$$(2.9) \quad Dx(t, y) = \eta + \int_a^t \{ \mathfrak{M}_0[\partial_x \bar{f}(s)] Dx(s, y) + \Delta[\bar{f}(s)] \} ds,$$

$$(2.10) \quad DJ(y) = \int_a^b \{ \mathfrak{M}_0[\partial_x \bar{b}^0(s)] Dx(s, y) + \Delta[\bar{b}^0(s)] \} ds,$$

$$(2.11) \quad Db(t, y) = \mathfrak{M}_0[\partial_x \bar{b}(t)] + \Delta[\bar{b}(t)] \quad \text{a.e.}$$

and

$$(2.12) \quad Dh(y) = \partial_{x_1} \bar{h} Dx(a, y) + \partial_{x_2} \bar{h} Dx(b, y).$$

Equation (2.9) can be solved for $Dx(t, y)$ to give

$$Dx(t, y) = \Phi(t)\eta + \Phi(t) \int_a^t \Phi^{-1}(s)\Delta[\bar{f}(s)] ds,$$

where Φ is the nonsingular $n \times n$ fundamental matrix solution to

$$\Phi(t) = I + \int_a^t \mathfrak{M}_0[\partial_x \bar{f}(s)] \Phi(s) ds,$$

and I is the $n \times n$ identity matrix. We note that Dx , DJ , Db , and Dh are all linear functions of y .

In the formulas above, as in the sequel, we have used a bar to indicate that any of the functions g , b , h , or f is evaluated along the trajectory x_0 of the optimal

curve C_0 . Thus

$$h(x_0(a), x_0(b)) \equiv \bar{h},$$

$$\mathfrak{M}_0[\partial_x b(t, x_0(t), u); t] \equiv \mathfrak{M}_0[\partial_x \bar{b}(t)]$$

and

$$\Delta[b(t, x_0(t), u); t] \equiv \Delta[\bar{b}(t)].$$

Let M denote the set of all elements y in \mathcal{Y} of the form $y = (\eta, \mathfrak{M} - \mathfrak{M}_0)$, where $|\eta| \leq 1$ and \mathfrak{M} is a relaxed control on U . Then we have the following fundamental theorem.

THEOREM 1. *Assume the notation and hypotheses introduced above. There exists a continuous linear functional λ defined on $R \times R^l \times L^\infty[T \rightarrow R^m]$ which separates the sets*

$$A = (DJ, Dh, Db)(M) \quad \text{and} \quad B = (-\infty, 0) \times \{0\} \times \{0\}.$$

Remark. Theorem 1 is the main result of this paper. The key to the proof is the observation that the set $Db(M)$ cannot contain a neighborhood of the origin unless the set $\{\Delta[b(t)] | (0, \Delta) \in M\}$ does also. This allows us to employ the implicit function theorem in an indirect proof parallel to the usual proofs of the maximum principle.

Proof. Let Z denote the set of all points $(Dh, Db)(y)$ with y in M and $DJ(y) < 0$, and let Z^* denote the closure of the cone

$$\bigcup_{\alpha > 0} [\alpha Z]$$

in the topology of $R^l \times L^\infty[T \rightarrow R^m]$. If Z^* is a proper subset of $R^l \times L^\infty[T \rightarrow R^m]$ then, since Z^* is a closed convex cone, there must exist a continuous linear functional ψ on $R^l \times L^\infty[T \rightarrow R^m]$ such that $\psi(Z^*) \geq 0$. This implies that the set $(DJ, \psi(Dh, Db))(M)$ in R^2 can be separated from $(-\infty, 0) \times (-\infty, 0)$, and hence there exists a nontrivial vector (α_1, α_2) such that $\alpha_1 \leq 0$ and

$$\alpha_1 DJ(y) + \alpha_2 \psi(Dh(y), Db(y)) \geq 0$$

for any y in M . Thus, if Z^* can be separated from zero the theorem holds.

We now suppose that $Z^* = R^l \times L^\infty[T \rightarrow R^m]$ and seek a contradiction. We will proceed through several steps to construct a control giving a solution to (2.1) which satisfies (2.3) and (2.4) and gives a value of J which is less than $J(C_0)$. The actual construction is carried out in Step 3 and Step 4 below. Step 1 and Step 2 are preliminary results.

Step 1. Let us begin by verifying the following claim.

$$(2.13) \quad \text{The set } P = \{\Delta \bar{b}(t) | y = (\eta, \Delta) \in M\} \text{ contains an open subset of } L^\infty[T \rightarrow R^m].$$

Suppose first that $m = 1$. Since U is compact, the functions b^* and b_* defined on T by

$$b^*(t) = \sup \{b(t, x_0(t), u) | u \in U\},$$

$$b_*(t) = \inf \{b(t, x_0(t), u) | u \in U\}$$

are both continuous, and both function values are attained on U at each $t \in T$.

Thus by the Flippov implicit function lemma [1, p. 78] (or by [11]) there exist (ordinary measurable) controls \mathfrak{M}^* and \mathfrak{M}_* such that

$$b^*(t) = \mathfrak{M}^*[\bar{b}(t)]$$

and

$$b_*(t) = \mathfrak{M}_*[\bar{b}(t)]$$

Therefore P contains every measurable function with values between $(b^* - \bar{b})$ and $(b_* - \bar{b})$. Now suppose that $b^*(t_0) = b_*(t_0)$ for some $t_0 \in T$. The function $\mathfrak{M}_0[\partial_x \bar{b}(t)]$ is measurable and hence approximately continuous ([9, p. 224]) on T except at points in a subset N with Lebesgue measure zero. The point t_0 must be in N since otherwise every function of the form

$$(2.14) \quad \mathfrak{M}_0[\partial_x \bar{b}(t)]Dx(t, y) + \Delta[\bar{b}(t)], \quad y \in M,$$

defined on T would be approximately continuous at t_0 and the closure of the linear span of such functions in $L^\infty[T \rightarrow R]$ would have the same property. This is impossible since no function equivalent to a step function with a jump at t_0 is approximately continuous at t_0 . Now let $\{t_k\}_{k=1}^\infty$ be a sequence of points from $(T - N)$ increasing (or decreasing) to t_0 , and define the function $q: T \rightarrow R$ by

$$q(t) = \begin{cases} 1, & t \in [t_k, (t_k + t_{k+1})/2], \quad k = 1, 2, \dots, \\ 0, & \text{otherwise.} \end{cases}$$

If $w: T \rightarrow R$ is a function of the form (2.14) which satisfies $\|q - w\|_\infty < \frac{1}{4}$, then w must have jumps of size at least $\frac{1}{2}$ at each point t_k and consequently there must be a constant $\alpha > 0$ such that $b^*(t_k) - b_*(t_k) > \alpha$ for all $k = 1, 2, \dots$. Since the sequence t_k converges to t_0 and both b^* and b_* are continuous, this means that $b^*(t_0) \geq b_*(t_0) + \alpha$, contradicting our assumption.

To extend this result to arbitrary values of m it suffices to show that the convex hull of the set $b(t, x_0(t), U)$ has nonempty interior for each t in T . This follows by observing that otherwise there exists a nonzero vector γ in R^m such that the inner product

$$\langle \gamma, b(t_0, x_0(t_0), u) \rangle = 0$$

for all $u \in U$. Applying the previous arguments to the function $\langle \gamma, b \rangle$ we see that $\langle \gamma, Z^* \rangle$ cannot equal $L^\infty[T \rightarrow R^1]$ which is a contradiction.

Step 2. In the constructions which follow we will be interested in a result like (2.13) for elements of M for which DJ is negative. We therefore establish the following result.

$$(2.15) \quad \begin{array}{l} \text{The set } Q = \{\Delta b(t) | y = (\eta, \Delta) \in M, DJ(y) < 0\} \\ \text{contains an open set in } L^\infty[T \rightarrow R^m]. \end{array}$$

We show this first under the hypothesis that $m = 1$. Suppose $y_0 = (\eta_0, \Delta_0) \in M$ and $DJ(y_0) = -\varepsilon < 0$. Since U is compact, $DJ(M)$ is bounded, and there exists a positive constant $c < \frac{1}{2}$ such that

$$|DJ(cM)| = |cDJ(M)| < \varepsilon/2.$$

Since the set P has nonempty interior, there exist elements $y^+ = (0, \Delta^+)$ and $y^-(0, \Delta^-)$ in M such that

$$\Delta^+[b(t)] = \Delta^-[b(t)] + \delta, \quad t \in T,$$

for some fixed positive δ . Now the set of all functions of the form $\Delta(\gamma)[b(t)]$, where we have

$$\Delta(\gamma) = \{(1 - c)\Delta_0 + c\{\gamma(t)\Delta^+ + (1 - \gamma(t))\Delta^-\}\}$$

defined for all measurable functions $\gamma: T \rightarrow [0, 1]$, is an open set in $L^\infty[T \rightarrow R]$. Moreover, setting $y(\gamma) = ((1 - c)\eta_0, \Delta(\gamma))$, we have

$$DJ(y(\gamma)) \in (1 - c)DJ(y_0) + cDJ(M)$$

and $DJ(y(\gamma)) < 0$. This proves our assertion when $m = 1$. The general result follows easily.

Step 3. The next part of our proof is the construction of a family of generalized curves $C(\alpha) = (\chi(\alpha), \mathfrak{M}(\alpha))$ satisfying (2.1) and (2.3), whose trajectories are differentiable with respect to α . We will do this in such a way that $C(\alpha_0) = C_0$ and $\partial_\alpha J(C(\alpha_0)) < 0$ for some positive α_0 . This will lead to a contradiction. For notational simplicity we suppose that $m = 1$ and $h = 0$. Let $y_1 = (\eta_1, \Delta_1)$ be an element of M with $\Delta_1 \bar{b}$ in the interior of Q and $DJ(y_1) < 0$. Since we are assuming that $Z^* = L^\infty[T \rightarrow R]$, for any $\varepsilon_1 > 0$ we may choose an element $y_2 = (\eta_2, \Delta_2)$ in y , with $cy_2 \in M$ for some positive c , and $DJ(y_2) < 0$, such that $\|Db(y_2) + Db(y_1)\|_\infty < \varepsilon_1$. Since $\Delta_1 \bar{b}$ is in the interior of Q there exist elements y^+ and y^- satisfying the same conditions as y_1 and in addition

$$(\Delta^+ - \Delta_1)[b(t)] = \delta_1, \quad (\Delta^- - \Delta_1)[b(t)] = -\delta_1 \quad \text{a.e. in } T$$

for some positive δ_1 . Define $y^* = (\eta^*, \Delta^*)$ and $y_* = (\eta_*, \Delta_*)$ by

$$\begin{aligned} \Delta^* &= \frac{1}{2}Ae^{B(t-a)}(\Delta^+ - \Delta_1), & \Delta_* &= \frac{1}{2}Ae^{B(t-a)}(\Delta^- - \Delta_1), \\ (2.16) \quad \eta^* &= \eta^+/B, & \eta_* &= \eta^-/B, \end{aligned}$$

where A and B are positive constants to be determined later. Now consider the expression

$$(2.17) \quad \{(\alpha e^\beta - \alpha_0)\Delta^* + (\alpha e^{-\beta} - \alpha_0)\Delta_* + (\alpha - \alpha_0)(\Delta_1 + \Delta_2) + \mathfrak{M}_0\}[b(t, \mathbf{x}, u); \mathbf{t}]$$

for $\alpha_0 > 0$, α, β real. The partial derivative of (2.17) with respect to β is given by

$$(\alpha e^\beta \Delta^* - \alpha e^{-\beta} \Delta_*)[b(t, \mathbf{x}, u); \mathbf{t}],$$

and this is not zero for $\beta = 0$, $\alpha = \alpha_0$, and x near $x_0(t)$. Hence there is a function $\beta(t, \mathbf{x}, \alpha)$ defined in a neighborhood of $\{(t, x_0(t), \alpha_0) | t \in T\}$ which makes (2.17) identically zero. The function β is measurable in t for fixed (x, α) and continuously differentiable in x and α for fixed t , and these partial derivatives are bounded measurable functions of t for fixed x and α . Moreover we have

$$(2.18) \quad \partial_\alpha \beta(t, x_0(t), \alpha_0) = - \frac{(\Delta_1 + \Delta_2)[\bar{b}(t)]}{2\alpha_0 d(t)}$$

and

$$(2.19) \quad \partial_x \beta(t, x_0(t), \alpha_0) = -\frac{\mathfrak{M}_0[\partial_x \bar{b}(t)]}{2\alpha_0 d(t)},$$

where

$$d(t) = \Delta^*[\bar{b}(t)] = -\Delta_*[\bar{b}(t)] = \delta_1 A e^{B(t-a)} > 0$$

for t in T . The system of equations

$$(2.20) \quad \begin{aligned} \chi(t, \alpha) = x_0(a) + (\alpha - \alpha_0)(\eta_1 + \eta_2) \\ + \int_a^t \{(\alpha e^{\beta(s, \chi(s, \alpha), \alpha)} - \alpha_0)\Delta^* + (\alpha e^{-\beta(s, \chi(s, \alpha), \alpha)} - \alpha_0)\Delta_* \\ + (\alpha - \alpha_0)(\Delta_1 + \Delta_2) + \mathfrak{M}_0\}[f(s, \chi(s, \alpha), u); s] ds \end{aligned}$$

defined for $\alpha \in R$ and $t \in T$, has a unique solution $\chi(t, \alpha)$ for α near α_0 , which satisfies $\chi(t, \alpha_0) = x_0(t)$. Define $\mathfrak{M}(\alpha)$ by

$$(2.21) \quad \begin{aligned} \mathfrak{M}(\alpha)[\cdot; t] = \{(\alpha e^{\beta(t, \chi(t, \alpha), \alpha)} - \alpha_0)\Delta^* + (\alpha e^{-\beta(t, \chi(t, \alpha), \alpha)} - \alpha_0)\Delta_* \\ + (\alpha - \alpha_0)(\Delta_1 + \Delta_2) + \mathfrak{M}_0\}[\cdot; t] \end{aligned}$$

and observe that $\mathfrak{M}(\alpha_0) = \mathfrak{M}_0$. Moreover,

$$(2.22) \quad \mathfrak{M}(\alpha)[b(t, \chi(t, \alpha), u); t] = 0 \quad \text{a.e. in } T.$$

Now by a standard theorem on the differentiability of solutions to systems of differential equations (e.g., [9, p. 356]) the function χ has a partial derivative with respect to α at $\alpha = \alpha_0$ given by the solution to the equation

$$(2.23) \quad \begin{aligned} D\chi(t) = (\eta_1 + \eta_2) \\ + \int_a^t \left\{ \mathfrak{M}_0 \left[\partial_x \bar{f}(s) - \frac{(\Delta^* - \Delta_*)}{2d(s)} [\bar{f}(s)] \partial_x \bar{b}(s) \right] D\chi(s) \right. \\ \left. + (\Delta_1 + \Delta_2) \left[\bar{f}(s) - \frac{(\Delta^* - \Delta_*)}{2d(s)} [\bar{f}(s)] \bar{b}(s) \right] + (\Delta^* + \Delta_*) [\bar{f}(s)] \right\} ds. \end{aligned}$$

By direct calculation or an appeal to [4, Appendix 6] we can verify that the difference quotient which defines $D\chi$ converges uniformly on T . Finally we observe that the functional

$$(2.24) \quad J(\alpha) = \int_a^b \mathfrak{M}(\alpha)[b^0(s, \chi(s, \alpha), u); s] ds$$

has a partial derivative with respect to α given at $\alpha = \alpha_0$ by

$$(2.25) \quad \begin{aligned} \partial_\alpha J(C(\alpha_0)) = \int_a^b \left\{ \mathfrak{M}_0 \left[\partial_x \bar{b}^0(s) - \frac{(\Delta^* - \Delta_*)}{2d(s)} [\bar{b}^0(s)] \partial_x \bar{b}(s) \right] D\chi(s) \right. \\ \left. + (\Delta_1 + \Delta_2) \left[\bar{b}^0(s) - \frac{(\Delta^* - \Delta_*)}{2d(s)} [\bar{b}^0(s)] \bar{b}(s) \right] + (\Delta^* + \Delta_*) [\bar{b}^0(s)] \right\} ds. \end{aligned}$$

Step 4. We complete the proof of Theorem 1, under the supplementary hypotheses $m = 1$ and $h = 0$, by showing that for proper choices of α , y_2 , A , and B the curve $C(\alpha) = (\chi(\alpha), \mathfrak{M}(\alpha))$ is a solution to the system (2.1)–(2.4) which has $J(C(\alpha)) < J(C_0)$. The operator $\mathfrak{M}(\alpha)$ is not automatically a relaxed control for $\alpha > \alpha_0$ unless

$$\begin{aligned}(\alpha e^{\beta(t, \chi(t, \alpha), \alpha)} - \alpha_0) &\geq 0, \\(\alpha e^{-\beta(t, \chi(t, \alpha), \alpha)} - \alpha_0) &\geq 0\end{aligned}$$

and

$$(\alpha e^{\beta(t, \chi(t, \alpha), \alpha)} - \alpha_0) + (\alpha e^{-\beta(t, \chi(t, \alpha), \alpha)} - \alpha_0) + 2(\alpha - \alpha_0) \leq 1.$$

The last inequality holds for α near α_0 since $\beta(t, \chi(t, \alpha), \alpha)$ converges to zero uniformly with $(\alpha - \alpha_0)$. The first two inequalities hold provided

$$(2.26) \quad e^{|\beta(t, \chi(t, \alpha), \alpha)|} \leq \frac{\alpha}{\alpha_0}.$$

To demonstrate that this is the case for judicious choices of α , y_2 , A and B , suppose that $t_0 \in T$ and $\beta(t_0, \chi(t_0, \alpha), \alpha) \geq 0$ for $\alpha - \alpha_0$ a small positive number. The reverse inequality leads to a similar argument. We have from (2.18), (2.19) and the boundedness of $\partial_x b$ that

$$\begin{aligned}(2.27) \quad |\partial_x e^{\beta(t_0, \chi(t_0, \alpha_0), \alpha_0)}| &= |1 \cdot [\partial_x \beta(t_0, x_0(t_0), \alpha_0) D\chi(t_0, \alpha_0) + \partial_x \beta(t_0, x_0(t_0), \alpha_0)]| \\&\leq \frac{1}{2\alpha_0 d(t_0)} \{ |\mathfrak{M}_0[\partial_x \bar{b}(t_0)] [D\chi(t_0) - Dx(t_0, y_1 + y_2)] \\&\quad + |\mathfrak{M}_0[\partial_x \bar{b}(t_0)] Dx(t_0, y_1 + y_2) + (\Delta_1 + \Delta_2)[\bar{b}(t_0)] \}| \\&\leq \frac{1}{2\alpha_0 d(t_0)} \{ K_1 |D\chi(t_0) - Dx(t_0, y_1 + y_2)| + \varepsilon_1 \}\end{aligned}$$

for some positive constant K_1 . On the other hand,

$$\begin{aligned}(2.28) \quad &|D\chi(t_0) - Dx(t_0, y_1 + y_2)| \\&= \left| \int_a^{t_0} \left\{ \mathfrak{M}_0 \left[\partial_x \bar{f}(s) - \frac{(\Delta^* - \Delta_*)}{2d(s)} [\bar{f}(s)] \partial_x \bar{b}(s) \right] (D\chi(s) - Dx(s, y_1 + y_2)) \right. \right. \\&\quad \left. \left. - \frac{(\Delta^* - \Delta_*)}{2d(s)} [\bar{f}(s)] [\mathfrak{M}_0[\partial_x \bar{b}(s)] Dx(s, y_1 + y_2) + (\Delta_1 + \Delta_2)[\bar{b}(s)]] \right. \right. \\&\quad \left. \left. + (\Delta^* + \Delta_*)[\bar{f}(s)] \right\} ds \right| \\&\leq K_2 \int_a^{t_0} |D\chi(s) - Dx(s, y_1 + y_2)| ds + \int_a^{t_0} \{ K_3 \varepsilon_1 + K_4 A e^{B(s-a)} \} ds\end{aligned}$$

so that Grönwall's lemma gives

$$\begin{aligned}|D\chi(t_0) - Dx(t_0, y_1 + y_2)| &\leq K_5 \{ \varepsilon_1 + (A/B) [e^{B(t_0-a)} - 1] \} \\&\leq K_5 \{ \varepsilon_1 + (A/B) e^{B(t_0-a)} \}\end{aligned}$$

for constants $K_2 - K_5$ which depend only on the functions f , b and b^0 and the operators Δ^+ and Δ^- . Thus we can choose $B > (2K_1 \cdot K_5)/\delta_1$, $A = e^{-B(b-a)}$, and

y_2 so that $\varepsilon_1 < \delta_1 A / (2K_5 K_1)$ and (2.27) gives

$$(2.29) \quad |\partial_\alpha e^{\beta(t_0, \chi(t_0, \alpha_0), \alpha_0)}| \leq \frac{1}{2\alpha_0 d(t_0)} \left(\frac{\delta_1 A}{2} + \frac{\delta_1 A}{2} e^{B(t_0 - a)} + \frac{\delta_1 A}{2K_5} \right) \leq \frac{1}{2\alpha_0}.$$

Moreover,

$$Ae^{B(t_0 - a)} = e^{-(B - t_0)} \leq 1.$$

Since the same inequalities hold when $\beta(t_0, \chi(t_0, \alpha), \alpha)$ is negative and since the function $\gamma: T \times R \rightarrow R$ defined by

$$\gamma(t, \alpha) = \begin{cases} \frac{\chi(t, \alpha) - (t, \alpha_0)}{\alpha - \alpha_0}, & \alpha \neq \alpha_0, \\ D\chi(t), & \alpha = \alpha_0, \end{cases}$$

is continuous on a compact neighborhood of $T \times \{\alpha_0\}$ in $T \times R$, convergence of the difference quotient is uniform in T , and for sufficiently small $(\alpha - \alpha_0)$, $\mathfrak{M}(\alpha)$ is a relaxed control.

Returning to (2.25) we see that an argument similar to (2.28) shows that by decreasing δ_1 and ε_1 if necessary we can assume that $\partial_\alpha J(C(\alpha_0))$ is arbitrarily close to $DJ(y_1 + y_2)$ and, in particular, close enough to be negative. This implies that for α near α_0 , $J(\alpha) < J(C_0)$ which is impossible.

To extend this conclusion to general m we introduce elements y_i^+ and y_i^- , $i = 1, \dots, m$, in (2.16) with the additional property that

$$(\Delta_i^\pm - \Delta_1)[\bar{b}^j(t)] = \begin{cases} \pm \delta_1, & i = j, \\ 0, & i \neq j. \end{cases}$$

We then write (2.17) as

$$(2.17') \quad \left\{ \sum_{i=1}^m [(\alpha e^{\beta_i} - \alpha_0) \Delta_i^* + (\alpha e^{-\beta_i} - \alpha_0) \Delta_{i*}] + (\alpha - \alpha_0)(\Delta_1 + \Delta_2) + \mathfrak{M}_0 \right\} [b(t, \mathbf{x}, u); \mathbf{t}].$$

The Fréchet derivative of (2.17') with respect to the vector β is nonsingular and we apply the implicit function theorem to obtain a vector function $\beta(t, \mathbf{x}, \alpha)$. The rest of the argument follows the outline above with straightforward modifications. To allow for nonzero h , let e_1, \dots, e_l be a basis of vectors in R^l , and for each $i = 1, 2, \dots, 2l$ choose elements y_1^i and y_2^i in the same manner as y_1 and y_2 were chosen in the paragraph before (2.16), but satisfying the additional relation

$$Dh(y_1^i + y_2^i) = \begin{cases} e^i, & i = 1, \dots, l, \\ -e^i, & i = l + 1, \dots, 2l. \end{cases}$$

We then proceed with the construction in Steps 3 and 4 above to obtain a function $C(\alpha, \sigma) = (\chi(\alpha, \sigma), \mathfrak{M}(\alpha, \sigma))$ depending on $\alpha \in R$ and $\sigma \in R^{2l}$, and a curve $\sigma(\alpha): R \rightarrow R^{2l}$ such that

$$\mathfrak{M}(\alpha, \sigma(\alpha))[b(t, \chi(t, \alpha, \sigma(\alpha)), u); t] \equiv 0$$

for almost all t in T ,

$$h(\chi(a, \alpha, \sigma(\alpha)), \chi(b, \alpha, \sigma(\alpha))) \equiv 0,$$

and the derivative of $J(C(\alpha, \sigma(\alpha)))$ with respect to α is negative. The conclusion of the theorem then follows as above.

Theorem 1 is somewhat unsatisfactory in that it gives a necessary condition involving elements in the dual space of $L^\infty[T \rightarrow R^m]$ which consists of integral operators with respect to finitely additive set functions. The following theorem gives a condition under which this difficulty can be circumvented. This condition was suggested by Warga [20, (3.1.2.9)] for constraints of type (2.2).

THEOREM 2. *Suppose we assume that, in addition to the hypotheses of Theorem 1, the convex hull of the set $b(t, x_0(t), U)$ contains an open set of fixed radius $\delta > 0$ about zero for almost all $t \in T$. Then the functional λ can be extended to be continuous on $R \times R^l \times L^1[T \rightarrow R^m]$.*

Proof. We prove this theorem by showing that under the hypotheses above the functional λ of Theorem 1 must separate zero from a subset of $R \times R^l \times L^\infty[T \rightarrow R^m]$ which is open in the topology of $R_1 \times R^l \times L^1[T \rightarrow R^m]$ on that set. Suppose for notational simplicity that $m = 1$. The general theorem follows by repeating the following argument for each dimension, but requiring $\xi \in L^\infty[T \rightarrow R^m]$ to be identically zero in all but one coordinate. By hypothesis and by (2.3) there exist elements $y^+ = (\eta^+, \Delta^+)$ and $y^- = (\eta^-, \Delta^-)$ in M such that

$$\begin{aligned} (\mathfrak{M}_0 + \Delta^+)[\bar{b}(t)] &= \mathfrak{M}_0[\bar{b}(t)] + \Delta^+[\bar{b}(t)] \\ &= 0 + \Delta^+[\bar{b}(t)] = \delta > 0, \end{aligned}$$

and similarly

$$(\mathfrak{M}_0 + \Delta^-)[\bar{b}(t)] = -\delta < 0$$

for almost all t in T . Let ξ be an arbitrary element of $L^\infty[T \rightarrow R]$ and define the function $\gamma(t, \chi): T \times R^n \rightarrow R$ by

$$(2.30) \quad \gamma(t, \chi) = (1/\delta)[\xi(t) - \mathfrak{M}_0[\partial_x \bar{b}(t)]\chi].$$

The function γ is measurable in t for fixed χ and Lipschitz continuous in χ for fixed t so that the functions

$$\begin{aligned} \gamma^+(t, \chi) &= \frac{1}{2}[\gamma(t, \chi) + |\gamma(t, \chi)|], \\ \gamma^-(t, \chi) &= \frac{1}{2}[-\gamma(t, \chi) + |\gamma(t, \chi)|] \end{aligned}$$

also have the same properties. It follows (see [16, Thm. 3.5, p. 96] for example) that the integral equation

$$(2.31) \quad \begin{aligned} D\chi(t, \xi) &= \int_a^t \{\mathfrak{M}_0[\partial_x \bar{f}(s)]D\chi(s, \xi) \\ &\quad + \gamma^+(s, D\chi(s, \xi))\Delta^+[\bar{f}(s)] + \gamma^-(s, D\chi(s, \xi))\Delta^-[\bar{f}(s)]\} ds \end{aligned}$$

has a unique solution. Now if we define $\Delta(\xi)$ by

$$\Delta(\xi)[\cdot; \mathbf{t}] = \gamma^+(\mathbf{t}, D\chi(\mathbf{t}, \xi))\Delta^+[\cdot; \mathbf{t}] + \gamma^-(\mathbf{t}, D\chi(\mathbf{t}, \xi))\Delta^-[\cdot; \mathbf{t}],$$

we can define a function $Y: L^\infty[T \rightarrow R] \rightarrow \mathcal{Y}$ by $Y(\xi) = (0, \Delta(\xi))$. Comparing (2.9) with (2.31) we see that

$$D\chi(t, \xi) = D\mathbf{x}(t, Y(\xi)).$$

From (2.30) we see that

$$\begin{aligned} \mathfrak{M}_0[\partial_x \bar{b}(t)]D\chi(t, \xi) + \gamma^+(t, D\chi(t, \xi))\Delta^+[\bar{b}(t)] + \gamma^-(t, D\chi(t, \xi))\Delta^-[\bar{b}(t)] \\ = \mathfrak{M}_0[\partial_x \bar{b}(t)]D\chi(t, \xi) + \gamma(t, D\chi(t, \xi)) \cdot \delta \\ = \xi(t). \end{aligned}$$

On the other hand the definition of γ and (2.31) imply that

$$\begin{aligned} |D\chi(t, \xi)| &\leq \int_a^t \{ |\mathfrak{M}_0[\partial_x \bar{f}(s)]| |D\chi(s, \xi)| \\ &\quad + (1/\delta)(|\xi(s)| + |\mathfrak{M}_0[\partial_x \bar{b}(s)]| |D\chi(s, \xi)|)(|\Delta^+[\bar{f}(s)]| + |\Delta^-[\bar{f}(s)]|) \} ds \\ &\leq \int_a^t \{ K_1 |D\chi(s, \xi)| + K_2 |\xi(s)| \} ds \end{aligned}$$

and an application of Grönwall's lemma implies

$$\begin{aligned} (2.32) \quad |D\chi(t, \xi)| &\leq K_3 \int_a^t |\xi(s)| ds \\ &\leq K_3 \|\xi\|_1, \end{aligned}$$

where K_1 , K_2 , and K_3 are constants independent of ξ . By (2.10) and the definition of $\Delta(\xi)$ we have

$$\begin{aligned} (2.33) \quad |DJ(Y(\xi))| &\leq \int_a^b \{ |\mathfrak{M}_0[\partial_x \bar{b}^0(s)]| |D\chi(s, \xi)| + |\Delta(\xi)[\bar{b}^0(s)]| \} ds \\ &\leq K_4 \|\xi\|_1 \end{aligned}$$

for some constant K_4 independent of ξ .

We may suppose without loss of generality that $\lambda((DJ, Dh, Db)(M)) \leq 0$. Since $\xi(t)$ is bounded, so is $\gamma(t, D\chi(t, \xi))$ and there is a positive constant c such that $cY(\xi)$ is in M . Thus

$$\lambda((DJ, Dh, Db)(Y(\xi))) = (1/c)\lambda((DJ, Dh, Db)(cY(\xi))) \leq 0.$$

Now let Γ denote the set of all elements in $L^\infty[T \rightarrow R]$ with L^1 -norm less than 1, and let λ_1 , λ_2 , and λ_3 be the projections of λ onto the spaces R , R^l , $L^\infty[T \rightarrow R^m]$, respectively. The functional λ_1 can be represented by multiplication by a non-positive number α_0 and λ_2 by an inner product with respect to a vector α . Let

$$Q = \{z \in R^l | \lambda_2(z) < 0\}$$

and define the convex sets Λ_1 and Λ_2 by

$$\Lambda_1 = (K_4, \infty), \quad \Lambda_2 = -\frac{K_3 \cdot K_5}{|\alpha|} \alpha + Q,$$

where $K_5 = |\partial_{x_1} \bar{h}| + |\partial_{x_2} \bar{h}|$, and consider the set $\Lambda_1 \times \Lambda_2 \times \Gamma$. We have

$$\begin{aligned} \lambda(\Lambda_1 \times \Lambda_2 \times \Gamma) &= \lambda_1(\Lambda_1) + \lambda_2(\Lambda_2) + \lambda_3(\Gamma) \\ &\leq \lambda_1(DJ(Y(\Gamma))) + \lambda_2(Dh(Y(\Gamma))) + \lambda_3(\Gamma) \\ &= \lambda((DJ, Dh, Db)(Y(\Gamma))) \leq 0. \end{aligned}$$

But $\Lambda_1 \times \Lambda_2 \times \Gamma$ is open in the $R \times R^l \times L^1[T \rightarrow R]$ topology on $R \times R^l \times L^\infty[T \rightarrow R]$, so λ must be continuous in that topology ([5, Thm. 5.4(v), p. 37]). Since the topology of $R \times R^l \times L^1[T \rightarrow R]$ is locally convex, λ extends to the entire space ([5, Thm. 14.1 (iii), p. 118]), proving the theorem.

We will have use for the following result, which is an obvious corollary of Theorem 2.

COROLLARY 1. *Assume the hypotheses of Theorem 1. Assume also that $b = (b^1, b^2)$, where $b^i: T \times G \times U \rightarrow R^{m_i}$ and $m_1 + m_2 = m$. Suppose there exists a fixed $\delta > 0$ and, for almost all t in T , there is a subset $U(t)$ of U such that $b^1(t, x_0(t), U(t))$ contains a neighborhood of radius δ about zero in R^{m_1} and $b^2(t, x_0(t), U(t)) = 0$. Then the functional λ can be extended to be continuous on $R \times R^l \times L^1[T \rightarrow R^{m_1}] \times L^\infty[T \rightarrow R^{m_2}]$.*

The hypotheses of Theorem 2 are somewhat stronger than we would like, since they cannot be verified without making some assumption about the nature of the optimal curve C_0 . It is tempting to hope that the same theorem might be true under the weaker assumption that the convex hull of $b(t, x_0(t), U)$ contains an open set of radius $\delta > 0$ for almost all t . This conjecture is false, however, as the following example shows. Suppose $U = \{(u_1, u_2) | -1 \leq u_1 \leq 1, -1 \leq u_2 \leq 1\}$ and set $f(t, x, u) = u_1$, $b^0(t, x, u) = x$, $b(t, x, u) = tx - (u_2)^2$, and $h(x_1, x_2) = x_1$. We then have the problem of minimizing the functional

$$J(C) = \int_0^1 x(s) ds,$$

where

$$x(t) = \int_0^t \mathfrak{M}[u_1; s] ds$$

subject to the constraint

$$tx(t) - \mathfrak{M}[(u_2)^2; t] = 0 \quad \text{a.e. in } [0, 1].$$

The minimum value of J under these conditions is given by setting $x(t) = 0$ on $[0, 1]$ and consequently $\mathfrak{M}_0[u_1; t] \equiv 0$; $\mathfrak{M}_0[(u_2)^2; t] \equiv 0$. For $y = (0, \Delta) \in \mathcal{Y}$ we have

$$\begin{aligned} Dx(t, y) &= \int_0^t \Delta[u_1; s] ds, \\ DJ(y) &= \int_0^1 \int_0^\sigma \Delta[u_1; s] ds dt, \\ Db(t, y) &= t \int_0^t \Delta[u_1; s] ds - \Delta[(u_2)^2; t]. \end{aligned}$$

If the conclusion of Theorem 2 is true, there exists a constant $\mu \leq 0$ and a bounded measurable function $v: [0, 1] \rightarrow R$ such that

$$\begin{aligned} 0 &\geq \mu \int_0^1 \int_0^\sigma \Delta[u_1; s] \, ds \, d\sigma \\ &\quad + \int_0^1 v(\sigma) \left\{ \sigma \int_0^\sigma \Delta[u_1; s] \, ds - \Delta[(u_2)^2; \sigma] \right\} d\sigma. \end{aligned}$$

Setting $\Delta[(u_2)^2; \sigma] \equiv 0$ we have

$$\begin{aligned} 0 &\geq \int_0^1 \int_0^\sigma (\mu + \sigma v(\sigma)) \Delta[u_1; s] \, ds \, d\sigma \\ &= \int_0^1 \left\{ \int_s^1 (\mu + \sigma v(\sigma)) \, d\sigma \right\} \Delta[u_1; s] \, ds \end{aligned}$$

by Fubini's theorem. Since $\Delta[u_1; s]$ can take on any value in $[-1, 1]$, the inequality above must in fact be an equality. Using the fundamental lemma of the calculus of variations, or the fact that the last integral on the right is a continuous linear functional on $L^\infty[[0, 1] \rightarrow R]$ which is zero on an open set of functions of the form $\Delta[u_1; t]$, we conclude that

$$\int_s^1 (\mu + \sigma v(\sigma)) \, d\sigma \equiv 0$$

and consequently

$$v(\sigma) = -\mu/\sigma$$

on $[0, 1]$. This is impossible since $-\mu/\sigma$ is not bounded (or even in L^p , $1 \leq p < \infty$) unless $\mu = 0$, and in that case v would be identically zero as well. Since $\Delta[(u_2)^2; s]$ can take on any value in $[0, 1]$, the hypotheses of the conjecture hold, showing it to be false.

3. Unilateral constraints. In §2 we indicated that Theorems 1 and 2 apply to optimizations with unilateral constraints, since these constraints can be reposed as bilateral constraints using Valetine's method [18]. We will demonstrate this for a simple problem. More general results follow the same line of argument. Suppose that the system defined by (2.1) and (2.2) has a single unilateral constraint of the form

$$(3.1) \quad x^{n+1}(t) \equiv g(x(a)) + \int_a^t \mathfrak{M}[b^1(s, x(s), u); s] \, ds \leq 0,$$

where $g(\mathbf{x}): G \rightarrow R$ is continuously differentiable on a compact neighborhood of G , and $b^1(\mathbf{t}, \mathbf{x}, \mathbf{u}): T \times G \times U \rightarrow R$ is continuous along with its first partial derivatives with respect to x on a compact neighborhood of $T \times G \times U$.

Let K be a bound for $|b^1(T, G, U)|$. Define $U^* = U \times [0, 2K]$, $f^*(\mathbf{t}, \mathbf{x}, \mathbf{u}) = (f(\mathbf{t}, \mathbf{x}, \mathbf{u}), b^1(\mathbf{t}, \mathbf{x}, \mathbf{u}))$, $b(\mathbf{t}, \mathbf{x}, \mathbf{u}) = \mathbf{x}^{n+1} + \mathbf{u}^{n+1}$ and $h(\mathbf{x}_1, \mathbf{x}_2) = g(x_1) - x_1^{n+1}$,

so that the problem posed in §2 is the following. Find a curve $(x^*, \mathfrak{M}^*) \equiv ((x, x^{n+1}), (\mathfrak{M}, \mathfrak{M}^{n+1}))$ which minimizes

$$(3.2) \quad J(C) = \int_a^b \mathfrak{M}[b^0(s, x(s), u); s] ds,$$

where

$$(3.3) \quad \begin{aligned} x(t) &= x(a) + \int_a^t \mathfrak{M}[f(s, x(s), u); s] ds, \\ x^{n+1}(t) &= x^{n+1}(a) + \int_a^t \mathfrak{M}^{n+1}[b^1(s, x(s), u); s] ds \end{aligned}$$

subject to

$$(3.4) \quad x^{n+1}(t) + \mathfrak{M}[u^{n+1}; t] = 0 \quad \text{a.e. on } T,$$

and

$$(3.5) \quad g(x(a)) - x^{n+1}(a) = 0.$$

Here \mathfrak{M} is a relaxed control on U and \mathfrak{M}^{n+1} is a relaxed control on $[0, 2K]$. The functions (2.10)–(2.13) now have the form

$$(3.6) \quad Dx(t, y^*) = \eta + \int_a^t \{ \mathfrak{M}_0[\partial_x \bar{f}(s)] Dx(s, y^*) + \Delta[\bar{f}(s)] \} ds,$$

$$Dx^{n+1}(t, y^*) = \eta^{n+1} + \int_a^t \{ \mathfrak{M}_0[\partial_x \bar{b}^1(s)] Dx(s, y^*) + \Delta[\bar{b}^1(s)] \} ds,$$

$$(3.7) \quad DJ(y^*) = \int_a^b \{ \mathfrak{M}_0[\partial_x \bar{b}^0(s)] Dx(s, y^*) + \Delta[\bar{b}^0(s)] \} ds,$$

$$(3.8) \quad Db(t, y^*) = Dx^{n+1}(t, y^*) + \Delta^{n+1}[u^{n+1}; t],$$

$$(3.9) \quad Dh(y^*) = (\partial_x \bar{g} \cdot \eta) - \eta^{n+1}$$

for $y^* = (y, \Delta^{n+1})$, where $y \in Y$ and $\Delta^{n+1} = \mathfrak{M}^{n+1} - \mathfrak{M}_0^{n+1}$. Applying Theorem 1 we see that there exists a nonpositive constant α_0 , a number α and a finitely additive set function μ such that for $y^* \in \mathfrak{M}^*$,

$$(3.10) \quad \lambda((DJ, Dh, Db)(y^*)) \equiv \alpha_0 DJ(y^*) + \alpha Dh(y^*) + \int_a^b Db(s, y^*) \mu(ds) \leq 0.$$

Now $Dx^{n+1}(t, y^*)$ is continuous, so for continuous $\Delta^{n+1}[u^{n+1}; t]$ we may replace μ by a countably additive measure μ^* . Indeed, the element $(\alpha_0, \alpha, -k)$ is in the interior of the set on which λ is positive, so that α_0, α , and μ^* do not all vanish identically. Setting $\eta, \eta^{n+1}, \Delta \bar{f}, \Delta \bar{b}^0$, and $\Delta \bar{b}^1$ all to zero shows that

$$\int_a^b \Delta^{n+1}[u^{n+1}; s] \mu^*(ds) \leq 0$$

for any nonnegative continuous function $\Delta^{n+1}[u^{n+1}; t]$, and consequently μ^* must be nonpositive. Moreover $\mu^* = 0$ on any interval on which $x^{n+1}(t) < 0$.

Thus we have shown that there exist numbers α_0, α with $\alpha_0 \leq 0$ and a nonpositive countably additive measure μ^* on T such that (for $y^* = (y, 0)$)

$$(3.11) \quad \alpha_0 DJ(y) + \alpha Dh(y) + \int_a^b Db(s, y) \mu^*(ds) \leq 0$$

for any $y \in M$. This is the abstract maximum principle for the unilateral constraint (3.1).

4. The multiplier rule. There are numerous references which take a result like Theorem 1 or Theorem 2 and derive concrete necessary conditions for an optimal curve (e.g., [10], [13], [14]). Similar arguments produce a maximum principle here, but we will not go through them. We will state the theorem for the problem posed in §1 under sufficient hypotheses to apply Theorem 2. A similar theorem holds under the weaker hypotheses of Theorem 1, but it yields a maximum principle in integrated form involving finitely additive measures. Considering our modest purposes we will simplify the problem by supposing that $m_1 = m_2 = l_2 = 1$ and $h^1 \equiv 0$.

Suppose we are given an interval $T = [a, b]$ in R , a bounded open set G in R^n and a compact set U in R^p . Further suppose that we have functions $f(\mathbf{t}, \mathbf{x}, \mathbf{u}): T \times G \times U \rightarrow R^n$, and $b^i(\mathbf{t}, \mathbf{x}, \mathbf{u}): T \times G \times U \rightarrow R$, $i = 0, 1, 2$, which are defined and continuous in a compact neighborhood of $T \times G \times U$, along with all first partial derivatives with respect to x . Let there also be given functions $g(\mathbf{x}): G \rightarrow R$ and $h(x_1, x_2): G \times G \rightarrow R$ which are continuously differentiable on a compact neighborhood of their domains. Let $C_0 = (x_0, \mathfrak{M}_0)$ minimize the functional

$$(4.1) \quad \int_a^b \mathfrak{M}[b(s, x(s), u); s] ds$$

over the class C of all generalized curves satisfying

$$(4.2) \quad x(t) = x(a) + \int_a^t \mathfrak{M}[f(s, x(s), u); s] ds$$

with

$$(4.3) \quad z(t) = g(x(a)) + \int_a^t \mathfrak{M}[b^1(s, x(s), u); s] ds \leq 0,$$

$$(4.4) \quad \mathfrak{M}[b^2(t, x(t), u); t] = 0$$

for almost all t in T , and

$$(4.5) \quad h(x(a), x(b)) = 0.$$

Further suppose that there exists a number $\delta > 0$ such that for almost all t in T ,

$$(4.6) \quad \begin{aligned} \sup_{u \in U} b^2(t, x_0(t), u) &\geq \delta, \\ \inf_{u \in U} b^2(t, x_0(t), u) &\leq -\delta. \end{aligned}$$

Then we have the following theorem. The proof is straightforward, except for the derivation of (iv) from the integrated form of the maximum principle, and this follows almost exactly the arguments given by McShane in [10, pp. 465–466].

THEOREM 3 (The maximum principle). *Assume the notation and assumptions above. Then there exist a number μ_0 , a vector μ^3 and measurable functions μ^1 , μ^2 and ψ satisfying (i)–(iv) below.*

(i) *The number μ^0 is nonpositive and the function μ^1 is bounded and monotone increasing on T with $\mu^1(b) = 0$. Furthermore μ^1 is constant on any interval on which $z(t)$ is negative.*

(ii) *The function ψ is absolutely continuous on T and satisfies*

$$\begin{aligned} \frac{d}{dt} \psi(t) = & -\mathfrak{M}_0[\partial_x \bar{f}(t)]\psi(t) + \mu_0 \mathfrak{M}_0[\partial_x \bar{b}^0(t)] \\ & + \mu^1(t) \mathfrak{M}_0[\partial_x \bar{b}^1(t)] + \mu^2(t) \mathfrak{M}_0[\partial_x \bar{b}^2(t)] \end{aligned}$$

almost everywhere.

(iii)

$$\psi(a) = \mu^1(a) \partial_x g(x(a)) + \mu^3 \partial_{x_1} \bar{h}^2$$

and

$$\psi(b) = -\mu^3 \partial_{x_2} \bar{h}^2.$$

(iv) *For all t in T except those in a negligible set A we have*

$$\begin{aligned} & -\psi(t)f(t, x_0(t), u_0) + \mu^0 b^0(t, x_0(t), u_0) \\ & \quad + \mu^1(t)b^1(t, x_0(t), u_0) + \mu^2(t)b^2(t, x_0(t), u_0) \\ & = \max_{u \in U} \{ -\psi(t)f(t, x_0(t), u) + \mu^0 b^0(t, x_0(t), u) \\ & \quad + \mu^1(t)b^1(t, x_0(t), u) + \mu^2(t)b^2(t, x_0(t), u) \}, \end{aligned}$$

where u_0 is any point in the support of \mathfrak{M}_0 at t .

Acknowledgment. The author would like to thank the referees, especially J. Warga, for careful reading and constructive criticism of earlier versions of this paper.

REFERENCES

- [1] A. F. FILIPPOV, *On certain questions in the theory of optimal control*, Vestnik Moskov. Univ. Ser. Mat. Meh. Astronom., 2 (1959), pp. 25–42; English transl., this Journal, 1 (1962), pp. 76–84.
- [2] R. A. GAMBILL, *Generalized curves and the existence of optimal controls*, this Journal, 1 (1963), pp. 246–260.
- [3] A. GHOUILA HOURI, *Sur la généralisation de la notion de commande d'un système guidable*, Rev. d'Information et de Recherche Operationelle, 4 (1967), pp. 7–32.
- [4] M. R. HESTENES, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.
- [5] J. L. KELLEY, I. NAMIOKA, et al., *Linear Topological Spaces*, Van Nostrand, Princeton, N.J., 1963.
- [6] E. J. MCSHANE, *Generalized curves*, Duke Math. J., 6 (1940), pp. 513–536.
- [7] ———, *Necessary conditions in generalized curve problems of the calculus of variations*, Ibid., 7 (1940), pp. 1–27.
- [8] ———, *Existence theorems for Bolza problems in the calculus of variations*, Ibid., 7 (1940), pp. 28–61.

- [9] ———, *Integration*, Princeton University Press, Princeton, N.J., 1944.
- [10] ———, *Relaxed controls and variational problems*, this Journal, 5 (1967), pp. 348–485.
- [11] E. J. MCSHANE AND R. B. WARFIELD, JR., *On Filippov's implicit functions lemma*, Proc. Amer. Math. Soc., 18 (1967), pp. 41–47.
- [12] L. W. NEUSTADT, *An abstract variational theory with applications to a broad class of optimization problems. I. General theory*, this Journal, 4 (1965), pp. 505–527.
- [13] ———, *An abstract variational theory with applications to a broad class of optimization problems. II. Applications*, this Journal, 5 (1967), pp. 90–138.
- [14] ———, *A general theory of extremals*, J. Comput. System Sci., 3 (1969), pp. 59–91.
- [15] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.
- [16] W. T. REID, *Ordinary Differential Equations*, John Wiley, New York, 1971.
- [17] A. B. SCHWARZKOPF, *Optimal controls for problems with a restricted state space*, this Journal, 10 (1972), pp. 487–511.
- [18] F. A. VALENTINE, *The problem of Lagrange with differential inequalities as side conditions*, Contributions to the Calculus of Variations 1933–37, University of Chicago Press, Chicago, 1937.
- [19] C. VİRSAN, *Necessary conditions for optimization problems with operational constraints*, this Journal, 8 (1970), pp. 527–558.
- [20] J. WARGA, *Minimizing variational curves restricted to a preassigned set*, Trans. Amer. Math. Soc., 112 (1964), pp. 432–455.
- [21] ———, *Unilateral variational problems defined by integral equations*, Michigan Math. J., 12 (1965), pp. 449–480.
- [22] ———, *Functions of relaxed controls*, this Journal, 5 (1967), pp. 628–641.
- [23] ———, *Unilateral minimax control problems defined by integral equations*, this Journal, 8 (1970), pp. 372–382.
- [24] ———, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
- [25] L. C. YOUNG, *Generalized curves and the existence of an attained absolute minimum in the calculus of variations*, Compt. Rend. Soc. Sci. et Lettres, Varsovie, Cl. III, 30 (1937), pp. 212–234.
- [26] ———, *Necessary conditions in the calculus of variations*, Acta Math., 69 (1938), pp. 239–258.
- [27] ———, *Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, 1969.

CONTROLLABILITY AND TENABILITY OF NONLINEAR SYSTEMS WITH STATE EQUALITY CONSTRAINTS*

A. B. SCHWARZKOPF†

Abstract. In this paper we consider the problem of maintaining a state equality constraint throughout an interval $[a, c]$ in spite of small perturbations of the dynamic and constraint equations. We obtain sufficient conditions for this to be possible in terms of the behavior of the associated linear variational systems. These results extend similar conditions for the local controllability of nonlinear systems obtained by Markus [2] and Yorke [7].

1. Introduction. In this paper we wish to examine certain controllability properties of a nonlinear differential system in terms of the controllability of the associated variational system. We approach the problem in the spirit of Markus [2] and Yorke [7], and our results are generalizations of Yorke's observations. The kind of controllability which we consider might perhaps be called "local controllability of a constraint along a solution," but rather than introduce another modification of the term controllable, we will introduce new terminology.

Let $T = [a, c]$, $G \subseteq R^n$, $U \subseteq R^p$ and let $f(t, x, u): T \times G \times U \rightarrow R^n$, $b(t, x, u): T \times G \times U \rightarrow R^m$ be bounded and continuous along with all partial derivatives with respect to x on an open neighborhood of their domain. Let $h(x_1, x_2): G \times G \rightarrow R^l$ be continuously differentiable on a neighborhood of its domain. Then the system

$$(1.1) \quad x(t) = x(a) + \int_a^t f(s, x(s), u(s)) ds,$$

with $x(t) \in G$, $u(t) \in U$, will be said to satisfy a *tenable* constraint

$$(1.2) \quad b(t, x(t), u(t)) = 0 \quad \text{a.e. in } T,$$

$$(1.3) \quad h(x(a), x(c)) = 0$$

if there exists a solution (x_0, u_0) of the system (1.1)–(1.3) and there exists a $\delta > 0$ such that if $\psi(t) \in L^\infty[T \rightarrow R^n]$, and $v \in R^l$ with $\|\psi(t)\|_\infty < \delta$, $\|v\| < \delta$, there exists a solution (x, u) to (1.1) such that

$$(1.4) \quad \int_a^t \{b(s, x(s), u(s)) - \psi(s)\} ds = 0, \quad t \in T,$$

$$(1.5) \quad h(x(a), x(c)) - v = 0.$$

The constraint will be called *approximately tenable* if given $\varepsilon > 0$, there is a $\delta > 0$ such that the left-hand sides of (1.4) and (1.5) can be kept within ε of zero for all ψ, v satisfying the above conditions. In general terms, a constraint is tenable if it can be maintained throughout the interval T in spite of small perturbations of the constraint equations.

* Received by the editors August 13, 1973, and in revised form April 25, 1974.

† Department of Mathematics, University of Oklahoma, Norman, Oklahoma 73069.

It is interesting to consider some special cases of this system. If $g(\mathbf{x}): G \rightarrow R^m$ is twice continuously differentiable, then the constraint

$$g(x(t)) = 0, \quad t \in T,$$

is equivalent to

$$\int_a^t \partial_x g(x(s)) f(s, x(s), u(s)) ds = 0, \quad g(x(a)) = 0,$$

and is therefore a special case of our system. The more usual controllability problem considered by Yorke [7] is obtained by omitting (1.2) and setting $h(x(a), x(c)) \equiv x(c)$. Another interesting problem is obtained by considering the system

$$x(t) = x(a) + \int_a^t w(s) ds,$$

where w is a measurable function with values in a ball W with radius larger than the bound for f , along with the constraints

$$\begin{aligned} \int_a^t \{f(s, x(s), u(s)) - w(s)\} ds &= 0, \\ \int_a^t b(s, x(s), u(s)) ds &= 0, \quad h(x(a), x(c)) = 0 \end{aligned}$$

where $U \otimes W$ is the new control set. Tenability of this system means that the constraint may be maintained at zero in spite of small perturbations of the driving system, as well as the constraint.

Suppose that (x_0, u_0) is a solution to the system (1.1)–(1.3). We will derive sufficient conditions for the constraints (1.2) and (1.3) to be tenable for the non-linear system in terms of the linear system

$$(1.6) \quad x(t) = x(a) + \int_a^t \{\partial_x f(s, x_0(s), u_0(s))x(s) + v(s)\} ds$$

and constraints

$$(1.7) \quad \int_a^t \{\partial_x b(s, x_0(s), u_0(s))x(s) + w(s)\} ds = 0,$$

$$(1.8) \quad \partial_{x_1} h(x_0(a), x_0(c))x(a) + \partial_{x_2} h(x_0(a), x_0(c))x(c) = 0,$$

where $(v(t), w(t))$ is in the convex hull of the set

$$(1.9) \quad \{(f(t, x_0(t), u), b(t, x_0(t), u)) | u \in U\}.$$

We will show (Theorem 2.6) that (1.2), (1.3) is an approximately tenable constraint for (1.1) with measurable controls $u(t)$, provided (1.7), (1.8) is tenable for (1.6). We can obtain tenability for (1.2), (1.3) if we admit relaxed controls.

The arguments in the sequel are based on the results of [4], which uses relaxed controls (chattering controls, sliding regimes). Rather than duplicate the arguments of that paper, we indicate several modifications which prove our results. Consequently we have maintained the notation of [4]. In particular, boldface letters \mathbf{t} , \mathbf{x} , \mathbf{u} indicate identity functions on the appropriate sets (i.e., T , G , U , respectively),

and the symbol $\mathcal{M}[\cdot; t]$ denotes a relaxed control function. The reader unfamiliar with relaxed controls is referred to [3] for an introduction (note also the additional references in [4]). Our decision to base the following arguments on relaxed controls is dictated by our desire to include problems with finite control sets U , and problems involving bang-bang controls. Similar arguments based on [5] yield the following result. Suppose U is convex. Then the constraint (1.2), (1.3) is tenable for (1.1) using ordinary measurable controls if the system

$$(1.10) \quad x(t) = x(a) + \int_a^t \{ \partial_x f(s, x_0(s), u_0(s))x(s) + \partial_u f(s, x_0(s), u_0(s))(u(s) - u_0(s)) \} ds$$

has tenable constraints:

$$(1.11) \quad \int_a^t \{ \partial_x b(s, x_0(s), u_0(s))x(s) + \partial_u b(s, x_0(s), u_0(s))(u(s) - u_0(s)) \} ds = 0,$$

$$(1.12) \quad \partial_{x_1} h(x_0(a), x_0(c))x(a) + \partial_{x_2} h(x_0(a), x_0(c))x(c) = 0.$$

Here $u(s) \in U$. We omit the proof.

2. Basic results. Continue the assumptions and notation introduced in § 1 and, further, suppose that G is bounded and open and U is compact. Let $\mathcal{C} = \{(x, \mathcal{M})\}$ be the set of all generalized curve solutions to the system (2.1) below, and suppose $C_0 = (x_0, \mathcal{M}_0) \in \mathcal{C}$ also satisfies (2.2) and (2.3). Set

$$(2.1) \quad x(t) = x(a) + \int_a^t \mathcal{M}[f(s, x(s), u); s] ds$$

and require that

$$(2.2) \quad \mathcal{M}[b(t, x(t), u); t] = 0 \quad \text{a.e. in } T,$$

$$(2.3) \quad h(x(a), x(c)) = 0.$$

Let $Y = \{(\eta, \Delta)\}$ be the linear span of

$$M = \{(\eta, \Delta) | \eta \in R^n, \|\eta\| \leq 1, \Delta = (\mathcal{M} - \mathcal{M}_0), \mathcal{M} \text{ a relaxed control}\}.$$

Note that M is a convex subset of Y . We wish to compare (2.1)–(2.3) to the system of variational equations (2.4)–(2.6) below:

$$(2.4) \quad Dx(t, y) = \eta + \int_a^t \{ \mathcal{M}_0[\partial_x \bar{f}(s)]Dx(s, y) + \Delta[\bar{f}(s)] \} ds,$$

with constraints

$$(2.5) \quad Db(t, y) \equiv \mathcal{M}_0[\partial_x \bar{b}(t)]Dx(t, y) + \Delta[\bar{b}(t)] = 0 \quad \text{a.e. in } T,$$

$$(2.6) \quad Dh(y) \equiv \partial_{x_1} \bar{h}Dx(a, y) + \partial_{x_2} \bar{h}Dx(c, y) = 0,$$

where y is in M . Here, as in [4], the overbar indicates that the designated function is evaluated at $x_0(t)$. Let $\tilde{M} = \bigcup_{\alpha > 0} \alpha M$.

LEMMA 2.1. *If the set $(Dh, Db(\mathbf{t}))(\tilde{M})$ contains a neighborhood of zero in $R^l \times L^\infty[T \rightarrow R^m]$, then the set*

$$(2.7) \quad P = \{\Delta[\bar{b}(\mathbf{t})] | y = (\eta, \Delta) \in M\}$$

has a nonempty interior in $L^\infty[T \rightarrow R^m]$.

Proof. This follows from the same arguments which establish [4, Thm. 1, Step 1, (2.13)].

LEMMA 2.2. *If the set $(Dh, Db(\mathbf{t}))(\tilde{M})$ contains a neighborhood of zero in $R^l \times L^\infty[T \rightarrow R^m]$, then there exists a positive number ε_1 such that for any ζ in R^l with $\|\zeta\| < \varepsilon_1$, there exists a $y(\zeta) = (\eta(\zeta), \Delta(\zeta))$ in M with $\Delta(\zeta)[\bar{b}(\mathbf{t})]$ bounded away from the complement of the set P in (2.7), and*

$$(Dh, Db(\mathbf{t}))(y(\zeta)) = (\zeta, 0).$$

Moreover, the functions $\eta(\zeta)$ and $\Delta(\zeta)[\phi(u); t]$ are Lipschitz continuous in ζ for fixed (t, ϕ) in $T \times C[T \rightarrow U]$.

Proof. Let $y_0 = (\eta_0, \Delta_0)$ be an element of M such that $\Delta_0[\bar{b}(\mathbf{t})] \in \text{int } P$. Since $0 \in \text{int } (Dh, Db(\mathbf{t}))(\tilde{M})$ implies that $(Dh, Db(\mathbf{t}))(\tilde{M}) = R^l \times L^\infty[Y \rightarrow R^m]$, there exists a point $y_* = (\eta_*, \Delta_*)$ in \tilde{M} such that $(Dh, Db(\mathbf{t}))(y_0 + y_*) = (0, 0)$. According to the definition of \tilde{M} , we can multiply y_0 and y_* by a constant from $[0, 1]$ so that $y_0, y_* \in M/3$ without changing the other properties above. Since P is convex, we also have that $(\Delta_0 + \Delta_*)[\bar{b}(\mathbf{t})]$ is in $\text{int } P$. Choose elements $v_i = (\eta_i, \Delta_i)$ from \tilde{M} , $i = 1, \dots, l + 1$ such that $Db(\mathbf{t}, v_i) = 0$ and the convex hull S of the points $Dh(v_i)$ contains a neighborhood of zero in R^l . We may suppose that $v_i \in M/3$, $i = 1, \dots, l + 1$. Thus we may express any point ζ in the simplex S as

$$\zeta = \sum_{i=1}^{l+1} \beta_i(\zeta) Dh(v_i) = Dh(y_0 + y_* + \sum_{i=1}^{l+1} \beta_i(\zeta) v_i),$$

where β_i are the barycentric coordinates of ζ and hence continuous (affine) in ζ , with $\beta_i(\zeta) \geq 0$, $\sum_{i=1}^{l+1} \beta_i(\zeta) = 1$. Moreover $(y_0 + y_* + \sum_{i=1}^{l+1} \beta_i(\zeta) v_i) \in M$ and

$$Db(\mathbf{t}, y_0 + y_* + \sum_{i=1}^{l+1} \beta_i(\zeta) v_i) = 0.$$

Now since P is convex,

$$(2.8) \quad \left(\Delta_0 = \Delta_* + \sum_{i=1}^{l+1} \beta_i(\zeta) v_i \right) [\bar{b}(t)] \in \text{int } P.$$

In fact, since for each t in T the points of the form (2.8) are all in the convex hull of a finite set of points, each of which is more than a fixed distance δ_1 from the complement of the set $\{\phi(t) | \phi \in P\}$, all of the functions obtained by replacing t by \mathbf{t} in (2.8) are also at least δ_1 from the complement of P .

LEMMA 2.3. *The set $(Dh, Db(\mathbf{t}))(\tilde{M})$ contains a neighborhood of zero in $R^l \times L^\infty[T \rightarrow R^m]$ if and only if $(Dh, Db(\mathbf{t}))(M)$ does.*

Proof. If $0 \in \text{int } (Dh, Db(\mathbf{t}))(M)$ then $0 \in \text{int } (Dh, Db(\mathbf{t}))(\tilde{M})$ since $M \subset \tilde{M}$. Conversely, if $(Dh, Db(\mathbf{t}))(\tilde{M})$ contains zero in its interior, then there exists an element $y(\zeta) \in M$ satisfying the conclusions of Lemma 2.2. Thus there exist vectors $w_i = (0, \Delta_i)$, $i = 1, \dots, m + 1$, in y such that $y(\zeta) + w_i \in M$ for all $i = 1, \dots, m + 1$ and the convex hull of the points $\Delta_i[\bar{b}(t)]$ contains a neighborhood of radius δ about zero in R^m for each $t \in T$. Thus given $\zeta(\mathbf{t}) \in L^\infty[T \rightarrow R^m]$, we can solve

$$(2.9) \quad \mathcal{M}_0[\partial_x \bar{b}(t)]_x + \Delta(\zeta)[\bar{b}(t)] + \sum_{i=1}^{m+1} \beta_i \Delta_i[\bar{b}(t)] = \zeta(t)$$

for nonnegative functions $\beta(t, x, \zeta)$ which are measurable in t for fixed $(x, \zeta) \in R^n \times R^l$, continuous in (x, ζ) for fixed t , and Lipschitz in x . Thus by [1, Thm. 4.1, p. 384], the equation

$$D\chi(t, \xi, \zeta)$$

$$= Dx(a, y(\zeta)) + \int_a^t \{ \mathcal{M}_0[\partial_x \bar{f}(s)] D\chi(s, \xi, \zeta) + \Delta(\zeta)[\bar{b}(s)] \\ + \sum_{i=1}^{m+1} \beta_i(s, D\chi(s, \xi, \zeta), \zeta) \Delta_i[\bar{b}(s)] \} ds$$

has a unique solution with $D\chi(t, 0, \zeta) = Dx(t, y(\zeta))$. Moreover, $D\chi(t, \xi, \zeta)$ converges uniformly to $Dx(t, y(\zeta))$ as $\|\xi\|_\infty \rightarrow 0$. It follows from (2.9) that for $\|\xi\|_\infty$ sufficiently small, the functions β_i are bounded by 1, and thus

$$\Delta(\xi, \zeta) = \Delta(\zeta)[\cdot; t] + \sum_{i=1}^{m+1} \beta_i(t, D\chi(t, \xi, \zeta), \zeta) \Delta_i[\cdot; t]$$

is a relaxed control such that

$$\mathcal{M}_0[\partial_x \bar{b}(t)] D\chi(t, \xi, \zeta) + \Delta(\xi, \zeta)[\bar{b}(t)] = \xi(t).$$

Furthermore, $D\chi(t, \xi, \zeta)$ is continuous in ζ , and since it converges to $Dx(t, y(\zeta))$ as $\|\xi\|_\infty \rightarrow 0$,

$$\partial_{x_1} \bar{h} D\chi(t, \xi, \zeta) + \partial_{x_2} \bar{h} D\chi(t, \xi, \zeta)$$

contains a neighborhood of radius ε_2 about zero for all ξ with $\|\xi\| < \delta$.

THEOREM 2.4. *If the set $(Dh, Db(\mathbf{t}))(\tilde{M})$ in $R^l \times L^\infty[T \rightarrow R^m]$ contains zero in its interior, then so does the set*

$$\{(h(x(a), x(c)), \mathcal{M}[b(\mathbf{t}, x(\mathbf{t}), u); \mathbf{t}]) | (x, \mathcal{M}) \in \mathcal{C}\}.$$

Proof. For notational simplicity, we will prove this theorem under the additional assumption that $l = 1$ and $m = 1$. The modifications necessary for extension to the general case will be indicated at the end of the proof. From Lemma 2.2 we see that there exists a number $\varepsilon_1 > 0$ and elements $y_1 = (\eta_1, \Delta_1)$, $y_2 = (\eta_2, \Delta_2)$ in M such that

$$Dh(y_1) = \varepsilon_1, \quad Dh(y_2) = -\varepsilon_1, \\ Db(\mathbf{t}, y_1) = 0, \quad Db(\mathbf{t}, y_2) = 0,$$

and both $\Delta_1[\bar{b}(\mathbf{t})]$ and $\Delta_2[\bar{b}(\mathbf{t})]$ are contained in the interior of P (see (2.7)). Thus there exists a number $\delta_1 > 0$ and elements $y_i^+ = (0, \Delta_i^+)$ and $y_i^- = (0, \Delta_i^-)$ in Y , $i = 1, 2$, such that

$$(2.10) \quad (\Delta_1^+ - \Delta_1)[\bar{b}(t)] = (\Delta_2^+ - \Delta_2)[\bar{b}(t)] = \delta_1 > 0, \\ (\Delta_1^- - \Delta_1)[\bar{b}(t)] = (\Delta_2^- - \Delta_2)[\bar{b}(t)] = -\delta_1 < 0,$$

and $(y_1 + y_1^+)$, $(y_1 + y_1^-)$, $(y_2 + y_2^+)$, $(y_2 + y_2^-)$ are all in M . Define

$$(2.11) \quad y(\tau) = (\tau\eta_1 + (1 - \tau)\eta_2, \tau\Delta_1 + (1 - \tau)\Delta_2) = (\eta(\tau), \Delta(\tau))$$

for $\tau \in [0, 1]$, and observe that

$$Dh(y(\tau)) = (2\tau - 1)\varepsilon_1 \in [-\varepsilon_1, \varepsilon_1], \quad Db(t, y(\tau)) = 0.$$

Fix $\alpha_0 > 0$, set

$$\Delta^*(\tau) = Ae^{B(t-a)}\Delta^+(\tau), \quad \Delta_*(\tau) = Ae^{B(t-a)}\Delta^-(\tau),$$

$$\Delta^\pm(\tau) = (\tau\Delta_1^\pm + (1 - \tau)\Delta_2^\pm),$$

and consider the equation

$$(2.12) \quad 0 = \{(\alpha e^\beta - \alpha_0)\Delta^*(\tau) + (\alpha e^{-\beta} - \alpha_0)\Delta_*(\tau) + (\alpha - \alpha_0)\Delta(\tau) + \mathcal{M}_0\} \\ \cdot [b(t, x, u); t] + (\alpha - \alpha_0)\xi.$$

This equation can be solved uniquely for $\beta(t, x, \alpha, \xi, \tau)$, where $t \in T$, $\tau \in [0, 1]$, $\alpha \in [0, 2\alpha_0]$, $\gamma \in [-1, 1]$, and x is in a neighborhood of radius δ_2 about $x_0(t)$. The function $\beta(t, x, \alpha, \xi, \tau)$ is measurable in t for fixed (x, α, ξ, τ) , continuously differentiable in (x, α, ξ, τ) for fixed t , and uniformly bounded along with its partial derivatives on its domain of definition. Moreover,

$$\partial_x \beta(t, x_0(t), \alpha_0, \xi, \tau) = \frac{-\mathcal{M}_0[\partial_x \bar{b}(t)]}{2\alpha_0 d(t)},$$

$$\partial_\alpha \beta(t, x_0(t), \alpha_0, \xi, \tau) = \frac{-\Delta(\tau)[\bar{b}(t)] + \xi}{2\alpha_0 d(t)},$$

$$\partial_\gamma \beta(t, x_0(t), \alpha_0, \gamma, \tau) = 0 = \partial_\tau \beta(t, x_0(t), \alpha_0, \xi, \tau).$$

where $d(t) \equiv Ae^{B(t-a)}\delta_1$. Let $\gamma(t) \in L^\infty[T \rightarrow [-1, 1]]$ and define

$$\mathcal{M}(\chi, \alpha, \gamma, \tau)[\cdot; t] = \{(\alpha e^{\beta(t, \chi, \alpha, \gamma, \tau)} - \alpha_0)\Delta^*(\tau) + (\alpha e^{-\beta(t, \chi, \alpha, \gamma, \tau)} - \alpha_0)\Delta_*(\tau) \\ + (\alpha - \alpha_0)\Delta(\tau) + \mathcal{M}_0\}[\cdot; t]$$

By [1, Thm. 4.1, p. 384] there exists a positive number σ such that if $|\alpha - \alpha_0| < \sigma$, then the equation

$$(2.13) \quad \chi(t) = x_0(a) + (\alpha - \alpha_0)\eta(\tau) + \int_a^t \mathcal{M}(\chi(s), \alpha, \gamma, \tau)[f(s, \chi(s), u); s] ds$$

has a unique solution $\chi(t) \equiv \chi(t, \alpha, \gamma, \tau)$ such that $\chi(t, \alpha_0, \gamma, \tau) = x_0(t)$ and there exists a constant C_1 such that

$$(2.14) \quad |\chi(t, \alpha, \gamma, \tau) - x_0(t)| \leq C_1 |\alpha - \alpha_0| \\ + C_1 \int_a^b \{|\alpha e^{\beta(s, x_0(s), \alpha, \gamma, \tau)} - \alpha_0|\Delta^*(\tau) + (\alpha e^{-\beta(s, x_0(s), \alpha, \gamma, \tau)} - \alpha_0)\Delta_*(\tau) \\ + (\alpha - \alpha_0)\Delta(\tau)\}|f(s, x_0(s), u); s| ds \\ \leq C_2 |\alpha - \alpha_0|,$$

where C_2 is a constant, perhaps larger than C_1 . The last inequality follows from (2.12) since $\beta > 0$ implies

$$(\alpha e^{\beta(t, x_0(t), \alpha, \gamma, \tau)} - \alpha_0) \leq d(t) \{ -(\alpha - \alpha_0)d(t) + (\alpha - \alpha_0)|\Delta(\tau)[\bar{b}(t)]| + (\alpha - \alpha_0)|\gamma| \},$$

and similarly for $\beta < 0$.

Now by [1, Thm. 6.1, p. 392] the function $\chi(t, \alpha, \gamma, \tau)$ is differentiable with respect to α , and if we set $\partial_\alpha \chi(t, \alpha_0, \gamma, \tau) \equiv D\chi(t, \gamma, \tau)$, we have

$$(2.15) \quad \begin{aligned} D\chi(t, \gamma, \tau) = \eta + \int_a^t \left\{ \mathcal{M}_0 \left[\partial_x \bar{f}(s) + \frac{(\Delta^*(\tau) - \Delta_*(\tau))[\bar{f}(s)]}{2\alpha_0 d(s)} \partial_x \bar{b}(s) \right] D\chi(s, \gamma, \tau) \right. \\ \left. + \Delta(\tau) \left[\bar{f}(s) + \frac{(\Delta^*(\tau) - \Delta_*(\tau))[\bar{f}(s)]}{2\alpha_0 d(s)} \bar{b}(s) \right] \right. \\ \left. - \frac{(\Delta^*(\tau) - \Delta_*(\tau))[\bar{f}(s)]}{2\alpha_0 d(s)} \gamma(s) + (\Delta^*(\tau) + \Delta_*(\tau))[\bar{f}(s)] \right\} ds, \end{aligned}$$

and the difference quotients converge uniformly for fixed (γ, τ) . It is tedious but not difficult to verify that the rate of convergence in (2.15) is independent of (γ, τ) . Using arguments similar to those which established [4, (2.27), (2.28) and (2.29)], we can show that

$$(2.16) \quad |D\chi(t, \gamma, \tau) - Dx(t, y(\tau))| \leq K_1 \{ \|\gamma(t)\|_\infty + (A/B) e^{B|c-a|} \}$$

for all $\tau \in [0, 1]$, and if $\beta > 0$, we have

$$(2.17) \quad \begin{aligned} |\partial_\alpha e^{\beta(t, \chi(t, \alpha_0, \gamma, \tau), \alpha_0, \gamma, \tau)}| &= |e^0 \partial_\alpha \beta(t, \chi(t, \alpha_0, \gamma, \tau), \alpha_0, \gamma, \tau)| \\ &\leq \frac{K_2}{2\alpha_0 d(t)} \{ |D\chi(t, \gamma, \tau) - Dx(t, y(t))| + \|\gamma\|_\infty \} \\ &\leq \frac{K_2}{2\alpha_0 A e^{B(t-a)}} \{ K_1 \{ \|\gamma(t)\|_\infty + (A/B) e^{B|c-a|} \} + \|\gamma_\infty\| \} \end{aligned}$$

so that for proper choices of $\|\gamma(t)\|_\infty$, A and B , this can be made arbitrarily small. A similar result holds for $\beta < 0$. Moreover, the difference quotients in (2.16) and (2.17) converge uniformly for all $\gamma(t)$ in $L^\infty[T \rightarrow [-1, 1]]$ and τ in $[0, 1]$. Thus (if $\beta > 0$)

$$\begin{aligned} \frac{d}{d\alpha} [\alpha e^{-\beta(t, \chi(t, \alpha, \gamma, \tau), \alpha, \gamma, \tau)} - \alpha_0]_{\alpha=\alpha_0} \\ = e^{-\beta(t, \chi(t, \alpha_0, \gamma, \tau), \alpha_0, \gamma, \tau)} [1 - \alpha_0 \partial_\alpha \beta(t, \chi(t, \alpha_0, \gamma, \tau), \alpha_0, \gamma, \tau)] \\ \geq 1 - \frac{\alpha_0 K_2 \{ K_1 \{ \|\gamma(t)\|_\infty + (A/B) e^{B|c-a|} \} + \|\gamma(t)\|_\infty \}}{2\alpha_0 A e^{B(t-a)}} > 0 \end{aligned}$$

for proper choices of $\|\gamma\|_\infty$, A and B . Thus $\mathcal{M}(\chi(t, \alpha_1, \gamma, \tau), \alpha_1, \gamma, \tau)[\cdot; t]$ in (2.13) is a legitimate relaxed control for some α_1 slightly larger than α_0 , all $\tau \in [0, 1]$, and $\gamma(t)$ in $L^\infty[T \rightarrow R]$ with $\|\gamma(t)\|_\infty$ sufficiently small. By (2.12) we have

$$(2.18) \quad \mathcal{M}(\chi(t, \alpha_1, \gamma, \tau), \alpha_1, \gamma, \tau)[b(t, \chi(t, \alpha_1, \gamma, \tau), u); t] = -(\alpha_1 - \alpha_0)\gamma(t)$$

for all $\tau \in [0, 1]$. On the other hand, by (2.16) and our choice of y_1 and y_2 , we may assume that for some $\varepsilon_2 > 0$

$$h(\chi(a, \alpha_1, \gamma, 1), \chi(c, \alpha_1, \gamma, 1)) \geq \varepsilon_2 > 0,$$

$$h(\chi(a, \alpha_1, \gamma, 0), \chi(c, \alpha_1, \gamma, 0)) \leq -\varepsilon_2 < 0$$

for all $\gamma(t)$ near 0. Thus given ζ in $[-\varepsilon_2, \varepsilon_2]$, there is a $\tau(\zeta)$ in $[0, 1]$ such that

$$(2.19) \quad h(\chi(a, \alpha_1, \gamma, \tau(\zeta)), \chi(c, \alpha_1, \gamma, \tau(\zeta))) = \zeta$$

and (2.18) holds. This proves the theorem for $l = 1, m = 1$.

To extend to arbitrary values of l and m , we must introduce functions $y_{1,k}$ and $y_{2,k}$, $k = 1, 2, \dots, l$ such that the k th coordinates of $Dh(y_{1,k})$ and $Dh(y_{2,k})$ are ε_1 and $-\varepsilon_1$, respectively, and zero everywhere else. Also introduce functions $\Delta_j^+(\tau)$, $\Delta_j^-(\tau)$, $j = 1, \dots, m$, such that the j th coordinate of $\Delta_j^+(\tau)[\bar{b}(t)]$ and $\Delta_j^-(\tau)[\bar{b}(t)]$ is identically δ_1 and $-\delta_1$, respectively, and zero otherwise. The argument then proceeds in a fashion similar to that above, introducing an m -dimensional vector function β which produces a relaxed control that satisfies (2.18) and (2.19) above.

COROLLARY 2.5. *If (1.7), (1.8) are tenable constraints for (1.6), then (1.2), (1.3) are tenable for (1.1) using relaxed controls, and approximately tenable for (1.1) using ordinary measurable controls.*

Proof. The system (1.6)–(1.8) is equivalent to the system (2.4)–(2.6) in the sense that if $x(t)$ is a solution to (1.6) with control values $(v(t), w(t))$, then $Dx(t, y) = x(t)$ for $y = (\eta, \Delta)$, $\eta \equiv x(a)$, $(u(t), w(t)) \equiv \Delta[(\bar{f}(t), \bar{b}(t))]$, and conversely. Tenableity of (1.7), (1.8) is equivalent to saying that $0 \in \text{int}(Dh, Db(t))(M)$, and by Lemma 2.3 this is equivalent to saying $0 \in \text{int}(Dh, Db(t))(\tilde{M})$. Thus Theorem 2.4 implies that (1.2), (1.3) is a tenable constraint for (1.1). Warga [6, Thm. 2.2] has shown that ordinary measurable controls (more precisely, piecewise constant controls) are weak*-dense in the collection of relaxed controls, and in particular there exist ordinary controls which solve (2.1) and give (2.2), (2.3) values arbitrarily close to their values with relaxed controls. This proves the approximate tenability of (1.2), (1.3).

Corollary 2.5 holds even though U might be a finite set or a set determining bang-bang controls. We do not even need to require that (x_0, \mathcal{M}_0) be an ordinary curve. On the other hand, we have only shown approximate tenability for (1.2) and (1.3). It is not possible to obtain tenability in this system with ordinary control functions without introducing additional hypotheses on U which assume the existence of “nearby” controls. We mentioned this in § 1. On the other hand, we are able to strengthen Corollary 2.5 sufficiently to include Yorke’s results in [7].

THEOREM 2.6. *If (1.7) and (1.8) are tenable constraints for (1.6), then (1.2), (1.3) are approximately tenable for (1.1) with ordinary piecewise constant controls. Moreover, we can insist that (1.3) holds identically.*

Proof. By hypothesis, $Dh(M)$ contains a neighborhood of zero, so that there exist elements $y_i = (\eta_i, \Delta_i)$, $i = 1, \dots, l + 1$, in M such that the convex hull of the points $Dh(y_i)$ contains a neighborhood of radius δ about zero. If $\Delta_i = (\mathcal{M}_i - \mathcal{M}_0)$ we have, by standard results, that

$$(2.20) \quad Dx(t, y_i) = \phi_0(t)\eta_i + \phi_0(t) \int_a^t \phi_0^{-1}(s) \mathcal{M}_i[\bar{f}(s)] ds - \phi_0(t) \int_a^t \phi_0^{-1}(s) \mathcal{M}_0[\bar{f}(s)] ds,$$

where

$$(2.21) \quad \phi_0(t) = I + \int_a^t \mathcal{M}_0[\partial_x \tilde{f}(s)] \phi_0(s) ds,$$

I being the $n \times n$ identity matrix. Since piecewise constant controls are weak*-dense in relaxed controls [6, Thm. 2.2] we may suppose that the controls \mathcal{M}_i are piecewise constant functions $v_i(\mathbf{t})$, $i = 1, \dots, l+1$, and consequently the first integral in (2.20) is a Riemann integral. Thus, for any $\varepsilon_1 > 0$, there exists a $\delta_1 > 0$ such that if $\Pi = \{a = s_0 < s_1 < \dots < s_N = c\}$ is a partition of $[a, c]$ with mesh less than δ_1 , then $Dh(y_i)$ differs from

$$(2.22) \quad \begin{aligned} & \partial_{x_1} \bar{h} \eta_i + \partial_{x_2} \bar{h} \left\{ \phi(c) \eta_i + \phi(c) \sum_{j=0}^{n-1} \phi^{-1}(s_j) f(s_j, x_0(s_j), v_i(s_j)) [s_{j+1} - s_j] \right. \\ & \quad \left. - \phi(c) \int_a^c \phi^{-1}(s) \mathcal{M}^0[\tilde{f}(s)] ds \right\} \end{aligned}$$

by less than ε_1 .

Let $\{\mathcal{M}_p\}_{p=1}^\infty$ be a sequence of piecewise constant right continuous controls given by control functions $u_p(\mathbf{t})$, respectively, and suppose that $\mathcal{M}_p \rightarrow \mathcal{M}_0$ in the weak* sense. For each p we obtain functions $x_p(\mathbf{t})$ by solving (2.1) with \mathcal{M}_p in place of \mathcal{M} , and $x_p(a) = x_0(a)$. Similarly we define $\phi_p(\mathbf{t})$ by replacing \mathcal{M}_0 and $x_0(\mathbf{t})$ in (2.21) with \mathcal{M}_p and $x_p(\mathbf{t})$, respectively, and observe that the weak*-convergence of \mathcal{M}_p to \mathcal{M}_0 shows that $x_p(\mathbf{t})$ and $\phi_p(\mathbf{t})$ converge uniformly to $x_0(\mathbf{t})$ and $\phi_0(\mathbf{t})$ as $p \rightarrow \infty$. Thus we may choose a sequence of partitions $\Pi_p = \{a = s_0 < \dots < s_{N(p)} = c\}$ of mesh less than δ_1 in such a way that $u_p(\mathbf{t})$ and $v_i(\mathbf{t})$, $i = 1, \dots, l+1$, are constant on every interval $[s_i, s_{i+1})$ and

$$(2.23) \quad \begin{aligned} & |\phi_p(c) \int_a^c \phi_p^{-1}(s) f(s, x_p(s), u_p(s)) ds \\ & - \phi_p(c) \sum_{j=0}^{N(p)-1} \phi_p^{-1}(s_j) f(s_j, x_p(s_j), u_p(s_j))| < \varepsilon_1. \end{aligned}$$

Let $\beta = (\beta^1, \dots, \beta^{l+1}) \in R^{l+1}$ be chosen so that $\beta^i \geq 0$ and $\sum_{i=1}^{l+1} \beta_i = 1$. For each $p = 1, 2, \dots$, and $\varepsilon > 0$, define the ordinary control $\mathcal{M}_p[\cdot; t, \beta, \varepsilon]$ with control $u_p(\mathbf{t}, \beta, \varepsilon)$ given by

$$(2.24) \quad u_p(t, \beta, \varepsilon) = \begin{cases} v_i(t), & t \text{ in the interval } S_{ij}, \\ u_p(t), & \text{otherwise,} \end{cases}$$

$$S_{ij} \equiv \left(s_j + \varepsilon[s_{j+1} - s_j] \sum_{k=1}^{i-1} \beta_k, s_j + \varepsilon[s_{j+1} - s_j] \sum_{k=1}^i \beta_k \right),$$

$$j = 0, \dots, N(p) - 1.$$

Let $x_p(t, \beta, \varepsilon)$ be the solution of (2.1) with $\mathcal{M}_p[\cdot; t, \beta, \varepsilon]$ replacing $\mathcal{M}[\cdot; t]$ and $x_p(a, \beta, \varepsilon) = x_0(a) + \varepsilon \sum_{i=1}^{l+1} \beta_i \eta_i$. This solution exists for small ε , is continuous in β , and converges to $x_p(\mathbf{t})$ as $\varepsilon \rightarrow 0$ uniformly in p , β and t . This last follows from

[1, Thm. 4.1, p. 384]. Furthermore,

$$\begin{aligned}
 D_p x(c, \beta) &\equiv \lim_{\varepsilon \rightarrow 0} \frac{x_p(c, \beta, \varepsilon) - x_p(c)}{\varepsilon} \\
 &= \phi_p(c) \sum_{i=1}^{l+1} \beta_i \eta_i + \phi_p(c) \sum_{i=1}^{l+1} \beta_i \sum_{j=0}^{N(p)-1} \phi_p^{-1}(s_j) f(s_j, x_p(s_j), v_i(s_j)) \\
 (2.25) \quad &\times [s_{j+1} - s_j] - \phi_p(c) \sum_{j=0}^{N(p)-1} \phi_p^{-1}(s_j) f(s_j, x_p(s_j), u_p(s_j)) \\
 &\times [s_{j+1} - s_j].
 \end{aligned}$$

A straightforward but tedious argument involving Gronwall's lemma and the uniform boundedness of $\partial_x f$ and f shows that the convergence in (2.25) is uniform in p . Thus we have

$$\begin{aligned}
 &\left| \frac{h(x_p(a, \beta, \varepsilon), x_p(c, \beta, \varepsilon))}{\varepsilon} - \sum_{i=1}^{l+1} \beta_i D h(y_i) \right| \\
 &\leq \left| \frac{h(x_p(a), x_p(c))}{\varepsilon} \right| \\
 &+ \left| \frac{h(x_p(a, \beta, \varepsilon), x_p(c, \beta, \varepsilon)) - h(x_p(a), x_p(c))}{\varepsilon} \right. \\
 &\quad \left. - \partial_{x_1} h(x_p(a), x_p(c)) D_p x(a, \beta) - \partial_{x_2} h(x_p(a), x_p(c)) D_p x(c, \beta) \right| \\
 &+ \left| \partial_{x_1} h(x_p(a), x_p(c)) D_p x(a, \beta) - \partial_{x_2} h(x_p(a), x_p(c)) D_p x(c, \beta) \right. \\
 &\quad \left. - \sum_{i=1}^{l+1} \beta_i D h(y_i) \right|.
 \end{aligned}$$

By (2.25) and the definition of $x_p(a, \beta, \varepsilon)$, we may fix $\varepsilon = \varepsilon_0$ so small that the second expression on the right in (2.26) is less than $\delta_1/4$ for all p , and for $\varepsilon = \varepsilon_0$ we may choose p sufficiently small that the first expression is also less than $\delta_1/4$. Using (2.22), (2.23) and (2.25), we see that the last expression can also be made less than $\delta_1/4$ for large p and small enough mesh in Π_p . Thus the function $h(x(a, \beta, \varepsilon_0), x(c, \beta, \varepsilon_0))$ defined for $\beta^i \geq 0$, $\sum \beta^i = 1$ covers an open set about the origin of radius $\varepsilon_0 \delta_1/4$. Since $x_p(t, \beta, \varepsilon)$ converges to $x_0(t)$ uniformly in t as $\varepsilon \rightarrow 0$, $p \rightarrow \infty$, the theorem follows.

3. Remarks. In [7], Yorke indicated three problems which he had not solved in his original work. The second asked if his controllability results hold for relaxed controls. This we have answered in the affirmative. He also asked whether these results hold for continuous (differentiable) control functions u . This we can also answer in the affirmative provided that U is convex and his equation L_Λ is replaced by (1.9). This follows by observing that the function β , obtained by solving the modified version of (2.12) implicitly, is continuous (differentiable) in t provided the right-hand side of (2.12) is.

REFERENCES

- [1] M. R. HESTENES, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.
- [2] L. MARKUS, *Controllability of nonlinear processes*, this Journal, 3 (1965), pp. 78–90.
- [3] E. J. MCSHANE, *Relaxed controls and variational problems*, this Journal, 5 (1967), pp. 438–485.
- [4] A. B. SCHWARZKOPF, *Relaxed control problems with state equality constraints*, this Journal, 13 (1975), pp. 677–694.
- [5] ———, *Optimal controls with equality state constraints*, J. Optimization Theory Appl., to appear.
- [6] J. WARGA, *Relaxed variational problems*, J. Math. Anal. Appl., 4 (1962), pp. 111–128.
- [7] J. A. YORKE, *The maximum principle and controllability of nonlinear systems*, this Journal, 10 (1972), pp. 334–338.

PROJECTIONS OF CONVEX PROGRAMS WITH UNATTAINED INFIMA*

ROBERT A. ABRAMS†

Abstract. Convex programs with closed objective function and closed feasible region are classified as degenerate if the objective function and the feasible region have a common direction of recession. For each degenerate program, a reduced form is defined by projecting the feasible region and the objective function epigraph on the orthogonal complement of the recession directions. A finite sequence of such reductions yields a nondegenerate problem for which the infimum is attained on a bounded set. Under a very mild condition the infimum of the reduced problem is equal to that of the original problem. It is shown that the objective and constraint functions of the "projected" problem may be obtained by calculating limits of the objective and constraint functions in the directions of recession. These results generalize the concept of degeneracy and reduction to canonical form which was originally developed for posynomial geometric programming.

1. Introduction. If a convex objective function is to be minimized over a closed but unbounded convex constraint region, the infimum may not be achieved, and if it is achieved, it may be achieved on an unbounded set. Following the terminology introduced by Duffin, Peterson and Zener in [3] for a special class of convex problems, we call such a program degenerate. The purpose of this paper is to extend the notion of degeneracy to a very general class of convex programs and to show how to reduce a degenerate program to an equivalent nondegenerate program, i.e., to one in which the infimum is attained on a bounded set. In the case of posynomial programs the results to be presented below reduce to those of [3], and in the case of quadratically constrained quadratic programs and l_p -approximation problems they reduce to and greatly simplify the results of [4].

We consider the problem

$$(1) \quad \begin{array}{ll} \text{Minimize} & f(x) \\ \text{subject to} & x \in C, \end{array}$$

where $f(x)$ is a closed convex function and C is a closed convex subset of finite-dimensional Euclidean space. If C and $f(x)$ have no direction of recession in common, it is known that the infimum of $f(x)$ over C will be attained. If, however, there is a common direction of recession, then the infimum may be achieved on an unbounded set or it may not be achieved, and the problem will be (by definition) degenerate. It will be shown (§ 3) that, under a very mild condition, an equivalent problem may be obtained by projecting the constraint set and the epigraph of the objective function on a subspace orthogonal to the directions of recession. If the new problem obtained is still degenerate, the projection process is repeated until a nondegenerate problem is obtained. The resulting problem should be

* Received by the editors July 5, 1973, and in revised form February 21, 1974.

† Department of Industrial Engineering and Management Sciences, The Technological Institute, Northwestern University, Evanston, Illinois 60201. This research was supported in part by the Air Force Office of Scientific Research under Grant 73-2516.

easier to solve because it will be in a space of smaller dimension than the original problem and because the infimum will be achieved on a bounded set.

Mathematical programs with unattained infima are most likely to arise when the original variables of a problem have undergone some kind of transformation as in the cases of the convex program resulting from a posynomial geometric program and the deterministic equivalent of a probabilistic program. For example, if the optimal value of a term in a posynomial program is zero, the corresponding convex program will have an unattained infimum. The computational value of the theory developed here will depend on the possibility of detecting and removing degeneracy before actually computing a solution. Many existing computer codes for posynomial programs do detect and remove degeneracy as an initial step in obtaining a solution. In the posynomial case the determination and removal of degeneracy is usually a by-product of the computation of an initial dual solution. Since posynomial problems are generally solved by dual methods, no additional effort is required to detect and remove degeneracy. It is likely that similar methods based on the extended notion of degeneracy will be useful for problems which are at least partially separable and include linear or polyhedral constraints.

The use of projections to reduce linear programs with unbounded feasible regions was studied by Charnes et al. in [2]. There it is shown, along with other results, that any linear program is equivalent to a linear program with a one-point feasible region. In [6], Rockafellar used a special case of the theory developed here to reduce a problem with an unbounded optimal set to a problem with a bounded optimal set.

In § 2 of this paper notation and definitions are given and a number of preliminary theorems are reviewed. In § 3, the reduced form of a degenerate program is presented and theorems relating the reduced and original problem are given. In § 4 it is shown that the present development reduces to the special cases considered in [3] and [4]. Also a simple example requiring a two-stage reduction is presented.

2. Notation, definitions and preliminaries.

2.1. Notation and definitions. R^n is n -dimensional Euclidean space.

For $x \in R^n$ and $y \in R^n$, (x, y) is the standard inner product of x and y , i.e., $(x, y) = y^T x$, and $\|x\| = (x, x)^{1/2}$.

The relative interior of a set C is denoted by $\text{ri } C$ and the closure by \bar{C} .

The epigraph of a function $f: C \rightarrow R$, where $C \subset R^n$, is

$$(2) \quad \text{epi } f = \{(x, \alpha) \in R^{n+1} : f(x) \leq \alpha, x \in C\}.$$

A function f defined on $C \subset R^n$ is convex if $\text{epi } f$ is a convex set or equivalently if C is convex and $f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)$ for all $0 \leq \lambda \leq 1$ and x and y in C .

A convex function is closed if $\text{epi } f$ is closed or equivalently if all its level sets are closed, i.e., if all sets $\mathcal{L}_\alpha = \{x : f(x) \leq \alpha\}$ for all $\alpha \in R$ are closed.

If $f: C \rightarrow R$ is convex, its domain of definition is extended to all of R^n by defining $f(x) = +\infty$ for $x \notin C$. Also define $\text{dom } f \equiv \{x : f(x) < \infty\}$.

A convex set C recedes in the direction of a vector y if for every $x \in C$, $x + \lambda y \in C$ for all $\lambda \geq 0$. To simplify terminology, such a vector y will be called a direction of recession of C . Directions of recession are sometimes referred to as directions of infinity, e.g., [7].

The recession cone of C , O^+C , is the set of all directions of recession of C .

A vector y is a direction of recession of a convex function $f(x)$ if y is a direction of recession of all nonempty level sets of f , i.e., of the sets $\mathcal{L}_\alpha = \{x: f(x) \leq \alpha\}$ for all $\alpha \in \mathbb{R}$. The set of all such vectors is O^+f .

For any $L \subset \mathbb{R}^n$, $L^\perp = \{x \in \mathbb{R}^n: (z, x) = 0 \text{ for } z \in L\}$ is the orthogonal complement of L and $L^* = \{x \in \mathbb{R}^n: (z, x) \geq 0 \text{ for } z \in L\}$ is the polar or dual cone of L .

2.2. Preliminaries. The following theorems contain all the information on directions of recession needed for this paper. Proofs may be found in [5].

THEOREM 2.1. *Let C be a closed convex set and suppose that for some $x \in C$, the ray $\{x + \lambda y: \lambda \geq 0\}$ is contained in C . Then y is a direction of recession of C .*

THEOREM 2.2. *The set of all directions of recession, O^+C , of a convex set C is a convex cone.*

THEOREM 2.3. *If y is a direction of recession of some nonempty level set \mathcal{L}_β of a closed convex function f , it is a direction of recession of every nonempty level set.*

THEOREM 2.4. *If y is a direction of recession of the convex function f , then $f(x + \lambda y)$ is a nonincreasing function of λ .*

THEOREM 2.5. *Let f be a closed convex function and D a closed convex set. If D and $f(x)$ have no common direction of recession, then the infimum of $f(x)$ over D is attained on a bounded set.*

3. Projections of degenerate convex programs. Consider problem (1), i.e., minimize $f(x)$ subject to $x \in C$, where $f(x)$ is a closed convex function and C is a closed convex set in \mathbb{R}^n . If y is a direction of recession of $f(x)$ and also a direction of recession of C , we say that y is a direction of recession of the problem (1), and that the recession cone of (1) is the set of all such y . If the recession cone of (1) contains a nonzero vector, then (1) is said to be *degenerate*. Otherwise it is said to be *canonical*. If a problem of the form (1) has a direction of recession, u , such that $\lim_{\lambda \rightarrow \infty} f(z + \lambda u) = -\infty$, then we say, following [4], that the problem is *totally degenerate*,¹ or, following the linear programming terminology, *unbounded*. In this case the problem has no finite infimum and hence has no solution. In what follows it is assumed that (1) is not totally degenerate.

For each degenerate problem of the form (1), a reduced problem will be defined by projecting the feasible region C and the epigraph of $f(x)$ onto the orthogonal complement of the recession cone of the problem. The reduced problem may be degenerate in which case the same projection procedure is repeated.

Let S be the recession cone of (1) and let P_{S^\perp} be the orthogonal projection on S^\perp , the orthogonal complement of S . Let $S_0 = \{(x, 0) \in \mathbb{R}^{n+1}: x \in S\}$ and let $P_{S_0^\perp}$ be the orthogonal projection on S_0^\perp . The objective function $h(x)$ of the reduced problem is defined by

$$(3) \quad \text{epi } h = \overline{P_{S_0^\perp}(\text{epi } f)}.$$

¹ This corresponds to total degeneracy as defined in [3], only if the objective function is taken to be the logarithm of the sum of the exponential functions.

The reduced problem obtained from (1) is

$$(4) \quad \begin{aligned} & \text{Minimize} \quad h(x) \\ & \text{subject to} \quad x \in \overline{P_{S^\perp} C}. \end{aligned}$$

The objective function and the feasible region of (4) are both closed and convex and hence (4) is of the same form as (1). Therefore if (4) is still degenerate, we may consider the problem as one defined in the space S^\perp rather than in R^n and the reduction process may be repeated. Since each reduction reduces the dimension of the feasible region by at least one, a canonical problem will be obtained after a sequence of at most n of the above reductions. The resulting problem will be called the canonical reduced problem.

In order to prove that the canonical reduced problem and the original problem have equal infima, we need the following lemma which gives sufficient conditions for the projection of an intersection to be equal to the intersection of the projections. The lemma is also important for another reason. Often a constraint set will be defined as an intersection of sets each defined by a class of functions, e.g., we may have convex and linear constraint functions in which case the feasible region is the intersection of the region defined by the convex constraints with the region defined by the linear constraints. When projecting the feasible region it is convenient to be able to project the sets separately. The following lemma permits this.

LEMMA 3.1. *Let C_1 and C_2 be convex sets. Let S be contained in $O^+ C_1 \cap O^+ C_2$. Then*

$$P_{S^\perp}(C_1 \cap C_2) = P_{S^\perp} C_1 \cap P_{S^\perp} C_2.$$

Proof. Let $\{v_1, \dots, v_m\}$ be a basis for the subspace generated by S and composed only of vectors in S .

Let $x \in P_{S^\perp} C_1 \cap P_{S^\perp} C_2$. Then there are scalars α_i , $i = 1, \dots, m$, and β_i , $i = 1, \dots, m$, such that $x + \sum_{i=1}^m \alpha_i v_i \in C_1$ and $x + \sum_{i=1}^m \beta_i v_i \in C_2$.

Since v_i , $i = 1, \dots, m$, are recession directions of both C_1 and C_2 , for any $\gamma_i \geq 0$ and $\delta_i \geq 0$,

$$x + \sum_{i=1}^m (\alpha_i + \gamma_i) v_i \in C_1 \quad \text{and} \quad x + \sum_{i=1}^m (\beta_i + \delta_i) v_i \in C_2.$$

Now we choose the γ_i and δ_i so that the vectors are equal and hence an element of $C_1 \cap C_2$. This can be done by letting

$$\gamma_i^* = \begin{cases} 0, & \beta_i \leq \alpha_i, \\ \beta_i - \alpha_i, & \alpha_i < \beta_i, \end{cases} \quad \text{and} \quad \delta_i^* = \begin{cases} 0, & \alpha_i \leq \beta_i, \\ \alpha_i - \beta_i, & \beta_i < \alpha_i. \end{cases}$$

Thus

$$x + \sum_{i=1}^m (\alpha_i + \gamma_i^*) v_i \in C_1 \cap C_2, \quad P_{S^\perp} \left(x + \sum_{i=1}^m (\alpha_i + \gamma_i^*) v_i \right) = x,$$

and $x \in P_{S^\perp}(C_1 \cap C_2)$.

The opposite inclusion is trivial and is true without the hypothesis regarding directions of recession.

THEOREM 3.2. *Suppose that (1) is not totally degenerate and that*

$$\text{ri}(\text{dom } f) \cap \text{ri } C \neq \emptyset.$$

Then the infimum of the canonical reduced problem corresponding to (1) is attained on a bounded set and it is equal to the infimum of (1).

Proof. The objective function and the feasible region of a canonical problem have no direction of recession in common. Hence by Theorem 2.5 the infimum is attained on a bounded set.

To prove the second part of the theorem we show that (1) and (4) have equal infima and that if $\text{ri}(\text{dom } f) \cap \text{ri } C \neq \emptyset$, then $\text{ri}(\text{dom } h) \cap \text{ri } \overline{P_{S^\perp} C} \neq \emptyset$. Since each problem in the sequence of reductions ending in a canonical problem is obtained from the previous one in exactly the same way, it will follow that all of the problems have equal infima.

If $w \in \text{ri } C \cap \text{ri}(\text{dom } f)$, then $P_{S^\perp} w \in \text{ri}(P_{S^\perp} C) \cap \text{ri}[P_{S^\perp}(\text{dom } f)]$. Hence $P_{S^\perp} w \in \text{ri}(\overline{P_{S^\perp} C})$ and $P_{S^\perp} w \in \text{ri}(P_{S^\perp} \text{dom } f)$. Since $P_{S^\perp}(\text{dom } f) = \text{dom } h$, it follows that $P_{S^\perp} w \in \text{ri}(\overline{P_{S^\perp} C}) \cap \text{ri}(\text{dom } h)$.

For any set $T \subset R^n$, define the indicator function $\delta(x, T)$ by $\delta(x, T) = 0$ for $x \in T$ and $\delta(x, T) = +\infty$ for $x \notin T$. It is clear that instead of considering problems (1) and (4), we may consider

$$(1') \quad \text{Minimize } F(x) \equiv f(x) + \delta(x, C)$$

and

$$(4') \quad \text{Minimize } H(z) \equiv h(z) + \delta(z, \overline{P_{S^\perp} C})$$

which have the same infima and feasible solutions as (1) and (4) respectively.

Letting $\gamma(Q) = \{(x, \alpha) \in R^{n+1} : x \in Q\}$ for any $Q \subset R^n$, it follows from the definition of $F(x)$ that

$$(5) \quad \text{epi } F = \text{epi } f \cap \gamma(C).$$

Using Lemma 3.1, (5) implies that

$$(6) \quad P_{S_0^\perp}(\text{epi } F) = P_{S_0^\perp}(\text{epi } f) \cap P_{S_0^\perp}[\gamma(C)].$$

Now $\text{ri}(\text{dom } f) \cap \text{ri } C \neq \emptyset$ implies $\text{ri}(\text{epi } f) \cap \text{ri}[\gamma(C)] \neq \emptyset$. Therefore

$$\text{ri}[P_{S_0^\perp}(\text{epi } f)] \cap \text{ri}[P_{S_0^\perp}[\gamma(C)]] \neq \emptyset$$

and thus the closure of the right-hand side of (6) is equal to the intersection of the closures. Therefore

$$\overline{P_{S_0^\perp}(\text{epi } F)} = \overline{P_{S_0^\perp}(\text{epi } f)} \cap \overline{P_{S_0^\perp}[\gamma(C)]}.$$

It is easily verified that

$$\overline{P_{S_0^\perp}[\gamma(C)]} = \gamma(\overline{P_{S^\perp} C})$$

and hence that

$$\overline{P_{S_0^\perp}(\text{epi } F)} = \overline{P_{S_0^\perp}(\text{epi } f)} \cap \gamma(\overline{P_{S^\perp} C}) = \text{epi } H.$$

Let α be the infimum of (4) and hence of (4'). Then for any $\varepsilon > 0$ there is a z^0 such that $H(z^0) - \alpha < \varepsilon/2$. Since

$$(z^0, H(z^0)) \in \text{epi } H = \overline{P_{S^\perp}(\text{epi } F)}$$

there is an $(x, \beta) \in \text{epi } F$ such that

$$(7) \quad \|(P_{S^\perp}x, \beta) - (z^0, H(z^0))\| < \varepsilon/2.$$

Thus

$$|\beta - H(z^0)| < \varepsilon/2 \quad \text{and} \quad |\alpha - \beta| < \varepsilon.$$

However $(x, \beta) \in \text{epi } F$ implies that $x \in C$ and that $f(x) \leq \beta \leq \alpha + \varepsilon$. Therefore $\inf_{x \in C} f(x) \leq \alpha$.

Let b be the infimum of (1). Then for any $\varepsilon > 0$ there is an $x \in C$ such that

$$|f(x) - b| < \varepsilon.$$

Because $P_{S^\perp}x \in \overline{P_{S^\perp}C}$ and $P_{S^\perp}(x, f(x)) \in \overline{P_{S^\perp}(\text{epi } f)}$, we have that

$$h(P_{S^\perp}x) \leq f(x) \leq b + \varepsilon.$$

Therefore the infimum of (4) is no greater than the infimum of (1). This completes the proof of the theorem.

Note that if the preceding theorem and proof were applied to the canonical reduced problem and the preceding problem of the reduction sequence, z^0 could be taken to be an optimal solution. Suppose an orthogonal basis for the space of the preceding problem is obtained by extending a basis for S^\perp . Then it follows from (7) that the components of z^0 will be close to the value of the corresponding components of an x for which $f(x)$, the objective function of the preceding problem, is close to its infimum. The same statement, but with z^0 a near optimal solution, holds for any two adjacent problems in the reduction sequence. Therefore with a properly constructed basis for R^n , the components of the optimal solution of the canonical reduced problem will be close to values of the corresponding components of near optimal solutions of the original problem (1).

Our next task is to give a convenient method for finding the function $h(z)$ and the projection of C when C is given implicitly by convex and linear constraint functions. In order to evaluate $h(z)$ we show that it may also be obtained by calculating the limits of $f(x)$ in the directions of recession of the problem. It is convenient to first prove two simple lemmas.

LEMMA 3.3. *Let C be convex in R^n . Let S be a convex cone contained in O^+C . Let $v \in \text{ri } S$. Then for any $x \in C$ there is an $M > 0$ such that $P_{S^\perp}x + \lambda v \in C$ for all $\lambda \geq M$.*

Proof. Let $x \in C$. Then $x = P_{S^\perp}x + z$, where z is an element of the subspace generated by S . Since S is a cone and contains the origin, the affine hull of S is the same as the subspace generated by S . Hence there is an M such that for $\lambda \geq M$ we have $v - z/\lambda \in S$, which implies $\lambda v - z \in S \subset O^+C$. Thus for $\lambda \geq M$,

$$x + (\lambda v - z) = P_{S^\perp}x + \lambda v \in C.$$

LEMMA 3.4. Let f be a closed convex function and let S be a convex cone contained in O^+f , the recession cone of f . Let $v \in \text{ri } S$. Then for each x , there is an $M > 0$ such that for $\lambda \geq M$,

$$f(P_{S^\perp}x + \lambda v) \leq f(x).$$

Proof. Fix x and consider the level set $\mathcal{L}_{f(x)}$, which being nonempty has recession cone O^+f . Since $x \in \mathcal{L}_{f(x)}$, by the previous lemma $P_{S^\perp}x + \lambda v \in \mathcal{L}_{f(x)}$ for all λ greater than some M . Thus $f(P_{S^\perp}x + \lambda v) \leq f(x)$ for all $\lambda \geq M$.

THEOREM 3.5. Let f be a closed convex function. Let S be a convex cone contained in O^+f . Let $v \in \text{ri } S$. Define the function $g: S^\perp \rightarrow R$ by $g(z) = \lim_{\lambda \rightarrow \infty} f(z + \lambda v)$. Then

$$\overline{\text{epi } g} = \overline{P_{S_0^\perp} \text{epi } f}.$$

Proof. It will first be shown that $\text{epi } g \supset P_{S_0^\perp}(\text{epi } f)$ and then that $\overline{\text{epi } g} \subset \overline{P_{S_0^\perp}(\text{epi } f)}$ from which the result will follow.

Let $(z, \alpha) \in P_{S_0^\perp}(\text{epi } f)$. Then for some x with $P_{S^\perp}x = z$ we have $(x, \alpha) \in \text{epi } f$. Thus $f(x) \leq \alpha$. From Lemma 3.4, $\lim_{\lambda \rightarrow \infty} f(z + \lambda v) \leq f(x) \leq \alpha$. Thus

$$g(z) \equiv \lim_{\lambda \rightarrow \infty} f(z + \lambda v) \leq \alpha$$

and $(z, \alpha) \in \text{epi } g$ which proves that $\text{epi } g \supset P_{S_0^\perp} \text{epi } f$.

Let $(z, \alpha) \in \text{epi } g$. Then there is a sequence $(z^i, \alpha^i) \in \text{epi } g$ with $(z^i, \alpha^i) \rightarrow (z, \alpha)$. Choose L_1 such that $i \geq L_1$ implies $\|(z^i, \alpha^i) - (z, \alpha)\| < \varepsilon/2$. Now

$$g(z^i) \equiv \lim_{\lambda \rightarrow \infty} f(z^i + \lambda v).$$

Choose L_2 such that $\lambda \geq L_2$ implies that $0 < f(z^i + \lambda v) - g(z^i) < \varepsilon/2$. For $i \geq L_1$ and $\lambda \geq L_2$, $f(z^i + \lambda v) < g(z^i) + \varepsilon/2 \leq \alpha^i + \varepsilon/2$ since $(z^i, \alpha^i) \in \text{epi } g$. Thus

$$(z^i + \lambda v, \alpha^i + \varepsilon/2) \in \text{epi } f$$

and therefore $(z^i, \alpha^i + \varepsilon/2) \in P_{S_0^\perp} \text{epi } f$. But $(z^i, \alpha^i + \varepsilon/2)$ is close to (z, α) , i.e.,

$$\begin{aligned} \|(z, \alpha) - (z^i, \alpha^i + \varepsilon/2)\| &\leq \|(z, \alpha) - (z^i, \alpha^i)\| + \|(z^i, \alpha^i) - (z^i, \alpha^i + \varepsilon/2)\| \\ &\leq \varepsilon/2 + \varepsilon/2 = \varepsilon \quad \text{for } i \geq L_1. \end{aligned}$$

Therefore (z, α) is a limit point of $P_{S_0^\perp} \text{epi } f$ and hence $\overline{\text{epi } g} \subset \overline{P_{S_0^\perp}(\text{epi } f)}$, thus completing the proof.

The function $h(z)$ in (4) is thus the closure of the function $g(z)$ defined in Theorem 3.5.

With the aid of Theorem 3.5 the function $h(z)$ can be calculated by taking limits. The following theorem shows that the "projected" constraint functions may be calculated in a similar manner if a point exists at which the constraints are satisfied as strict inequalities.

THEOREM 3.6. Let $g: R^n \rightarrow R^m$ be convex and let $C = \{x: g(x) \leq 0\}$. Let S be a convex cone contained in the recession cone of C and let $v \in \text{ri } S$. Suppose there is an x^0 such that $g(x^0) < 0$. Then

$$\{z \in S^\perp: \lim_{\lambda \rightarrow \infty} g(z + \lambda v) \leq 0\} = \overline{P_{S^\perp} C}.$$

Proof. We shall show that

$$(8) \quad P_{S^\perp}C \subset \left\{ z \in S^\perp : \lim_{\lambda \rightarrow \infty} g(z + \lambda v) \leq 0 \right\} \subset \overline{P_{S^\perp}C}$$

from which the result will follow.

To prove the first inclusion let $z \in P_{S^\perp}C$. Then from Lemma 3.3, $z + \lambda v \in C$ for all $\lambda \geq 0$ greater than some M . Therefore $g(z + \lambda v) \leq 0$ for $\lambda \geq M$, implying that $\lim_{\lambda \rightarrow \infty} g(z + \lambda v) \leq 0$, and the first inclusion follows.

To prove the second inclusion in (8), let $z \in S^\perp$ and assume

$$\lim_{\lambda \rightarrow \infty} g(z + \lambda v) \leq 0.$$

We shall show that given $\varepsilon > 0$, we can find a $w \in C$ (and a $\lambda \geq 0$) such that $\|w - (z + \lambda v)\| < \varepsilon$. Since $\|P_{S^\perp}w - z\| = \|P_{S^\perp}[w - (z + \lambda v)]\| \leq \|w - (z + \lambda v)\| < \varepsilon$, it will follow that z is an element of the closure of $P_{S^\perp}C$.

By hypothesis $g(x^0) < 0$ and we assume for some $\alpha > 0$, $g(x^0) \leq -\alpha e$, where $e = (1, 1, \dots, 1)^T$. Since v is a direction of recession of every level set of $g(x)$, $g(x^0 + \lambda v) \leq -\alpha e$ for all $\lambda \geq 0$.

Because $\lim_{\lambda \rightarrow \infty} g(z + \lambda v) \leq 0$, for any $\delta > 0$ there is a $\lambda^0(\delta)$ such that for $\lambda \geq \lambda^0(\delta)$,

$$(9) \quad g(z + \lambda v) < \delta e.$$

Let $M = \|x^0 - z\|$ and let $\varepsilon > 0$. Choose a positive β^0 less than ε/M , and let $\delta^0 = \beta^0/(1 - \beta^0)\alpha$. Define

$$w = \beta^0(x^0 + \lambda v) + (1 - \beta^0)(z + \lambda v)$$

for some fixed $\lambda > \lambda^0(\delta^0)$. Then

$$\begin{aligned} g(w) &\leq \beta^0 g(x^0 + \lambda v) + (1 - \beta^0)g(z + \lambda v) \\ &\leq -\beta^0 \alpha e + (1 - \beta^0)\delta^0 e = [-\beta^0 \alpha + \beta^0 \alpha] e = 0. \end{aligned}$$

Thus $w \in C$ and

$$\begin{aligned} \|w - (z + \lambda v)\| &= \|(z + \lambda v) + \beta^0[(x^0 + \lambda v) - (z + \lambda v)] - (z + \lambda v)\| \\ &= \beta^0 M < \varepsilon \end{aligned}$$

which completes the proof.

Remark. The assumption of a feasible point x^0 with $g(x^0) < 0$ is necessary as shown by the function defined by $g(x) = (x_2 - \alpha)^4/x_1^2$ for $x_1 \geq 1$ and by $g(x) = +\infty$ for $x_1 < 1$. The vector $\begin{pmatrix} 1 \\ 0 \end{pmatrix} \equiv v$ is a recession direction and clearly $\lim_{\lambda \rightarrow \infty} g(x + \lambda v) = 0$ for any fixed x_2 . However, the set $\{x : g(x) \leq 0\}$ is the line $x_2 = \alpha$ so the conclusion of Theorem 3.6 does not hold.

In some problems of interest, $\lim_{\lambda \rightarrow \infty} g(z + \lambda v)$ is strictly less than zero or $-\infty$ for all $z \in S^\perp$. In these cases the assumption of an interior point required in Theorem 3.6 is satisfied automatically. It then follows from Theorem 3.6 that if

$\lim g(z + \lambda v) < 0$ for all $z \in S^\perp$, then $\overline{P_{S^\perp} C} = S^\perp$ and, since the reduced problem is defined in S^\perp , the constraint may simply be dropped.

4. Examples.

4.1. Posynomial geometric programming [3]. Define a posynomial

$$\sum_i u_i(t) = \sum_i c_i \prod_{j=1}^m t_j^{a_{ij}},$$

where all c_i are positive and consider the problem

$$\begin{aligned} & \text{Minimize} \quad \sum_{i \in J(0)} u_i(t) \\ (10) \quad & \text{subject to} \quad \sum_{i \in J(k)} u_i(t) \leq 1, \quad k = 1, \dots, n, \\ & \quad t_j > 0, \quad i = 1, \dots, m, \end{aligned}$$

where the index sets $J(k)$, $k = 0, 1, \dots, n$, are defined so that $J(0) = \{1, 2, \dots, m_0\}$, $J(1) = \{m_0 + 1, \dots, m_1\}$, \dots , $J(n) = \{m_{n-1} + 1, \dots, m_n\}$ and $m_n = n$. Making the change of variables $t_j = e^{z_j}$ and $x_i = \sum a_{ij} z_j$, (10) becomes

$$\begin{aligned} & \text{Minimize} \quad \sum_{i \in J(0)} c_i e^{x_i} \\ (11) \quad & \text{subject to} \quad \sum_{i \in J(k)} c_i e^{x_i} \leq 1, \quad k = 1, \dots, K, \\ & \quad x \in R(A) \equiv \text{range of } A, \end{aligned}$$

where A is the matrix of exponents $\{a_{ij}\}$. Because $R(A)$ is a subspace, any $\bar{x} \in R(A)$ is a direction of recession of $R(A)$. A vector \bar{x} is a recession direction of the set defined by

$$\left\{ x: \sum_{i \in J(k)} c_i e^{x_i} \leq 1, k = 1, \dots, n \right\}$$

if and only if $x_i \leq 0$, $i \in J(k)$, $k = 1, \dots, n$. A vector \bar{x} is a recession direction of the objective function if and only if $\bar{x}_i \leq 0$, $i \in J(0)$. Thus \bar{x} is a direction of recession of the problem (11) if and only if $\bar{x} \in R(A)$ and $\bar{x}_i \leq 0$, $i = 1, \dots, n$, and the recession cone of (11) is the intersection of $R(A)$ and the nonpositive orthant. To apply Theorems 3.5 and 3.6 we must find a v in the relative interior of the recession cone. Such a v is characterized in the following lemma.

LEMMA 4.1. *Let S be the recession cone of (11) and let $v \in S$. Then $v \in \text{ri } S$ if and only if the number of negative components is maximal.*

Proof. Suppose $v \in S$ and $u \in S$ with $u_i < 0$ for some i with $v_i = 0$. (We can assume $v_i < 0$ implies $u_i < 0$ since otherwise u can be replaced with $u + v$.) Consider the line segment starting at u and extending slightly past v , namely,

$$u + \lambda(v - u) \quad \text{for } 1 < \lambda < 1 + \varepsilon.$$

The i th component of this point will be positive and hence it will not be an element of S . Hence v is not a relative interior point of S .

Now suppose $v \in S$ but $v \notin \text{ri } S$. Then for any $w \in \text{ri } S$,

$$(12) \quad w + (1 + \varepsilon)(v - w) \notin S \quad \text{for all } \varepsilon > 0.$$

Since w and v are both elements of $R(A)$, the linear combination (12) is also. Therefore some component of (12) is positive. Thus for some i , $w_i + (1 + \varepsilon)(v_i - w_i) = v_i + \varepsilon(v_i - w_i) > 0$ for all $\varepsilon > 0$. Hence $v_i = 0$ and $w_i < 0$ which shows that v_i has fewer nonzero components than $v + w$. The lemma is proved.

Therefore we seek a nonpositive $v \in R(A)$ with a maximum number of negative components. Such a vector, which will be denoted by \bar{v} , may be easily found by solving a finite sequence of linear programming problems. With Theorems 3.5 and 3.6, \bar{v} may be used to determine the objective and constraint functions of the reduced problem. Let $T = \{i: \bar{v}_i = 0\}$ and note that

$$\lim_{\lambda \rightarrow \infty} \sum_{i \in J(0)} c_i e^{x_i + \lambda \bar{v}_i} = \sum_{i \in J(0) \cap T} c_i e^{x_i}$$

for any $x \in R^n$. The exponential constraint functions are reduced similarly and the reduced problem becomes

$$(13) \quad \begin{aligned} & \text{Minimize} \quad \sum_{i \in J(0) \cap T} c_i e^{x_i} \\ & \text{subject to} \quad \sum_{i \in J(k) \cap T} c_i e^{x_i} \leq 1, \quad i = 1, \dots, n, \\ & \quad \quad \quad x \in P_{S^\perp}[R(A)]. \end{aligned}$$

Problem (13) is canonical and therefore no further reductions are required.

Note that the objective function and the exponential constraints depend only on x_i for $i \in T$. Write any vector $x \in R(A)$ as a sum of vectors in the subspace generated by $R(A) \cap R_-^n$ and $[R(A) \cap R_-^n]^\perp$, i.e., let

$$x = z + y \quad \text{with} \quad z \in [R(A) \cap R_-^n]^\perp, \quad \text{and} \quad y \in [R(A) \cap R_-^n]^{\perp\perp}.$$

Then z is the desired projection of x onto $[R(A) \cap R_-^n]^\perp$. Because y may be written as a linear combination of points in $R(A) \cap R_-^n$, $y_i = 0$ for $i \in T$. Therefore

$$x_i = z_i \quad \text{for } i \in T.$$

Since the objective function and constraint functions depend only on these components ($i \in T$), we may replace $x \in P_{S^\perp}(R(A))$ by $x \in R(A)$ and then for each $i \notin T$, the i th row of the matrix A may be eliminated. The resulting reduced problem is then as presented in [3]. However, in [3], the special structure of the problem is used to obtain the reduced form without the assumption of a "strict interior" point as required by Theorem 3.6.

4.2. Quadratic programming and l_p -approximation problems. In [4] quadratically constrained quadratic programming problems and l_p -constrained l_p -approximation are both shown to be transformable to problems of the form

$$(14) \quad \begin{aligned} & \text{Minimize} \quad G_0(x) \\ & \text{subject to} \quad G_k(x) \leq 0, \\ & \quad \quad \quad x \in P, \end{aligned}$$

where P is a subspace of R^n and

$$G_k(x) = \sum_{[k]} \frac{1}{p_i} |x_i - b_i|^{p_i} + x_{[k]} - b_{[k]},$$

$$[k] = \{m_k, m_{k+1}, \dots, n_k\}, \quad k = 0, 1, \dots, r,$$

$$]k[= n_{k+1}, \quad k = 0, 1, \dots, r,$$

$$m_0 = 1, \quad m_1 = n_0 + 2, \dots, m_r = n_{r-1} + 2, \quad n_r + 1 = n.$$

b is a fixed vector in R^n and $p_i > 1$.

In order for $v \in R^n$ to be a recession direction of the problem, it is necessary that for any feasible point x^0 , $x^0 + sv$ remain feasible for $s > 0$ and that the objective function not increase as s increases. Therefore $v \in P$ is a recession direction of (14) if and only if

$$v_i = 0, \quad i \in [k], \quad k = 0, 1, \dots, r,$$

$$v_{[k]} \leq 0, \quad k = 0, 1, \dots, r.$$

The relative interior of the recession cone of (14) is characterized in a manner similar to that of (11). A vector v in the recession cone of (14) is an element of the relative interior if and only if the number of negative components is maximal. The proof of this statement is analogous to that of Lemma 4.1 and is therefore omitted.

To determine the reduced form of (14), we seek a nonpositive vector $v \in P$ with $v_i = 0$, $i \in [k]$, $k = 0, 1, 2, \dots, r$, and with as many as possible of the components $v_{[k]}$, $k = 0, 1, \dots, r$, less than zero. Let \bar{v} be such a vector. If $v_{[0]} < 0$, the objective function $g_0(x)$ may be made as small as desired using the feasible vector $(x^0 + \lambda \bar{v})$, $\lambda > 0$, where x^0 is an arbitrary feasible vector. Thus the problem does not have a finite infimum and hence is totally degenerate.

Computing the feasible region of the reduced problem is particularly simple in this case, since

$$\lim_{\lambda \rightarrow \infty} G_k(x + \lambda v) = \begin{cases} G_k(x) & \text{if } \bar{v}_{[k]} = 0, \\ -\infty & \text{if } \bar{v}_{[k]} < 0. \end{cases}$$

Therefore it follows from Theorem 3.6 that the constraints of the canonical reduced problem are obtained by deleting the constraints corresponding to negative components of \bar{v} and leaving the others unaltered. An argument similar to that of § 4.1 shows that the projection of vector $x \in P$ leaves the components corresponding to $\{i: \bar{v}_i = 0\}$ unchanged. The reduced problem is independent of x_i for $i \in \{]k[\cup [k]: \bar{v}_{[k]} < 0\}$. Therefore if P is given as the range of a matrix A , the subspace for the reduced problem may be obtained simply by deleting rows of A corresponding to $i \in \{]k[\cup [k]: \bar{v}_{[k]} < 0\}$.

A fairly lengthy proof is required to show that the above canonical reduced form is the same as that obtained in [4]. However, once this has been done the final part of the Key Theorem 3.1 of [4] follows directly from the above definition of degeneracy, and the original long and complicated proof is no longer necessary.

A proof of the equivalence of the canonical reduced forms and a more thorough discussion of the preceding remark is given in [1].

4.3. An example requiring a two-stage reduction. Consider the following problem in R^2 :

$$(15) \quad \begin{array}{ll} \text{Minimize} & 1/x_1 \\ \text{subject to} & x_1^2 - x_2 \leq 0, \quad 1 - x_1 \leq 0. \end{array}$$

The recession cone of (15) is the ray generated by $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$. Using Theorem 3.5 for the objective function, and Theorem 3.6 for the constraints, the reduced problem defined in S^\perp becomes

$$(16) \quad \begin{array}{ll} \text{Minimize} & 1/x_1 \\ \text{subject to} & 1 - x_1 \leq 0. \end{array}$$

Problem (16) also has a nonzero recession cone, namely, the ray generated by $x_1 = 1$. A second reduction gives the canonical reduced form of minimizing the zero function subject to no constraints.

5. Conclusions. It has been shown that any convex programming problem with closed objective function and feasible region can be reduced to a canonical problem for which the infimum is attained on a bounded set. The reduction process is accomplished by projection of the objective function and feasible region on the orthogonal complement of the recession directions of the problem or equivalently by calculating limits of the objective and constraint functions in the direction of recession.

In the two special cases, § 4.1 and § 4.2, in which the reduction procedure was already known, it is the special structure of the problem which allows easy detection of degeneracy and reduction to canonical form. In both cases, with the exception of the linear constraint, the problems are completely separable and some or all of the terms are nonincreasing in their variables. It appears probable that straightforward reduction to canonical form will be possible for many classes of problems with similar structure, e.g., linear equality or inequality constraints in addition to monotonicity along a ray in other constraints and the objective function.

Acknowledgment. The author wishes to thank Professor Adi Ben. Israel for several helpful discussions during this research.

REFERENCES

- [1] R. A. ABRAMS, *Degeneracy in quadratic programming and l_p -approximation*, Department of Industrial Engineering Rep., Northwestern Univ., Evanston, Ill., 1972.
- [2] A. CHARNES, W. W. COOPER AND G. L. THOMPSON, *Some properties of redundant constraints and extraneous variables in direct and dual linear programming problems*, *Operations Res.*, 10 (1962), pp. 711-723.
- [3] R. J. DUFFIN, E. L. PETERSON AND C. ZENER, *Geometric Programming*, John Wiley, New York, 1967.

- [4] E. L. PETERSON AND J. G. ECKER, *Geometric programming: Duality in quadratic programming and l_q -approximation. III (Degenerate programs)*, J. Math. Anal. Appl., 29 (1970), pp. 365–383.
- [5] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, N.J., 1970.
- [6] ———, *Helly's theorem and minima of convex functions*, Duke Math. J., 32 (1965), pp. 381–397.
- [7] J. STOER AND C. WITZGALL, *Convexity and Optimization on Finite Dimensions. I*, Springer-Verlag, New York, 1970.

GENERALIZED RANDOM PROCESSES: A THEORY AND THE WHITE GAUSSIAN PROCESS*

HIROSHI INABA† AND BYRON D. TAPLEY‡

Abstract. A class of generalized random processes is defined in the framework of vector-valued distributions. Then two subclasses of generalized random processes are defined and their relationships are discussed. The so-called white Gaussian process is defined as a generalized random process, and its characterization is discussed.

1. Introduction. Shortly after L. Schwartz's work on distributions [1], Ito [2], and Gel'fand [3], [4] introduced independently the concept of generalized random processes (random distributions in Ito's definition). Their definitions were based on the work by Schwartz, and the processes were studied primarily as they relate to the correlation analysis of stationary generalized random processes. In principle, the generalized random process is defined to be a continuous linear mapping from a topological vector space of test functions to a topological vector space of random variables. Therefore, as presented in [5, Appendix], there are a number of possible classes of generalized random processes, depending upon the choice of the space of test functions and the choice of the space of random variables. However, it should be noted also that a generalized random process has been discussed by Urbanik [6] and Hida [7] from different viewpoints. Urbanik's definition is somewhat different from that mentioned above. However, the class defined by Hida [7, p. 73] can be shown to be a subclass of that defined by Gel'fand and Vilenkin [4, p. 243]. For a detailed discussion of this relation, see [8, pp. 7–11].

In the present investigation, we will introduce a class of generalized random processes, study a particular subclass as the basis for various engineering applications of the generalized random processes used in this study in a way somewhat different from the ordinary one, and discuss a stochastic integral (the Bochner integral) which plays an essential role in the following discussion. In §3 we will discuss some of the relationships between random processes and generalized random processes, and then the derivative of generalized random processes will be discussed. In §4 we will study the space of generalized random processes defined on the space $L^2(T)$ of square integrable functions on an interval T . Section 5 will then define the correlation operator of a generalized random process, which is a generalized notion of correlation functions of random processes. Finally, in §6 we will define and study the so-called white Gaussian process and the white process in the framework of generalized random processes.

* Received by the editors September 20, 1973. This investigation was supported by the Air Force Office of Scientific Research under Grant 72-2233.

† Department of Aerospace Engineering and Engineering Mechanics, University of Texas at Austin, Austin, Texas. Now at Department of Electrical Communications Engineering, Tokyo Denki Daigaku (Tokyo Denki Institute of Technology), Tokyo 101, Japan.

‡ Department of Aerospace Engineering and Engineering Mechanics, University of Texas at Austin, Austin, Texas 78712.

In this investigation, we will not discuss the applications of the generalized random processes to engineering problems (for example, the innovation process [11], the linear stochastic differential equation [14], the Kalman–Bucy filter [12], etc.). These will be studied in separate papers. We also note that the argument developed in §§ 4 and 5 can easily be extended to more general spaces of generalized random processes (see [8, Chaps. IV and V]).

2. Basic definitions. Throughout this investigation it is assumed that a probability space (Ω, A, P) is given and fixed, where Ω is a set, A is a σ -algebra of subsets of Ω and P is a probability measure on the measurable space (Ω, A) . The expectation operator is denoted by E . Further, let Z denote the Hilbert space of all P -equivalence¹ classes of real random variables defined on (Ω, A, P) having finite second moments, where the inner product $(\cdot, \cdot)_Z$ and norm $\|\cdot\|_Z$ are given, respectively, by

$$(2.1) \quad (\xi, \eta)_Z = E(\xi\eta) \quad \text{and} \quad \|\xi\|_Z = (\xi, \xi)_Z^{1/2}.$$

The space Z is assumed to be separable, and hence there is a complete orthonormal system $\{\xi_i\}$ of Z , and every ξ in Z can be expressed uniquely as

$$\xi = \sum_{i=1}^{\infty} a_i \xi_i,$$

where $a_i = (\xi, \xi_i)_Z$. Furthermore, let R denote the real line, T an interval in R , $M(T)$ the σ -algebra of Lebesgue measurable sets in T , and, finally, λ the Lebesgue measure. With these notations, the random process used in this study is defined as follows.

DEFINITION 2.1. A function $x: T \rightarrow Z$ is called a *finitely-valued random process on T* if there is a finite number of disjoint sets S_1, S_2, \dots, S_n in $M(T)$ with $\lambda(S_i) < \infty$ such that $x(t)$ is a constant on each S_i and $x(t) = 0$ on $T - \bigcup_{i=1}^n S_i$. \square

DEFINITION 2.2. A function $x: T \rightarrow Z$ defined for almost every t in T is called a *random process on T* if there exists a sequence $\{x_j\}$ of finitely-valued random processes on T such that for almost every² t in T $\{x_j(t)\}$ converges in Z to $x(t)$. \square

The above definition seems to be too restrictive. However, it is wide enough to cover most of the interesting processes appearing in both theory and application. For example, any function $x: T \rightarrow Z$ which is continuous in the mean-square sense is a random process on T in the sense of Definition 2.2, and furthermore, any $(M(T) \times A)$ -measurable function $x: T \times \Omega \rightarrow R$ which is square-integrable over $T \times \Omega$ is a random process in the above sense. (For the proof of this statement, see [8, Appendix 7].)

Continuity and differentiability of a random process of the above type can be defined in the usual manner. Namely, a random process x on T is said to be *continuous over T* if, for every t in T ,

$$(2.2) \quad \lim_{h \rightarrow 0} \|x(t+h) - x(t)\|_Z = 0,$$

¹ Two random variables ξ and η are said to be P -equivalent if $P(\{\omega \in \Omega: \xi(\omega) \neq \eta(\omega)\}) = 0$. For an arbitrary measure μ , μ -equivalence is defined analogously.

² Hereafter “almost every” will be abbreviated by *a.e.* Sometimes *a.e.* will mean “almost everywhere”.

and *differentiable* over T if, for every t in T , there exists a random variable $y(t)$ in Z such that

$$(2.3) \quad \lim_{h \rightarrow 0} \left\| \frac{1}{h} (x(t+h) - x(t)) - y(t) \right\|_z = 0.$$

When x is differentiable over T , the function $y: T \rightarrow Z$ can be shown to be a random process on T (see [9, p. 74]). The random process y and the value $y(t)$ will be denoted by dx/dt and $dx(t)/dt$, respectively.

In the discussion following, an integral of random processes is needed. It is clear that the ordinary mean-square integral cannot be defined for the random processes defined above and that the integral to be employed must be of Lebesgue type. There have been two such integrals proposed, the Bochner integral and the Pettis integral [9], [10]. The latter is more general than the former, but the Bochner integral is more suitable for applications. Thus we will employ the Bochner integral. The integral is defined in a manner analogous to the Lebesgue integral of real-valued functions, and has most of the properties which hold for the Lebesgue integral. For the definition and its properties the reader is referred to [9] and [10]. All the integrals used in the sequel are of Lebesgue type for both real-valued functions and Z -valued functions, and so the ordinary integral notation $\int \cdot dt$ will be used to mean the integrals $\int \cdot d\lambda(t)$ of Lebesgue type.

3. Generalized random processes. First, we need to define a set of test functions and a topology on the set. Though there are a number of possible sets and topologies to be adopted, we follow the celebrated work by Schwartz [1]. Let T be an interval in R , as defined previously. Then the set of *test functions* is defined to be the collection of all infinitely differentiable functions on T to R such that the support of each function φ , i.e., the closure of the set $\{t \in T: \varphi(t) \neq 0\}$, is a compact set contained in \dot{T} , the interior of T . This set with the addition and the scalar multiplication defined pointwise is a real vector space, denoted by $C_0^\infty(T)$. A sequence $\{\varphi_j\}$ of functions in $C_0^\infty(T)$ is then said to converge to the identically zero function (the zero element of $C_0^\infty(T)$) if (i) there exists a compact set K in \dot{T} such that K contains the support of every φ_j , and (ii) the derivatives $d^\alpha \varphi_j / dt^\alpha$ of each order $\alpha \geq 0$ converge to 0 uniformly on K ($d^\alpha \varphi / dt^\alpha = \varphi$). Then the vector space $C_0^\infty(T)$ with the topology induced by the notion of convergence defined above becomes a topological vector space, and the space will be denoted by $D(T)$. Spaces $C_0^\infty(T)$ and $D(T)$ may be used interchangeably when the topology is not in question.

Throughout this study, we will denote by $L(V, W)$ the vector space of all continuous linear mappings from a topological vector space V to another topological vector space W , where the addition and the scalar multiplication are defined pointwise. We can now state our definition of generalized random processes as follows.

DEFINITION 3.1. A *generalized random process* on T is a continuous linear mapping from $D(T)$ to Z . Namely, $D(T)^* = L(D(T), Z)$ is the vector space of all generalized random processes on T . The value of a generalized random process Φ at φ will be denoted by $\langle \varphi, \Phi \rangle$. \square

An example of the generalized random processes is as follows. Let $x: T \rightarrow Z$ be a locally Bochner integrable random process on T , i.e., x is a random process

on T such that for any compact set S in T the Bochner integral $\int_S x(t) dt$ exists. Then, for every φ in $D(T)$, the Bochner integral $\int_S \varphi(t)x(t) dt$ exists since φ is continuous on T and zero outside some compact set in \hat{T} . Moreover, the mapping $\Phi_x: D(T) \rightarrow Z$ defined by

$$(3.1) \quad \langle \varphi, \Phi_x \rangle = \int_T \varphi(t)x(t) dt, \quad \varphi \in D(T)$$

is linear and continuous. In fact, the linearity follows from the linearity of the Bochner integral, and the continuity can be proved as follows. Let $\{\varphi_j\}$ be a sequence from $D(T)$ converging in $D(T)$ to 0. Then by virtue of [9, Thm. 3.7.6, p. 82], we obtain

$$(3.2) \quad \begin{aligned} \|\langle \varphi_j, \Phi_x \rangle\|_Z &\leq \int_K |\varphi_j(t)| \|x(t)\|_Z dt \\ &\leq \max \{|\varphi_j(t)| : t \in K\} \int_K \|x(t)\|_Z dt, \end{aligned}$$

where K is a compact set in \hat{T} which contains the support of every φ_j . Since $\varphi_j \rightarrow 0$ in $D(T)$ implies $\max \{|\varphi_j(t)| : t \in K\} \rightarrow 0$ as $j \rightarrow \infty$, and since the local integrability of x and [9, Thm. 3.7.4, p. 80] ensure the existence of the integral in the last expression, it follows that $\langle \varphi_j, \Phi_x \rangle \rightarrow 0$ in Z as $j \rightarrow \infty$. Hence Φ_x is continuous and is a generalized random process in $D(T)^*$.

From the above example it can be seen that every locally Bochner integrable random process on T determines a generalized random process of $D(T)^*$ in the form of (3.1). Moreover, the following theorem holds.

THEOREM 3.1. *Let x and y be locally Bochner integrable random processes on T , and let Φ_x and Φ_y be the corresponding generalized random processes in $D(T)^*$. Then $\Phi_x = \Phi_y$ if and only if x and y are λ -equivalent, i.e., $x(t) = y(t)$ for a.e. t in T .*

Proof. Suppose first that $x(t) = y(t)$ for a.e. t in T , and let $z(t) = x(t) - y(t)$. Then we have $E[z(t)z(s)] = 0$ a.e. on $T \times T$, and, from [9, Thm. 3.7.13, p. 84] and Fubini's theorem, it follows that for every φ in $D(T)$,

$$\begin{aligned} \|\langle \varphi, \Phi_x - \Phi_y \rangle\|_Z^2 &= \int_T \|\varphi(t)(x(t) - y(t))\|_Z^2 dt \\ &= \int_T \int_T \varphi(t)E[z(t)z(s)]\varphi(s) dt ds = 0; \end{aligned}$$

so $\Phi_x = \Phi_y$. To show the converse, suppose $\Phi_x = \Phi_y$, and consider the Fourier series

$$(3.3) \quad z(t) = \sum_{i=1}^{\infty} g_i(t)\xi_i,$$

where $\{\xi_i\}$ is a complete orthonormal system of Z and $g_i(t) = (z(t), \xi_i)_Z$ are locally Lebesgue integrable real-valued functions on T (see e.g., [9, p. 72]). Then, for every φ in $C_0^\infty(T)$, we obtain

$$(3.4) \quad 0 = \|\langle \varphi, \Phi_x - \Phi_y \rangle\|_Z^2 = \left\| \int_T \varphi(t)z(t) dt \right\|_Z^2 = \sum_{i=1}^{\infty} \left\{ \int_T \varphi(t)g_i(t) dt \right\}^2.$$

Since, for any compact set K in T , $C_0^\infty(K)$ is dense in the Banach space $L^1(K)$ of all Lebesgue absolutely integrable functions on K , (3.4) implies $g_i(t) = 0$ a.e. for all i . Thus from (3.3) the desired result $z(t) = x(t) - y(t) = 0$ a.e. is obtained. \square

From the above theorem it can be seen that every λ -equivalence class of locally Bochner integrable random processes $\{x\}$ on T is identified uniquely with the corresponding generalized random process Φ_x of $D(T)^*$. Thus any locally Bochner integrable random process x on T may be called a generalized random process in $D(T)^*$, and Φ_x may be written simply as x (i.e., $\langle \varphi, \Phi_x \rangle$ may be written as $\langle \varphi, x \rangle$). Moreover, the space $B(T)$ of all λ -equivalence classes of locally Bochner integrable random processes on T can be regarded as a subset of $D(T)^*$ (i.e., $B(T) \subset D(T)^*$). It can easily be shown that $B(T)$ is a proper subset of $D(T)^*$. Namely, $D(T)^*$ contains those generalized random processes that cannot be expressed by any random process in the form of (3.1). In fact, the generalized random process Φ defined by

$$(3.5) \quad \langle \varphi, \Phi \rangle = \langle \varphi, \delta \rangle \xi \quad \text{for all } \varphi \text{ in } D(T)$$

is such a process, where δ is the Dirac δ -functional and ξ is a random variable in Z . Furthermore, since R is a subspace of Z as the space of all constant random variables, the δ -functional itself is a generalized random process in $D(T)^*$ and the space $D(T)' = L(D(T), R)$ of all distributions on T is contained in $D(T)^*$. Thus the notion of generalized random processes is more general than that of distributions. (In fact, the generalized random processes can be treated in the framework of vector-valued distributions [13].)

We now discuss derivatives of the generalized random processes. Suppose that x is a differentiable random process on a finite interval $[a, b]$ such that dx/dt belongs to $B([a, b])$. Then, as was discussed above, dx/dt determines a unique generalized random process $\Phi_{dx/dt}$ in $D([a, b])^*$ by

$$\langle \varphi, \Phi_{dx/dt} \rangle = \int_a^b \varphi(t) \frac{dx(t)}{dt} dt \quad \text{for all } \varphi \text{ in } D([a, b]).$$

Since every φ is continuously differentiable and $\varphi(a) = \varphi(b) = 0$, it follows from [9, Cor. 2, p. 88] that

$$\begin{aligned} \langle \varphi, \Phi_{dx/dt} \rangle &= \int_a^b \left\{ \frac{d}{dt}(\varphi(t)x(t)) - \frac{d\varphi(t)}{dt}x(t) \right\} dt \\ &= - \int_a^b \frac{d\varphi(t)}{dt}x(t) dt = - \langle d\varphi/dt, \Phi_x \rangle. \end{aligned}$$

Based on this equality, we make the following definition.

DEFINITION 3.2. The *derivative* Φ' of a generalized random process Φ in $D(T)^*$ is defined by

$$\langle \varphi, \Phi' \rangle = - \langle d\varphi/dt, \Phi \rangle \quad \text{for all } \varphi \text{ in } D(T). \quad \square$$

It should be verified that the mapping $\varphi \rightarrow \langle d\varphi/dt, \Phi \rangle$ is really a continuous linear mapping from $D(T)$ to Z . However, the verification is straightforward, and so it is omitted here. From the above definition it easily follows that every generalized

random process Φ in $D(T)^*$ has the derivative $\Phi^{(\alpha)}$ of any order $\alpha \geq 1$, and it is given by

$$(3.6) \quad \langle \varphi, \Phi^{(\alpha)} \rangle = (-1)^\alpha \langle d^\alpha \varphi / dt^\alpha, \Phi \rangle \quad \text{for all } \varphi \text{ in } D(T).$$

Moreover, we can easily see that any random process x in $B(T)$ has its derivative $x^{(\alpha)}$ of every order $\alpha \geq 1$ as a generalized random process in $D(T)^*$, and it is given by

$$(3.7) \quad \langle \varphi, x^{(\alpha)} \rangle = (-1)^\alpha \langle d^\alpha \varphi / dt^\alpha, x \rangle \quad \text{for all } \varphi \text{ in } D(T).$$

In order to distinguish $d^\alpha x / dt^\alpha$ and $x^{(\alpha)}$, we will call $x^{(\alpha)}$ the α th generalized derivative of x . Of course, if dx/dt belongs to $B(T)$, it follows from the definition of x' that x' is the same as dx/dt in the sense that

$$\langle \varphi, x' \rangle = \langle \varphi, dx/dt \rangle \quad \text{for all } \varphi \text{ in } D(T).$$

Thus the generalized derivative is a notion more general than the ordinary derivative.

Finally, we note that a multidimensional generalized random process can be defined in an analogous manner. Namely, an n -dimensional generalized random process on T is defined to be a continuous linear mapping from $D(T)_n$ to Z , where $D(T)_n$ is the product topological vector space of n spaces $D(T)$'s. It can be shown that any n -dimensional generalized random process is uniquely decomposed into n one-dimensional generalized random processes [8, p. 97].

4. Generalized random processes defined on $L^2(T)$. This section is concerned with a class of generalized random processes which are extended to the space $L^2(T)$ of test functions. Here $L^2(T)$ is the Hilbert space of λ -equivalence classes of Lebesgue square integrable functions on T with its inner product $(\cdot, \cdot)_{L^2(T)}$ and norm $\|\cdot\|_{L^2(T)}$ defined by

$$(4.1) \quad (\varphi, \psi)_{L^2(T)} = \int_T \varphi(t)\psi(t) dt, \quad \|\varphi\|_{L^2(T)} = (\varphi, \varphi)_{L^2(T)}^{1/2}.$$

Let $L^2(T)^*$ denote the Banach space of all continuous linear mappings from $L^2(T)$ to Z (i.e., $L^2(T)^* = L(L^2(T), Z)$) with the norm defined by³

$$(4.2) \quad \|\Phi\|_{L^2(T)^*} = \sup \{ \|\langle \varphi, \Phi \rangle\|_Z : \varphi \in L^2(T), \|\varphi\|_{L^2(T)} \leq 1 \},$$

where $\langle \varphi, \Phi \rangle$ indicates the value of Φ in $L^2(T)^*$ at φ in $L^2(T)$. Further, let a sequence $\{\Phi_j\}$ of generalized random processes in $D(T)^*$ be said to converge to an element Φ in $D(T)^*$ if for each φ in $D(T)$,

$$(4.3) \quad \lim_{j \rightarrow \infty} \langle \varphi, \Phi_j \rangle = \langle \varphi, \Phi \rangle \quad \text{in } Z.$$

Then the following theorem holds.

THEOREM 4.1. *Let Φ be in $L^2(T)^*$ and let $\Phi_{D(T)}$ denote the restriction of Φ to $D(T)$. Then $\Phi_{D(T)}$ belongs to $D(T)^*$, and mapping $\Phi \rightarrow \Phi_{D(T)}$ of $L^2(T)^*$ to $D(T)^*$ is linear, continuous and one-to-one (i.e., $L^2(T)^*$ can be imbedded in $D(T)^*$ linearly and continuously).*

Proof. Since any convergent sequence in $D(T)$ is a convergent sequence in $L^2(T)$ and since the linearity of $\Phi_{D(T)}$ is obvious, the first assertion follows. To

³ Note that every function f in $L^2(T)$ is an element of $L^2(T)^*$ and $\|f\|_{L^2(T)^*} = \|f\|_{L^2(T)}$.

show the second assertion, let Φ and Ψ be distinct elements in $L^2(T)^*$ and suppose $\Phi_{D(T)} = \Psi_{D(T)}$ in $D(T)^*$. Also, let φ be in $L^2(T)$ and $\{\varphi_j\}$ be a sequence from $C_0^\infty(T)$ converging in $L^2(T)$ to φ . (Note that since $C_0^\infty(T)$ is dense in $L^2(T)$, such a sequence always exists.) Then we get

$$\begin{aligned} \langle \varphi, \Phi \rangle - \langle \varphi, \Psi \rangle &= \lim_{j \rightarrow \infty} \langle \varphi_j, \Phi - \Psi \rangle \\ (4.4) \qquad &= \lim_{j \rightarrow \infty} \langle \varphi_j, \Phi_{D(T)} - \Psi_{D(T)} \rangle = 0. \end{aligned}$$

Since φ was an arbitrary function in $L^2(T)$, (4.4) contradicts the assumption $\Phi \neq \Psi$, and hence the mapping $\Phi \rightarrow \Phi_{D(T)}$ is one-to-one. To show the continuity of the mapping, let $\{\Phi_j\}$ be a Cauchy sequence from $L^2(T)^*$ converging in $L^2(T)^*$ to Φ in $L^2(T)^*$, and let φ be an arbitrary function in $D(T)$. Then

$$\begin{aligned} \|\langle \varphi, \Phi \rangle - \langle \varphi, \Phi_{D(T)} \rangle\|_Z &= \|\langle \varphi, \Phi_{D(T)} - \Phi_{D(T)} \rangle\|_Z \\ &\leq \|\varphi\|_{L^2(T)} \|\Phi_j - \Phi\|_{L^2(T)^*} \rightarrow 0 \quad \text{as } j \rightarrow \infty. \end{aligned}$$

Thus $\Phi_{D(T)} \rightarrow \Phi_{D(T)}$ in $D(T)^*$, and so the mapping $\Phi \rightarrow \Phi_{D(T)}$ of $L^2(T)^*$ to $D(T)^*$ is continuous. Since the linearity of the mapping is obvious, this completes the proof of the theorem. \square

Based on the above theorem, each element Φ of $L^2(T)^*$ can be identified uniquely with the corresponding generalized random process $\Phi_{D(T)}$ in $D(T)^*$. Therefore we will call any element of $L^2(T)^*$ a generalized random process on T , and write $L^2(T)^* < D(T)^*$ to mean that $L^2(T)^*$ is imbedded in $D(T)^*$ linearly and continuously.

We next consider another class of generalized random processes. Let $Y(T)$ denote the vector space of infinitely differentiable random processes on T having compact supports contained in \dot{T} . Then since any random process in $Y(T)$ belongs to $B(T)$, each x in $Y(T)$ determines a unique generalized random process Φ_x in $L^2(T)^*$ by

$$(4.5) \qquad \langle \varphi, \Phi_x \rangle = \int_T \varphi(t)x(t) dt \quad \text{for all } \varphi \text{ in } L^2(T).$$

Moreover, it is obvious that $x \neq y$ in $Y(T)$ implies $\Phi_x \neq \Phi_y$ in $L^2(T)^*$. Therefore by identifying each x of $Y(T)$ with Φ_x of $L^2(T)^*$, we can write $Y(T) < L^2(T)^*$. Now let $X(T)$ denote the completion of $Y(T)$ with respect to norm $\|\cdot\|_{L^2(T)^*}$. Then since for any φ in $L^2(T)$ and any y in $Y(T)$

$$(4.6) \qquad \|\langle \varphi, y \rangle\|_Z \leq \|\varphi\|_{L^2(T)} \|y\|_{L^2(T)^*},$$

each element x of $X(T)$ determines a unique generalized random process by

$$(4.7) \qquad \langle \varphi, x \rangle = \lim_{j \rightarrow \infty} \langle \varphi, x_j \rangle \quad \text{for all } \varphi \text{ in } L^2(T),$$

where $\{x_j\}$ is a Cauchy sequence from $Y(T)$ converging in $X(T)$ to x . It is easily checked that $\langle \varphi, x \rangle$ is independent of the choice of sequence $\{x_j\}$ and satisfies

$$(4.8) \qquad \|\langle \varphi, x \rangle\|_Z \leq \|\varphi\|_{L^2(T)} \|x\|_{L^2(T)^*}.$$

Moreover, since $Y(T) < L^2(T)^*$ and $X(T)$ is the completion of $Y(T)$ with respect to $\|\cdot\|_{L^2(T)^*}$, we obtain $Y(T) < X(T) < L^2(T)^*$. Thus using the relation $L^2(T)^* < D(T)^*$, we get $Y(T) < X(T) < L^2(T)^* < D(T)^*$.

Let $\{\xi_j\}$ be a complete orthonormal system of Z , and for any f in $L^2(T)$ and any ξ in Z , let $f\xi$ denote the generalized random process $x(t) = f(t)\xi$, i.e.,

$$(4.9) \quad \langle \varphi, f\xi \rangle = (\varphi, f)_{L^2(T)} \xi \quad \text{for all } \varphi \text{ in } L^2(T).$$

Then the following theorem can be proved (see Appendix A).

THEOREM 4.2. *Let x be a generalized random process on T . Then a necessary and sufficient condition for x to belong to $X(T)$ is that x is expressible in the form*

$$(4.10) \quad x = \sum_{i=1}^{\infty} f_i \xi_i$$

for some functions in $L^2(T)$, and the sequence $\{x_n\}$ defined by

$$(4.11) \quad x_n = \sum_{i=1}^n f_i \xi_i$$

is a Cauchy sequence converging in $L^2(T)^*$ to x .

Based on this theorem, we will prove the following.

THEOREM 4.3. *$X(T)$ is a proper closed subspace of $L^2(T)^*$.*

Proof. That $X(T)$ is a closed subspace follows from the facts that $X(T)$ is a Banach space with the norm $\|\cdot\|_{L^2(T)^*}$ and $X(T) < L^2(T)^*$. To see that $X(T)$ is a proper subset of $L^2(T)^*$, consider the mapping $\Phi: L^2(T) \rightarrow Z$ defined by $\Phi = \sum_{i=1}^{\infty} e_i \xi_i$, where $\{e_i\}$ is a complete orthonormal system of $L^2(T)$. Then for any φ in $L^2(T)$,

$$\|\langle \varphi, \Phi \rangle\|_Z^2 = \sum_{i=1}^{\infty} (\varphi, e_i)_{L^2(T)}^2 = \|\varphi\|_{L^2(T)}^2,$$

so Φ is continuous. Since the linearity of Φ is obvious, Φ is a generalized random process belonging to $L^2(T)^*$. However, it is not in $X(T)$. In fact, if we define a sequence $\{x_n\}$ by $x_n = \sum_{i=1}^n e_i \xi_i$, then for any $n > m$,

$$\|x_n - x_m\|_{L^2(T)^*} = \sup \left\{ \sum_{j=m+1}^n (\varphi, e_j)_{L^2(T)}^2 : \varphi \in L^2(T) : \|\varphi\|_{L^2(T)} \leq 1 \right\} = 1.$$

Thus $\{x_n\}$ is not a Cauchy sequence converging in $L^2(T)^*$ to Φ . So, by Theorem 4.2, Φ does not belong to $X(T)$, and hence $X(T)$ is a proper subset of $L^2(T)^*$. \square

It should be noted from this theorem that a generalized random process x that belongs to $L^2(T)^*$, but not to $X(T)$, cannot be approximated by a random process. In other words, for a sufficiently small $\varepsilon > 0$, there is no random process y on T such that $\|x - y\|_{L^2(T)^*} < \varepsilon$. This is an essential distinction of the generalized random processes from the generalized functions. Namely, for the case of generalized functions, any generalized function f belonging to $L(L^2(T), R)$ can be approximated by a function g in $C_0^\infty(T)$ with any given accuracy ε :

$$\|f - g\|_{L^2(T)^*} = \|f - g\|_{L^2(T)} < \varepsilon.$$

5. Correlation operators. For a random process x on T its (auto) correlation function $C(t, s)$ is defined by $C(t, s) = E[x(t)x(s)]$. The present section is concerned with an extended notion of the correlation function. Let $\Gamma(T)$ denote the vector space of all continuous linear operators from $L^2(T)$ to itself, i.e., $\Gamma(T) = L(L^2(T), L^2(T))$. Furthermore, if $K(t, s)$ is a Lebesgue square integrable function over $T \times T$, i.e., K is a function in $L^2(T \times T)$, we denote by K the operator in $\Gamma(T)$ defined by

$$(5.1) \quad (K\varphi)(t) = \int_T K(t, s)\varphi(s) ds \quad \text{for all } \varphi \text{ in } L^2(T).$$

Now let x be a generalized random process in $L^2(T)^*$. Then x determines a unique operator C_x belonging to $\Gamma(T)$ by

$$(5.2) \quad (\varphi, C_x\psi)_{L^2(T)} = (\langle\psi, x\rangle, \langle\varphi, x\rangle)_Z \quad \text{for all } \varphi, \psi \text{ in } L^2(T).$$

In fact, the linearity is obvious, and the continuity follows from

$$\begin{aligned} |(\varphi, C_x\psi)_{L^2(T)}| &\leq \|\langle\varphi, x\rangle\|_Z \|\langle\psi, x\rangle\|_Z \\ &\leq \|\varphi\|_{L^2(T)} \|\psi\|_{L^2(T)} \|x\|_{L^2(T)^*}^2. \end{aligned}$$

Finally, the uniqueness follows from the fact that, for any C in $\Gamma(T)$, $(\varphi, C\psi)_{L^2(T)} = 0$ for all φ, ψ in $L^2(T)$ implies $C = 0$.

DEFINITION 5.1. For each generalized random process x in $L^2(T)^*$, the bounded linear operator C_x in $\Gamma(T)$ defined by (5.2) is called the *correlation operator* of x . \square

When x is a random process in $L^2(T)^*$, the correlation operator C_x is expressed as

$$\begin{aligned} (\varphi, C_x\psi)_{L^2(T)} &= E \left[\int_T \varphi(t)x(t) dt \int_T \psi(s)x(s) ds \right] \\ &= \int_T \int_T \varphi(t) E[x(t)x(s)] \psi(s) dt ds \\ &= \int_T \varphi(t) \left\{ \int_T K_x(t, s) \psi(s) ds \right\} dt = (\varphi, K_x\psi)_{L^2(T)}, \end{aligned}$$

where $K_x(t, s) = E[x(t)x(s)]$. Thus $C_x = K_x$, and hence C_x is the operator determined by the correlation function $K_x(t, s)$ of x . Therefore, the correlation operator is an extended notion of the correlation functions.

6. The white Gaussian process. Although there are a number of ways to define the white Gaussian process as a generalized random process, we will define it using a Brownian process (or Wiener process). Throughout this section, assume $T = [0, \infty)$.

DEFINITION 6.1. A *Brownian process* β on T is a function from T to Z such that (i) β has independent increments, (ii) for any t and s in T , $\beta(t) - \beta(s)$ is a Gaussian random variable with zero mean and the variance $q|t - s|$, where $q > 0$ is called the *magnitude* of β , and (iii) $\beta(0) = 0$ in Z . \square

It can be shown easily that the Brownian process is continuous, and hence it is a random process in the sense of Definition 2.2. However, β is not differentiable on T . In fact, for each t in T ,

$$\left\| \frac{1}{h}(\beta(t+h) - \beta(t)) \right\|_Z^2 = q|h|/|h|^2 \rightarrow \infty \quad \text{as } h \rightarrow 0,$$

and hence $(\beta(t+h) - \beta(t))/h$ does not have a limit in Z . However, the generalized derivative β' is well-defined. Based on this fact, a white Gaussian process is defined as follows.

DEFINITION 6.2. A white Gaussian process w on T is a generalized random process in $D(T)^*$ such that there exists a Brownian process β on T satisfying $w = \beta'$, i.e.,

$$\langle \varphi, w \rangle = \langle \varphi, \beta' \rangle = -\langle d\varphi/dt, \beta \rangle \quad \text{for all } \varphi \in D(T).$$

In this case, the magnitude $q > 0$ of β is called also the magnitude of w . \square

Note that for every φ in $D(T)$, $\langle \varphi, w \rangle$ is a Gaussian random variable, and also that, for every finite number of functions $\varphi_1, \varphi_2, \dots, \varphi_n$ in $D(T)$, the random variables $\{\langle \varphi_j, w \rangle : 1 \leq j \leq n\}$ are jointly Gaussian. (This is the reason that w is called a white "Gaussian" process.) Furthermore, the following theorem holds.

THEOREM 6.1. Let w be a white Gaussian process in $D(T)^*$ with its magnitude $q > 0$. Then

$$E[\langle \varphi, w \rangle] = 0 \quad \text{for all } \varphi \text{ in } D(T)$$

and

$$E[\langle \varphi, w \rangle \langle \psi, w \rangle] = q \int_T \varphi(t) \psi(t) dt \quad \text{for all } \varphi, \psi \text{ in } D(T).$$

Proof. Let β be a Brownian process satisfying $\beta' = w$, and φ be in $D(T)$. Then by Definition 6.2 and [9, Thm. 3.7.12, p. 83],

$$E[\langle \varphi, w \rangle] = E\left[-\int_T \frac{d\varphi(t)}{dt} \beta(t) dt\right] = -\int_T \frac{d\varphi(t)}{dt} E[\beta(t)] dt.$$

Since $E[\beta(t)] = 0$ for all t in T and φ was arbitrary, the first desired equality is obtained. To show the second equality, let φ and ψ be in $D(T)$. Then by [9, Thms. 3.7.12 and 3.7.13], we obtain

$$\begin{aligned} E[\langle \varphi, w \rangle \langle \psi, w \rangle] &= E\left[\int_T \frac{d\varphi(t)}{dt} \beta(t) dt \int_T \frac{d\psi(s)}{ds} \beta(s) ds\right] \\ (6.1) \quad &= \int_T \int_T \frac{d\varphi(t)}{dt} E[\beta(t)\beta(s)] \frac{d\psi(s)}{ds} dt ds. \end{aligned}$$

From Definition 6.1 we have $E[\beta(t)\beta(s)] = q \min(t, s)$, and substituting this into (6.1) and applying integration by parts to the resultant expression yields the desired equality. \square

Furthermore, the following theorem can be proved.

THEOREM 6.2. *Every white Gaussian process w on T can be extended uniquely by continuity to all of $L^2(T)$. The extended white Gaussian process, also denoted by w , belongs to $L^2(T)^*$, but not to $X(T)$.*

Proof. Let φ be a function in $L^2(T)$ and let $\{\varphi_j\}$ be a Cauchy sequence from $C_0^\infty(T)$ converging to φ . (Note that since $C_0^\infty(T)$ is dense in $L^2(T)$, such a sequence exists for any function in $L^2(T)$.) Then from Theorem 6.1 we obtain

$$\begin{aligned}\|\langle \varphi_i, w \rangle - \langle \varphi_j, w \rangle\|_Z^2 &= \|\langle \varphi_i - \varphi_j, w \rangle\|_Z^2 \\ &= q \|\varphi_i - \varphi_j\|_{L^2(T)}^2 \rightarrow 0 \quad \text{as } i, j \rightarrow \infty,\end{aligned}$$

where $q > 0$ is the magnitude of w . Thus $\{\langle \varphi_j, w \rangle\}$ is a Cauchy sequence of Gaussian random variables in Z , and so it has a unique limit in Z , denoted by $\langle \varphi, w \rangle$. (Note that $\langle \varphi, w \rangle$ is Gaussian also.) The limit $\langle \varphi, w \rangle$ is clearly independent of the choice of the approximating sequence $\{\varphi_j\}$, and hence w is extended uniquely by continuity to all φ of $L^2(T)$.

To show that w belongs to $L^2(T)^*$, we first note that for any φ in $L^2(T)$ and for any $\{\varphi_j\}$ in $C_0^\infty(T)$ such that $\varphi_j \rightarrow \varphi$ in $L^2(T)$,

$$(6.2) \quad \|\langle \varphi, w \rangle\|_Z^2 = \lim_{j \rightarrow \infty} \|\langle \varphi_j, w \rangle\|_Z^2 = \lim_{j \rightarrow \infty} q \|\varphi_j\|_{L^2(T)}^2 = q \|\varphi\|_{L^2(T)}^2.$$

Thus w is continuous. Since the linearity of w on $L^2(T)$ is obvious, w belongs to $L^2(T)^*$. We will next show that w is not in $X(T)$. First, it will be proved that the set $Z_w = \{\langle \varphi, w \rangle \in Z : \varphi \in L^2(T)\}$ is a closed subspace of Z . Let $\{\eta_j\}$ be a Cauchy sequence from Z_w and let $\{\varphi_j\}$ be functions in $L^2(T)$ such that $\eta_j = \langle \varphi_j, w \rangle$. Then by virtue of (6.2),

$$\|\eta_i - \eta_j\|_Z^2 = \|\langle \varphi_i - \varphi_j, w \rangle\|_Z^2 = q \|\varphi_i - \varphi_j\|_{L^2(T)}^2 \rightarrow 0 \quad \text{as } i, j \rightarrow \infty.$$

Thus $\{\varphi_j\}$ is a Cauchy sequence in $L^2(T)$, and if $\varphi_j \rightarrow \varphi$ and $\eta_j \rightarrow \eta$,

$$\eta = \lim_{j \rightarrow \infty} \eta_j = \lim_{j \rightarrow \infty} \langle \varphi_j, w \rangle = \langle \varphi, w \rangle.$$

So η belongs to Z_w . Since Z_w is obviously a vector space, Z_w is a closed subspace of Z . Now, let $\{e_i\}$ be a complete orthonormal system of $L^2(T)$ and let $\xi_j = \langle e_j, w \rangle / \sqrt{q}$ for all $j \geq 1$. Then $\{\xi_j\}$ can easily be shown to be a complete orthonormal system of Z_w , and for any φ in $L^2(T)$ we have

$$\begin{aligned}\langle \varphi, w \rangle &= \sum_{j=1}^{\infty} (\varphi, e_j)_{L^2(T)} e_j, w \\ &= \sqrt{q} \sum_{j=1}^{\infty} \{(\varphi, e_j)_{L^2(T)} \langle e_j, w \rangle / \sqrt{q}\} = \left\langle \varphi, \sqrt{q} \sum_{j=1}^{\infty} e_j \xi_j \right\rangle.\end{aligned}$$

Thus w can be expressed as $w = \sqrt{q} \sum_{j=1}^{\infty} e_j \xi_j$. Now define a sequence $\{w_n\}$ of random processes on T by $w_n(t) = \sqrt{q} \sum_{j=1}^n e_j(t) \xi_j$. Then each w_n is clearly in $X(T)$, but $\{w_n\}$ is not a Cauchy sequence in $L^2(T)^*$ because for $n > m \geq 1$,

$$\begin{aligned}\|w_n - w_m\|_{L^2(T)^*}^2 &= \sup \left\{ q \sum_{j=m+1}^n (\varphi, e_j)_{L^2(T)}^2 : \varphi \in L^2(T), \|\varphi\|_{L^2(T)} \leq 1 \right\} \\ &= q (e_n, e_n)_{L^2(T)}^2 = q > 0.\end{aligned}$$

Therefore by virtue of Theorem 4.2, w does not belong to $X(T)$. \square

For the extended white Gaussian process w , the value $\langle \varphi, w \rangle$ can be shown to be the Wiener integral for all φ in $L^2(T)$ (see [8, p. 94]). Moreover, it can be seen from the above theorem that any white Gaussian process w on T does not have an approximating random processes in the sense of norm $\|\cdot\|_{L^2(T)^*}$. That is to say, for sufficiently small $\varepsilon > 0$, there is no random process y on T such that for all φ in $L^2(T)$

$$(6.3) \quad \|\langle \varphi, w - y \rangle\|_Z \leq \varepsilon \|\varphi\|_{L^2(T)}.$$

However, we have the following theorem, which may give an intuitive characterization of a white Gaussian process. (The proof is given in Appendix B.)

THEOREM 6.3. *Let β be a Brownian process on T , and define a sequence $\{w_n\}$ of random processes on T by*

$$w_n(t) = (\beta(t + h_n) - \beta(t))/h_n \quad \text{for all } t \in T,$$

where $\{h_n\}$ is a sequence of positive numbers converging to 0. Then for each fixed φ in $L^2(T)$,

$$\lim_{n \rightarrow \infty} \langle \varphi, w_n \rangle = \langle \varphi, \beta' \rangle,$$

that is, $\{w_n\}$ converges to β' pointwise.

We now consider the correlation operator C_w of a white Gaussian process w in $L^2(T)^*$. First, note that since the extension of a white Gaussian process to $L^2(T)$ is continuous, Theorem 6.1 holds for all functions of $L^2(T)$. Thus, by Definition 5.1,

$$(6.4) \quad (\varphi, C_w \psi)_{L^2(T)} = E[\langle \varphi, w \rangle \langle \psi, w \rangle] = (\varphi, qI\psi)_{L^2(T)}$$

for all φ and ψ in $L^2(T)$, where q is the magnitude of w and I indicates the identity operator from $L^2(T)$ to itself. So we obtain $C_w = qI$. In engineering applications, a white Gaussian process w may be described as

$$(6.5) \quad E[w(t)] = 0 \quad \text{and} \quad E[w(t)w(s)] = q\delta(t - s)$$

where $\delta(t - s)$ is the Dirac δ -functional. This description may be interpreted as follows. If we write the Wiener integral $\langle \varphi, w \rangle$ in the integral form

$$(6.6) \quad \langle \varphi, w \rangle = \int_T \varphi(t)w(t) dt$$

and if we suppose the order of E and $\int_T \cdot dt$ can be interchanged, then Theorem 6.1 implies that for all φ in $L^2(T)$,

$$(6.7) \quad E\left[\int_T \varphi(t)w(t) dt\right] = \int_T \varphi(t)E[w(t)] dt = 0,$$

and for all φ and ψ in $L^2(T)$,

$$(6.8) \quad \begin{aligned} E\left[\int_T \varphi(t)w(t) dt \int_T \psi(s)w(s) ds\right] &= \int_T \int_T \varphi(t)E[w(t)w(s)]\psi(s) dt ds \\ &= \int_T \int_T \varphi(t)q\delta(t - s)\psi(s) dt ds. \end{aligned}$$

Since (6.7) and (6.8) hold for all φ and ψ in $L^2(T)$, equations (6.5) are obtained formally, where $\delta(t - s)$ is used as a formal representation of the identity operator I from $L^2(T)$ to itself (i.e., $q\delta(t - s) = qI = C_w$).

A nonstationary white Gaussian process can be defined in the framework of generalized random processes.

DEFINITION 6.3. A generalized random process v on T is called a *nonstationary white Gaussian process on T* if for all φ in $L^2(T)$, $\langle \varphi, v \rangle = \langle \varphi \sqrt{q}, w \rangle$, where w is a white Gaussian process on T with the magnitude of unity and q is a Lebesgue measurable, nonnegative and bounded function on T which is called the *magnitude function* of v ($q(\varphi \sqrt{q})(t) = \varphi(t) \sqrt{q(t)}$). \square

It is clear that $\langle \varphi, w \rangle$ is a Gaussian random variable for all φ in $L^2(T)$. Moreover from Theorem 6.1 it follows that for all φ in $L^2(T)$,

$$(6.9) \quad E[\langle \varphi, v \rangle] = 0,$$

and for all φ and ψ in $L^2(T)$,

$$(6.10) \quad E[\langle \varphi, v \rangle \langle \psi, v \rangle] = \int_T \varphi(t) q(t) \psi(t) dt.$$

Finally, we will define the so-called *white process* [14] as a generalized random process. Let $\{\eta_i\}$ be an orthonormal set of vectors in Z (i.e., $E[\eta_i^2] = 1$ and $E[\eta_i \eta_j] = 0$ for $i \neq j$) and $\{e_i\}$ a complete orthonormal system of $L^2(T)$. Define v by

$$(6.11) \quad v = \sum_{i=1}^{\infty} e_i \eta_i.$$

Then for all φ in $L^2(T)$, $\langle \varphi, v \rangle$ defined by

$$(6.12) \quad \langle \varphi, v \rangle = \sum_{i=1}^{\infty} (\varphi, e_i)_{L^2(T)} \eta_i$$

is meaningful because

$$(6.13) \quad \|\langle \varphi, v \rangle\|_Z = \|\varphi\|_{L^2(T)}.$$

Thus v belongs to $L^2(T)^*$, and hence it is a generalized random process. Moreover, it can be easily shown that for all φ and ψ in $L^2(T)$,

$$(6.14) \quad E[\langle \varphi, v \rangle \langle \psi, v \rangle] = \int_T \varphi(t) \psi(t) dt.$$

Based on the above argument, we can now make the following definition (also see Definition 6.3 and (6.10)).

DEFINITION 6.4. A generalized random process v on T is called a *white process on T* if for all φ and ψ in $L^2(T)$,

$$E[\langle \varphi, v \rangle \langle \psi, v \rangle] = \int_T \varphi(t) q(t) \psi(t) dt,$$

where $q(t)$ is a Lebesgue measurable, nonnegative and bounded function on T which is named the *magnitude function* of v . \square

It is obvious that the correlation operator C_v of the white process v is given by

$$(6.15) \quad (C_v \varphi)(t) = q(t) \varphi(t),$$

or formally,

$$(6.16) \quad C_v = q(t)\delta(t-s).$$

7. Conclusions. A class $D(T)^*$ of generalized random processes has been defined, and then the spaces $L^2(T)^*$ and $X(T)$ of generalized random processes defined on $L^2(T)$ have been studied. It was shown that $X(T)$ is a proper subspace of $L^2(T)^*$, i.e., that there are generalized random processes in $L^2(T)^*$ which cannot be approximated by random processes in the uniform norm $\|\cdot\|_{L^2(T)^*}$ of $L^2(T)^*$. (This is the essential distinction between generalized random processes and generalized functions.) The so-called white Gaussian process (also the white process) was discussed in the framework of generalized random processes; it was shown to belong to $L^2(T)^*$, but not to $X(T)$. The nonstationary white Gaussian process and the white process were discussed as generalized random processes.

Appendix A. The proof of Theorem 4.2. The sufficiency will be proved first. Since $C_0^\infty(T)$ is dense in $L^2(T)$, for every f_i there exists a sequence $\{f_{ij}: j = 1, 2, \dots\}$ from $C_0^\infty(T)$ converging in $L^2(T)$ to f_i . These sequences can be chosen to satisfy the conditions

$$(A.1) \quad \|f_i - f_{ij}\|_{L^2(T)} < 1/j, \quad 1 \leq i \leq j.$$

Now define the sequence $\{y_n\}$ in $Y(T)$ by

$$(A.2) \quad y_n(t) = \sum_{i=1}^n f_{in}(t)\xi_i.$$

Then using (4.11), (A.2) and (A.1), we have

$$\begin{aligned} \|x - y_n\|_{L^2(T)^*} &\leq \|x - x_n\|_{L^2(T)^*} + \|x_n - y_n\|_{L^2(T)^*} \\ &= \|x - x_n\|_{L^2(T)^*} \\ &\quad + \left[\sup \left\{ \sum_{i=1}^n (\varphi, f_i - f_{in})_{L^2(T)}^2 : \varphi \in L^2(T), \|\varphi\|_{L^2(T)} \leq 1 \right\} \right]^{1/2} \\ &\leq \|x - x_n\|_{L^2(T)^*} + \left[\sum_{i=1}^n \|f_i - f_{in}\|_{L^2(T)}^2 \right]^{1/2} \\ &\leq \|x - x_n\|_{L^2(T)^*} + 1/\sqrt{n}. \end{aligned}$$

Thus from the condition $\|x - x_n\|_{L^2(T)^*} \rightarrow 0$ as $n \rightarrow \infty$, it follows that $\{y_n\}$ is a Cauchy sequence from $Y(T)$ converging in $L^2(T)^*$ to x . Thus x belongs to $X(T)$.

To show the necessity, let x be in $X(T)$. Then for each i , the mapping $\varphi \rightarrow (\xi_i, \langle \varphi, x \rangle)_Z$ from $L^2(T)$ to R (real numbers) is linear and continuous. In fact, the linearity is obvious, and the continuity follows from

$$|(\xi_i, \langle \varphi, x \rangle)_Z| \leq \|\xi_i\|_Z \|\langle \varphi, x \rangle\|_Z \leq \|\xi_i\|_Z \|\varphi\|_{L^2(T)} \|x\|_{L^2(T)^*}.$$

Thus, by Riesz's representation theorem (see, e.g., [10, p. 90]), there exists a unique function f_i in $L^2(T)$ such that

$$(\xi_i, \langle \varphi, x \rangle)_Z = (\varphi, f_i)_{L^2(T)} \quad \text{for all } \varphi \text{ in } L^2(T).$$

So $\langle \varphi, x \rangle$ can be expressed uniquely in the form

$$\langle \varphi, x \rangle = \sum_{i=1}^{\infty} (\varphi, f_i)_{L^2(T)} \xi_i = \langle \varphi, \sum_{i=1}^{\infty} f_i \xi_i \rangle$$

for all φ of $L^2(T)$, and hence x can be written uniquely as

$$(A.3) \quad x = \sum_{i=1}^{\infty} f_i \xi_i, \quad \text{with } f_i \in L^2(T), \quad i = 1, 2, \dots$$

Now it will be shown that the sequence $\{x_n\}$ in $X(T)$ defined by $x_n = \sum_{i=1}^n f_i \xi_i$ is a Cauchy sequence in $L^2(T)^*$ converging to x . To do this, let $\{y_j\}$ be a sequence from $Y(T)$ such that $\|x - y_j\|_{L^2(T)^*} \rightarrow 0$ as $j \rightarrow \infty$, and express each y_j in the form $y_j(t) = \sum_{i=1}^{\infty} f_{ji}(t) \xi_i$, where $f_{ji}(t) = (\xi_i, y_j(t))_z$. Then note that since y_j belongs to $Y(T)$, we have $\int_T \|y_j(t)\|_Z^2 dt < \infty$, and thus

$$(A.4) \quad \sum_{i=1}^{\infty} \|f_{ji}\|_{L^2(T)}^2 = \int_T \|y_j(t)\|_Z^2 dt < \infty.$$

Now observe that for every j and every n ,

$$(A.5) \quad \|x - x_n\|_{L^2(T)^*} \leq \|x - y_j\|_{L^2(T)^*} + \|y_j - x_n\|_{L^2(T)^*},$$

$$\begin{aligned} \|y_j - x_n\|_{L^2(T)^*}^2 &= \sup \left\{ \sum_{i=1}^n (\varphi, f_{ji} - f_i)_{L^2(T)}^2 \right. \\ &\quad \left. + \sum_{i=n+1}^{\infty} (\varphi, f_{ji})_{L^2(T)}^2 : \varphi \in L^2(T), \|\varphi\|_{L^2(T)} \leq 1 \right\} \end{aligned}$$

$$\begin{aligned} (A.6) \quad &\leq \sup \left\{ \sum_{i=1}^n (\varphi, f_{ji} - f_i)_{L^2(T)}^2 : \varphi \in L^2(T), \|\varphi\|_{L^2(T)} \leq 1 \right\} \\ &\quad + \sum_{i=n+1}^{\infty} \|f_{ji}\|_{L^2(T)}^2 \\ &\leq \|x - y_j\|_{L^2(T)^*}^2 + \sum_{i=n+1}^{\infty} \|f_{ji}\|_{L^2(T)}^2. \end{aligned}$$

Therefore, by (A.5) and (A.6), for all j we have

$$(A.7) \quad \|x - x_n\|_{L^2(T)^*} \leq \|x - y_j\|_{L^2(T)^*} + \left\{ \|x - y_j\|_{L^2(T)^*}^2 + \sum_{i=n+1}^{\infty} \|f_{ji}\|_{L^2(T)}^2 \right\}^{1/2}.$$

Since $\|x - y_j\|_{L^2(T)^*} \rightarrow 0$, for any $\varepsilon > 0$ there is a j such that $\|x - y_j\|_{L^2(T)^*} < \varepsilon$. Furthermore, (A.4) implies that there is an n_0 such that for all $n \geq n_0$,

$$\sum_{i=n+1}^{\infty} \|f_{ji}\|_{L^2(T)}^2 < \varepsilon^2.$$

Thus it follows from (A.7) that for all $n \geq n_0$

$$\|x - x_n\|_{L^2(T)^*} \leq \varepsilon + \{\varepsilon^2 + \varepsilon^2\}^{1/2} = (1 + \sqrt{2})\varepsilon.$$

Since ε was arbitrary, this implies $\|x - x_n\|_{L^2(T)^*} \rightarrow 0$ as $n \rightarrow \infty$, and hence $\{x_n\}$ defined by (A.3) is a Cauchy sequence converging in $L^2(T)^*$ to x . This completes the proof of the theorem.

Appendix B. The proof of Theorem 6.3. Since $C_0^\infty(T)$ is dense in $L^2(T)$, it suffices to show the equality only for functions φ of $C_0^\infty(T)$. To do this, let φ be in $C_0^\infty(T)$ and let $[a, b] \subset T$ contain the support of φ . Then

$$\begin{aligned} \langle \varphi, w_n \rangle &= \frac{1}{h_n} \int_a^b \varphi(t) \{ \beta(t + h_n) - \beta(t) \} dt \\ &= \frac{1}{h_n} \left[\int_{a+h_n}^{b+h_n} \varphi(s - h_n) \beta(s) ds - \int_a^b \varphi(t) \beta(t) dt \right] \\ &= - \int_a^b \frac{1}{h_n} \{ \varphi(t) - \varphi(t - h_n) \} \beta(t) dt + \frac{1}{h_n} \int_b^{b+h_n} \varphi(s - h_n) \beta(s) ds \\ &\quad - \frac{1}{h_n} \int_a^{a+h_n} \varphi(s - h_n) \beta(s) ds. \end{aligned}$$

The last two terms evidently approach 0 as $n \rightarrow \infty$ because $(d^\alpha \varphi(t)/dt^\alpha)$ equals 0 at $t = a$ and b for every $\alpha \geq 0$. Since $\{ \varphi(t) - \varphi(t - h_n) \}/h_n$ converges to $d\varphi(t)/dt$ uniformly on $[a, b]$, we have the desired equality

$$\lim_{n \rightarrow \infty} \langle \varphi, w_n \rangle = - \int_a^b \frac{d\varphi(t)}{dt} \beta(t) dt = \langle \varphi, \beta' \rangle.$$

Since φ was an arbitrary function in $C_0^\infty(T)$, this completes the proof.

Acknowledgments. The authors wish to thank Dr. R. E. Showalter and Dr. J. R. Dickenson, both at the University of Texas at Austin, for their many useful discussions.

REFERENCES

- [1] L. SCHWARTZ, *Théorie des Distributions*, Hermann, Paris, 1950.
- [2] K. ITO, *Stationary random distributions*, Mem. Coll. Sci. Kyoto Univ. Ser. A, 28 (1954), no. 3, pp. 209–223.
- [3] I. M. GEL'FAND, *Generalized random processes*, Dokl. Akad. Nauk SSSR, 100 (1956), pp. 853–856 (in Russian).
- [4] I. M. GEL'FAND AND N. YA. VILENKIN, *Generalized Functions*, vol. 4, Academic Press, New York, 1964.
- [5] A. M. YAGLOM, *An Introduction to the Theory of Stationary Random Functions*, Prentice-Hall, Englewood Cliffs, N.J., 1962.
- [6] K. URBANIK, *Stochastic processes whose sample functions are distributions*, Theory Probability Appl., 1 (1956), pp. 132–134.
- [7] T. HIDA, *Stationary Stochastic Processes*, Princeton University Press, Princeton, N.J., 1970.
- [8] H. INABA, *A theory of generalized random processes and its applications*, Appl. Mech. Res. Lab. Rep. AML-1039, Univ. of Texas at Austin, 1972.
- [9] E. HILLE AND R. PHILLIPS, *Functional analysis and semigroups*, Colloquium Publications, No. 31, Amer. Math. Soc. Providence, R. I., 1957.
- [10] K. YOSHIDA, *Functional Analysis*, 2nd ed., Springer-Verlag, New York, 1968.
- [11] T. KAILATH, *An innovations approach to least-squares estimation. Part I: Linear filtering in additive white noise*, IEEE Trans. Automatic Control, AC-13, (1968), pp. 646–654.

- [12] R. E. KALMAN AND R. S. BUCY, *New Results in Linear Filtering and Prediction Theory*, Trans. ASME Ser. D. J. Basic Engrg., 83 (1961), pp. 96–107.
- [13] R. E. CARROLL, *Abstract Analysis in Partial Differential Equations*, Harper and Row, New York, 1969.
- [14] E. WONG, *Stochastic Processes in Information and Dynamical Systems*, McGraw-Hill, New York, 1971.

STOCHASTIC SCHEDULING DESIGN IN MULTISERVER SYSTEMS*

MICHA HOFRI†

Abstract. The problem of assigning traffic in a multiserver system, where customer-server relationships are determined beforehand, can be regarded as a problem in system design. This is done, and designs are identified which lead to optimization of several alternative criteria. The notion of most efficient set of servers is presented; some ramifications, into different forms of scheduling, are discussed.

Introduction.

1. A variety of real-life systems can be thus schematically described: A large set of sources generates service requests. These are of similar types, though not identical, usually, and can be handled by any of a given set of servers. The servers are interchangeable, at least to some extent, but will usually have different capabilities. For operational reasons, however (dictated by the nature of the application), a relationship between each source and a server must be established beforehand, so that when a source needs a service, only one¹ server of the set can perform it reasonably well. We intend to discuss a situation where the characteristics of the sources and the servers are known—to a degree we shall later specify. In this setup we investigate optimal ways of establishing the mentioned source-server relationship; optimal is considered to be that which minimizes waiting or response times.

2. Once such a relationship is established, service requests from a given source have a unique destination. Define p_i as the fraction of the total load allotted to the i th server. We assume that different sources generate service requests independently. From a point of view that only distinguishes the destinations of the service requests (and not their origins), an incoming request has a probability p_i of requiring the i th server. This is the source of the term stochastic scheduling, even though no scheduler is actually employed.

3. Examples are: (a) *Computer system*, with programs as sources of requests for data storage and retrieval on different auxiliary memories such as magnetic tape, discs and drums (the essentially one-way traffic associated with card readers and punches, paper tape devices and line printers can also be accommodated within this description). If data are stored at one run on a given device, it is reasonable to use the same device at the next run, rather than request a prior transfer—and certainly this is inadvisable during one run. The different devices are then the servers of our model. (b) *Medical clinic*, with GP's and specialists, each familiar with certain sets of families, patients and disorders, to the point of inadvisability of switching. Another, more complex example is (c) *Telephone exchange*. Here service is establishing a connection, and the mentioned relationship represents

* Received by the editors May 17, 1973, and in final revised form February 28, 1974.

† Department of Computer Science, State University of Pennsylvania, University Park, Pennsylvania. Now at P.O.B. 7063, Hakirya, Tel-Aviv, Israel.

¹ In § 28 we give a brief discussion of a more general type of system.

the selectors available to each incoming line, through the wiring of the exchange and its routing procedures; these are usually so designed that only part of the outgoing lines are available to any customer. The results of our analysis can be adapted to this application only in case of a delay system. Systems that clear blocked calls will naturally require different optimization criteria. Note that usually the relationships have to be established on the basis of estimates of future activity. The natural extension to subsequent observation and recalibration is not taken up here.

4. We formulate a mathematical model of the foregoing, suggest reasonable object functions to serve as optimization criteria, derive the appropriate optimal allocations and study their properties. Some elaborations and extensions are presented in conclusion.

5. The problem was suggested to us by Chen [1], [2]. Reference [2] contains a very similar solution (with identical results) to the one we produce in § 21, embedded in the context of example (a) above, and the proof of optimality uses a similar approach. In [3, §§ 4.4, 4.5] Kleinrock considers a somewhat comparable situation, but from a different objective. He investigates optimal assignment of server capabilities under a given load pattern, constrained by a given total service capacity, and then also optimal load assignment for a given total capacity (individual server capabilities are *not* specified then). Only the case of exponentially-distributed service durations is considered. It is interesting to note that the results obtained there, for optimal capacity assignment, differ from an “intuitive” assignment in a manner similar to the one we encounter here.

Mathematical model.

6. The aggregate demand for service is represented by a homogeneous Poisson process of arrival of single requests at rate Λ .

7. Each request is shunted to the server associated with its origin. If the server is free, service starts at once. If the server is busy, it joins the queue. Within each queue service is in the order of arrival.

8. Since we assume many sources to be associated with each server, and that service requests originate independently, the input process to each server is a homogeneous Poisson process, of rate $\lambda_i = p_i \Lambda$.

9. The durations of service given by the i th server are assumed to form independent realizations of a random variable S_i . We further assume these realizations do not depend on any feature of the state of the system or on the input process.

10. Under the assumptions in §§ 7, 8 and 9, each server and its population of potential customers can be regarded as a standard $M/G/1$ queueing system.

11. *Notation:* The number of servers is m , and we shall find it convenient to number them in order of nondecreasing value of $E(S_i)$. Also, we use the following notation: $\tau_i = E(S_i)$, $\eta_i = E(S_i^2)$, $\sigma_i^2 = V(S_i) = \eta_i - \tau_i^2$, W_i = the time a request has to wait for service at the i th server, $R_i = W_i + S_i$ is the response time of the i th subsystem.

12. The known Khintchine–Pollaczek formula gives us the following values, for stable subsystems, at statistical equilibrium (“steady state”):

$$(1) \quad E(W_i) = \lambda_i \frac{\eta_i}{2(1 - \lambda_i \tau_i)},$$

and immediately

$$(2) \quad E(R_i) = \frac{2\tau_i + \lambda_i(\sigma_i^2 - \tau_i^2)}{2(1 - \lambda_i \tau_i)}.$$

The expected waiting and response times for a random request that enters the system are given by

$$(3) \quad W = \sum_{i=1}^m p_i E(W_i) = \frac{1}{2\Lambda} \sum_{i=1}^m \frac{\lambda_i^2 \eta_i}{1 - \lambda_i \tau_i}$$

and

$$(4) \quad R = \sum_{i=1}^m p_i E(R_i) = \frac{1}{2\Lambda} \sum_{i=1}^m \frac{2\lambda_i \tau_i + \lambda_i^2(\sigma_i^2 - \tau_i^2)}{1 - \lambda_i \tau_i}.$$

The condition for (1)–(4) to hold is that

$$(5) \quad \lambda_i \tau_i < 1, \quad i = 1, 2, \dots, m;$$

otherwise the queues diverge and “steady state” formulation is meaningless.

Optimization.

13. We proceed to calculate the allocation of service requirements to servers—and determine λ_i in this manner—so as to bring our object function to a minimum.

14. Consider W defined in (3) as an object function. We shall minimize it over a set of all possible values of λ_i that satisfy the constraints (5) and

$$(6) \quad \lambda_i \geq 0, \quad i = 1, 2, \dots, m,$$

$$(7) \quad \sum \lambda_i = \Lambda.$$

15. We note at this stage that (5), (6) and (7) are all linear in λ_i , and so concave in λ_i .

16. The approach we found most natural to this nonlinear programming problem was to try to solve it using explicitly as few of the constraints as possible, and then to test for total feasibility. Minimizing (W) without any constraints leads immediately to

$$(8) \quad \lambda_i^* \tau_i = 2,$$

in direct contradiction with (5). Combining (3) and (7) we form the Lagrangian

$$(9) \quad L = W + \beta(\Lambda - \sum \lambda_i);$$

the equations $\partial L / \partial \lambda_i = 0$ are satisfied by

$$(10) \quad \lambda_i^* = \frac{1}{\tau_i} (1 \pm [\eta_i / (\eta_i + 2\beta\Lambda\tau_i)]^{1/2}).$$

Choosing the $(-)$ sign in (10) assures that (5) holds. The constant β is determined by constraint (7), which yields the equation

$$(11) \quad \Lambda = \sum_{i=1}^m \frac{1}{\tau_i} \left(1 - \left[\frac{\eta_i}{\eta_i + 2\beta\Lambda\tau_i} \right]^{1/2} \right).$$

17. Before we discuss the solution of (11), notice that (6) will only be satisfied if the square root in (10) is less than 1, which immediately reduces to

$$(12) \quad \beta \geq 0.$$

Now we observe that the derivative of the R.H.S. of (11) with respect to β is always positive—of constant sign, anyway; therefore, when β is varied in the interval $(0, \infty)$, the R.H.S. of (11) varies monotonically in the interval $(0, \sum 1/\tau_i)$. Under constraints (5) we have $\Lambda < \sum 1/\tau_i$, and thus (11), when considered an equation in β , has a unique real positive root $\bar{\beta}$.

18. The value of $\bar{\beta}$ could not be obtained analytically, but (11) is easily solved using any standard root-finding algorithm. We used the simple Newton–Raphson iteration procedure and found for many combinations of parameters that convergence was fast.

19. We point out that W is *convex* in λ_i , for all values that satisfy (5); this and § 15 supply the sufficient conditions for (10) to define the global minimum of W ; see [4, pp. 89–90].

20. From (10) we obtain for the optimal unused capacity of the servers (the complement of the utilization factor $\lambda_i^*\tau_i$ to 1) the values

$$(13) \quad \zeta_i^* = [\eta_i/(\eta_i + 2\Lambda\bar{\beta}\tau_i)]^{1/2}.$$

Here the way the allocation (10) differs for two servers according to their respective values of η and τ is more transparent. Note, however, that as Λ increases to its largest allowed value, $\sum 1/\tau_i$, $\bar{\beta} \rightarrow \infty$ and the utilization factors approach 1 uniformly in τ .

21. Consider now R , as defined in (4), as an object function to be minimized. We again consider all λ_i that satisfy (5)–(7). The same procedure that yielded (10) now gives

$$(14) \quad \lambda_i^* = \frac{1}{\tau_i} \left(1 \pm \left[\frac{\eta_i}{2\alpha\Lambda\tau_i + \sigma_i^2 - \tau_i^2} \right]^{1/2} \right), \quad i = 1, \dots, m,$$

where α is the appropriate Lagrange multiplier. Again, the $(-)$ sign in (14) is the one used to satisfy (5). In order to satisfy (7) we determine α through the equation obtained by summing (14) over all i :

$$(15) \quad \sum_{i=1}^m \frac{1}{\tau_i} - \Lambda = \sum_{i=1}^m \frac{1}{\tau_i} \left[\frac{\eta_i}{2\alpha\Lambda\tau_i + \sigma_i^2 - \tau_i^2} \right]^{1/2}.$$

Constraints (6) and the need for real λ_i^* require that $\alpha\Lambda \geq \tau_i$, and by § 11 this can be expressed as

$$(16) \quad \alpha\Lambda \geq \tau_m.$$

This is a restriction on α unlike anything we had for β , and actually, it is not always satisfied! This may be explicitly shown when all $S_i \sim \exp(\mu_i)$, since then $\tau_i^2 = \sigma_i^2 = 1/\mu_i^2$, and α can be computed directly as

$$(17) \quad \alpha_{\text{exp}} = \frac{1}{\Lambda} \left(\frac{\sum_{i=1}^m \sqrt{\mu_i}}{\sum \mu_j - \Lambda} \right)^2.$$

That this does not necessarily satisfy (16) is clear by inspection. This corresponds to the situation where some servers are so slow that it is not worthwhile to request any work of them; it is better to assign “their” load to the faster servers.

22. We now show that this interpretation is correct. That is, we consider the NLP posed as {Minimize R ; subject to (5), (6), (7)}, and show that (14), perhaps after removal of the “worst” servers, is the optimal allocation.

Consider the following procedure:

Step 1. Check if Λ is in the interval $[0, \sum_{i=1}^m 1/\tau_i]$; if not, no allocation that would result in stable operation is feasible, and the problem is outside the present context. If the inequality $0 \leq \Lambda < \sum 1/\tau_i$ holds, we define $k \leftarrow m$ and proceed.

Step 2. Solve the following equation for α :

$$(18) \quad \sum_{i=1}^k \frac{1}{\tau_i} - \Lambda = \sum_{i=1}^k \frac{1}{\tau_i} \left[\frac{\eta_i}{2\alpha\Lambda\tau_i + \sigma_i^2 - \tau_i^2} \right]^{1/2}.$$

Call the solution α_k . If this α_k satisfies the inequality

$$(19) \quad \alpha_k \Lambda \geq \tau_k,$$

then go to Step 4. Otherwise, go to Step 3.

Step 3. Reduce k by 1, and repeat Step 2.

Step 4. This is the final step. The current k and α_k are used to compute the allocated traffic intensities through (14), and we end up with the allocation

$$(20) \quad \lambda_i^* = \begin{cases} \frac{1}{\tau_i} \left(1 - \left[\frac{\eta_i}{2\alpha_k \Lambda \tau_i + \sigma_i^2 - \tau_i^2} \right]^{1/2} \right), & 1 \leq i \leq k, \\ 0, & k < i \leq m. \end{cases}$$

Since these values satisfy the set of constraints (5)–(7) which is linear, and therefore concave, and also since R is convex in λ_i (for $\lambda_i \tau_i < 1$), and α_k —as well as λ_i^* —is continuous in Λ , we have that (20) is indeed the optimal allocation. The linearity assures us that the Kuhn–Tucker constraint qualification holds, and then by [4], (20) is optimal for allocation on the k servers $1, \dots, k$. The continuity and monotonicity in Λ assure us that (20) is also optimal for $k = m$, and thus, when Step 4 is reached in the procedure we indeed obtain our aim.

It remains to show that Step 4 is always reached.

23. This amounts to showing that neither of the following two results can be obtained while performing the procedure described in § 22.

(i) For a certain k , (19) is not satisfied, i.e., we have

$$(21) \quad \alpha_k \Lambda < \tau_k$$

and on iterating Step 2 we find that

$$(22) \quad \sum_{i=1}^{k-1} \frac{1}{\tau_i} \leq \Lambda,$$

so that an allocation that would result in a stable system is impossible.

(ii) Step 3 is realized with $k = 1$, without the event described in (i) occurring before.

First we show that (i) is impossible. If Λ was found to satisfy $\Lambda < \sum_{i=1}^k 1/\tau_i$ and later $\Lambda \geq \sum_{i=1}^{k-1} 1/\tau_i$, this would mean that the R.H.S. of (18), when solved for k servers, is $< 1/\tau_k$. Hence

$$(23) \quad \sum_{i=1}^k \frac{1}{\tau_i} \left[\frac{\eta_i}{2\alpha\Lambda\tau_i + \sigma_i^2 - \tau_i^2} \right]^{1/2} < \frac{1}{\tau_k},$$

in particular, since all the terms are positive

$$(24) \quad \frac{1}{\tau_k} \left[\frac{\eta_k}{2\alpha\Lambda\tau_k + \sigma_k^2 - \tau_k^2} \right]^{1/2} < \frac{1}{\tau_k}$$

which immediately reduces to

$$(25) \quad \alpha_k\Lambda > \tau_k,$$

in contradiction with the assumed (21).

Case (ii) is likewise shown to be impossible, since it entails solving (18) for $k = 1$, i.e.,

$$\frac{1}{\tau_1} \left[\frac{\eta_1^2}{2\alpha\Lambda\tau_1 + \sigma_1^2 - \tau_1^2} \right]^{1/2} = \frac{1}{\tau_1} - \Lambda < \frac{1}{\tau_1}$$

from which (25) follows immediately for $k = 1$, and the algorithm is terminated in Step 4, as postulated.

Discussion.

24. Some remarks on (20) and the way it compares with (11) are in order.

The phenomenon that assigning work to more servers actually degrades overall system performance occurs only when the performance is evaluated using R . W , as is intuitively clear, is monotonically decreasing in the number of servers, unlike R .

Another interesting feature is the role of the variances of the service times. As § 22 and § 23 clearly show, only the expected values of S_i determine the order at which servers are “dropped” from the system when Λ declines (although the values of Λ , where this occurs, do depend on the variances). Also, the ratios of different λ_i^* depend on the variances, and when variance ratios differ markedly from 1 the dependence is quite conspicuous. (This is illustrated in Table 4.)

For the case mentioned in § 21, that of “exponential” servers, we obtain a simple ratio between the utilization factors of different servers. From (14) and (17), we have

$$\lambda_i^* = \mu - \sqrt{\mu_i} \theta$$

with

$$(26) \quad \theta = (\sum \mu_j - \Lambda) / \sum \sqrt{\mu_k}.$$

Thus, calling $1 - \rho$ the “unused capacity” of the server we have $\zeta_i = 1 - \rho_i = \theta / \sqrt{\mu_i}$, which gives explicitly the rate of increase of the loading of a server when his rate of service increases. Moreover, the average queue size (Q_i) seen by a customer arriving at the i th server is larger the “better” (i.e., faster) is the server! Again, in the exponential case,

$$(27) \quad E(Q_i) = \frac{(\theta - \sqrt{\mu_i})^2}{\theta \sqrt{\mu_i}},$$

and this quantity increases with increasing μ_i , if $\mu_i > \theta^2$, which would be the case if the system is highly loaded, for example.

25. We note in particular that both R and W are not minimized when the “classically intuitive” allocation, namely: equal utilizations for all the servers, is employed. This allocation turns out to minimize the probability of waiting. Let b_i be the probability that a request for the i th server is delayed. Then

$$(28) \quad b_i = 1 - \rho_i = 1 - \lambda_i \tau_i;$$

the average of b_i over the m servers is

$$(29) \quad \pi = \sum b_i \frac{\lambda_i}{\Lambda} = \frac{1}{\Lambda} \sum_{i=1}^m \lambda_i (1 - \lambda_i \tau_i).$$

The minimum of this, which satisfies (5)–(7), is immediately given, by the now well-rehearsed method, as

$$(30) \quad \lambda_i = \frac{\phi}{\tau_i}, \quad \phi = \Lambda / \sum_j \frac{1}{\tau_j},$$

where ϕ , which is always in the interval $(0, 1)$, may be called the total utilization of the system. This is a global minimum, as all the conditions we detailed in § 22 are satisfied here too.

26. The three optimal allocations described were numerically evaluated for a variety of combinations of service parameters. The quantities which were of special interest were the relative values of R and W under the various assignments and the dependence of the allotted load to each server on its parameters, particularly the variance ratio. Tables 1–4 describe the behavior of a few of the combinations investigated. Tables 1 and 2 specify the allocation over two servers, according to all three mentioned assignments. The main difference between the data of the two tables is in the variance ratios of the service durations. Table 3 describes such allocation for three servers, in Part I; Part II contains the allocation when Λ is such that the assignment that minimizes R “drops” the slowest server (the rest are also calculated then for the two remaining servers). Table 4 shows explicitly the dependence of the allocation, according to (20), on the variance ratio of the faster server. A qualitative assessment of the results from those tables and further calculations is that assignments (11) and (20) gave comparable values for R and W , for systems whose total utilization figure (ϕ , as in (30)) was 0.5 and

TABLE 1
 $m = 2, \tau_1 = 1.0, \sigma_1 = 5.0, \tau_2 = 5.0, \sigma_2 = 0.0$

Proportional Allocation						Alloc. to Minimize R					Alloc. to Minimize W				
λ	$\lambda_1 = \rho_1$	λ_2	ρ_2	W	R	$\lambda_1 = \rho_1$	λ_2	ρ_2	W	R	$\lambda_1 = \rho_1$	λ_2	ρ_2	W	R
1.194	0.9950	0.1990	0.9950	2239.0	2240.0	0.9945	0.1995	0.9976	2124.0	2126.0	0.9945	0.1995	0.9975	2124.0	2126.0
1.164	0.9700	0.1940	0.9700	363.7	365.4	0.9969	0.1970	0.9855	344.2	345.9	0.9669	0.1970	0.9855	344.2	345.9
1.116	0.9300	0.1860	0.9300	149.5	151.1	0.9228	0.1932	0.9660	140.8	142.5	0.9228	0.1932	0.9660	140.8	142.5
1.020	0.8500	0.1700	0.8500	63.75	65.42	0.8347	0.1853	0.9624	59.44	61.17	0.8347	0.1853	0.9266	59.44	61.17
0.9600	0.8000	0.1600	0.8000	45.00	46.67	0.7798	0.1801	0.9007	41.67	43.42	0.7797	0.1802	0.9014	41.67	43.42
0.9000	0.7500	0.1500	0.7500	33.75	35.42	0.7252	0.1748	0.8740	31.01	32.79	0.7249	0.1751	0.8754	31.01	32.79
0.8400	0.7000	0.1400	0.7000	26.25	27.92	0.6708	0.1692	0.8460	23.92	25.73	0.6703	0.1697	0.8486	23.92	25.73
0.7200	0.6000	0.1200	0.6000	16.87	18.54	0.5630	0.1569	0.7841	15.09	16.96	0.5618	0.1582	0.7909	15.08	16.96
0.6000	0.5000	0.1000	0.5000	11.25	12.92	0.4580	0.1420	0.7101	9.834	11.78	0.4548	0.1451	0.7257	9.820	11.79
0.4800	0.4000	0.0800	0.4000	7.500	9.167	0.3568	0.1231	0.6157	6.390	8.417	0.3502	0.1298	0.6490	6.362	8.443
0.4200	0.3500	0.0700	0.3500	6.058	7.724	0.3087	0.1113	0.5564	5.098	7.158	0.2991	0.1209	0.6043	5.050	7.202
0.3600	0.3000	0.0600	0.3000	4.821	6.488	0.2630	0.0970	0.4852	4.023	6.101	0.2492	0.1108	0.5539	3.943	6.173
0.3000	0.2500	0.0500	0.2500	3.750	5.417	0.2205	0.0795	0.3977	3.139	5.200	0.2008	0.0992	0.4962	3.000	5.323
0.2640	0.2200	0.0440	0.2200	3.173	4.840	0.1970	0.0670	0.3353	2.699	4.715	0.1725	0.0915	0.4573	2.501	4.887
0.2400	0.2000	0.0400	0.2000	2.812	4.479	0.1823	0.0577	0.2887	2.445	4.407	0.1541	0.0858	0.4291	2.194	4.625
0.2160	0.1800	0.0360	0.1800	2.470	4.136	0.1685	0.0475	0.2376	2.226	4.106	0.1362	0.0798	0.3990	1.906	4.384
0.1920	0.1600	0.0320	0.1600	2.143	3.810	0.1556	0.0364	0.1818	2.048	3.805	0.1186	0.0734	0.3668	1.635	4.163
0.1680	0.1400	0.0280	0.1400	1.831	3.498	0.1438	0.0242	0.1209	1.919	3.495	0.1016	0.0664	0.3322	1.380	3.962
0.1440	0.1200	0.0240	0.1200	1.534	3.200	0.1330	0.0109	0.0546	1.855	3.159	0.0850	0.0590	0.2949	1.141	3.780
0.1320	0.1100	0.0220	0.1100	1.390	3.057	0.1280	0.0039	0.0195	1.855	2.973	0.0770	0.0550	0.2752	1.028	3.695
0.1272	0.1060	0.0212	0.1060	1.334	3.001	0.1262	0.0012	0.0051	1.862	2.894	0.0738	0.0534	0.2670	0.9833	3.663

TABLE 2
 $m = 2, \tau_1 = 1.0, \sigma_1 = 1.0, \tau_2 = 5.0, \sigma_2 = 5.0$

Proportional Allocation						Alloc. to Minimize R					Alloc. to Minimize W				
λ	$\lambda_1 = \rho_1$	λ_2	ρ_2	W	R	$\lambda_1 = \rho_1$	λ_2	ρ_2	W	R	$\lambda_1 = \rho_1$	λ_2	ρ_2	W	R
1.194	0.9950	0.1990	0.9950	331.7	333.3	0.9959	0.1981	0.9907	289.0	290.7	0.9959	0.1981	0.9907	289.0	290.7
1.164	0.9700	0.1940	0.9700	53.89	55.56	0.9751	0.1889	0.9444	46.61	48.26	0.9751	0.1889	0.9444	46.61	48.26
1.116	0.9300	0.1860	0.9300	22.14	23.81	0.9420	0.1740	0.8702	18.93	20.55	0.9418	0.1742	0.8708	18.93	20.55
1.020	0.8500	0.1700	0.8500	9.444	11.11	0.8756	0.1444	0.7219	7.880	9.447	0.8745	0.1455	0.7277	7.878	9.449
0.9600	0.8000	0.1600	0.8000	6.667	8.333	0.8342	0.1258	0.6292	5.483	7.007	0.8314	0.1286	0.6428	5.477	7.013
0.9000	0.7500	0.1500	0.7500	5.000	6.666	0.7927	0.1073	0.5365	4.058	5.535	0.7875	0.1125	0.5626	4.046	5.546
0.8400	0.7000	0.1400	0.7000	3.889	5.556	0.7512	0.0888	0.4438	3.122	4.545	0.7424	0.0975	0.4880	3.101	4.566
0.7200	0.6000	0.1200	0.6000	2.500	4.166	0.6683	0.0517	0.2583	1.995	3.282	0.6486	0.0714	0.3571	1.938	3.335
0.6000	0.5000	0.1000	0.5000	1.667	3.333	0.5854	0.0146	0.0729	1.387	2.485	0.5496	0.0504	0.2518	1.259	2.595
0.5640	0.4700	0.0940	0.4700	1.478	3.145	0.5605	0.0035	0.0173	1.268	2.293	0.5190	0.4500	0.2250	1.108	2.428

TABLE 3-Part I:
 $m = 3, \tau_1 = 1.0, \sigma_1 = 0.5, \tau_2 = 2.0, \sigma_2 = 1.0, \tau_3 = 10.0, \sigma_3 = 10.0$

λ	$\lambda_1 = \rho_1$	λ_2	ρ_2	λ_3	ρ_3	W	R
			Proportional Allocation				
1.592	0.995	0.4975	0.995	0.0995	0.995	279.8	281.7
1.552	0.970	0.4850	0.970	0.0970	0.970	45.47	47.34
1.488	0.930	0.4650	0.930	0.0930	0.930	18.68	20.56
1.440	0.900	0.4500	0.900	0.0900	0.900	12.66	14.53
1.360	0.850	0.4250	0.850	0.0850	0.850	7.969	9.844
1.280	0.800	0.4000	0.800	0.0800	0.800	5.625	7.500
1.200	0.750	0.3750	0.750	0.0750	0.750	4.219	6.094
1.120	0.700	0.3500	0.700	0.0700	0.700	3.281	5.156
1.072	0.670	0.3350	0.670	0.0670	0.670	2.855	4.730
			Allocation to Minimize R				
1.592	0.9962	0.4973	0.9946	0.0985	0.9845	215.1	216.9
1.552	0.9772	0.4839	0.9678	0.0909	0.9089	34.43	36.27
1.488	0.9469	0.4624	0.9248	0.0788	0.7876	13.82	15.61
1.440	0.9241	0.4462	0.8925	0.0697	0.6969	9.209	10.95
1.360	0.8862	0.4192	0.8384	0.0546	0.5464	5.653	7.323
1.280	0.8483	0.3920	0.7840	0.0397	0.3973	3.910	5.495
1.200	0.8105	0.3645	0.7290	0.0250	0.2500	2.896	4.388
1.120	0.7728	0.3368	0.6735	0.0105	0.1048	2.253	3.638
1.072	0.7502	0.3199	0.6399	0.0019	0.0189	1.977	3.219
			Allocation to Minimize W				
1.592	0.9962	0.4973	0.9946	0.0985	0.9848	215.1	216.9
1.552	0.9772	0.4839	0.9678	0.0909	0.9092	34.43	36.27
1.488	0.9466	0.4623	0.9246	0.0791	0.7908	13.82	15.61
1.440	0.9234	0.4460	0.8920	0.0706	0.7062	9.206	10.96
1.360	0.8839	0.4185	0.8369	0.0576	0.5764	5.643	7.332
1.280	0.8432	0.3904	0.7809	0.0464	0.4638	3.886	5.517
1.200	0.8010	0.3620	0.7240	0.0370	0.3696	2.849	4.428
1.120	0.7574	0.3333	0.6666	0.0293	0.2928	2.172	3.705
1.072	0.7305	0.3160	0.6321	0.0254	0.2543	1.869	3.377

TABLE 3 Part II:
 $m = 2$, first two servers as Part I

Λ	Proportional Allocation					Alloc. to Minimize R					Alloc. to Minimize W				
	$\lambda_1 = \rho_1$	λ_2	ρ_2	W	R	$\lambda_1 = \rho_1$	λ_2	ρ_2	W	R	$\lambda_1 = \rho_1$	λ_2	ρ_2	W	R
1.493	0.9950	0.4975	0.9850	165.8	167.2	0.9956	0.4969	0.9938	161.0	162.4	0.9956	0.4969	0.9938	161.0	162.4
1.455	0.9700	0.4850	0.9700	26.94	28.28	0.9736	0.4814	0.9627	26.13	27.46	0.9736	0.4814	0.9627	26.13	27.46
1.395	0.9300	0.4650	0.9300	11.07	12.40	0.9385	0.4565	0.9130	10.71	12.04	0.9384	0.4566	0.9131	10.71	12.04
1.350	0.9000	0.4500	0.9000	7.500	8.833	0.9122	0.4378	0.8757	7.241	8.565	0.9120	0.4380	0.8760	7.241	8.565
1.275	0.8500	0.4250	0.8500	4.722	6.056	0.8685	0.4065	0.8130	4.544	5.863	0.8677	0.4073	0.8140	4.544	5.864
1.200	0.8000	0.4000	0.8000	3.333	4.667	0.8249	0.3751	0.7501	3.197	4.510	0.8231	0.3769	0.7537	3.197	4.511
1.125	0.7500	0.3750	0.7500	2.500	3.833	0.7817	0.3433	0.6867	2.391	3.696	0.7781	0.3469	0.6937	2.389	3.697
1.050	0.7000	0.3500	0.7000	1.944	3.278	0.7387	0.3113	0.6226	1.854	3.151	0.7326	0.3174	0.6347	1.852	3.154
0.9000	0.6000	0.3000	0.6000	1.250	2.583	0.6540	0.2460	0.4921	1.189	2.463	0.6397	0.2603	0.5206	1.181	2.471
0.7500	0.5000	0.2500	0.5000	0.8333	2.167	0.5713	0.1787	0.3573	0.8001	2.038	0.5436	0.2064	0.4128	0.7814	2.057
0.6000	0.4000	0.2000	0.4000	0.5556	1.889	0.4913	0.1087	0.2173	0.5572	1.738	0.4437	0.1563	0.3125	0.5167	1.777
0.5250	0.3500	0.1750	0.3500	0.4487	1.782	0.4525	0.0725	0.1450	0.4745	1.613	0.3922	0.1328	0.2655	0.4156	1.669
0.4500	0.3000	0.1500	0.3000	0.3571	1.690	0.4145	0.0354	0.0710	0.4151	1.494	0.3396	0.1103	0.2207	0.3294	1.575
0.4200	0.2800	0.1400	0.2800	0.3241	1.657	0.3996	0.0204	0.0408	0.3983	1.447	0.3183	0.1017	0.2034	0.2984	1.541
0.3825	0.2550	0.1275	0.2550	0.2852	1.619	0.3811	0.0014	0.0027	0.3835	1.387	0.2913	0.0912	0.1823	0.2778	1.611

TABLE 4
 $m = 2, \Lambda = 0.8, \tau_1 = 1.0, \tau_2 = 5.0, \sigma_2 = 2.0$

σ_1	$\lambda_1 = \rho_1$	λ_2	ρ_2
0.1	0.7404	0.0596	0.2980
0.2	0.738	0.0610	0.305
0.3	0.736	0.0633	0.317
0.4	0.733	0.0663	0.331
0.6	0.726	0.0738	0.369
0.8	0.717	0.0823	0.411
1.0	0.709	0.0908	0.454
1.2	0.701	0.0988	0.494
1.4	0.693	0.106	0.531
1.5	0.690	0.109	0.548
2.0	0.676	0.124	0.621
2.5	0.665	0.135	0.675
3.0	0.656	0.143	0.717
4.0	0.644	0.155	0.775
5.0	0.637	0.162	0.814
6.0	0.631	0.168	0.842
10.0	0.619	0.180	0.901
15.0	0.613	0.186	0.932

$> \rho_1$

higher. They differed for lower loads considerably, and the relative efficiency of these two was quite sensitive to the variances of the service durations. Both assignments performed *much* better than the proportional allocation, at all loads (except, of course, at *very* close to saturation). That this improvement was sensitive to the variance is natural; we found, however, that the improvement declined with increasing variance ratios, at all loads (but was still always larger than the difference between the results of the allocations (11) and (20)). Cutoff values of the load (that is, values of total utilization where (20) recommends dropping the slowest server from service), can be computed directly from the solution of (15) for Λ with $\alpha\Lambda = \max(\tau_i)$. The solution is immediate and its dependence on the parameters is clear from (15).

27. We consider now briefly a situation where value is placed on assuring regular, as well as reasonably fast service. This means that we undertake to minimize not just the expected response time but some dispersion measure as well. We define an object function for each server

$$(31) \quad U_i = E(R_i) + \zeta_i V(R_i), \quad V(R_i) = \frac{\lambda_i E(S_i^3)}{1 - \lambda_i \tau_i} + \frac{\lambda_i^2 E^2(S_i^2)}{4(1 - \lambda_i \tau_i)^2} + \sigma_i^2,$$

where ζ_i is a positive constant, probably different for different servers. We want to minimize

$$(32) \quad U = \sum \frac{\lambda_i}{\Lambda} U_i.$$

A routine check verifies that U is indeed convex in λ_i under constraints (5). We remark that we preferred using U , to the perhaps more natural U' defined through

$$(33) \quad U'_i = E(R_i) + \zeta_i [V(R_i)]^{1/2};$$

the reason being that U' is not convex in λ_i , making the assurance of an optimum a very different (and considerably more difficult, usually) problem. On the other hand, by calibrating ζ_i judiciously we can compensate for the change of units. Applying the above methods to U does not give us as simple an equation for the λ_i ; actually we end up with a system of $m + 1$ equations which have to be solved simultaneously. The first m are fourth degree algebraic equations in λ_i (one for every λ_i)² and the last one is similar to (18). We did not investigate the properties of the solutions of this system. Note, however, that the case presented in § 21 is obtained from this one if we let all $\zeta_i \rightarrow 0$.

Conclusions and extensions.

28. The physical situations that we considered, characterized by a unique source-server relationship, can be regarded from the point of view of the system as stochastic scheduling. That this scheduling, when such a relationship does not exist, is not the most efficient, is evident. Indeed, even when limitations are still placed on the way customers join the system, more efficient regimes can be found. We intend to further consider systems where the server that will serve a given customer is decided only when the customer reaches the system, not beforehand. Even then, we shall find that situations where different amounts of information about the state of the system are available, dictate wholly different allocation schemes. When no information at all is available—except the capabilities of the servers—the above stochastic scheduling is the best we can do. Even a slight increase in the information available can improve performance remarkably. As an example, consider a system with two identical servers. Stochastic scheduling, under any criterion, would designate each customer with probability $\frac{1}{2}$ to join each server. Now we imagine the situation where customers are assigned to the two servers alternately; that is, the information as to which server the *last* arrival was assigned is available. For service durations of any distribution, standard results in [5, p. 224] inform us that the second setup results in shorter waiting times, and that when the rate of input increases to saturate the system, the *difference* in the mean waiting times that would be obtained under the two different regimes goes to infinity! This is appealing to our intuition and we consider pursuing this in a subsequent paper.

REFERENCES

- [1] PETER P. S. CHEN, *Optimal partition of input load to parallel exponential servers*, Fifth Annual Southeastern Symposium on System Theory, North Carolina State University, Raleigh, 1973.
- [2] ———, *Optimal file allocation*, Doctoral thesis, Harvard University, Cambridge, Mass., 1973.
- [3] LEONARD KLEINROCK, *Communications Nets: Stochastic Message Flow and Delay*, Dover, New York, 1972.
- [4] ANTHONY V. FIACCO AND GARTH P. MCCORMICK, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley, New York, 1968.
- [5] ROBERT B. COOPER, *Introduction to Queueing Theory*, Macmillan, New York, 1972.

² Thus these equations can yield analytic expressions for λ_i ; we do not believe, however, that it would be transparent enough for interpretation.

STRUCTURAL STABILITY FOR THE RICCATI EQUATION*

R. S. BUCY†

Abstract. We consider a matrix Riccati equation with respect to changes in the phase portrait induced by continuously differential deformations of the coefficient matrices. Necessary and sufficient conditions are given which characterize structural stability of an equilibrium. Further examples of birth, death and coalescence of equilibria are given.

Introduction. Recently, certain symmetric equilibrium solutions of the matrix Riccati equation, which are indefinite, have been discovered (see [1] and [2]). Furthermore, in certain cases, r -parameter families of such indefinite solutions can exist. The indefinite solutions have a saddle-type phase portrait and consequently cannot be determined numerically by numerical integration of the Riccati equation. Thom, in [3] and its references, has proposed a modified structural stability for equilibria of differential equations and has investigated the bifurcation of equilibria in the case in which they are structurally unstable, under the assumption that the right-hand side of the differential equation is the gradient of a potential function. His definition is interesting in that only those equilibria which are structurally stable possess numerical reality, i.e., could be determined by calculations, subject to roundoff error.

In this paper, we give necessary and sufficient conditions for the structural stability of equilibria of the Riccati equation. We further show that, when an infinity of equilibria of the Riccati equation exists, they are structurally unstable. Finally, in the case of a two-dimensional Riccati equation, several examples of bifurcation are given which illustrate the richness of the theory. A deeper aim of this paper is to suggest to workers in control theory that certain algebraic pathologies which arise in linear autonomous control theory may not even exist in the real world where no calculation is exact.

1. Notation. In general, capital italic letters will refer to $d \times d$ real-entried matrices, while lowercase boldface letters are real column vectors. By A' we denote the transpose of the matrix A , while $A \geq 0$ denotes that the symmetric matrix A is positive semidefinite (i.e., $\mathbf{x}'A\mathbf{x} \geq 0$ for all \mathbf{x}). We consider the algebraic matrix Riccati equation

$$(1.0) \quad S(P) \equiv FP + PF' - PMP + C = 0,$$

with $M = M' \geq 0$ and $C = C'$, where F, M, C are given. A symmetric P satisfying (1.0) is sought. We consider the natural map j from $d \times d$ matrices to R^{d^2} , $P \mapsto \mathbf{p}$, which explicitly is: if $P = \{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_d\}'$, then

$$\mathbf{p} = \begin{pmatrix} \mathbf{p}_1 \\ \vdots \\ \mathbf{p}_d \end{pmatrix}.$$

* Received by the editors August 2, 1973, and in revised form March 8, 1974.

† Aerospace Engineering Department, University of Southern California, Los Angeles, California 90007 and Laboratoire d'Automatique Toulouse, France. This research was supported in part by the United States Air Force Office of Aerospace Research under Grant AF OSR-71-2141.

Using this map, (1.0) becomes

$$(1.1) \quad (F \otimes I + I \otimes F)\mathbf{p} - (P \otimes P)\mathbf{m} + \mathbf{c} = 0,$$

where $\mathbf{p}, \mathbf{m}, \mathbf{c}$ are the images under j of P, M, C , respectively, and $A \otimes B$ is the Kronecker product $(a_{ij}B)$. We recall that the eigenvalues of $F \otimes I + I \otimes \bar{F}$ are $\lambda_i(F) + \lambda_j(\bar{F})i, j = 1, \dots, d$. Consider now the matrix Riccati equation

$$(1.2) \quad dP/dt = S(P)$$

and a symmetric equilibrium point $P^*, S(P^*) = 0$. Equation (1.2) can be considered in translated coordinates $D = P - P^*$ and takes the form

$$(1.3) \quad dD/dt = \bar{F}_* D + D \bar{F}'_* - DMD,$$

with $\bar{F}_* = F - P^*M$, or as a vector equation

$$(1.4) \quad \frac{d\mathbf{d}}{dt} = (\bar{F}_* \otimes I + I \otimes \bar{F}'_*)\mathbf{d} - (D \otimes D)\mathbf{m}.$$

With each equilibrium solution P^* , we can associate the triplet of integers (n_1^*, n_0^*, n_{-1}^*) , the index of P^* , where

n_1^* = the number of eigenvalues of J with real part positive,

n_0^* = the number of eigenvalues of J with real part zero,

n_{-1}^* = the number of eigenvalues of J with real part negative,

where $J = \bar{F}_* \otimes I + I \otimes \bar{F}'_*$. We can now proceed to our main results.

2. Results. We consider a variant of (1.0):

$$(2.0) \quad S(P, t) = F(t)P + PF'(t) - PM(t)P + C(t),$$

with $F(\cdot), M(\cdot)$ and $C(\cdot)$ continuously differentiable matrix-valued functions of t , where $M'(t) = M(t)$ and $C'(t) = C(t)$.

DEFINITION 1. P^* , a symmetric equilibrium point of (1.0), is *structurally stable* if and only if for every C^1 choice of $F(\cdot), M(\cdot)$ and $C(\cdot)$, with $F(t_0) = F, M(t_0) = M$ and $C(t_0) = C$ for $t - t_0$ sufficiently small, there exists a $P(t) = P'(t)$, an equilibrium of (2.0), unique in a sufficiently small neighborhood of P^* , and further, $\text{index}(P^*) = \text{index}(P(t))$.

In other words, a small diffeomorphic perturbation of the coefficients of the Riccati equation topologically preserves the local phase portrait about a structurally stable equilibrium point. Note, if P is a solution of (2.0), then P' is also.

THEOREM 1. *A symmetric equilibrium P^* of (1.0) is structurally stable if and only if*

$$\det(\bar{F}_* \otimes I + I \otimes \bar{F}'_*) \neq 0.$$

Proof. Suppose $\det(\bar{F}_* \otimes I + I \otimes \bar{F}'_*) \neq 0$. Now, viewing (2.0) as n^2 equations $j(S(P, t)) = \mathbf{s}$ and $j(P) = \mathbf{p}$, we find

$$(2.1) \quad ds/d\mathbf{p} = \bar{F}(t) \otimes I + I \otimes \bar{F}(t),$$

where $\bar{F}(t) = F(t) - PM(t)$.

Hence

$$\left. \frac{ds}{d\mathbf{p}} \right|_{P=P^*, t=t_0} = \bar{F}_* \otimes I + I \otimes \bar{F}_*,$$

so that by assumption the Jacobian is nonsingular, and consequently the implicit function theorem shows that for $t - t_0$ sufficiently small, a unique $\mathbf{p}(t)$ exists in a neighborhood of \mathbf{p}^* . Consequently, $j^{-1}(\mathbf{p}(t)) = P(t)$ is a solution of (2.0), and is symmetric, since otherwise $P'(t)$ is a solution and hence $j^{-1}(P'(t)) = \mathbf{p}'(t)$ is a solution in any neighborhood of P^* containing $\mathbf{p}(t)$, contradicting local uniqueness. By assumption, $\text{index}(P^*) = (n_1^*, 0, n_{-1}^*)$, and for $t - t_0$ sufficiently small, $\text{index}(P^*) = \text{index}(P(t))$ by continuity of the eigenvalues of $P(t)$.

Conversely, suppose P^* is structurally stable but $\det(\bar{F}_* \otimes I + I \otimes \bar{F}_*) = 0$; then consider the following deformation:

$$F(t) = F + \sigma I, \quad C(t) = C - 2\sigma P^*, \quad M(t) = M,$$

with $\sigma = t - t_0$. The index of P^* has $n_0^* \neq 0$, and it is easy to see that by choosing σ however small but nonzero, we can achieve $\text{index}(P^*) \neq \text{index}(P(t))$, which is a contradiction.

Remarks. One can show that structural stability of P_+ and P_- is implied by detectability and identifiability of the associated model for the relevant control problem at least when C is nonnegative definite; see, in [1], the discussion following Theorem 2.5 in Remark 2.5. However, it is clear that structural stability of the P_θ is much more complex; in fact, Example 2 below shows that even a completely controllable and completely observable system can have structurally unstable indefinite equilibria. Our condition $\det(\bar{F}_* \otimes I + I \otimes \bar{F}_*) \neq 0$ requires that no repeated characteristic roots occur among the characteristic roots of \bar{F}_* , and all P_θ 's are consequently structurally stable if the characteristic roots of the Hamiltonian are simple, since the roots of $F - P_\theta M$ are roots of the Hamiltonian; see [1, Thm. 2.6].

We recall the results of Canabal [1]: the generalized Bass–Roth theorem says that every real equilibrium solution of the Riccati equation P is a solution of $(-P, I)\Delta(H) = 0$, where H is Hamiltonian and Δ is a *real* polynomial of degree d possessing roots all of which are eigenvalues of H . Different choices of d of the $2d$ roots of H to form Δ give rise to different P 's. For convenience, positive semidefinite, indefinite and negative semidefinite solutions of (1.0) will be labeled as P_+ , P_θ and P_- , respectively. In the following examples, $d = 2$, and one can see the possible bifurcation behavior from structurally unstable P_θ 's.

Example 1. Let

$$F = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad M = C = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix},$$

$$\text{all } P_+ = I, \quad P_- = -I; \quad P_\theta = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix},$$

Consider the perturbed problem

$$F = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad M = C = \begin{bmatrix} \varepsilon & 1 \\ 1 & 1 \end{bmatrix},$$

$$P_+ = I, \quad P_- = -I, \quad P_{\theta_1} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad P_{\theta_2} = \begin{bmatrix} \frac{4 - (1 - \varepsilon)^2}{4 + (1 - \varepsilon)^2} & \frac{4(1 - \varepsilon)}{4 + (1 - \varepsilon)^2} \\ \frac{4(1 - \varepsilon)}{4 + (1 - \varepsilon)^2} & \frac{(1 - \varepsilon)^2 - 4}{4 + (1 - \varepsilon)^2} \end{bmatrix}.$$

We see here that under a small perturbation, the double point P_θ bifurcates into P_{θ_1} and P_{θ_2} .

Example 2. If

$$F = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad M = C = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

then

$$P_+ = I, \quad P_- = -I; \quad P_\theta = \begin{pmatrix} \cos \theta & \sin \theta \\ \sin \theta & -\cos \theta \end{pmatrix},$$

so that there is a one-parameter family of indefinite equilibria. Consider the perturbation

$$F = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad M = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad C = \begin{pmatrix} \varepsilon & 0 \\ 0 & 1 \end{pmatrix}.$$

If $\varepsilon > 0$, then

$$P_+ = \begin{pmatrix} \sqrt{\varepsilon} & 0 \\ 0 & 1 \end{pmatrix}, \quad P_- = \begin{pmatrix} -\sqrt{\varepsilon} & 0 \\ 0 & -1 \end{pmatrix}, \quad P_{\theta_1} = \begin{pmatrix} -\sqrt{\varepsilon} & 0 \\ 0 & 1 \end{pmatrix}, \quad P_{\theta_2} = \begin{pmatrix} \sqrt{\varepsilon} & 0 \\ 0 & -1 \end{pmatrix}.$$

If $\varepsilon < 0$, then there is no solution.

Example 3.

$$F = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad M = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad C = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix},$$

$$P_+ = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad P_- = \begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \text{no } P_\theta.$$

The perturbation is $C = \begin{pmatrix} 1 & 0 \\ 0 & \varepsilon \end{pmatrix}$. For $\varepsilon > 0$,

$$P_+ = \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{\varepsilon} \end{pmatrix}, \quad P_- = \begin{pmatrix} -1 & 0 \\ 0 & -\sqrt{\varepsilon} \end{pmatrix}, \quad P_{\theta_1} = \begin{pmatrix} 1 & 0 \\ 0 & -\sqrt{\varepsilon} \end{pmatrix}, \quad P_{\theta_2} = \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{\varepsilon} \end{pmatrix},$$

For $\varepsilon < 0$, no equilibria exist.

Example 4.

$$F = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad M = \begin{pmatrix} \varepsilon & 0 \\ 0 & 0 \end{pmatrix}, \quad \varepsilon \geq 0,$$

$$\alpha^2 = 2\beta - 2 + \varepsilon, \quad \beta^2 = 1 + \varepsilon,$$

where α, β are positive solutions. For $0 < \varepsilon < 8$,

$$P_+ = \frac{1}{\varepsilon} \begin{pmatrix} \alpha & \beta - 1 \\ \beta - 1 & \alpha\beta \end{pmatrix}, \quad P_- = \frac{1}{\varepsilon} \begin{pmatrix} -\alpha & \beta - 1 \\ \beta - 1 & -\alpha\beta \end{pmatrix}, \quad \text{no } P_\theta.$$

For $\varepsilon = 8$,

$$P_+ = \frac{1}{4} \begin{pmatrix} \sqrt{3} & 1 \\ 1 & 3\sqrt{3} \end{pmatrix}, \quad P_- = \frac{1}{4} \begin{pmatrix} -\sqrt{3} & 1 \\ 1 & -3\sqrt{3} \end{pmatrix}, \quad P_\theta = \begin{pmatrix} 0 & -\frac{1}{2} \\ -\frac{1}{2} & 0 \end{pmatrix},$$

For $\varepsilon > 8$,

$$P_+ = \frac{1}{\varepsilon} \begin{pmatrix} \alpha & \beta - 1 \\ \beta - 1 & \alpha\beta \end{pmatrix}, \quad P_- = \frac{1}{\varepsilon} \begin{pmatrix} -\alpha & \beta - 1 \\ \beta - 1 & -\alpha\beta \end{pmatrix},$$

$$P_{\theta_1} = \frac{1}{\varepsilon} \begin{pmatrix} \sqrt{\alpha^2 - 4\beta} & -\beta - 1 \\ -\beta - 1 & -\beta\sqrt{\alpha^2 - 4\beta} \end{pmatrix}, \quad P_{\theta_2} = \frac{1}{\varepsilon} \begin{pmatrix} -\sqrt{\alpha^2 - 4\beta} & -\beta - 1 \\ -\beta - 1 & \beta\sqrt{\alpha^2 - 4\beta} \end{pmatrix}.$$

For $\varepsilon \leq 0$, there are no solutions.

3. Conclusions. We have given necessary and sufficient conditions for the structural stability for equilibria of the Riccati equation. In particular, P_θ 's which form an r -parameter family of equilibria are structurally unstable, as the Hamiltonian has repeated roots and hence the determinant in Theorem 1 vanishes. The physical meaning of P_+ and P_- is clear in both filtering and control; however, the P_θ 's seem to defy physical interpretation. Of course, they correspond to spectral factorizations which have both stable and unstable modes, but this type of factorization seems artificial. Finally, Thom's modified structural stability seems to provide a convenient way to separate algebraic pathology from algebra, which is relevant to digital computation.

REFERENCES

- [1] J. RODRIGUEZ CANABAL, *Geometry of the Riccati equation*, Stochastics, 1 (1973), pp. 129–149.
- [2] J. C. WILLEMS, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automatic Control, 16 (1971), pp. 621–634.
- [3] R. THOM, *Topological models in biology*, Topology, 8 (1969), pp. 313–335.

CONVEX CONTROL PROBLEMS OF BOLZA IN HILBERT SPACES*

VIOREL BARBU†

Abstract. Conditions characterizing optimal arcs for certain problems of Bolza in Hilbert spaces are obtained by using methods of convex analysis and of maximal monotone operators.

1. Introduction. This paper concerns optimality conditions for a control problem of the form

$$(1.1) \quad \text{Minimize} \quad \int_0^T L(x(t), u(t)) dt + l(x(0), x(T))$$

subject to

$$(1.2) \quad x'(t) + A(t)x(t) = u(t), \quad 0 < t < T,$$

$$(1.3) \quad x(t) \in K \quad \text{for } 0 \leq t \leq T,$$

where $A(t)$ is a family of linear, closed and densely defined operators in a Hilbert space H and K is a closed convex subset of H . The functions L and l are lower semicontinuous and convex from $H \times H$ to $]-\infty, +\infty]$.

Many of the optimal control problems in Hilbert spaces considered in the literature (see Lions [3]) are special cases of this problem.

The main result of this paper, Theorem 1 (formulated in § 3), corresponds to some sharp results (not entirely covered by the results of this paper) recently given by R. T. Rockafellar [9], [11] for convex problems of Bolza in finite-dimensional spaces (see also [8], [10]). However, the approach we use to derive optimality conditions is technically different from Rockafellar's and somewhat simpler and more direct.

The proof of Theorem 1 (given in § 4) is based on an existence result for a nonlinear boundary problem associated with (1.1), (1.2) and (1.3). This existence result is obtained by exploiting the monotonicity properties of the subgradient mappings.

For simplicity we restrict ourselves to the case where L is independent of t , but the main part of Theorem 1 remains true for $L(t, \cdot, \cdot)$ under certain measurability assumptions similar to those used in [9].

In § 5 we discuss a class of control problems for linear parabolic equations to which the general result is immediately applicable.

In order to avoid making the exposition too ponderous, we assume familiarity with the main results in convex analysis and in the theory of nonlinear maximal monotone operators. However, in § 2 we recall some definitions and we refer to lecture notes of Moreau [5], Brézis [1] and the survey of Rockafellar [6] for detailed information on the subject.

* Received by the editors September 15, 1973, and in final revised form February 11, 1974.

† Faculty of Mathematics, University of Iasi, Iasi, Romania.

2. Preliminaries. Let X be a real Banach space and X' its dual. For $x \in X$ and $x' \in X'$, we write (x, x') instead of $x'(x)$. If $\varphi: X \rightarrow]-\infty, +\infty]$ is any proper convex function on X we set

$$(2.1) \quad D(\varphi) = \{x \in X | \varphi(x) < +\infty\}$$

and denote by $\partial\varphi(x)$ the set of all $x' \in X'$ such that

$$\varphi(x) < \varphi(y) + (x - y, x') \quad \text{for all } y \in X.$$

The multivalued mapping $\partial\varphi: X \rightarrow X'$ is called the *subdifferential* of φ , and every $x' \in \partial\varphi(x)$ is said to be a subgradient of φ at x . If φ is lower semicontinuous on X , then $\partial\varphi$ is a maximal monotone subset of $X \times X'$. Moreover, $\text{int } D(\varphi) \subset D(\partial\varphi)$ (see, e.g., [1], [5]). In particular, if φ is the indicator function of a closed convex subset $K \subset X$, then $D(\varphi) = D(\partial\varphi) = K$ and $\partial\varphi(x)$ coincides with the cone of normals to K at the point x .

Let $[0, T]$ be a fixed real interval ($0 < T < +\infty$) and let $C(0, T; X)$ denote the usual Banach space of continuous functions from $[0, T]$ to X . Let $\mathcal{M}(0, T; X')$ be the dual of $C(0, T; X)$, i.e., the Banach space of all finite, X' -valued regular measures on $]0, T[$. If K is any closed convex subset of X , we denote by \mathcal{K} the subset of $C(0, T; X)$ defined by

$$\mathcal{K} = \{x \in C(0, T; X) | x(t) \in K \text{ for every } t \in [0, T]\}.$$

For every $x \in \mathcal{K}$, we denote by $\mathcal{N}(x, K)$ the cone of normals to \mathcal{K} at x , i.e.,

$$(2.2) \quad \mathcal{N}(x, K) = \{\mu \in \mathcal{M}(0, T; X') | \mu(x - y) \geq 0 \text{ for all } y \in \mathcal{K}\}.$$

Next, we assume that X is a reflexive Banach space. Let

$$L: X \times X \rightarrow]-\infty, +\infty]$$

be a lower semicontinuous convex function on the product space $X \times X$, non-identically $+\infty$. The Hamiltonian function corresponding to L is the conjugate of the convex function $v \rightarrow L(x, v)$ for each $x \in X$. In other words,

$$(2.3) \quad H(x, p) = \sup \{(v, p) - L(x, v) | v \in X\}.$$

Since the function $v \rightarrow L(x, v)$ is lower semicontinuous, one has (see [5])

$$(2.4) \quad L(x, v) = \sup \{(v, p) - H(x, p) | p \in X'\}.$$

Moreover, the function $H(x, p)$ is concave as a function of x and convex lower semicontinuous as a function of p . We denote by

$$\partial H(x, p) = \{-\partial_x H(x, p), \partial_p H(x, p)\}$$

the set of all subgradients of the concave-convex function H at the point (x, p) . In other words,

$$(2.5) \quad \partial_x H(x, p) = \{u \in X' | H(x, p) \geq H(y, p) + (x - y, u) \text{ for all } y \in X\},$$

$$(2.6) \quad \partial_p H(x, p) = \{v \in X | H(x, p) \leq H(x, q) + (v, p - q) \text{ for all } q \in X'\}.$$

Since H is the *partial conjugate* of the lower semicontinuous proper convex function L it follows from a result due to Rockafellar (see [7, Thm. 9]) that the mapping $(x, p) \rightarrow \partial H(x, p)$ from $X \times X'$ into $X' \times X$ is *maximal monotone*.

We finally notice that by the properties of the conjugacy correspondence, one has

$$(2.7) \quad \partial_p H(x, p) = (\partial L_x)^{-1}(p),$$

where ∂L_x denotes the subdifferential of the convex function $L(x, \cdot)$.

3. Statement of the main result. We are given real Hilbert spaces V and H which satisfy

$$V \subset H \subset V'$$

with each inclusion mapping continuous and densely defined. Here V' denotes the dual of V , and H is identified with its own dual. Denote by (v, v') the pairing between v' in V' and v in V ; if $v, v' \in H$, this is the ordinary inner product in H . The norms in V and H will be denoted by $\|\cdot\|$ and $|\cdot|$ respectively. The norm in V' will be denoted by $\|\cdot\|_*$.

Let $[0, T]$ be a fixed real interval ($0 < T < +\infty$), and let $W(0, T)$ denote the space

$$(3.1) \quad W(0, T) = \{x \in L^2(0, T; V) | x' \in L^2(0, T; V')\},$$

where $d/dt = '$ denotes the derivative in the sense of V' -valued distributions on $]0, T[$ (i.e. in $\mathcal{D}'(0, T, V')$). It is well known (see [4, Chap. 1]) that $C(0, T; H) \supset W(0, T)$.

We now list the basic assumptions:

(A) We are given a family $\{A(t) | 0 < t < T\}$ of linear continuous operators from V into V' such that:

(a) For all u, v in V the function $t \rightarrow (A(t)u, v)$ is measurable on $]0, T[$ and

$$(3.2) \quad \|A(t)u\|_* \leq C\|u\| \quad \text{for all } u \in V,$$

with some positive constant C independent of t .

(b) There are α real and $\omega > 0$ such that

$$(3.3) \quad (A(t)u, u) + \alpha|u|^2 \geq \omega\|u\|^2 \quad \text{for all } u \in V.$$

(B) L and l are lower semicontinuous convex functions defined from $H \times H$ into $] -\infty, +\infty]$ and nonidentically $+\infty$. For every $(x, p) \in H \times H$ the Hamiltonian function

$$(3.4) \quad H(x, p) = \sup \{(p, v) - L(x, v) | v \in H\}$$

is finite.

(C) K is a closed convex subset of H . There exists $x \in W(0, T)$ such that $x' + A(t)x \in L^2(0, T; H)$, $L(x, x' + A(t)x) \in L^1(0, T)$, $(x(0), x(T)) \in D(l)$ and

$$(3.5) \quad x(t) \in \text{int } K \quad \text{for every } t \in [0, T].$$

Let C_L denote the set of all pairs $(x(0), x(T))$ arising from arcs $x \in W(0, T)$ such that $x' + A(t)x \in L^2(0, T; H)$, $L(x, x' + A(t)x) \in L^1(0, T)$ and $x(t) \in K$ for all $t \in [0, T]$.

The last assumption may be formulated as follows:

(D) There exists $(x_1, x_2) \in D(l) \cap C_L$ such that one of the following two conditions holds:

$$(3.6) \quad x_2 \in \text{int} \{x \in H \mid (x_1, x) \in D(l)\},$$

$$(3.7) \quad x_2 \in \text{int} \{x \in H \mid (x_1, x) \in C_L\}.$$

In particular, assumption (D) holds if $(\text{int } D(l)) \cap C_L$ or $D(l) \cap \text{int } C_L$ is nonempty.

Let $W_A(0, T)$ denote the space of all functions $x \in W(0, T)$ such that $x' + A(t)x \in L^2(0, T; H)$. Obviously, $W_A(0, T)$ is a Banach space with norm defined by

$$(3.8) \quad \|x\|_1^2 = |x(0)|^2 + \int_0^T |x' + A(t)x|^2 dt.$$

We consider the problem of minimizing a function of the form

$$(3.9) \quad F(x) = \int_0^T L(x, x' + A(t)x) dt + l(x(0), x(T))$$

subject to $x \in \mathcal{K}_0$, where

$$(3.10) \quad \mathcal{K}_0 = \{x \in W_A(0, T) \mid x(t) \in K \text{ for every } t \in [0, T]\}.$$

THEOREM 1. *Assume that hypotheses (A), (B), (C) and (D) hold. Then, in order that $x \in W_A(0, T)$ be an arc minimizing F on \mathcal{K}_0 it is necessary and sufficient that there exist a function $p \in L^\infty(0, T; H)$ and a measure $\mu \in \mathcal{N}(x, K)$ such that*

$$(3.11) \quad x' + A(t)x - \partial_p H(x, p) \ni 0 \quad \text{a.e. } t \in]0, T[,$$

$$(3.12) \quad p' - A^*(t)p + \partial_x H(x, p) - \mu \ni 0 \quad \text{in } \mathcal{D}'(0, T; V'),$$

$$(3.13) \quad p' - A^*(t)p - \mu \in L^2(0, T; H),$$

$$(3.14) \quad p \in L^2(0, T; V) \cap \text{BV}(0, T; V'),$$

$$(3.15) \quad \{p(0), -p(T)\} \in \partial l(x(0), x(T)).$$

In Theorem 1, A^* denotes the adjoint of A , i.e.,

$$(A^*(t)u, v) = (u, A(t)v) \quad \text{for all } u, v \in V,$$

and $\text{BV}(0, T; V')$ is the space of all functions $p: [0, T] \rightarrow V'$ of bounded variation on $[0, T]$. It is well known that for such a function p there exist $p(t^-)$ and $p(t^+)$ for all t . For simplicity we have written $p(0)$ instead of $p(0^+)$ and $p(T)$ instead of $p(T^-)$.

The conditions (3.11), (3.12) together with the boundary condition (3.15) may be regarded as Euler-Lagrange conditions for optimality. If $K = H$ (i.e., there are no state constraints), then $\mathcal{N}(x, K) = 0$ (see § 2), so that (3.13) and (3.14) imply that $p \in W(0, T)$ and $p' - A^*(t)p \in L^2(0, T; H)$. Thus in this case the condition (3.12) will be satisfied a.e. on $]0, T[$ with ordinary derivative $p'(t)$.

In the general case, as $p(t)$ is a function of bounded variation on $[0, T]$, its weak derivative \dot{p} defined by

$$\frac{d}{dt}(p(t), v) = (\dot{p}(t), v) \quad \text{for all } v \in V$$

exists a.e. on $]0, T[$ and belongs to $L^1(0, T; V')$. Then p' can be written as $p' = p \, dt + dp_0$, where dp_0 is the *singular part* of the measure p' with respect to Lebesgue measure. On the other hand, by (3.13) it follows that there exists $q \in \mathbf{BV}(0, T; V')$ such that $q' = \mu$. Since $\mu \in \mathcal{M}(0, T; H)$, it follows by using a standard device that $q \in \mathbf{BV}(0, T; H)$. Consequently, we can write μ as $\mu = q \, dt + dq_0$, where

$$q, q_0 \in \mathbf{BV}(0, T; H).$$

Since, in virtue of (3.13), $dp_0 - dq_0 \in L^1(0, T; V')$ we conclude that $dp_0 = dq_0$. This equality shows that the *singular part* of the measure p' coincides with the “singular contribution” dq_0 of the state constraints. It can also be proved that

$$\dot{q}(t) \in \partial I_K(x(t)) \quad \text{a.e. } t \in]0, T[,$$

where $\partial I_K(x)$ denotes the cone of normals to $K \subset H$ at x .

Moreover, it turns out that the functions p which appear in the optimality conditions (3.11)–(3.15) can be interpreted as minimizing arcs of a certain dual problem associated with (3.9) and (3.10). We expect to give details in a later paper.

4. Proof of Theorem 1. The main step in the proof of Theorem 1 is the following existence result.

PROPOSITION 1. *Assume that the hypotheses of Theorem 1 are satisfied. If we are given*

$$f \in L^\infty(0, T; H), \quad x_1 \in H,$$

then there exists a unique pair of functions $(x, p) \in W_A(0, T) \times L^\infty(0, T; H)$ and a measure $\mu \in \mathcal{M}(0, T; H)$ which satisfy

$$(4.1) \quad \mu \in \mathcal{N}(x, K), \quad p \in L^2(0, T; V) \cap \mathbf{BV}(0, T; V'),$$

$$(4.2) \quad p' - A^*(t)p - \mu \in L^2(0, T; H),$$

$$(4.3) \quad \{p' - A^*(t)p - \mu, p - x' - A(t)x + f(t)\} \in \partial L(x; x' + A(t)x) \\ \text{a.e. } t \in]0, T[,$$

$$(4.4) \quad \{p(0) - x(0) + x_1, -p(T)\} \in \partial l(x(0), x(T)).$$

Proof. For every $\lambda > 0$ let us denote by L_λ and l_λ the convex functions from $H \times H$ to $]-\infty, +\infty]$ defined by

$$L_\lambda(x, v) = \inf \{(|x - y|^2 + |v - u|^2)/2\lambda + L(y, u) \mid (y, u) \in H \times H\},$$

$$l_\lambda(y_1, y_2) = \inf \{(|y_1 - z_1|^2 + |y_2 - z_2|^2)/2\lambda + l(z_1, z_2) \mid (z_1, z_2) \in H \times H\}.$$

We set $\varphi = I_K$ (the indicator function of K) and in a similar manner define φ_λ , i.e.,

$$\varphi_\lambda(x) = \inf \{|x - y|^2/2\lambda \mid y \in K\}.$$

We recall (see [1], [5]) that L_λ , l_λ and φ_λ are Fréchet differentiable on $H \times H$ (respectively on H) and

$$(4.5) \quad \partial L_\lambda = \lambda^{-1}(1 - (1 + \lambda \partial L)^{-1}), \quad \partial l_\lambda = \lambda^{-1}(1 - (1 + \lambda \partial l)^{-1}),$$

$$(4.6) \quad \partial \varphi_\lambda = \lambda^{-1}(1 - (1 + \lambda \partial \varphi)^{-1}).$$

Here, we have written 1 instead of the identity function I on H .

(i) *Approximate equations.* For every λ , the function

$$F_\lambda(x) = \int_0^T (L_\lambda(x, x' + A(t)x) + \varphi_\lambda(x) - (x' + A(t)x, f(t)) + \frac{1}{2}|x' + A(t)x|^2) dt \\ + l_\lambda(x(0), x(T)) + \frac{1}{2}|x(0)|^2 - (x(0), x_1)$$

is continuous and convex from $W_A(0, T)$ to $]-\infty, +\infty]$. Since every nonempty level set $\{x \in W_A(0, T) | F_\lambda(x) \leq \beta\}$ is strictly convex and weakly compact in $W_A(0, T)$, there exists a unique $x_\lambda \in W_A(0, T)$ such that

$$(4.7) \quad F_\lambda(x_\lambda) \leq F_\lambda(y) \quad \text{for all } y \in W_A(0, T).$$

As the functions φ_λ , L_λ and l_λ are Fréchet differentiable on $H \times H$, (4.7) gives (suppressing the t)

$$(4.8) \quad \int_0^T [(u_\lambda^1, y) + (u_\lambda^2, y' + Ay) + (\partial \varphi_\lambda(x_\lambda), y) + (x'_\lambda + Ax_\lambda - f, y' + Ay)] dt \\ + (v_\lambda^1 + x_\lambda(0) - x_1, y(0)) + (v_\lambda^2, y(T)) = 0$$

for every $y \in W_A(0, T)$. Here we have used the notation

$$[u_\lambda^1, u_\lambda^2] = \partial L_\lambda(x_\lambda, x'_\lambda + Ax_\lambda), \quad [v_\lambda^1, v_\lambda^2] = \partial l_\lambda(x_\lambda(0), x_\lambda(T)).$$

Let $p_\lambda \in W(0, T)$ be such that

$$p'_\lambda - A^*p_\lambda = \partial \varphi_\lambda(x_\lambda) + u_\lambda^1 \quad \text{on }]0, T[$$

and

$$p_\lambda(T) = -v_\lambda^2.$$

The existence of such p_λ is implied by assumption (A) (see Lions [3, Chap. 3]). Then, by (4.8) we have

$$\int_0^T (x'_\lambda + Ax_\lambda - p_\lambda - f + u_\lambda^2, y' + Ay) dt + (x_\lambda(0) - p_\lambda(0) - x_1 + v_\lambda^1, y(0)) = 0$$

for all y in $W_A(0, T)$. Since the mapping $y \rightarrow \{y' + Ay, y(0)\}$ is a linear isomorphism from $W_A(0, T)$ onto $L^2(0, T; H) \times H$ (this is again a consequence of (A)) we deduce that

$$x'_\lambda + A(t)x_\lambda - p_\lambda - f + u_\lambda^2 = 0 \quad \text{a.e. } t \in]0, T[$$

and

$$x_\lambda(0) - p_\lambda(0) - x_1 + v_\lambda^1 = 0.$$

We have therefore proved that there exists a unique pair $(x_\lambda, p_\lambda) \in W_A(0, T) \times W_{A^*}(0, T)$ satisfying

$$(4.9) \quad \{p'_\lambda - A^*(t)p_\lambda - \partial\varphi_\lambda(x_\lambda), p_\lambda - x'_\lambda - A(t)x_\lambda + f(t)\} = \partial L_\lambda(x_\lambda, x'_\lambda + A(t)x_\lambda)$$

and the transversality conditions

$$(4.10) \quad \{p_\lambda(0) - x_\lambda(0) + x_1, -p_\lambda(T)\} = \partial l_\lambda(x_\lambda(0), x_\lambda(T)).$$

Here $W_{A^*}(0, T)$ denotes the space $p \in W(0, T)$, $p' - A^*(t)p \in L^2(0, T; H)$.

(ii) *A priori estimates.* Since L_λ , l_λ and φ_λ are uniformly bounded below by affine functions and there exists at least one arc $x \in D(F) \cap \mathcal{X}$, by (4.7) we deduce that

$$(4.11) \quad \|x_\lambda\|_1^2 = \int_0^T |x'_\lambda + A(t)x_\lambda|^2 dt + |x_\lambda(0)|^2 \leq C_1 \quad \text{for all } \lambda > 0.$$

In particular, the estimate (4.11), assumption (A) and the Gronwall inequality imply that

$$\{x_\lambda\} \text{ is bounded in } C(0, T; H) \cap L^2(0, T; V).$$

On the other hand, from (4.5) and (4.9) it follows that

$$\begin{aligned} & \{\lambda p'_\lambda - \lambda A^*(t)p_\lambda - \lambda \partial\varphi_\lambda(x_\lambda), \lambda p_\lambda - \lambda x'_\lambda - \lambda A(t)x_\lambda + \lambda f(t)\} \\ &= \{x_\lambda, x'_\lambda + A(t)x_\lambda\} - (1 + \lambda \partial L)^{-1}(x_\lambda, x'_\lambda + A(t)x_\lambda). \end{aligned}$$

Consequently for all $\lambda, \mu > 0$, we have

$$\frac{d}{dt}(\lambda p_\lambda - \mu p_\mu, x_\lambda - x_\mu) \geq (\lambda(x'_\lambda + A(t)x_\lambda - f(t)) - \mu(x'_\mu + A(t)x_\mu - f(t)),$$

$$x'_\lambda + A(t)x_\lambda - x'_\mu - A(t)x_\mu) \quad \text{a.e. } t \in]0, T[,$$

because $(1 + \lambda \partial L)^{-1}$ and $(1 + \lambda \partial \varphi)^{-1}$ are nonexpansive on $H \times H$ and H respectively. We notice that every $(x, p) \in W(0, T) \times W(0, T)$ the function

$$t \rightarrow (x(t), p(t))$$

is absolutely continuous on $[0, T]$. Integrating the above inequality from 0 to T and using the conditions (4.10), we find that (again suppressing t)

$$(4.12) \quad \int_0^T (\lambda(x'_\lambda + Ax_\lambda - f) - \mu(x'_\mu + Ax_\mu - f), x'_\lambda + Ax_\lambda - x'_\mu - Ax_\mu) dt \\ + (\lambda(x_\lambda(0) - x_1) - \mu(x_\mu(0) - x_1), x_\lambda(0) - x_\mu(0)) \leq 0.$$

We set

$$Y_\lambda = \{x'_\lambda + Ax_\lambda - f, x_\lambda(0) - x_1\}$$

and regard Y_λ as an element of the Hilbert space $L^2(0, T; H) \times H$ with the natural inner product $\langle \cdot, \cdot \rangle$. Then this inequality may be written as

$$\langle \lambda Y_\lambda - \mu Y_\mu, Y_\lambda - Y_\mu \rangle \leq 0 \quad \text{for all } \lambda, \mu > 0.$$

Since Y_λ is bounded, this inequality implies (see [2, Lemma 2.4]) that a subsequence (denoted again by x_λ) can be extracted from $\{x_\lambda\}$ such that for $\lambda \rightarrow 0$,

$$(4.13) \quad x_\lambda \rightarrow x \quad \text{strongly in } W_A(0, T).$$

In particular, it follows that

$$(4.14) \quad x_\lambda(t) \rightarrow x(t) \quad \text{uniformly on } [0, T] \text{ in } H.$$

On the other hand, we have

$$\varphi_\lambda(x) = \lambda |\partial \varphi_\lambda(x)|^2 / 2 \quad \text{for all } x \in H.$$

As $\{\int_0^T \varphi_\lambda(x_\lambda) dt\}$ is bounded, we deduce that

$$\lambda \partial \varphi_\lambda(x_\lambda) \rightarrow 0 \quad \text{strongly in } L^2(0, T; H).$$

Therefore

$$(1 + \lambda \partial \varphi)^{-1} x_\lambda \rightarrow x \quad \text{strongly in } L^2(0, T; H).$$

Since $D(\partial \varphi) = K$ and $x(t)$ is an H -valued continuous function on $[0, T]$, this implies that

$$(4.15) \quad x(t) \in K \quad \text{for every } t \in [0, T].$$

Further a priori estimates are obtained as follows. By assumption (C) there exists $x_0 \in W_A(0, T)$ such that $x_0(t) \in \text{int } K$ for every $t \in [0, T]$, $(x_0(0), x_0(T)) \in D(l)$ and

$$(4.16) \quad \int_0^T L(x_0, x'_0 + Ax_0) dt < +\infty.$$

Thus, there is $\rho > 0$ such that

$$(4.17) \quad x_0(t) + \rho w \in K \quad \text{for } t \in [0, T], \quad |w| = 1.$$

In the inequality

$$(\partial \varphi_\lambda(x_\lambda), x_\lambda - x_0 - \rho w) \geq \varphi_\lambda(x_\lambda) - \varphi_\lambda(x_0 + \rho w),$$

we take $w = \partial \varphi_\lambda(x_\lambda) / |\partial \varphi_\lambda(x_\lambda)|$. But $\varphi_\lambda(x_0 + \rho w) = 0$ by (4.17). Hence

$$(4.18) \quad \rho \int_0^T |\partial \varphi_\lambda(x_\lambda)| dt \leq \int_0^T (\partial \varphi_\lambda(x_\lambda), x_\lambda - x_0) dt.$$

On the other hand, by (4.9) we have

$$(p'_\lambda - A^* p_\lambda, x_\lambda - x_0) - (\partial \varphi_\lambda(x_\lambda), x_\lambda - x_0) + (p_\lambda - x'_\lambda - Ax_\lambda + f, x'_\lambda + Ax_\lambda - x'_0 - Ax_0) \geq L_\lambda(x_\lambda, x'_\lambda + Ax_\lambda) - L_\lambda(x_0, x'_0 + Ax_0) \quad \text{a.e. } t \in]0, T[.$$

Consequently

$$\begin{aligned} (\partial \varphi_\lambda(x_\lambda), x_\lambda - x_0) &\leq L_\lambda(x_0, x'_0 + Ax_0) + \frac{d}{dt}(p_\lambda, x_\lambda - x_0) \\ &\quad - L_\lambda(x_\lambda, x'_\lambda + Ax_\lambda) - (x'_\lambda + Ax_\lambda - f, x'_\lambda + Ax_\lambda - x'_0 - Ax_0). \end{aligned}$$

Integrating this inequality from 0 to T and using (4.11), (4.16) gives

$$\int_0^T (\partial\varphi_\lambda(x_\lambda), x_\lambda - x_0) dt \leq C_2 \quad \text{for all } \lambda > 0,$$

because $L_\lambda(y, v) \leq L(y, v)$ for all $\lambda > 0$ and every $(y, v) \in H \times H$. Thus, by (4.18) we obtain

$$(4.19) \quad \rho \int_0^T |\partial\varphi_\lambda(x_\lambda)| dt \leq C_2 \quad \text{for all } \lambda > 0.$$

According to assumption (B), $D(\partial H) = H \times H$. Since the mapping

$$(x, p) \rightarrow (-\partial_x H(x, p), \partial_p H(x, p)) = \partial H(x, p)$$

is maximal monotone from $H \times H$ into itself (see § 2), it follows by a well-known result due to Rockafellar (see [6]) that $\partial H(x, p)$ is locally bounded at every point $(x, p) \in H \times H$. Therefore, there are positive constants ρ , δ and C_3 such that

$$(4.20) \quad \sup \{|z| \mid z \in \partial_p H(x(t) + \rho w, 0)\} \leq C_3$$

and

$$(4.21) \quad -H(x(t) + \rho w, 0) \leq C_3$$

for every $t \in [T - \delta, T]$ and all $w \in H$ with $|w| = 1$. Let v_0 be an arbitrary point in $\partial_p H(x(t) + \rho w, 0)$. The equality (2.3) then implies that

$$L(x(t) + \rho w, v_0) = -H(x(t) + \rho w, 0),$$

so that

$$(4.22) \quad L_\lambda(x(t) + \rho w, v_0) \leq C_3 \quad \text{for } t \in [T - \delta, T] \quad \text{and} \quad |w| = 1.$$

In the inequality

$$\begin{aligned} & (p'_\lambda - A^*p_\lambda + \partial\varphi_\lambda(x_\lambda), x_\lambda - x - \rho w) + (p_\lambda - x'_\lambda - Ax_\lambda + f, x'_\lambda + Ax_\lambda - v_0) \\ & \geq L_\lambda(x_\lambda, x'_\lambda + Ax_\lambda) - L_\lambda(x + \rho w, v_0), \end{aligned}$$

we take

$$w = \frac{p'_\lambda - A^*p_\lambda - \partial\varphi_\lambda(x_\lambda)}{|p'_\lambda - A^*p_\lambda - \partial\varphi_\lambda(x_\lambda)|}.$$

Then, by (4.11) and (4.22) one obtains that

$$(4.23) \quad (\rho - |x_\lambda - x|)|p'_\lambda - A^*p_\lambda - \partial\varphi_\lambda(x_\lambda)| \leq |p_\lambda + f||x'_\lambda + Ax_\lambda - v_0| + C_4$$

a.e. $t \in]T - \delta, T[$

because $\partial_p H$ is locally bounded. As $x_\lambda(t)$ converges to $x(t)$ uniformly on $[0, T]$, we finally obtain

$$|p'_\lambda - A^*p_\lambda| \leq C_5(|\partial\varphi_\lambda(x_\lambda)| + |p_\lambda + f||x'_\lambda + Ax_\lambda - v_0| + 1).$$

The estimate (4.19) together with Schwarz's inequality then yields

$$(4.24) \quad \left(\int_t^T |p'_\lambda - A^*p_\lambda| ds \right)^2 \leq M \left(1 + \int_t^T |p_\lambda|^2 ds \right), \quad T - \delta \leq t \leq T,$$

for all sufficiently small $\lambda > 0$.

For the time being, let us assume that

$$(4.25) \quad \{p_\lambda(T)\} \text{ is bounded in } H.$$

By hypothesis (A) we have

$$(4.26) \quad \frac{d}{dt}|p_\lambda|^2 - 2\omega\|p_\lambda\|^2 \geq -2\alpha|p_\lambda|^2 + 2(p'_\lambda - A^*p_\lambda, p_\lambda) \quad \text{a.e. } t \in]0, T[.$$

Hence

$$|p_\lambda(t)| + \omega \int_t^T \|p_\lambda\| ds \leq |p_\lambda(T)| + \alpha \int_t^T |p_\lambda| ds + \int_t^T |p'_\lambda - A^*p_\lambda| ds.$$

The estimates (4.24), (4.25) and Gronwall's inequality then yield

$$(4.27) \quad |p_\lambda(t)| \leq C_6 \quad \text{for } T - \delta \leq t \leq T,$$

and λ sufficiently small. Let $v \in]0, T[$ be the lower bound of all $t_0 \in]0, T[$ with the property that $p_\lambda(t)$ is uniformly bounded on $[t_0, T]$ for all sufficiently small λ . By choosing ρ and δ such that (4.20) and (4.21) hold in the interval $[v - \delta, v + \delta]$, one deduces by the same procedure as before that (4.27) holds globally, i.e., there exists $C > 0$ independent of λ such that

$$(4.28) \quad |p_\lambda(t)| \leq C \quad \text{for } t \in [0, T]$$

and λ sufficiently small. Thus, from (4.19) and (4.23) one deduces in a similar manner that

$$(4.29) \quad \{p'_\lambda - A^*p_\lambda - \partial\varphi_\lambda(x_\lambda)\} \text{ is bounded in } L^2(0, T; H),$$

$$(4.30) \quad \{p'_\lambda - A'p_\lambda\} \text{ is bounded in } L^1(0, T; H).$$

Then the inequality (4.26) yields

$$(4.31) \quad \{p_\lambda\} \text{ is bounded in } L^2(0, T; V).$$

It remains to show (4.25). For every $(y_1, y_2) \in H \times H$ we denote by $\phi(y_1, y_2)$ the infimum of

$$G(y) = \int_0^T [L(y, y' + Ay) + \frac{1}{2}|y' + Ay|^2 - (f, y' + Ay)] dt$$

over all $y \in W_A(0, T)$ such that $y(t) \in K$ for every $t \in [0, T]$ and $y(0) = y_1, y(T) = y_2$. Clearly, the function ϕ is convex, lower semicontinuous on $H \times H$ and nowhere $-\infty$. Moreover, $D(\phi) = C_L$, and for every $(y_1, y_2) \in C_L$ the infimum defining $\phi(y_1, y_2)$ is attained. We first assume that condition (3.7) in hypothesis (D) holds. Then, there exist $y \in W_A(0, T)$ and $\rho > 0$ such that $(y(0), y(T)) \in D(l)$ and

$$(4.32) \quad \phi(y(0), y(T) + \rho w) \leq M_1 \quad \text{for } |w| = 1.$$

On the other hand, from (4.9) we see that

$$\begin{aligned} & (p_\lambda(T), x_\lambda(T) - y(T) - \rho w) - (p_\lambda(0), x_\lambda(0) - y(0)) \\ & \geq G_\lambda(x_\lambda) - \phi(y(0), y(T) + \rho w), \end{aligned}$$

where G_λ has the same form as G but with L_λ instead of L . By taking $w = p_\lambda(T)/|p_\lambda(T)|$ it follows by (4.32) that

$$\begin{aligned} \rho|p_\lambda(T)| &\leq \text{const.} + (p_\lambda(T), x_\lambda(T) - y(T)) \\ &\quad - (p_\lambda(0), x_\lambda(0) - y(0)) \quad \text{for all } \lambda > 0. \end{aligned}$$

By (4.10), the right-hand side of this inequality is bounded by $l_\lambda(y(0), y(T)) - l_\lambda(x_\lambda(0), x_\lambda(T)) + (x_1 - x_\lambda(0), x_\lambda(0) - y(0))$, so that

$$|p_\lambda(T)| \quad \text{is bounded}$$

because $l_\lambda(y(0), y(T)) \leq l(y(0), y(T))$ for all $\lambda > 0$.

Next, we assume that condition (3.6) holds and choose $y \in W_A(0, T)$ such that $(y(0), y(T)) \in C_L \cap D(l)$ and $y_\lambda(T) \in \text{int} \{u \in H | (y(0), u) \in D(l)\}$. By again using (4.10), we get

$$\begin{aligned} (p_\lambda(0), x_\lambda(0) - y(0)) - (p_\lambda(T), x_\lambda(T) - y(T) - \rho w) \\ \geq l_\lambda(x_\lambda(0), x_\lambda(T)) - l_\lambda(y(0), y(T) + \rho w) + (x_\lambda(0) - x_1, x_\lambda(0) - y(0)). \end{aligned}$$

Hence for sufficiently small λ we have

$$\rho|p_\lambda(T)| \leq \text{const.} + (p_\lambda(0), x_\lambda(0) - y(0)) - (p_\lambda(T), x_\lambda(T) - y(T))$$

because the function $u \rightarrow l(y(0), u)$ is locally bounded at $u = y(T)$. Again using (4.9) one easily deduces that the right-hand side of the above relation is bounded. Thus (4.25) is completely proved.

(iii) *Convergence of the approximate solutions.* It follows from (4.28), (4.29) and (4.30) that there exists a subsequence again denoted by $\{p_\lambda\}$ such that

$$\begin{aligned} p_\lambda &\rightarrow p \quad \text{weak* in } L^\infty(0, T; H), \\ p_\lambda &\rightarrow p \quad \text{in the weak topology of } L^2(0, T; V), \\ p'_\lambda &\rightarrow p' \quad \text{in } \mathcal{D}'(0, T; V'), \\ p'_\lambda - A^*p_\lambda - \partial\varphi_\lambda(x_\lambda) &\rightarrow q \quad \text{in the weak topology of } L^2(0, T; H), \\ \{p_\lambda(0), p_\lambda(T)\} &\rightarrow \{y_1, y_2\} \quad \text{in the weak topology of } H \times H. \end{aligned}$$

Now we pass to the limit in (4.9) and (4.10). As ∂L and ∂l are maximal monotone from $H \times H$ into itself, by (4.13) it follows that $\{x, x' + Ax\} \in D(\partial L)$ (see [1]) and

$$\begin{aligned} (4.33) \quad \{q, p - x' - Ax + f\} &\in \partial L(x, x' + Ax) \quad \text{a.e. } t \in]0, T[, \\ \{y_1 - x(0) + x_1, -y_2\} &\in \partial l(x(0), x(T)). \end{aligned}$$

Now, extracting a further subsequence, if necessary, by (4.30) we have

$$p'_\lambda - A^*p_\lambda \rightarrow p' - A^*p \quad \text{weak* in } \mathcal{M}(0, T; H).$$

The inequality

$$(\partial\varphi_\lambda(x_\lambda(t)), x_\lambda(t) - y(t)) \geq 0 \quad \text{for all } y \in C(0, T; H), \quad y \in \mathcal{K},$$

then implies that

$$(p' - A^*p - q)(x - y) \geq 0 \quad \text{for all } y \in \mathcal{K}.$$

In other words,

$$(4.34) \quad p' - A^*p - q \in \mathcal{N}(x, K).$$

Since $\{p'_\lambda\}$ is a bounded subset of $L^1(0, T; V')$ it follows that $p \in \text{BV}(0, T; V')$ (see, e.g., [1, Prop. A.5]). Furthermore, the following identity holds:

$$p'(\psi) = (y_2, \psi(T)) - (y_1, \psi(0)) - \int_0^T (p(t), \psi'(t)) dt$$

for every $\psi \in W(0, T)$.

Consequently,

$$p(0) = y_1, \quad p(T) = y_2.$$

Thus we have shown that x and p found as above satisfy the conditions (4.1), (4.2), (4.3) and (4.4). The uniqueness follows by the usual procedure. Thus the proof of Proposition 1 is complete.

Remark 1. From the inequality (4.24) it is clear that the mapping $f \rightarrow p$ is bounded from $L^2(0, T; H)$ to $L^\infty(0, T; H)$ on every bounded subset (in L^2 -norm) of $L^\infty(0, T; H)$.

Proof of Theorem 1. Let us denote again by F the convex function defined on $W_A(0, T)$ by

$$F(x) = \begin{cases} \int_0^T L(x, x' + Ax) dt + l(x(0), x(T)) & \text{if } x(t) \in K \text{ for } t \in [0, T], \\ +\infty & \text{otherwise.} \end{cases}$$

By assumption (D), $F \not\equiv +\infty$, and a standard argument based on Fatou's lemma shows that F is lower semicontinuous on $W_A(0, T)$.

We note that the dual space of $W_A(0, T)$ can be identified with $L^2(0, T; H) \times H$ under the pairing

$$(x(0), y_0) + \int_0^T (x' + Ax, y) dt = \langle x, [y_0, y] \rangle.$$

Let $\mathcal{A}: W_A(0, T) \rightarrow L^2(0, T; H) \times H$ be the mapping defined by

$$\mathcal{A}x = \{f, y_0\} \quad \text{for } x \in D(\mathcal{A}),$$

where $D(\mathcal{A})$ is the set of all $x \in W_A(0, T)$ such that $x(t) \in K$ for every $t \in [0, T]$ and there are $p \in L^\infty(0, T; H) \cap L^2(0, T; V)$ and $\mu \in \mathcal{N}(x, K)$ satisfying

$$(4.35) \quad p \in \text{BV}(0, T; V'), \quad p' - A^*p - \mu \in L^2(0, T; H),$$

$$(4.36) \quad \begin{aligned} x' + Ax - \partial_p H(x, p + f) &\ni 0, \\ p' - A^*p + \partial_x H(x, p, +f) - \mu &\ni 0, \end{aligned}$$

$$(4.37) \quad \{p(0) + y_0, -p(T)\} \in \partial l(x(0), x(T)).$$

From the properties of $H(x, p)$ we see that (4.36) may be written in the following form:

$$(4.38) \quad \{p' - A^*p - \mu, p + f\} \in \partial L(x, x' + Ax),$$

and by a simple calculation involving (4.37) and (4.38) it follows that $\mathcal{A} \subset \partial F$.

Let $x \rightarrow \Lambda x = \{x' + Ax, x(0)\}$ be the duality mapping (canonical isomorphism) from $W_A(0, T)$ onto its dual space $L^2(0, T; H) \times H$.

For every $(f, x_1) \in L^2(0, T; H) \times H$, the equation

$$\Lambda x + \mathcal{A}x = (f, x_1)$$

is obviously equivalent to problem (4.3), (4.4) with the conditions (4.1) and (4.2). Therefore, referring to Proposition 1, $R(\Lambda + \mathcal{A}) \supset L^\infty(0, T; H) \times H$. This implies that $R(\Lambda + \mathcal{A})$ is the whole space $L^2(0, T; H) \times H$, i.e., $\overline{\mathcal{A}}$ is maximal monotone from $W_A(0, T)$ into $L^2(0, T; H) \times H$. Here $\overline{\mathcal{A}}$ denotes the closure of \mathcal{A} . Hence $\overline{\mathcal{A}} = \partial F$. To conclude the proof, it suffices to show that

$$(4.39) \quad \overline{\mathcal{A}}^{-1}(0) = \mathcal{A}^{-1}(0).$$

Let $x \in W_A(0, T)$ be such that $0 \in \overline{\mathcal{A}}x = \partial F(x)$.

We have

$$x = (\Lambda + \overline{\mathcal{A}})^{-1}\Lambda x.$$

Let $\{f_n\} \subset L^\infty(0, T; H)$ be such that

$$f_n \rightarrow x' + Ax \quad \text{strongly in } L^2(0, T; H).$$

As $(\Lambda + \overline{\mathcal{A}})^{-1} = (\Lambda + \mathcal{A})^{-1}$ on $R(\Lambda + \mathcal{A})$ we deduce that

$$x_n = (\Lambda + \mathcal{A})^{-1}(f_n, x(0)) \rightarrow x \quad \text{strongly in } W_A(0, T)$$

and

$$(4.40) \quad \Lambda x_n + \mathcal{A}x_n \ni (f_n, x(0)).$$

Condition (4.40) may be written as

$$(4.41) \quad \begin{aligned} x'_n + Ax_n + \partial_p H(x_n, p_n - x'_n - Ax_n + f_n) &\ni 0, \\ p'_n - A^*p_n + \partial_x H(x_n, p_n - x'_n - Ax_n + f_n) - \mu_n &\ni 0 \end{aligned}$$

and

$$(4.42) \quad \{p_n(0) - x_n(0) + x(0), -p_n(T)\} \in \partial l(x_n(0), x_n(T)).$$

By the proof of Proposition 1 (see Remark 1) it follows that

$$(4.43) \quad \begin{aligned} \{p_n(0), p_n(T)\} &\text{ is bounded in } H \times H, \\ \{p_n\} &\text{ is bounded in } L^\infty(0, T; H) \times L^2(0, T; V). \end{aligned}$$

We have

$$\begin{aligned} (p'_n - A^*p_n - \mu_n, x_n - x - \rho w) \\ \geq -H(x_n, p_n - x'_n - Ax_n + f_n) + H(x + \rho w, p_n - x'_n - Ax_n + f_n) \end{aligned}$$

for all $w \in H$ and $\rho > 0$. On the other hand, by the first equation we obtain

$$H(x_n, p_n - x'_n - Ax_n + f_n) \leq H(x_n, 0) - (x'_n + Ax_n, p_n - x'_n - Ax_n + f_n).$$

We notice also the inequality

$$\begin{aligned} H(x + \rho w, 0) &\leq H(x + \rho w, p_n - x'_n - Ax_n + f_n) \\ &\quad - (\partial_p H(x + \rho w, 0), p_n - x'_n - Ax_n + f_n). \end{aligned}$$

By taking $w = (p'_n - A^*p_n - \mu_n)/|p'_n - A^*p_n - \mu_n|$ we deduce by the same argument as in the proof of Proposition 1 that

$$(4.44) \quad \{p'_n - A^*p_n - \mu_n\} \text{ is bounded in } L^2(0, T; H).$$

Let $x_0 \in W_A(0, T)$ be such that $x_0(t) \in \text{int } K$ for every $t \in [0, T]$ and

$$\int_0^T L(x_0, x'_0 + Ax_0) dt < +\infty.$$

Thus there exists $\rho > 0$ such that

$$(4.45) \quad \mu_n(x_n - x_0 - \rho w) \geq 0 \quad \text{for all } w \in S,$$

where S denotes the unit ball in $C(0, T; H)$. On the other hand, as in the proof of Proposition 1 one obtains

$$(4.46) \quad \mu_n(x_n - x_0) \leq \text{const.} \quad \text{for all } n.$$

By (4.45) and (4.46) it follows that

$$(4.47) \quad \{\mu_n\} \text{ is bounded in } \mathcal{M}(0, T; H).$$

It follows from (4.43), (4.44) and (4.47) that without loss of generality we may assume

$$\begin{aligned} p_n &\rightarrow p \quad \text{weak* in } L^\infty(0, T; H), \\ p'_n - A^*p_n &\rightarrow p' - A^*p \quad \text{weak* in } \mathcal{M}(0, T; H), \\ \mu_n &\rightarrow \mu \quad \text{weak* in } \mathcal{M}(0, T; H), \\ p'_n - A^*p_n - \mu_n &\rightarrow q \quad \text{weakly in } L^2(0, T; H). \end{aligned}$$

Since $\{p'_n\}$ is bounded in $\mathcal{M}(0, T; V')$ we deduce that $p \in \text{BV}(0, T; V')$. Thus one finally obtains that

$$\begin{aligned} x' + Ax - \partial_p H(x, p) &\ni 0, \\ p' - A^*p + \partial_x H(x, p) - \mu &\ni 0, \\ \{p(0), -p(T)\} &\in \partial l(x(0), x(T)). \end{aligned}$$

Hence $0 \in \mathcal{A}x$. This completes the proof of Theorem 1.

5. Examples. The main difficulty which arises in applications is to verify hypotheses (C) and (D) stated earlier. We begin by considering some important cases where these basic assumptions could be easily verified. The notation is that used in § 2 and § 3.

A weaker form of (C) is:

(C₀) There exists a function $x \in W_A(0, T)$ such that

$$(5.1) \quad x' + A(t)x - \partial_p H(x(t), 0) \ni 0 \quad \text{a.e. } t \in]0, T[,$$

$$(5.2) \quad x(t) \in \text{int } K \quad \text{for every } t \in [0, T], \quad (x(0), x(t)) \in D(l).$$

Indeed, if (5.1) holds, then

$$L(x(t), x'(t) + A(t)x(t)) = -H(x(t), 0) \in L^1(0, T),$$

because of (2.3) and (B).

Similarly, Assumption (D) holds if:

(D₀) There exist $x \in W_A(0, T)$ and $\rho > 0$ such that

$$(5.3) \quad x' + A(t)x - \partial_p H(x(t), 0) \ni 0 \quad \text{a.e. } t \in]0, T[,$$

$$x(t) \in K \quad \text{for } t \in [0, T] \quad (x(0), x(T)) \in D(l),$$

and one of the following two conditions holds:

(a) For all $w \in H, |w| = 1$, $(x(0), x(T) + \rho w)$ is an endpoint pair for the solutions $y(t)$ of (5.3) satisfying $y(t) \in K$ for every $t \in [0, T]$.

(b) $(x(0), x(T) + \rho w) \in D(l)$ for all w in $H, |w| = 1$.

Now, we consider the particular case where $D(L) = H \times H$ and $L(y(t), v(t)) \in L^1(0, T)$ for every pair $(y, v) \in L^2(0, T; H)$. Assume that A is independent of t and denote again by A its restriction to H .

In addition, suppose that A is positive and denote by $S(t)$ the semigroup of linear contractions generated by A on H . We shall assume that

$$(5.4) \quad S(t)K \subset K \quad \text{for all } t > 0.$$

Then both assumptions (C) and (D) are implied by the following:

(H₀) There exist an arc $x \in W_A(0, T)$ and $(x_1, x_2) \in (K \times K) \cap D(l)$ such that $(x(0), x(T)) \in D(l)$, $x(t) \in \text{int } K$ for all $t \in [0, T]$ and

$$(5.5) \quad x_2 \in \text{int } K.$$

We note that in this case C_L coincides with the set of all endpoint pairs $(x(0), x(T))$, where $x \in W_A(0, T)$ and $x(t) \in K$ for every $t \in [0, T]$. Since (C) is obviously satisfied we restrict attention to verifying (D). Let $\rho > 0$ be such that $x_2 + \rho w \in K$ for all $w \in H, |w| = 1$. Let

$$y(t) = (1 - t/T)S(t/T)x_1 + (t/T)S(1 - t/T)(x_2 + \rho w), \quad 0 \leq t \leq T.$$

By (5.4) it follows that $y(t) \in K$ for every $t \in [0, T]$. Moreover, $y \in W(0, T)$ and $dy/dt + Ay \in L^2(0, T, H)$. Thus we have proved that

$$x_2 \in \text{int } \{y \in H | (x_1, y) \in C_L\}$$

as claimed.

Finally, we consider an example from partial differential equations. For other problems in this field to which our result is immediately applicable we refer to the book of Lions [3].

Let Ω be a bounded open subset of R^n and let

$$Ay = - \sum_{i,j=1}^n \frac{\partial}{\partial x_i} \left(a_{ij}(x) \frac{\partial y}{\partial x_j} \right)$$

be the partial differential operator from $H_0^1(\Omega)$ into $H^{-1}(\Omega)$ defined by

$$(A\varphi, \psi) = \sum_{i,j=1}^n \int_{\Omega} a_{ij}(x) \frac{\partial \varphi}{\partial x_i} \frac{\partial \psi}{\partial x_j} dx, \quad \varphi, \psi \in H_0^1(\Omega),$$

where $a_{ij} \in L^\infty(\Omega)$. Here $H_0^1(\Omega)$ and $H^{-1}(\Omega)$ are the usual Sobolev spaces. We shall assume that A is strongly elliptic on Ω .

Consider the following optimal control problem:

$$(5.6) \quad \text{Minimize } F(y, u) = \int_0^T L(y(t), u(t)) dt + f(y(0), y(T))$$

in $y \in L^2(0, T; H_0^1(\Omega))$ and $u \in L^2(0, T; L^2(\Omega))$ subject to the constraints

$$(5.7) \quad \partial y / \partial t + Ay = u \quad \text{in } Q = \Omega \times]0, T[,$$

$$(5.8) \quad y(x, 0) \in Y_0, \quad y(x, T) \in Y_T \quad \text{in } \Omega,$$

$$(5.9) \quad y(t, \cdot) \in K, \quad t \in [0, T],$$

where L and f are lower semicontinuous proper convex functions from $L^2(\Omega) \times L^2(\Omega)$ to $] -\infty, +\infty]$ and K, Y_0, Y_T are nonempty closed convex subsets of $L^2(\Omega)$. More precisely, the following conditions will be assumed:

(a) $D(f) = L^2(\Omega) \times L^2(\Omega)$ and $L(y(t), u(t)) \in L^1(0, T)$ for every pair $(y, u) \in L^2(Q) \times L^2(Q)$, $Q = \Omega \times]0, T[$.

(b) For every $y \in L^2(\Omega)$,

$$L(y, v) / \|v\|_{L^2(\Omega)} \rightarrow +\infty \quad \text{as } \|v\|_{L^2(\Omega)} \rightarrow +\infty.$$

(c) $Y_0 \cap K \neq \emptyset$, $Y_T \cap \text{int } K \neq \emptyset$ and

$$(5.10) \quad (I + \lambda A)^{-1} K \subset K \quad \text{for all } \lambda > 0,$$

where A denotes the restriction to $L^2(\Omega)$ of the above elliptic operator.

(d) There exists an arc $y \in W_A(0, T)$ such that $y(0, x) \in Y_0$, $y(T, x) \in Y_T$ and

$$(5.11) \quad y(t, \cdot) \in \text{int } K \quad \text{for all } t \in [0, T].$$

We note that conditions (A), (B) and (H₀) hold. Then we can apply Theorem 1 where $V = H_0^1(\Omega)$ and $H = L^2(\Omega)$. F is the function defined by (5.6) and

$$l(y_1, y_2) = \begin{cases} f(y_1, y_2) & \text{if } y_1 \in Y_0 \text{ and } y_2 \in Y_T, \\ +\infty & \text{otherwise.} \end{cases}$$

In this case $\partial l(y_1, y_2) = f(y_1, y_2) + \{\mathcal{N}_1(y_1), \mathcal{N}_2(y_2)\}$ for $y_1 \in Y_0$, $y_2 \in Y_T$. Here $\mathcal{N}_1(y_1)$ and $\mathcal{N}_2(y_2)$ denote the cone of normals to Y_0 (respectively Y_T) at the point y_1 (respectively y_2).

Therefore, $y \in L^2(0, T; H_0^1(\Omega))$ and $u \in L^2(Q)$ form an extremal pair for the

above control problem if there exist a function

$$p \in L^\infty(0, T; L^2(\Omega)) \cap L^2(0, T; H_0^1(\Omega)) \cap BV(0, T; H^{-1}(\Omega))$$

and an $L^2(\Omega)$ -valued measure μ on $]0, T[$ such that

$$(5.12) \quad \partial y / \partial t + Ay = u \quad \text{a.e. in } Q,$$

$$(5.13) \quad \partial p / \partial t - A^*p + \partial_y H(y, p) \ni \mu \quad \text{in } \mathcal{D}'(Q),$$

$$(5.14) \quad u \in \partial_p H(y, p) \quad \text{in } Q,$$

$$(5.15) \quad y \in K, \quad \mu(y - \varphi) \geq 0 \quad \text{for all } \varphi \in C(0, T; L^2(\Omega)),$$

$$\varphi(t) \in K, \quad t \in [0, T]$$

and

$$(p(0, x), -p(T, x)) \in \partial f(y(0, x), y(T, x)) + (\mathcal{N}_1(y(0, x)), \mathcal{N}_2(y(T, x))).$$

Furthermore,

$$\partial p / \partial t - A^*p - \mu \in L^2(Q).$$

In particular, if K is the ball with center 0 and radius r in $L^2(\Omega)$, the conditions (5.10) and (d) are obviously satisfied and (5.15) is equivalent to

$$(5.16) \quad \mu(y) = r \|\mu\|_{\mathcal{M}(0, T; L^2(\Omega))}.$$

Now we shall take $A = -d^2/dx^2$, $\Omega = (a, b)$ and

$$(5.17) \quad K = \{y \in L^2(\Omega) | y \geq 0 \text{ a.e. on } \Omega\}.$$

Let K_0 be the subset of K defined by

$$K_0 = \{y \in L^\infty(\Omega) | y \geq 0 \text{ a.e. on } \Omega\}$$

and let \mathring{K}_0 be the interior of K_0 with respect to the L^∞ -topology on Ω .

Suppose that (a), (b) hold and, in addition,

$$(c') \quad Y_0 \cap K_0 \neq \emptyset, \quad Y_T \cap \mathring{K}_0 \neq \emptyset.$$

(d') There exists $y \in W_A(0, T)$ such that $y(0, x) \in Y_0$, $y(T, x) \in Y_T$ and $y(t, \cdot) \in \mathring{K}_0$ for all $t \in [0, T]$. By the maximum principle we deduce immediately that $(I + \lambda A)^{-1} K_0 \subset K_0$ for all $\lambda > 0$. Thus (5.4) holds and, arguing as above, one deduces that assumption (C) holds with \mathring{K}_0 instead of $\text{int } K$ and the condition (3.7) in assumption (D) is satisfied with $L^\infty(\Omega)$ interior instead of $L^2(\Omega)$ interior. Under these assumptions, in the proof of Theorem 1, the inequality (4.17) holds for all $w \in L^\infty(\Omega)$ such that $\|w\|_{L^2(\Omega)} = 1$. But for the proof this is enough because $\partial \varphi_\lambda(x_\lambda) / \|\partial \varphi_\lambda(x_\lambda)\|_{L^2(\Omega)}$ belongs to $L^\infty(\Omega)$. In fact, $x_\lambda \in H_0^1(\Omega)$ and

$$\partial \varphi_\lambda(x_\lambda) = \lambda^{-1}(x_\lambda - P_K x_\lambda),$$

where $P_K x = \max(0, x)$ for any $x \in L^2(\Omega)$. Similarly, as $p_\lambda(T) \in H_0^1(\Omega)$ it suffices to have (4.32) only for $w \in L^\infty(\Omega)$.

Thus the equations (5.12)–(5.15) characterize the minimizing arcs of the optimal control problem (5.6)–(5.9) where the state constraints are given by (5.17).

Interpreting μ as a scalar measure on $Q =]0, T[\times]a, b[$ we observe that (5.15) can be written as

$$\begin{aligned} y &\geq 0, \quad \mu \geq 0 \quad \text{on } Q, \\ \mu &= 0 \quad \text{on } \{(t, x) \in Q \mid y(t, x) > 0\}. \end{aligned}$$

Thus, the equations (5.12)–(5.14) can be written as

$$\begin{aligned} y_t - y_{xx} - \partial_p H(y, p) &\ni 0 \quad \text{in } Q, \\ p_t + p_{xx} + \partial_y H(y, p) &\geq 0 \quad \text{in } Q, \\ p_t + p_{xx} + \partial_y H(y, p) &\ni 0 \quad \text{in } \{(t, x) \in Q \mid y > 0\}, \\ y(t, x) &\geq 0 \quad \text{in } Q. \end{aligned}$$

REFERENCES

- [1] H. BREZIS, *Opérateurs maximaux monotones et semigroupes de contractions dans les espaces de Hilbert*, Cours de 3^{ème} cycle, Paris, 1971; Math. Studies, no. 5, North-Holland, Amsterdam, 1973.
- [2] M. G. CRANDALL AND A. PAZY, *Semigroups of nonlinear contractions and dissipative sets*, J. Functional Anal., 3 (1969), pp. 376–418.
- [3] J. L. LIONS, *Contrôle Optimal de Systèmes Gouvernés par des Équations aux Dérivées Partielles*, Dunod Gauthier–Villars, Paris, 1968.
- [4] J. L. LIONS AND E. MAGENES, *Problèmes aux Limites Nonhomogènes et Applications*, vol. I, Dunod Gauthier–Villars, Paris, 1967.
- [5] J. MOREAU, *Fonctionnelles convexes*, Lecture notes, Séminaire sur les équations aux dérivées partielles, Collège de France, 1966–1967.
- [6] R. T. ROCKAFELLAR, *Convex functions, monotone operators and variational inequalities*, Theory and Applications of Monotone Operators, A. Ghizzetti, ed., Oderisi, Gubbio, 1969, pp. 35–66.
- [7] ———, *Saddle-points and convex analysis*, Differential Games and Related Topics, H. W. Kuhn and G. P. Szegö, eds., North-Holland, Amsterdam, 1971, pp. 109–128.
- [8] ———, *Conjugate convex functions in optimal control and the calculus of variations*, J. Math. Anal. Appl., 32 (1970), pp. 174–222.
- [9] ———, *Existence and duality theorems for convex problems of Bolza*, Trans. Amer. Math. Soc., 159 (1971), pp. 1–40.
- [10] ———, *Dual problems of optimal control*, Technique of Optimization, A. V. Balakrishnan, ed., Academic Press, New York, 1972, pp. 423–432.
- [11] ———, *State constraints in convex control problems of Bolza*, this Journal, 10 (1972), pp. 691–715.

UNCONSTRAINED LAGRANGIANS IN NONLINEAR PROGRAMMING*

O. L. MANGASARIAN†

Abstract. The main purpose of this work is to associate a wide class of Lagrangian functions with a nonconvex, inequality and equality constrained optimization problem in such a way that *unconstrained* stationary points and local saddle points of each Lagrangian are related to Kuhn–Tucker points or local or global solutions of the optimization problem. As a consequence of this we are able to obtain duality results and two computational algorithms for solving the optimization problem. One algorithm is a Newton algorithm which has a local superlinear or quadratic rate of convergence. The other method is a locally linearly convergent method for finding stationary points of the Lagrangian and is an extension of the method of multipliers of Hestenes and Powell to inequalities.

1. Introduction. In 1970 Rockafellar [21] introduced a Lagrangian for inequality constrained *convex* programming problems for which an *unconstrained* saddle-point corresponds to a solution of the convex programming problem. Moreover, this Lagrangian was once differentiable everywhere if the objective and constraint functions of the convex programming problem were also differentiable everywhere. In 1971 Arrow, Gould and Howe [1] considered a general class of Lagrangians (including Rockafellar's) for nonconvex programming problems and established *local* saddle-point properties for this class of Lagrangians. For their class of Lagrangians, however, the saddle-point was in general nonnegatively constrained just at it is in the classical Kuhn–Tucker [11] Lagrangian for nonlinear programming. The local saddle-point property was obtained by the presence of a convexifying parameter in their Lagrangian which made the Hessian of the Lagrangian positive definite for large enough, but finite, values of the parameter. This elegant idea of local convexification was first introduced by Arrow and Solow in 1958 [2] in connection with equality constrained problems and was later independently reconsidered in a different algorithmic context by Hestenes [8], [9] and Powell [19] in 1969 and by Haarhoff and Buys [7] in 1970. Miele, Moseley and Cragg [14], [15] have conducted numerical experiments on these ideas for equality constrained problems. More recently Rockafellar [22] gave an illuminating derivation of his Lagrangian for inequality constrained problems from the Arrow–Solow Lagrangian for equality constrained problems by the use of slack variables.

A primary purpose of this work is to relate Kuhn–Tucker points of nonconvex, inequality and equality constrained nonlinear programming problems to *unconstrained* stationary points of a wide class of Lagrangian functions. Such a relation is important because it can bring to bear all the algorithms and results of nonlinear equations theory [17], [18] on nonlinear programming. As a consequence of this relationship we present in this work local and global duality results (§3), a new superlinearly or quadratically convergent algorithm (Algorithm 4.7 and Theorem 4.8), and a linearly convergent extension to inequality constraints

* Received by the editors May 3, 1973, and in revised form February 2, 1974.

† Computer Sciences Department, University of Wisconsin, Madison, Wisconsin 53706. This work was supported by the National Science Foundation under Grant GJ 35292.

and to more general Lagrangians of the method of multipliers (Algorithm 4.9 and Theorem 4.10).

The difference between our approach and that of Rockafellar [21], [22], [23], is that Rockafellar's results are valid only for convex problems, whereas in our approach convexity plays only a minor role in some of the peripheral results. In [24] Rockafellar extends the results for his specific Lagrangian to nonconvex optimization problems and relates *global* solutions of the optimization problem to *global* saddle-points of his Lagrangian. Our results are principally aimed at related *local* stationary points of the two problems and are established for a general class of Lagrangians. Also Rockafellar's Lagrangian is differentiable only once globally, whereas ours are twice differentiable globally. This is an important distinction in the application of Newton-type algorithms which require twice differentiability. In obtaining this twice differentiability property we lose the general concavity of Rockafellar's Lagrangian in the dual variable y . However our Lagrangians are concave in y for primal feasible points (see Remark 2.13 below). The difference between our approach and that of Arrow, Gould and Howe [1] is that for their general result the Lagrangian saddle-point is constrained by nonnegativity constraints, whereas the stationary points of our Lagrangians are completely unconstrained. Also, the conditions imposed on our Lagrangians are different from their conditions. In addition we give a new general formulation for unconstrained Lagrangians together with new concrete realizations.

We shall be concerned throughout this paper with the following problem:

$$(1.1a) \quad \text{Minimize } f(x)$$

subject to

$$(1.1b) \quad \begin{aligned} g_i(x) &\leq 0, & i &= 1, \dots, m, \\ g_i(x) &= 0, & i &= m+1, \dots, k, \end{aligned}$$

where f and g_i , $i = 1, \dots, k$, are functions from R^n into R . We shall associate with this problem a real-valued Lagrangian function L in such a way that Kuhn–Tucker points of (1.1) are related to *unconstrained* stationary and saddle-points of L . This is done in §2 of the paper where, in addition, we give sufficient conditions different from those of [1] for the Hessian of L with respect to x to be positive definite. This latter result is important in establishing the local duality results of §3 and the convergence of the algorithms of §4. In §3 we establish duality results between problem (1.1) and an *equality* constrained dual problem, problem (3.2). We establish a weak duality theorem 3.3 in the presence of convexity, a duality theorem 3.4 and a converse duality theorem 3.10 in which convexity plays a secondary role. In particular, we relate, among other things, points satisfying Kuhn–Tucker conditions to points satisfying second order optimality conditions, without any convexity assumptions. In §4 we present two computational algorithms for the solution of (1.1) based upon finding stationary points of a Lagrangian M obtained by augmenting L . Algorithm 4.7 is a Newton method for finding a zero of the gradient of M and for which we establish under suitable conditions a superlinear or quadratic rate of convergence. Algorithm 4.9 is an extension of the method of multipliers [8], [9], [19], [7] to inequalities and for which we establish a local linear rate of convergence.

We shall make use of the following notation. For the point \bar{x} satisfying the constraints of problem (1.1) we shall define the following sets:

$$I = \{i | g_i(\bar{x}) = 0, i = 1, \dots, m\}, \quad J = \{i | g_i(\bar{x}) < 0, i = 1, \dots, m\}, \\ E = \{i | i = m + 1, \dots, k\}$$

and $K = I \cup E$. For the Lagrangian $L: R^n \times R^k \times (0, \infty) \rightarrow R$, $\nabla L(x, y, \alpha)$ will denote its $(n + k)$ -dimensional gradient with respect to (x, y) , $\nabla_1 L(x, y, \alpha)$ its n -dimensional gradient with respect to x , $\nabla_2 L(x, y, \alpha)$ its k -dimensional gradient with respect to y , and $\nabla^2 L(x, y, \alpha)$ its $(n + k) \times (n + k)$ Hessian matrix with respect to (x, y) . The submatrices of $\nabla^2 L(x, y, \alpha)$ will be denoted by $\nabla_{11} L(x, y, \alpha)$, $\nabla_{12} L(x, y, \alpha)$, $\nabla_{21} L(x, y, \alpha)$ and $\nabla_{22} L(x, y, \alpha)$. All vectors are either row or column vectors depending on the context. A superscript T will denote the transpose and will be used only in denoting the transpose of a matrix or the tensor product of two vectors.

2. Equivalence of Kuhn–Tucker points and unconstrained stationary and saddle points. A primary objective of this work is to relate points that satisfy the Kuhn–Tucker conditions for problem (1.1) to unconstrained stationary points and saddle points of an appropriately defined Lagrangian L . For that purpose we begin by defining such a Lagrangian as follows:

$$(2.1) \quad L(x, y, \alpha) = f(x) + \sum_{i=1}^m (\psi(\alpha g_i(x) + y_i)_+ - \psi(y_i)) + \sum_{i=m+1}^k (\psi(\alpha g_i(x) + y_i) - \psi(y_i)),$$

where $\alpha > 0$,

$$\psi: R \rightarrow R, \quad \psi(\zeta)_+ = \begin{cases} \psi(\zeta) & \text{if } \zeta \geq 0, \\ 0 & \text{if } \zeta < 0, \end{cases}$$

and ψ satisfies the following conditions:

- (a) ψ is twice differentiable on R and $\psi''(\zeta) > 0$ for $\zeta \neq 0$;
 (2.2) (b) ψ' maps R onto R and $\psi'(0) = 0$;
 (c) $\psi(0) = 0$.

It immediately follows from the above conditions that

- (2.2) (d) ψ' is a strictly increasing function on R ,
 (e) ψ is a nonnegative convex function on R .

The motivation behind the above Lagrangian is the following. For the case of equality constraints *only*, it is easy to see that for

$$L(x, y, \alpha) = f(x) + \sum_{i=m+1}^k (\psi(\alpha g_i(x) + y_i) - \psi(y_i)),$$

the condition $\nabla L(x, y, \alpha) = 0$ is equivalent to

$$\nabla f(x) + \sum_{i=m+1}^k \alpha \psi'(\alpha g_i(x) + y_i) \nabla g_i(x) = 0$$

and

$$\psi'(\alpha g_i(x) + y_i) - \psi'(y_i) = 0, \quad i = m + 1, \dots, k.$$

Since ψ' is strictly increasing, the last equality gives that $\alpha g_i(x) + y_i = y_i$ or $g_i(x) = 0$, $i = m + 1, \dots, k$, and the gradient with respect to x becomes

$$\nabla f(x) + \sum_{i=m+1}^k \alpha \psi'(y_i) \nabla g_i(x) = 0.$$

These are precisely the Kuhn–Tucker conditions for the equality constrained problem with classical Lagrange multipliers $u_i = \alpha \psi'(y_i)$, $i = m + 1, \dots, k$. The case of inequality constraints $g_i(x) \leq 0$, $i = 1, \dots, m$, is handled by introducing the slack variable variables z_i and writing $g_i(x) + z_i^2 = 0$, $i = 1, \dots, m$. Using the Lagrangian just introduced for equality constraints we have

$$L(x, z, y, \alpha) = f(x) + \sum_{i=1}^m (\psi(\alpha g_i(x) + \alpha z_i^2 + y_i) - \psi(y_i)).$$

The variable z can be eliminated now by setting the gradient of L with respect to z equal to zero, a condition which must be satisfied by the Lagrangian for equality constraints. This gives the condition that z_i must satisfy $2\alpha z_i \psi'(\alpha g_i(x) + \alpha z_i^2 + y_i) = 0$, $i = 1, \dots, m$. This condition is satisfied if we set $z_i = \alpha^{-1/2}(-\alpha g_i(x) - y_i)^{1/2}$ when $\alpha g_i(x) + y_i < 0$ and $z_i = 0$ when $\alpha g_i(x) + y_i \geq 0$. The Lagrangian $L(x, z, y, \alpha)$ becomes then

$$f(x) + \sum_{i=1}^m (\psi(\alpha g_i(x) + y_i)_+ - \psi(y_i)),$$

which is what is given in (2.1) for the inequality constraints.

Because $\psi'(0) = 0$, we have that $(\psi(\zeta)_+)' = \psi'(\zeta)_+$ for all ζ in R . On the other hand, $(\psi(\zeta)_+)' = \psi''(\zeta)_+$ only for $\zeta > 0$ in R , with equality holding for $\zeta \leq 0$ if we assume in addition that $\psi''(0) = 0$. This extra assumption will be explicitly made where needed. For notational simplicity here and elsewhere we have used the same ψ -function for both inequality and equality constraint functions g_i , $i = 1, \dots, k$. In fact, different ψ -functions may be used for each constraint function g_i , $i = 1, \dots, k$. Occasionally we shall write ψ_i to denote the ψ -function used with a specific constraint function g_i , $i = 1, \dots, k$. Typical ψ -functions which satisfy all of conditions (2.2) are

$$(2.3a) \quad \psi(\zeta) = \frac{1}{\alpha t} |\zeta|^t, \quad \alpha \in R, \quad \alpha > 0, \quad t \geq 2,$$

$$(2.3b) \quad \psi(\zeta) = \cosh \zeta - (\zeta^2/2) - 1,$$

$$(2.3c) \quad \psi(\zeta) = \frac{1}{2}(\cosh \zeta - 1)^2.$$

If the ψ -function of (2.3a) with $t = 2$ were used for both inequality and equality constraints, we would obtain Rockafellar's Lagrangian [21]–[24] which is not twice differentiable globally because $\psi''(0) > 0$. However, every other ψ -function given in (2.3) has the property that $\psi''(0) = 0$ and hence $(\psi(\zeta)_+)' = \psi''(\zeta)_+$ for all ζ in R . This property $\psi''(0) = 0$ that leads to global twice differentiability will be exploited in some of the subsequent results such as Theorems 3.4, 3.10 and Algorithms 4.7 and 4.9. A more general Lagrangian formulation is given in [13].

For the sake or explicitness we give below a Lagrangian based on the ψ -

function of (2.3a) with $t = 4$ for the inequality constraints and $t = 2$ for equality constraints:

$$\begin{aligned} L(x, y, \alpha) &= f(x) + \frac{1}{4\alpha} \sum_{i=1}^m ((\alpha g_i(x) + y_i)_+^4 - y_i^4) + \frac{1}{2\alpha} \sum_{i=m+1}^k ((\alpha g_i(x) + y_i)^2 - y_i^2) \\ (2.4) \quad &= f(x) + \frac{1}{4\alpha} \sum_{i=1}^m ((\alpha g_i(x) + y_i)_+^4 - y_i^4) + \sum_{i=m+1}^k \left(\frac{\alpha}{2} g_i(x)^2 + y_i g_i(x) \right). \end{aligned}$$

Almost every result obtained in this paper applies but is not limited to this specific Lagrangian which is twice (thrice) differentiable globally if the functions f and $g_i, i = 1, \dots, k$, are also twice (thrice) differentiable globally.

We begin by relating Kuhn–Tucker points of problem (1.1) and stationary points of L , that is, (x, y) such that $\nabla L(x, y, \alpha) = 0$.

EQUIVALENCE THEOREM 2.5. *Let f and $g_i, i = 1, \dots, k$, be differentiable at \bar{x} and let α be any positive number. If (\bar{x}, \bar{u}) is a Kuhn–Tucker point of (1.1), then \bar{x} and \bar{y} defined by (2.6) below constitute a stationary point of L . Conversely, if (\bar{x}, \bar{y}) is a stationary point of L , then \bar{x} and \bar{u} defined by (2.8) below constitute a Kuhn–Tucker point of (1.1).*

Proof. Suppose that (\bar{x}, \bar{u}) satisfies the Kuhn–Tucker conditions of problem (1.1). Define \bar{y} in R^k as follows:

$$(2.6) \quad \psi'(\bar{y}_i) = \bar{u}_i/\alpha, \quad i = 1, \dots, k.$$

The existence of a unique \bar{y} satisfying (2.6) is assured by assumption (2.2b). Hence

$$\begin{aligned} \nabla_1 L(\bar{x}, \bar{y}, \alpha) &= \nabla f(\bar{x}) + \sum_{i \in I} \alpha \psi'(\bar{y}_i)_+ \nabla g_i(\bar{x}) + \sum_{i \in J} \alpha \psi'(\alpha g_i(\bar{x}) + \bar{y}_i)_+ \nabla g_i(\bar{x}) \\ &\quad + \sum_{i \in E} \alpha \psi'(\bar{y}_i) \nabla g_i(\bar{x}) \\ &= \nabla f(\bar{x}) + \sum_{i=1}^k \bar{u}_i \nabla g_i(\bar{x}) = 0, \end{aligned}$$

$$\frac{\partial L}{\partial y_i}(\bar{x}, \bar{y}, \alpha) = \psi'(\alpha g_i(\bar{x}) + \bar{y}_i)_+ - \psi'(\bar{y}_i) = \psi'(\bar{y}_i) - \psi'(\bar{y}_i) = 0, \quad i \in I,$$

$$\frac{\partial L}{\partial y_i}(\bar{x}, \bar{y}, \alpha) = \psi'(\alpha g_i(\bar{x}))_+ - \psi'(0) = 0, \quad i \in J,$$

$$\frac{\partial L}{\partial y_i}(\bar{x}, \bar{y}, \alpha) = \psi'(\alpha g_i(\bar{x}) + \bar{y}) - \psi'(\bar{y}_i) = \psi'(\bar{y}_i) - \psi'(\bar{y}_i) = 0, \quad i \in E.$$

Hence $\nabla_2 L(\bar{x}, \bar{y}, \alpha) = 0$ and (\bar{x}, \bar{y}) is a stationary point of L . To prove the converse it will be convenient to establish the following key lemma first.

LEMMA 2.7. *For any $\alpha > 0$ and L defined by (2.1) we have that*

$$\begin{aligned} \left\{ \frac{\partial L}{\partial y_i}(x, y, \alpha) = 0, i = 1, \dots, m \right\} &\Leftrightarrow \{g_i(x) \leq 0, y_i \geq 0, y_i g_i(x) = 0, i = 1, \dots, m\}, \\ \left\{ \frac{\partial L}{\partial y_i}(x, y, \alpha) = 0, i = m+1, \dots, k \right\} &\Leftrightarrow \{g_i(x) = 0, i = m+1, \dots, k\}. \end{aligned}$$

Proof. For $i = 1, \dots, m$,

$$\left\{ \frac{\partial L}{\partial y_i}(x, y, \alpha) = \psi'(\alpha g_i(x) + y_i)_+ - \psi'(y_i) = 0 \right\} \Leftrightarrow$$

$$\left\{ \begin{array}{c} \alpha g_i(x) + y_i \geq 0, \alpha g_i(x) + y_i = y_i \\ \text{or} \\ \alpha g_i(x) + y_i < 0, y_i = 0 \end{array} \right\} \Leftrightarrow \left\{ \begin{array}{c} g_i(x) = 0, y_i \geq 0 \\ \text{or} \\ g_i(x) < 0, y_i = 0 \end{array} \right\} \Leftrightarrow \left\{ \begin{array}{c} g_i(x) \leq 0 \\ y_i \geq 0 \\ y_i g_i(x) = 0 \end{array} \right\}.$$

For $i = m + 1, \dots, k$,

$$\left\{ \frac{\partial L}{\partial y_i}(x, y, \alpha) = \psi'(\alpha g_i(x) + y_i) - \psi'(y_i) = 0 \right\} \Leftrightarrow \{\alpha g_i(x) + y_i = y_i\} \Leftrightarrow \{g_i(x) = 0\}. \quad \square$$

To complete the proof of Theorem 2.5 now, suppose that (\bar{x}, \bar{y}) is a stationary point of L . Define $\bar{u} \in R^k$ as follows:

$$(2.8) \quad \bar{u}_i = \begin{cases} \alpha \psi'(\alpha g_i(\bar{x}) + \bar{y}_i)_+, & i = 1, \dots, m, \\ \alpha \psi'(\alpha g_i(\bar{x}) + \bar{y}), & i = m + 1, \dots, k. \end{cases}$$

Hence $\bar{u}_i \geq 0, i = 1, \dots, m$, and

$$\nabla f(\bar{x}) + \sum_{i=1}^k \bar{u}_i \nabla g_i(\bar{x}) = \nabla_1 L(\bar{x}, \bar{y}, \alpha) = 0.$$

By Lemma 2.7 we have that, since $\nabla_2 L(\bar{x}, \bar{y}, \alpha) = 0, g_i(\bar{x}) \leq 0, y_i \geq 0, y_i g_i(\bar{x}) = 0, i = 1, \dots, m$, and $g_i(\bar{x}) = 0, i = m + 1, \dots, k$. Hence for $i \in J, \bar{y}_i = 0, \bar{u}_i = \alpha \psi'(\alpha g_i(\bar{x}))_+ = 0$, and so $\bar{u}_i g_i(\bar{x}) = 0, i = 1, \dots, m$. Hence (\bar{x}, \bar{u}) satisfies the Kuhn-Tucker conditions for problem (1.1). \square

The significance of Theorem 2.5 lies in the fact that the problem of finding a Kuhn-Tucker point of a nonlinear programming problem has been reduced to that of finding solutions of the nonlinear equations $\nabla L(x, y, \alpha) = 0$ for *any* positive α . In §4 we shall describe two computational methods for finding Kuhn-Tucker points of problem (1.1) based on solving these nonlinear equations.

We establish a result now which is essentially due to Arrow, Gould and Howe [1], but under different assumptions from theirs. This result is important in establishing some of the duality and computational results to follow. We shall need the *second order sufficiency* conditions for problem (1.1)[5]. A point $(\bar{x}, \bar{u}) \in R^n \times R^k$ is said to satisfy the second order sufficiency conditions for problem (1.1) if (a) it satisfies the Kuhn-Tucker conditions of problem (1.1), (b) if $x \nabla^2 L^0(\bar{x}, \bar{u}) x > 0$ for each nonzero x in R^n satisfying $\nabla g_i(\bar{x}) x = 0$ for $i \in \{i | \bar{u}_i > 0, g_i(\bar{x}) = 0, i = 1, \dots, m\} \cup E$, and $\nabla g_i(\bar{x}) x \leq 0$ for $i \in \{i | \bar{u}_i = 0, g_i(\bar{x}) = 0, i = 1, \dots, m\}$, where L^0 is the classical Lagrangian defined by

$$L^0(x, u) = f(x) + \sum_{i=1}^k u_i g_i(x).$$

We shall also use the concept of *strict complementarity* at (\bar{x}, \bar{u}) , that is, $\bar{u}_i \neq 0$ for each $i = 1, \dots, k$ for which $g_i(\bar{x}) = 0$.

THEOREM 2.9 (Positive definiteness of $\nabla_{11} L(\bar{x}, \bar{y}, \alpha)$ and saddle-point result).

(a) Let f and $g_i, i = 1, \dots, k$, be twice differentiable at \bar{x} , let (\bar{x}, \bar{u}) satisfy the second order sufficiency conditions for problem (1.1), and let strict complementarity

hold at (\bar{x}, \bar{y}) . Then for \bar{x} and \bar{y} defined by (2.6) and for some large enough but finite α , $\nabla_{11}L(\bar{x}, \bar{y}, \alpha)$ is positive definite and

$$(2.10) \quad L(\bar{x}, y, \alpha) \leq L(\bar{x}, \bar{y}, \alpha) < L(x, \bar{y}, \alpha) \quad \text{for all } y \in R^k, x \in N(\bar{x}), x \neq \bar{x},$$

where $N(\bar{x})$ is some open neighborhood of \bar{x} . If f and $g_i, i = 1, \dots, m$, are convex, and $g_i, i = m + 1, \dots, k$, are affine, then $N(\bar{x}) = R^n$.

(b) Conversely, if (2.10) holds, with $<$ possibly replaced by \leq , for some $\alpha > 0$, then \bar{x} is a solution of (1.1) subject to the extra restriction that $x \in N(\bar{x})$.

Proof. (a) By strict complementarity we have that for $i \in I$, $\bar{u}_i > 0$, and hence by (2.6), $\bar{y}_i > 0$ and $\alpha g_i(\bar{x}) + \bar{y}_i > 0$. So $(\psi(\zeta)_+)' = \psi''(\zeta)$ for $\zeta = \alpha g_i(\bar{x}) + \bar{y}_i, i \in I$. From (2.6) we have that for $i \in J$, $\bar{y}_i = 0$ and hence $\alpha g_i(\bar{x}) + \bar{y}_i < 0$. So $\psi(\zeta)_+ = (\psi(\zeta)_+)' = (\psi(\zeta)_+)' = 0$ for $\zeta = \alpha g_i(\bar{x}) + \bar{y}_i, i \in J$.

Thus

$$(2.11) \quad \begin{aligned} \nabla_{11}L(\bar{x}, \bar{y}, \alpha) &= \nabla^2 f(\bar{x}) + \sum_{i \in I \cup E} \alpha^2 \psi''(\alpha g_i(\bar{x}) + \bar{y}_i) \nabla g_i(\bar{x}) \nabla g_i(\bar{x})^T \\ &\quad + \sum_{i \in I \cup E} \alpha \psi'(\alpha g_i(\bar{x}) + \bar{y}_i) \nabla^2 g_i(\bar{x}) \\ &= \left[\nabla_{11}L^0(\bar{x}, \bar{u}) + \sum_{i \in I \cup E} \alpha^2 \psi''(\bar{y}_i) \nabla g_i(\bar{x}) \nabla g_i(\bar{x})^T \right] \quad (\text{by 2.6}), \end{aligned}$$

where $L^0(x, u)$ is the standard Lagrangian. Note that by (2.2a) and strict complementarity we have that $\psi''(\bar{y}_i) > 0$ for $i \in I \cup E$. Hence by Debreu's theorem [4, Thm. 3] which states that

$$\{x \neq 0, Mx = 0 \Rightarrow xLx > 0\} \Leftrightarrow \left\{ \begin{array}{l} L + \gamma M^T M \text{ is positive} \\ \text{definite for } \gamma \text{ sufficiently large} \end{array} \right\}$$

and the second order sufficiency conditions, it follows that the term in the square bracket in (2.11) is positive definite for α sufficiently large. Hence for α large enough $\nabla_{11}L(\bar{x}, \bar{y}, \alpha)$ is positive definite and the second inequality of (2.10) holds for x , different from \bar{x} , in some open neighborhood $N(\bar{x})$ of \bar{x} .

To establish the first inequality of (2.10) we have from (2.6) that $\bar{y}_i \geq 0$ for $i \in I$ and $\bar{y}_i = 0$ for $i \in J$. Hence

$$\begin{aligned} L(\bar{x}, y, \alpha) - L(\bar{x}, \bar{y}, \alpha) &= \sum_{i=1}^m (\psi(\alpha g_i(\bar{x}) + y_i)_+ - \psi(y_i)) \\ &\quad + \sum_{i=m+1}^k (\psi(\alpha g_i(\bar{x}) + y_i) - \psi(y_i)) \leq 0, \end{aligned}$$

where the last inequality follows from the fact that $\psi(\alpha g_i(\bar{x}) + y_i)_+ \leq \psi(y_i)_+ \leq \psi(y_i)$ for $i = 1, \dots, m$, since $g_i(\bar{x}) \leq 0$, ψ is nonnegative and $\psi(0) = 0$, and from the fact that $g_i(\bar{x}) = 0$ for $i = m + 1, \dots, k$. Hence $L(\bar{x}, y, \alpha) \leq L(\bar{x}, \bar{y}, \alpha)$ for all y in R^k . If, in addition, $f, g_i, i = 1, \dots, m$, are convex, and $g_i, i = m + 1, \dots, k$, are affine, then it follows from the convexity and monotonicity of $\psi(\cdot)_+$, the convexity of ψ and affinity of $g_i, i = m + 1, \dots, k$, that $L(x, y, \alpha)$ is convex in x for each fixed y and α . Since $\nabla_1 L(\bar{x}, \bar{y}, \alpha) = 0$, it follows that $L(\bar{x}, \bar{y}, \alpha) < L(x, \bar{y}, \alpha)$ for all x in R^n .

(b) Suppose now that (2.10) holds. From the second inequality of (2.10) we get that $\nabla_2 L(\bar{x}, \bar{y}, \alpha) = 0$ and from Lemma 2.7 we obtain that $\bar{y}_i \geq 0$, $g_i(\bar{x}) \leq 0$, $\bar{y}_i g_i(\bar{x}) = 0$ for $i = 1, \dots, m$, and $g_i(\bar{x}) = 0$ for $i = m + 1, \dots, k$. Hence \bar{x} is feasible. For any other feasible point x which is also in $N(\bar{x})$ we have that

$$\begin{aligned} 0 &\leq L(x, \bar{y}, \alpha) - L(\bar{x}, \bar{y}, \alpha) \quad (\text{by 2.10}) \\ &= f(x) - f(\bar{x}) + \sum_{i=1}^m (\psi(\alpha g_i(x) + \bar{y}_i)_+ - \psi(\alpha g_i(\bar{x}) + \bar{y}_i)_+) \\ &\quad + \sum_{i=m+1}^k (\psi(\alpha g_i(x) + \bar{y}_i) - \psi(\alpha g_i(\bar{x}) + \bar{y}_i)) \\ &= f(x) - f(\bar{x}) + \sum_{i \in I} (\psi(\alpha g_i(x) + \bar{y}_i)_+ - \psi(\bar{y}_i)_+) + \sum_{i \in J} (\psi(\alpha g_i(x))_+ - \psi(\alpha g_i(\bar{x}))_+) \\ &\quad + \sum_{i \in E} (\psi(\bar{y}_i) - \psi(\bar{y}_i)) \leq f(x) - f(\bar{x}). \end{aligned}$$

Hence $f(\bar{x}) \leq f(x)$ for all $x \in N(\bar{x})$ which are feasible. \square

Remark 2.12. If it is further assumed that $\psi''(0) > 0$, which is the case in Rockafellar's Lagrangian, then the strict complementarity requirement at (\bar{x}, \bar{u}) for Theorem 2.9 above can be slightly weakened to the following: $\bar{u}_i > 0$ for $i = 1, \dots, m$ for which $g_i(\bar{x}) = 0$.

Remark 2.13. We observe that $L(x, y, \alpha)$ is concave in y for each fixed α and fixed feasible x if we assume that $\psi''(0) = 0$ and $\psi''(\zeta)$ is nondecreasing for $\zeta \geq 0$. This follows from the facts that

$$\frac{\partial^2 L}{\partial y_i \partial y_j}(x, y, \alpha) = 0 \quad \text{for } i \neq j,$$

that for $g_i(x) \leq 0$, $i = 1, \dots, m$,

$$\begin{aligned} \frac{\partial^2}{\partial y_i^2} L(x, y, \alpha) &= \psi''(\alpha g_i(x) + y_i)_+ - \psi''(y_i) \\ &\leq \psi''(\alpha g_i(x) + y_i)_+ - \psi''(y_i)_+ \leq 0, \end{aligned}$$

and that for $g_i(x) = 0$, $i = m + 1, \dots, k$,

$$\frac{\partial^2}{\partial y_i^2} L(x, y, \alpha) = \psi''(\alpha g_i(x) + y_i) - \psi''(y_i) = 0.$$

It was also shown in the proof of Theorem 2.9(a) above that if f, g_i , $i = 1, \dots, m$, are convex and g_i , $i = m + 1, \dots, k$, are affine, then L is convex in x for each fixed y and α .

3. Duality. We observe first that as a consequence of Lemma 2.7 the primal problem (1.1) is equivalent to:

$$(3.1a) \quad \underset{x, y}{\text{Minimize}} \quad L(x, y, \alpha)$$

subject to

$$(3.1b) \quad \nabla_2 L(x, y, \alpha) = 0,$$

where L is defined by (2.1) or more specifically by (2.4) and α is any positive number. We shall associate with this problem the following dual problem:

$$(3.2a) \quad \underset{x, y}{\text{Maximize}} \quad L(x, y, \alpha)$$

subject to

$$(3.2b) \quad \nabla_1 L(x, y, \alpha) = 0.$$

We shall assume no convexity in many of the following results, and hence the standard techniques of deriving duality results such as the use of the min-max theorem [26], [10], [27] will not apply, nor will the elegant conjugate function theory of Rockafellar [20] apply directly; however, see [24].

The results of this section consist of a weak duality theorem 3.3 (for which convexity is needed) and a duality theorem 3.4. This relates a Kuhn–Tucker point of (1.1) to a Kuhn–Tucker point of the dual problem (3.2), to a second order maximum of (3.2) under no convexity assumptions, and finally to a global solution of (3.2) under convexity. The converse duality theorem 3.10 similarly relates a local solution of the dual problem (3.2) to a Kuhn–Tucker point of the primal problem (1.1), to a second order minimum under no convexity assumptions, and finally to a global solution of (1.1) under convexity.

Probably the most important features of these duality theorems are the absence of inequality constraints from the dual problem (3.2) and the relations between second order optima of the dual problems obtained in Theorems 3.4 and 3.10 without any convexity assumptions. Related local results for a specific L have also been given by Buys [3].

WEAK DUALITY THEOREM 3.3. *Let \hat{x} satisfy the constraints of the primal problem (1.1), or equivalently let (\hat{x}, \hat{y}) satisfy the constraints of (3.1). Let (x, y) satisfy the constraints of the dual problem (3.2), let $f, g_i, i = 1, \dots, m$, be differentiable and convex on R^n , and let $g_i, i = m + 1, \dots, k$, be affine functions. Then $f(\hat{x}) \geq L(x, y)$.*

Proof.

$$\begin{aligned} f(\hat{x}) &\geq f(x) + \nabla f(x)(\hat{x} - x) && \text{(by convexity of } f) \\ &= f(x) - \sum_{i=1}^m \alpha \psi'(\alpha g_i(x) + y_i)_+ \nabla g_i(x)(\hat{x} - x) \\ &\quad - \sum_{i=m+1}^k \alpha \psi'(\alpha g_i(x) + y_i) \nabla g_i(x)(\hat{x} - x) \quad (\text{since } \nabla_1 L(x, y, \alpha) = 0) \\ &\geq f(x) + \sum_{i=1}^m \alpha \psi'(\alpha g_i(x) + y_i)_+ (g_i(x) - g_i(\hat{x})) \\ &\quad + \sum_{i=m+1}^k \alpha \psi'(\alpha g_i(x) + y_i) (g_i(x) - g_i(\hat{x})) \quad \begin{array}{l} \text{(by convexity of } g_i, i = 1, \dots, m, \\ \text{and affineness of } g_i, i = m + 1, \\ \dots, k) \end{array} \\ &\geq f(x) + \sum_{i=1}^m \alpha \psi'(\alpha g_i(x) + y_i)_+ g_i(x) \\ &\quad + \sum_{i=m+1}^k \alpha \psi'(\alpha g_i(x) + y_i) g_i(x) && \text{(by primal feasibility of } \hat{x}) \\ &\geq f(x) + \sum_{i=1}^m (\psi(\alpha g_i(x) + y_i)_+ - \psi(y_i)_+) \\ &\quad + \sum_{i=m+1}^k (\psi(\alpha g_i(x) + y_i) - \psi(y_i)) && \text{(by convexity of } \psi(\cdot)_+ \text{ and } \psi) \\ &= L(x, y, \alpha). \end{aligned}$$

□

DUALITY THEOREM 3.4. *Let $f, g_i, i = 1, \dots, k$, be differentiable at \hat{x} .*

(a) *If (\bar{x}, \bar{u}) is a Kuhn–Tucker point of (1.1), and if either strict complementarity holds at (\bar{x}, \bar{u}) or $\psi_i''(0) = 0$ for $i \in I$, then (\bar{x}, \bar{y}) defined by (2.6) satisfies the following Kuhn–Tucker conditions of the dual problem (3.2):*

$$(3.5) \quad \begin{aligned} \nabla_2 L(\bar{x}, \bar{y}, \alpha) + \bar{v} \nabla_{12} L(\bar{x}, \bar{y}, \alpha) &= 0, \\ \nabla_1 L(\bar{x}, \bar{y}, \alpha) + \bar{v} \nabla_{11} L(\bar{x}, \bar{y}, \alpha) &= 0, \\ \nabla_1 L(\bar{x}, \bar{y}, \alpha) &= 0 \end{aligned}$$

with $\bar{v} = 0$.

(b) *If $f, g_i, i = 1, \dots, k$, are twice differentiable at \bar{x} , if the second order sufficiency conditions and strict complementarity hold at (\bar{x}, \bar{u}) , if $\nabla g_i(\bar{x}), i \in I \cup E$, are linearly independent, then for sufficiently large α , (\bar{x}, \bar{y}) determined by (2.6) forms an isolated local maximum of the dual problem (3.2) if $\psi_i''(0) > 0$ for $i \in I \cup J$. If $\psi_i''(0) = 0$ for $i \in I \cup J$, then (\bar{x}, \bar{y}) forms an isolated local maximum of (3.2) subject to the additional constraints that $y_i = 0, i \in J$.*

(c) *If in addition to the assumptions of part (a) above, $g_i, i = 1, \dots, m$, are differentiable and convex on R^n and $g_i, i = m + 1, \dots, k$, are affine, then (\bar{x}, \bar{y}) solves the dual problem (3.2) and the extrema $f(\bar{x})$ and $L(\bar{x}, \bar{y}, \alpha)$ are equal.*

Proof. (a) By Theorem 2.5 we have that $\nabla_1 L(\bar{x}, \bar{y}, \alpha) = 0$ and $\nabla_2 L(\bar{x}, \bar{y}, \alpha) = 0$ which are the Kuhn–Tucker conditions (3.5) with $\bar{v} = 0$. Strict complementarity or $\psi_i''(0) = 0$ for $i \in I$ are imposed here so that the second derivatives of (3.5) are well-defined.

(b) By part (a) of this theorem, (\bar{x}, \bar{y}) satisfies the Kuhn–Tucker conditions (3.5) of the dual problem (3.2) with $\bar{v} = 0$. To show that (\bar{x}, \bar{y}) is an isolated local maximum of (3.2) we need to show that the second order sufficiency conditions for (3.2) are satisfied at (\bar{x}, \bar{y}) , that is,

$$(3.6) \quad (\nabla_{11} L(\bar{x}, \bar{y}, \alpha) \quad \nabla_{12} L(\bar{x}, \bar{y}, \alpha)) \begin{pmatrix} x \\ y \end{pmatrix} = 0, \quad (x \ y) \neq 0,$$

implies that

$$(3.7) \quad (x \ y) \nabla^2 L(\bar{x}, \bar{y}, \alpha) \begin{pmatrix} x \\ y \end{pmatrix} < 0,$$

where

$$(3.8) \quad \nabla^2 L(\bar{x}, \bar{y}, \alpha) = \begin{bmatrix} \nabla_{11} L(\bar{x}, \bar{y}, \alpha) & \nabla_{12} L(\bar{x}, \bar{y}, \alpha) \\ \nabla_{21} L(\bar{x}, \bar{y}, \alpha) & \nabla_{22} L(\bar{x}, \bar{y}, \alpha) \end{bmatrix} \\ = \begin{bmatrix} \nabla_{11} L(\bar{x}, \bar{y}, \alpha) & \alpha \psi''(\bar{y}_i) \nabla g_i(\bar{x}) & 0 \\ & (i \in I \cup E) & \\ \alpha \psi''(\bar{y}_i) \nabla g_i(\bar{x}) & 0 & 0 \\ & (i \in I \cup E) & \\ 0 & 0 & -\psi_i''(0) \\ & & i \in J \end{bmatrix}.$$

But for (x, y) satisfying (3.6) we have that

$$(3.9) \quad (x \ y) \nabla^2 L(\bar{x}, \bar{y}, \alpha) \begin{pmatrix} x \\ y \end{pmatrix} = -x \nabla_{11} L(\bar{x}, \bar{y}, \alpha) x + y \nabla_{22} L(\bar{x}, \bar{y}, \alpha) y.$$

For the case where $\psi_i''(0) > 0$ for $i \in I \cup J$, we obtain the negativity of (3.9) for $(x, y_j) \neq 0$ from the positive definiteness of $\nabla_{11}L(\bar{x}, \bar{y}, \alpha)$ for large α (by Theorem 2.9(a)) and from $-\psi_j''(0) < 0$. The case $(x, y_j) = 0$ is excluded because by (3.6) and (3.8), $y_{I \cup E} \neq 0$ and

$$\sum_{i \in I \cup E} \alpha \psi_i''(\bar{y}_i) \nabla g_i(\bar{x}) y_i = 0,$$

which contradicts the linear independence of $\nabla g_i(\bar{x})$, $i \in I \cup E$, since by strict complementarity $\psi_i''(\bar{y}_i) > 0$ for $i \in I \cup E$. We have thus established that (3.6) implies (3.7) if $\psi_i''(0) > 0$ for $i \in I \cup J$. When $\psi_i''(0) = 0$ for $i \in I \cup J$, we establish that (3.6) implies (3.7) under the added assumption that $y_J = 0$. For this case, (3.9) becomes

$$(x \quad y) \nabla^2 L(\bar{x}, \bar{y}, \alpha) \begin{pmatrix} x \\ y \end{pmatrix} = -x \nabla_{11} L(\bar{x}, \bar{y}, \alpha) x,$$

which is negative for $x \neq 0$ by the positive definiteness of $\nabla_{11}L(\bar{x}, \bar{y}, \alpha)$ for large α . The case $x = 0$ is excluded because then $y_{I \cup E} \neq 0$ and by (3.6) and (3.8),

$$\sum_{i \in I \cup E} \alpha \psi_i''(\bar{y}_i) \nabla g_i(\bar{x}) y_i = 0,$$

which contradicts the linear independence of $\nabla g_i(\bar{x})$, $i \in I \cup E$.

(c) By part (a) of this theorem (\bar{x}, \bar{y}) determined from (2.6) satisfies (3.5) with $\bar{v} = 0$. Hence $\nabla_1 L(\bar{x}, \bar{y}, \alpha) = 0$ and (\bar{x}, \bar{y}) is a feasible point for the dual problem (3.2). For any dual feasible point (x, y) we have by the weak duality theorem 3.3 that $f(\bar{x}) \geq L(x, y, \alpha)$. But

$$\begin{aligned} L(\bar{x}, \bar{y}, \alpha) &= f(\bar{x}) + \sum_{i \in I} (\psi(\bar{y}_i)_+ - \psi(\bar{y}_i)) + \sum_{i \in J} (\psi(\alpha g_i(\bar{x}))_+ - \psi(0)) \\ &\quad + \sum_{i=m+1}^k (\psi(\bar{y}_i) - \psi(\bar{y}_i)) \\ &= f(\bar{x}) + \sum_{i \in I} (\psi(\bar{y}_i) - \psi(\bar{y}_i)) \quad (\text{since } \bar{y}_i \geq 0, i \in I, \text{ and } g_i(\bar{x}) < 0, i \in J) \\ &= f(\bar{x}). \end{aligned}$$

Hence $L(\bar{x}, \bar{y}, \alpha) = f(\bar{x}) \geq L(x, y, \alpha)$ for any dual feasible point (x, y) , and (\bar{x}, \bar{y}) is a global solution of (3.2). \square

CONVERSE DUALITY THEOREM 3.10. *Let (\bar{x}, \bar{y}) be a local or global solution of the dual problem 3.2, let $f, g_i, i = 1, \dots, k$, be twice continuously differentiable at \bar{x} , and let either $\bar{y}_i > 0$ for $i \in I$ or $\psi_i''(0) = 0$ for $i \in I$.*

(a) *If the matrix $\nabla_{11}L(\bar{x}, \bar{y}, \alpha)$ is nonsingular, then \bar{x} and $\bar{u} \in R^k$ determined by (2.8) satisfy the Kuhn–Tucker conditions for the primal problem (1.1).*

(b) *If, in addition, $\nabla_{11}L(\bar{x}, \bar{y}, \alpha)$ is positive definite and $\bar{y}_i > 0$ for $i \in I$, then \bar{x} and $\bar{u} \in R^k$ determined by (2.8) satisfy the second order sufficient optimality conditions for the primal problem (1.1).*

(c) *If in addition to the assumption of part (a) above, f is convex or pseudoconvex at \bar{x} , $g_i, i = 1, \dots, m$, are convex or quasiconvex at \bar{x} , and $g_i, i = m+1, \dots, k$, are affine or simultaneously quasiconvex and quasiconcave at \bar{x} , then \bar{x} is a global solution of the primal problem (1.1).*

Proof. (a) Since (\bar{x}, \bar{y}) is a solution of the dual problem (3.2), (\bar{x}, \bar{y}) and some $(\bar{v}_0, \bar{v}) \in R \times R^n$ satisfy the following Fritz John conditions [12, p. 170]:

$$\begin{aligned} \bar{v}_0 \nabla_1 L(\bar{x}, \bar{y}, \alpha) + \bar{v} \nabla_{11} L(\bar{x}, \bar{y}, \alpha) &= 0, \\ \bar{v}_0 \nabla_2 L(\bar{x}, \bar{y}, \alpha) + \bar{v} \nabla_{12} L(\bar{x}, \bar{y}, \alpha) &= 0, \\ \nabla_1 L(\bar{x}, \bar{y}, \alpha) &= 0, \\ (\bar{v}_0, \bar{v}) &\neq 0. \end{aligned} \quad (3.11)$$

From the first and third equations above and the nonsingularity of $\nabla_{11} L(\bar{x}, \bar{y}, \alpha)$ it follows that $\bar{v} = 0$ and hence $\bar{v}_0 \neq 0$. So $\nabla_1 L(\bar{x}, \bar{y}, \alpha) = 0$ and $\nabla_2 L(\bar{x}, \bar{y}, \alpha) = 0$ and by Theorem 2.5, \bar{x} and \bar{u} defined by (2.8) satisfy the Kuhn–Tucker conditions for (1.1).

(b) By part (a) above \bar{x} and \bar{u} determined by (2.8) satisfy the Kuhn–Tucker conditions for (1.1). As in the proof of Theorem 2.9 we have since $\bar{y}_i > 0$ for $i \in I$ that

$$(3.12) \quad \nabla_{11} L(\bar{x}, \bar{y}, \alpha) = \nabla_{11} L^0(\bar{x}, \bar{u}) + \sum_{i \in I \cup E} \alpha^2 \psi''(\bar{y}_i) \nabla g_i(\bar{x}) \nabla g_i(\bar{x})^T,$$

where L^0 is the standard Lagrangian. We establish now the implication

$$(3.13) \quad \{\nabla g_i(\bar{x})x = 0, x \neq 0, i \in I \cup E\} \Rightarrow x \nabla_{11} L^0(\bar{x}, \bar{u})x > 0.$$

For if not, then for some $\hat{x} \neq 0$, $\nabla g_i(\bar{x})\hat{x} = 0$, $i \in I \cup E$, and $\hat{x} \nabla_{11} L^0(\bar{x}, \bar{u})\hat{x} \leq 0$ which by (3.12) gives that $\hat{x} \nabla_{11} L(\bar{x}, \bar{y}, \alpha)\hat{x} \leq 0$, contradicting the positive definiteness of $\nabla_{11} L(\bar{x}, \bar{y}, \alpha)$. Implication (3.13), which because of (2.8) and $\bar{y}_i > 0$ for $i \in I$, is the second order sufficient optimality condition for (1.1).

(c) This part follows from the sufficiency theorem of the Kuhn–Tucker conditions [12, Thm. 2, p. 162]. \square

4. Local computational algorithms. We shall present in this section two local algorithms for the solution of problem (1.1) which are based on reducing problem (1.1) to that of finding solutions of the $n + k$ nonlinear equations $\nabla L(x, y, \alpha) = 0$. The first algorithm 4.7 is a Newton algorithm for which we establish, under suitable conditions, local superlinear or quadratic convergence rates. The second method is an extension of the method of multipliers investigated by Arrow–Solow [2], Hestenes [8], [9], Powell [19], Haarhoff and Buys [7], and Miele, Moseley and Cragg [14], [15] for the case of equality constraints. Our extension is to inequality constraints and to a general Lagrangian. We establish linear convergence for the algorithm and indicate under what sorts of conditions we may expect fast or slow convergence of the method. In [3], Buys gives, for a specific Lagrangian, a dual algorithm which is related to our stationary-point problem. One specific implementation of his algorithm, for equality constraints only, turns out to be the method of multipliers [8], [9] for which he establishes local convergence. For inequalities, however, a particular case of his algorithm gives a relative to a special case of our Algorithm 4.9. He does not however establish convergence nor a rate of convergence for that algorithm.

In both of the computational methods to be considered here, in order to establish convergence we need to have a nonsingular Hessian at the solution point. For that purpose we shall employ another Lagrangian $M(x, y, \alpha)$ which is

obtained by augmenting $L(x, y, \alpha)$ in such a way that $\nabla^2 M(\bar{x}, \bar{y}, \alpha)$ is nonsingular and $\nabla M(\bar{x}, \bar{y}, \alpha) = 0$ is equivalent to $\nabla L(\bar{x}, \bar{y}, \alpha) = 0$. (The feasibility of augmenting L to obtain another Lagrangian M , which has identical stationary points as L , presents the intriguing possibility of generating a still wider class of unconstrained Lagrangians with possibly better properties than L . This possibility has not been investigated in depth here.) For functions ψ and $\bar{\psi}$ satisfying (2.2) define

$$\begin{aligned}
 M(x, y, \alpha) &= L(x, y, \alpha) - \sum_{i=1}^m y_i^2 \bar{\psi}(-g_i(x))_+ \\
 (4.1) \quad &= f(x) + \sum_{i=1}^m (\psi(\alpha g_i(x) + y_i)_+ - \bar{\psi}(y_i) - y_i^2 \psi(-g_i(x))_+) \\
 &\quad + \sum_{i=m+1}^k (\psi(\alpha g_i(x) + y_i) - \psi(y_i)).
 \end{aligned}$$

We establish now immediately the equivalence of stationary points of L and M .

LEMMA 4.2. $\nabla L(\bar{x}, \bar{y}, \alpha) = 0 \Leftrightarrow \nabla M(\bar{x}, \bar{y}, \alpha) = 0$.

Proof. (\Rightarrow) By Lemma 2.7 we have that $\bar{y}_i g_i(\bar{x}) = 0, i = 1, \dots, m$. Hence

$$\begin{aligned}
 \nabla_1 M(\bar{x}, \bar{y}, \alpha) &= \nabla_1 L(\bar{x}, \bar{y}, \alpha) + \sum_{i=1}^m \bar{y}_i^2 \bar{\psi}'(-g_i(\bar{x}))_+ \nabla g_i(\bar{x}) = 0, \\
 \frac{\partial}{\partial y_i} M(\bar{x}, \bar{y}, \alpha) &= \frac{\partial}{\partial y_i} L(\bar{x}, \bar{y}, \alpha) - 2\bar{y}_i \bar{\psi}(-g_i(\bar{x}))_+ = 0, \quad i = 1, \dots, m, \\
 \frac{\partial}{\partial y_i} M(\bar{x}, \bar{y}, \alpha) &= \frac{\partial}{\partial y_i} L(\bar{x}, \bar{y}, \alpha) = 0, \quad i = m+1, \dots, k.
 \end{aligned}$$

(\Leftarrow) We first show that $\nabla_2 L(\bar{x}, \bar{y}, \alpha) = 0$:

$$\begin{aligned}
 \frac{\partial M}{\partial y_i}(\bar{x}, \bar{y}, \alpha) &= \frac{\partial L}{\partial y_i}(\bar{x}, \bar{y}, \alpha) - 2\bar{y}_i \bar{\psi}(-g_i(\bar{x}))_+ \\
 &= \psi'(\alpha g_i(\bar{x}) + \bar{y}_i)_+ - \psi'(\bar{y}_i) - 2\bar{y}_i \bar{\psi}(-g_i(\bar{x}))_+ = 0, \quad i = 1, \dots, m.
 \end{aligned}$$

We now show that $\bar{y}_i \bar{\psi}(-g_i(\bar{x}))_+ = 0$ for $i = 1, \dots, m$, and hence $\nabla_2 L(\bar{x}, \bar{y}, \alpha) = 0$. Suppose not. Then for some $i, \bar{y}_i \neq 0$ and $g_i(\bar{x}) < 0$. Two cases can arise.

$$\begin{aligned}
 \text{Case 1. } \{\bar{y}_i < 0, g_i(\bar{x}) < 0\} &\Rightarrow \left\{ \begin{array}{l} \psi'(\alpha g_i(\bar{x}) + \bar{y}_i)_+ = 0 \\ \psi'(\bar{y}_i) < 0 \\ 2\bar{y}_i \bar{\psi}(-g_i(\bar{x}))_+ < 0 \end{array} \right\} \\
 &\Rightarrow \{\psi'(\alpha g_i(\bar{x}) + \bar{y}_i)_+ - \psi'(\bar{y}_i) - 2\bar{y}_i \bar{\psi}(-g_i(\bar{x}))_+ > 0\} \Rightarrow \\
 &\text{Contradicts } (\partial M / \partial y_i)(\bar{x}, \bar{y}, \alpha) = 0.
 \end{aligned}$$

$$\begin{aligned}
 \text{Case 2. } \{\bar{y}_i > 0, g_i(\bar{x}) < 0\} &\Rightarrow \left\{ \begin{array}{l} \psi'(\alpha g_i(\bar{x}) + \bar{y}_i)_+ < \psi'(\bar{y}_i) \\ 2\bar{y}_i \bar{\psi}(-g_i(\bar{x}))_+ > 0 \end{array} \right\} \\
 &\Rightarrow \{\psi'(\alpha g_i(\bar{x}) + \bar{y}_i)_+ - \psi'(\bar{y}_i) - 2\bar{y}_i \bar{\psi}(-g_i(\bar{x}))_+ < 0\} \Rightarrow \\
 &\text{Contradicts } (\partial M / \partial y_i)(\bar{x}, \bar{y}, \alpha) = 0.
 \end{aligned}$$

Hence $\nabla_2 L(\bar{x}, \bar{y}, \alpha) = 0$. By Lemma 2.7 we have that $\bar{y}_i g_i(\bar{x}) = 0, i = 1, \dots, m$, and hence

$$\nabla_1 L(\bar{x}, \bar{y}, \alpha) = \nabla_1 M(\bar{x}, \bar{y}, \alpha) + \sum_{i=1}^m \bar{y}_i^2 \bar{\psi}'(-g_i(\bar{x}))_+ \nabla g_i(\bar{x}) = 0. \quad \square$$

We establish next the positive definiteness of $\nabla_{11} M(\bar{x}, \bar{y}, \alpha)$, the nonsingularity of $\nabla^2 M(\bar{x}, \bar{y}, \alpha)$ and the saddle-point property of M .

LEMMA 4.3.

(a) *Let the assumptions of Theorem 2.9(a) hold and $\bar{\psi}''(0) = 0$. Then for sufficiently large but finite α , $\nabla_{11} M(\bar{x}, \bar{y}, \alpha)$ is positive definite, $\nabla^2 M(\bar{x}, \bar{y}, \alpha)$ is nonsingular and*

$$(4.4) \quad M(\bar{x}, y, \alpha) \leq M(\bar{x}, \bar{y}, \alpha) < M(x, \bar{y}, \alpha) \quad \text{for all } y \in R^k, \quad x \in N(\bar{x}), \quad x \neq \bar{x},$$

where $N(\bar{x})$ is some open neighborhood of \bar{x} .

(b) *Conversely, if (4.4) holds, with $<$ possibly replaced by \leq , then \bar{x} is a solution of (1.1) subject to the extra restriction that $x \in N(\bar{x})$.*

Proof. (a) As in the proof of Theorem 2.9(a) we have that

$$\begin{aligned} \nabla_{11} M(\bar{x}, \bar{y}, \alpha) &= \nabla_{11} L(\bar{x}, \bar{y}, \alpha) \\ &= \nabla_{11} L^0(\bar{x}, \bar{u}) \\ &\quad + \sum_{i \in I \cup E} \alpha^2 \psi''(\bar{y}_i) \nabla g_i(\bar{x}) \nabla g_i(\bar{x})^T. \end{aligned}$$

It follows again by strict complementarity, the second order sufficiency conditions and Debreu's theorem that for large enough α , $\nabla_{11} M(\bar{x}, \bar{y}, \alpha)$ is positive definite and hence the second inequality of (4.4) holds. The first inequality of (4.4) holds because

$$M(\bar{x}, y, \alpha) \leq L(\bar{x}, y, \alpha) \leq L(\bar{x}, \bar{y}, \alpha) = M(\bar{x}, \bar{y}, \alpha).$$

To show that $\nabla^2 M(\bar{x}, \bar{y}, \alpha)$ is nonsingular we observe that

$$(4.5) \quad \nabla^2 M(\bar{x}, \bar{y}, \alpha) = \begin{bmatrix} \nabla_{11} M(\bar{x}, \bar{y}, \alpha) & \alpha \psi''(\bar{y}_i) \nabla g_i(\bar{x}) & 0 \\ & (i \in I \cup E) & \\ \alpha \psi''(\bar{y}_i) \nabla g_i(\bar{x}) & 0 & 0 \\ (i \in I \cup E) & & \\ 0 & 0 & -\psi_i''(0) - 2\bar{\psi}_i(-g_i(\bar{x})) \\ & & (i \in J) \end{bmatrix}.$$

The nonsingularity of $\nabla^2 M(\bar{x}, \bar{y}, \alpha)$ for a large α follows from the positive definiteness of $\nabla_{11} M(\bar{x}, \bar{y}, \alpha)$, the linear independence of $\nabla g_i(\bar{x})$, $\psi''(\bar{y}_i) > 0, i \in I \cup E$, and $-\psi_i''(0) - 2\bar{\psi}_i(-g_i(\bar{x})) < 0, i \in J$.

(b) From (4.4) we have that $\nabla M(\bar{x}, \bar{y}, \alpha) = 0$. By Lemma 4.2 it follows that $\nabla L(\bar{x}, \bar{y}, \alpha) = 0$ and by Lemma 2.7 we have that $\bar{y}_i g_i(\bar{x}) = 0, i = 1, \dots, m$. Hence $M(\bar{x}, \bar{y}, \alpha) = L(\bar{x}, \bar{y}, \alpha)$. From (4.4) we also have that for $x \in N(\bar{x})$,

$$0 \leq M(x, \bar{y}, \alpha) - M(\bar{x}, \bar{y}, \alpha) \leq L(x, \bar{y}, \alpha) - L(\bar{x}, \bar{y}, \alpha).$$

The rest of the proof is identical to that of Theorem 2.9(b). \square

We observe that if in (4.1), $\bar{\psi}''(0) = 0$ for the ψ -function explicitly stated therein and for the $\psi_i, i \in \{1, \dots, m\}$, associated with the inequality constraints, then M is globally twice differentiable provided that f and $g_i, i = 1, \dots, k$, are also twice differentiable. To be specific we state explicitly a recommended globally twice differentiable M -function associated with problem (1.1):

$$(4.6) \quad \begin{aligned} M(x, y, \alpha) = & f(x) + \frac{1}{4\alpha} \sum_{i=1}^m ((\alpha g_i(x) + y_i)_+^4 - y_i^4 - y_i^2(-g_i(x))_+^4) \\ & + \sum_{i=m+1}^k \left(\frac{\alpha}{2} g_i(x)^2 + y_i g_i(x) \right). \end{aligned}$$

We are prepared now to state and establish the convergence and rates of convergence of our algorithms.

NEWTON ALGORITHM 4.7. Choose $\alpha > 0$ and $(x^0, y^0) \in R^n \times R^k$. Determine (x^{j+1}, y^{j+1}) from (x^j, y^j) as follows: Linearize $\nabla M(x, y, \alpha) = 0$ around the point (x^j, y^j) and solve for (x^{j+1}, y^{j+1}) .

LOCAL CONVERGENCE AND RATE OF CONVERGENCE OF THE NEWTON ALGORITHM 4.8.

(a) Let $(\bar{x}, \bar{u}) \in R^n \times R^k$ satisfy the Kuhn–Tucker conditions of (1.1), let f and $g_i, i = 1, \dots, k$, be twice continuously differentiable at each point of an open neighborhood of \bar{x} , let $\nabla g_i(\bar{x}), i \in I \cup E$, be linearly independent and let the assumptions of Theorem 2.9(a) hold. Then for large enough but finite α there exists an open neighborhood $N(\bar{x}, \bar{y})$ of (\bar{x}, \bar{y}) in $R^n \times R^k$, where \bar{y} is determined from \bar{u} by (2.6), such that for every $(x^0, y^0) \in N(\bar{x}, \bar{y})$, the Newton algorithm iterates of 4.7 are well-defined and converge superlinearly to (\bar{x}, \bar{y}) in the sense that

$$\lim_{j \rightarrow \infty} \frac{\|z^{j+1} - \bar{z}\|}{\|z^j - \bar{z}\|} = 0,$$

where $z = (x, y)$.

(b) If, in addition, f and $g_i, i = 1, \dots, k$, are three times differentiable on $N(\bar{x})$, ψ is three times differentiable on R and $\psi_i'''(0) = 0$ for $i \in \{1, \dots, m\}$, then the Newton iterates converge quadratically to (\bar{x}, \bar{y}) , that is, for some constant d and some integer j_0 depending on z^0 ,

$$\|z^{j+1} - \bar{z}\| \leq d \|z^j - \bar{z}\|^2 \quad \text{for } j \geq j_0.$$

Proof. Since $(\bar{x}, \bar{y}) \in R^n \times R^k$ is obtained from (\bar{x}, \bar{u}) by (2.6), it follows by Theorem 2.5 that $\nabla L(\bar{x}, \bar{y}, \alpha) = 0$, by Lemma 4.2 that $\nabla M(\bar{x}, \bar{y}, \alpha) = 0$ and by Lemma 4.3(a) that $\nabla^2 M(\bar{x}, \bar{y}, \alpha)$ is nonsingular. The convergence properties stated in the theorem follow then from the local convergence theorem of Newton's method [16, p. 148]. \square

We present now a second method which is an extension to inequality constraints and to more general Lagrangians of the method of multipliers. Originally this method was proposed for equality constraints by Arrow and Solow [2] by using differential equations to determine a small step size algorithm. Later and independently of Arrow and Solow and of each other, Hestenes [8], [9], Powell [19] and Haarrhoff and Buys [7] used a similar Lagrangian approach for equality constraints and proposed a large step size method. Miele, Moseley and Cragg [14],

[15] made numerical tests of the algorithm and variants of it. More recently Buys [3] and Wierzbicki [28] considered extensions to inequality constraints. Buys suggested a dual problem approach for a specific Lagrangian function but did not give any convergence rates. Wierzbicki considered another specific but different Lagrangian.

ALGORITHM 4.9 (Method of Lagrange multipliers). Choose $\alpha > 0$, $\beta > 0$, $y^0 \in R^k$ and $x^0 \in R^n$ satisfying $\nabla_1 M(x^0, y^0, \alpha) = 0$. Let (x^j, y^j) determine (x^{j+1}, y^{j+1}) as follows:

(a) Determine x^{j+1} such that

$$M(x^{j+1}, y^j, \alpha) = \min_{x \in R^n} M(x, y^j, \alpha)$$

or $\nabla_1 M(x^{j+1}, y^j, \alpha) = 0$. If x^{j+1} is not unique, take a closest x^{j+1} , in any norm, to x^j .

(b) $y^{j+1} = y^j + \beta \nabla_2 M(x^{j+1}, y^j, \alpha)$.

THEOREM 4.10 (Local convergence and rate of convergence of the method of Lagrange multipliers). *Let the assumptions of Theorem 4.8(a) hold, and let \bar{y} be determined from \bar{u} by (2.6). Then for large but finite α , there exist open neighborhoods $N_0(\bar{x})$ and $N_0(\bar{y})$ in R^n and R^k , respectively, such that for each $y^0 \in N_0(\bar{y})$ there exists a unique x^0 in the closure $\bar{N}_0(\bar{x})$ of $N_0(\bar{x})$ satisfying $\nabla_1 M(x^0, y^0, \alpha) = 0$. Also the iterates of Algorithm 4.9 are well-defined and converge linearly to (\bar{x}, \bar{y}) for $\beta \in (0, \bar{\beta})$ for some $\bar{\beta} > 0$.*

Proof. As in the proof of Theorem 4.8 we have that $\nabla M(\bar{x}, \bar{y}, \alpha) = 0$ by Theorem 2.5 and Lemma 4.2. By Lemma 4.3(a), $\nabla_{11} M(\bar{x}, \bar{y}, \alpha)$ is positive definite for sufficiently large but finite α . It follows by the implicit function theorem [17, p. 128] that for some open neighborhoods $N(\bar{y})$ in R^k and $N(\bar{x})$ in R^n , there exists a function $e: R^k \rightarrow R^n$ which is continuously differentiable on $N(\bar{y})$ and such that:

$$(4.11) \quad \begin{aligned} &\text{For } y \in N(\bar{y}), \quad x = e(y) \text{ is a unique solution of } \nabla_1 M(x, y, \alpha) = 0 \\ &\text{in } \bar{N}(\bar{x}); \quad \bar{x} = e(\bar{y}) \text{ and } e(y) \in \bar{N}(\bar{x}). \end{aligned}$$

Define

$$(4.12) \quad N_0(\bar{x}) = N(\bar{x}).$$

For $y^j \in N(\bar{y})$, Algorithm 4.9 is well-defined and is equivalent to

$$(4.13) \quad y^{j+1} = y^j + \beta \nabla_2 M(e(y^j), y^j, \alpha)$$

if we assume for the time being that x^{j+1} of step 4.9(a) is unique.

Consider now the mapping $G: R^k \rightarrow R^k$ underlying the iteration (4.13) and defined by

$$(4.14) \quad G(y) = y + \beta \nabla_2 M(e(y), y, \alpha)$$

and its gradient evaluated at \bar{y} ,

$$(4.15) \quad \nabla G(\bar{y}) = I + \beta \nabla_{21} M(\bar{x}, \bar{y}, \alpha) \nabla e(\bar{y}) + \beta \nabla_{22} M(\bar{x}, \bar{y}, \alpha).$$

Differentiating $\nabla_1 M(e(y), y, \alpha) = 0$ with respect to y and evaluating at \bar{y} gives

$$\nabla_{11} M(\bar{x}, \bar{y}, \alpha) \nabla e(\bar{y}) + \nabla_{12} M(\bar{x}, \bar{y}, \alpha) = 0$$

and hence

$$\nabla e(\bar{y}) = -\nabla_{11}M(\bar{x}, \bar{y}, \alpha)^{-1}\nabla_{12}M(\bar{x}, \bar{y}, \alpha).$$

Substitution in (4.15) gives

(4.16)

$$\nabla G(\bar{y}) = I - \beta[\nabla_{21}M(\bar{x}, \bar{y}, \alpha)\nabla_{11}M(\bar{x}, \bar{y}, \alpha)^{-1}\nabla_{12}M(\bar{x}, \bar{y}, \alpha) - \nabla_{22}M(\bar{x}, \bar{y}, \alpha)].$$

By referring to (4.5) this expression can be rewritten as

(4.17)

$$\nabla G(\bar{y}) = I - \beta \begin{bmatrix} \alpha^2 \psi''(\bar{y}_i)^2 \nabla g_i(\bar{x}) \nabla_{11}M(\bar{x}, \bar{y}, \alpha)^{-1} \nabla g_i(\bar{x}) & 0 \\ (i \in I \cup E) & \\ 0 & \psi''_i(0) + 2\bar{\psi}_i(-g_i(\bar{x})) \\ & (i \in J) \end{bmatrix}.$$

It follows from the linear independence of $\nabla g_i(\bar{x})$, $i \in I \cup E$, the positive definiteness of $\nabla_{11}M(\bar{x}, \bar{y}, \alpha)$ and $\psi''_i(0) + 2\bar{\psi}_i(-g_i(\bar{x})) > 0$, $i \in J$, that the matrix in the square brackets is positive definite. Hence for some $\bar{\beta} > 0$, the eigenvalues of $\nabla G(\bar{y})$ are less than one in magnitude for $\beta \in (0, \bar{\beta})$ and hence the spectral radius $\rho(\nabla G(\bar{y})) < 1$. It follows by Ostrowki's point of attraction [16, p. 145] that there exist open neighborhoods $N_1(\bar{y})$, $N_2(\bar{y})$ with $N_1(\bar{y}) \subset N_2(\bar{y}) \subset N(\bar{y})$ and such that when $y^0 \in N_1(\bar{y})$ the iterates of (4.13) remain in $N_2(\bar{y})$ and converge linearly to \bar{y} . Since e is differentiable on $N_2(\bar{y})$ the iterates $\{x^j\}$ defined by $x^j = e(y^{j-1})$ are also well-defined and converge linearly to $\bar{x} = e(\bar{y})$ if we choose any $N_0(\bar{y}) \subset N_1(\bar{y})$. Hence the sequence $\{(x^j, y^j)\}$ converges linearly to (\bar{x}, \bar{y}) and the theorem is established for the case when x^{j+1} of Algorithm 4.9(a) is unique.

Suppose now that x^{j+1} of step 4.9(a) is not unique and that there also exists an $\hat{x}^{j+1} \neq x^{j+1} = e(y^j)$ such that $\nabla_1 L(\hat{x}^{j+1}, y^j, \alpha) = 0$. We will show that x^{j+1} is closer than \hat{x}^{j+1} to x^j and hence \hat{x}^{j+1} will not appear in the sequence $\{x^j\}$ generated by the algorithm. We have from (4.11) that $\hat{x}^{j+1} \notin \bar{N}(\bar{x})$ and hence

$$(4.18) \quad \|\hat{x}^{j+1} - \bar{x}\| > \delta,$$

where δ is the radius of some open ball $B_\delta(\bar{x})$ around \bar{x} which is contained in $\bar{N}(\bar{x})$. It follows again by Ostrowski's point of attraction theorem that for the sequence $\{y^j\}$ obtained from (4.13) and starting with any $y^0 \in N_1(\bar{y})$,

$$(4.19) \quad \begin{aligned} \|y^j - \bar{y}\| &\leq c\gamma^j \|y^0 - \bar{y}\|, \\ \|y^{j+1} - y^j\| &\leq 2c\gamma^j \|y^0 - \bar{y}\|, \end{aligned}$$

where $\gamma = \rho(\nabla G(\bar{y})) + 2\varepsilon < 1$ for some $\varepsilon > 0$ and c is a positive constant depending on ε and the norm employed. Also since e is differentiable we also have for $x^j = e(y^{j-1})$ and $x^{j+1} = e(y^j)$ that

$$(4.20) \quad \begin{aligned} \|x^j - \bar{x}\| &\leq c'\gamma^j \|y^0 - \bar{y}\|, \\ \|x^{j+1} - x^j\| &\leq 2c'\gamma^j \|y^0 - \bar{y}\|, \end{aligned}$$

where c' is some positive constant. Now define

$$N_0(\bar{y}) = \{y | y \in N_1(\bar{y}), \|y - \bar{y}\| < \delta/(4c')\}.$$

Hence for $y^0 \in N_0(\bar{y})$,

$$(4.21) \quad \|x^j - \bar{x}\| < \delta/2 \quad \text{for all } j \geq 1$$

and

$$(4.22) \quad \|x^{j+1} - x^j\| < \delta/2 \quad \text{for all } j \geq 1.$$

It follows from (4.18), (4.21) and (4.22) that

$$\|\hat{x}^{j+1} - x^j\| \geq \|\hat{x}^{j+1} - \bar{x}\| - \|x^j - \bar{x}\| > \delta - (\delta/2) = \delta/2 > \|x^{j+1} - x^j\|.$$

Hence x^{j+1} is closer than \hat{x}^{j+1} to x^j . So $x^{j+1} = e(y^j)$ will be picked up in step 4.9(a) rather than \hat{x}^{j+1} and iteration (4.13) will again represent Algorithm 4.9 for the nonunique case also. The remainder of the proof is the same as for the unique case. \square

We make some remarks here about the relation between the size of β and the speed of convergence. Since the size of β was determined from the requirement that some norm of $\nabla G(\bar{y})$ as given by (4.17) is less than one, it follows from (4.17) and this requirement that for some $\delta \in (0, 1)$,

$$\delta \geq \|\nabla G(\bar{y})\| \geq 1 - \beta v$$

and

$$\delta \geq \|\nabla G(\bar{y})\| \geq \beta v - 1,$$

where v is the norm of the matrix in the square brackets of (4.17). Hence

$$2/v > (1 + \delta)/v \geq \beta \geq (1 - \delta)/v > 0 \quad \text{for some } \delta \in (0, 1).$$

Fast convergence is obtained when δ is close to zero and hence $\beta \cong 1/v$. On the other hand, when $\beta \cong 0$ or $\beta \cong 2/v$, δ will be close to one and hence slow convergence is to be expected.

Another possible source of slow convergence is the condition number of $\nabla_{11}M(\bar{x}, \bar{y}, \alpha)$ which affects step (a) of Algorithm 4.9. It can be shown that if $\bar{k} < n$ where \bar{k} is the number of active inequality and equality constraints (that is, the solution \bar{x} of (1.1) lies on a manifold), then the condition number of $\nabla_{11}M(\bar{x}, \bar{y}, \alpha)$ approaches ∞ as α approaches ∞ . However, for the special case when $\bar{k} = n$ (that is, the solution \bar{x} lies on a "vertex") then the condition number of $\nabla_{11}M(\bar{x}, \bar{y}, \alpha)$ remains finite when α approaches ∞ . On the other hand, when α is not large enough, $\nabla_{11}M(\bar{x}, \bar{y}, \alpha)$ may not be positive definite then and again it may be difficult to implement step (a) of Algorithm 4.9. In summary it may be stated that *in general, convergence problems may be expected for small values of either α and β and also for large values of either α and β . Best results should be at intermediate values of α and β .* Numerical results of Miele, Mosley and Cragg [14, Table 2, Exs. 6.2 and 6.3], where $\beta = \alpha$ (k in their notation), slow convergence occurred for both small and large values of β and fast convergence occurred for intermediate values of β .

In conclusion we make some remarks about possible enlargement of the region of convergence. Basically we are finding an unconstrained saddle point of $M(x, y, \alpha)$ or a root of $\nabla M(x, y, \alpha) = 0$. Since there are no global methods for solving these problems in the absence of convexity, there is little hope in the present state

of the art for a foolproof global algorithm for solving our problem here. However, practical improvements may be achieved by introducing a step size which is determined by either minimizing the function $\|\nabla M(x, y, \alpha)\|^2$ or using an Armijo procedure [17, p. 491] on the same function along the direction $(x^{j+1} - x^j, y^{j+1} - y^j)$ or along a related direction such that $\|\nabla M(x, y, \alpha)\|^2$ decreases sufficiently along that direction. Such a procedure would lead to a point satisfying $\nabla^2 M(\bar{x}, \bar{y}, \alpha) \nabla M(\bar{x}, \bar{y}, \alpha) = 0$. If $\nabla M(\bar{x}, \bar{y}, \alpha) = 0$, we are done; if not, the procedure would have to be restarted again either at (\bar{x}, \bar{y}) or elsewhere.

Another acceleration procedure is to ignore all ε -inactive inequality constraints, that is, for some $\varepsilon > 0$ delete at iteration j the inequality constraints such that $g_i(x^j) < -\varepsilon$, $i \in \{1, \dots, m\}$, from consideration. It can be shown [13, Lemma A.17] that for some neighborhood of \bar{x} and for some ε such a procedure would not remove any of the active constraints g_i , $i \in I$, from the problem.

Acknowledgment. I am indebted to my student S. P. Han for valuable discussion and for his reading of the manuscript.

REFERENCES

- [1] K. J. ARROW, F. J. GOULD AND S. M. HOWE, *A general saddle point result for constrained optimization*, Mathematical Programming, 5 (1973), pp. 225–234.
- [2] K. J. ARROW AND R. M. SOLOW, *Gradient methods for constrained maxima with weakened assumptions*, Studies in Linear and Nonlinear Programming, K. J. Arrow, L. Hurwicz and H. Uzawa, eds., Stanford University Press, Stanford, Calif., 1958, pp. 166–176.
- [3] J. D. BUYS, *Dual algorithms for constrained optimization problems*, Doctorate dissertation, University of Leiden, 1972.
- [4] G. DEBREU, *Definite and semidefinite quadratic forms*, Econometrica, 20 (1952), pp. 295–300.
- [5] A. V. FIACCO AND G. P. MCCORMICK, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley, New York, 1968.
- [6] T. N. E. GREVILLE, ed., *Theory and Applications of Spline Functions*, Academic Press, New York, 1969.
- [7] P. C. HAARHOFF AND J. D. BUYS, *A new method for the optimization of a nonlinear function subject to nonlinear constraints*, Comput. J., 13 (1970), pp. 178–184.
- [8] M. R. HESTENES, *Multiplier and gradient methods*, Computing Methods in Optimization Problems, vol. 2, L. A. Zadeh, L. W. Neustadt and A. V. Balakrishnan, eds., Academic Press, New York, 1969, pp. 143–164.
- [9] ———, *Multiplier and gradient methods*, J. Optimization Theory Appl., 4 (1969), pp. 303–320.
- [10] S. KARAMARDIAN, *Strictly quasi-convex (concave) functions and duality in mathematical programming*, J. Math. Anal. Appl., 20 (1967), pp. 344–358.
- [11] H. W. KUHN AND A. W. TUCKER, *Nonlinear programming*, Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, J. Neyman, ed., University of California Press, Berkeley, 1951, pp. 481–492.
- [12] O. L. MANGASARIAN, *Nonlinear Programming*, McGraw-Hill, New York, 1969.
- [13] ———, *Unconstrained Lagrangians in nonlinear programming*, Computer Sciences Department Rep. 174, Univ. of Wisconsin, Madison, 1973.
- [14] A. MIELE, P. E. MOSELEY AND E. E. CRAGG, *Numerical experiments on Hestenes' method of multipliers for mathematical programming problems*, Aero-Astronautics Rep. 85, Rice University, Houston, Tex., 1971.
- [15] ———, *A modification of the method of multipliers for mathematical programming problems*, Techniques of Optimization, A. V. Balakrishnan, ed., Academic Press, New York, 1972, pp. 247–260.
- [16] J. M. ORTEGA, *Numerical Analysis, A Second Course*, Academic Press, New York, 1972.
- [17] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

- [18] A. OSTROWSKI, *Solution of Equations and Systems of Equations*, 2nd ed., Academic Press, New York, 1966.
- [19] M. J. D. POWELL, *A method for nonlinear constraints in minimization problems*, Optimization, R. Fletcher, ed., Academic Press, New York, 1969, pp. 283–298.
- [20] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, N.J., 1970.
- [21] ———, *New applications of duality in convex programming*, Proc. Fourth Conference on Probability, Brasov, Romania, 1971.
- [22] ———, *A dual approach to solving nonlinear programming problems by unconstrained optimization*, Mathematics Dept., Washington University, Seattle, 1972.
- [23] ———, *The multiplier method of Hestenes and Powell applied to convex programming*, Mathematics Dept., Washington University, Seattle, 1972.
- [24] ———, *Augmented Lagrange multiplier functions and duality in nonconvex programming*, this Journal, 12 (1974), pp. 268–285.
- [25] I. J. SCHOENBERG, ed., *Approximations with Special Emphasis on Spline Functions*, Academic Press, New York, 1969.
- [26] J. STOER, *Duality in nonlinear programming and the minimax theorem*, Numer. Math., 5 (1963), pp. 371–379.
- [27] J. STOER AND C. WITZGALL, *Convexity and Optimization in Finite Dimensions. I*, Springer-Verlag, Heidelberg, 1970.
- [28] A. P. WIERZBICKI, *A penalty function shifting method in constrained static optimization and its convergence properties*, Arch. Automat. Telemekh., 16 (1971), pp. 395–416.

NONNEGATIVITY OF A QUADRATIC FUNCTIONAL*

B. P. MOLINARI†

Abstract. The question of the nonnegativity of an integral $\int_{t_0}^{t_f} q(x, u) dt$, where $q(x, u)$ is a quadratic form defined on solutions of the linear system

$$\dot{x} = Ax + Bu, \quad x(t_0) = 0,$$

arises in optimal control, in optimal filtering, and in system passivity. This paper derives sufficient conditions and necessary conditions for such an integral to be nonnegative. The conditions have the same form, and the “gap” between them can be considered as small. The existing theory for particular classes of $q(x, u)$ is recovered as a special case.

1. Introduction. This paper is concerned with necessary conditions and sufficient conditions for the nonnegativity of the quadratic functional

$$(1.1) \quad J[x, u] = \int_{t_1}^{t_f} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix}' \begin{bmatrix} Q(t) & C'(t) \\ C(t) & R(t) \end{bmatrix} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix} dt + x'(t_f)Sx(t_f),$$

where $u(t)$ is any m -tuple function satisfying

$$(1.2) \quad u(t) \text{ is piecewise continuous on } [t_0, t_f],$$

and $x(t)$ is an n -tuple function satisfying

$$(1.3) \quad x(t) = A(t)x(t) + B(t)u(t) \quad x(t_0) = 0 \quad Dx(t_f) = 0.$$

The matrices $Q(t)$, $C(t)$, $R(t)$, $A(t)$ and $B(t)$ have dimensions consistent with x and u , and are continuous on $[t_0, t_f]$. Further, S and D are constant matrices of dimensions $n \times n$ and $r \times n$ respectively. Without loss of generality $Q(t)$, $R(t)$ and S are symmetric and D is of full rank $r \leq n$. Finally, the transpose of a matrix M is denoted by M' .

The problem has relevance in several fields. It is the “accessory minimization problem” of optimal control [1, p. 182], [2, § 25]. It is intimately associated with the general least-squares regulator problem [3, Lemma 1], [4, Thms. 1, 2]. In the dual problem of optimal filtering, the nonnegativity of a similar quadratic functional is essentially a covariance condition [5, § III]. Finally, in two special cases the nonnegativity of $J[x, u]$ is used as a definition for passivity [6, p. 30].

A well-known necessary condition for $J[x, u] \geq 0$ for all function pairs (x, u) satisfying (1.2) and (1.3) is $R(t) \geq 0$ for all $t \in [t_0, t_f]$. For each of the two cases $R(t) > 0$ and $R(t) = 0$, necessary conditions and sufficient conditions for $J[x, u] \geq 0$ are available (Jacobson [7] provides a recent survey paper). Here we develop a generalization of these results. Equivalent results have recently been obtained by Anderson [8], by Krasner and Kailath [5], [9] and by Coppel [10], all using different approaches.

* Received by the editors December 12, 1972, and in final revised form April 29, 1974.

† School of Electrical Engineering, University of New South Wales, Kensington, Sydney, Australia. Now at Computer Science Group, Statistics Department, Australian National University, Canberra ACT 2600, Australia.

For the special case of the time-invariant problem (Q , C , R , A and B are constant), the nonnegativity of $J[x, u]$ has been characterized for general $R \geq 0$ [4, Thms. 2, 3]. With nontrivial modification, that approach has proved workable for the general problem.

2. Sufficient condition. The basic necessary conditions and sufficient conditions stated in this paper involve an integral inequality, where integration is interpreted in the Riemann–Stieltjes sense [11, Chap. VI]. In particular, the following identity is used at several places in the discussion.

LEMMA 1. Consider any $n \times n$ symmetric matrix $P(t)$ of bounded variation on $[t_1, t_2] \subset [t_0, t_f]$ and any function pair (x, u) satisfying (1.2) and (1.3)₁. Then

$$(2.1) \quad \int_{t_1}^{t_2} \begin{bmatrix} x \\ u \end{bmatrix}' \begin{bmatrix} dP + (A'P + PA) dt & PB dt \\ B'P dt & 0 \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} = \left[x'Px \right]_{t_1}^{t_2}.$$

Proof. The standard integration by parts formula is valid [11, Thms. 29.5.7 and 29.5.1]:

$$\int_{t_1}^{t_2} x' dPx + \int_{t_1}^{t_2} (dx)'Px + \int_{t_1}^{t_2} x'P dx = [x'Px]_{t_1}^{t_2}.$$

The piecewise continuity of \dot{x} provides [11, Thm. 30.3.1]

$$\int_{t_1}^{t_2} x'P dx = \int_{t_1}^{t_2} x'P \dot{x} dt.$$

Noting (1.3)₁ and collecting the integrals provides (2.1).

For convenience, write the integral (1.1) as

$$(2.2) \quad J[x, u] = \int_{t_0}^{t_f} q(x, u) dt + x'(t_f)Sx(t_f).$$

Further, consider any $n \times (n - r)$ matrix Z (of full rank) whose columns span the nullspace of D . Then

$$(2.3) \quad Dx = 0 \quad \text{if and only if} \quad x = Z\alpha \quad \text{for some } \alpha.$$

Sufficient conditions for the nonnegativity of $J[x, u]$ are easily obtained.

THEOREM 1. Assume an $n \times n$ symmetric matrix $P(t)$ of bounded variation on $[t_0, t_f]$ satisfying

$$(2.4) \quad Z'(P(t_f) - S)Z \leq 0$$

and

$$(2.5) \quad \int_{t_1}^{t_2} \begin{bmatrix} x \\ u \end{bmatrix}' \begin{bmatrix} dP + (A'P + PA + Q) dt & (C' + PB) dt \\ (C + B'P) dt & R dt \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} \geq 0,$$

for all (x, u) satisfying (1.2) and (1.3)₁ and all $[t_1, t_2] \subset [t_0, t_f]$. Then $J[x, u] \geq 0$ for all (x, u) satisfying (1.2) and (1.3).

Proof. Substitution of (2.1) in (2.5) provides

$$\int_{t_1}^{t_2} q(x, u) dt + [x'Px]_{t_1}^{t_2} \geq 0.$$

For $[t_1, t_2] = [t_0, t_f]$ this provides

$$J[x, u] \geq [x'Px]^{t_0} + [x'(S - P)x]^{t_f}.$$

Noting (1.3) and (2.3) gives

$$J[x, u] \geq \alpha' Z'(S - P(t_f))Z\alpha \quad \text{for all } \alpha.$$

Condition (2.4) completes the proof.

Explicit conditions of greater utility are directly provided.

COROLLARY 1.1. *Let $P(t)$ be an $n \times n$ symmetric matrix of bounded variation on $[t_0, t_f]$ such that (2.4) holds and the matrix*

$$(2.6) \quad \psi(t) = \begin{bmatrix} P(t) + \int_{t_0}^t A'P + PA + Q \, d\tau & \int_{t_0}^t C' + PB \, d\tau \\ \int_{t_0}^t C + B'P \, d\tau & \int_{t_0}^t R \, d\tau \end{bmatrix}$$

is nondecreasing on $[t_0, t_f]$. Then $J[x, u] \geq 0$ for all (x, u) satisfying (1.2) and (1.3).

Proof. Consider any (x, u) satisfying (1.2) and (1.3)₁. Trivially,

$$\int_{t_1}^{t_1} \begin{bmatrix} x \\ u \end{bmatrix}' d\psi \begin{bmatrix} x \\ u \end{bmatrix} \geq 0.$$

This is condition (2.5), which completes the proof.

COROLLARY 1.2. *Let $P(t)$ be an $n \times n$ symmetric matrix continuously differentiable on $[t_0, t_f]$ such that (2.4) holds and*

$$(2.7) \quad \Lambda(t) = \begin{bmatrix} \dot{P} + A'P + PA + Q & C' + PB \\ C + B'P & R \end{bmatrix} \geq 0.$$

Then $J[x, u] \geq 0$ for all (x, u) satisfying (1.2) and (1.3).

Proof. The matrix $P(t)$ is of bounded variation on $[t_0, t_f]$ [11, § 28.5 D]. Further, the matrix

$$\psi(t) = \int_{t_0}^t \Lambda \, dt$$

satisfies (2.6) and is nondecreasing, which completes the proof.

Corollary 1.1 has been stated by Coppel [10], while Corollary 1.2 is due to Jacobson [12, Thm. 1]. The explicit sufficient conditions given by Anderson [8] and by Krasner and Kailath [9] are essentially the same as Corollary 2.1.

3. Necessary conditions. The approach involves the study of an optimal control problem. For $t_1 \in (t_0, t_f)$ consider

$$(3.1) \quad V(\xi, t_1) = \inf_u \int_{t_1}^{t_f} q(x, u) \, dt + x'(t_f)Sx(t_f)$$

subject to

$$(3.2) \quad u(t) \text{ is piecewise continuous on } [t_1, t_f],$$

and

$$(3.3) \quad \dot{x} = Ax + Bu, \quad x(t_1) = \xi, \quad Dx(t_f) = 0.$$

We begin by establishing the essential properties of $V(\xi, t)$. Denote the transition matrix [13, § 3] of (1.3)₁ by $\Phi(\lambda, \zeta)$.

PROPOSITION 1. *Let*

$$(3.4) \quad \int_{t_0}^{t_1} \Phi(t_0, t)B(t)B'(t)\Phi'(t_0, t) dt > 0 \quad \text{for all } t_1 \in (t_0, t_f],$$

and let

$$(3.5) \quad D \left\{ \int_{t_1}^{t_f} \Phi(t_f, t)B(t)B'(t)\Phi'(t_f, t) dt \right\} D' > 0 \quad \text{for all } t_1 \in [t_0, t_f].$$

If $J[x, u] \geq 0$ for all (x, u) satisfying (1.2) and (1.3), then $V(\xi, t_1)$ is finite for all ξ and all $t_1 \in (t_0, t_f)$.

Proof. Consider any ξ and any $t_1 \in (t_0, t_f)$. Condition (3.5) is a weakened controllability condition and is sufficient (see Lemma 1, Appendix) for the existence of a function pair (x_c, u_c) satisfying (3.2) and (3.3). Hence the optimization problem is not vacuous, and it remains to show that V is finite. Condition (3.4) is equivalent to the standard notion of complete reachability over $[t_0, t_1]$ [14, Prop. 2.3]. This guarantees a pair (x_r, u_r) where $u_r(t)$ is continuous and where $x_r(t)$ satisfies

$$\dot{x} = Ax + Bu, \quad x(t_0) = 0, \quad x(t_1) = \xi.$$

Further, consider any function pair (x, u) satisfying (3.2) and (3.3). The concatenated control $(u_r; u)$ on $[t_0, t_f]$ and the resulting state $(x_r; x)$ satisfy conditions (1.2) and (1.3). The nonnegativity condition provides

$$(3.6) \quad \int_{t_0}^{t_1} q(x_r, u_r) dt + \int_{t_1}^{t_f} q(x, u) dt + x'(t_f)Sx(t_f) \geq 0.$$

The cost functional of (3.1) has a lower bound and $V(\xi, t_1)$ is finite, which completes the proof.

PROPOSITION 2. *The function $V(\xi, t)$ guaranteed by Proposition 1 admits the representation*

$$(3.7) \quad V(\xi, t) = \xi'P(t)\xi$$

for some $n \times n$ symmetric matrix $P(t)$ on (t_0, t_f) .

Proof. Consider any $t_1 \in (t_0, t_f)$ and, where convenient, write $V(\xi, t_1)$ as $V(\xi)$. The proof consists of demonstrating the conditions

$$(3.8) \quad |V(\xi)| \leq c\xi'\xi \quad \text{for some positive } c,$$

and

$$(3.9) \quad V(\xi_1 + \xi_2) + V(\xi_1 - \xi_2) = 2\{V(\xi_1) + V(\xi_2)\},$$

which are shown in the Appendix to be sufficient for (3.7).

Consider the function pairs (x_r, u_r) and (x_c, u_c) used in the proof of Proposition 1. By (3.1) and (3.6),

$$-\int_{t_0}^{t_1} q(x_r, u_r) dt \leq V(\xi, t_1) \leq \int_{t_1}^{t_f} q(x_c, u_c) dt + x'_c(t_f)Sx_c(t_f).$$

The standard construction for $u_r(t)$ [13, Thm. 13.1] provides that both $u_r(t)$ and $x_r(t)$ are continuous on $[t_0, t_1]$ and are linear in ξ . The quadratic nature of $q(x, u)$ provides that

$$\int_{t_0}^{t_1} q(x_r, u_r) dt \leq c_1 \xi' \xi$$

for some positive number c_1 (dependent on t_1). Similarly, both $u_c(t)$ and $x_c(t)$ are continuous on $[t_1, t_f]$ and are linear in ξ (see Lemma 2, Appendix). Again,

$$\int_{t_1}^{t_f} q(x_c, u_c) dt + x'_c(t_f)Sx_c(t_f) \leq c_2 \xi' \xi,$$

for some positive number c_2 . Together, then,

$$-c_1 \xi' \xi \leq V(\xi, t_1) \leq c_2 \xi' \xi,$$

which provides (3.8).

For (x, u) satisfying (3.2) and (3.3), denote

$$J[\xi; u] = \int_{t_1}^{t_f} q(x, u) dt + x'(t_f)Sx(t_f).$$

Let (x_1, u_1) satisfy (3.2) and (3.3) with $x_1(t_1) = \xi_1$ and (x_2, u_2) satisfy (3.2) and (3.3) with $x_2(t_1) = \xi_2$. The linearity of (3.3) and the quadratic nature of $q(x, u)$ provide

$$(3.10) \quad J[\xi_1 + \xi_2; u_+] + J[\xi_1 - \xi_2; u_-] = 2\{J[\xi_1; u_1] + J[\xi_2; u_2]\},$$

where

$$(3.11) \quad u_+ = u_1 + u_2, \quad u_- = u_1 - u_2.$$

Assume that the parallelogram identity (3.9) does not hold, that is,

$$(3.12) \quad V(\xi_1 + \xi_2) + V(\xi_1 - \xi_2) = 2[V(\xi_1) + V(\xi_2)] + \varepsilon.$$

First consider the case $\varepsilon > 0$. By definition (3.1), there exist controls \hat{u}_1 and \hat{u}_2 such that

$$J[\xi_1; \hat{u}_1] \leq V(\xi_1) + \varepsilon/8,$$

$$J[\xi_2; \hat{u}_2] \leq V(\xi_2) + \varepsilon/8.$$

Adding, and noting (3.10)–(3.12), provides

$$\begin{aligned} J[\xi_1 + \xi_2; \hat{u}_+] + J[\xi_1 - \xi_2; \hat{u}_-] &\leq V(\xi_1 + \xi_2) + V(\xi_1 - \xi_2) - \varepsilon + \varepsilon/2, \\ &< V(\xi_1 + \xi_2) + V(\xi_1 - \xi_2), \end{aligned}$$

which is impossible. Now consider (3.12) for the case $\varepsilon < 0$. By definition, there

exist controls \tilde{u}_+ and \tilde{u}_- such that

$$J[\xi_1 + \xi_2; \tilde{u}_+] \leq V(\xi_1 + \xi_2) - \varepsilon/4,$$

$$J[\xi_1 - \xi_2; \tilde{u}_-] \leq V(\xi_1 - \xi_2) - \varepsilon/4.$$

Adding, and noting (3.10)–(3.12), provides a contradiction exactly as before. Hence $\varepsilon = 0$, and the proof is complete.

Remark 1. As indicated in the proof of Proposition 1, condition (3.5) is a weakened controllability condition. If D is nonsingular, it is equivalent to the standard notion of complete controllability over $[t_1, t_f]$ [14, Prop. 2.3], whereas if $D = 0$ the condition is vacuous.

It is convenient to define

$$(3.13) \quad P(t_f) = S.$$

This is consistent with (3.7) and (3.1). We now establish the essential properties of the matrix $P(t)$.

PROPOSITION 3. *The matrix $P(t)$ guaranteed by Proposition 2 satisfies*

$$(3.14) \quad \int_{t_1}^{t_2} q(x, u) dt + [x'Px]_{t_1}^{t_2} \geq 0,$$

for all $[t_1, t_2] \subset (t_0, t_f)$ and all (x, u) satisfying (1.2) and (1.3).

Proof. The condition (3.14) is immediate from the “dissipation inequality”

$$V(x(t_1), t_1) \leq \int_{t_1}^{t_2} q(x, u) dt + V(x(t_2), t_2),$$

which is obtained directly from (3.1).

The objective is to use Lemma 1 in (3.14), reversing the logic of Theorem 1. To this end, $P(t)$ has to be shown to be sufficiently well-behaved.

PROPOSITION 4. *Consider the matrix*

$$(3.15) \quad \tilde{P}(t) = \Phi'(t, t_f)\{P(t) - W(t)\}\Phi(t, t_f),$$

where $W(t)$ is the well-defined solution of the linear differential equation

$$(3.16) \quad \dot{W} + A'W + WA + Q = 0, \quad W(t_f) = S.$$

Then $\tilde{P}(t)$ is nondecreasing on (t_0, t_f) , and $Z'\tilde{P}(t)Z \leq 0$.

Proof. Consider (1.3)₁ for the special case of the zero control function, $u(t) = 0$. The resulting state satisfies

$$(3.17) \quad x(t_2) = \Phi(t_2, t_1)x(t_1).$$

It is a standard result [13, § 11] that

$$(3.18) \quad \int_{t_1}^{t_f} q(x, 0) dt + x'(t_f)Sx(t_f) = [x'Wx]_{t_1}^{t_f},$$

where $W(t)$ is the solution of (3.16). Directly,

$$(3.19) \quad \int_{t_1}^{t_2} q(x, 0) dt = [-x'Wx]_{t_1}^{t_2}.$$

Firstly, consider the inequality (3.14) for the special case of zero control. Equation (3.19) provides

$$[x'(P - W)x]_{t_1}^{t_2} \geq 0.$$

Noting (3.17) provides, in turn,

$$\Phi'(t_2, t_1)\{P(t_2) - W(t_2)\}\Phi(t_2, t_1) \geq P(t_1) - W(t_1).$$

Noting the relation $\Phi(t_2, t_1) = \Phi(t_2, t_f)\Phi^{-1}(t_1, t_f)$ provides the nondecreasing condition on $\tilde{P}(t)$.

Secondly, definition (3.11) provides

$$x'(t_1)P(t_1)x(t_1) \leq \int_{t_1}^{t_f} q(x, 0) dt + x'(t_f)Sx(t_f)$$

for all $x(t_f)$ satisfying $Dx(t_f) = 0$. Noting (3.18), (3.17) and (2.3) provides

$$x'Z'\Phi'(t_1, t_f)\{P(t_1) - W(t_1)\}\Phi(t_1, t_f)Z\alpha \leq 0 \quad \text{for all } \alpha.$$

That is, $Z'\tilde{P}(t_1)Z \leq 0$, which completes the proof.

Remark 2. If $D = 0$, then it is immediate that the inequality (3.14) holds for all $[t_1, t_2] \subset (t_0, t_f]$, and consequently $\tilde{P}(t)$ is nondecreasing on $(t_0, t_f]$.

The passage to the basic sufficiency result is now easy.

THEOREM 2. Assume conditions (3.4) and (3.5). If $J[x, u] \geq 0$ for all (x, u) satisfying (1.2) and (1.3), then there exists an $n \times n$ symmetric matrix $P(t)$ of bounded variation on all $[t_1, t_2] \subset (t_0, t_f)$ which satisfies

$$(3.20) \quad \lim_{t \uparrow t_f} Z'\{\Phi'(t, t_f)P(t)\Phi(t, t_f) - S\}Z \leq 0,$$

and

$$(3.21) \quad \int_{t_1}^{t_2} \begin{bmatrix} x \\ u \end{bmatrix}' \begin{bmatrix} dP + (A'P + PA + Q) dt & (C' + PB) dt \\ (C + B'P) dt & R dt \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} \geq 0$$

for all (x, u) satisfying (1.2) and (1.3)₁.

Proof. The existence of $P(t)$ is provided by Proposition 2. From Proposition 4 we have the representation

$$P(t) = \Phi'(t_f, t)\tilde{P}(t)\Phi(t_f, t) + W(t),$$

where $\Phi(t_f, t)$ and $W(t)$ are continuously differentiable and $\tilde{P}(t)$ is nondecreasing. Standard theory [11, § 28] provides that $P(t)$ is of bounded variation. Proposition 4 further provides that $Z'\tilde{P}(t)Z$ is nondecreasing and has 0 as an upper bound. Thus

$$\lim_{t \uparrow t_f} Z'\tilde{P}(t)Z \leq 0,$$

where the limit is guaranteed to exist [11, § 10.2F]. This is precisely (3.20). Finally (3.21) is obtained by combining Lemma 1 with the inequality (3.14).

In general, condition (3.20) does not imply that $P(t_{f-})$ exists.

Remark 3. If $D = 0$, it is clear that condition (3.5) is vacuous and in Theorem 2, $P(t)$ is of bounded variation and (3.21) holds, on all $[t_1, t_2] \subset (t_0, t_f]$. Moreover,

(3.20) reduces to

$$(3.22) \quad P(t_f-) - S \leq 0.$$

The following explicit conditions parallel those of § 2, and are of greater utility.

COROLLARY 2.1. Consider the matrix

$$(3.23) \quad \psi(t) = \begin{bmatrix} P(t) + \int_{t_3}^t A'P + PA + Q \, d\tau & \int_{t_3}^t C' + PB \, d\tau \\ \int_{t_3}^t C + B'P \, d\tau & \int_{t_3}^t R \, d\tau \end{bmatrix},$$

for any $t_3 \in (t_0, t_f)$, where $P(t)$ is the matrix guaranteed by Theorem 2. Then $\psi(t)$ is nondecreasing on (t_0, t_f) . If $D = 0$, it is nondecreasing on $(t_0, t_f]$.

Proof. In (3.21) consider the special case of $u(t) = \eta$, constant on $[t_1, t_2]$, and denote $x(t_1) = \xi$. Then

$$\begin{bmatrix} x(t) \\ u(t) \end{bmatrix} = \Sigma(t) \begin{bmatrix} \xi \\ \eta \end{bmatrix},$$

where

$$\Sigma(t) = \begin{bmatrix} \Phi(t, t_1) & \int_{t_1}^t \Phi(t, \tau) B(\tau) \, d\tau \\ 0 & I \end{bmatrix}.$$

Note that $\Sigma(t)$ is nonsingular on $[t_0, t_f]$. Condition (3.21) then provides

$$\int_{t_1}^{t_2} \Sigma'(t) \, d\psi \, \Sigma(t) \geq 0.$$

Equivalently, the matrix

$$(3.24) \quad \bar{P}(t) = \int_{t_3}^t \Sigma'(t) \, d\psi \, \Sigma(t)$$

is nondecreasing on (t_0, t_f) . The substitution rule [11, Thm. 29.6.1] provides

$$\int_{t_1}^{t_2} d\psi = \int_{t_1}^{t_2} \Sigma'^{-1}(t) \, d\bar{P} \, \Sigma^{-1}(t) \geq 0,$$

by the nondecreasing property of $\bar{P}(t)$. In other words, $\psi(t)$ is nondecreasing. The modification for $D = 0$ follows from the last remark.

COROLLARY 2.2. Consider the matrix $P(t)$ guaranteed by Theorem 2. Then

$$(3.25) \quad \begin{bmatrix} \dot{P} + A'P + PA + Q & C' + PB \\ C + B'P & R \end{bmatrix} \geq 0 \quad \text{a.e.}$$

Proof. It is a standard result that $(d\psi(t)/dt)$ exists almost everywhere [15, p. 100]. Condition (3.25) is simply the fact that $d\psi/dt \geq 0$.

Remark 4. The gap between Theorems 1 and 2, and the corresponding Corollaries 1.1 and 2.1 is limited to that of the endpoint behavior of $P(t)$, namely the limits $t \downarrow t_0$ and $t \uparrow t_f$. The gap between Corollaries 1.2 and 2.2 is somewhat larger.

Corollary 2.1 has also been derived by Coppel [10], by a limiting argument on the known necessary conditions for the nonsingular ($R > 0$) case. The more complicated condition of $\bar{P}(t)$ nondecreasing (see (3.24)) has also been derived by Anderson [8], and by Krasner and Kailath [9], by transformation arguments on necessary conditions for the singular ($R = 0$) case. As these conditions were obtained in turn by limiting arguments on the nonsingular conditions [16], the logical chain is rather long.

4. Special cases. The relation of Theorem 2 to earlier necessary conditions is easily established.

The Legendre necessary condition, that $R(t) \geq 0$ for all $t \in [t_0, t_f]$, is immediate from (3.25) and the assumption of continuity on R .

If nonsingularity of R obtains over an interval, then (3.25) reduces to

$$\dot{P} + A'P + PA + Q - (C' + PB)R^{-1}(C + B'P) \geq 0, \quad \text{a.e.}$$

In fact, far more can be said.

COROLLARY 2.3. *Assume that $R(t) \geq 0$ on some $[t_1, t_2] \subset (t_0, t_f)$. Then the matrix $P(t)$ guaranteed by Theorem 2 is continuously differentiable on $[t_1, t_2]$, and satisfies*

$$(4.1) \quad \dot{P} + A'P + AP + Q - (C' + PB)R^{-1}(C + B'P) = 0.$$

Proof. Consider the Riccati differential equation

$$(4.2) \quad \dot{X} + A'X + XA + Q - (C' + XB)R^{-1}(C + B'X) = 0, \quad X(t_2) = P(t_2).$$

A solution exists at least on a neighborhood of t_2 , say $[t_3, t_2]$. Consider any (x, u) satisfying

$$\dot{x} = Ax + Bu, \quad x(t_3) = \xi, \quad Dx(t_f) = 0.$$

The standard "completion-of-squares" lemma [13, § 21] provides

$$(4.3) \quad \int_{t_3}^{t_2} q(x, u) dt + x'(t_2)X(t_2)x(t_2) = \xi'X(t_3)\xi + \int_{t_3}^{t_2} \|u + R^{-1}(C + B'X)x\|_R^2 dt.$$

Substitution of this identity into the inequality (3.14) provides

$$\xi'P(t_3)\xi \leq \xi'X(t_3)\xi,$$

once the special feedback control $u = -R^{-1}(C + B'X)x$ is considered over $[t_3, t_2]$.

Conversely, it follows from (4.3) that

$$\begin{aligned} \int_{t_3}^{t_f} q(x, u) dt + x'(t_f)Sx(t_f) &\geq \int_{t_3}^{t_2} q(x, u) dt + [x'Px]^{t_2}, \\ &\geq \int_{t_3}^{t_2} q(x, u) dt + [x'Xx]^{t_2}, \\ &\geq \xi'X(t_3)\xi. \end{aligned}$$

Hence

$$\xi'P(t_3)\xi \geq \xi'X(t_3)\xi.$$

Together, then,

$$(4.4) \quad P(t_3) = X(t_3).$$

This excludes any possibility of a finite-time escape phenomenon in the solution of (4.2). Hence (4.4) holds for all $t_3 \in [t_1, t_2]$ and the proof is complete.

If $D = 0$, the argument remains valid for the extension to $[t_1, t_2] \subset (t_0, t_f]$. Finally, if R is nonsingular over the interval $[t_0, t_f]$, these results match those usually obtained by the Jacobi conjugate-point analysis [2, p. 123], [17].

On the other hand, if $R = 0$ on some open interval $(t_1, t_2) \subset (t_0, t_f)$, then (3.25) provides

$$C + B'P = 0 \quad \text{a.e.}$$

In fact, an argument of Jacobson [16, Thm. B.1] shows that this equality cannot fail on the open interval (t_1, t_2) . If $R = 0$ on the entire interval (t_0, t_f) , then the necessary conditions of Jacobson [7] are obtained. It is worth noting that Jacobson invokes the classical concept of normality [18, § 8.2], [1, p. 164] instead of the controllability condition (3.5). Finally, other necessary conditions such as higher order Legendre–Clebsch conditions and Jacobi conditions can be derived, somewhat in the spirit of Jacobson [7].

5. Example. A simple example [18, § 5.11] may help to illuminate the necessary conditions of Theorem 2. In the terminology of § 1, consider

$$J[x, u] = \int_0^{t_f} -x^2 + u^2 dt,$$

where u is piecewise continuous and x satisfies

$$\dot{x} = u, \quad x(0) = 0, \quad x(t_f) = 0.$$

For $t_f > \pi$, J can be negative (consider $u = \cos(\pi t/t_f)$), whereas for $t_f = \pi$, J is nonnegative [18, § 5.11], [19]. Consider the optimal control problem of § 3, namely,

$$(5.1) \quad V(\xi, t_1) = \min_u \int_{t_1}^{\pi} -x^2 + u^2 dt,$$

where u is piecewise continuous and

$$(5.2) \quad \dot{x} = u, \quad x(t_1) = \xi, \quad x(\pi) = 0.$$

The Riccati differential equation,

$$(5.3) \quad \dot{p} = p^2 + 1,$$

is easily checked to have solutions on $[t_1, \pi]$, namely,

$$(5.4) \quad p_\varepsilon(t) = -\cot(t - \varepsilon) \quad \text{for any } 0 < \varepsilon < t_1.$$

Since

$$\begin{bmatrix} \dot{p} - 1 & p \\ p & 1 \end{bmatrix} \geq 0,$$

Theorem 1 provides

$$\int_{t_1}^{\pi} -x^2 + u^2 dt \geq -\xi^2 \cot(t - \varepsilon) \quad \text{for any } 0 < \varepsilon < t_1,$$

for any (x, u) satisfying (5.2). Considering $\lim_{\varepsilon \downarrow 0}$ (to provide the tightest bound) gives

$$\int_{t_1}^{\pi} -x^2 + u^2 \geq -\xi^2 \cot t.$$

Finally note that this lower bound is in fact achieved by the control

$$u = (\xi/\sin t_1) \cos t.$$

Thus

$$V(\xi, t) = -\xi^2 \cot t,$$

$$P(t) = -\cot t.$$

The conditions of Theorem 2 and its corollaries can be checked by inspection. Note that $P(t)$ is unbounded as $t \downarrow 0$ and $t \uparrow \pi$, and is thus of bounded variation only on the open interval $(0, \pi)$. The necessary conditions of Theorem 2 thus cannot be improved in general. Finally note that for $0 < t_f < \pi$,

$$V(\xi, t) = -\xi^2 \cot(t - t_f + \pi).$$

In this case, then $\lim_{t \downarrow 0} P(t)$ exists while $\lim_{t \uparrow t_f} P(t)$ does not.

Now consider the same problem for unrestricted $x(t_f)$. For $t_f > \pi/2$, J can be negative (again consider $u(t) = \cos(\pi t/t_f)$), whereas for $t_f = \pi/2$, J is non-negative [18, § 5.11]. By the same argument as before

$$V(\xi, t) = -\xi^2 \cot t,$$

$$P(t) = -\cot t.$$

Note that $\lim_{t \uparrow \pi/2} P(t) = 0$ and that $P(t)$ is of bounded variation on $(0, \pi/2]$, consistent with Remark 3. Again Theorem 2 cannot be improved in general.

6. Conclusion. This paper has derived sets of sufficient conditions and necessary conditions for the nonnegativity of the quadratic functional (1.1). These conditions complement the well-known theory for the special nonsingular and singular cases.

In as far as one set of such conditions has already appeared in the literature [8], the main contribution of the paper is seen to be the method used. It directly handles the general problem in a relatively straightforward fashion, and is self-contained in comparison with the other limiting/transformation approaches to the problem.

Appendix.

LEMMA 2. *If condition (3.5) holds, then there exists a pair (x, u) satisfying*

$$(1) \quad \dot{x} = Ax + Bu, \quad x(t_1) = \xi, \quad Dx(t_f) = 0.$$

Proof. It is sufficient [13, Thm. 13.1] to demonstrate the existence of an η satisfying

$$(2) \quad W\eta = \xi - \Phi(t_1, t_f)Z\alpha \quad \text{for some } \alpha,$$

where

$$W = \int_{t_1}^{t_f} \Phi(t_1, t)B(t)B'(t)\Phi'(t_1, t) dt.$$

Firstly, consider the $n \times n$ matrix $[Z|D']$. This matrix is nonsingular and thus

$$[Z|D']\lambda = \Phi(t_f, t_1)\xi$$

has a unique solution λ . An obvious partitioning of λ gives

$$(3) \quad \Phi(t_f, t_1)\xi = Z\alpha + D'\beta.$$

Secondly, note that (3.5) is equivalent to the condition

$$(4) \quad D\Phi(t_f, t_1)W \quad \text{is of full rank.}$$

Thus the matrix equation

$$D\Phi(t_f, t_1)W\eta = DD'\beta$$

has a solution η_0 . Further, the vector

$$(5) \quad \gamma = \Phi(t_f, t_1)W\eta_0 - D'\beta$$

satisfies $D\gamma = 0$, or equivalently,

$$(6) \quad \gamma = Z\hat{\alpha} \quad \text{for some } \hat{\alpha}.$$

Collecting these results gives

$$\Phi(t_f, t_1)W\eta_0 = \Phi(t_f, t_1)\xi - Z(\alpha + \hat{\alpha}),$$

which provides (2) and completes the proof.

The remainder of the Appendix is concerned with conditions that are sufficient for a function $V(x)$ to be a quadratic form, that is,

$$V(x) = x'Px.$$

The first condition imposed is the parallelogram identity

$$(7) \quad V(x + y) + V(x - y) = 2\{V(x) + V(y)\} \quad \text{for all } x, y.$$

Immediate consequences for $V(x)$ are

$$V(0) = 0, \quad V(-x) = V(x), \quad V(2x) = 4V(x).$$

Now consider the associated function $W(x, y)$ defined by

$$(8) \quad W(x, y) = V(x + y) - V(x - y).$$

In view of the immediate identities,

$$4V(x) = W(x, x), \quad W(y, x) = W(x, y),$$

it will suffice [20, p. 244] to establish that $W(x, y)$ is a bilinear form. This is “almost” implied by the parallelogram identity.

LEMMA 3. Assume (7). Then $W(x, y)$ satisfies

$$(9) \quad W(x_1 + x_2, y) = W(x_1, y) + W(x_2, y),$$

and

$$(10) \quad W(kx, y) = kW(x, y) \quad \text{for all rational } k.$$

Proof. Trivially, we have

$$(11) \quad W(0, y) = 0, \quad W(-x, y) = -W(x, y).$$

Now

$$\begin{aligned} 2\{W(x_1, y) + W(x_2, y)\} &= 2\{V(x_1 + y) - V(x_1 - y) + V(x_2 + y) - V(x_2 - y)\}, \\ &= \{V(x_1 + x_2 + 2y) + V(x_1 - x_2)\} \\ &\quad - \{V(x_1 + x_2 - 2y) + V(x_1 - x_2)\}, \\ &= W(x_1 + x_2, 2y). \end{aligned}$$

Putting $x_2 = 0$ provides

$$2W(x_1, y) = W(x_1, 2y) \quad \text{for all } x_1.$$

That is,

$$W(x_1 + x_2, 2y) = 2W(x_1 + x_2, y),$$

and (9) follows. Secondly, induction on (9) provides

$$W(nx, y) = nW(x, y) \quad \text{for all integers } n.$$

Then

$$mW\left(\frac{n}{m}x, y\right) = W(nx, y) = nW(x, y) \quad \text{for all } m, n,$$

which is equation (10).

It remains to find a convenient condition for (10) to hold for all real k , and not just rational k .

LEMMA 4. *If (7) holds and*

$$(12) \quad W(\lambda x, y) \text{ is continuous in } \lambda \text{ at } \lambda = 0,$$

then (10) holds for all real k , and $V(x)$ is a quadratic form.

Proof. Consider any real k . There exists a sequence of rational numbers r_n such that $\lim_{n \rightarrow \infty} r_n = k$. Writing $k = r_n + \lambda_n$, it follows from (9) and (10) that

$$\begin{aligned} W(kx, y) &= W(\lambda_n x, y) + W(r_n x, y), \\ &= W(\lambda_n x, y) + r_n W(x, y). \end{aligned}$$

Taking limits, the continuity assumption provides

$$\begin{aligned} W(kx, y) &= W(0x, y) + kW(x, y), \\ &= 0 + kW(x, y), \end{aligned}$$

which completes the lemma.

Note that the limited continuity property (12) is certainly implied if $V(x)$ is continuous in x . This set of sufficient conditions are given in [20, pp. 244–246], and apparently are implicitly used in [21, pp. 24–25].

Of more relevance to this paper, however, is the following result.

LEMMA 5. *If (7) holds and*

$$(13) \quad |V(x)| \leq cx'x \quad \text{for some positive } c,$$

then $V(x)$ is a quadratic form

Proof. Immediately

$$\begin{aligned} |W(x, y)| &\leq |V(x + y)| + |V(x - y)| \\ &\leq 2c\{x'x + y'y\}. \end{aligned}$$

From condition (10), for all integers n and real numbers λ ,

$$\begin{aligned} |nW(\lambda x, y)| &= |W(n\lambda x, y)| \\ &\leq 2c\{n^2\lambda^2 x'x + y'y\}. \end{aligned}$$

Consider any $\varepsilon > 0$, and choose a positive integer N and positive real δ to satisfy

$$\frac{1}{\delta} \geq N \geq \frac{2c}{\varepsilon} \{x'x + y'y\}.$$

Then for all λ satisfying $|\lambda| \leq \delta$, it follows that

$$N^2\lambda^2 \leq N^2\delta^2 \leq 1,$$

and

$$\begin{aligned} N|W(\lambda x, y)| &\leq 2c\{x'x + y'y\} \\ &\leq N\varepsilon, \end{aligned}$$

by choice of N . Hence $W(\lambda x, y)$ is continuous in λ at $\lambda = 0$, and $V(x)$ is a quadratic form by Lemma 4.

REFERENCES

- [1] A. E. BRYSON AND Y. C. HO, *Applied Optimal Control*, Ginn, Waltham, Mass., 1969.
- [2] J. M. GEL'FAND AND S. V. FOMIN, *Calculus of Variations*, Prentice-Hall, Englewood Cliffs, N.J., 1963.
- [3] J. B. MOORE AND B. D. O. ANDERSON, *Extensions of quadratic minimization theory: I. Finite time results*, Internat. J. Control, 7 (1968), pp. 465–472.
- [4] J. C. WILLEMS, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 621–634.
- [5] N. KRASNER AND T. KAILATH, *A stochastic interpretation of singular quadratic minimization theory. I: General conditions for minimality*, submitted for publication.
- [6] M. R. WOHLERS, *Lumped and Distributed Passive Networks*, Academic Press, New York, 1969.
- [7] D. H. JACOBSON, *Totally singular quadratic minimization problems*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 651–659.
- [8] B. D. O. ANDERSON, *Partially-singular linear-quadratic control problems*, IEEE Trans. Automatic Control, AC-18 (1973), pp. 407–409.
- [9] N. KRASNER AND T. KAILATH, *A stochastic interpretation of singular quadratic minimization theory. II: Generalized Legendre–Clebsch conditions, Riccati differential equations, and partially singular problems*, submitted for publication.
- [10] W. A. COPPEL, *Notes on linear quadratic optimal control*, in preparation.
- [11] R. A. RANKIN, *An Introduction to Mathematical Analysis*, Pergamon Press, Oxford, England, 1963.
- [12] D. H. JACOBSON, *A general sufficiency theorem for the second variation*, J. Math. Anal. Appl., 34 (1971), pp. 578–589.
- [13] R. W. BROCKETT, *Finite-Dimensional Linear Systems*, John Wiley, New York, 1970.
- [14] R. E. KALMAN, *Lectures on controllability and observability*, Centro Internazionale Matematico Estivo, Bologna, Italy, 1968.
- [15] H. L. ROYDEN, *Real Analysis*, Macmillan, New York, 1968.
- [16] D. H. JACOBSON AND J. L. SPEYER, *Necessary and sufficient conditions for optimality for singular control problems: A limit approach*, J. Math. Anal. Appl., 34 (1971), pp. 239–266.
- [17] W. A. COPPEL, *Disconjugacy*, Lecture Notes in Mathematics, vol. 220, Springer-Verlag, Berlin, 1971.
- [18] L. A. PARS, *An Introduction to the Calculus of Variations*, Heinemann, London, 1962.
- [19] R. TAPIA, *An extremum problem*, SIAM Rev., 15 (1973), pp. 386–387.
- [20] W. H. GREUB, *Linear Algebra*, Springer-Verlag, Berlin, 1967.
- [21] B. D. O. ANDERSON AND J. B. MOORE, *Linear Optimal Control*, Prentice-Hall, Englewood Cliffs, N.J., 1971.

FINITE TIME CONTROLLABILITY OF NONLINEAR CONTROL PROCESSES*

ETHELBERT N. CHUKWU†

Abstract. A control process is globally finite time null controllable if it is globally asymptotically stable and locally controllable to the origin. Sufficient conditions are stated for the system

$$\dot{x} = f(t, x, u) \quad \text{in } C^1(R \times R^n \times R^m)$$

to be globally finite time null controllable. The conditions are stated in terms of the Jacobian of f and the controllability of a related linear equation.

1. Introduction. Consider the control process

$$(1.1) \quad dx/dt = f(t, x, u),$$

where $f: R \times R^n \times R^m \rightarrow R^n$ is such that $f(t, x, u)$ and $\partial f / \partial x(t, x, u)$ are continuous in (t, x, u) . It is assumed that u belongs to the set of all bounded measurable control functions with values in Ω , a subset of R^m . Such control functions are called admissible.

The domain \mathcal{C} of null controllability of (1.1) is defined as the set of initial points $x_0 \in R^n$ such that if $u(t)$ is an admissible control and $x(t)$ is a solution of

$$(1.2) \quad \frac{dx}{dt} = f(t, x, u(t)) \quad \text{with } x(t_0) = x_0,$$

then at some finite time $t_1 \geq t_0$ we have $x(t_1) = 0$. If \mathcal{C} contains an open neighborhood of zero, then (1.1) is said to be locally controllable to the origin.

The system (1.1) is globally asymptotically stable if, for each $\varepsilon > 0$ there exists $\delta = \delta(t_0, \varepsilon)$ such that $\|x_0\| < \delta$ implies that every solution $x(t)$ of (1.1) with some admissible control $u^*(t)$, initiating at $x(t_0) = x_0$ satisfies $\|x(t)\| < \varepsilon$ on $t_0 \leq t < \infty$, $\lim_{t \rightarrow \infty} x(t) = 0$, and every solution of (1.2) with this u^* admissible can be extended over $t_0 \leq t < \infty$ and tends toward the origin as $t \rightarrow \infty$.

The system (1.1) is globally finite time null controllable if it is globally asymptotically stable and locally controllable to the origin.

This paper formulates sufficient conditions on f which guarantee the global finite time null controllability of the control process (1.1). The first result, Theorem 1 is on the global asymptotic stability of (1.1). Theorem 2 indicates when the domain of null controllability of (1.1) is open in R^n . Thus both theorems unite to yield the required global finite time null controllability of the process (1.1).

We note that the usual definition of stability very often requires that $f(t, 0, u) \equiv 0$, $t \in [t_0, \infty)$, so that the equilibrium $x = 0$ is the solution which is desired to be "stable". But the definition of stability given above does not require an equilibrium of (1.1) to exist: $f(t, 0, u)$ is not necessarily zero. The point $x = 0$ is the point chosen about which trajectories of the system are "stable", but even if an equilibrium exists, $x = 0$ is not necessarily the equilibrium. The term stability is used above for lack of a better word. The essential fact in the above definition

* Received by the editors January 28, 1974, and in revised form May 1, 1974.

† Department of Mathematics, Cleveland State University, Cleveland, Ohio 44115.

is that every solution of $\dot{x} = f(t, x, u^*(t))$ with $u^* \in \Omega$ arrive in a neighborhood of the origin. When the system is locally controllable every solution is then brought to the origin in finite time, a justification of the term "finite time null controllable".

For the origins and importance of these notions see [2, p. 78] and [3, p. 397]. Indeed Theorem 17.6 of [2] is a linear analogue of our Theorem 3 while [3, p. 397] and Markus [4] is the autonomous version of our result. But the result of Markus is quite different in spirit from ours since we rely on the far deeper and interesting paper [5] for the proof of Theorem 2.

It is clear that Theorem 1 reduces to Ezeilo [1; Thm. 1] on the equation

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

Consequently the fundamental notion here, global finite time controllability could as well be called global finite time stability. But the stability studied here differs from the studies of Weiss and Infante [7] and [8]: our concept of finite time stability is more in line with the classical Liapunov theory of asymptotic stability, though in this case everything happens in finite time.

Notation. In what follows we denote the Jacobian matrix $\partial f / \partial f$ by J and its transpose by J^T .

2. Statement of results.

THEOREM 1. Suppose in (1.1), for some admissible control $u^*(t) \in \Omega$,

(i) there exists a symmetric positive definite $n \times n$ constant matrix A such that the eigenvalues $\lambda_k(x, u^*, t)$, $k = 1, 2, \dots, n$, of the matrix

$$\frac{1}{2}(AJ + J^T A)$$

satisfy

$$(2.1) \quad \lambda_k \leq -\delta < 0, \quad k = 1, 2, \dots, n,$$

for all $(x, t) \in \mathbb{R}^{n+1}$, where δ is a constant;

(ii) there are constants $r > 0$ and ρ , $1 \leq \rho \leq 2$, such that

$$(2.2) \quad \int_t^{t+r} \|f(\tau; 0, u^*)\| \rightarrow 0 \quad \text{as } t \rightarrow \infty.$$

Then the differential equation (1.1) is globally asymptotically stable about the origin. Hence for each $x_0 \in \mathbb{R}^n$ the solution $x(t)$ of (1.1), with $u(t) = u^*(t)$ and $x(t_0) = x_0$, tends toward $x_1 = 0$ as $t \rightarrow \infty$.

In the special case

$$f(t, x, u) \equiv f(t, x),$$

Theorem 1 yields

$$(2.3) \quad x(t) \rightarrow 0 \quad \text{as } t \rightarrow \infty,$$

and this is Ezeilo's Theorem 1 in [1]. Also when $f(t, x, u) \equiv f(x, u)$ we have from Theorem 1 a result similar to Lee and Marcus [3, pp. 397–398] obtained under less generous conditions.

The next result gives the local controllability condition. Consider the linear equation

$$(2.4) \quad \dot{x}(t) = L(t)x(t) + q(t), \quad q(t) \in Q(t) \subseteq R^n,$$

where $L(t) = f_x(t, 0, 0)$ is an $n \times n$ matrix which is continuous in t . Here $Q(t) = K(\Lambda(t))$, where $K(\Lambda)$ is the unbounded closed convex cone of Λ with $\Lambda(t) = f(t, 0, \Omega)$; that is, $K(\Lambda)$ is the smallest closed convex set containing Λ such that $v \in K(\Lambda)$ implies $\delta v \in K(\Lambda)$ for all $\delta > 0$.

DEFINITION. Consider the control process (2.4). The *reachable set* of (2.4) is the set

$$\mathcal{R}(t, t_0) = \left\{ \int_{t_0}^t X^{-1}(s, t_0)q(s) ds : q(t) \in Q(t), q \text{ summable} \right\},$$

where $X(t, t_0)$ is the transition matrix of (2.4).

DEFINITION (see [2]). The system (2.4) is *proper* at t_0 if and only if the origin is an interior point of the reachable set $\mathcal{R}(t, t_0)$ of (2.4) for each $t > t_0$. The system (2.4) is called *proper* if and only if it is proper for each $t_0 \geq 0$.

THEOREM 2. Assume that $0 \in \Omega$ and $f(t, 0, 0) = 0$. Suppose the linear system (2.4) is *proper*. Then (1.1) is *locally controllable (near the origin)*.

THEOREM 3. Assume all the conditions of Theorem 1 and Theorem 2; then (1.1) is *globally finite time null controllable*.

For completeness it is important to describe conditions on L and Q which insure that the property "proper" assumed in the hypothesis of Theorem 2 be achieved. Such a description is available in two cases:

(i) The usual one, it is required that Q be time-invariant and have a non-empty interior in R^n with the origin contained in this interior. In this case if $q(t) = B(t)u(t)$, $u(t) \in Q$, then (2.4) is proper if and only if for each $t_1 > t_0$, there is a $t \in (t_0, t_1)$ such that

$$\text{rank} [B(t), \Gamma B(t), \dots, \Gamma^{n-1} B(t)] = n,$$

where $\Gamma = -L(t) + D$, with D the differential operator, is as defined in [2, p. 75]. Here $L(t)$ and $B(t)$ are assumed analytic. For a slightly more general description when Q is not necessarily time-invariant, but is assumed compact see Chukwu [11].

(ii) Here the origin is not assumed to be interior to the range of the set of admissible controls and

$$q(t) = Bu(t), \quad u \in Q,$$

where L, B, Q are all time-invariant. The following characterization is given by Brammer [9].

THEOREM (Brammer). Suppose the control system

$$(M) \quad \dot{x} = Lx + Bu, \quad u \in Q \subseteq R^n,$$

has the following properties:

(a) The set Q contains a vector in the kernel of B (i.e., there exists a $u \in Q$ satisfying $Bu = 0$).

(b) The convex hull of $Q = CH(Q) = Q$ has nonempty interior in R^n . The following conditions are necessary and sufficient for the system (M) to be proper.

(c) $\text{rank}[B, LB, \dots, L^{n-1}B] = n$.

(d) There is no eigenvalue v of L^T satisfying $(vBu) \leq 0$ for all $u \in Q$. The parentheses denote the scalar product in R^n .

The following slightly more general result is recently given by Brammer [10] under the assumption that the origin is a boundary point of Q .

THEOREM (Brammer). The control system

$$\dot{x} = Lx + q(t), \quad q \in Q \subset R^n,$$

where L is constant, is proper if and only if there is no cyclic subspace of L^T which contains a relatively open cone of exterior normals to Q .

The phrase relatively open means that the cone Q is open in the subspace topology of the cyclic linear subspace. By a cyclic subspace of a matrix A we mean a subspace generated by a single vector v under the action of A (i.e., the subspace spanned by the vectors $V, Av, \dots, A^{n-1}v$).

3. Preliminary results. We require three preliminary results contained in Lemma 1 and Propositions 1 and 2 below for the proof of Theorem 1. Both of these results are refinements of [1].

LEMMA 1. Let $g: R \times R^n \times R^m \rightarrow R^n$ be a function such that $g(t, x, u)$ and $\partial g_i / \partial x_j$, $1 \leq i, j \leq n$, are continuous in (t, x, u) . Suppose there is a constant M such that $-\infty < M$, and the characteristic roots of the matrix

$$(3.1) \quad \frac{1}{2} \left(\frac{\partial g_i}{\partial x_j} + \frac{\partial g_j}{\partial x_i} \right)$$

satisfy

$$V_k \leq M, \quad k = 1, 2, \dots, n, \quad \text{for some } u^*,$$

uniformly in x and $t \geq t_0$. Then the scalar product G defined by

$$G = (g(t; x + h, u^*(t)) - g(t; x, u^*(t)), h)$$

satisfies

$$(3.2) \quad G \leq M \|h\|^2 \quad \text{for all } t \geq t_0.$$

Proof. Define by D the $n \times n$ matrix d_{ij} , where

$$(3.3) \quad d_{ij} = \frac{\partial g_i}{\partial x_j}(t; x + \theta_i h, u(t))$$

and $\theta_i = \theta_i(x, t)$ satisfies $0 < \theta_i < 1$. Then the mean value theorem yields the identity

$$G \equiv (Dh, h) = (D^*h, h)$$

for all x, u, h and t , where D^* is the symmetric $n \times n$ matrix $\frac{1}{2}(d_{ij} + d_{ji})$ and d_{ij} is given in (3.3). Because D^* is symmetric

$$(D^*h, h) \leq M(h, h) = M \|h\|^2,$$

$t \geq t_0$, and this is our lemma.

PROPOSITION 1. Suppose in (1.1) assumption (i) of Theorem 1 holds. Then every solution of (1.1) satisfies

$$(3.4) \quad \|x\|^\rho \leq \left\{ e^{-(1/2)(\rho\delta/2)t} \left[D_1 + D_2 \int_{t_0}^t \|f(\tau; 0, u^*(\tau))\|^\rho e^{(1/2)(\rho\delta/2)\tau} d\tau \right] \right\}$$

for all $t \geq t_0$ where $D_1 = D_1(\delta, t_0, A, x(t_0)) > 0$ and $D_2 = D_2(\delta, A) > 0$ are constants depending only on the arguments shown; α is the largest eigenvalue of A and ρ is any constant in the range $1 \leq \rho \leq 2$.

Proof. Let $x = x(t)$ be any solution of (1.1), and $V(t)$ be a function defined by

$$V(t) = (Ax(t), x(t)),$$

where A is the matrix defined in Theorem 1. Recall that A was required to be symmetric and positive definite, so that

$$(3.5) \quad \alpha \|x(t)\|^2 \geq V(t) \geq \alpha' \|x(t)\|^2$$

for all t , where $\alpha > 0$, $\alpha' > 0$ are the greatest and least eigenvalues of A . Now A symmetric implies that

$$(3.6) \quad \begin{aligned} \frac{1}{2} \frac{dV}{dt} &= (Af(t; x, u^*(t)), x) \\ &= (Af(t; x, u^*) - Af(t; 0, u^*), x) + (Af(t; 0, u^*), x) \\ &\equiv U_1 + U_2. \end{aligned}$$

Utilizing the lemma we deduce that

$$\begin{aligned} U_1 &\leq -\delta \|x\|^2 \quad (t \geq t_0) \\ &\leq -\delta/\alpha V(t) \end{aligned}$$

by (3.5). To estimate U_2 we appeal to Cauchy's inequality to obtain

$$\begin{aligned} |U_2| &\leq \left\{ \sum_{1 \leq i, j \leq n} a_{ij}^2 \right\}^{1/2} \|f(t; 0, u^*(t))\| \|x\| \\ &\leq \left\{ \sum_{1 \leq i, j \leq n} a_{ij}^2 \right\}^{1/2} \|f(t; 0, u^*(t))\| (V/\alpha')^{1/2}, \end{aligned}$$

by (3.5). Now substitute these estimates into (3.6); then

$$(3.7) \quad dV/dt \leq -2\delta_0 V + C_3 \|f(t; 0, u^*(t))\| V^{1/2}$$

for all $t \geq t_0$, where $\delta_0 = \frac{1}{2}\delta/\alpha$ and C_3 with $0 < C_3 < \infty$ is a constant depending only on A . The inequality (3.7) plays a crucial role in what follows.

Now let us recast (3.7) as follows:

$$(3.8) \quad dV/dt + \delta_0 V = -\delta_0 V + C_3 \|f(t; 0, u^*(t))\| V^{1/2}.$$

If ρ is any constant such that $1 \leq \rho \leq 2$, set

$$(3.9) \quad \beta = 1 - \frac{1}{2}\rho;$$

then $0 \leq \beta \leq \frac{1}{2}$ and

$$(3.8') \quad dV/dt + \delta_0 V \leq V^\beta U,$$

where

$$(3.10) \quad U \equiv V^{(1/2-\beta)} \{C_3 \|f(t; 0, u^*(t))\| - \delta_0 V^{1/2}\}.$$

From (3.8') and (3.10), it is almost immediate that

$$(3.11) \quad U \leq C_4 \|f(t; 0, u^*(t))\|^\rho$$

for all t , where $C_4 = C_4(C_3, \delta_0) > 0$ is a constant.

Indeed, if for example

$$C_3 \|f(t; 0, u^*(t))\| \leq \delta_0 V^{1/2},$$

then (3.10) yields $U \leq 0$, which is included in (3.11). On the other hand if

$$C_3 \|f(t; 0, u^*(t))\| > \delta_0 V^{1/2},$$

then (3.10) gives

$$\begin{aligned} U &\leq V^{(1/2-\beta)} C_3 \|f(t; 0, u^*(t))\| \\ &< [\{C_3 \|f(t; 0, u^*(t))\|/\delta_0\}^2]^{(1/2-\beta)} C_3 \|f(t; 0, u^*(t))\| \\ &= C_3^\rho \delta_0^{2\beta-1} \|f(t; 0, u^*)\|^\rho \end{aligned}$$

by (3.9). Set $C_4 \equiv C_3^\rho \delta_0^{2\beta-1}$ to obtain (3.11). Since (3.11) holds for all t , substitute this in (3.8') to obtain

$$(3.12) \quad dV/dt + \delta_0 V \leq C_4 V^\beta \|f(t; 0, u^*)\|^\rho$$

for all $t \geq t_0$.

Next, set $\xi = (1 - \beta)\delta_0$ and multiply the inequality (3.12) by $e^{\xi t}$; rearrange to obtain

$$\frac{d}{dt} \{V^{(1-\beta)} e^{\xi t}\} \leq (1 - \beta) C_4 \|f(t; 0, u^*)\|^\rho e^{\xi t}.$$

On integrating both sides of this on the interval $[t_0, T]$ we deduce

$$\{V(T)\}^{1-\beta} \leq e^{-\xi T} \left\{ e^{\xi t_0} [V(t_0)]^{(1-\beta)} + (1 - \beta) C_4 \int_{t_0}^T \|f(t; 0, u^*(t))\|^\rho e^{\xi t} dt \right\}.$$

When we insert the values

$$1 - \beta = \frac{1}{2}\rho, \quad \xi = \frac{1}{2}\rho\delta/\alpha$$

into this and use (3.5) the result (3.4) is immediate. The proposition is proved.

PROPOSITION 2. *Suppose all the conditions of Proposition 1 hold; and further there are constants $r > 0$ and ρ , $1 \leq \rho \leq 2$, such that*

$$(3.13) \quad \int_t^{t+r} \|f(\tau; 0, u^*(\tau))\|^\rho d\tau \rightarrow 0 \quad \text{as } t \rightarrow \infty.$$

Then every solution of (1.1) satisfies

$$(3.14) \quad x(t) \rightarrow 0 \quad \text{as } t \rightarrow \infty.$$

Proof. First we shall show that every solution $x(t)$ of (1.1) satisfies

$$(3.15) \quad \sup_{t \geq t_0} \|x(t)\| \leq D_0 \sup_{t \geq t_0} \left\{ \int_t^{t+r} \|f(\tau; 0, u^*(\tau))\|^\rho d\tau \right\}^{1/\rho}$$

provided the latter supremum is finite. With (3.15), (3.14) follows from (3.13).

The verification of (3.15) is similar to the technique in [1] except for numerous but minor changes. It is being reproduced here for ease of presentation. Assume the hypothesis of Proposition 1. Suppose $r > 0$ and ρ , $1 \leq \rho \leq 2$, are such that

$$(3.16) \quad \Phi_0 = \sup_{t \geq t_0} \int_t^{t+r} \|f(\tau; 0, u^*(\tau))\|^\rho d\tau < \infty.$$

If $x(t)$ is a solution of (1.1), then by (3.4)

$$(3.17) \quad \|x(t)\|^\rho \leq e^{-\lambda t} D_1 + D_2 \int_{t_0}^t \|f(\tau; 0, u^*(\tau))\|^\rho e^{-\lambda(t-\tau)} d\tau$$

for all $t \geq t_0$, where $\lambda = \frac{1}{2}\rho\delta\alpha^{-1}$. For of all $t \geq t_0$ suppose m is the nonnegative greatest integer not exceeding $(t - t_0)r^{-1}$, that is $m = [(t - t_0)r^{-1}]$. In the integration at the left-hand side of (3.17) split the interval (t_0, t) into the following subintervals:

$$(t_0, t - mr), (t - mr, t - (m - 1)r), \dots, (t - r - 1, t - r), (t - r, t).$$

It is clear that for $t \geq t_0$,

$$\begin{aligned} \int_{t_0}^t \|f(\tau; 0, u^*(\tau))\|^\rho e^{\lambda\tau} d\tau &= \left\{ \int_{t_0}^{t-mr} + \sum_{j=1}^m \int_{t-jr}^{t-(j-1)r} \right\} \|f(\tau; 0, u^*)\|^\rho e^{\lambda\tau} d\tau \\ &\leq \left\{ \sum_{j=0}^m e^{\lambda(t-jr)} \right\} \Phi_0 \end{aligned}$$

by (3.16). Therefore,

$$\int_{t_0}^t \|f(\tau; 0, u^*(\tau))\|^\rho e^{-\lambda(t-\tau)} d\tau \leq \left\{ \sum_{j=0}^m e^{-\lambda jr} \right\} \Phi_0.$$

When we combine this with (3.17) we deduce that

$$(3.18) \quad \|x(t)\|^\rho \leq e^{-\lambda t} D_1 + D_2(1 - e^{-\lambda r})^{-1} \Phi_0, \quad t \geq t_0.$$

Next we observe that

$$(3.19) \quad \sup_{t \geq t_0} \|x(t)\|^\rho \leq 2D_2(1 - e^{-\lambda r})^{-1} \Phi_0.$$

Suppose on the contrary that

$$\sup_{t \geq t_0} \|x(t)\|^\rho > 2D_2(1 - e^{-\lambda r})^{-1} \Phi_0;$$

then there exists a sequence

$$t_0 < t_1 < t_2 < \cdots < t_n < \cdots$$

such that $t_n \rightarrow \infty$ as $n \rightarrow \infty$, and such that

$$\|x(t_n)\|^\rho > 3D_2\{2(1 - e^{-\lambda r})\}^{-1}\Phi_0, \quad n = 1, 2, \dots.$$

But (3.18) would imply that

$$3D_2\{2(1 - e^{-\lambda r})\}^{-1}\Phi_0 < e^{-\lambda t_n}D_1 + D_2(1 - e^{-\lambda r})^{-1}\Phi_0, \quad n = 1, 2, \dots,$$

and on letting $n \rightarrow \infty$,

$$3D_2\{2(1 - e^{-\lambda r})\}^{-1}\Phi_0 \leq D_2(1 - e^{-\lambda r})^{-1}\Phi_0$$

which is a contradiction. The inequality (3.19) is now established. In (3.19) let $\lambda = \frac{1}{2}\rho\delta\alpha^{-1}$; then

$$(1 - e^{-\lambda r})^{-1} \leq \{1 - \exp(-1/2r\delta\alpha^{-1})\}^{-1}$$

for all ρ ; $1 \leq \rho \leq 2$. From (3.19) it is clear that

$$\sup \|x(t)\|^\rho \leq D_3\Phi_0,$$

where $D_3 = D_3(r, \delta, A) > 0$. This verifies (3.15) and thus (3.14).

4. Proofs of theorems. Assume all the conditions of Theorem 1. Then by Proposition 2 every solution $x(t)$ of (1.1) satisfies

$$x(t) \rightarrow x_1 = 0 \quad \text{as } t \rightarrow \infty.$$

The theorem is proved, and as a consequence, there is a finite time T such that every solution $x(t)$ with $x(t_0) = x_0$ has $x(T)$ in an open neighborhood of the origin.

Proof of Theorem 2. The proof of Theorem 2 is an immediate consequence of the main Theorem in [5]: The reachable set of (1.1) contains a neighborhood of 0 whenever the reachable set of (2.4) contains a neighborhood of 0. But this is equivalent to (2.4) being proper [2, p. 78].

Proof of Theorem 3. From Theorem 1 there is a finite time T such that every solution $x(t)$ with $x(t_0) = x_0$ has $x(T)$ in an arbitrary open neighborhood of the origin. Since (1.1) is locally controllable the point $x(T)$ is indeed brought to the origin in finite time.

5. Example. Consider the model of a mass spring system

$$(5.1) \quad \ddot{x} + a\dot{x} + bx = g(u),$$

where

$$a > 0, \quad b > 0, \quad u: R \rightarrow \Omega \subset R \text{ admissible, } g: R \rightarrow R \text{ with}$$

$$(5.2) \quad g(0) = 0, \quad g(-b) < 0 < g(b), \quad \int_0^\infty g(u(\tau)) d\tau < \infty.$$

Let $\mathbf{x} = (x_1, x_2)$ where $x_1 = x$, $x_2 = \dot{x}$, so the mass spring equation becomes

$$(5.3) \quad \dot{\mathbf{x}} = \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} x_2 \\ -bx_1 - ax_2 + g(u) \end{bmatrix} = f(\mathbf{x}, u).$$

Hence

$$L = L(t) = \begin{bmatrix} 0 & 1 \\ b & -a \end{bmatrix}.$$

Then

$$f(0, \Omega) = \{0\} \times g(\Omega).$$

We may write (2.4) as follows:

$$(5.4) \quad \dot{\mathbf{x}} = L\mathbf{x} + BV,$$

where $B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ and $V \in K(g(\Omega)) \subset R$.

We calculate the controllability matrix

$$[B, LB] = \left\{ \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 1 \\ -a \end{bmatrix} \right\}$$

and note that

$$(5.5) \quad \text{rank } [B, LB] = 2.$$

Assume that $\Omega = \{-b, 0, b\}$ for some $b > 0$. Then from (5.2), $K(g(\Omega)) = R$. From (5.5) it follows that (5.4) is proper. Hence Theorem 2 gives that (5.1) is locally controllable.

Let $A = I$ and then $J = \begin{bmatrix} 0 & 1 \\ -b & -a \end{bmatrix}$, and $\frac{1}{2}(J + J^T)$ has its characteristic values strictly negative. From (5.2), (2.2) is satisfied; thus Theorem 1 holds for (5.1). It follows that the system (5.1) is globally stable in finite time.

Acknowledgments. The author is very grateful to Professor O. Hajek and Professor J. A. Yorke for their valuable criticisms of this paper.

REFERENCES

- [1] J. O. C. EZEILO, *An estimate for the solutions of a certain system of differential equations*, Nigerian J. Sci., 1 (1966), pp. 5-10.
- [2] H. HERMES AND J. P. LASALLE, *Functional Analysis and Time Optimal Control*, Academic Press, New York, 1969.
- [3] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
- [4] L. MARKUS, *Controllability of nonlinear processes*, this Journal, 3 (1965), pp. 78-90.
- [5] J. A. YORKE, *The maximum principle and controllability of nonlinear equations*, Tech. Note BN-645, Univ. of Maryland, IFDAM, College Park, 1970.
- [6] L. WEISS, *Lectures on controllability and observability*, Tech. Rep. BM-590, Univ. of Maryland, IFDAM, College Park, 1969.
- [7] L. WEISS AND E. F. INFANTE, *On the stability of systems defined over a finite time interval*, Proc. Nat. Acad. Sci., 54 (1965), pp. 44-48.

- [8] ———, *Finite time stability under perturbing forces and on product spaces*, IEEE Trans. Automat. Control, AC-12 (1967), pp. 54–59.
- [9] R. F. BRAMMER, *Controllability in linear autonomous systems with positive controllers*, this Journal, 10 (1972), pp. 339–353.
- [10] ———, *Differential controllability and the solution of linear inequalities* (Code 816) NASA/Goddard Space Flight Center, Greenbelt, Md., 1973.
- [11] E. N. CHUKWU, *On the reachable sets of non-autonomous linear control systems*, Tech. Rep. CSU Md. 10, Dept. of Mathematics, Cleveland State Univ., Cleveland, Ohio, 1973.

SOME NEW ANALYTIC AND COMPUTATIONAL RESULTS FOR OPERATOR RICCATI EQUATIONS*

J. CASTI† AND L. LJUNG‡

Abstract. The operator Riccati equation associated with a distributed parameter quadratic cost-linear dynamics control process is considered.

Making use of ideas from transport theory, a derivation of a generalized version of the X - Y functions of radiative transfer is given, and it is seen that, under commonly occurring conditions, the new equations may be easier to numerically resolve than the original Riccati equation. In particular, the new equations express directly the optimal gain function. The analytic results are illustrated by a numerical example of heat regulation on a rod.

1. Introduction. One of the foremost topics of contemporary interest in the control theory community is the study of systems governed by functional or partial differential equations—the so-called distributed parameter control problems. These problems arise in a number of diverse areas ranging from the flow of non-Newtonian fluids [1] and problems in plasticity [2] to the study of chemical reactions occurring in the body [3].

Despite their obvious importance in practical physical systems, the systematic mathematical study of distributed parameter problems is of very recent origin, essentially dating from the late 1960's with the work by J. L. Lions and his colleagues [4]–[6]. Since that time, a large number of research articles have appeared detailing various aspects of distributed parameter problems, both from a theoretical and a computational point of view. A representative bibliography of this work may be found in [7].

The literature on distributed parameter problems seems to support the thesis that the major stumbling block between theoretical results and their application to practical problems is a computational one. There appear to be no fundamentally new mathematical concepts necessary to formally transfer most of the standard control-theoretic results such as the Pontryagin maximum principle, Bellman's principle of optimality, Lagrange multipliers, etc., to the infinite-dimensional setting. Of course, several technical hurdles need to be overcome to rigorously justify such a transfer, but they seem to involve work more in the line of attention to details and mathematical virtuosity than upon new foundational results. However, the infinite-dimensionality of the state space of such control processes has made the practical implementation of the theoretical results rather limited since, generally speaking, the equations which must be numerically resolved are finite systems of partial differential equations or infinite systems of ordinary differential equations of boundary value type. Needless to say, this is already a formidable part of numerical analysis, and no routine methods analogous to the integration schemes for ordinary differential equations exist.

* Received by the editors January 24, 1974. This work was supported in part by the National Science Foundation under Grant GK-31420.

† Departments of Mathematics and Systems and Industrial Engineering, University of Arizona, Tucson, Arizona 85721.

‡ Division of Automatic Control, Lund Institute of Technology, Lund, Sweden.

Our objective in this paper is to single out a particular type of problem—the linear regulator problem—and to show how some recent results obtained in another context [8]–[10] can be utilized to effect a significant reduction in the computational effort required to obtain the optimal control. It is well known that all of the essential information about the linear regulator problem is carried in an operator Riccati equation which, in the classical finite-dimensional setting, reduces to a matrix Riccati equation that may be readily handled by standard computational techniques. However, in the infinite-dimensional setting, treatment of the relevant operator Riccati equation requires various approximation and truncation techniques which take its solution outside the bounds of routine computation. We shall show that under very reasonable assumptions, the solution of the appropriate operator Riccati equation may be replaced by the solution of a pair of lower-dimensional operator equations and that the solution to the original Riccati equation may be expressed as an algebraic combination of these lower-dimensional operators. More importantly, it will be seen that the original operator Riccati equation need never have been introduced! It suffices to consider only the new operators which, following terminology introduced in transport theory, we call the “generalized X and Y ” operators.

The paper is divided into six sections. Section 2 briefly reviews the standard linear regulator problem in its infinite-dimensional setting and introduces the Riccati equation formulation for the optimal control law. In § 3 we present our main theorem detailing the new operator equation satisfied by the X - Y operators and its connection with the traditional Riccati equation. A numerical example illustrating the computational efficacy of the new equations is given in § 4. The example involves the classical heat equation, so that we may easily compare our approach with other analytical and computational studies. Following the example, in § 5 we discuss the implications of our results for infinite-dimensional linear filtering and the relevance of our basic assumptions in this context. Finally, § 6 presents several areas which we feel merit further study, such as the infinite interval problem, computational refinements, and so forth.

2. The linear regulator problem. To describe the infinite-dimensional regulator problem, we shall make the following definitions:

DEFINITION 1. $L^2[0, T; E]$ = the set of all functions f defined on $[0, T]$ with values in a Hilbert space E such that $\int_0^T \|f(t)\|^2 dt < \infty$.

DEFINITION 2. $W[0, T] = \{f: f \in L^2[0, T; V], df/dt \in L^2[0, T; V^*]\}$.

We assume that Hilbert spaces V and H are given with V^* the dual of V . Further, let $V \subset H$ and let the injection of V into H be continuous, with V being dense in H . Introduce also two separable Hilbert spaces E and F . Let $\mathcal{L}(H_1, H_2)$ denote the set of linear maps from H_1 to H_2 . Consider the evolution equation

$$(1) \quad \frac{d}{dt}y(t) + Ay(t) = f(t) + Bv(t), \quad z(t) = Cy(t),$$

with initial condition

$$(2) \quad y(0) = y_0, \quad y_0 \in H.$$

Suppose that

$$\begin{aligned} y(t) \in V, \quad \frac{d}{dt}y(t) \in V^*, \quad A \in \mathcal{L}(V, V^*), \quad f(t) \in V^*, \\ v(t) \in E, \quad z(t) \in F, \quad B \in \mathcal{L}(E, V^*), \quad C \in \mathcal{L}(V, F). \end{aligned}$$

All functions are square integrable over the time interval $[0, T]$.

It may be shown [5] that, under certain regularity conditions on A , (1) has a unique solution in $W[0, T]$ which is continuously dependent upon the data of the problem, f , v and y_0 .

With the above existence result at our disposal, we formulate the optimal control problem as that of minimizing the cost function

$$(3) \quad J(v) = \int_0^T \|z(t)\|_F^2 dt + \int_0^T (Nv(t), v(t))_E dt,$$

where $N \in \mathcal{L}(E, E)$ with $(Nv, v) > 0$, and where $z(t)$ satisfies (1) for a given v .

Letting Λ_E, Λ_F denote the canonical isomorphisms of E onto E^* and F onto F^* , respectively, it can be shown that there exists a unique $u \in U$ minimizing J characterized by

$$(4) \quad u(t) = -N^{-1}\Lambda_E^{-1}B^*p(t),$$

where the element $p \in L^2[0, T; V]$ satisfies the boundary value problem

$$(5) \quad \frac{dy(t)}{dt} + Ay(t) = f(t) - BN^{-1}\Lambda_E^{-1}B^*p(t), \quad y(0) = y_0,$$

$$(6) \quad -\frac{dp(t)}{dt} + A^*p(t) = C^*\Lambda_F C y(t), \quad p(T) = 0.$$

Set

$$D_1 = BN^{-1}\Lambda_E^{-1}B^* \quad \text{and} \quad D_2 = C^*\Lambda_F C.$$

Clearly, $D_1 \in \mathcal{L}(V, V^*)$, $D_2 \in \mathcal{L}(V, V^*)$, $D_1 = D_1^*$ and $D_2 = D_2^*$. Then the system (5)–(6) has the form

$$(5') \quad \frac{dy(t)}{dt} + Ay(t) + D_1p(t) = f(t), \quad y(0) = y_0,$$

$$(6') \quad -\frac{dp(t)}{dt} + A^*p(t) - D_2y(t) = 0, \quad p(T) = 0.$$

Since the representation (4) shows that the costate p is of primary importance, we may use the linearity of (5)–(6) to write

$$(7) \quad p(t) = P(t)y(t) + r(t),$$

where P and r satisfy the equations [5]:¹

$$P \in \mathcal{L}(H, H),$$

¹ In [5], the additional assumption $B \in \mathcal{L}(E, H)$ is made. This assumption, however, is said probably not to be necessary.

$$(8) \quad -\frac{dP}{dt} + PA + A^*P + PD_1P = D_2,$$

$$(9) \quad -\frac{dr}{dt} + A^*r + PD_1r = Pf,$$

$$(10) \quad P(T) = 0, \quad r(T) = 0.$$

Remark. The operator equation (8) should be interpreted as $(-dP/dt + PA + A^*P + PD_1P)\eta = D_2\eta$ for all $\eta \in \text{domain of } A$.

Hence, we see that $P(t)$ satisfies a nonlinear partial differential equation of Riccati type which essentially carries all the information necessary to completely resolve the regulator problem.

With separability assumptions on V and H , the Riccati equation (8) may be thought of as a matrix differential equation with a countably infinite number of rows and columns. Of course, to numerically resolve such a problem requires some type of closure or truncation technique. In the next section we shall exhibit new equations which, under certain finiteness assumptions on E and F , insure that the doubly infinite Riccati matrix equation (8) may be represented by an algebraic combination of simpler functions which can be computed by standard algorithms. In fact, it will be shown that, in order to find the optimal control, the Riccati function P may be dispensed with altogether!

In the following section, we shall have occasion to use the Schwartz kernel theorem [11] to represent $P(t)$ in the form

$$(P(t)\phi)(x) = \int_{\Omega} p(x, \xi, t)\phi(\xi) d\xi$$

for all $\phi \in D(\Omega)$, where Ω is the spatial region of our problem and $P(x, \xi, t)$ is the unique distribution on $\Omega_x \times \Omega_{\xi}$ defined by $P(t)$.

Consider the case with zero boundary conditions in (1), and suppose that E and F are finite-dimensional. Let the column vector functions $b(x)$ and $c(x)$ be the kernels of B and C , respectively. Then the kernel $p(x, \xi, t)$ satisfies the equation

$$(11) \quad \begin{aligned} & \frac{\partial p}{\partial t} p(x, \xi, t) + (A_x^* + A_{\xi}^*)p(x, \xi, t) \\ & + \iint p(x, \zeta_1, t)b^T(\zeta_1)N^{-1}b(\zeta_2)p(\zeta_2, \xi, t)d\zeta_1 d\zeta_2 = c^T(x)c(\xi), \end{aligned}$$

with

$$\begin{aligned} p(x, \xi, t) &= p(\xi, x, t), \\ p(x, \xi, t) &= 0 \quad \text{if } \xi \in \Omega, \quad x \in \partial\Omega, \\ p(x, \xi, T) &= 0. \end{aligned}$$

We note that $D(\Omega)$ is the space of infinitely differentiable functions in Ω with compact support in Ω , endowed with the inductive limit topology of Schwartz.

3. New representation of optimal control. In case there is no forcing term $f(t)$ in (1), the optimal input $u(t)$ is obtained as pure state feedback :

$$u(t) = -N^{-1}(t)\Lambda_E^{-1}B^*(t)P(t)y(t).$$

Denote the feedback operator by $K(t)$:

$$u(t) = -K(t)y(t).$$

Using the same technique as in [8]–[10], it will now be shown that we can directly solve for the operator $K(t)$. Thus it is not necessary to first solve the Riccati equation (8). The possibility of extending the results to the infinite-dimensional problems is pointed out by Kailath in [10].

THEOREM. *The optimal feedback operator $K(t)$ is obtained as the solution of²*

$$(12a) \quad \frac{d}{dt}K(t) = -N^{-1}\Lambda_E^{-1}B^*L^*(t)\Lambda_F L(t),$$

$$(12b) \quad \frac{d}{dt}L(t) = L(t)A + L(t)BK(t),$$

where $K(t) \in \mathcal{L}(V^*, E)$, $L(t) \in \mathcal{L}(V^*, F)$ and $K(T) = 0$, $L(T) = C$. The operator $P(t)$ is found as the solution of

$$(12c) \quad P(t)A + A^*P(t) = -K^*(t)K(t) + C^*\Lambda_F C - L^*(t)\Lambda_F L(t).$$

Proof. Since A is time-invariant, it is possible to choose the element η time-invariant [5]. Therefore, differentiation of (8) with respect to t gives

$$(13) \quad (-P_{tt} + A^*P_t + P_tA + P_tBN^{-1}\Lambda_E^{-1}B^*P + PBN^{-1}\Lambda_E^{-1}B^*P_t)\eta = 0.$$

From (8) we also obtain

$$(14) \quad P_t(T)\eta = -C^*\Lambda_F C_\eta.$$

Now consider (13) as a differential equation for P_t with initial condition (14). We will show that the solution to this linear equation can be written $P_t(t) = -L^*(t)\Lambda_F L(t)$, where $L(t) \in \mathcal{L}(H, F)$ and satisfies (12b). Equation (14) is then trivially satisfied. The left member of (13) is

$$\begin{aligned} & (-L_t^*\Lambda_F L - L^*\Lambda_F L_t + A^*L^*\Lambda_F L^* + L^*\Lambda_F LA + L^*\Lambda_F LBK + K^*B^*L^*\Lambda_F L)\eta \\ & = L^*\Lambda_F(-L + LA^* + LBK)\eta + (-L_t + LA + LBK)^*\Lambda_F L\eta, \end{aligned}$$

which is zero according to (12b). Thus, $P_t(t)$ can be written as above, and (12a) now follows from the definition of $K(t)$. Equation (12c) is a rewrite of (8). This concludes the proof.

The operators $K(t)$ and $L(t)$ correspond to the generalized X - and Y -functions discussed in [10] and [9].

A common situation in practice is that the spaces E and F are finite-dimensional. This means that there are only a finite number of observations and control variables. In that case, the decomposition of the theorem considerably

² For interpretation of the operator identities (12a), (12b), (12c), see the remark of previous section.

reduces the complexity of the problem. A similar reduction for the finite-dimensional approximation of the Riccati equation is given in [15] in the special case when A is diagonal. The kernel $k(t, x)$ of the operator $K(t)$ is then an m_1 -dimensional vector, where m_1 is the number of control variables. The kernel $l(t, x)$ of the operator $L(t)$ is in the same way an m_2 -dimensional vector, m_2 being the number of components in $z(t)$. The kernels are the solution of

$$\begin{aligned}\frac{\partial}{\partial t}k(t, x) &= -\left[\int_{\Omega} N^{-1}b(\xi)l(t, \xi) d\xi\right]l^T(t, x), \\ \frac{\partial}{\partial t}l(t, x) &= A^*l(t, x) + k^T(t, x)\left[\int_{\Omega} b(\xi)l(t, \xi) d\xi\right], \\ k(T, x) &= 0, \quad l(T, x) = c(x).\end{aligned}$$

The boundary conditions on $\partial\Omega$ are the same as those of the state equation. The functions $b(x)$ and $c(x)$ are defined from the operators B and C by

$$(Bu(t))(x) = b^T(x)u(t), \quad Cy(t) = \int_{\Omega} c(\xi)y(t, \xi) d\xi.$$

The optimal control is calculated as

$$u(t) = -\int_{\Omega} k(t, x)y(t, x) dx.$$

These equations should be compared with (11) for the kernel $P(t, x, \xi)$. A significant simplification has been achieved by applying the theorem.

If the forcing term $f(t)$ in (1) is not zero, (9) must also be solved for $r(t)$ in order to obtain the optimal control (4) from (7). The operator $P(t)D_1$ can be expressed in terms of $K(t)$, and $P(t)f(t)$ can be found as follows, without solving for $P(t)$. Suppose $f(t) = Dw(t)$, where $D \in \mathcal{L}(G, V^*)$. Then the operator $\tilde{K}(t) = P(t)D$ can be found as the solution of

$$\frac{d}{dt}\tilde{K}(t) = -L^*(t)\Lambda_F L(t)D, \quad \tilde{K}(T) = 0,$$

and the right-hand side of (9) is then expressed as $\tilde{K}(t)w(t)$.

If G is finite-dimensional, say of dimension m_3 , consequently the optimal control problem can be solved from $m_1 + m_2 + m_3$ partial differential equations in one space variable, instead of from the operator Riccati equation (11), which is a partial differential equation in two space variables.

4. A numerical example. The heat equation for a one-dimensional heat diffusion process will be used to show the applicability of the theorem.

Consider a heat rod of length A with diffusion constant κ . Its endpoints are kept at (say) zero temperature. The temperature at distance x from the endpoint at time t is denoted by $y(t, x)$. The control variable $v(t)$ is the heat flow at the midpoint of the rod. Then

$$\frac{\partial}{\partial t}y(t, x) = \kappa \frac{\partial^2}{\partial x^2}y(t, x) + \delta(x - A/2) \cdot v(t)$$

In the terminology of § 2, the spaces are

$$H = \{y(x), 0 \leq x \leq A | y \in L^2[0, A; R], y(0) = 0, y(A) = 0\},$$

$$V = \{y(x), 0 \leq x \leq A | y \in H, y' \in L^2[0, A; R]\} = H_0^1(0, A).$$

Obviously $E = F = R$.

Let the observation be the average temperature over certain intervals of the heat rod:

$$z(t) = \int_0^A y(t, x) c(x) dx.$$

The control variable $v(t)$ shall be chosen to minimize

$$J = \int_0^T (z^2(t) + v^2(t)) dt.$$

According to the theorem, the optimal control $u(t)$ is given by

$$u(t) = - \int_0^A k(t, x) y(t, x) dx,$$

where $k(t, x)$ is the solution of

$$\begin{aligned} \frac{\partial}{\partial t} k(t, x) &= -l(t, A/2)l(t, x), \\ (15) \quad \frac{\partial}{\partial k} l(t, x) &= -\kappa \frac{\partial^2}{\partial x^2} l(t, x) + l(t, A/2)k(t, x), \\ k(t, 0) &= k(t, A) = l(t, 0) = l(t, A) = 0, \\ k(T, x) &= 0, \quad l(T, x) = c(x). \end{aligned}$$

It is an easy task to solve (15) numerically. In the present paper we will not elaborate on various methods. Here the solution of (15) is readily obtained using a simple difference approximation method. The results for some different choices of $c(x)$ are shown in Fig. 1.

5. The filtering problem. Consider the system

$$\begin{aligned} (16) \quad \frac{d}{dt} y(t) + Ay(t) &= Bv(t) + \xi(t), \\ z(t) &= Cy(t) + \eta(t), \end{aligned}$$

with spaces and operators as in § 2. The variables $\xi(t)$ and $\eta(t)$ are noise acting on the state and measurement variables, respectively.

The filtering problem, i.e., obtaining the minimum variance estimate of the state vector $y(t)$ from measurements $z(s)$, has been considered by Bensoussan [12]. He is able to show that a well-posed filtering problem is, in fact, dual to the optimal control problem. The optimal feedback operator corresponds to the

optimal filter gain. Thus the operator Riccati equation is also of fundamental importance in filtering theory, and our theorem applies without change.

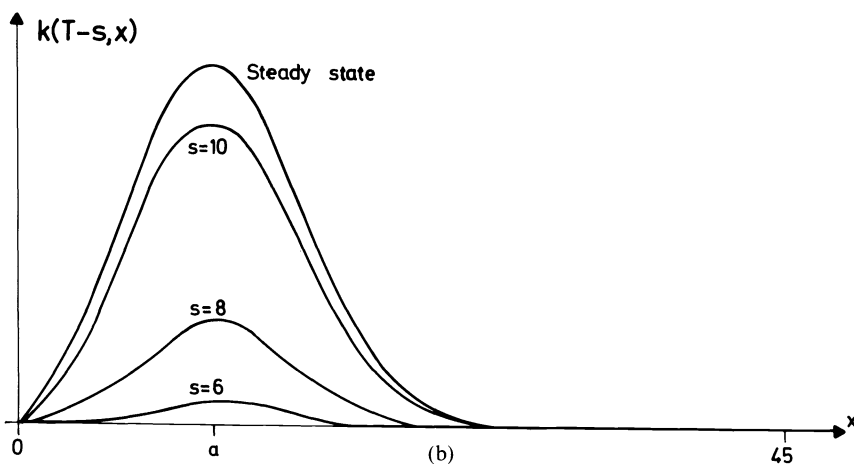
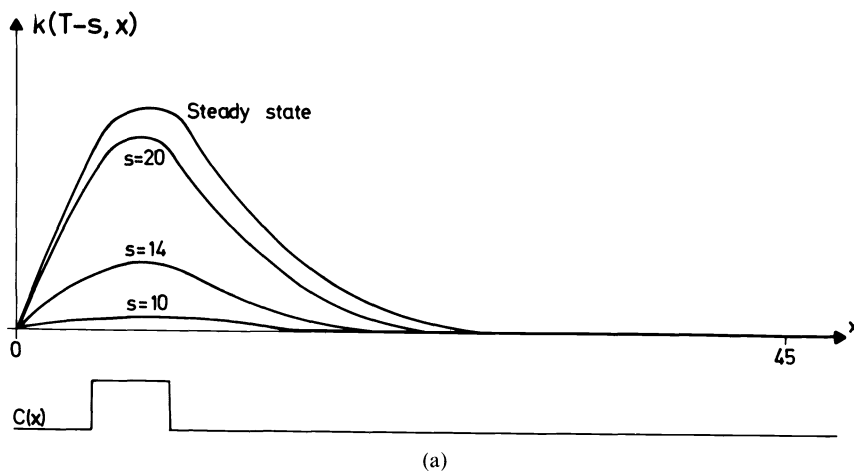


FIG. 1. Optimal feedback kernel $k(t, x)$ for a heat rod of length 45.0 cm and with $\kappa = 1.16 \text{ cm}^2/\text{sec}$. s denotes the time $T - t$. In Fig. 1(a) the observation is the average temperature over the indicated interval. In Fig. 1(b) the observation is the temperature in the point $a = 45/4 \text{ cm}$.

The case E and F being finite-dimensional for the regulator problem, corresponds to the case when the measurement $z(t)$ is a finite-dimensional vector and the system noise $\zeta(t)$ is finitely generated, i.e., $\zeta(t) = D(t)\zeta(t)$, with $\zeta(t)$ finite-dimensional. Consequently, for this case the theorem implies considerable simplification.

It is very reasonable to assume finite-dimensional observations. The case with finitely generated system noise deserves some discussion.

If the state representation has a physical interpretation which we want to retain when filtering, then the question of how the noise is generated must be

answered by physical considerations. In many cases, however, in applying the separation principle, the estimated state vector is used for feedback purposes. Then the state representation is used only as an intermediate stage in computing the input from the output. Without affecting the input-output relation, the system noise $\zeta(t)$ can then be replaced by an equivalent noise $D(t)\zeta(t)$, where $\zeta(t)$ has the same dimension as the output. For the finite-dimensional case see, e.g., [13, Chap. 6].

The following very general situation is treated in [14]. Consider a system with finite-dimensional input $v(t)$ and finite-dimensional output $z(t)$, which are related as follows:

$$z(t) = y(t) + \zeta_2(t), \quad y(t) = \int_0^t B(t-s)v(s) ds + \int_0^t F(t-s)\zeta_1(s) ds,$$

where ζ_i , $i = 1, 2$, is white Gaussian noise.

The input $v(t)$ shall be determined from feedback control based on $y(s)$, $0 \leq s \leq t$, so as to minimize

$$J = \int_0^T E(Qy(t), y(t)) dt + \int_0^T E(v(t), v(t)) dt,$$

where E denotes mathematical expectation.

This formulation covers most practical situations. In [14], it is shown that the solution is essentially obtained from two Riccati equations, one for the filtering problem and one for the deterministic optimal control problem. Both equations have a structure that allows reduction to finite-dimensional expressions as in § 3.

6. Discussion and remarks. The foregoing results raise many additional questions meriting further investigation. In this section, we shall mention just a few of the numerous possibilities that seem promising. Some of these areas have already been treated in the finite-dimensional situation [10], while others are peculiar to the infinite-dimensional case. Extensions to boundary control problems are treated in [16].

Of basic importance in many investigations is the question of steady state behavior, i.e., the limiting behavior of $u(t)$, $y(t)$ as $t \rightarrow \infty$. It is an interesting problem to consider different possibilities to calculate $K = \lim K(t)$ either from the algebraic Riccati equation or from the reduced equations (12).

Another point worth noting is that we have assumed that the terminal cost of the optimization problem is zero. This means that $P(T) = 0$. It does not seem to be possible to generalize the results to completely arbitrary initial conditions. However, if the terminal cost is of the form $(Q_0 z(T), z(T))_F$, i.e., that only the output vector is penalized, then the results can be generalized by letting the operator $L(t)$ belong to $\mathcal{L}(V^*, F \times F)$. Notice also that in the event $\dim F = \infty$, we may still obtain lower dimensional operators if $P(T)$ is chosen properly. This means that by adding an appropriate terminal cost, the case $\dim F = \infty$ may still be handled by the above methods.

Another item worthy of serious future investigation is the problem of computational methods for solving the X - Y equations. In our example, we have

chosen a very simple and classical differencing scheme. However, it is rather clear that the central importance of the X - Y operators to many areas demands that special integration schemes be developed, taking account of the particular type of quadratic nonlinearity exhibited by these equations.

All of these topics are under current study and will be reported in forthcoming papers.

Acknowledgment. The authors wish to thank professor K. J. Åström, who suggested the possibility of reduction for the operator Riccati equation.

REFERENCES

- [1] J. DUVAUT AND J. L. LIONS, *Sur les Inéquations en Mécanique et en Physique*, Dunod, Paris, 1972.
- [2] W. PRAJER AND P. G. HODGE, *Theory of Perfectly Plastic Solids*, John Wiley, New York, 1961.
- [3] J. KERNEVEZ, Thesis, University of Paris, 1972.
- [4] J. L. LIONS, *Some aspects of the optimal control of distributed parameter systems*, SIAM Regional Conf. Series in Appl. Math., no. 6, 1972.
- [5] ———, *The Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.
- [6] ———, *Optimal control of deterministic distributed parameter systems*, IFAC Symp. on the Control of Distributed Parameter Systems, Banff, Canada, 1971.
- [7] S. K. MITTER, *Optimal control of distributed parameter systems*, Control of Distributed Parameter Systems, ASME JACC, 1969.
- [8] J. CASTI, *Matrix Riccati equations, dimensionality reduction and generalized X - Y functions*, Utilitas Mathematica, (1974), to appear.
- [9] ———, *Reduction of dimensionality for systems of linear two-point boundary value problems with constant coefficients*, J. Math. Anal. Appl., 45 (1974), no. 2.
- [10] T. KAILATH, *Some new algorithms for recursive linear estimation in constant linear systems*, IEEE Trans. Information Theory, IT-19 (1973), no. 6.
- [11] L. SCHWARTZ, *Théorie des Noyaux*, Proc. Internat. Congress Mathematicians, 1 (1950), pp. 220–230.
- [12] A. BENSOUSSAN, *Filtrage Optimal des Systèmes Linéaires*, Dunod, Paris, 1971.
- [13] K. J. ÅSTRÖM, *Introduction to Stochastic Control Theory*, Academic Press, New York, 1970.
- [14] A. V. BALAKRISHNAN, *Identification and stochastic control of non-dynamic systems*, Preprints of the 3rd IFAC Symp. on Identification and Parameter Estimation, the Hague, 1973.
- [15] M. J. E. MAYHEW AND A. J. PRITCHARD, *Reduction of the Riccati equation for distributed parameter systems*, Elec. Letters, 7 (1971), no. 20.
- [16] L. LJUNG AND J. CASTI, *Reduction of the operator Riccati equation*, Proc. Internat. Symp. on Control Theory, Numerical Methods and Computer Systems Modelling, Paris, Springer-Verlag, Berlin, 1974.

BILINEAR AND NONLINEAR REALIZATIONS OF INPUT-OUTPUT MAPS*

ARTHUR J. KRENER†

Abstract. Given a nonlinear realization of an input-output map, sufficient conditions are given for the existence of an equivalent bilinear realization for small t . It is also shown that every nonlinear realization can be approximated by a bilinear realization, with an error that grows like an arbitrary power of t .

1. Introduction. In recent years there has been considerable interest in bilinear control systems. This interest can be attributed to the fact that this class of systems is general enough to model many physical and biological processes and at the same time, it is specific enough to support a rich mathematical structure [1], [2], [3], [4]. We would like to propose another reason for considering such systems, namely, that in a sense to be made precise later, every nonlinear system with control entering linearly is locally almost bilinear.

Given an input-output map, $u(t) \mapsto w(t)$, a *bilinear realization* of this is

$$\begin{aligned} \dot{x}(t) &= \left(A_0 + \sum_{i=1}^h u_i(t) A_i \right) x(t), \\ (1.1) \quad w(t) &= Cx(t), \\ x(0) &= x^0, \quad u(t) \in \Omega, \end{aligned}$$

where $x = (x_1, \dots, x_m)$, $w = (w_1, \dots, w_l)$. A_0, \dots, A_h are $m \times m$ matrices, C is an $l \times m$ matrix and $u(t) = (u_1(t), \dots, u_h(t))$ is a measurable control with values in $\Omega = \{u: |u_i| \leq 1, i = 1, \dots, h\}$. The differential equation and map $x \mapsto w$ are called the *dynamics* and the *output map* of the realization and in this case are *bilinear* and *linear* respectively.

A *nonlinear realization* of an input-output map, $u(t) \mapsto z(t)$, is

$$\begin{aligned} \dot{y}(t) &= b_0(y) + \sum_{i=1}^h u_i(t) b_i(y), \\ (1.2) \quad z(t) &= f(y(t)), \\ y(0) &= y^0, \quad u(t) \in \Omega, \end{aligned}$$

where $y = (y_1, \dots, y_n)$, $z = (z_1, \dots, z_l)$, $b_0(y), \dots, b_h(y)$ are C^∞ n -dimensional vector fields, f is a $C^\infty \mathbb{R}^l$ -valued function and $u(t) = (u_1(t), \dots, u_h(t))$ a measurable control with values in $\Omega = \{u: |u_i| \leq 1, i = 1, \dots, h\}$. Here both the *dynamics* and output map, $z = f(y)$, are *nonlinear*.

In Theorem 1, a necessary and sufficient condition is given for there to exist a change of state which bilinearizes the dynamics of (1.2) for small t . As a corollary using a technique of Brockett, we obtain sufficient conditions for the existence of a

* Received by the editors March 7, 1974, and in revised form June 17, 1974.

† Department of Mathematics, University of California-Davis, Davis, California 95616.

bilinear realization, (1.1), such that every input, $u(t)$, gives the same output, $w(t) = z(t)$, for small t .

Theorem 2 shows that for any integer, $\mu \geq 0$, there exists a system with bilinear dynamics which approximates the dynamics of (1.2) with error $O(t^{\mu+1})$. As a corollary there exists for every nonlinear realization, (1.2), a bilinear realization, (1.1), such that every input gives approximately the same output, $w(t) = z(t) + O(t^{\mu+1})$, for small t .

2. Preliminaries. Instead of considering (1.1), it is useful to consider the matrix bilinear system

$$\begin{aligned} \dot{X}(t) &= \left(A_0 + \sum_{i=1}^h u_i A_i \right) X(t), \\ (2.1) \quad W(t) &= C X(t), \\ X(0) &= I, \quad u(t) \in \Omega, \end{aligned}$$

where $X(t)$ takes values in the group, $Gl(m, \mathbb{R})$, of all invertible $m \times m$ matrices.

Each column of the matrix equation (2.1) is a system of the form (1.1). Therefore instead of considering the problems of replacing or approximating (1.2) by (1.1), we study the equivalent problem of replacing or approximating (1.2) by (2.1).

The advantage of considering (2.1) over (1.1) is that $Gl(m, \mathbb{R})$ is a Lie group and each A_j defines a right invariant vector field, $A_j X$, on this group, hence a member of the associated Lie algebra, $gl(m, \mathbb{R})$, of all $m \times m$ real matrices. This algebra is finite-dimensional over the field, \mathbb{R} , and the multiplication is defined by the Lie bracket

$$[A_i, A_j] = A_j A_i - A_i A_j.$$

This is a noncommutative and nonassociative operation which instead satisfies the skew-symmetry and Jacobi relations,

$$[A_i, A_j] = -[A_j, A_i],$$

and

$$[A_i, [A_j, A_k]] = [[A_i, A_j] A_k] + [A_j [A_i, A_k]].$$

For further discussion of Lie groups and algebras we refer the reader to [5], [6], [7].

There is a unique subalgebra, g , of $gl(m, \mathbb{R})$ generated by $\{A_0, \dots, A_h\}$ under bracketing and corresponding to this is a closed Lie subgroup, G , of $Gl(m, \mathbb{R})$. This subgroup is the set of all products of the form

$$\exp(t_{i_1} A_{i_1}) \cdots \exp(t_{i_k} A_{i_k})$$

for all $k \geq 0$ and $t_{i_j} \in \mathbb{R}$, [8]. Another characterization of G is that it is the set of all accessible matrices of

$$\begin{aligned} \dot{X}(t) &= \left(\sum_{i=0}^h u_i(t) A_i \right) X(t), \\ X(0) &= I, \quad |u_i| \leq 1, \quad i = 0, \dots, h. \end{aligned}$$

This follows from the theorem of Chow [9].

The dimension of G as a submanifold of $Gl(m, \mathbb{R})$ is precisely the dimension of the Lie subalgebra, \mathfrak{g} . Furthermore, it has been shown [10], [11] that the set of accessible matrices of (2.1) is a subset of G with nonempty interior in the relative topology of G , hence G is the smallest subgroup of $Gl(m, \mathbb{R})$ containing all accessible matrices of (2.1). For this reason G is said to *carry* (2.1).

The corresponding situation for (1.2) is more complicated because of the nonlinearity. We restrict our discussion of this system to some neighborhood, \mathcal{V} , of y^0 in \mathbb{R}^n . If $b_i(y)$, $b_j(y)$ are C^∞ -vector fields defined on \mathcal{V} , then the Lie bracket, $[b_i, b_j](y)$, is another C^∞ -vector field defined on \mathcal{V} by

$$[b_i, b_j](y) = \frac{\partial b_j}{\partial y}(y)b_i(y) - \frac{\partial b_i}{\partial y}(y)b_j(y).$$

Once again the skew symmetry and Jacobi relations hold.

The set, $V(\mathcal{V})$, of all C^∞ -vector fields on \mathcal{V} becomes a Lie algebra over \mathbb{R} with this definition, however it, in general, is infinite-dimensional. Let $W(\mathcal{V})$ denote the smallest subalgebra of $V(\mathcal{V})$ containing $\{b_0, \dots, b_h\}$. In many cases, but not in general, there is a submanifold \mathcal{N} of \mathcal{V} corresponding to $W(\mathcal{V})$, and containing y^0 . To be more precise, let $W(y)$ be the linear subspace of \mathbb{R}^n formed by evaluating the vector fields of $W(\mathcal{V})$ at y . A submanifold \mathcal{N} of \mathcal{V} is an integral manifold of $W(\mathcal{V})$ if for every $y \in \mathcal{N}$, $W(y)$ is precisely the tangent space to \mathcal{N} at y . We define the rank of $W(\mathcal{V})$ at y to be the dimension of $W(y)$. Then there exists an integral manifold \mathcal{N} of $W(\mathcal{V})$ containing y^0 if the rank of $W(\mathcal{V})$ is constant (Frobenius) [12] or if $b_0(y), \dots, b_h(y)$ are analytic [13]. Other sufficient conditions are found in [12] and [14].

Henceforth we shall assume that \mathcal{N} exists, the dimension of \mathcal{N} is the same as the rank of $W(\mathcal{V})$ at y^0 and by Chow's theorem, is the set of all points in \mathcal{N} accessible from y^0 under the system

$$\begin{aligned} \dot{y}(t) &= \sum_{i=0}^h u_i(t)b_i(y), \\ y(0) &= y^0, \quad |u_i| \leq 1, \quad i = 0, \dots, h. \end{aligned}$$

The set of all points in \mathcal{N} accessible from y^0 by (1.2) is again a subset of \mathcal{N} with nonempty relative interior [10], so \mathcal{N} is said to *locally carry* (1.2).

3. Bilinearization. The problem of replacing a nonlinear realization by a bilinear one can be broken into two parts. The first is: when does there exist a change of state which linearizes the vector fields $b_0(y), \dots, b_h(y)$, resulting in a system with bilinear dynamics and nonlinear output map? The second is: given a realization of this hybrid type, when can it be converted into a bilinear realization?

As for the first question, Guillemin and Sternberg [15] have shown that a family of vector fields, $b_0(y), \dots, b_h(y)$, can be converted to linear vector fields, A_0x, \dots, A_hx , by a change of coordinates, $x = x(y)$, in some neighborhood of y^0 if the vector fields are analytic, all vanish at y^0 and generate a finite-dimensional semisimple Lie algebra. Hermann [16] gave a formal power series construction of the change of coordinates. However, these results are not directly applicable to our questions, since if all the vector fields vanish at y^0 , then the system, (1.2), is trivial.

Asking for a change of coordinates to linearize the vector fields in some neighborhood of y^0 is actually too restrictive for our purposes. Assuming (1.2) is carried locally by \mathcal{N} , what we would like is a system (2.1) carried by G , a neighborhood, \mathcal{M} , of I in G and a differentiable map $\lambda: \mathcal{M} \rightarrow \mathcal{N}$ which *preserves solutions*, that is, $\lambda(X(t)) = y(t)$ for each $u(t)$. The map need not be a local diffeomorphism from \mathcal{M} onto \mathcal{N} , for the dimension of \mathcal{M} could be greater than that of \mathcal{N} ; however, it should be onto since \mathcal{N} carries (1.2). Hartman dealt with a similar question in studying the structural stability of a single vector field about a critical point [21].

If such a λ exists, then its differential, λ_* , is a Lie algebra homomorphism from the Lie algebra, \mathfrak{g} , generated by A_0, \dots, A_h onto the Lie algebra, $W(\mathcal{N})$, generated by b_0, \dots, b_h restricted to \mathcal{N} . Therefore a necessary condition for λ to exist is that $W(\mathcal{N})$ be a finite-dimensional Lie algebra. This also turns out to be sufficient and we have the following theorem.

THEOREM 1. *Suppose that $b_0(y), \dots, b_h(y)$ of (1.2) are analytic and the system is carried locally by \mathcal{N} . There exists a system (2.1) carried locally by \mathcal{M} in $Gl(m, \mathbb{R})$ and an analytic map $\lambda: \mathcal{M} \rightarrow \mathcal{N}$ preserving solutions if and only if the Lie algebra generated by $b_0(y), \dots, b_h(y)$ is finite-dimensional when restricted to \mathcal{N} .*

Proof. Assume $W(\mathcal{N})$ is finite-dimensional. Then by Ado's theorem [17] there exists a Lie subalgebra, \mathfrak{g} , of $gl(m, \mathbb{R})$ for some m and a Lie algebra isomorphism $\varphi: W(\mathcal{N}) \rightarrow \mathfrak{g}$. Define a system with matrix bilinear dynamics, (2.1), by letting $A_i = \varphi(b_i)$. Let e be the evaluation map, $e: W(\mathcal{N}) \rightarrow W(y^0)$, defined by $e(c) = c(y^0)$ for $c \in W(\mathcal{N})$. Then the map $l = e \circ \varphi^{-1}$ satisfies the following

$$l([A_{i_1} \cdots [A_{i_{v-1}}, A_{i_v}] \cdots]) = [b_{i_1} \cdots [b_{i_{v-1}}, b_{i_v}] \cdots](y^0)$$

for any v and $0 \leq i_1, \dots, i_v \leq h$.

It follows from a theorem of the author [18] (generalized by Sussmann [19]) that there exist a neighborhood \mathcal{M} of I and a map $\lambda: \mathcal{M} \rightarrow \mathcal{N}$ preserving solutions. Q.E.D.

Remark 1. In general, the map λ is locally a projection from \mathcal{M} onto \mathcal{N} . However, if the evaluation map, $e: W(\mathcal{N}) \rightarrow W(y^0)$, is a vector space isomorphism, then so is l and the abovementioned theorem implies λ is a local diffeomorphism.

Remark 2. A Lie algebra homomorphism $\varphi: W(\mathcal{N}) \rightarrow gl(m, \mathbb{R})$ is called a *representation* of $W(\mathcal{N})$ and is said to be *faithful* if φ is 1-1, and hence an isomorphism onto its range. If $W(\mathcal{N})$ is of dimension m , then the adjoint representation, $\text{ad}: W(\mathcal{N}) \rightarrow gl(m, \mathbb{R})$, can always be constructed as follows. Choose a basis d_1, \dots, d_m for $W(\mathcal{N})$, and for each $c \in W(\mathcal{N})$ let $\text{ad}(c)$ be the matrix $B = [B_{ij}]$ defined by

$$[c, d_j] = \sum_{i=1}^m B_{ij} d_i.$$

The Jacobi relation implies this is a Lie algebra homomorphism.

The kernel of ad is the *center* of $W(\mathcal{N})$, i.e., the set of all c such that $[c, d] = 0$ for all $d \in W(\mathcal{N})$. If the center is empty, then ad is faithful and this representation can be used in Theorem 1. If $W(\mathcal{N})$ is semisimple, then the center is empty.

Remark 3. If the center is not empty but is contained in the kernel of the evaluation map, e , then the adjoint representation can still be used. In this case, l is constructed by the standard homomorphism theorem as illustrated in Fig. 1.

$$\begin{array}{ccc}
 W(\mathcal{N}) & \xrightarrow{e} & W(y^0) \\
 \text{ad} \downarrow & \nearrow l & \\
 \text{ad}(W(\mathcal{N})) & \subseteq & gl(m, \mathbb{R})
 \end{array}$$

FIG. 1

As for the second step in bilinearization, we have a theorem of Brockett [4] which states that every realization with bilinear dynamics and polynomial output map is equivalent to a realization with bilinear dynamics and linear output map. This results in the following.

COROLLARY 1. *Given: any nonlinear realization (1.2) of the input-output, $u(t) \mapsto z(t)$, satisfying the hypothesis of Theorem 1. If the map $f \circ \lambda: X \rightarrow z$ is a polynomial, then there exists a bilinear realization (1.2) of $u(t) \mapsto w(t)$ and a constant $T > 0$ such that for any input, $u(t)$, the corresponding outputs satisfy $w(t) = z(t)$ for $t \in [0, T]$. (Polynomial here means each component of z is a polynomial in the components of X .)*

4. Approximation of nonlinear systems by bilinear systems. If the Lie algebra, $W(\mathcal{N})$, is not finite-dimensional, then Theorem 1 does not hold; however, we can ask whether (1.2) can be approximated by systems of type (2.1). To be more precise, given (2.1) carried locally by \mathcal{M} and (1.2) carried locally by \mathcal{N} , a C^∞ -map $\lambda: \mathcal{M} \rightarrow \mathcal{N}$ preserves solutions to order μ if there exists a $T > 0$ and $K \geq 0$ such that for any solution, $X(t)$ and $y(t)$, of (2.1) and (1.2) using the same control

$$|\lambda(X(t)) - y(t)| \leq Kt^{\mu+1}$$

for $t \in [0, T]$.

THEOREM 2. *Suppose that $b_0(y), \dots, b_h(y)$ of (1.2) are C^∞ and the system is carried locally by \mathcal{N} . Then for any $\mu \geq 0$ there exists a system (2.1) carried locally by \mathcal{M} in $Gl(m, \mathbb{R})$ and a C^∞ -map $\lambda: \mathcal{M} \rightarrow \mathcal{N}$ preserving solutions to order μ .*

Proof. An abstract Lie algebra, g , is a vector space over \mathbb{R} with a multiplication which satisfies the skew symmetry and Jacobi relations. Suppose a_0, \dots, a_h are elements of g ; then we call $[a_{i_1} \cdots [a_{i_{v-1}}, a_{i_v}] \cdots]$ a *bracket of order v of a_0, \dots, a_h* . One way to construct an abstract Lie algebra, g , is to consider a_0, \dots, a_h to be elements of the algebra and linearly independent over \mathbb{R} . Then treat all the brackets of these up to and including order v as new elements of g which are linearly independent except for those relations implied by the skew symmetry and Jacobi relations. All brackets of order greater than v are taken to be 0. The result is a finite-dimensional Lie algebra which we shall call the *canonical algebra of order v with $h + 1$ generators*.

By Ado's theorem, this algebra is isomorphic to a subalgebra of $gl(m, \mathbb{R})$ which we also denote by g . Under this identification, each a_i becomes a $m \times m$ matrix, A_i , and these are used to construct (2.1). We call the resulting system the *canonical system of order μ with h controls*.

Next we define a linear map $l: g \rightarrow \mathbb{R}^n$ by setting

$$l([A_{i_1} \cdots [A_{i_{v-1}}, A_{i_v}] \cdots]) = [b_{i_1} \cdots [b_{i_{v-1}}, b_{i_v}] \cdots](y^0).$$

It then follows from a theorem of the author [20] that there exists a neighborhood, \mathcal{M} , of I in the subgroup, G , of $Gl(m, \mathbb{R})$ carrying (2.1), a neighborhood, \mathcal{N} , of y^0 in

the submanifold carrying (1.2) and a C^∞ -map $\lambda: \mathcal{M} \rightarrow \mathcal{N}$ which preserves solutions to order μ . Q.E.D.

Remark 1. Once again λ is locally a projection; however, it need not be onto \mathcal{N} unless the brackets of b_1, \dots, b_h up to order μ span the tangent space to \mathcal{N} at y^0 . Of course if l is 1–1, then so is λ .

Remark 2. The adjoint representation of the canonical algebra of order μ is not a faithful representation because the center consists of all brackets of a_0, \dots, a_h of order μ . However, for precisely this reason, the adjoint representation of the canonical algebra of order $\mu + 1$ is isomorphic to the canonical algebra of order μ and can be used to construct (2.1).

Remark 3. Since the canonical algebra is nilpotent, the algebra generated by A_0, \dots, A_h will be nilpotent, and hence any matrix exponential solution of (2.1) for piecewise constant u is a finite series with no terms of order greater than μ .

Remark 4. The dimension of (2.1) is the dimension of the Lie algebra generated by A_0, \dots, A_h and not m or m^2 . The system (2.1) as constructed in the theorem may not be of minimal dimension among bilinear systems which preserve solutions of (1.2) to order μ . A smaller system can be constructed as follows.

Since l is only a linear map, the kernel of l need not be an ideal of g . However, it is a subalgebra of g because l preserves brackets to order μ and all higher order brackets are 0 in g . Let h denote the largest ideal of g contained in the kernel of l . Then the nilpotent Lie algebra g/h can be used in place of g in the construction of the theorem. There may also be Lie algebras of smaller dimension than g/h which need not be nilpotent that can be used to construct (1.1), for example, if (1.2) generates a finite-dimensional Lie algebra.

COROLLARY 2. *Given any nonlinear realization, (1.2), of the input-output map, $u(t) \mapsto z(t)$, and any integer $\mu \geq 0$, there exists a bilinear realization (1.1) of $u(t) \mapsto w(t)$ and constants M and $T > 0$ such that for any input, $u(t)$, the corresponding outputs satisfy*

$$|w(t) - z(t)| \leq Mt^{\mu+1} \quad \text{for } t \in [0, T].$$

Proof. Using Theorem 2, we construct a system with the matrix bilinear dynamics and a map $\lambda: \mathcal{M} \rightarrow \mathcal{N}$ which preserves solutions to order μ . We define a polynomial output map, ψ , for this system by letting ψ be the power series expansion around I of $f \circ \lambda$ up to and including terms of order μ . Using Brockett's technique [4], an equivalent system with bilinear dynamics and linear output map can always be constructed, so all we need show is that our system with bilinear dynamics and polynomial output map approximates (1.2) as required.

By passing to smaller neighborhoods if necessary, we can assume \mathcal{M} and \mathcal{N} are compact; then there exist constants K_1 and K_2 such that

$$|f \circ \lambda(X) - \psi(X)| \leq K_1 |X - I|^{\mu+1} \quad \text{for any } X \in \mathcal{M}$$

$$|f(y^1) - f(y^2)| \leq K_2 |y^1 - y^2| \quad \text{for any } y^1, y^2 \in \mathcal{N}.$$

Let $X(t)$ and $y(t)$ be the solutions of our matrix bilinear system and (1.2) for the same control $|u(t)|$. Then since λ preserves solution to order μ , there exists a

K_3 and $T > 0$ such that for $t \in [0, T]$,

$$|\lambda(X(t)) - y(t)| \leq K_3 t^{\mu+1}.$$

By a standard argument there exists a constant, K_4 , such that for $t \in [0, T]$,

$$|X(t) - I| \leq K_4 t.$$

Putting it all together, we have

$$\begin{aligned} |\psi(X(t)) - f(y(t))| &\leq |\psi(X(t)) - f \circ \lambda(X(t))| + |f \circ \lambda(X(t)) - f(y(t))| \\ &\leq K_1 |X(t) - I|^{\mu+1} + K_2 |\lambda(X(t)) - y(t)| \\ &\leq (K_1 K_4^{\mu+1} + K_2 K_3) t^{\mu+1}. \end{aligned} \quad \text{Q.E.D.}$$

Acknowledgment. I would like to thank Roger Brockett for suggesting this problem to me.

REFERENCES

- [1] R. R. MOHLER, *Bilinear Control Processes with Applications to Engineering, Ecology and Medicine* Academic Press, New York, 1973.
- [2] C. BRUNI, G. DIPILLO AND G. KOCH, *Bilinear systems: An appealing class of "nearly linear" systems in theory and applications*, IEEE Trans. Automatic Control, AC 19 (1974), to appear.
- [3] P. D'ALESSANDRO, A. ISIDORI AND A. RUBERTI, *Realization and structure theory of bilinear dynamical systems*, this Journal, 12 (1974), pp. 517–535.
- [4] R. W. BROCKETT, *On the algebraic structure of bilinear systems*, Theory and Applications of Variable Structure Systems, R. R. Mohler and A. Ruberti, eds., Academic Press, New York, 1972, pp. 153–168.
- [5] J. BELINFANTE AND B. KOLMAN, *A Survey of Lie Groups and Lie Algebras with Applications and Computational Methods*, Society for Industrial and Applied Mathematics, Philadelphia, 1972.
- [6] H. SAMELSON, *Notes on Lie Algebras*, Van Nostrand, New York, 1969.
- [7] J. E. HUMPHREYS, *Introduction to Lie Algebras and Representation Theory*, Springer-Verlag, New York, 1972.
- [8] R. W. BROCKETT, *Systems theory on group manifolds and coset spaces*, this Journal, 10 (1972), pp. 265–284.
- [9] W. L. CHOW, *Über Systeme von linearen partiellen Differentialgleichungen erster Ordnung*, Math. Ann., 117 (1939), pp. 98–105.
- [10] A. J. KRENER, *A generalization of Chow's theorem and the bang-bang theorem to nonlinear control problems*, this Journal, 12 (1974), pp. 43–52.
- [11] H. J. SUSSMANN AND V. J. JURDJEVIC, *Controllability of non-linear systems*, J. Differential Equations, 12 (1972), pp. 95–116.
- [12] R. HERMANN, *On the accessibility problem in control theory*, International Symposium, Nonlinear Differential Equations and Nonlinear Mechanics, Academic Press, New York, 1963, pp. 325–332.
- [13] T. NAGANO, *Linear differential systems with singularities and an application to transitive Lie algebras*, J. Math. Soc. Japan, 18 (1966), pp. 398–404.
- [14] C. LOBRY, *Contrôlabilité des systèmes non linéaires*, this Journal, 8 (1970), pp. 573–605.
- [15] V. W. GUILLEMIN AND S. STERNBERG, *Remarks on a paper of Hermann*, Trans. Amer. Math. Soc., 130 (1968), pp. 110–116.
- [16] R. HERMANN, *The formal linearization of a semisimple Lie algebra of vector fields about a singular point*, Ibid., 130 (1968), pp. 105–109.
- [17] I. D. ADO, *The representation of Lie algebras by matrices*, Uspehi Mat. Nauk (22), 3 (1947), no. 6, pp. 159–173; English transl., Amer. Math. Soc. Transl. no. 2, 1949.

- [18] A. J. KRENER, *On the equivalence of control systems and the linearization of nonlinear systems*, this Journal, 11 (1973), pp. 670–676.
- [19] H. J. SUSSMANN, *An extension of a theorem of Nagano on transitive Lie algebras*, Proc. Amer. Math. Soc., to appear.
- [20] A. J. KRENER, *Local approximation of control systems*, J. Differential Equations, to appear.
- [21] P. HARTMAN, *A lemma in the structural stability of differential equations*, Proc. Amer. Math. Soc., 11 (1960), pp. 610–620.

DIFFERENTIAL GAMES AND A NASH EQUILIBRIUM SEARCHING ALGORITHM*

L. F. PAU†

Abstract. Nonlinear nonzero sum differential games with fixed playing time are considered. They are transformed into a two-level hierarchical game, with a suprenal referee optimizing an overall objective function in order to improve the respect of the Nash playing rule. We propose an algorithm, based on constraint coordination, in order to approximate open-loop Nash equilibrium controls. The theory illustrating this algorithm is presented.

Introduction. In the present paper, we are concerned with N -player nonzero sum differential games of fixed duration, in which mutual conflicting interests will play a part as in real life problems; these games have been considered in relation to optimal control theory, i.e., by Ho and Starr [14], Berkovitz [5], [7], Blaquière, Gérard and Leitmann [8]. No coalitions are allowed here, and the constraints on the control vectors of the N players are assumed to be decentralized; if this is not the case, Bensoussan [3], [4] explains how to use decomposition in order to be in the former case.

The type of equilibrium controls studied here are conflicting Nash–Cournot equilibriums, as investigated for example, by Case [10], Friedman [11], Telser [25]. In few words, the Nash control is the best control for each player if it is assumed that all the other players are holding firm to their own Nash controls. The Nash solution is therefore safe against any unilateral attempts to optimize by individual players.

Considerable attention has always been given to the theoretical study of differential games, but very little to the development of computational techniques to solve such problems. As far as the author is aware even less literature may be found about the computation of Nash equilibriums in nonlinear nonzero sum differential games. All known references investigate open loop controls based upon necessary Nash equilibrium conditions.

Simaan and Cruz [23] have improved the treatment of the Riccati equations related to the classic linear-quadratic game. Quintana and Davison [22] have investigated the Newton–Raphson method and a gradient optimization of the Hamiltonian functions in a pursuit-evasion game with quadratic criterion. Nonlinear dynamics are introduced by Gourishankar and Salama [12] in a specific two-player differential game of the minimax type. Holt and Mukundan [15] describe an application of the “ping-pong” algorithm proposed by Advani and Mukundan [2] to produce Nash solutions; it is based upon a perturbation of necessary Nash equilibrium conditions, and on the computation of the optimal controls for the individual players; a similar approach was suggested by Starr [24].

This report concentrates on some theoretical aspects of a Nash equilibrium

* Received by the editors September 7, 1973, and in revised form February, 25, 1974.

† IMSOR, Technical University of Denmark, Lundtoftevej, Denmark. Now at Laboratoire d'Automatique, ENS des Télécommunications, 46 rue Barrault, F 75634 PARIS Cedex 13, France, and Laboratoire d'Automatique théorique, Tour 14-24(v), Université Paris VII, 2 Place Jussieux, F 75005 Paris 5, France.

searching algorithm (A). The equilibria are to be found within the functional space F of piecewise continuous control functions on the fixed time-horizon $(0, T)$. It should be noticed that these Nash equilibrium controls in F are open-loop controls, which do not depend on the state of the game. The differential game is namely transformed into a hierarchical two-level game (see definitions in Jumarie [17]), with a suprenal referee administrating the Nash–Cournot playing rule. The coordination strategy maximizing necessarily the satisfaction of this rule is then computed, using constraint coordination with the optimal control problems of the individual players. These problems are solved by means of the generalized reduced gradient algorithm. The computation of open loop Nash equilibrium controls is certainly quite complex, and in general it is surprising that one can arrive at a solution at all in F . Such a solution would usually only be expected if the state equations were separable and the criteria were concave (Varaiya [26]). However, the (A) algorithm has successfully been applied to finding Nash equilibrium controls in a large dynamic macroeconomic model (Pau [20]).

Our problem is stated formally in §1, before explaining more precisely the principles of the (A) algorithm in §2. A necessary Nash equilibrium condition is stated in §3 in relation to an apparent overall objective function for the referee. The coordination process and the decomposition of the optimal control problem for the referee, are described in §4 and §5 respectively. The (A) algorithm is given in §6, and §7 contains some elements about the numerical application [20].

1. Statement of the problem.¹ In this section, we describe the notation and the formulation which will be used throughout this paper.

Let there be an N -person nonzero sum differential game, with players $j = 1, N$, having the state evolution equation (1.1) on the fixed time horizon $(0, T)$, and with a given initial state $Z(0)$:

$$\begin{aligned}
 dZ(t)/dt &= F(Z(t), U(\cdot), t) \triangleq \{g_1(U(\cdot), Z(0), t), \\
 &\quad \dots, g_N(U(\cdot), Z(0), t); f_1(X(t), U(\cdot), t), \\
 &\quad \dots, f_n(X(t), U(\cdot), t)\}, \\
 (1.1) \quad Z(t) &\triangleq (X_0(t), X(t)) \in E^{N+n}, \\
 X_0(t) &\triangleq (x_{01}(t), \dots, x_{0N}(t)) \in E^N, \\
 X(t) &\triangleq (x_1(t), \dots, x_n(t)) \in \chi(t), \\
 u_j(\cdot) &\in \mu_j(\cdot) \subset F_j, \quad j = 1, N, \quad F \triangleq F_1 \times \dots \times F_N, \\
 U(\cdot) &= (u_1(\cdot), \dots, u_N(\cdot)) \in F,
 \end{aligned}$$

where:

E^k : Euclidean real space of dimension k , $E \triangleq E^1$;

F_j : reflexive real functional Banach space on $(0, T)$, which has the pre-Hilbertian scalar product represented by $\langle \cdot, \cdot \rangle_j: F_j \times F_j \rightarrow E^1$.

The corresponding norm $\|\cdot\|_j$ is supposed to be continuous on F_j ; let $\|\cdot\|$ be the continuous norm in F induced by all norms $\|\cdot\|_j$, $j = 1, N$, and defining the topology in F ;

$\chi(t)$: given compact subset of E^n , for any $t \in (0, T)$, such that the complementary set of $\chi(t)$ in E^n is connex;

¹ Throughout, the notations $j = 1, N$, $j \in (1, N)$ and $j = 1, 2, \dots, N$ are equivalent.

$X(t)$: is the state vector in E^n , for which we specify the state constraints $\chi(t)$ for all $t \in (0, T)$; the initial state $X(0) \in \chi(0)$ is given; the target set is $\chi(T)$;

$Z(0)$: given $(N + n)$ -dimensional vector, such that $X(0) \in \chi(0)$;

$\mu_j(\cdot)$: given closed compact, bounded subset of F_j , such that the complementary set of $\mu_j(\cdot)$ in F_j is connex, for any $j = 1, \dots, N$;

$u_j(\cdot)$: is the control functional of the player j , $j = 1, N$, defined at all instants t within the playing time: $u_j(\cdot): (0, T) \rightarrow E$; $u_j(\cdot)$ is assumed to be piecewise continuous on $(0, T)$, and $u_j(t)$ is its value at time $t \in (0, T)$; the control constraints are obtained by specifying $u_j(t) \in \mu_j(t)$ for all $t \in (0, T)$,

$$u_j(\cdot) \in \mu_j(\cdot) \Leftrightarrow \forall t \in (0, T) |r_j(u_j(t), t) \leq 0,$$

where $r_j: E \times (0, T) \rightarrow E^-$ is a scalar continuous and differentiable function; f_k : are for all $k \in (1, n)$ continuously differentiable functions: $\chi(\cdot) \times \mu_1(\cdot) \times \dots \times \mu_N(\cdot) \times (0, T) \rightarrow E$; these functions therefore fulfill the Lipschitz condition; for given $Z(0)$, and with the assumptions of this section, the state $X(t)$ computed from (1.1) is thus unique;

g_j : are for all $j = 1, N$ two-times continuous and differentiable functions for all $U(\cdot) \in \mu_1(\cdot) \times \dots \times \mu_N(\cdot)$, $X(0) \in \chi(0)$ and $t \in (0, T)$; it is moreover important to stress that g_j will usually depend explicitly on the state $X(t)$, $t \in (0, T)$; since the state is unique for given $X(0)$ and $U(\cdot)$, we have preferred to abbreviate the notation for g_j .

Accordingly, the functions f_k , $k = 1, n$, and g_j , $j = 1, N$, will be respectively once and twice continuously Fréchet differentiable with respect to the control $U(\cdot)$ in the set $\mu_1(\cdot) \times \dots \times \mu_N(\cdot)$, for the topology on $F_1 \times \dots \times F_N$. These properties are assumed to hold uniformly with respect to $t \in (0, T)$ and $X(0) \in \chi(0)$.

The N players want to minimize with respect to $U(\cdot)$ their own terminal costs $x_{0j}(T)$, $j = 1, N$, at time T , rewritten as

$$\begin{aligned} x_{0j}(T) &\triangleq x_{0j}(U(\cdot), Z(0), T), \quad j = 1, N, \\ (1.2) \quad x_{0j}(U(\cdot), Z(0), t) &\triangleq x_{0j}(0) + \int_0^t g_j(U(\cdot), Z(0), \tau) d\tau, \\ t \in (0, T), u_j(\cdot) &: \text{specific control of player } j \text{ on } (0, T). \end{aligned}$$

The terminal costs $x_{0j}(U(\cdot), Z(0), T)$ are moreover twice continuously Fréchet differentiable with respect to $U(\cdot) \in \mu_1(\cdot) \times \dots \times \mu_N(\cdot)$, according to our previous assumptions. It should be recalled that the functions g_j , $j = 1, N$, may depend explicitly on the state $X(t)$, $t \in (0, T)$.

DEFINITION 1.1. The control $U(\cdot) \in F$ is *playable* at $X(0)$ iff $U(\cdot) \in \mu_1(\cdot) \times \dots \times \mu_N(\cdot)$ and if $X(t)$ computed by (1.1) for given $X(0)$ satisfies $X(t) \in \chi(t)$ for all $t \in (0, T)$.

DEFINITION 1.2. The control $U^*(\cdot) \in F$ is an *open loop Nash optimal control* at $X(0)$, iff $U^*(\cdot)$ is playable at $X(0)$ and if it satisfies the Nash equilibrium conditions for the terminal costs:

$$(1.3) \quad \forall l \in (1, N) \begin{cases} \forall u_l(\cdot) \in \mu_l(\cdot), \\ U_l^0 \triangleq (u_1^*, \dots, u_{l-1}^*, u_l, u_{l+1}^*, \dots, u_N^*), \\ x_{0l}(U^*(\cdot), Z(0), T) \leq x_{0l}(U_l^0(\cdot), Z(0), T). \end{cases}$$

2. Principles of the (A) algorithm. In the interest of pedagogical clarity, this section presents the proposed algorithm in idealized form. Technical matters pertaining to assumptions on the functions to be considered, and to the decomposition-coordination scheme used, are all postponed to the following sections.

Our basic idea is that our present differential game among N players with imperfect information, may be viewed as a two-level system (S) in which (see Fig. 1):

(a) each infimal subsystem $j = 1, N$ is a player having to minimize his terminal cost $x_{0j}(U(\cdot), Z(0), T)$;

(b) the single supramal subsystem is a fictive or real referee ($j = 0$) for which the criterion function g expresses both the Nash rules accepted by the N infimal players, and an overall cost minimization; g though, is, only related to local necessary conditions for having a Nash equilibrium.

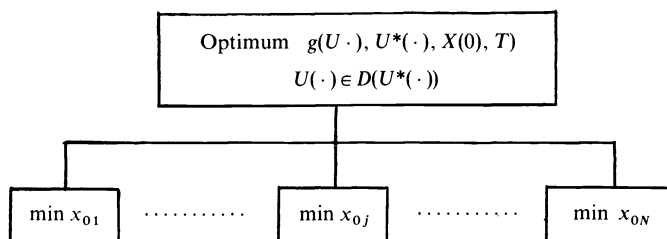


FIG. 1. The hierarchical system (S)

The system (S) as a whole is hierarchical in the sense that the constraint coordination performed by the $N + 1$ players must first aim at satisfying player 0's interest, before taking care of the N others.

If there exists a Nash optimal control $U^*(\cdot) \in F$, the algorithm (A) performs a direct search where the referee carries the coordination role, and the N infimal players perform their individual constrained optimizations.

The coordination is carried out iteratively by a coordination function $a(\cdot) \triangleq (a_1(\cdot), \dots, a_N(\cdot)) \in F$ representing the best current approximation of $a(\cdot) = U^*(\cdot)$. Each player $j, j = 1, N$, responds by giving some local information concerning the decisions of all remaining $N - 1$ players that player j would regard as being optimal for himself. The referee will then try to revise the current approximation of $U^*(\cdot)$ by minimizing the overall cost g with respect to $a(\cdot)$ chosen close to the decisions previously requested by the infimal players. The improvement of the overall cost g will express a better agreement with the Nash playing rule. Notice that each time a player $j, j = 1, N$, is invited to give his opinion, the referee obliges him to play $a_j(\cdot)$ imposed from the supramal level.

In the above procedure, the minimization of g is achieved by reducing the slack in the only nondecentralized constraint resulting from the decomposition of the Lagrangian function L for the referee.

It is clear that a similar approach with coordination and hierarchy may be used in order to find other types of equilibrium controls in nonzero sum differential games, especially in the case of conflicting interests among the N players (see i.e., Friedman [11], Telser [25]).

As already mentioned, this algorithm (A) allows for incomplete information as it may be understood from the former description:

(a) since the (A) algorithm performs a constraint coordination, the individual infimal players $j = 1, N$ are not required to know about one another's terminal cost functionals (about this topic, see Harsanyi [13]);

(b) the infimal players $j = 1, N$ must be kept informed of the constraints on all controls $U(\cdot)$ and states $X(\cdot)$ in the system (S). They must also know the evolution equation governing the state vector $X(\cdot)$, and the initial state $X(0)$;

(c) the infimal players are assumed to be able to build the control functions of the other players into their own terminal cost functionals;

(d) the referee is the only player to require a full knowledge of the whole system (S), including all terminal cost functionals; he may either be fictive, or represented by one of the infimal players, or a third body: in [20] it was the State.

3. Necessary Nash equilibrium conditions. Candidates for Nash optimal controls are usually found by necessary conditions: e.g., see Case [10], Starr and Ho [14], Holt and Mukundan [15]; these conditions are usually deduced from (1.3) and related to the Hamiltonian functions of the N respective players.

The idea used here is to try to find an interlevel performance functional aggregating the infimal costs $x_{0j}(T)$ of the N infimal players $j = 1, N$, together with the features of a Nash equilibrium. Using the language of Mesarovic [19, p. 121], the problem is then to find a control $U^*(\cdot)$ for the overall system (S), which is overall optimal for the system in the sense that it satisfies a necessary Nash equilibrium condition when the interlevel performance functional is minimum.

Usually, the Nash optimal control $U^*(\cdot)$ is not known, and the N infimal players only observe an actual nonequilibrium control $V(\cdot)$ for which the system is not Nash-balanced. One may then introduce an apparent overall objective function g for the system (S), which represents the interlevel performance as it appears to the infimal players for the actual nonbalanced coordination control $V(\cdot)$.

DEFINITION 3.1. Let $U(\cdot)$ and $V(\cdot)$ be playable controls from F at $X(0)$. We then define the *apparent overall objective function* g as being

$$g: (\mu_1(\cdot) \times \cdots \times \mu_N(\cdot))^2 \times \chi(0) \times (0, T) \rightarrow E,$$

$$(3.1) \quad g(U(\cdot), V(\cdot), X(0), t) \triangleq \prod_{l=1, N} \left\langle \frac{\partial x_{0l}(V_l^0(\cdot), Z(0), t)}{\partial u_l(\cdot)}, u_l(\cdot) - v_l(\cdot) \right\rangle_l,$$

$$V \triangleq (v_1, \dots, v_N), \quad V_l^0 \triangleq (v_1, \dots, v_{l-1}, u_l, v_{l+1}, \dots, v_N),$$

where the arguments indicate the place where the partial derivative mapping is computed in the sense of Fréchet.

We may point out already that for $t = T$, the above g function fulfills for given $U(\cdot)$ and $Z(0)$ neither the monotonicity assumption [19, pp. 121–123] nor the additivity with respect to the infimal criterion functions. Usually the intralevel harmony [19, p. 124] will not be satisfied either, because it is not reasonable to assume that the playable controls $V(\cdot)$ close to a Nash optimal control are able to increase simultaneously all infimal terminal costs. The absence of conflict is namely illustrated by the monotonicity property and unrestricted harmony properties.

We will now see how the Nash equilibrium coordination is related to the local extremality, in a certain sense, of the apparent overall objective function g .

DEFINITION 3.2. Given a function $w : s \geq 0 \rightarrow w(s) \in E$, we say that $w(s)$ is an infinitesimal quantity of rank $p > 0$ with respect to s , iff

$$\begin{aligned} \exists M > 0, \quad \exists \eta > 0 : \forall s, 0 \leq s \leq \eta, |w(s)| \leq M \cdot s^p, \\ \lim_{s \rightarrow 0+} w(s) = 0, \quad \lim_{s \rightarrow 0+} (w(s)/s^p) = 0. \end{aligned}$$

LEMMA 3.1. Let $U^*(\cdot) \in F$ be a Nash optimal control at $X(0)$. Assume that $N \geq 2$, and that the following property holds for all $l = 1, N$:

$$(3.2) \quad \begin{aligned} &\forall \varepsilon_l > 0 \quad \exists b_l(\varepsilon_l) > 0: \\ &\forall U(\cdot) \quad \text{such that } \|U_l^0(\cdot) - U^*(\cdot)\| < \varepsilon_l, \end{aligned}$$

then,

$$(x_{0l}(U(\cdot), Z(0), T) - x_{0l}(U^*(\cdot), Z(0), T)) \geq b_l(\varepsilon_l) \Omega_{1l} \geq 0,$$

where: (i) $U_l^0(\cdot)$ is defined as in Definition 1.2;

(ii) $\Omega_{1l} \triangleq \Omega_{1l}(\|U(\cdot) - U^*(\cdot)\|)$ is an infinitesimal quantity of rank strictly less than 2 with respect to $\|U(\cdot) - U^*(\cdot)\|$, and $\Omega_{1l} \geq 0$.

Then there exists a neighborhood $D(U^*(\cdot))$ of $U^*(\cdot)$, $D(U^*(\cdot)) \subset F$, $D(U^*(\cdot)) \subset \mu_1(\cdot) \times \cdots \times \mu_N(\cdot)$, so that $g(\cdot, U^*(\cdot), X(0), T)$ is extremal at $U^*(\cdot)$ (see Fig. 2):

$$(3.3) \quad \begin{aligned} &\text{(i) } \forall U(\cdot) \in D(U^*(\cdot)): g(U(\cdot), U^*(\cdot), X(0), T) \geq 0, \\ &\text{(ii) } g(U^*(\cdot), U^*(\cdot), X(0), T) = 0, \\ &\text{(iii) } \forall l \in (1, N), \partial g(U^*(\cdot), U^*(\cdot), X(0), T) / \partial u_l(\cdot) = 0. \end{aligned}$$

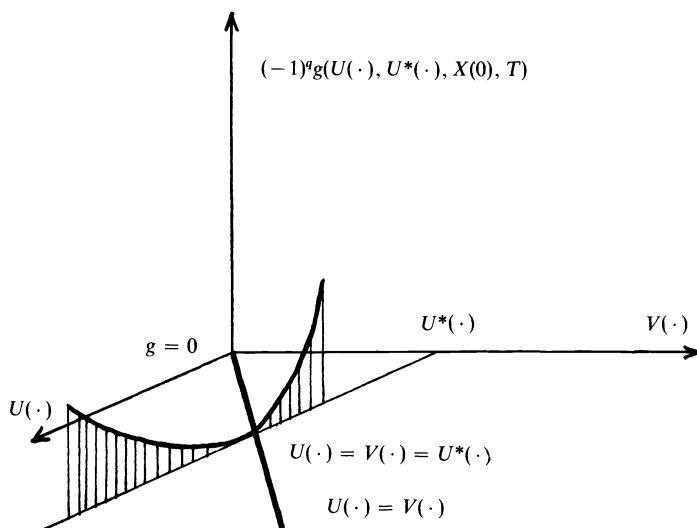


FIG. 2. Section of the g -surface by the "plane" $V(\cdot) = U^*(\cdot)$

Proof. (i) According to our assumptions from §1, we may write limited developments of $x_{0l}(\cdot, Z(0), T)$ for all $l = 1, N$ in a neighborhood of the playable control $U(\cdot)$:

$$(3.4) \quad \begin{aligned} & x_{0l}(U(\cdot), Z(0), T) - x_{0l}(U^*(\cdot), Z(0), T) \\ &= \sum_{j=1, N} \left\langle \frac{\partial x_{0l}(U(\cdot), Z(0), T)}{\partial u_j(\cdot)}, u_j(\cdot) - u_j^*(\cdot) \right\rangle_j + \Omega_{2l}, \end{aligned}$$

where $\Omega_{2l} \triangleq \Omega_{2l}(\|U(\cdot) - U^*(\cdot)\|)$ is an infinitesimal quantity having a rank larger or equal to 2 with respect to $\|U(\cdot) - U^*(\cdot)\|$.

Using the assumption (3.2) of Lemma 3.1, define $\bar{\varepsilon}_l > 0$ as the largest positive number fulfilling

$$(3.5) \quad \forall U(\cdot), \|U_l^0(\cdot) - U^*(\cdot)\| < \bar{\varepsilon}_l: \exists b_l(\bar{\varepsilon}_l) > 0$$

so that

$$\Omega_{2l} \leq b_l(\bar{\varepsilon}_l) \Omega_{1l}, \quad \Omega_{1l} \geq 0.$$

$b_l(\bar{\varepsilon}_l) > 0$ exists because of the difference in the ranks of the two infinitesimal quantities Ω_{1l}, Ω_{2l} .

The relations (3.2), (3.4), (3.5) then result in

$$(3.6) \quad \forall U(\cdot), \|U_l^0(\cdot) - U^*(\cdot)\| < \bar{\varepsilon}_l,$$

then

$$\left\langle \frac{\partial x_{0l}(U(\cdot), Z(0), T)}{\partial u_l(\cdot)}, u_l(\cdot) - u_l^*(\cdot) \right\rangle_l \geq 0.$$

Define then the parallelotope $D(U^*(\cdot))$ as

$$(3.7) \quad D(U^*(\cdot)) \triangleq \{U(\cdot) | \forall l \in (1, N) \|U_l^0(\cdot) - U^*(\cdot)\| < \bar{\varepsilon}_l \text{ and } U(\cdot) \text{ playable}\}.$$

Writing the product of N positive scalar products (3.6) we finally get

$$(3.8) \quad \forall U(\cdot) \in D(U^*(\cdot)): g(U(\cdot), U^*(\cdot), X(0), T) \geq 0.$$

(ii) This relation is evident.

(iii) $g(\cdot, U^*(\cdot), X(0), T)$ is in turn once continuously Fréchet differentiable; let us compute the linear mappings which are the partial derivatives of $g(\cdot, U^*(\cdot), X(0), T)$:

$$(3.9) \quad \begin{aligned} \frac{\partial g(U(\cdot), U^*(\cdot), X(0), T)}{\partial u_j(\cdot)} &= \left(\prod_{\substack{l=1, N \\ l \neq j}} \left\langle \frac{\partial x_{0l}(U_l^0(\cdot), Z(0), T)}{\partial u_l(\cdot)}, u_l(\cdot) - u_l^*(\cdot) \right\rangle_l \right) \\ &\cdot \left(\frac{\partial x_{0j}(U_j^0(\cdot), Z(0), T)}{\partial u_j(\cdot)} \right. \\ &\quad \left. + \left\langle \frac{\partial^2 x_{0j}(U_j^0(\cdot), Z(0), T)}{\partial^2 u_j(\cdot)}, u_j(\cdot) - u_j^*(\cdot) \right\rangle_j \right), \\ U_j^0 &\triangleq (u_1^*, \dots, u_{j-1}^*, u_j, u_{j+1}^*, \dots, u_N^*). \end{aligned}$$

If $N \geq 2$, $(\partial g(U(\cdot), U^*(\cdot), X(0), T)/\partial u_j(\cdot))$ is zero for $U = U^*$, simply because the first scalar term in the relation (3.9) will be zero. This remark completes the proof of Lemma (3.1):

$$(3.10) \quad \text{if } N \geq 2, \forall j \in (1, N): \partial g(U^*(\cdot), U^*(\cdot), X(0), T)/\partial u_j(\cdot) = 0.$$

COROLLARY 3.1. *We assume that the requirements of Lemma 3.1 are fulfilled; we moreover assume that the terminal costs $x_{0j}(U(\cdot), Z(0), T)$, $j = 1, N$, of the N players are quasi-convex functions of $U(\cdot)$ in the neighborhood $D(U^*(\cdot))$ found in Lemma 3.1; the conclusions of this Lemma 3.1 may then be completed by*

(iv)

$$(3.11) \quad \forall l \in (1, N) \quad \text{and} \quad \forall U(\cdot) \in D(U^*(\cdot)),$$

$$\left\langle \frac{\partial g(U(\cdot), U^*(\cdot), X(0), T)}{\partial u_j(\cdot)}, u_j(\cdot) - u_j^*(\cdot) \right\rangle_j \geq 0.$$

Proof. Because of the convexity of the terminal cost functions with respect to a playable control $U(\cdot)$ in $D(U^*(\cdot))$,

$$(3.12) \quad \forall U(\cdot) \in D(U^*(\cdot)) \quad \text{and} \quad \forall j \in (1, N),$$

$$\left\langle \frac{\partial^2 x_{0j}(U_j^0(\cdot), Z(0), T)}{\partial^2 u_j(\cdot)}, u_j(\cdot) - u_j^*(\cdot) \right\rangle_j \geq 0.$$

According to (3.6) we then get (3.11); all scalar coefficients of the linear mappings in (3.9) are namely nonnegative for $N \geq 2$; besides, the scalar products by $(u_j(\cdot) - u_j^*(\cdot))$ of the two linear mappings appearing as such in (3.9), are non-negative following (3.6) and (3.12) if $U(\cdot) \in D(U^*(\cdot))$.

Remark 3.1. It should be noted that $g(U(\cdot), V(\cdot), X(0), T)$ is separable with respect to the functions $u_l(\cdot)$, $l = 1, N$, coordinates of $U(\cdot)$. We could also have expressed g as a sum of scalar products, instead of the product form used here: Lemma 3.1 would then not have been valid any more (because of (iii)). Besides, the g function used here is an infinitesimal quantity having a rank at least N times larger than the infinitesimal quantities $|u_l(\cdot) - u_l^*(\cdot)|$; the practical consequence is a noticeable acceleration in the convergence of the (A) algorithm (§6).

Remark 3.2. The necessary extremality of $g(U(\cdot), U^*(\cdot), X(0), T)$ as stated in Lemma 3.1 is naturally only a local necessary condition for having a Nash optimal control $U^*(\cdot)$. This is due to the use of partial derivatives of the terminal costs.

Remark 3.3. Lemma 3.1 is not concerned about the variations of $g(U(\cdot), V(\cdot), X(0), T)$ with $V(\cdot)$. These variations are most important, since the (A) algorithm described later replaces $V(\cdot)$ with the best current approximation of $U^*(\cdot)$; (A) is attempting to use the uniform continuity of $g(U(\cdot), V(\cdot), X(0), T)$ with respect to $V(\cdot) \in D(U^*(\cdot))$ when optimizing g for $U(\cdot)$. This requires usually that $U(\cdot)$ can only be a perturbed value of $V(\cdot)$, and close to it.

Remark 3.4. It may happen that the assumption (3.2) of Lemma 3.1 is not satisfied for some index $l \in (1, N)$. If there is a scalar $b_l(\varepsilon_l) < 0$ so that the third line in (3.2) may be replaced by

$$(3.13) \quad b_l(\varepsilon_l)\Omega_{1l} \geq (x_{0l}(U(\cdot), Z(0), T) - x_{0l}(U^*(\cdot), Z(0), T)),$$

$$b_l(\varepsilon_l) < 0, \quad \Omega_{1l} \geq 0,$$

then Lemma 3.1 still holds, with the following modification:

(i) must be replaced by: $\forall U(\cdot) \in D(U^*(\cdot))$, the sign of $g(U(\cdot), U^*(\cdot), X(0), T)$ is the same as the sign of $(-1)^q$, where q is the number of indices $l \in (1, N)$ for which (3.13) holds.

The proof is straightforward.

Remark 3.5. During the optimization of $g(U(\cdot), V(\cdot), X(0), T)$, the (A) algorithm will reduce the numerical values of those scalar products

$$|\langle \partial x_{0l}(V_l^0(\cdot), Z(0), T) / \partial u_l(\cdot), u_l(\cdot) - v_l(\cdot) \rangle_l|$$

which are largest in absolute value either:

- (a) because the descent rate of the terminal cost is too large in norm; or
- (b) because $U(\cdot)$ is too distant from the best current approximation $V(\cdot)$ of $U^*(\cdot)$, even if it minimizes the terminal costs $x_{0l}(U(\cdot), Z(0), T)$, $l = 1, N$.

4. The coordination process. In the present section, we regard the assumptions of Lemma 3.1 as being fulfilled. Then let $U^*(\cdot)$ be the Nash optimal control at $X(0)$, and $D(U^*(\cdot))$ the corresponding neighborhood.

DEFINITION 4.1. The *coordination control* $a(\cdot)$ is such that

$$(i) \quad \begin{aligned} a(\cdot) &\triangleq (a_1(\cdot), \dots, a_N(\cdot)) \in F_1 \times \dots \times F_N, \\ a_l(\cdot) &: (0, T) \rightarrow F_l, \quad l = 1, N, \end{aligned}$$

(ii) $a_l(\cdot)$ is at any time the coordination input imposed by the referee $j = 0$ to the infimal player l , $l \in (1, N)$: $u_l(\cdot) = a_l(\cdot)$.

DEFINITION 4.2. The *optimal control problem of the infimal player* l , $l = 1, N$, in the hierarchical system (S) with the coordination strategy $a(\cdot)$ is

$$(4.1) \quad \begin{aligned} x_{0l}(U_l(\cdot), Z(0), T) &\triangleq \min_{U(\cdot)} x_{0l}(U(\cdot), Z(0), T), \\ dx_{0l}/dt &= g_l(U(\cdot), Z(0), T), \\ dX/dt &= (f_1(X(t), U(\cdot), t), \dots, f_n(X(t), U(\cdot), t)), \\ U(\cdot) &\triangleq (u_1(\cdot), \dots, u_N(\cdot)) \in (\mu_1(\cdot) \times \dots \times \mu_N(\cdot)) \cap D(U^*(\cdot)), \\ \forall t \in (0, T): X(t) &\in \chi(t); \quad X(0), x_{0l}(0) \text{ given}, \\ \text{Coordination constraint: } u_l(\cdot) &= a_l(\cdot). \end{aligned}$$

If the optimal control $U_l(\cdot)$ exists for given $a(\cdot)$, we may use the following notation:

$$(4.2) \quad \begin{aligned} U_l(\cdot) &\triangleq (u_{l1}(\cdot), \dots, u_{lN}(\cdot)), \\ X_l(\cdot) &\triangleq \text{optimal trajectory associated to the optimal} \\ &\quad \text{control } U_l(\cdot) \text{ and } Z(0), \\ X_l(\cdot) &\triangleq (x_{l1}(\cdot), \dots, x_{ln}(\cdot)), \quad u_{il}(\cdot) = a_l(\cdot). \end{aligned}$$

DEFINITION 4.3. The *optimal control problem for the referee* $j = 0$ in the hierarchical system (S) with the coordination control $a(\cdot)$ is at $U^*(\cdot)$:

$$\begin{aligned}
 (-1)^q g(\hat{a}(\cdot), U^*(\cdot), X(0), T) &= \min_{a(\cdot)} (-1)^q g(a(\cdot), U^*(\cdot), X(0), T), \\
 dX/dt &= (f_1(X(t), a(\cdot), t), \dots, f_n(X(t), a(\cdot), t)), \\
 dX_l/dt &= (f_1(X_l(t), U_l(\cdot), t), \dots, f_n(X_l(t), U_l(\cdot), t)), \\
 a(\cdot) &\in (\mu_1(\cdot) \times \dots \times \mu_N(\cdot)) \cap D(U^*(\cdot)), \\
 \text{sgn}(g(a(\cdot), U^*(\cdot), X(0), T)) &= \text{sgn}(-1)^q, \\
 u_{il}(\cdot) &= a_l(\cdot), \quad X(0), x_{0l} \text{ given} \quad \forall l \in (1, N), \\
 \forall t &\in (0, T).
 \end{aligned}
 \tag{4.3}$$

This problem includes, besides the overall objective function g , all constraints fulfilled by the coordination control $a(\cdot)$ in all the infimal control problems (Definition 4.2).

Using the interaction balance principle (Mesarovic [19, p. 117]), we will solve the infimal control problems (Definition 4.2), $l = 1, N$, for different coordination controls $a(\cdot)$, in the constraint coordination mode. These coordination controls are modified iteratively while the optimal control problem for the referee $j = 0$ is solved (Definition 4.3).

DEFINITION 4.4. The hierarchical system (S) is *optimally Nash coordinated* (or *balanced*), iff there exists a feasible coordination control $a(\cdot) \in (\mu_1(\cdot) \times \dots \times \mu_N(\cdot)) \cap D(U^*(\cdot))$ which is a Nash optimal control at $X(0)$ for the N players (Definition 1.2).

THEOREM 4.1. *Let there be a Nash optimal control $U^*(\cdot)$ at $X(0)$ for the system (S). If the assumptions of Lemma 3.1 (and Remark 3.4) are fulfilled, then $U^*(\cdot)$ is necessarily a solution of the optimal control problem for the referee $j = 0$, i.e., $\hat{a}(\cdot) = U^*(\cdot)$. Moreover, (S) will then be optimally Nash coordinated.*

Proof. The proof proceeds directly from Lemma 3.1. The constraints

$$\begin{aligned}
 dX_l/dt &= (f_1(X_l(t), U_l(\cdot), t), \dots, f_n(X_l(t), U_l(\cdot), t)), \\
 u_{il}(\cdot) &= a_l(\cdot) \quad \forall l \in (1, N)
 \end{aligned}
 \tag{4.4}$$

are required because they say that $g(\cdot, U^*(\cdot), X(0), T)$ being extremal does not require the terminal costs of the infimal players to be reduced in some way. The simple optimization of $g(\cdot, U^*(\cdot), X(0), T)$ may very well lead to increasing systematically the infimal terminal costs. This is why the optimal control problem for the referee requires that the coordination control $a(\cdot)$ must be "close" to all infimal optimal controls $U_l(\cdot)$, $l = 1, N$.

5. Decomposition of the optimal control problem for the referee $j = 0$. In this section, we assume that Theorem 4.1 holds. In order to ease the resolution of the optimal control problem for the referee (Definition 4.3), we will try to decompose it through the corresponding Lagrangian function L estimated in $a(\cdot) = U^*(\cdot)$. This approach is close to the decomposition schemes for nonlinear dynamic systems used recently.

L will include the imbalance between:

(i) the actual coordination control $a(\cdot)$ and the desired Nash optimal coordination control $U^*(\cdot)$;

(ii) the actual optimal states $X_l(\cdot)$ of the infimal players given $a(\cdot)$, and the desired common equilibrium state $X^*(\cdot)$ associated with $U^*(\cdot)$ by the evolution equation (1.1).

$$\begin{aligned}
 L(a(\cdot)) &\triangleq \frac{d}{dt} g(a(\cdot), U^*(\cdot), X(0), t)(-1)^q \\
 (5.1) \quad &+ \sum_{j=1, n} \psi_{jk} \left(\frac{dx_k^*}{dt} - f_k(X_j(t), U_j(\cdot), t) \right) \\
 &+ \sum_{j=1, N} (\lambda_j(a_j(t) - u_j^*(t)) + \rho_j r_j(a_j(t), t)), \\
 X^*(\cdot) &\triangleq (x_1^*(\cdot), \dots, x_n^*(\cdot)),
 \end{aligned}$$

where ψ_{jk} , λ_j , ρ_j are scalar Lagrange multipliers, and $\rho_j \geq 0$, $t \in (0, T)$. All functions in (5.1) being once continuously Fréchet–Gateaux differentiable, we may write the necessary Euler–Lagrange optimality conditions for L at $U^*(\cdot)$, provided that $a(\cdot)$ is not on the boundary of $(\mu_1(\cdot) \times \dots \times \mu_N(\cdot)) \cap D(U^*(\cdot))$.

$$\begin{aligned}
 a(\cdot) &= U^*(\cdot), \quad k = 1, n, \quad l = 1, N, \quad j = 1, N, \quad t \in (0, T), \\
 (5.2) \quad \frac{\partial L}{\partial x_k^*} - \frac{d}{dt} \frac{\partial L}{\partial \dot{x}_k^*} &= \frac{\partial}{\partial x_k^*} \frac{d}{dt} g(a(\cdot), U^*(\cdot), X(0), t)(-1)^q - \sum_{j=1, N} \frac{d\psi_{jk}}{dt} = 0,
 \end{aligned}$$

$$(5.3) \quad \frac{\partial L}{\partial u_l^*(\cdot)} - \frac{d}{dt} \frac{\partial L}{\partial \dot{u}_l^*(\cdot)} = \frac{\partial}{\partial u_l^*(\cdot)} \frac{d}{dt} g(a(\cdot), U^*(\cdot), X(0), t)(-1)^q - \lambda_l = 0,$$

$$\begin{aligned}
 (5.4) \quad \frac{\partial L}{\partial a_l(\cdot)} - \frac{d}{dt} \frac{\partial L}{\partial \dot{a}_l(\cdot)} &= \lambda_l - \sum_{k=1, n} \psi_{lk} \frac{\partial}{\partial a_l(\cdot)} f_k(X_l(t), U_l(\cdot), t) \\
 &+ (-1)^q \frac{\partial}{\partial a_l(\cdot)} \frac{d}{dt} g(a(\cdot), U^*(\cdot), X(0), t) + \rho_l \frac{\partial}{\partial a_l(\cdot)} r_l(a_l(t), t) = 0,
 \end{aligned}$$

$$(5.5) \quad \psi_{jk} \left(\frac{dx_k^*}{dt} - f_k(X_j(t), U_j(\cdot), t) \right) = 0,$$

$$(5.6) \quad \lambda_l(a_l(t) - u_l^*(t)) = 0, \quad \rho_l r_l(a_l(t), t) = 0, \quad \rho_l \geq 0.$$

We may then conclude that if the Theorem 4.1 holds, and if $U^*(\cdot)$ is not on the boundary of $(\mu_1(\cdot) \times \dots \times \mu_N(\cdot)) \cap D(U^*(\cdot))$, then $a(\cdot) = U^*(\cdot)$ will necessarily satisfy (5.2)–(5.6). Among these relations, only (5.4) is not decentralized, i.e., it may not be solved by a lonely infimal player. This property will be used in the following section.

6. Algorithm (A). This Nash optimal control searching algorithm (A) is viewed as an iterative coordination process in the hierarchical system (S), as in [19, pp. 226–228].

The optimal control problem for the referee $j = 0$ is decentralized as much as possible, leaving at the suprenal level mainly the last nondecentralized relation (5.4). Since it will usually not be equal to zero as long as the Nash optimal control $U^*(\cdot)$ has not been reached, the referee will have to minimize the first member of

(5.4), for example, in the sense of the L^2 -norm on $E^N \times (0, T)$. If the minimum is zero, this should insure the optimality of the apparent overall objective function.

In this iterative coordination process with the apparent overall objective function $g(U(\cdot), V(\cdot), X(0), T)$, $V(\cdot)$ will be at each iteration, the best current approximation of $U^*(\cdot)$. It will be sequentially revised, using the best playable control $U(\cdot)$ minimizing the L^2 -norm of the first member of (5.4) for given $V(\cdot)$.

All optimal control problems, especially the infimal optimal control problems for given $V(\cdot)$, may be solved numerically using an heuristic algorithm (G), such as the generalized reduced gradient (Abadie [1]). (G) finds an optimal control, given decentralized constraints on the control variables; it requires also a feasible start control to be given.

Step A1. Let there be the playable control $a^0(\cdot) \in F$, $a^0(\cdot) \in \mu_1(\cdot) \times \cdots \times \mu_N(\cdot)$ for (S) at $Z(0)$; $\forall a(\cdot)$ close to $a^0(\cdot)$, $\text{sign}(g(a(\cdot), a^0(\cdot), X(0), T)) = \text{sign}((-1)^q)$, and we also require that $g(a(\cdot), a^0(\cdot), X(0), T)$ is small in absolute value. Select $\varepsilon > 0$ depending upon the required precision on $U^*(\cdot)$. Define $m = 0$.

Step A2. (i) Use the (G) algorithm in order to solve iteratively the optimal control problem (A2.1) related to the minimization of the L^2 -norm of (5.4):

$$\begin{aligned} \|T(\hat{a}(\cdot))\|_m &\triangleq \min_{a(\cdot)} \|T(a(\cdot))\|_m, \quad j = 1, N, a(\cdot) \in F, \\ e_j^m(a(\cdot), t) &\triangleq \frac{\partial}{\partial u_j(\cdot)} \frac{d}{dt} g(a(\cdot), a^m(\cdot), X(0), t)(-1)^q \\ &\quad - \sum_{k=1, n} \psi_{jk}^m(a) f_k(X_j^m(t), U_j^m(\cdot), t) \\ &\quad + \rho_j^m(a) \frac{\partial}{\partial u_j(\cdot)} r_j(a_j(t), t), \\ (A2.1) \quad U_j^m &\triangleq (u_{j1}^m, \dots, u_{j(j-1)}^m, a_j, u_{j(j+1)}^m, \dots, u_{jN}^m), \\ \|T(a(\cdot))\|_m^2 &\triangleq \sum_{j=1, N} \int_0^T (e_j^m(a(\cdot), t))^2 dt, \\ a(\cdot) &\in (\mu_1(\cdot) \times \cdots \times \mu_N(\cdot)) \cap D(U^*(\cdot)), \quad t \in (0, T), \\ \text{sgn}(g(a(\cdot), a^m(\cdot), X(0), T)) &= \text{sgn}(g(\hat{a}(\cdot), a^m(\cdot), X(0), T)). \end{aligned}$$

(ii) (A2.1) is solved by calling the Step A3 for any new candidate strategy $a(\cdot)$ appearing in (A2.1); for each given $a(\cdot)$, (A2.1) receives from Steps A3 and A4,

$$(A2.2) \quad \left[\begin{array}{l} \psi_{jk}^m(a), \rho_j^m(a) \\ X_j^m(\cdot), U_j^m(\cdot) \end{array} \right] \left| \begin{array}{l} \text{for } j = 1, N, \quad k = 1, n, \\ t \in (0, T). \end{array} \right.$$

(iii) (A2.1) tells us, according to §5, that $\hat{a}(\cdot)$ optimizes $g(a(\cdot), a^m(\cdot), X(0), T)$, where $a^m(\cdot)$ is the best current approximation of $U^*(\cdot)$. $a^m(\cdot)$ is next used to compute in (A2.3) the corresponding state $X^{*m}(\cdot)$ at the end of the iteration m

$$(A2.3) \quad \begin{aligned} dX^{*m}/dt &= (f_1(X^{*m}(t), a^m(\cdot), t), \dots, f_n(X^{*m}(t), a^m(\cdot), t)), \\ X^{*m}(0) &= X(0), \quad X^{*m} \triangleq (x_1^{*m}, \dots, x_n^{*m}). \end{aligned}$$

(iv) (a) If there exists no playable control $\hat{a}(\cdot) \in F$, $\hat{a}(\cdot) \neq a^m(\cdot)$ as defined in (A2.1), then go to Step A6.

§ $|g(\hat{a}(\cdot), a^m(\cdot), X(0), T)| \geq \varepsilon > 0$: Define $a^{m+1}(\cdot) = \hat{a}(\cdot)$;
 increment m to $(m + 1)$;
 go to (i) and
 continue therefrom
 inside Step A2.

Step A3: infimal players $j = 1, N$. Using (G), solve, given $a(\cdot)$, the N optimal control problems of the infimal players $j = 1, N$ (Definition 4.2). (G) may use $U(\cdot) = a^m(\cdot)$ as a feasible start control for these problems. If it converges, (G) provides the Steps A2 and A4 with the optimal controls $U_j^m(\cdot)$ and the optimal states $X_i^m(\cdot)$ for $j = 1, N$ (see Fig. 3). Go to Step A4.

(i) The referee $j = 0$ receives $X_j^m(\cdot)$, $U_j^m(\cdot)$, $j = 1, N$, corresponding to $a(\cdot)$ from Step A3, and computes $\|T(a(\cdot))\|_m$ by (A2.1). $\lambda_i^m(a)$, $\rho_i^m(a)$, $\psi_{ik}^m(a)$ are computed as required by decentralized equations similar to (5.2), (5.3), (5.5), (5.6), and restated in (A4.1). The transversality conditions of the maximum principle should be used for $t = 0$, $t = T$, and if the control $a(\cdot)$ has a singular arc lying on the boundary of $(\mu_1(\cdot) \times \cdots \times \mu_N(\cdot)) \cap D(U^*(\cdot))$ —see Blaquiere [9] or Leitmann and Stalford [18].

(ii) Return to Step A2 in order to reduce $\|T(a(\cdot))\|_m$ by the coordination at the referee's level.

$$(A5.1) \quad U^*(\cdot) \stackrel{\Delta}{=} \hat{a}(\cdot) \in F$$

Step A6. We are in one of the following circumstances:

(ii) (A2.1) has a solution $\hat{a}(\cdot) \in F$, $\hat{a}(\cdot) \neq a^m(\cdot)$, which becomes a playable control if the constraint $\text{sgn}((-1)^q) = \text{sgn}(g(a(\cdot), a^m(\cdot), X(0), T))$ is relaxed;

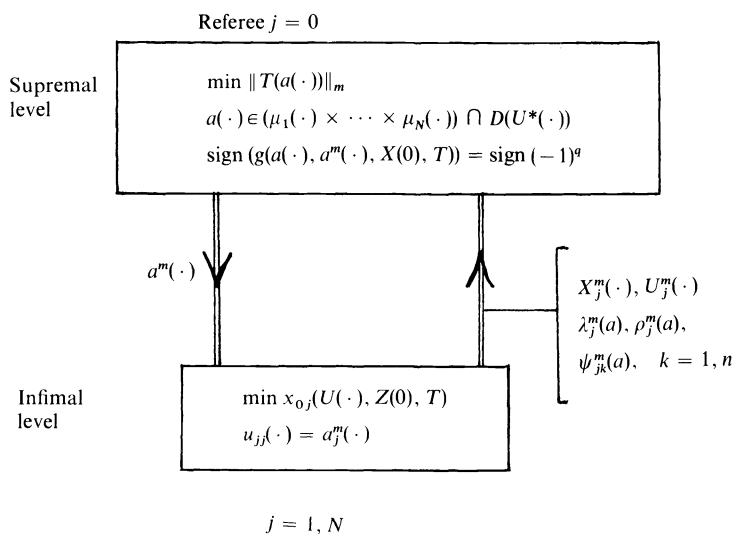


FIG. 3. Information flows in the (A) algorithm

then

$$(A6.1) \quad \text{sgn}(g(\hat{a}(\cdot), a^m(\cdot), X(0), T)) = -\text{sgn}((-1)^q).$$

The best corrective action should be to start up at Step A1 with another initial control $a^0(\cdot)$, and $q = q + 1$.

(iii) The (A) algorithm does not converge, because there exists no Nash optimal control for our problem in F at $Z(0)$ and in a neighborhood of $a^0(\cdot)$. See Remark 6.3.

THEOREM 6.1. Assume that we dispose over a Nash optimal control $U^*(\cdot) \in F$, and that the assumptions of Theorem 4.1 are satisfied. We moreover require that $U^*(\cdot)$ has no singular arc lying on the boundary of $(\mu_1(\cdot) \times \cdots \times \mu_N(\cdot)) \cap D(U^*(\cdot))$, so that the necessary optimality conditions of §5 may be used as such. We may then conclude that if the initial control $a^0(\cdot) \in D(U^*(\cdot))$, then the (A) algorithm will necessarily converge towards $U^*(\cdot)$, if it does converge.

Proof. It is a direct consequence of our previous results.

Remark 6.1. In most cases of practical interest, we will usually not know a priori if there exists a Nash optimal control in F , in a neighborhood of $a^0(\cdot)$. Moreover, all results sustaining the (A) algorithm are mathematically necessary, but usually not sufficient, optimality or equilibrium conditions. Therefore, if the (A) algorithm converges, i.e., stops at Step A5, one may only conclude that $U^*(\cdot)$ is an ε -approximate Nash optimal control, after having verified whether it satisfies a sufficient Nash optimality condition, or Definition 1.2.

A number of sufficiency conditions may be found in the literature as “verification results”: Isaacs [16], Blaqui re [8], Case [10]; one of the most readily verifiable sufficiency theorems is given in Leitmann and Stalford [18]: it uses the adjoint functions for the N players, and these are provided by the (G) algorithm during the last Step A3.

Remark 6.2. The (A) algorithm may be in trouble sometimes because of the

constraints $\chi(\cdot)$ and $\mu_1(\cdot) \times \cdots \times \mu_N(\cdot)$, often because the singular arcs are handled unsatisfactorily by the (G) algorithm. Another major difficulty is the usual large number of jumps of $U^*(\cdot)$, the consequence of it being a large number of short disconnected control arcs, some of them being singular.

In the numerical examples which have been solved, (A) converges fairly well thanks to the “forcing” function g . (A) achieves some kind of direct convergence towards $U^*(\cdot)$ along a thalweg for g acceptable by all infimal players. $U^*(\cdot)$ is therefore not obtained by perturbation methods leading to find the intersection of control surfaces parametrized with respect to the other players’ decisions; this “intersection” approach is probably unreliable in large systems (Advani-Mukundan [2]).

Remark 6.3. It may very well be that there exists no Nash optimal control in F for a specific application problem having the structure specified in §1 for the topologies used on E, F . Sometimes, one may find out that the solution obtained is a Pareto equilibrium or noninferior equilibria. Much assistance is required from existence theorems for Nash equilibria, but these are few: Varaiya [26], Friedman [11]. One must also admit that, in many cases with practical relevance, human experience will not clarify the problem of the existence of such equilibria.

7. Numerical application. The (A) algorithm has successfully been applied to a large nonlinear nonconvex dynamic sectoral model of the Danish economy in the years 1947–52 (model described in detail in [20]). All differential equations from the previous sections were however replaced by difference equations, the sampling period being one year. Six interrelated sectors $j = 1, \dots, 6$ were considered: agriculture, industry, transportation, State and public works, services, housing (rentals). The criterion for each of the private sectors $j \neq 4, j = 1, 6$, was the present value of the cash-flow generated by the sector after taxes. The criterion of the State was the present value of the surplus of public services and help to the other sectors, as financed partly via taxes earned from households and private enterprises. The 21 controls are:

- (i) for each private sector: investments, labor, write-offs;
- (ii) for the State $j = 4$; total import for free consumption, marginal tax-rates on wages and profits for the private sectors, and the State’s own investments, labor, write-offs.

The 2 state variables were: foreign debt, State budget excess. Production functions are used for each sector, depending on investments, write-offs, and labor. The sectors are tied together namely by an input-output relation. Exogeneous constraints were introduced on the state and control variables, although only few of them were active in most computer runs. Substitution constraints were also introduced on the self-financing ratio in each sector, on the nationwide unemployment rate, on the smoothness in year-to-year labor changes in each sector, and on the write-offs in each sector.

The initial controls $a^0(\cdot)$ of A1 were the actual historical values for 1947–52. Depending on the constraints introduced into the model, the computer time required was 0,8–2,0 hours on IBM 370-65.

The numerical results (Figs. 4, 5, 6) indicate that the Nash optimal state $X^*(\cdot)$ is somehow closer to the actual historical evolutions than the one yielded by

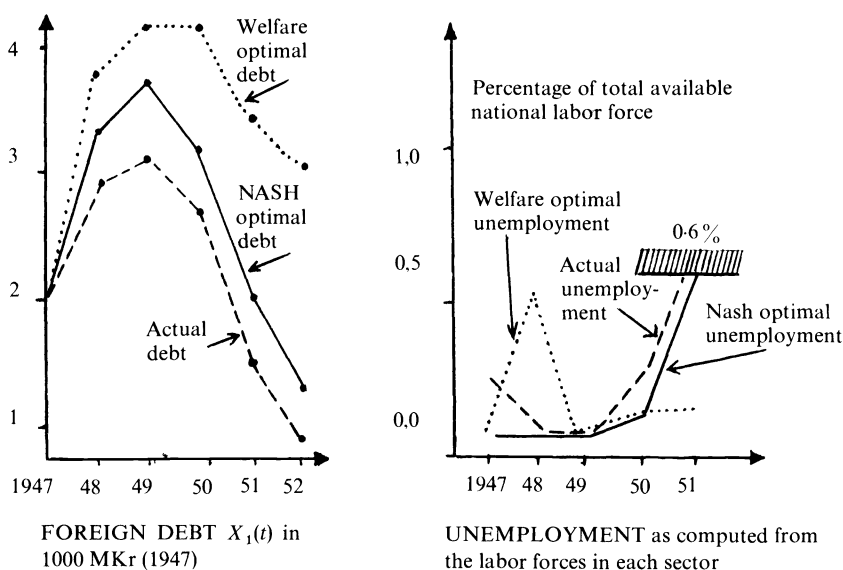


FIG. 4. Foreign debt (state variable) and unemployment for 3 different equilibria: actual historical values, Nash equilibrium, and maximization of the consumption welfare functional of Houthakker (direct addilog utility index)

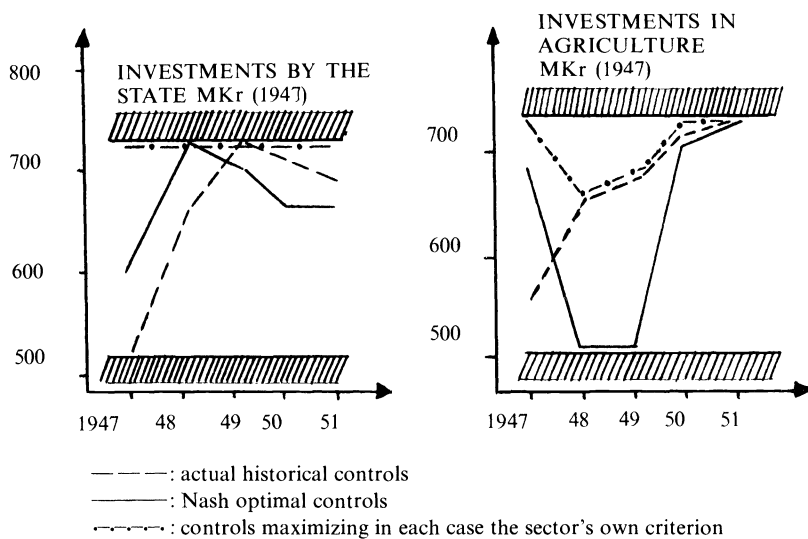


FIG. 5. Investments in the public and agricultural sectors

maximizing a single classical welfare criterion ([20]) based on the aggregation of the consumptions from all sectors by the households.

The Nash equilibrium control $U^*(\cdot)$ seems for example to secure an efficient defense of the agricultural labor force, as opposed to the reduction requested by this sector's own present value criterion.

How well the (A) algorithm would stand up under further numerical research is still open to question, and much work is still required (see Remarks from §6).

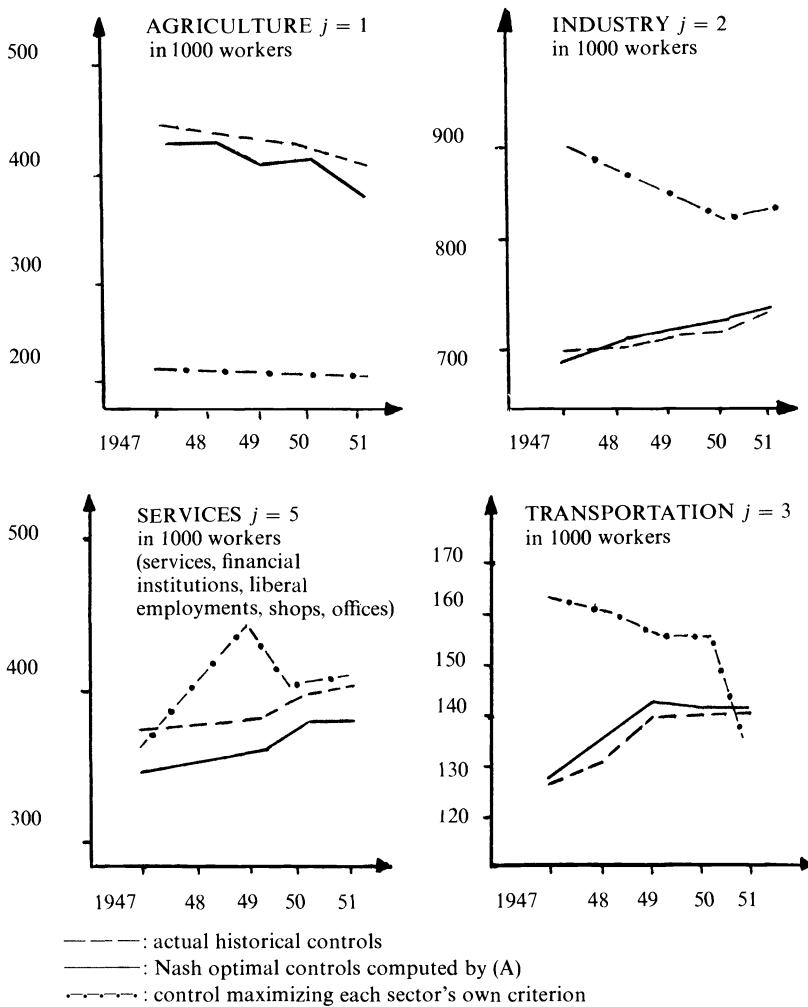


FIG. 6. Labor force evolutions in 4 sectors; the maximum relative year-to-year change for all these controls is 15%

Besides improving the (G) and (A) algorithms, it is necessarily better to investigate the properties of the local equilibria $U^*(\cdot)$ which they compute. It is moreover necessary to find new methods in order to handle the discontinuities of the equilibrium control $U^*(\cdot)$.

Acknowledgments. The author wishes to thank Professor A. Blaqui re, Laboratoire d'Automatique th orique, Universit  Paris VII. This work has been in cooperation with this institution.

REFERENCES

- [1] J. ABADIE, *Application of the GRG algorithm to optimal control problems*, Integer and Non-linear Programming, J. Abadie, ed., North-Holland, Amsterdam, 1970, pp. 191-193.
- [2] R. J. ADVANI AND R. MUKUNDAN, *Optimal control methods applied to economic systems*, International J. Systems Sci., 1 (1970-71), no. 2, pp. 153-172.
- [3] A. BENSOUSSAN, J. L. LIONS AND R. TENAM, *M thodes num riques d'analyse de systemes, Tome 2: m thodes de d composition*, Cahier no 11, IRIA, Paris, 1972, pp. 5-190.

- [4] A. BENSOUSSAN, *Price decentralization in the case of interrelated payoffs*, Techniques for Optimization, A. V. Balakrishnan, ed., Academic Press, New York, 1972, pp. 391–406.
- [5] L. D. BERKOVITZ, *A variational approach to differential games*, *Advances in Game Theory*, *Annals of Mathematical Studies* no. 52, Princeton University Press, Princeton, 1964, pp. 127–174.
- [6] ———, *Necessary conditions for optimal strategies in a class of differential games and control problems*, this Journal, 5 (1967), pp. 1–24.
- [7] ———, *A survey of differential games*, Mathematical Theory of Control, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, New York, 1967.
- [8] A. BLAQUIÈRE, F. GÉRARD AND G. LEITMANN, *Quantitative and Qualitative Games*, Academic Press, New York, 1969.
- [9] A. BLAQUIÈRE, L. JURICEK AND K. E. WIESE, *Geometry of Pareto equilibria and a maximum principle in N-person differential games*, J. Math. Anal. Appl., 38 (1972), pp. 223–243.
- [10] J. H. CASE, *Toward a theory of many player differential games*, this Journal, 7 (1969), pp. 179–197.
- [11] J. W. FRIEDMAN, *A non-cooperative equilibrium for supergames*, Rev. Economic Studies (1), 38 (1971), no. 113, pp. 1–12.
- [12] V. GOURISHANKAR AND A. SALAMA, *A technique for solving a class of differential games*, Internat. J. Control, 15 (1972), no. 3, pp. 529–539.
- [13] J. C. HARSANYI AND R. SELTEN, *A generalized Nash solution for two-person bargaining games with incomplete information*, Management Sci., 18 (1972), pp. 80–106.
- [14] Y. C. HO AND A. W. STARR, *Non-zero sum differential games*, J. Optimization Theory Appl., 3 (1969), pp. 144–152.
- [15] D. HOLT AND R. MUKUNDAN, *A Nash algorithm for a class of non-zero sum differential games*, Internat. J. Systems Sci., 2 (1972), no. 4, pp. 379–387.
- [16] R. ISAACS, *Differential Games*, John Wiley, New York, 1965.
- [17] G. JUMARIE, *Towards a theory of multilevel hierarchical games and its perspectives in economic systems*, IFAC/IFORS International Conference on Dynamic Modelling and Control of National Economies, Conference publication no. 101, Institution of Electrical Engineers, London, 1973, pp. 270–281.
- [18] G. LEITMANN AND H. STALFORD, *Sufficiency for optimal strategies in Nash equilibrium games*, Techniques of Optimization, A. V. Balakrishnan, ed., Academic Press, New York, 1972, pp. 279–285.
- [19] M. D. MESAROVIC, D. MACKO AND Y. TAKAHARA, *Theory of hierarchical multilevel systems*, Mathematics in Science and Engineering, no. 68, Academic Press, New York, 1970.
- [20] L. F. PAU, *Differential game in macroeconomy*, IFAC/IFORS International Conference on Dynamic Modelling and Control of National Economies, Conference publication no. 101, Institution of Electrical Engineers, London, 1973, pp. 254–269.
- [21] ———, *Control, Conflict and Cooperation in Evolutive Economic Systems*, IMSOR, Technical University of Denmark, Copenhagen, 1971.
- [22] V. H. QUINTANA AND E. J. DAVISON, *Two numerical techniques to solve differential game problems*, Internat. J. Control, 16 (1972), no. 3, pp. 465–474.
- [23] M. SIMAAN AND J. B. CRUZ, *On the solution of the open-loop Nash–Riccati equations in linear quadratic differential games*, Ibid., 18 (1973), no. 1, pp. 57–63.
- [24] A. W. STARR, *Computation of Nash equilibria for non-linear non-zero sum differential games*, Proc. 1st Internat. Conference on the Theory and Applications of Differential Games, Univ. of Massachusetts, Amherst, Mass., 1969, pp. IV 13–18.
- [25] L. G. TELSER, *Competition, Collusion and Game Theory*, Aldine-Atherton, Chicago, 1972.
- [26] P. P. VARAIYA, *On the existence of solutions to a differential game*, this Journal, 5 (1967), pp. 153–162.

REGULATION OF LINEAR STOCHASTIC SYSTEMS*

J. SNYDERS† AND W. M. WONHAM‡

Abstract. Linear constant dynamic systems with noisy measured output y and a second output z (the one to be regulated), are considered. Sufficient conditions for the existence of an estimated-state feedback that assures mean-square boundedness on $[0, \infty)$ of z are derived, the estimation being based on y . The part of the problem related to boundedness of the estimation error is more thoroughly investigated, and necessary and sufficient conditions are obtained. Necessary and sufficient conditions for boundedness of the solution to a Riccati equation are included in this result.

1. Introduction. Systems considered in this paper are represented by a state n -vector $x(\cdot)$, an input m -vector $u(\cdot)$, an observed output p -vector $y(\cdot)$ and an output q -vector $z(\cdot)$ satisfying the (formal) Itô equations

$$(1) \quad \begin{aligned} dx(t) &= Ax(t) dt + Bu(t) dt + Gdw(t), & x(0) &= x_0, \\ dy(t) &= Cx(t) dt + Hdw(t), & y(0) &= 0, \\ z(t) &= Dx(t). \end{aligned}$$

Here $w(\cdot)$ is a standard r -dimensional Brownian motion, and the matrices A, B, C, D, G and H are constant with real entries. Based on the observations $\{y(t); t \geq 0\}$ an input $u(\cdot)$ is to be found for regulation of $z(\cdot)$. In problems dealing with noiseless systems described by (1) with $G = 0$ and $H = 0$ it is customary [1], [2] to adopt the following requirement as the meaning of regulation: $z(t) \rightarrow 0$ as $t \rightarrow \infty$ for every initial state x_0 . In the present context we must take into account that x_0 is a random variable. We say that x_0 is an *admissible* initial condition if x_0 is a second-order random variable independent of the $w(\cdot)$ process. Furthermore, as the requirement $z(t) \rightarrow 0$ is unrealistic in the presence of dynamic noise, we replace it by the more modest criterion that the mean square values of the components of $z(\cdot)$ be bounded on $[0, \infty)$. Finally, instead of admitting a priori all nonanticipative functionals $u(\cdot)$ of the observations $y(\cdot)$, we consider only controls $u(\cdot)$ such that $u(t)$ is a linear transformation of the estimated value of $x(t)$. This simplifying restriction is motivated by the separation theorem [3] of control and filtering, although we do not claim that the separation assumption is actually nonrestrictive in the present situation as well.

In §2 the problem is posed in terms of matrix differential equations: a Riccati equation and a linear equation with forcing term depending on the Riccati solution. Sufficient conditions for regulation are obtained in §3. It is shown in §4 that, in case the state-noise and observation-noise processes are independent, certain of the sufficiency conditions are also necessary. Necessary and sufficient conditions for boundedness of the solution to a Riccati equation are included in this result. In the extensive literature dealing with Riccati equations very few necessary

* Received by the editors February 1, 1974. This research was supported in part by the National research Council of Canada under Grant A-7399.

† System Science Department, School of Engineering and Applied Science, University of California, Los Angeles, California 90024.

‡ Department of Electrical Engineering, University of Toronto, Toronto, Canada M5S 1A4.

conditions concerning the existence of properties of solutions are derived [4], and only a few results on boundedness exist [5], [6]. For ease of reference certain known results are listed in the Appendix.

2. Preliminaries. Linear spaces are denoted by script capitals, and Roman capitals stand both for maps (i.e., linear transformations) and their matrix representations. Unless otherwise stated or indicated by appropriate notation, all spaces, maps and matrices are defined over the real numbers. Nevertheless, we shall frequently deal with the extended spaces defined over the complex field (i.e., spaces obtained as a linear span, with complex coefficients, of any basis in the original spaces), thus permitting operation of maps on complex vectors. Unless otherwise stated, all spaces and subspaces which appear explicitly are nonzero. All identity maps are represented by I . The image of a map B is written $\text{Im } B$, and $\text{Ker } B = \{x: Bx = 0\}$. $\text{Re } \lambda$ and $\bar{\lambda}$ stand, respectively, for the real part and complex conjugate of a number λ . Transpose of a matrix B is denoted by B' , and $B^* = \bar{B}'$. A matrix M is called semisimple if it has a block-diagonal structure $M = \text{diag}(J_1, J_2, \dots, J_k)$, where each J_i is either diagonal or has the form

$$J_i = \begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix}$$

with real α and β . For a semisimple M there exists a unitary matrix T such that TMT^* is diagonal (with complex entries). A semisimple map is a map that is representable by a semisimple matrix. If M and N are self-adjoint matrices, then $M \leq N$ means that $N - M$ is nonnegative definite.

Throughout the paper \mathcal{X} is an n -space and $A: \mathcal{X} \rightarrow \mathcal{X}$ is a fixed map. Let \mathcal{U} , \mathcal{Y} and \mathcal{F} be linear spaces, $\mathcal{F} \subset \mathcal{X}$, and let $B: \mathcal{U} \rightarrow \mathcal{X}$ and $C: \mathcal{X} \rightarrow \mathcal{Y}$ be fixed maps. Let \mathcal{X}/\mathcal{F} be the factor space mod \mathcal{F} , and denote by T , $T: \mathcal{X} \rightarrow \mathcal{X}/\mathcal{F}$ the canonical projection. If $\mathcal{F} \subset \text{Ker } C$, then C/\mathcal{F} stands for the unique map $C/\mathcal{F}: \mathcal{X}/\mathcal{F} \rightarrow \mathcal{Y}$ satisfying $C = (C/\mathcal{F})T$. If $A\mathcal{F} \subset \mathcal{F}$, then A/\mathcal{F} is the restriction of A to \mathcal{F} and A/\mathcal{F} is the unique (induced) map $A/\mathcal{F}: \mathcal{X}/\mathcal{F} \rightarrow \mathcal{X}/\mathcal{F}$ satisfying $(A/\mathcal{F})T = TA$. The map $B/\mathcal{F}: \mathcal{U} \rightarrow \mathcal{X}/\mathcal{F}$ is defined as $B/\mathcal{F} = TB$. Let $\phi^+(\lambda)$, $\phi^0(\lambda)$ and $\phi^-(\lambda)$ be factors of the minimal polynomial of A having roots exclusively with positive, zero and negative real parts, respectively. Then \mathcal{X} is decomposable into the direct sum $\mathcal{X} = \mathcal{X}^+(A) \oplus \mathcal{X}^0(A) \oplus \mathcal{X}^-(A)$, where $\mathcal{X}^\times(A) = \text{Ker } \phi^\times(A)$ with $^\times$ standing for any superscript. We write $\mathcal{X}^{0+}(A)$ for $\mathcal{X}^0(A) \oplus \mathcal{X}^+(A)$. The unobservable subspace $\mathcal{N}(C, A)$ of (C, A) is defined as $\mathcal{N}(C, A) = \bigcap_{i=1}^n \text{Ker } (CA^{i-1})$. The pair (C, A) is called observable or detectable if $\mathcal{N}(C, A) = 0$ or $\mathcal{N}(C, A) \subset \mathcal{X}^-(A)$, respectively. $\langle A | \text{Im } B \rangle$ denotes the controllable subspace associated with (A, B) , i.e., $\sum_{j=1}^n A^{j-1}(\text{Im } B)$.

Let λ be the characteristic value of A associated with some Jordan block of dimension k . An ordered set of generalized characteristic vectors $(Ax^1 = \lambda x^1, Ax^j = \lambda x^j + x^{j-1}, k \geq j \geq 2)$ is called a characteristic chain related to λ . The ordered subsets (x^1, x^2, \dots, x^j) , $k \geq j \geq 1$, of this characteristic chain are called characteristic subchains associated with λ . It is easily checked that every A -invariant subspace is spanned by the real and imaginary parts of suitable characteristic subchains of A . In particular, $\mathcal{N}(C, A)$ is spanned by the real and imaginary parts of those characteristic subchains of A which are nullified by C [7].

Let $x(\cdot)$, $y(\cdot)$ and $z(\cdot)$ satisfy (1) and write $\hat{x}(t) = \mathcal{E}\{x(t)|y_0'\}$, where $\mathcal{E}\{\cdot|y_0'\}$ means expectation conditional on the σ -algebra generated by $\{y(s); 0 \leq s \leq t\}$. It is well known that, if x_0 is admissible, HH' is invertible and under certain restrictions [3] on $u(\cdot)$, $\hat{x}(\cdot)$ is determined by the stochastic differential equation

$$d\hat{x}(t) = A\hat{x}(t) dt + Bu(t) dt + [P(t)C' + GH'] \cdot (HH')^{-1}[dy(t) - C\hat{x}(t) dt], \quad \hat{x}(0) = \mathcal{E}x_0,$$

where $P(t) = \mathcal{E}\{[x(t) - \hat{x}(t)][x(t) - \hat{x}(t)]'|y_0'\}$. Furthermore, $P(\cdot)$ is given by the Riccati equation

$$\frac{dP}{dt} = AP + PA' + GG' - (PC' + GH')(HH')^{-1}(CP + HG'), \quad P(0) = P_0, \quad (2)$$

where $P_0 = \mathcal{E}\{(x_0 - \mathcal{E}x_0)(x_0 - \mathcal{E}x_0)'\}$. We admit only controls of the form $u(t) = F\hat{x}(t)$, where F is a constant linear map. Taking into account that

$$dy(t) - C\hat{x}(t) dt = (HH')^{-1/2} dw_1(t),$$

where $w_1(\cdot)$ is standard Brownian motion [3], there follows

$$d\hat{x}(t) = (A + BF)\hat{x}(t) dt + (PC' + GH')(HH')^{-1/2} dw_1(t), \quad \hat{x}(0) = \mathcal{E}x_0.$$

Writing $Q(t) = \mathcal{E}\{\hat{x}(t)\hat{x}(t)'\}$ we have

$$\begin{aligned} \frac{dQ}{dt} &= (A + BF)Q + Q(A + BF)' \\ &+ (PC' + GH')(HH')^{-1}(CP + HG'), \quad Q(0) = Q_0, \end{aligned} \quad (3)$$

where $Q_0 = \mathcal{E}x_0(\mathcal{E}x_0)'$. Let $e(t) = x(t) - \hat{x}(t)$. Then evidently $P(t) = \mathcal{E}\{e(t)e(t)'\}$ and also

$$\mathcal{E}\{e(t)\hat{x}(t)'\} = \mathcal{E}\{\mathcal{E}\{[x(t) - \hat{x}(t)]\hat{x}(t)'\}|y_0'\} = 0.$$

Consequently $\mathcal{E}\{x(t)x(t)'\} = P(t) + Q(t)$ and, setting $M(t) = \mathcal{E}\{z(t)z(t)'\}$, it follows that

$$M(t) = DP(t)D' + DQ(t)D'. \quad (4)$$

Regulation of the system (1) is defined as assuring boundedness of $M(\cdot)$ on $[0, \infty)$ for every admissible initial condition x_0 . Accordingly, we shall derive conditions for the existence of a feedback map F such that, in view of (2) and (3), $M(\cdot)$ is bounded on $[0, \infty)$ for every $P_0 \geq 0$ and $Q_0 \geq 0$. Obviously, boundedness of both $DP(\cdot)D'$ and $DQ(\cdot)D'$ is equivalent to boundedness of $M(\cdot)$. In the next section sufficient conditions for regulation are obtained. Necessary and sufficient conditions for boundedness of $DP(\cdot)D'$ are derived in §4 under the assumption that the state-noise and observation-noise processes are independent, i.e., $GH' = 0$ (if $GH' \neq 0$ then these conditions are sufficient). Evidently, boundedness of the solution to Riccati equations is included in the results.

3. Sufficient conditions. Let us write (2) in the form

$$\frac{dP}{dt} = \tilde{A}P + P\tilde{A}' + \tilde{G}\tilde{G}' - P\tilde{C}'\tilde{C}P, \quad P(0) = P_0, \quad (5)$$

where

$$(6) \quad \begin{aligned} \tilde{A} &= A - GH'(HH')^{-1}C, \\ \tilde{G} &= G[I - H'(HH')^{-1}H], \\ \tilde{C} &= (HH')^{-1/2}C. \end{aligned}$$

LEMMA 1. *The conditions listed below imply that $M(\cdot)$ is bounded on $[0, \infty)$ for all $P_0 \geq 0$ and $Q_0 \geq 0$:*

- (a) $\mathcal{N}(C, A) \cap \mathcal{X}^{0+}(A) \subset \text{Ker } D$,
 (b) F is a constant map such that

$$\begin{aligned} \mathcal{X}^{0+}(A + BF) &\subset \text{Ker } D, \\ \mathcal{N}(C, A) \cap \mathcal{X}^{0+}(A) &\subset \text{Ker } F. \end{aligned}$$

To interpret this result vis-à-vis the processes defined by (1), it is helpful to keep in mind that $\text{Ker } D$ in condition (a), and $\text{Ker } F$, are replaceable by $\mathcal{N}(D, A)$ and $\mathcal{N}(F, A)$ respectively. It is also noteworthy that condition (a) is equivalent to detectability of the pair $(C/R, A/R)$, where R may stand for either $\mathcal{N}(C, A) \cap \mathcal{N}(D, A)$ or $\mathcal{N}(C, A) \cap \mathcal{X}^{0+}(A)$.

Proof. By (a), with a suitable basis for \mathcal{X} ,

$$C = (C_1 \ 0), \quad D = (D_1 \ 0), \quad A = \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix},$$

where $(\tilde{C}_1, \tilde{A}_{11})$ is detectable. Thus the maps introduced in (6) admit the representation

$$\tilde{C} = (\tilde{C}_1 \ 0), \quad \tilde{A} = \begin{pmatrix} \tilde{A}_{11} & 0 \\ \tilde{A}_{21} & \tilde{A}_{22} \end{pmatrix}, \quad \tilde{G} = \begin{pmatrix} \tilde{G}_1 \\ \tilde{G}_2 \end{pmatrix},$$

where $(\tilde{C}_1, \tilde{A}_{11})$ is detectable and $\tilde{A}_{22} = A_{22}$. With the corresponding representation of P , (5) becomes

$$\frac{dP_{11}}{dt} = \tilde{A}_{11}P_{11} + P_{11}\tilde{A}_{11}' + \tilde{G}_1\tilde{G}_1' - P_{11}\tilde{C}_1'\tilde{C}_1P_{11}.$$

Corollary 4 (Appendix) now assures boundedness of $P_{11}(\cdot)$ on $[0, \infty)$. Obviously $Q(\cdot)$ satisfies

$$\begin{aligned} Q(t) &= e^{(A+BF)t}Q_0e^{(A+BF)'t} \\ &+ \int_0^t e^{(A+BF)s}[P(s)C' + GH'](HH')^{-1}[CP(s) + HG']e^{(A+BF)'s} ds. \end{aligned}$$

According to condition (b), $F = (F_1 \ 0)$ in the chosen basis, and setting $B' = (B'_1 \ B'_2)$ we have

$$(7) \quad De^{(A+BF)s}P(s)C' = D_1e^{(A_{11}+B_1F_1)s}P_{11}(s)C'_1.$$

Since $P_{11}(\cdot)$ is bounded on $[0, \infty)$ and $D \exp(A + BF)t \rightarrow 0$ as $t \rightarrow \infty$, it thus follows that $DQ(\cdot)D'$ is bounded on $[0, \infty)$.

By Theorem 1 of [1], (b) implies (a), and this gives the following.

PROPOSITION 1. $M(\cdot)$ is bounded on $[0, \infty)$ for every $P_0 \geq 0$ and $Q_0 \geq 0$ provided

$$\mathcal{X}^{0+}(A + BF) \subset \text{Ker } D$$

and

$$\mathcal{N}(C, A) \cap \mathcal{X}^{0+}(A) \subset \text{Ker } F.$$

Furthermore, Corollary 1.2 of [1] yields an algorithmic procedure for checking the existence of a map F that satisfies the above conditions. In view of this result we have the following theorem.

THEOREM 1. There exists F such that $M(\cdot)$ is bounded on $[0, \infty)$ for every $P_0 \geq 0, Q_0 \geq 0$ if

$$(8) \quad \mathcal{N}(C, A) \cap \mathcal{X}^{0+}(A) \subset \text{Ker } D$$

and

$$(9) \quad \mathcal{X}^{0+}(A) \subset \langle A | \text{Im } B \rangle + \mathcal{V}^*.$$

Here \mathcal{V}^* is computed as follows:

$$\begin{aligned} \mathcal{V}^0 &= \text{Ker } D, \\ \mathcal{V}^j &= \text{Ker } D \cap A^{-1}(\mathcal{V}^{j-1} + \text{Im } B), \quad j = 1, 2, \dots, \\ \mathcal{V}^* &= \mathcal{V}^n \end{aligned}$$

with A^{-1} standing for the functional inverse of A , i.e., $A^{-1} \mathcal{K} = \{x : Ax \in \mathcal{K}\}$.

Condition (8) (equivalently (a)) states that the unstable y -unobservable modes of A are also z -unobservable, while condition (9) means, roughly, that the unstable modes of A are either controllable or z -unobservable.

Write now

$$\bar{\mathcal{X}} = \mathcal{X}/\mathcal{R}, \quad \bar{A} = A/\mathcal{R}, \quad \bar{G} = G/\mathcal{R} \quad \text{and} \quad \bar{C} = C/\mathcal{R},$$

where

$$\mathcal{R} = \mathcal{N}(C, A) \cap \mathcal{N}(D, A),^1$$

and consider the following conditions:

- (c₁) $\bar{A} | \mathcal{N}(\bar{C}, \bar{A}) \cap \mathcal{X}^0(\bar{A})$ is semisimple,
- (c₂) $(\bar{C} | \mathcal{N}(\bar{C}, \bar{A}) \cap \mathcal{X}^0(\bar{A}), \bar{A} | \mathcal{N}(\bar{C}, \bar{A}) \cap \mathcal{X}^0(\bar{A}))$ is detectable,
- (c₃) there exists an \bar{A} invariant complement $\bar{\mathcal{L}}$ to $\mathcal{N}(\bar{C}, \bar{A}) \cap \mathcal{X}^0(\bar{A})$ in $\bar{\mathcal{X}}$ such that $\text{Im } \bar{G} \subset \bar{\mathcal{L}}$ (or, equivalently, $\langle \bar{A} | \text{Im } \bar{G} \rangle \subset \bar{\mathcal{L}}$).

¹ Again, no result would change if we defined instead

$$\mathcal{R} = \mathcal{N}(C, A) \cap \mathcal{N}(D, A) \cap \mathcal{X}^{0+}(A).$$

It is shown in §4 (Corollary 1) that (c_1) – (c_3) imply boundedness of $DP(\cdot)D'$ on $[0, \infty)$ for every $P_0 \geq 0$. Consequently, if (c_1) – (c_3) are satisfied and F is such that

$$\mathcal{X}^{0+}(A + BF) \subset \text{Ker } D$$

and

$$\mathcal{N}(C, A) \cap \mathcal{N}(D, A) \subset \text{Ker } F,$$

then regulation is achieved. Indeed, this results in view of (7), where A_{11} represents $A/\mathcal{N}(C, A) \cap \mathcal{N}(D, A)$ and by application of Lemma 2 (Appendix) to demonstrate that $P_{11}(\cdot)$ is bounded on $[0, \infty)$.

The set of conditions (c_1) – (c_3) is obviously less restrictive than (a): it allows the existence of purely oscillatory and constant (unstable) modes of A which are y -unobservable but z -observable provided, roughly, that these modes are not directly driven by the dynamic noise. We shall not proceed along this line of tightening the sufficient conditions further, but merely point out that there is room for improvement even if $GH' = 0$, although in that case (c_1) – (c_3) are also necessary according to Theorem 2. The last statement is justified by the following scalar example: $A = B = G = 0$, $C = D = 1$. Evidently, there is no F such that $\mathcal{X}^{0+}(A + BF) \subset \text{Ker } D$; nevertheless $P(t) + Q(t) = P_0 + Q_0$ is bounded.

4. Boundedness of $DP(\cdot)D'$. Independence of the state-noise and observation-noise processes is expressed by the condition $GH' = 0$. This and the normalization $HH' = I$ yield

$$(10) \quad \frac{dP}{dt} = AP + PA' + GG' - PC'CP, \quad P(0) = P_0.$$

In the next theorem x stands for the last (j th) member of a characteristic subchain of A associated with characteristic value λ , and it is assumed that all members of this subchain preceding x (if any) are nullified both by C and D .

THEOREM 2. $DP(\cdot)D'$, where $P(\cdot)$ satisfies (8), is bounded on $[0, \infty)$ for every $P_0 \geq 0$ if and only if the conditions (c_1) – (c_3) listed above or, equivalently, the conditions (d_1) – (d_3) below, hold for each x :

- (d_1) if $\text{Re } \lambda > 0$ and $Cx = 0$ then $Dx = 0$;
- (d_2) if $\text{Re } \lambda = 0$, $Cx = 0$ and there exists a (complex) vector w such that $Aw = \lambda w + x$, then $Dx = 0$;
- (d_3) if $\text{Re } \lambda = 0$, $Cx = 0$ and there exist (complex) vectors y and z such that

$$(11) \quad \begin{pmatrix} -A' & C'C \\ GG' & A \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} = \lambda \begin{pmatrix} y \\ z \end{pmatrix} + \begin{pmatrix} 0 \\ x \end{pmatrix},$$

then $Dx = 0$.

It is easily seen that conditions (d_1) – (d_3) are actually constructive. To check them, select any set of characteristic chains of A such that the real and imaginary parts of their elements provide a basis for \mathcal{X} . $DP(\cdot)D'$ is bounded on $[0, \infty)$ for every $P_0 \geq 0$ if and only if each of these selected chains satisfies the stated conditions.

Proof. We shall follow the scheme: $(c_1)-(c_3) \Rightarrow$ boundedness $\Rightarrow (d_1)-(d_3) \Rightarrow (c_1)-(c_3)$. To show the sufficiency of $(c_1)-(c_3)$ select a basis for \mathcal{X} in which

$$(12) \quad \begin{aligned} C &= (C_1 \quad 0 \quad 0), \quad D = (D_1 \quad D_2 \quad 0), \\ A &= \begin{pmatrix} A_{11} & 0 & 0 \\ 0 & A_{22} & 0 \\ A_{31} & A_{32} & A_{33} \end{pmatrix}, \quad G = \begin{pmatrix} G_1 \\ 0 \\ G_3 \end{pmatrix}, \end{aligned}$$

where (C_1, A_{11}) is detectable, $\text{Ker } D_2 = 0$ and A_{22} is semisimple with all its characteristic values having zero real part. Set $K' = (K'_1 \quad 0 \quad 0)$, where K'_1 is such that $A_{11} + K'_1 C_1$ is stable. The result now follows by Lemma 2.

To prove the necessity of (d_1) and (d_2) it is enough to consider the case where $G = 0$ and P_0 has complex entries; this follows by Lemma 3 and the observation that if $P_0 \geq 0$ then $P_0 + \bar{P}_0 \geq P_0$. The corresponding solution to (10) is given by²

$$(13) \quad P(t) = e^{At} P_0 \left[I + \int_0^t e^{A's} C' C e^{As} P_0 ds \right]^{-1} e^{A't}$$

and we may set $P_0 = xx^*$. Simple manipulations yield

$$\begin{aligned} x^* \left(I + \int_0^t e^{A's} C' C e^{As} x x^* ds \right) &= \left(1 + x^* C' C x \int_0^t e^{(\lambda + \bar{\lambda})s} ds \right) x^*, \\ DP(t)D' &= \frac{e^{(\lambda + \bar{\lambda})t}}{1 + x^* C' C x \int_0^t e^{(\lambda + \bar{\lambda})s} ds} D x x^* D' \end{aligned}$$

and (d_1) follows. Now suppose that $\text{Re } \lambda = 0$ and $Aw = \lambda w + x$, and set $P_0 = ww^*$ in (13). Then

$$DP(t)D' = \left[1 + \int_0^t (w + sx)^* C' C (w + sx) ds \right]^{-1} D(w + tx)(w + tx)^* D'$$

and (d_2) easily follows. As for necessity of (d_3) , assume that $\text{Re } \lambda = 0$, $Cx = 0$ and (9) is satisfied but $Dx \neq 0$. Following the technique of [4], let us premultiply the first and second rows of (11) by z^* and y^* , respectively. Then

$$\begin{aligned} -z^* A' y + z^* C' C z &= \lambda z^* y, \\ y^* B B' y + y^* A z &= \lambda y^* z + y^* x. \end{aligned}$$

Therefore

$$z^* C' C z + y^* B B' y = y^* x \geq 0$$

and in fact

$$(14) \quad y^* x > 0$$

² The matrix inversion in (13) is justified by the relation $\det(I + YX) = \det(I + XY)$, which holds whenever both products are defined.

since otherwise $B'y = 0$, $Az = \lambda z + x$ and (d_2) implies $Dx = 0$. Noticing that (11) holds with z replaced by $\tilde{z} = z + \alpha x$, where α is any number, and in view of $y^* \tilde{z} = y^* z + \alpha y^* x$ and (14), we conclude that there is no loss of generality in assigning an arbitrary value to $y^* z$. Set $z^* y = 1$ and $P_0 = zz^*$. Applying Lemma 4, and using the notation introduced there, it follows that

$$\begin{aligned} [\Lambda_{11}(t) + \Lambda_{12}(t)P_0]y &= \Lambda_{11}(t)y + \Lambda_{12}(t)z = e^{\lambda t}y, \\ [\Lambda_{21}(t) + \Lambda_{22}(t)P_0]y &= \Lambda_{21}(t)y + \Lambda_{22}(t)z = e^{\lambda t} \left(z + tx + \frac{t^2}{2!} x^{j-1} + \cdots + \frac{t^j}{j!} x^1 \right), \end{aligned}$$

where $(x^1, x^2, \dots, x^{j-1}, x)$ is the characteristic subchain considered, and consequently

$$P(t)y = z + tx + \frac{t^2}{2!} x^{j-1} + \cdots + \frac{t^j}{j!} x^1.$$

It may happen that y does not belong to the complex $\text{Im } D'$, but the existence of u and v such that $y + u = D'v$ and $x^*u = 0$ is always guaranteed. Thus

$$\begin{aligned} v^*DP(t)D'v &= v^*DP(t)y + v^*DP(t)u \\ &= v^*D(z + tx) + (y^* + u^*)P(t)u \\ &= (y^* + u^*)(z + tx) + u^*P(t)u + \left(z + tx + \frac{t^2}{2!} x^{j-1} + \cdots + \frac{t^j}{j!} x^1 \right)^* u \\ &= y^*z + u^*z + z^*u + ty^*x + u^*P(t)u + \left(\frac{t^2}{2!} x^{j-1} + \cdots + \frac{t^j}{j!} x^1 \right)^* u. \end{aligned}$$

We shall show that the last term is zero, and the conclusion will follow according to (14). Since $u = y - D'v$ this term equals

$$\left(\frac{t^2}{2!} x^{j-1} + \cdots + \frac{t^j}{j!} x^1 \right)^* y.$$

Applying the relation $(A - \lambda I)x = x^{j-1}$ and the first row in (11), we get

$$y^* x^{j-1} = y^*(A - \lambda I)x = z^* C' C x = 0.$$

Similarly

$$y^* x^i = y^*(A - \lambda I)x^{i+1} = z^* C' C x^{i+1} = 0, \quad j - 2 \geq i \geq 1.$$

It is straightforward to check that (d_1) and (d_2) imply (c_1) , (c_2) and the existence of an \bar{A} -invariant complement to $N(\bar{C}, \bar{A}) \cap \mathcal{X}^0(\bar{A})$ in $\bar{\mathcal{X}}$. Thus A , C and D are representable as in (12), but $G' = (G'_1 \quad G'_2 \quad G'_3)$, where in general $G_2 \neq 0$. To complete the proof it is enough to verify the existence of a matrix S satisfying

$$(15) \quad SG_1 + G_2 = 0$$

and

$$(16) \quad SA_{11} - A_{22}S = 0,$$

since in that case the transformation

$$\begin{pmatrix} I & 0 & 0 \\ -S & I & 0 \\ 0 & 0 & I \end{pmatrix}$$

applied to the previously adopted basis produces a representation having all the properties of (12). Let $\operatorname{Re} \lambda = 0$ and $-A'_{22}y_2 = \lambda y_2$. Then there exists y_1, z_1, x_2 and x_3 such that

$$(17) \quad \begin{pmatrix} -A'_{11} & 0 & -A'_{31} & C'_1C_1 & 0 & 0 \\ 0 & -A'_{22} & -A'_{32} & 0 & 0 & 0 \\ 0 & 0 & -A'_{33} & 0 & 0 & 0 \\ G_1G'_1 & G_1G'_2 & G_1G'_3 & A_{11} & 0 & 0 \\ G_2G'_1 & G_2G'_2 & G_2G'_3 & 0 & A_{22} & 0 \\ G_3G'_1 & G_3G'_2 & G_3G'_3 & A_{31} & A_{32} & A_{33} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ 0 \\ z_1 \\ 0 \\ 0 \end{pmatrix} = \lambda \begin{pmatrix} y_1 \\ y_2 \\ 0 \\ z_1 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ x_2 \\ x_3 \end{pmatrix}.$$

This claim is justified by rewriting the first and fourth rows in the form

$$(18) \quad \begin{pmatrix} -A'_{11} - \lambda I & C'_1C_1 \\ G_1G'_1 & A_{11} - \lambda I \end{pmatrix} \begin{pmatrix} y_1 \\ z_1 \end{pmatrix} = \begin{pmatrix} 0 \\ -G_1G'_2y_2 \end{pmatrix}$$

and applying Lemma 5. Also, with the aid of the Jordan canonical form, it is verified that for properly selected z_2 and z_3 ,

$$\begin{pmatrix} A_{22} - \lambda I & 0 \\ A_{32} & A_{33} - \lambda I \end{pmatrix} \begin{pmatrix} x_2 + (A_{22} - \lambda I)z_2 \\ x_3 + A_{32}z_2 + (A_{33} - \lambda I)z_3 \end{pmatrix} = \begin{pmatrix} 0 \\ p \end{pmatrix},$$

where p is either a zero vector or a member of a characteristic chain of A_{33} associated with λ . Furthermore, (17) also holds with

$$\begin{pmatrix} z_1 \\ 0 \\ 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 \\ x_2 \\ x_3 \end{pmatrix}$$

replaced, respectively, by

$$\begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 \\ x_2 + (A_{22} - \lambda I)z_2 \\ x_3 + A_{32}z_2 + (A_{33} - \lambda I)z_3 \end{pmatrix}.$$

Since $\operatorname{Ker} D_2 = 0$, it follows by (d₃) that $x_2 + (A_{22} - \lambda I)z_2 = 0$. Consequently $y_2^*x_2 = -y_2^*(A_{22} - \lambda I)z_2 = 0$ and application of this result to the fifth row in (17) implies

$$(19) \quad y_2^*(G_2G'_1y_1 + G_2G'_2y_2) = 0.$$

The first row of (17) reads

$$(20) \quad y_1^*(A_{11} - \lambda I) = z_1^* C_1' C_1,$$

and in light of the fourth row in (17),

$$(21) \quad y_1^*(G_1 G_1^* y_1 + G_1 G_2' y_2) + z_1^* C_1' C_1 z_1 = 0.$$

Combining (21) with (19) we get

$$(22) \quad (y_1^* G_1 + y_2^* G_2)(G_1' y_1 + G_2' y_2) + z_1^* C_1' C_1 z_1 = 0,$$

therefore $y_1^* G_1 + y_2^* G_2 = 0$. This equation holds for each characteristic vector y_2 of A_{22} and corresponding y_1 , computed according to (18). Now suitable arrangement of the real and imaginary parts of the characteristic vectors of A_{22} yields a matrix S that satisfied (15). According to (22), $C_1 z_1 = 0$, and by (20), $y_1^* A_{11} = \lambda y_1^*$. This result assures that (16) is also satisfied.

The following conclusion is easily obtained with the aid of (6).

COROLLARY 1. *Let $P(\cdot)$ be determined by (2). Conditions (c_1) – (c_3) or, equivalently, conditions (d_1) – (d_3) imply that $DP(\cdot)D'$ is bounded on $[0, \infty)$ for every $P_0 \geq 0$.*

COROLLARY 2. *$P(\cdot)$ satisfying (10) is bounded on $[0, \infty)$ for every $P_0 \geq 0$ if and only if $A|_{\mathcal{N}(C, A) \cap \mathcal{X}^0(A)}$ is semisimple, $(C|_{\mathcal{N}(C, A) \cap \mathcal{X}^0(A)}, A|_{\mathcal{N}(C, A) \cap \mathcal{X}^0(A)})$ is detectable and there exists an A -invariant complement \mathcal{L} to $\mathcal{N}(C, A) \cap \mathcal{X}^0(A)$ such that $\text{Im } G \subset \mathcal{L}$.*

COROLLARY 3. *Let $P(\cdot)$ be determined by (10), and let x be a vector such that $Ax = \lambda x$ for some number λ . $P(\cdot)$ is bounded on $[0, \infty)$ for every $P_0 \geq 0$ if and only if the following three conditions hold:*

- (e₁) *if $\text{Re } \lambda > 0$ and $Cx = 0$, then $x = 0$;*
- (e₂) *if $\text{Re } \lambda = 0$, $Cx = 0$ and there exists a vector w such that $Aw = \lambda w + x$, then $x = 0$;*
- (e₃) *if $\text{Re } \lambda = 0$, $Cx = 0$ and there exist vectors y and z satisfying*

$$\begin{pmatrix} -A' & C'C \\ GG' & A \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} = \lambda \begin{pmatrix} y \\ z \end{pmatrix} + \begin{pmatrix} 0 \\ x \end{pmatrix},$$

then $x = 0$.

Appendix. Matrices and spaces in the sequel are defined over the complex field. For $t \geq 0$ let

$$\frac{dP}{dt} = AP + PA^* + GG^* - PC^*CP, \quad P(0) = P_0 \geq 0,$$

$$\frac{dQ}{dt} = (A + KC)Q + Q(A + KC)^* + GG^* + KK^*, \quad Q(0) = Q_0 \geq 0,$$

$$\frac{dR}{dt} = AR + RA^* + G_1 G_1^* - RC^*CR, \quad R(0) = R_0 \geq 0,$$

where the coefficients are constant. Then

$$\frac{d(Q - P)}{dt} = (A + KC)(Q - P) + (Q - P)(A + KC)^* + (PC^* + K)(CP + K^*)$$

and the next two results [5] follow immediately.

LEMMA 2. If $Q_0 \geq P_0$, then $Q(t) \geq P(t)$.

COROLLARY 4. If (C, A) is detectible, then $P(\cdot)$ is bounded on $[0, \infty)$.

A well-known technique [8] is to write

$$\begin{aligned} \frac{d(P - R)}{dt} &= (A - PC^*C)(P - R) + (P - R)(A - PC^*C)^* + GG^* - G_1G_1^* \\ &\quad + (P - R)C^*C(P - R). \end{aligned}$$

This equation yields the following statement.

LEMMA 3. If $GG^* \geq G_1G_1^*$ and $P_0 \geq R_0$, then $P(t) \geq R(t)$.

LEMMA 4. $P(t) = [\Lambda_{21}(t) + \Lambda_{22}(t)P_0][\Lambda_{11}(t) + \Lambda_{12}(t)P_0]^{-1}$, where $\Lambda_{ij}(\cdot)$ are obtained by partitioning

$$\exp \left[\begin{pmatrix} -A^* & C^*C \\ GG^* & A \end{pmatrix} t \right] = \begin{pmatrix} \Lambda_{11}(t) & \Lambda_{12}(t) \\ \Lambda_{21}(t) & \Lambda_{22}(t) \end{pmatrix}.$$

Indeed, $\Lambda_{11}(t) + \Lambda_{12}(t)P_0$ is the transition matrix of a differential equation [9] and is therefore invertible. Substitution verifies the validity of the expression for $P(t)$.

LEMMA 5.

$$\text{Im} \begin{pmatrix} 0 \\ G \end{pmatrix} \subset \text{Im} \begin{pmatrix} -A^* & C^*C \\ GG^* & A \end{pmatrix}.$$

Proof. Let x and y satisfy

$$\begin{pmatrix} -A & GG^* \\ C^*C & A^* \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = 0.$$

Premultiplication of the first and second rows by y^* and x^* , respectively, yields

$$\begin{aligned} -y^*Ax + y^*GG^*y &= 0, \\ x^*C^*Cx + x^*A^*y &= 0. \end{aligned}$$

Thus $x^*C^*Cx + y^*GG^*y = 0$, implying $G^*y = 0$. Consequently

$$\begin{pmatrix} 0 & G^* \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = 0.$$

REFERENCES

- [1] W. M. WONHAM, *Tracking and regulation in linear multivariable systems*, this Journal, 11 (1973), pp. 424-437.
- [2] S. P. BHATTACHARYYA, J. B. PEARSON AND W. M. WONHAM, *On zeroing the output of a linear system*, Information and Control, 20 (1972), pp. 135-141.
- [3] W. M. WONHAM, *Random differential equations in control theory*, Probabilistic Methods in Applied Mathematics, vol. 2, A. T. Bharucha-Reid, ed., Academic Press, New York, 1970, pp. 131-212.

- [4] V. K. KUČERA, *A contribution to matrix Riccati equations*, IEEE Trans. Automatic Control, AC-12 (1972), pp. 344–347.
- [5] W. M. WONHAM, *On a matrix Riccati equation of stochastic control*, this Journal, 6 (1968), pp. 681–697.
- [6] R. S. BUCY, *The Riccati equation and its bounds*, J. Comput. System Sci., 6 (1972), pp. 343–353.
- [7] J. SNYDERS AND M. ZAKAI, *On nonnegative solutions of the equation $AD + DA' = -C$* , SIAM J. Appl. Math., 18 (1970), pp. 704–714.
- [8] K. MÅRTENSSON, *On the matrix Riccati equation*, Information Sci., 3 (1971), pp. 17–49.
- [9] R. E. KALMAN AND R. S. BUCY, *New results in linear filtering and prediction theory*, Trans. ASME, J. of Basic Engineering, 83 (1961), pp. 95–108.

WEAK CONVERGENCE OF SET-VALUED FUNCTIONS AND CONTROL*

ZVI ARTSTEIN†

Abstract. Weak convergence on the space of integrably bounded set-valued functions is defined. Generalizations of results of weak convergence of real-valued integrable functions are obtained in the set-valued case. The results are applied to the characterization of the continuous dependence of the attainable set of a linear control system on the restraint set. We show that the weak convergence of the restraint set is a sufficient condition for the uniform convergence of the attainable set, and under the additional condition of uniform integrability the weak convergence is also a necessary condition.

1. Introduction. The analysis of set-valued functions has many applications in various mathematical fields and, in particular, in the study of linear control systems (see [5], [11], [12], [14], [16]). In this paper, we shall develop a theory of weak convergence of set-valued functions and apply it to characterize the continuous dependence of the attainable set of a linear control system on the restraint set.

Let T be an interval in the real line. For every t in T let $F(t)$ be a subset of the n -dimensional Euclidean space E_n . The integral $\int_U F(t) dt$ of the set-valued function F over the measurable subset U of T is defined by

$$\int_U F(t) dt = \left\{ \int_U f(t) dt : f \text{ is an integrable selector of } F \right\}.$$

Here a selector means a point-valued measurable function f such that $f(t) \in F(t)$ dt —almost everywhere. The definition was stated in this form by Aumann [3]. It is related to summation of sets as the Lebesgue integral of real-valued functions is related to summation of numbers. We shall use this point of view in the definition of the weak convergence in §4. Briefly speaking, the dual operation on a set-valued function is the operation of a bounded point-valued function (which is the typical dual element of L_1) on the set of integrable selectors of this set-valued function.

We shall give several equivalent definitions for the weak convergence. The main result is that the sequence F_k converges weakly to F if and only if the integrals $\int_U F_k$ converge in the Hausdorff metric to $\int_U F$ for every measurable subset of U of T . This theorem is analogous to the theorem for real-valued functions (see [7, Chap. IV, 8.7]). The theorem establishes the relationship between the convergence of set-valued functions and the convergence of their integrals. This has a direct implication to linear control systems of the form $dx/dt = A(t)x + B(t)u$. A (well-known) substitution can be done in order to show that the attainable set of the system is a “nice” image of the integral of a certain set-valued function which is closely related to the restraint set-valued function of the system. Thus we show (in §6 below) that weak convergence of the restraints implies the uniform convergence of the attainable set, and under the additional assumption of uniform

* Received by the editors January 7, 1974.

† Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. This research was supported by the Office of Naval Research under Contract NONR N00014-67-A-0191-0009 and the National Science Foundation under Contract GP 28931X2.

integrability the weak convergence is also a necessary condition. This gives a characterization of the continuous dependence of the attainable set on the restraint set.

In less technical terms our conditions show what kind of an error in the description of the constraints is allowed in order to assure that the resulting error in the attainable set will be small. We shall also discuss the similar question for the dependence of a trajectory of the system on the control function.

The theory of weak convergence will be developed for set-valued functions which are Borel-measurable and which have closed values. We shall see in §2 that these assumptions are not as restrictive as it seems, as long as integrals are concerned.

In the applications to control theory, t in the interval T stands for "time" and T has the structure of an interval in the real-time line. During the rest of the paper we shall only use the nonatomicity of T . All the results are valid for a general nonatomic finite measure space.

The paper is organized as follows. In §2, we give our notations and some preliminary results. The space IB of integrably bounded set-valued functions is defined in §3. The weak convergence is investigated in §4. The convex and compact sets can be embedded in a Banach space and the integration theory for convex-valued functions can be regarded as integration into this space (see [6], [2]). In §5 we shall compare our weak convergence with the "natural" weak convergence for convex-valued functions arising by this embedding. The last two sections deal with the applications to control theory. In §6, we shall discuss the dependence of the attainable set on the restraint set and in §7 we treat the continuous dependence of the solution to the control equation on the controls.

2. Notations and preliminaries. The scalar product of the two vectors $x = (x^1, \dots, x^n)$ and $p = (p^1, \dots, p^n)$ in E_n is $\sum_{i=1}^n x^i p^i$ and is denoted by $p \cdot x$. The norm $|x|$ of x is its Euclidean norm, i.e., $|x| = (x \cdot x)^{1/2}$. If $p \in E_n$ and A is a subset of E_n , then $p \cdot A = \{p \cdot a : a \in A\}$ is a subset of the real line. The distance $\delta(A, B)$ of the two subsets A and B of E_n is their Hausdorff distance, i.e., $\delta(A, B) = \max(\sup_A \inf_B |a - b|, \sup_B \inf_A |a - b|)$. The distance function δ is a semi-metric. If only closed and bounded subsets of E_n are concerned, then δ is a metric and the space of closed and bounded subsets is complete with respect to δ . Bounded subsets of this space are precompact (see [9]). The norm of a subset A of E_n is denoted $|A|$ and defined by $|A| = \delta(A, \{0\})$ which is equal to $\max\{|a| : a \in A\}$. The support function $s(p, A)$ of a subset A of E_n is defined for every $p \in E_n$ by $s(p, A) = \sup\{p \cdot a : a \in A\}$. It is a convex function (see [9]). If A and B are convex sets, then

$$\delta(A, B) = \sup\{|s(p, A) - s(p, B)| : |p| = 1\}.$$

This equality, which is easily verified, goes back to Minkowski. A simple consequence of it is that if A_k , $k = 1, 2, \dots$, are convex sets in E_n , then $\delta(A_k, A)$ tends to zero if and only if A is convex and $s(p, A_k)$ tends to $s(p, A)$ for every $p \in E_n$.

A set-valued function F is called *Borel-measurable* if for every closed subset C of E_n the set $F^-(C) = \{t : F(t) \cap C \neq \emptyset\}$ is a Borel subset of T . The graph of F is the subset of $T \times E_n$ defined by the graph $F = \{(t, x) : x \in F(t)\}$. A set-valued function with closed values is Borel-measurable if and only if its graph is a Borel

subset of $T \times E_n$. The “only if” direction was proved by Debreu [6, (3.4)]; see also [17, Thm. 2]. The “if” part (which is not widely known) follows from Novikoff [15].

A selector of F is a vector-valued measurable function such that $f(t)$ belongs to $F(t)$ for almost every t .

PROPOSITION 2.1. *If F is a set-valued function with a Borel-measurable graph, then a selector of F exists.*

The proposition is true also for functions with analytic graphs and the proof is due to von Neumann (see [3, p. 3]). Aumann [4] gave a generalization to abstract measurable spaces. For another approach see [13].

PROPOSITION 2.2. *If F has a Borel-measurable graph and if $p(t)$ is a measurable E_n -valued function, then $s(p(t), F(t)) = \sup \{p(t) \cdot x : x \in F(t)\}$ is a measurable function.*

The proof is implied by the following observation:

$$\begin{aligned} \{t : s(p(t), F(t)) \geq \alpha\} &= \text{proj}_T \{(t, x) : p(t) \cdot x \geq \alpha, x \in F(t)\} \\ &= \text{proj}_T (\{t, x) : p(t) \cdot x \geq \alpha\} \cap \text{graph } F), \end{aligned}$$

where proj_T means projection on T .

COROLLARY 2.3. *If F has a Borel-measurable graph, then the functions $s(p, F(t))$ for $p \in E_n$ and $m(t) = |F(t)|$ are measurable functions.*

The statement for the support function $s(p, F(t))$ is a particular case of Proposition 2.2. The function $m(t)$ is measurable since $m(t) = \sup_j s(p_j, F(t))$, where $\{p_j\}_{j=1}^\infty$ is a dense sequence in $\{p : |p| = 1\}$.

PROPOSITION 2.4. *Suppose that F has a Borel-measurable graph and that $\int_U F$ is not empty. Then $\int_U s(p, F(t)) = s(p, \int_U F)$ for every $p \in E_n$.*

The result is widely known. It is an easy consequence of Proposition 2.1 and Corollary 2.3. A complete proof can be found in [2, Lemma 2.2].

PROPOSITION 2.5. *If F has closed values and a measurable graph, then $m(t) = |F(t)|$ is integrable implies that $\int_U F$ is closed for every measurable $U \subset T$.*

The result was proved by Aumann [3, Thm. 4].

PROPOSITION 2.6. *$\int_U F$ is a convex set for every measurable subset U of T .*

This result is an easy consequence of Lyapunov's convexity theorem. The proof for this case is due to Richter (see [3, Thm. 1]).

As was noted in the Introduction, we shall develop the theory of weak convergence for set-valued functions which have closed values and are Borel-measurable. The theorem which follows this paragraph shows that if closedness is required then without loss of generality the functions are Borel-measurable (or equivalently have a Borel-measurable graph). Thus there is no need to develop the theory for general closed-valued functions. The “without loss of generality” above corresponds to the fact that we integrate the functions. If one drops the closedness assumption, then one has to deal with the semimetric δ , where the equivalence classes of δ are sets with the same closure. Since any “reasonable” measurability condition, for instance the analyticity of the graph, leads to $\int \text{closure } F(t) \subset \text{closure } \int F$ it follows that it is enough to deal with closed-valued functions.

THEOREM 2.7. *Let F be a set-valued function with closed values. Then there exists a Borel-measurable set-valued function G with closed values, such that $G(t) \subset F(t)$ a.e. and such that if f is a selector of F then f is also a selector of G . In particular, $\int_U F = \int_U G$ for every measurable subset U of T .*

Proof. Denote \mathcal{F} the collection of selectors of F . The equivalence classes under equality a.e. of measurable functions form a separable metric space under convergence in measure. A metric can be found in [10, §42, exercise 4], while the separability follows easily from the separability of T . Let f_k be a dense subset of the Borel-measurable functions in \mathcal{F} . Define $H(t) = \{f_1(t), f_2(t), \dots\}$. It is easily seen that H is Borel-measurable. Define $G(t) = \text{closure } H(t)$. In view of [17, Lemma 5.2], G is also Borel-measurable. Obviously, $G(t) \subset F(t)$ for almost every t . Let f be in \mathcal{F} . Then a subsequence of f_k converges in measure to f . Therefore, a sub-subsequence is converging to f a.e. and hence f is a selector of G .

3. The space IB . A set-valued function F is integrably bounded if there exists an integrable real-valued function $m(t)$ such that for almost every t if $x \in F(t)$ then $|x| \leq m(t)$. For a set-valued function F with a Borel-measurable graph, Corollary 2.3 implies that F is integrably bounded if and only if $m(t) = |F(t)|$ is integrable.

The space IB consists of all the integrably bounded set-valued functions that are Borel-measurable and have closed values. Instead of Borel-measurable we could require that the graph of a member in IB will be a Borel-measurable set. See the second paragraph of §2.

We shall deal with equivalence classes of members of IB where the equivalence is defined by equality almost everywhere, and as usual we shall not distinguish between a function and its equivalence class. We shall define a metric on IB which will be analogous to the L_1 metric on integrable functions, and get similar results. A complete analogy cannot be achieved since the space of compact subsets in E_n is not a vector space.

DEFINITION AND PROPOSITION 3.1. For the two elements F and G in IB define

$$d(F, G) = \int_T \delta(F(t), G(t)) dt.$$

Then d is a metric on IB .

Proof. First we have to show that d is well-defined. Let \mathcal{F} and \mathcal{G} be the collections of selectors of F and G respectively. Let $\{f_k\}_{k=1}^\infty$ and $\{g_k\}_{k=1}^\infty$ be two dense-in-convergence-in-measure sequences in \mathcal{F} and \mathcal{G} . (See the proof of Theorem 2.7.) From the definition of the Hausdorff metric δ it follows that a.e.

$$\delta(F(t), G(t)) = \max \left(\sup_k \inf_j |g_k(t) - f_j(t)|, \sup_j \inf_k |g_k(t) - f_j(t)| \right).$$

Since each of $|g_k(t) - f_j(t)|$ is measurable it follows that $\delta(F(t), G(t))$ is measurable. The latter is integrable since it is dominated by the integrable function $2 \max(|F(t)|, |G(t)|)$, and therefore d is well-defined. Since δ is a metric it follows that d is a metric.

PROPOSITION 3.2. The space IB with the metric d is a complete space. If the sequence F_k in IB converges in the metric d to F , then there is a subsequence of F_k which converges to F almost everywhere where the convergence is with respect to δ .

Proof. If F_k is a Cauchy sequence, then one can find a subsequence, say F_m , such that $d(F_m, F_{m+1}) \leq 1/2^{m+1}$. For this subsequence it is easy to show that almost everywhere $\delta(F_m(t), F_{m+k}(t))$ converges to zero as $m \rightarrow \infty$, i.e., $F_m(t)$ is a.e. a δ -Cauchy sequence. Since δ is complete it follows that there is a.e. a set $F(t)$ such that $\delta(F_m(t), F(t)) \rightarrow 0$. It is easy to show that $d(F_k, F)$ converges to zero.

Some other results similar to what is known about L_1 -spaces can be proved. We shall need in the sequel the concept of uniform integrability which is also adopted from the L_1 motivation.

DEFINITION 3.3. The collection F_α of elements in IB is *uniformly integrable* if for every $\varepsilon > 0$ there is an $\eta > 0$ such that if the Lebesgue measure of U is smaller than η then $|\int_U F_\alpha| \leq \varepsilon$ for every F_α in the collection.

PROPOSITION 3.4. *The collection F_α is uniformly integrable if and only if the collection $m_\alpha(t) = |F_\alpha(t)|$ of integrable real-valued functions is uniformly integrable.*

Proof. Suppose that m_α is a uniformly integrable collection. Let $\eta(\varepsilon)$ be given for the collection m_α . Let f be a selector of F . Then $|\int_U f| \leq \int_U |f| \leq \int_U m_\alpha$ and therefore $\eta(\varepsilon)$ is valid also for the collection F_α . This proves the "if" part. Suppose now that m_α is not uniformly integrable. Therefore, an $\varepsilon > 0$ exists such that for every $\eta > 0$ there is a U with measure $\leq \eta$ and there is an α such that $\int_U m_\alpha \geq \varepsilon$. Let f_α be a selector of the corresponding F_α such that $|f_\alpha(t)| = m_\alpha(t)$. Such a selector exists in view of Proposition 2.1 and Corollary 2.3. Let U_1, \dots, U_{2^n} be the partition of U such that the values of f_α restricted to each U_i are in the same orthant of E_n . Then the measure of each U_i is still $\leq \eta$, but at least for one i it is true that $|\int_{U_i} f_\alpha| \geq \varepsilon/2^n$. This contradicts the existence of $\eta = \eta(\varepsilon 2^{-n})$ for establishing the uniform integrability of F_α .

COROLLARY 3.5. *The collection F_α is uniformly integrable if and only if the collection of all selectors of all its members is uniformly integrable.*

4. Weak convergence in IB . The following theorem gives the equivalence of three conditions which will serve as the definition of weak convergence. Recall from §2 that the support function $s(p, A)$ is defined by $s(p, A) = \sup \{p \cdot a : a \in A\}$.

THEOREM 4.1. *Let F_k be a sequence of elements in IB and let F belong to IB . The following three statements are equivalent.*

(i) *For every bounded and measurable E_n -valued function $p(t)$ the sequence $\int_T p(t) \cdot F_k(t)$ of sets in E_1 converges to $\int_T p(t) \cdot F(t)$.*

(ii) *For every bounded and measurable E_n -valued function $p(t)$ the sequence $\int_T s(p(t), F_k(t))$ of real numbers converges to $\int_T s(p(t), F(t))$.*

(iii) *For every $p \in E_n$ and measurable subset U of T the sequence $\int_U s(p, F_k(t))$ converges to $\int_U s(p, F(t))$.*

Proof. (i) \Rightarrow (ii). If $\int_T p(t) \cdot F_k(t)$ converges to $\int_T p(t) \cdot F(t)$, then, in particular, the sequence of numbers $a_k = \sup (\int_T p(t) \cdot F_k(t))$ converges to the number $a = \sup (\int_T p(t) \cdot F(t))$. We shall show that $a_k = \int_T s(p(t), F_k(t))$ and that

$$a = \int_T s(p(t), F(t)),$$

and this will complete the proof. Proposition 2.2 shows that $s(p(t), F(t))$ is measurable; thus we can change it on a null set and get a Borel-measurable function, say $s(t)$. Let

$$F'(t) = \{x \in F(t) : p(t) \cdot x = s(t)\}.$$

Then F' has a Borel-measurable graph and by Proposition 2.1 a selector $f(t)$ of F' exists. Obviously,

$$\int_T s(p(t), F(t)) = \int_T p(t) \cdot f(t).$$

On the other hand, $p(t) \cdot x(t) \geq r$ for every $r \in p(t) \cdot F(t)$; therefore $\int_T p(t) \cdot f(t) = a$.

The claim for a_k is proved by adding the subscript k whenever F and a appear in the proof.

(i) \Rightarrow (iii). Obvious.

(iii) \Rightarrow (i). Condition (iii) implies that the sequence of real-valued functions $s(p, F_k(t))$ is weakly convergent to $s(p, F(t))$ (see [7, IV.8.7]). (The arguments in [7] require the boundedness in L_1 of the sequence, but it is easy to verify this in our case where the measure space has a finite measure.) Therefore, $s(p, F_k(t))$ are uniformly integrable [7, IV.8.9]. Let p_1, \dots, p_{2n} be the $2n$ vectors $(0, \dots, 0, \pm 1, 0, \dots, 0)$. Then the collection of functions $s(p_i, F_k(t))$ where $i = 1, \dots, 2n, k = 1, 2, \dots$, is a uniformly integrable collection. Since the support function $s(p, A)$ is convex in p and since $|F_k(t)| = \sup \{(x/|x|) \cdot x : x \in F_k(t)\}$ it follows that $|F_k(t)| \leq c \{\sup p_i \cdot x : x \in F_k(t), i = 1, \dots, 2n\}$ for a certain fixed c ($c = n^{1/2}$ will fit). Thus the collection $m_k(t) = |F_k(t)|$ is uniformly integrable. In particular, $\int_T m_k(t)$ is bounded, say by M .

Let $p(t)$ be a measurable and bounded E_n -valued function. For each $\varepsilon > 0$ we can find an E_n -valued function $q(t)$, such that q has only a finite number of values, q is measurable and $|p(t) - q(t)| < \varepsilon$. Then

$$\begin{aligned} \delta\left(\int_T p(t) \cdot F_k(t), \int_T q(t) \cdot F_k(t)\right) &\leq \left|\int_T (p(t) - q(t)) \cdot F_k(t)\right| \\ &\leq \int_T |p(t) - q(t)| |F_k(t)| \leq \varepsilon M. \end{aligned}$$

Condition (iii) implies that for every U and a fixed vector p the sets $\int_U p \cdot F_k(t)$ converge to $\int_U p \cdot F(t)$. Indeed, $\int_U p \cdot F_k(t)$ is the interval whose extremes are given by $\int_U s(-p, F_k(t))$ and $\int_U s(p, F_k(t))$ while $\int_U p \cdot F(t)$ is the interval with extremes $\int_U s(-p, F(t))$ and $\int_U s(p, F(t))$. Since $q(t)$ has only a finite number of values also $\int_U q(t) \cdot F_k(t)$ converges to $\int_U q(t) \cdot F_k(t)$. Since ε was arbitrarily small the inequalities above show the desired convergence also for $p(t)$.

DEFINITION. If for the sequence F_k of elements in IB and for $F \in IB$ condition (i) (and hence also (ii) and (iii)) of Theorem 4.1 holds, we say that the sequence F_k converges weakly to F .

Remark. Note that for singleton-set-valued functions, the definition coincides with the usual definition of weak convergence of point-valued functions.

Remark 4.2. Each of (i), (ii) and (iii) can be stated in terms of Cauchy sequences (of $\int_T p(t) \cdot F_k(t)$, $\int_T s(p(t), F_k(t))$ and $\int_U s(p, F_k(t))$ respectively) rather than in terms of convergence. It is also clear that the proof shows the equivalence also in this presentation. Therefore, we have also the concept of a *weak Cauchy sequence*, namely a sequence of elements such that one of (i), (ii) or (iii) (and hence the other two) is satisfied in the Cauchy formulation. We shall see below that every weak Cauchy sequence has a weak limit.

Remark 4.3. Notice that the limit of a weak convergent sequence is not unique. Indeed, the support functions of two sets with the same convex hulls are the same and the convergence criterion (iii) is given in terms of the support functions. Thus, if F_k converges weakly to F , then it converges weakly to every G provided the convex hull of $G(t)$ equals a.e. the convex hull of $F(t)$. However, it is clear that the convex-valued limit is unique. Indeed, the support function $s(p, F(t))$ of the limit

is the weak limit in L_1 of the functions $s(p, F_k(t))$ and in L_1 the weak limit is unique.

The last paragraph of the proof of Theorem 4.1 contains the proof to the following proposition.

PROPOSITION 4.4. *If F_k converges weakly to F or if F_k is a weak Cauchy sequence, then the functions $m_k(t) = |F_k(t)|$ are uniformly integrable.*

COROLLARY 4.5. *A weakly convergent sequence is bounded in IB and is uniformly integrable. Furthermore, the collection of all selectors of the members of the sequence is uniformly integrable.*

Proof. The uniform integrability of $|F_k(t)|$ where F_k is the sequence implies the boundedness. The rest of the claims follow from Proposition 3.4 and Corollary 3.5.

THEOREM 4.6. *Let F_k be a sequence in IB and let F belong to IB . The sequence F_k converges weakly to F if and only if for every measurable subset U of T the sets $\int_U F_k$ converge in the δ -metric to $\int_U F$.*

Proof. Proposition 2.6 implies that the sets $\int_U F_k$ and $\int_U F$ are convex. In this case the convergence of $\int_U F_k$ to $\int_U F$ is equivalent to the convergence of $s(p, \int_U F_k)$ to $s(p, \int_U F)$ for every $p \in E_n$; see the discussion about the support function in §2. By Proposition 2.4 the equalities

$$s\left(p, \int_U F_k\right) = \int_U s(p, F_k(t)) \quad \text{and} \quad s\left(p, \int_U F\right) = \int_U s(p, F(t))$$

hold, but the convergence of the right-hand side of these equalities is the definition via condition (iii) of the weak convergence.

Remark 4.7. It is clear that in terms of Cauchy sequences the last proposition should be read as F_k is a weak Cauchy sequence if and only if for every U the sequence of sets $\int_U F_k$ is a δ -Cauchy sequence.

The following propositions give some results which are analogous to the properties of weak convergences in L_1 (see [7, IV.8]).

PROPOSITION 4.8. *The space IB is weakly sequentially complete, i.e., every weak Cauchy sequence has a weak limit.*

Proof. Suppose that F_k is a weak Cauchy sequence. In view of Remark 4.7 for every measurable $U \subset T$ the sets $\int_U F_k$ form a δ -Cauchy sequence. The metric δ is complete and therefore $\int_U F_k$ has a limit, denoted by $\Phi(U)$. Since $\int_U F_k$ is additive in U it follows that $\Phi(U)$ is additive too. Since the F_k are uniformly integrable (see Propositions 4.4 and 3.4) it follows that $\Phi(U)$ is countably additive and bounded, i.e., it is a set-valued measure in the terminology of [1]. Since Φ has closed values it follows from [1, Thm. 9.1] that Φ has a Borel-measurable Radon–Nikodym derivative with closed values, i.e., a Borel-measurable set-valued function F such that for every $U \subset T$ measurable $\int_U F = \Phi(U)$. In view of Theorem 4.6, F is a weak limit of F_k .

PROPOSITION 4.9. *A set in IB is weakly sequentially precompact, i.e., every sequence has a weakly convergent subsequence, if and only if its members are uniformly integrable.*

Proof. If the set is not uniformly integrable, then a sequence can be found in it such that every subsequence of it is not uniformly integrable and hence (Corollary 4.5) no subsequence converges weakly.

Suppose the set is uniformly integrable and let F_k be a sequence in it. Let U_1, U_2, \dots be a sequence of Borel sets in T that are dense in the Borel field of T . The sequence $\int_{U_1} F_k$ is bounded, and therefore (see §2) has a convergent subsequence, say $\int_{U_1} F_{l_1}$. The sequence $\int_{U_2} F_{l_1}$ is bounded, and thus has a convergent subsequence, say $\int_{U_2} F_{m_1}$. We can continue building these sub-sub- \dots -subsequences, and by a standard diagonal procedure get a subsequence F_{i_j} of F_k such that for every U_j the sequence $\int_{U_j} F_{i_j}$ for $j = 1, 2, \dots$ will converge. The uniform integrability together with the density of U_j implies that the $\int_U F_{i_j}$ converge for every U . In view of Remark 4.7 the sequence F_{i_j} is a weak Cauchy sequence, and therefore (Proposition 4.8) has a limit.

The following three propositions give the connection between the weak convergence of the set-valued functions and the weak convergence of their selectors, as integrable functions.

PROPOSITION 4.10. *If F_k converges weakly to F and if f_k is a selector of F_k for $k = 1, 2, \dots$, then the sequence f_k has a weakly convergent subsequence.*

Proof. F_k converges weakly implies (Corollary 3.5) that the f_k are uniformly integrable and therefore a weakly convergent subsequence exists.

PROPOSITION 4.11. *If F_k converges weakly to F and if f_k is a sequence of selectors of F_k for $k = 1, 2, \dots$, such that f_k converges weakly to f , then a.e. $f(t)$ belongs to the convex hull of $F(t)$.*

Proof. For every $p \in E_n$ the inequality $p \cdot f_k(t) \leq s(p, F_k(t))$ holds. Since $s(p, F_k(t))$ converges weakly to $s(p, F(t))$ and since $p \cdot f_k(t)$ converges weakly to $p \cdot f(t)$ it follows that almost everywhere and for every p the inequality $p \cdot f(t) \leq s(p, F(t))$ holds. (We can deduce "almost everywhere for every p ..." from "for every p , almost everywhere ..." since it is enough to show the inequality for a dense sequence of vectors.) Since for every set A the convex hull of A is $\{a: p \cdot a \leq s(p, A) \text{ for every } p \in E_n\}$ it follows that $f(t)$ a.e. belongs to the convex hull of $F(t)$.

PROPOSITION 4.12. *Let F_k converge weakly to F . Let f be a selector of F . Then there exists a sequence f_k of selectors of F_k , $k = 1, 2, \dots$, such that f_k converges weakly to f .*

Proof. Without loss of generality $T = [0, 1]$. Let $[(j-1)2^{-m}, j2^{-m}]$ for $j = 1, \dots, 2^m$ be the m th dyadic partition of $[0, 1]$. There are 2^m intervals in the partition, and denote them by U_1, \dots, U_{2^m} . Since the F_k converge weakly to F it follows from Theorem 4.6 that for k large enough, say $k \geq k(m)$, the δ distance between $\int_{U_j} F_k$ and $\int_{U_j} F$ is less than $1/m2^m$ for every $j = 1, \dots, 2^m$. Therefore a selector $f_{k,m}$ of F_k for $k \geq k(m)$ exists such that

$$\left| \int_{U_j} f_{k,m} - \int_{U_j} f \right| \leq \frac{1}{m2^m} \quad \text{for every } j = 1, \dots, 2^m.$$

Notice that by this choice if U is a union of members of the m th partition then

$$\left| \int_U f_{k,m} - \int_U f \right| \leq \frac{1}{m}.$$

Without loss of generality the sequence $k(m)$ for $m = 1, 2, \dots$ is strictly increasing. Define now the selector f_k of F_k by $f_k = f_{k,m}$ if $k(m) \leq k \leq k(m+1)$. It is clear that

for every finite union U of dyadic intervals the sequence $\int_U f_k$ converges to $\int_U f$. The uniform integrability of f_k (see Corollary 4.5) implies that $\int_U f_k$ converges to $\int_U f$ for every measurable U and hence by [7, IV.8] f_k converges weakly to f .

5. Weak convergence via embedding. The space of compact subsets of E_n cannot be embedded in a vector space although summation and multiplication by a nonnegative scalar are defined on it by $A + B = \{a + b : a \in A, b \in B\}$ and $\alpha A = \{\alpha a : a \in A\}$. The reason is that the cancellation law is not valid, for instance, $[0, 1] + \{0, 1\} = [0, 1] + [0, 1]$. If only convex sets are concerned, the embedding can be done and for instance the mapping that associates every set with its support function is an embedding into the continuous functions on the compact set $\{p : |p| = 1\}$ with the sup norm. As was noted in §2, this embedding goes back to Minkowski. Thus the set-valued functions with convex and compact values can be regarded as vector-valued functions into a Banach space. This approach was established by G. Debreu [6] who showed that the integration of these set-valued functions can be done as integration into the Banach space (see also [2]). For integrable functions into a Banach space the concept of weak convergence is given by the classical theory. The purpose of this section is to show that the weak convergence for the embedded functions coincides with the weak convergence defined in this paper.

In order to show this, we will need a precise representation of the embedding and we shall use the one mentioned above, i.e., a convex and compact set B is associated with the restriction of the support function $s(p, B)$ to $K = \{p : |p| = 1\}$. Thus the Banach space is $C = C(K)$, the continuous real-valued functions on K . Its dual C^* is the space of regular measures μ on K (see [7, IV. 6.3]), where the dual operation is

$$\mu(h) = \int_K h(p) d\mu(p).$$

Denote the integrable functions from T into C by $L_1(T, C)$. The dual space of $L_1(T, C)$ is $L_\infty(T, C^*)$, the bounded functions into C^* (see [8, 8.18.1 and 8.18.2]). (I want to thank H. T. Banks for this reference.) The embedding of a convex compact set in C is an isometry; therefore the embedding of the convex-valued elements of IB in $L_1(T, C)$ is also an isometry, and every set-valued function such that its embedding is integrable (i.e., belongs to $L_1(T, C)$) is in IB . Finally, we shall use the fact that bounded collections of compact sets are precompact in the Hausdorff metric.

THEOREM 5.1. *Let F and F_k for $k = 1, 2, \dots$, be set-valued functions with convex and compact values. Suppose that they belong to IB (or equivalently that they belong to $L_1(T, C)$). Then the sequence F_k converges weakly to F in IB if and only if F_k w-converges to F in $L_1(T, C)$.*

Proof. The “if” part. Let $p(t)$ be a measurable and bounded E_n -valued function. It is easy to verify that the operation $\int_T s(p(t), G(t)) dt$ is linear on the members of IB . It is also bounded since

$$\left| \int_T s(p(t), G(t)) dt \right| \leq \sup_t |p(t)| \cdot \int_T |G(t)| dt.$$

Therefore, if F_k converges weakly to F in $L_1(T, C)$ then $\int_T s(p(t), F_k(t)) dt$ converges to $\int_T s(p(t), F(t)) dt$, and in view of condition (ii) of Theorem 4.1, F_k w -converges to F in IB .

The "only if" part. Suppose that F_k w -converges to F in IB . Then F_k is uniformly integrable (Corollary 4.5) and therefore $m_k(t) = |F_k(t)|$ is a uniformly integrable sequence (Proposition 3.4). In particular, if $\eta > 0$ is given, then a number M can be found such that for every k if $T_k = \{t: m_k(t) \leq M\}$ then the Lebesgue measure of T_k is less than η . Consider the embedding of the convex and compact sets in E_n with norm less than or equal to M and denote this collection of functions by D . The set D is a compact set in $C(K)$ and therefore equicontinuous [7, IV.6.7]. Let K_1, \dots, K_l be a partition of K and let p_1, \dots, p_l be points in K_1, \dots, K_l respectively. If the maximum diameter of the K_j is small and if μ is a regular measure on K , then the functional $\sum_{j=1}^l \mu(K_j) f(p_j)$ is an approximation to $\int_K f(p) d\mu(p)$ and the approximation is uniform with respect to all members of D . This is a direct consequence of the equicontinuity.

We have to show that for every bounded and measurable C^* -valued function μ_t the numbers $\int_T \int_K s(p, F_k(t)) d\mu_t(p) dt$ converge to $\int_T \int_K s(p, F(t)) d\mu_t(p) dt$. We proceed with the following approximations. Let $\varepsilon > 0$. Since F_k is uniformly integrable it follows that an $\eta > 0$ exists such that if the Lebesgue measure of T_k is less than η then

$$\left| \int_{T_k} \int_K s(p, F_k(t)) d\mu_t(p) \right| \leq \varepsilon \sup \|\mu_t\|.$$

For this η let M be given by the preceding discussion. For this M let K_1, \dots, K_l and p_1, \dots, p_l be given such that $\sum_{j=1}^l \mu_t(K_j) f(p_j)$ is a D -uniform ε -approximation to $\int_K f(p) d\mu_t(p)$ for a t -set with complement that has Lebesgue measure less than η . The uniform integrability implies that on this complement, and denote it by T' , the dt -integral of $\int_K s(p, F_k(t)) d\mu_t(p)$ is less than $\varepsilon \sup \|\mu_t\|$. The weak convergence of F_k in IB implies that the

$$(5.2) \quad \int_{T \setminus (T_k \cup T')} \sum_{j=1}^l \mu_t(K_j) s(p_j, F_k(t)) \quad \text{converge to} \\ \int_{T \setminus (T_k \cup T')} \sum_{j=1}^l \mu_t(K_j) s(p_j, F(t)).$$

Recall that we have to show that the

$$(5.3) \quad \int_T \int_K s(p, F_k(t)) d\mu_t(p) dt \quad \text{converge to} \quad \int_T \int_K s(p, F(t)) d\mu_t(p) dt.$$

Since the integral on $T_k \cup T'$ is less than $2\varepsilon \sup \|\mu_t\|$, and since the summation in (5.2) ε -approximates the inner integral in (5.3) and since ε is arbitrarily small we conclude that (5.3) holds.

6. Continuous dependence of the attainable set of a linear control system. Consider the linear control equation

$$(6.1) \quad \frac{dx}{dt} = A(t)x + B(t)u,$$

where $x \in E_n$, $u \in E_m$, the matrix-valued functions $A(t)$ and $B(t)$ are measurable with $A(t)$ integrable on bounded intervals and $B(t)$ bounded on bounded intervals. The admissible control functions are the selectors of a restraint set-valued function $\Omega(t)$ and we shall assume that Ω has closed values and is integrably bounded on bounded intervals.

For every admissible control $\mathbf{u} = u(t)$ there is a unique solution to (6.1) with the initial condition $x(t_0) = x_0$. Denote it by $x(t, \mathbf{u})$. By using the variation of parameters formula we can get an explicit form of $x(t, \mathbf{u})$, namely,

$$x(t, \mathbf{u}) = X(t)x_0 + X(t) \int_{t_0}^t X^{-1}(\tau)B(\tau)u(\tau) d\tau,$$

where $X(t)$ is the fundamental solution to the homogeneous equation $dx/dt = A(t)x$ and $X(t_0)$ is the identity matrix.

The attainable set $K(t)$ consists of all the vectors $z = x(t, \mathbf{u})$, where \mathbf{u} is an admissible control, i.e., all the vectors to which x_0 can be steered in the time t . The variation of parameters formula shows that

$$K(t) = \left\{ X(t)x_0 + X(t) \int_{t_0}^t X^{-1}(\tau)B(\tau)u(\tau) dt : \mathbf{u} = u(\tau) \text{ is an admissible control} \right\}.$$

Motivated by this formula, we define a set-valued function F by

$$(6.2) \quad F(\tau) = \{X^{-1}(\tau)B(\tau)u : u \in \Omega(\tau)\}.$$

The following equality gives the representation of $K(t)$ in terms of integration of the set-valued function F . The result is widely known and used but since the proof of it is short, we shall give it here.

$$K(t) = X(t)x_0 + X(t) \int_{t_0}^t F(\tau) d\tau.$$

In order to prove this identity notice that if u is an admissible control then obviously $\int_{t_0}^t X^{-1}(\tau)B(\tau)u(\tau) d\tau$ belongs to $\int_{t_0}^t F$. In order to show the converse direction let $f(t)$ be a selector of F and without loss of generality assume f is Borel-measurable. We will show the existence of an admissible \mathbf{u} such that a.e. $f(\tau) = X^{-1}(\tau)B(\tau)u(\tau)$. Define

$$\Omega'(\tau) = \{u \in \Omega(\tau) : f(\tau) = X^{-1}(\tau)B(\tau)u\}.$$

Then Ω' has a.e. nonempty values. Moreover, its graph is the intersection of the graph Ω with the inverse image of $\{0\}$ by the function $(\tau, u) \rightarrow X^{-1}(\tau)B(\tau)u - f(\tau)$. Since without loss of generality both f and B are Borel-measurable it follows that the graph of Ω' is Borel-measurable and by Proposition 2.1 a selector \mathbf{u} of Ω' exists. This \mathbf{u} is the desired admissible control.

We showed that the attainable set $K(t)$ is a translation (by $X(t)x_0$) of a regular image (by $X(t)$) of the integral of the set-valued function F . This substitution enables us to translate the theory of calculus of set-valued functions into the theory of linear control systems. In the sequel we shall apply the theory of weak convergence which we developed in this paper to the problem of continuous dependence of the attainable set $K(t)$ on the restraint set $\Omega(t)$.

We consider the equation (6.1) with the fixed initial value $x(t_0) = x_0$. We shall see what happens to the attainable set if the restraint set is changed. If Ω or Ω_k are restraint set-valued functions for the system (6.1), then the corresponding set-valued functions F or respectively F_k are defined by (6.2). We shall consider the system for a finite time interval $T = [t_0, t_1]$ and all the functions $A(t)$, $B(t)$ etc. are assumed to be defined only on T .

PROPOSITION 6.3. *If Ω_k converges weakly to Ω , then F_k converges weakly to F . The converse is also true provided that the Ω_k are uniformly integrable and that $B(\tau)$ is almost everywhere one-to-one.*

Proof. Suppose that Ω_k converges weakly to Ω . Let $p(\tau)$ be a bounded measurable E_n -valued function. Define $q(\tau) = p(\tau)X^{-1}(\tau)B(\tau)$. Then $q(\tau)$ is a bounded measurable E_m -valued function. The weak convergence of Ω_k implies that $\int_T q(\tau) \cdot \Omega_k(\tau)$ converges to $\int_T q(\tau) \cdot \Omega(\tau)$, and the equalities

$$\int_T q(\tau) \cdot \Omega_k(\tau) = \int_T p(\tau) \cdot F_k(\tau)$$

show that (i) of Theorem 4.1 holds also for F_k . This proves the first statement.

Suppose now that F_k converges weakly to F , that the Ω_k are uniformly integrable and that $B(\tau)$ is one-to-one. We have to show that for every bounded $q(\tau)$ the integrals $\int_T q(\tau) \cdot \Omega_k(\tau)$ converge to $\int_T q(\tau) \cdot \Omega(\tau)$, but the uniform integrability implies that it is enough to show that the $\int_{U_j} q(\tau) \cdot \Omega_k(\tau)$ converge to $\int_{U_j} q(\tau) \cdot \Omega(\tau)$ for an increasing sequence U_j such that $\bigcup_{j=1}^{\infty} U_j = T$. Denote $B^{-1}(\tau)$ an inverse of $B(\tau)$. (Since $B(\tau)$ is one-to-one it follows that it has an inverse defined on its range and $B^{-1}(\tau)$ is an extension of this inverse to all E_n .) Define $p(\tau) = q(\tau)B^{-1}(\tau)X(\tau)$ and define $U_j = \{\tau: |p(\tau)| \leq j\}$ for $j = 1, 2, \dots$. Then $\bigcup_{j=1}^{\infty} U_j = T$. Also for U_j the weak convergence of F_k implies that the

$$\int_{U_j} p(\tau) \cdot F_k(\tau) = \int_{U_j} q(\tau) \cdot \Omega_k(\tau) \quad \text{converge to} \quad \int_{U_j} p(\tau) \cdot F(\tau) = \int_{U_j} q(\tau) \cdot \Omega(\tau)$$

and this completes the proof.

Remark. Neither the uniform integrability nor that $B(\tau)$ is one-to-one can be removed from the conditions of the proposition.

In the following, we shall consider several restraint set-valued functions to the equation (6.1). The attainable set that corresponds with the restraint Ω_j will be denoted K_j . (We changed the index from k to j in order to avoid the inconvenient K_k symbol.)

THEOREM 6.4. *If Ω_j converges weakly to Ω , then $K_j(t)$ converges to $K(t)$ uniformly in $t \in T$. If the Ω_j are uniformly integrable and if $B(\tau)$ is a.e. one-to-one, then the convergence of $K_j(t)$ to $K(t)$ for every t implies that Ω_k converges weakly to Ω .*

Proof. The fundamental matrix $X(t)$ is continuous and has a continuous inverse; therefore it is enough to prove the theorem for the set-valued function $L(t) = \int_{t_0}^t F(\tau) d\tau$ instead of the attainable set $K(t)$. If Ω_j converges weakly, then by Proposition 6.3 the sequence F_j converges weakly to F and Theorem 4.6 implies that $L_j(t)$ converges to $L(t)$ for every t . The uniform convergence follows from the observation that $\delta(L_j(t), L_j(\tau)) = |\int_t^\tau F_j|$ and that the sequence F_j is uniformly integrable.

Suppose now that the Ω_j are uniformly integrable. Since $F_j(\tau) = X^{-1}(\tau)B(\tau) \cdot \Omega_j(\tau)$ and since $X(\tau)$ and $B(\tau)$ are bounded it follows that the F_j are also uniformly integrable. If $L_j(t)$ converges to $L(t)$ for every t , it follows that for every interval $[t, \tau]$ the sets $\int_t^\tau F_j$ converge to $\int_t^\tau F$ and together with the uniform integrability this implies that the F_j converge weakly to F . If, in addition, $B(\tau)$ is almost everywhere one-to-one, then Proposition 6.3 implies that the sequence Ω_j converges weakly to Ω .

COROLLARY. *If the Ω_j are uniformly integrable and if $K_j(t)$ converges to $K(t)$ for every t , then the convergence of the attainable sets K_j is uniform in $t \in T$.*

Remark. In many applications the restraint set-valued function Ω is described in terms of inequalities $q_j(t) \cdot u \leq \alpha_j(t)$ for $j = 1, \dots, k$, where the q_j are E_m -valued functions and $\alpha_j(t)$ are real-valued functions. (In many applications Ω is a constant polygon, i.e., q_j and α_j are constants.) Theorem 6.4 gives the answer to the following question. What errors in the description of the $\alpha_j(t)$ are allowed in order to have only small errors in the attainable set. Criterion (iii) for the weak convergence implies that the "right" changes are small changes with respect to the weak L_1 -topology. By "right" we mean that under the additional assumption of uniform integrability the weak L_1 -convergence is also a necessary condition, i.e., the weak L_1 -topology is the weakest topology with respect to which the attainable set is still continuous.

7. Continuous dependence of the trajectories on the controls. In the particular case where $\Omega(t) = \{u(t)\}$ is a singleton, Theorem 6.4 gives a characterization of the continuous dependence of the trajectory (the solution) $x(t, u)$ on the control function $u = u(t)$. It is worthwhile to formulate it again for this case.

THEOREM 7.1. *If the sequence u_k of controls converges weakly to the control function u , then the solutions $x(t, u_k)$ converge to $x(t, u)$ uniformly on T . Under the additional assumptions that the u_k are uniformly integrable and $B(\tau)$ is a.e. one-to-one, the convergence of $x(t, u_k)$ to $x(t, u)$ for every t implies that u_k converges weakly to u .*

The first part of the theorem is known and was used in the literature (see [15]). It is easy to give direct proofs to both the statements of the theorem. The common tool which is used in the literature is the continuous dependence of the solution on the control functions when the latter are endowed with the weak topology of L_2 . The tool which is suggested by Theorem 7.1, namely the weak topology of L_1 , has the "advantage" that it is also a necessary condition. In particular, convergence in weak L_2 implies convergence in L_1 but not vice versa.

Remark. Theorem 7.1 does not hold for linear systems of the form $dx/dt = A(x) + \phi(t, u)$, i.e., when the control does not appear linearly. Indeed, in general the functional $\int_T \phi(u(t)) dt$ is not weakly continuous.

REFERENCES

- [1] Z. ARTSTEIN, *Set-valued measures*, Trans. Amer. Math. Soc., 165 (1972), pp. 103–125.
- [2] ———, *On the calculus of closed set-valued functions*, Indiana Univ. Math. J., 24 (1974), pp. 433–441.
- [3] R. J. AUMANN, *Integrals of set-valued functions*, J. Math. Anal. Appl., 12 (1965), pp. 1–12.
- [4] ———, *Measurable utility and the measurable choice theorem*, Proc. Internat. Colloq. La Decision, C.N.R.S. Aix-en-Provence, 1967, pp. 15–26.
- [5] T. F. BRIDGLAND, JR., *Trajectory integrals of set-valued functions*, Pacific J. Math., 33 (1970), pp. 43–68.

- [6] G. DEBREU, *Integration of correspondences*, Proc. Fifth Berkeley Symposia on Mathematical Statistics and Probability, vol. II, Univ. of California Press, Berkeley, 1967, pp. 351–372.
- [7] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Part I*, Interscience, New York, 1958.
- [8] R. E. EDWARDS, *Functional Analysis, Theory and Applications*, Holt, Rinehart and Winston, New York, 1965.
- [9] H. G. EGGELSTON, *Convexity*, Cambridge University Press, Cambridge, 1958.
- [10] P. R. HALMOS, *Measure Theory*, Van Nostrand, Princeton, N.J., 1950.
- [11] H. HERMES, *Calculus of set-valued functions and control*, J. Math. Mech., 18 (1968), pp. 47–60.
- [12] H. HERMES AND J. P. LASALLE, *Functional Analysis and Time Optimal Control*, Academic Press, New York, 1969.
- [13] L. J. HIMMELBERG, M. Q. JACOBS AND F. S. VAN VLECK, *Measurable multifunctions, selectors, and Filippov's implicit functions lemma*, J. Math. Anal. Appl., 25 (1969), pp. 276–284.
- [14] M. Q. JACOBS, *On the approximation of integrals of multivalued functions*, this Journal, 7 (1969), pp. 158–177.
- [15] P. NOVIKOFF, *Sur les projections de certains ensembles mesurables B*, C. R. Acad. Sci. URSS, 23 (1939), pp. 864–865.
- [16] C. OLECH, *Convexity in existence theory of optimal solutions*, Proc. International Congress of Mathematicians, vol. III, Nice, 1970, pp. 187–192.
- [17] R. T. ROCKAFELLAR, *Measurable dependence of convex sets and functions on parameters*, J. Math. Anal. Appl., 28 (1969), pp. 4–25.

GLOBAL BILINEARIZATION OF SYSTEMS WITH CONTROL APPEARING LINEARLY*

JAMES TING-HO LO†

Abstract. A necessary and sufficient condition that a nonlinear system, with control appearing linearly, be dynamically equivalent to an observable bilinear system is derived. When the condition is satisfied, a procedure to construct an observability canonical form of such a bilinear system is provided in the proof of the sufficiency part of the condition.

1. Introduction. In this paper, we will be concerned with the following nonlinear control system with control (u, v) appearing linearly: for $t \in [0, T]$ $= T \subset R_1$,

$$(1) \quad \begin{aligned} \dot{x}(t) &= f(x(t)) + G(x(t))u(t), \\ z(t) &= h(x(t)) + Q(x(t))v(t), \end{aligned}$$

where x (state), z (output), u , and v (inputs) are n , p , m , and q -dimensional vector-valued functions of time, respectively; f and h are n , p -dimensional vector-valued functions of $x(t)$; and G and Q are matrix-valued functions of $x(t)$ of appropriate sizes.

Controllability and optimal control problems for this system when $z = x$ have been studied in literature, e.g., Lee and Markus [13], Hermes and Haynes [9], Hermes [10], Haynes [7], Haynes and Hermes [8]. The main result of this paper is a necessary and sufficient condition for this system to have a dynamically equivalent bilinear system of the following form: for $t \in T$,

$$(2) \quad \begin{aligned} \dot{y}(t) &= \left(A + \sum_{i=1}^m B_i u_i(t) \right) y(t), \\ z(t) &= \left(C + \sum_{i=1}^q D_i v_i(t) \right) y(t), \end{aligned}$$

where A, B_i, C, D_i , are constant matrices of appropriate sizes, and for some positive integers $M_i, i = 0, \dots, q$,

$$(3) \quad \begin{aligned} \text{rank } [C', A'C', \dots, (A')^{M_0-1}C', D'_1, A'D'_1, \dots, (A')^{M_1-1}D'_1, \\ \dots, D'_q, A'D'_q, \dots, (A')^{M_q-1}D'_q] = \dim A. \end{aligned}$$

Bilinear systems have been extensively studied in recent years for three primary reasons. First, it has been shown that bilinear systems are feasible mathematical models for large classes of problems of practical importance. Second, bilinear systems provide higher order approximations to nonlinear systems than do linear systems. (Linear systems are special cases of bilinear systems.) Third, bilinear systems have rich geometric and algebraic structures which promise a

* Received by the editors March 1, 1974, and in revised form June 20, 1974.

† Division of Mathematics and Physics, University of Maryland, Baltimore County, Baltimore, Maryland 21228. This work was supported in part by a UMBC Summer Research Fellowship and by the USAF Office of Scientific Research under Grant AFOSR-74-2671.

fruitful field of research. We refer the interested readers to Mohler and Ruberti [16], d'Alessandro et al. [6], Brockett et al. [2], Mayne and Brockett [4], Mohler [17]. Most related articles can be traced from these references.

As a final remark, we note that some necessary and sufficient conditions were given in Krener [12] for two systems of the form (1) with $z = x$ to be locally equivalent.

Some of the results on stochastic systems, which are analogous to some of those given here, can be found in Lo [15].

2. A necessary and sufficient condition. In a recent paper by Brockett [3], it was reported that bilinear systems are capable of representing a wide variety of highly nonlinear models. Motivated by his results, we will in this section derive a necessary and sufficient condition for a nonlinear system (1) to have the same input-output mapping as does a bilinear system (2) and (3). When the condition is satisfied, a procedure to construct such a bilinear system is provided in the proof of the sufficiency part of the condition.

To initiate the mathematical development, several definitions are in order. The reader is referred to Kalman et al. [11] and Brockett [1] for terminologies unspecified here.

DEFINITION. Given an initial state $x(0) = x_0$ and an input function (u, v) on T , the control system (1) produces a corresponding output function z on T . Thus for an initial state x_0 , the system (1) defines an *input-output mapping*. Two control systems are said to be *dynamically equivalent* if, for appropriate initial states, they have the same input-output mappings.

Remarks. We note that we do not take into consideration the notion of the minimal dynamical state in the above definition, as opposed to the common practice in defining system isomorphism in the linear system theory (Kalman et al. [11]).

DEFINITION. Let L be the differentiation operator defined by

$$L(g(x)) = \sum_{i=1}^n f_i(x) \frac{\partial}{\partial x_i} (g(x)) = g_x(x) f(x),$$

for a differentiable scalar function $g(x)$, where $g_x(x)$ is the gradient (a row vector) of g . When g is a vector function, $L(g)$ denotes applying L to each component of g and g_x denotes taking the gradient of each component of g . We note that $L(g)$ is then a vector function and g_x a matrix function.

If h is infinitely differentiable, the set of functions $\{h(x(t)), Lh(x(t)), L^2h(x(t)), \dots\} \cup [\cup_{i=1}^q \{Q_i(x(t)), LQ_i(x(t)), L^2Q_i(x(t)), \dots\}]$, where Q_i denotes the i th column of Q , is called the *sensor orbit* of the system (1) at the time t (whatever the input is).

Remark. The notion of sensor orbits for stochastic systems was introduced in Bucy and Joseph [5, p. 62] to deduce a set of suboptimal filtering equations. Its application to optimal estimation was discussed in Lo [14]. In defining sensor orbits for stochastic systems, the differentiation operator corresponding to L above is a Kolmogorov backward operator.

DEFINITION. The system (1) is said to have a *finite-dimensional sensor orbit*, if there exist integers M_i , $i = 0, \dots, q$, such that for $k = 1, \dots, q$ and all state trajectories $x(t)$, $t \in T$,

$$L^{M_0}h(x(t)) = \sum_{i=0}^{M_0-1} A(0, 0, i+1)L^i h(x(t)) + \sum_{j=1}^q \sum_{i=0}^{M_q-1} A(0, j, i+1)L^i Q_j(x(t)),$$

$$L^{M_k}Q_k(x(t)) = \sum_{i=0}^{M_0-1} A(k, 0, i+1)L^i h(x(t)) + \sum_{j=1}^q \sum_{i=0}^{M_q-1} A(k, j, i+1)L^i Q_j(x(t)),$$

where $A(i, j, k)$ are constant $p \times p$ matrices, and every column of $(L^i h(x(t)))_x G(x(t))$ and $(L^i Q_j(x(t)))_x G(x(t))$, $i = 0, \dots, M_j - 1$, $j = 1, \dots, q$, lies on the sensor orbit, i.e., for $k = 1, \dots, m$, and all state trajectories $x(t)$, $t \in T$, the k th column of

$$(L^i h(x(t)))_x G(x(t)) = \sum_{j=1}^{M_0} B_0(k, i, 0, j)L^{j-1}h(x(t)) + \sum_{j=1}^q \sum_{l=1}^{M_q} B_0(k, i, j, l)L^{l-1}Q_j(x(t)),$$

the k th column of

$$(L^i Q_j(x(t)))_x G(x(t)) = \sum_{l=1}^{M_0} B_j(k, i, 0, l)L^{l-1}h(x(t)) + \sum_{r=1}^q \sum_{l=1}^{M_q} B_j(k, i, r, l)L^{l-1}Q_r(x(t)).$$

We are now in a position to state our main result.

THEOREM. The nonlinear system (1), with control appearing linearly, is dynamically equivalent to the bilinear system (2) and (3), if and only if the nonlinear system has a finite-dimensional sensor orbit.

When this condition is satisfied, a procedure for constructing a dynamically equivalent bilinear system in an observability canonical form is suggested in the following proof of sufficiency.

Proof. Sufficiency. Set $y_{0i}(t) = L^{i-1}h(x(t))$ and

$$y_{ji}(t) = L^{i-1}Q_j(x(t)).$$

By the chain rule of differentiation, for $i = 1, \dots, M_j$, $j = 1, \dots, q$,

$$\dot{y}_{ji} = y_{ji+1}(t) + (L^{i-1}Q_j(x(t)))_x G(x(t))u(t),$$

$$\dot{y}_{0i}(t) = y_{0i+1}(t) + (L^{i-1}h(x(t)))_x G(x(t))u(t).$$

Since the system (1) has a finite-dimensional sensor orbit, for $j = 0, \dots, q$,

$$\dot{y}_{jM_j}(t) = \sum_{k=0}^q \sum_{i=1}^{M_q} A(j, k, i)y_{ki}(t) + \sum_{i=1}^m \sum_{k=0}^q \sum_{l=1}^{M_k} B_j(i, M_j, k, l)y_{kl}(t)u_i(t)$$

and, for $r = 1, \dots, M_j - 1$, $j = 0, \dots, q$,

$$\dot{y}_{jr}(t) = y_{jr+1}(t) + \sum_{i=1}^m \sum_{k=0}^q \sum_{l=1}^{M_k} B_j(i, r, k, l)y_{kl}(t)u_i(t),$$

where $A(j, k, i)$ and $B_j(i, r, k, l)$ are constant matrices. Let

$$y = [y'_{01}, \dots, y'_{0M_0}, \dots, y'_{q1}, \dots, y'_{qM_q}]'.$$

Simple observation yields the first equation of (2), where

$$\begin{aligned}
 A &= [A_{ij}], \\
 A_{ii} &= \begin{bmatrix} O & I & & \\ & \ddots & \ddots & \\ & & O & I \\ A(i, i, 1) & A(i, i, 2) & \cdots & A(i, i, M_i) \end{bmatrix} \quad \text{for } i = 0, \dots, q, \\
 A_{ij} &= \begin{bmatrix} O \\ A(i, j, 1) & A(i, j, 2) & \cdots & A(i, j, M_j) \end{bmatrix} \quad \text{for } i, j = 0, \dots, q \text{ and } i \neq j, \\
 B_i &= \begin{bmatrix} B_0(i, 0) & B_0(i, 1) & \cdots & B_0(i, q) \\ B_1(i, 0) & B_1(i, 1) & \cdots & B_1(i, q) \\ \vdots & & & \vdots \\ B_q(i, 0) & \cdots & \cdots & B_q(i, q) \end{bmatrix}, \\
 B_j(i, k) &= \begin{bmatrix} B_j(i, 1, k, 1) & B_j(i, 1, k, 2) & \cdots & B_j(i, 1, k, M_k) \\ B_j(i, 2, k, 1) & \cdots & \cdots & B_j(i, 2, k, M_k) \\ \vdots & & & \vdots \\ B_j(i, M_k, k, 1) & \cdots & \cdots & B_j(i, M_k, k, M_k) \end{bmatrix}.
 \end{aligned}$$

It follows immediately from the second equation of (1) that the second equation of (2) holds, where

$$\begin{aligned}
 C &= [I \quad O \quad \cdots \quad O], \\
 D &= \underbrace{[O \quad \cdots \quad O]}_{\left(\sum_{j=0}^{i-1} M_j\right)p} \quad I \quad O \quad \cdots \quad O].
 \end{aligned}$$

Straightforward calculation shows that (3) is true. This completes the proof of sufficiency.

Necessity. As the bilinear and the nonlinear systems are dynamically equivalent, we have, for all control functions (u, v) ,

$$z(t) = h(x(t)) + Q(x(t))v(t) = \left(C + \sum_{i=1}^p D_i v_i(t) \right) y(t).$$

Setting $v = 0$, we have for all state trajectories $x(t)$, $t \in T$,

$$(4) \quad h(x(t)) = Cy(t).$$

Setting $v_i = 1$, we have for all state trajectories $x(t)$, $t \in T$,

$$(5) \quad Q_i(x(t)) = D_i y(t), \quad i = 1, \dots, p.$$

Differentiating (4) and (5) with respect to time, we obtain, in view of (2), for $i = 2, 3, \dots$, and $k = 1, \dots, q$,

$$\begin{aligned} L^i h(x(t)) &= CA^i y(t), \\ (L^{i-1} h(x(t)))_x G(x(t)) &= CA^{i-1} [B_1 y(t), \dots, B_m y(t)], \\ L^i Q_k(x(t)) &= D_k A^i y(t), \\ (L^{i-1} Q_k(x(t)))_x G(x(t)) &= D_k A^{i-1} [B_1 y(t), \dots, B_m y(t)]. \end{aligned}$$

By the Cayley–Hamilton theorem, there exist constants $c_i, i = 0, \dots, N$, such that $c_N \neq 0$ and $\sum_{i=0}^N c_i A^i = 0$. Hence for $k = 1, \dots, q$,

$$(6) \quad \begin{aligned} \sum_{i=0}^N c_i L^i h(x(t)) &= \sum_{i=0}^N c_i CA^i y(t) = 0, \\ \sum_{i=0}^N c_i L^i Q_k(x(t)) &= \sum_{i=1}^N c_i D_k A^i y(t) = 0. \end{aligned}$$

Consider the j th column, say $CA^i B_j y(t)$, of $(L^i h(x(t)))_x G(x(t))$. Because of (3), the k th column of $B_j(A')^i C'$ can be expressed as

$$\sum_{l=1}^{M_0} (A')^{l-1} C' B'_0(j, i, 0, l, k) + \sum_{r=1}^q \sum_{l=1}^{M_r} (A')^{l-1} D'_r B'_0(j, i, r, l, k),$$

for some constant p -vectors $B'_0(j, i, r, l, k)$. Hence

$$B_j(A')^i C' = \sum_{l=1}^{M_0} (A')^{l-1} C' B'_0(j, i, 0, l) + \sum_{r=1}^q \sum_{l=1}^{M_r} (A')^{l-1} D'_r B'_0(j, i, r, l),$$

where $B_0(j, l, r, i) = [B'_0(j, i, r, l, 1), \dots, B'_0(j, i, r, l, p)]'$ and

$$(7) \quad CA^i B_j y(t) = \sum_{l=1}^{M_0} B_0(j, i, 0, l) L^{l-1} h(x(t)) + \sum_{r=1}^q \sum_{l=1}^{M_r} B_0(j, i, r, l) L^{l-1} Q_r(x(t)).$$

Similarly, the k th column of $(L^i Q_j(x(t)))_x G(x(t))$,

$$D_j A^{i-1} B_k y(t) = \sum_{l=1}^{M_0} B_j(k, i, 0, l) L^{l-1} h(x(t)) + \sum_{r=1}^q \sum_{l=1}^{M_q} B_j(k, i, r, l) L^{l-1} Q_r(x(t)),$$

for some constant matrices $B_j(k, i, r, l)$. This together with (6) and (7) completes the proof of necessity.

In the following, we will look at an interesting special case of this theorem for which the proof follows directly from the theorem and is omitted.

Let V_1, \dots, V_r and W_1, \dots, W_s be vector spaces over the same field, and let the maps

$$\Phi_r: V_1 \times \dots \times V_r \rightarrow W_1,$$

$$\Phi_r: V_1 \times \dots \times V_r \rightarrow W_1 \times \dots \times W_s$$

be r -linear, i.e., for all α and β in the field, and all $i = 1, 2, \dots, r$,

$$\begin{aligned} & \Phi_r(v_1, \dots, \alpha v_i + \beta v_i^*, \dots, v_r) \\ &= \alpha \Phi_r(v_1, \dots, v_i, \dots, v_r) + \beta \Phi_r(v_1, \dots, v_i^*, \dots, v_r), \\ & \Phi_r(v_1, \dots, \alpha v_i + \beta v_i^*, \dots, v_r) \\ &= \alpha \Phi_r(v_1, \dots, v_i, \dots, v_r) + \beta \Phi_r(v_1, \dots, v_i^*, \dots, v_r) \end{aligned}$$

respectively. We can now state a corollary, which is more general than Theorem 2, of Brockett [3].

COROLLARY. *The nonlinear system,*

$$\begin{aligned} \dot{x}(t) &= \left(F + \sum_{i=1}^m G_i u_i(t) \right) x(t), \\ z(t) &= \sum_{r=1}^p \Phi_r(x(t), \dots, x(t)) + \sum_{r=1}^q \Phi_r(x(t)) v(t), \end{aligned}$$

where F and G_i are constant matrices, has a finite-dimensional sensor orbit and is dynamically equivalent to the bilinear system,

$$\begin{aligned} \dot{y}(t) &= \left(A + \sum_{i=1}^m B_i u_i(t) \right) y(t), \\ z(t) &= \left(C + \sum_{i=1}^s D_i v_i(t) \right) y(t), \end{aligned} \tag{8}$$

where A, B_i, C, D_i are some constant matrices of appropriate sizes and, for some positive integers $M_i, i = 0, \dots, q$, (3) holds.

Remark. A way to construct a bilinear system (8) is to follow the procedure given in Theorem 1. An alternative way is to employ an idea used in Theorem 2 of Brockett [3]. We may set $y' = [x^{[1]}, x^{[2]}, \dots, x^{[\max(p,q)]}]$, where $x^{[1]} = x$, $x^{[2]} = [x_1^2, x_1 x_2, \dots, x_2^2, x_2 x_3, \dots, x_2 x_{\max(p,q)}, \dots, x_{\max(p,q)}^2]$, \dots , etc. Then by straightforward calculation and substitution, a dynamically equivalent bilinear system can be obtained. This way is somewhat simpler than the procedure given in Theorem 1. However, it sometimes leads to a nonobservable bilinear system (when $(u, v) = 0$), as can be seen easily from the scalar system $\dot{x}(t) = ax(t)$, $z(t) = x^2(t)$.

REFERENCES

- [1] R. W. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.
- [2] R. W. BROCKETT ET AL., *Differential geometric methods in control*, Tech. Rep. 628, Div. of Engineering and Applied Physics, Harvard Univ., Cambridge, Mass., 1971.
- [3] R. W. BROCKETT, *On the algebraic structure of bilinear systems*, Proc. Conf. on Variable Structure Systems, Academic Press, New York, 1972.
- [4] D. Q. MAYNE AND R. W. BROCKETT, eds., *Geometric Methods in System Theory*, D. Reidel, Boston, Mass., 1973.
- [5] R. S. BUCY AND P. D. JOSEPH, *Filtering for Stochastic Processes with Application to Guidance*, John Wiley, New York, 1968.
- [6] P. D'ALESSANDRO, A. ISIDORI AND R. RUBERTI, *Lectures on Bilinear System Theory*, Springer-Verlag, New York, 1972.

- [7] G. HAYNES, *On the optimality of a totally singular vector control: An extension of the Green's theorem approach to higher dimensions*, this Journal, 4 (1966), pp. 662–677.
- [8] G. HAYNES AND H. HERMES, *Nonlinear-controllability via Lie theory*, this Journal, 8 (1970), pp. 450–460.
- [9] H. HERMES AND G. HAYNES, *On the nonlinear control problem with control appearing linearly*, this Journal, 1 (1963), pp. 85–108.
- [10] H. HERMES, *Controllability and the singular problem*, this Journal, 2 (1965), pp. 241–260.
- [11] R. E. KALMAN, P. L. FALB AND M. A. ARBIB, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1969.
- [12] A. J. KRENER, *On the equivalence of control systems and the linearization of nonlinear systems*, this Journal, 11 (1973), pp. 670–676.
- [13] E. B. LEE AND L. MARKUS, *Optimal control for nonlinear processes*, Arch. Rational Mech. Anal., 8 (1961), pp. 36–58.
- [14] J. T. LO, *Finite dimensional sensor orbits and optimal nonlinear filtering*, IEEE Trans. Information Theory, IT-18 (1972), pp. 583–588.
- [15] ———, *Bilinear stochastic systems and finite dimensional sensor orbits*, Math. Res. Rep. 74-2, Div. of Mathematics and Physics, Univ. of Maryland Baltimore County, Baltimore, Md., 1974.
- [16] R. R. MOHLER AND A. RUBERTI, eds., *Theory and Application of Variable Structure Systems*, Academic Press, New York, 1972.
- [17] R. R. MOHLER, *Bilinear Control Processes*, Academic Press, New York, 1973.

STOCHASTIC CONTROL OF ROTATIONAL PROCESSES WITH ONE DEGREE OF FREEDOM*

JAMES TING-HO LO† AND ALAN S. WILLSKY‡

Abstract. A class of bilinear stochastic control problems involving single-degree-of-freedom rotation is formulated and resolved. Both synchronization control and orientation control are considered. In each case, the measurement data is first processed through a nonlinear transformation. The transformed process then goes through an ordinary estimator, such as a Kalman–Bucy filter. After another nonlinear processing of the output of the ordinary estimator, the desired optimal control is yielded. A generalization of the approach illustrated by these results to control problems on arbitrary Abelian Lie groups is included.

1. Introduction. In this paper we will study several classes of stochastic control problems associated with single-degree-of-freedom rotation. As we shall see, the relevant state and sensor dynamic equations are bilinear in nature.

In the past, such stochastic control problems have been studied strictly in a vector space setting. While such techniques have been most useful in the study of linear systems, these methods have not yielded closed form optimal synthesis techniques for large classes of nonlinear systems, such as the bilinear systems considered here.

It is the purpose of this paper to use an alternative technique to the vector space approach. The motivation for this is to study the bilinear equations of interest with the aid of algebraic and analytical tools that are as natural to these problems as the vector space methods are to the linear problems. In this sense, one should view the present work as being motivated not only by the failure of vector space theory to handle some nonlinear problems adequately, but also by the success of vector space theory in effectively utilizing the structure of linear systems.

Very recently, the theory of Lie groups and Lie algebras has been successfully applied to a number of bilinear systems problems. Specifically, the results of Wei and Norman [11], [12] on differential equations, Brockett [1], Sussman, and Jurdjevic [5] on the structures of bilinear control systems, and Lo and Willsky [8] on estimation of rotational processes with one degree of freedom indicate that, much as in the theory of linear systems, the differential geometric structure of some bilinear systems may be used to obtain simple, explicit solutions. It is in this spirit that this paper is written.

Specifically, we will concern ourselves with the study of stochastic processes on the circle, S^1 , and its extensions to higher dimensions. Topics such as FM modulation, frequency stability, single-degree-of-freedom gyroscopic analysis, and satellite attitude control are well-known examples in this framework.

* Received by the editors June 25, 1973, and in revised form June 20, 1974.

† Division of Mathematics, University of Maryland, Baltimore, Maryland 21228. The work of this author was supported in part by the U.S. Office of Naval Research under the Joint Services Electronics Program by Contract N00014-67-A-0298-0006, while he was with the Division of Engineering and Applied Physics, Harvard University, Cambridge, Massachusetts.

‡ Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139. The work of this author was done in part while he was a Fannie and John Hertz Foundation Fellow in the Department of Aeronautics and Astronautics, M.I.T.

In the next section, a class of stochastic control problems on the unit circle will be formulated. The state and sensor dynamic processes are constructed by taking the projection modulo 2π of the corresponding typical 1-dimensional processes. The stochastic differential equations which govern their evolution are bilinear in form. The control function and the observational noise can be viewed as entering multiplicatively.

In §3, we will briefly discuss two kinds of control criteria on the circle, namely synchronization control and orientation control criteria. An effective optimal control procedure for each of these two kinds of control problems will then be deduced with the aid of the optimal estimation schemes derived in Lo and Willsky [8]. In each case, the measurement data is first processed through a nonlinear transformation. The transformed process then goes through an ordinary estimator, such as a Kalman–Bucy filter. After another nonlinear processing of the output of the ordinary estimator, the desired control is yielded. The approach illustrated by these results can be extended to a large class of problems—those involving processes evolving on Abelian Lie groups. This will be discussed at the end of §3.

Section 2 is relatively abstract, since it describes the mathematical setting of the problems to be considered. The authors wish to point special attention to §3, in which we explicitly solve several nonlinear stochastic control problems.

The reader is referred to Lo and Willsky [9] for some examples, which illustrate the application of results in this paper to a number of important practical problems. Among them are a control problem of the synchronous rotation of a prime mover in a hydraulic plant, a feedback frequency modulation problem, and a satellite attitude control problem.

2. Stochastic control systems. In this section, we will formulate a stochastic model of a control system for continuous rotational processes with one degree of freedom. This model consists of equations for the state and the sensor dynamics.

A natural state space for single-degree-of-freedom rotational processes is the circle group, S^1 . It has been shown (Ito and McKean [4]) that the circular Brownian motion on S^1 can be constructed by taking the projection modulo 2π of the standard 1-dimensional Brownian motion onto the unit circle S^1 . This method will now be used to construct the continuous state and sensor dynamics to be used in this paper.

We will adopt the following notation:

- (Ω, \mathcal{A}, P) = a probability space;
- s = a positive real number;
- C_1^s = the family of real-valued continuous functions, a , on $[0, s]$ such that $a(0) = 0$;
- C_2^s = the family of 2×2 orthogonal-matrix-valued continuous functions, A , on $[0, s]$ such that $A(0) = I$, the identity matrix;
- B_i^s = the Borel σ -field of C_i^s with respect to the uniform topology of C_i^s , for $i = 1$ and 2 .

Lower case letters denote elements in C_1^s and upper case letters denote elements in C_2^s .

Let $J: C_1^s \rightarrow C_2^s$ be defined by

$$(1) \quad \begin{aligned} (J(a))(t) &= \exp(a(t)R) = \begin{bmatrix} \cos a(t) & \sin a(t) \\ -\sin a(t) & \cos a(t) \end{bmatrix}, \\ R &= \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \end{aligned}$$

for $a \in C_1^s$ and $t \in [0, s]$. It is easily seen that J is B_1^s -measurable and bijective. This bijective operator will play a key role in this paper. The reader is referred to Lo and Willsky [8] for a physically appealing argument concerning the bijectivity of J .

Thus a continuous stochastic process Y on S^1 corresponds to a continuous stochastic process y on R in the sense that one can be induced by the other via the bijective operator J . In the following we will refer to y as the C_1 -representation and Y as the C_2 -representation of the continuous stochastic process under consideration.

2.1. State dynamic equations. We first formulate a state dynamic equation on S^1 as the following scalar Ito differential equation viewed as its C_1 -representation:

$$(2) \quad \begin{aligned} dx &= a(t) dt + F(t)x(t) dt + G(t)u(t) dt + Q^{1/2}(t) dw(t), \\ x(0) &= 0, \end{aligned}$$

where a , F , G , and $Q^{1/2}$ are scalar functions, w is a standard Brownian motion on (Ω, \mathcal{A}, P) , and u is the scalar control function. In considering rotational processes with one degree of freedom, the dynamic state of the process is specified by the 2×2 orthogonal matrix representation of the process. In this sense, the C_1 -representation above is *not* a dynamic-state-space representation of the process. Injecting x into S^1 via the operator J , we obtain, with the aid of the Ito differential rule, the following Ito matrix differential equation satisfied by the C_2 -representation $X = J(x)$. This is a dynamic-state-space representation of the state dynamics.

$$(3) \quad \begin{aligned} dX(t) &= [(A_1(t) + A_2(t)) dt + B(t)u(t) dt + C(t) dw(t) + D(t)x(t) dt]X(t), \\ X(0) &= I, \\ x(t) &= \left[\int_0^t (dX(s))X^{-1}(s) \right]_{12}, \end{aligned}$$

where

$$(4) \quad A_2(t) = a(t)R,$$

$$(5) \quad B(t) = G(t)R,$$

$$(6) \quad C(t) = Q^{1/2}(t)R,$$

$$(7) \quad D(t) = F(t)R,$$

$$(8) \quad A_1(t) = \frac{1}{2}C^2(t).$$

We note that in this equation A_1 is introduced to keep the evolution of X on S^1 when the equation is interpreted as an Ito differential equation. In fact the term $A_1 X dt$ is precisely the second order correction term that arises in Ito differential calculus. $A_2(t)$ as well as $B(t)$, $C(t)$, and $D(t)$ are skew symmetric matrices.

If we set $C = D = A_1 = 0$, equation (3) then becomes

$$\dot{X}(t) = (A_2(t) + B(t)u(t))X(t),$$

which is a well-known deterministic model (Brockett [1]) for control systems on S^1 . This indicates that our formulation (3) introduces randomness (in the form of white Gaussian noise) into the above well-known deterministic model in a very natural way. In addition, the terms involving the coefficient D in (3) allow the physical quantity $x(t)$, the total angle that the considered rotational process has swept, to enter the state dynamics directly.

2.2. Sensor dynamic equations. We will now formulate a sensor dynamic equation on S^1 . The C_1 -representation of the sensor dynamics is given by the following scalar Ito differential equation:

$$(9) \quad dz(t) = H(t)x(t) dt + R^{1/2}(t) dv(t), \quad z(0) = 0,$$

where v is a standard scalar Brownian motion independent of w , $R^{1/2}(t)$ and $H(t)$ are scalar functions. Injecting z into S^1 via J , we obtain the following Ito matrix differential equation satisfied by the C_2 -representation $Z = J(z)$ of the sensor dynamic equation:

$$(10) \quad \begin{aligned} dZ(t) &= [\tfrac{1}{2}N(t) dt + S(t)x(t) dt + E(t) dv(t)]Z(t), \\ Z(0) &= I, \end{aligned}$$

where

$$(11) \quad N(t) = E^2(t),$$

$$(12) \quad S(t) = H(t)R,$$

$$(13) \quad E(t) = R^{1/2}(t)R.$$

We note that this is a dynamic-state-space representation of the sensor dynamics. The term $\frac{1}{2}NZ dt$ in (10) plays the same role as $\frac{1}{2}A_1 X dt$ did in (3). The matrices $S(t)$ and $E(t)$ are skew-symmetric.

We note that because J is a bijective operator and $Z = J(z)$, the σ -field in (Ω, \mathcal{A}, P) generated by $Z' = \{Z(s), 0 \leq s \leq t\}$ is the same as that generated by $z' = \{z(s), 0 \leq s \leq t\}$. In other words, Z' and z' carry the same amount of information about X . This enables the C_1 -representation (9) to serve as an extremely useful auxiliary equation in the analysis of detection, estimation, and control. While the detection and the estimation problems were treated in Lo and Willsky [8], and Lo [17], the application of the C_1 -representation to control problems will be considered in this paper.

Since a sensor cannot take measurement of future state evolution, the observation process Z (the output of the sensor) must be nonanticipative with respect to state evolution. More specifically $Z(t)$ must be a function of $X' = \{X(s), 0 \leq s \leq t\}$

or equivalently x^t (from the mathematical viewpoint) since $X^t = J(x^t)$ and J is bijective. We note that the sensor dynamic equation (10) does, in fact, have this essential feature.

From a physical viewpoint the sensors used to observe single-degree-of-freedom rotational processes can be classified into two kinds by the way in which the measurement is taken. The first type measures the orientation $X(t)$ directly (as in the measurement of a gimbal angle in an inertial navigation system (Wrigley, Hollister, and Denhard [14])). A sensor of the second kind measures the total angle swept $x(t)$ directly (e.g., an integrating gyroscope).

3. Cost criteria and feedback control. A cost criterion for a control system operating over some time period T is usually defined as a real-valued functional η on the direct product of the space of state trajectories and the space U (to be specified later) of admissible control functions over T . As shown in the previous section, there are two representations of the space of state trajectories— C_1 and C_2 , which are related by the bijective operator J . Therefore we may define the cost criterion as a real-valued functional on either $C_1 \times U$ or $C_2 \times U$. One form of the criterion can be easily obtained from the other via the operator J . A cost criterion in the form of a function on $C_i \times U$ will be called its C_i -representation for $i = 1$ and 2 .

Just as with the classification of sensors from a physical viewpoint, the C_1 - and the C_2 -representations of the cost criterion have different physical interpretations. When the cost is induced directly by the time history of the deviation of the total swept angle of the controlled rotational process from some desired total swept angle (or, alternatively, when it is induced by the deviation of the angular velocity of the controlled process from some desired rotational rate), it is obviously physically more natural to first write down the C_1 -representation of the cost criterion. A notable example of this kind is the control of synchronous rotation such as the control of a rotor in a motor or electric generator, or in the adjustment of a high-accuracy clock or an oscillator used for frequency modulation.

On the other hand, when the cost is induced directly by the time history of the “deviation” (a measure of angular deviation will be specified later) of the orientation of the controlled rotational process from the desired orientation, it is then physically more natural to write down the C_2 -representation of the cost criterion. A notable example of this kind is the satellite attitude control problem (Leondes [7]). In the following we will study the control problems for these two kinds of cost criteria. They will be referred to as synchronization control and orientation control respectively.

In the following, we will consider control systems defined on the fixed interval $T = [0, t_*]$. The space U of admissible control functions is defined as follows: let the mapping $\pi_t: C_2 \rightarrow C_2$ be defined by

$$\begin{aligned} (\pi_t A)(s) &= A(s), & 0 \leq s \leq t, \\ &= A(t), & t \leq s \leq t_*, \end{aligned}$$

for $A \in C_2$. Let $\|\cdot\|_s$ denote the supremum norm in C_2 defined by

$$\|A\|_s = \sup_{t \in T} (\text{tr } A(t)A'(t))$$

and let $\psi: T \times C_2 \rightarrow U$ (a convex subset of R^1) be a mapping with the following properties: $\psi(t, A)$ is Hölder continuous in t for each $A \in C_2$ and satisfies a uniform Lipschitz condition

$$|\psi(t, A_1) - \psi(t, A_2)| < c_3 \|A_1 - A_2\|_s$$

for $t \in T$ and $A_1, A_2 \in C_2$. Let Ψ denote the family of functions ψ . We call a control u admissible, and write $u \in U$, if

$$u(t) = \psi(t, \pi_t Z), \quad t \in T,$$

for some $\psi \in \Psi$. These conditions ensure the causality of the control.

The control problem to be studied in the following subsections is: given a cost function η , find $u^* \in U$ such that

$$\eta[u^*] = \min \{ \eta[u] : u \in U \}.$$

The corresponding function ψ^* will be called an optimal control law.

3.1. Synchronization control. Let the C_1 -representation of a desired rotational process be continuous and denoted by $\phi(t)$. Assume that the cost criterion η can be expressed as follows:

$$(14) \quad \eta[u] = E \left[\int_0^{t_*} (x(s) - \phi(s))^2 W(s) ds + \int_0^{t_*} u_2(s) V(s) ds \right],$$

where $W(t)$ and $V(t)$ are respectively nonnegative and positive-valued functions with $V^{-1}(t)$ bounded on T . We have mentioned that since linear control theory is better established than bilinear control theory, the C_1 -representation of the dynamic-state-space representation (2) and (9) serves as a very useful auxiliary equation. This is best shown by the following derivation of the optimal control law for the synchronization control problem.

Given the dynamic-state-representation, (3) and (10), of the control system, we first write down the C_1 -representations, (2) and (9), of the system with the coefficients $a(t)$, $F(t)$, $G(t)$, $Q^{1/2}(t)$, $H(t)$, $R^{1/2}(t)$ being determined by (4) ~ (8), (12), and (13).

In addition, we now define the set U_1 of admissible control functions, defined with respect to the C_1 system representation (as opposed to the set U , which was defined earlier with respect to the C_2 -representation).

Let π_t , defined earlier, also denote the mapping from C_1 into C_1 defined by

$$(\pi_t(a))(s) = \begin{cases} a(s), & 0 \leq s \leq t, \\ a(t), & t \leq s \leq t_*, \end{cases}$$

for $a \in C_1$. Let $\|\cdot\|_s$ also denote the supremum norm in C_1 and let $\psi_1: T \times C_1 \rightarrow R$ be a mapping with the properties: $\psi_1(t, a)$ is Hölder continuous in t for each $a \in C_1$ and satisfies a uniform Lipschitz condition

$$|\psi_1(t, a_1) - \psi_1(t, a_2)| < c_4 \|a_1 - a_2\|_s$$

for $t \in T$ and $a_1, a_2 \in C_1$. Let Ψ_1 denote the class of functionals ψ_1 . We call the

control u admissible and write $u \in U_1$ if

$$u(t) = \psi_1(t, \pi_t z), \quad t \in T,$$

for some $\psi_1 \in \Psi_1$. An element $\psi_1^0 \in \Psi_1$ is called an optimal control law if

$$\eta[u^0] = \min \{ \eta[u] : u \in U_1 \},$$

where $u^0(t) = \psi_1^0(t, \pi_t z)$.

By either completing squares or applying Lemma 5.1 (optimality criterion) of Wonham [13], the following lemma can be easily obtained.

LEMMA 1. *Consider the cost criterion (14) and the control system described by (2) and (9), with $x(t)$ regarded as the dynamic state. Then the optimal control law, $u^0(t) = \psi_1^0(t, \pi_t z)$, is given by*

$$(15) \quad u^0(t) = -V^{-1}(t)G(t)(P_1(t)[\hat{x}(t) - \phi(t)] + b(t)),$$

$$(16) \quad \begin{aligned} \dot{b}(t) &= P_1(t)G^2(t)V^{-1}(t)b(t) - F(t)b(t) - P_1(t)(\dot{a}(t) - \dot{\phi}(t)), \\ b(t_*) &= 0, \end{aligned}$$

$$(17) \quad \begin{aligned} \dot{P}_1(t) &= -F(t)P_1(t) - A(t)P_1(t) + G^2(t)V^{-1}(t)P_1^2(t) - W(t), \\ P_1(t_*) &= 0, \end{aligned}$$

$$(18) \quad \begin{aligned} \hat{x}(t) &= E(x(t)|z^t), \\ z^t &= \{z(s), 0 \leq s \leq t\}, \end{aligned}$$

$$(19) \quad \begin{aligned} d\hat{x}(t) &= a(t) dt + F(t)\hat{x}(t) dt + G(t)u^0(t) dt \\ &\quad + P_2(t)H(t)R^{-1}(t)(dz(t) - H(t)\hat{x}(t) dt), \\ \hat{x}(0) &= 0, \end{aligned}$$

$$(20) \quad \begin{aligned} \dot{P}_2(t) &= 2F(t)P_2(t) + Q(t) - H^2(t)R^{-1}(t)P_2^2(t), \\ P_2(0) &= 0. \end{aligned}$$

Using Lemma 1, we can now determine the optimal synchronization control law. We observe that the σ -subfield of \mathcal{A} generated by z^t is the same as that generated by $Z^t = \{Z(s), 0 \leq s \leq t\}$, because $Z^t = J(z^t)$ and J is bijective. In other words, z and Z are causally equivalent. Let this σ -subfield be denoted by \mathcal{A}_z^t . Then the conditional expectation $E(x(t)|\mathcal{A}_z^t)$ is both a B_1 -measurable functional f_1 of z^t and a B_2 -measurable functional f_2 of Z^t , and

$$f_2(Z^t) = f_1(J^{-1}(Z^t)).$$

Let $\hat{x}(t)$ and $\hat{x}(t|t)$ denote $f_1(z^t) \triangleq E(x(t)|z^t)$ and $f_2(Z^t) \triangleq E(x(t)|Z^t)$, respectively. Note that this notation is consistent with (18). Referring to Lo and Willsky [8], it is easily seen that

$$(21) \quad \begin{aligned} d\hat{x}(t|t) &= a(t) dt + F(t)\hat{x}(t|t) dt + G(t)\psi_1^0(t, \pi_t(J^{-1}(Z^t))) dt \\ &\quad + P^2(t)H(t)R^{-1}(t)\{[(dZ(t))Z'(t) - \tfrac{1}{2}N(t) dt]_{12} \\ &\quad \quad - H(t)\hat{x}(t|t) dt\}, \\ \hat{x}(0|0) &= 0. \end{aligned}$$

Again because z and Z are causally equivalent, we may define a $\psi_1 \in \Psi_1$ for each $\psi \in \Psi$ by

$$\psi_1(t, \pi_t z) = \psi[t, \pi_t(J(z))],$$

and we may define $\psi^0 \in \Psi$ by

$$\psi^0(t, \pi_t Z) = \psi_1^0[t, \pi_t(J^{-1}(Z))].$$

We note here that the properties of Ψ and Ψ_1 do not give rise to trouble in the above argument. Since ψ_1^0 is optimal, we have

$$\eta[\psi^0(t, \pi_t Z)] = \eta[\psi_1^0(t, \pi_t(J^{-1}(Z)))] \leq \eta[\psi_1(t, \pi_t(z))] = \eta[\psi(t, \pi_t Z)]$$

for all $\psi \in \Psi$. Hence ψ^0 is optimal. Summarizing what has been shown, we obtain the following theorem.

THEOREM 2. *Consider the control system of rotational processes described by the bilinear matrix Ito differential equations (3) and (10) and consider the cost criterion (14). The optimal control law ψ^* is given by*

$$u^*(t) = \psi^*(t, \pi_t Z) = -V^{-1}(t)G(t)(P_1(t)(\hat{x}(t|t) - \phi(t)) + b(t))$$

and

$$\begin{aligned} d\hat{x}(t|t) &= a(t) dt + F(t)\hat{x}(t|t) dt + G(t)u^*(t) dt + P_2(t)H(t)R^{-1}(t) \\ &\quad \cdot \{[dZ(t)Z'(t) - \frac{1}{2}N(t) dt]_{12} - H(t)\hat{x}(t|t) dt\} \\ \hat{x}(0|0) &= 0, \end{aligned}$$

where $a, F, G, R, H, P_1, b, P_2$ are determined by (4), (7), (5), (13), (12), (17), (16), (20), respectively.

3.2. Orientation control. The standard distance function (Riemannian metric) on the circle—i.e., the distance, ρ , between two points on the circle is the arc length of the shortest path (geodesic line) joining them. Any valid mathematical expression for the “distance” between two orientations must be a positive-valued function $\lambda: S^1 \times S^1 \rightarrow R^1$, which is nondecreasing with respect to ρ , i.e.,

$$\rho(\Theta_1, \Theta_2) > \rho(\Theta_1, \Theta_2) \Leftrightarrow \lambda(\Theta_1, \Theta_2) > \lambda(\Theta_1, \Theta_3)$$

for $\Theta_i \in S^1, i = 1, 2$. In this subsection, we will consider only

$$\lambda(\Theta_1, \Theta_2) = \frac{1}{2}(2 - \text{tr } \Theta_1 \Theta_2')$$

to avoid complexity in illustrating the approach.

Let $\Phi \in C_2$ be the desired evolution of the orientation. Then a cost criterion η for orientation control can be expressed as follows:

$$(22) \quad \eta[u] = E \left[\int_0^{t^*} \frac{1}{2}(2 - \text{tr } X(s)\Phi'(s)) ds + \int_0^{t^*} \gamma(s)u^2(s) ds \right],$$

where ϕ is a nonnegative scalar function over T .

Let y be the C_1 -representation of Φ . It is easily seen that the C_1 -representation of η can be written as follows:

$$(23) \quad \eta[u] = E \left[\int_0^{t^*} (1 - \cos(x(s) - \phi(s))) ds + \int_0^{t^*} \gamma(s)u^2(s) ds \right].$$

We note that the function $1 - \cos x$ was used in estimation problems in Bucy and Mallinckrodt [16].

In view of the C_1 -representation (2) and (9), setting $y(t) = x(t) - \phi(t)$, we have

$$\begin{aligned} dy &= (a - \dot{\phi} + F\phi) dt + Fy dt + Gu dt + Q^{1/2} dw, \\ y(0) &= 0, \\ dz &= H\phi dt + Hy dt + R^{1/2} dv, \\ z(0) &= 0, \\ \eta[u] &= E \left[\int_0^{t^*} (1 - \cos y(s)) ds + \int_0^{t^*} \gamma(s) u^2(s) ds \right]. \end{aligned}$$

Thus the Bellman functional equation (Kushner [6] and Wonham [13]) is

$$\min_{u \in U} [V_t(t, \xi) + \frac{1}{2} P^2 H^2 R^{-1} V_{\xi\xi}(t, \xi) + (F\xi + a - \dot{\phi} + F\phi + Gu) V_\xi(t, \xi) + 1 + \gamma u^2 - \exp(-\frac{1}{2}P) \cos \xi] = 0,$$

$$V(t_*, \xi) = 0,$$

where

$$(24) \quad \dot{P} = 2FP - H^2 R^{-1} P^2 + Q, \quad P(0) = 0.$$

We set

$$u = -\gamma^{-1} G V_\xi(t, \xi).$$

Then the control law $-\gamma^{-1}(t)G(t)V_\xi(t, \hat{y}(t))$ is optimal in U , if there exists a solution to the following partial differential equation (see Lemma 5.1 of Wonham [13]):

$$(25) \quad \begin{aligned} V_t(t, \xi) + \frac{1}{2} P^2 H^2 R^{-1} V_{\xi\xi}(t, \xi) + (F\xi + a - \dot{\phi} + F\phi) V_\xi(t, \xi) + 1 \\ - \exp(-\frac{1}{2}P) \cos \xi = 0, \\ V(t_*, \xi) = 0. \end{aligned}$$

Let

$$L(\cdot) \triangleq \frac{\partial}{\partial t}(\cdot) + \frac{1}{2} P^2 H^2 R^{-1} \frac{\partial^2}{\partial \xi^2}(\cdot) + (F\xi + a - \dot{\phi} + F\phi) \frac{\partial}{\partial \xi}(\cdot).$$

We note that L is a Kolmogorov backward operator (Doob [3, p. 275]). It is well known that there exists one and only one solution $V(t, \xi)$ to (25) and it can be written as

$$(26) \quad V(t, \xi) = \int_t^{t^*} \left\{ \int_{-\infty}^{\infty} g(t, \xi; s, \zeta) \left[1 - \exp\left(-\frac{P}{2}\right) \cos \zeta \right] d\zeta \right\} ds,$$

where g is a Green's function which satisfies

$$(27) \quad L[g(t, \xi; s, \zeta)] = 0, \quad g(s, \xi; s, \zeta) = \delta(\xi - \zeta),$$

δ being the Dirac delta function. It can be checked by simple calculation that the solution to (27) is

$$(28) \quad g(t, \xi; s, \eta) = \frac{\beta(t; s)}{\sqrt{2\pi\alpha(t; s)}} \exp \left[-\frac{(\xi - \mu(t; s, \eta))^2}{2\alpha(t; s)} \right],$$

where

$$(29) \quad \beta(t; s) = \exp \left[- \int_t^s F(\tau) d\tau \right],$$

$$(30) \quad \mu(t; s, \eta) = \beta(t; s)\eta + \int_t^s \beta(t; \tau)(a(\tau) - \dot{\phi}(\tau) + F(\tau)\phi(\tau)) d\tau,$$

$$(31) \quad \alpha(t; s) = \int_t^s \beta^2(t; \tau)P^2(\tau)H^2(\tau)R^{-1}(\tau) d\tau.$$

Substituting (28) into (26) yields

$$V(t, \xi) = (t_* - t) - \int_t^{t_*} \exp \left\{ - \frac{1}{2} \left[\frac{\alpha(t; s)}{\beta^2(t; s)} + P(s) \right] \right\} \cos \left[\frac{\xi - \mu(t; s, 0)}{\beta(t; s)} \right] ds.$$

Thus,

$$V_\xi(t, \xi) = \int_t^{t_*} \frac{1}{\beta(t; s)} \exp \left\{ - \frac{1}{2} \left[\frac{\alpha(t; s)}{\beta^2(t; s)} + P(s) \right] \right\} \sin \left[\frac{\xi - \mu(t; s, 0)}{\beta(t; s)} \right] ds.$$

Summarizing what has been shown, we obtain the following theorem for optimal orientation control.

THEOREM 3. *Consider the control system of rotational processes described by (3), (10) and consider the cost criterion (14). The optimal control law ψ^* is given by*

$$u^*(t) = \psi^*(t, \pi_t Z) = \int_y^{t_*} K(t, s) \sin \left[\frac{\hat{x}(t|t) - \phi(t) - \mu(t; s, 0)}{\beta(t; s)} \right] ds,$$

$$K(t, s) = - \frac{G(t)}{\gamma(t)\beta(t; s)} \exp \left\{ - \frac{1}{2} \left[\frac{\alpha(t; s)}{\beta^2(t; s)} + P(s) \right] \right\},$$

where β, μ, α are determined by (29) ~ (31), and

$$d\hat{x}(t|t) = a(t) dt + F(t)\hat{x}(t|t) dt + G(t)u^*(t) dt$$

$$+ P(t)H(t)R^{-1}(t)[(dZ(t))Z'(t)]_{12} - H(t)\hat{x}(t|t) dt),$$

$$\hat{x}(0|0) = 0,$$

where a, F, G, R, H, P are determined by (4), (7), (5), (13), (12), (24), respectively.

We remark that $K(t, s)$, $\beta(t; s)$, and $\mu(t; s, 0)$ can be precomputed and stored in the feedback controller. Hence it is believed that the optimal control scheme of the previous theorem can easily be implemented.

When x does not directly enter the state dynamics (3), i.e., when $D \equiv 0$, the optimal orientation control law takes a very simple and interesting form. We state it in the following corollary.

COROLLARY. *Consider the control problem in the previous theorem. If $D \equiv 0$, the optimal control law ψ^* is given by*

$$u^*(t) = \psi^*(t, \pi_t Z)$$

$$= c_1(t) \cos(x(t|t) - \phi(t)) + c_2(t) \sin(\hat{x}(t|t) - \phi(t)),$$

where

$$c_1(t) = - \int_t^{t^*} \exp \left\{ -\frac{1}{2} \left[\int_t^\tau P^2(s) H^2(s) R^{-1}(s) ds + P(\tau) \right] \right\} \sin \left[\int_t^\tau (a(s) - \dot{\phi}(s) + F(s)\phi(s)) ds \right] d\tau,$$

$$c_2(t) = \int_t^{t^*} \exp \left\{ -\frac{1}{2} \left[\int_t^\tau P^2(s) H^2(s) R^{-1}(s) ds + P(\tau) \right] \right\} \cos \left[\int_t^\tau (a(s) - \dot{\phi}(s) + F(s)\phi(s)) ds \right] d\tau,$$

and $\hat{x}(t|t)$, a , F , G , R , H , P are determined as in the previous theorem.

In Lo and Willsky [8], orientation estimation of rotational processes with one degree of freedom was studied. It was shown that the optimal orientation estimate $\hat{X}(t|t)$ of $X(t)$ given observation Z^t is

$$\hat{X}(t|t) = \exp(R\hat{x}(t|t)).$$

Hence the optimal control law in the previous corollary is in fact linear:

$$u^*(t) = [c_1(t), c_2(t)]\Phi'(t)\hat{X}(t|t)[1, 0]',$$

where $\Phi(t) = \exp(R\phi(t))$ is the C_2 -representation of $\phi(t)$.

3.3. Control on Abelian Lie groups. The results of the previous subsections can be extended to a large class of problems—those involving processes evolving on Abelian Lie groups. It is well known (Warner [10]) that a given Abelian Lie group G is isomorphic to the direct product of a number of copies of the real line and a number of copies of the unit circle, i.e.,

$$G \approx R^n \times (S^1)^m.$$

The diffusion processes on this type of space have been used to model some interesting satellite and pendulum systems in Ku and Sheporaitis [15]. Following Lo and Willsky [8], a bijective mapping $J_{nm}:(C_1^s)^{n+m} \rightarrow (C_1^s)^n \times (C_2^s)^m$ is defined by

$$(J_{nm}(a))(t) = [a_1(t), \dots, a_n(t), (J(a_{n+1}))(t), \dots, (J(a_{n+m}))(t)]$$

for $a \in (C_1^s)^{n+m}$, a_i being the i th component of a . Thus a continuous random signal process on G which is described by an \mathcal{A} -measurable function $X:\Omega \rightarrow (C_1^s)^n \times (C_2^s)^m$ corresponds to a unique continuous random signal process on R^{n+m} which is described by an \mathcal{A} -measurable function $x:\Omega \rightarrow (C_1^s)^{n+m}$ such that

$$X(t) = (J_{nm}(x))(t), \quad t \in [0, s].$$

The mathematical model for a control system on G can be obtained by first using J_{nm} to inject the following $(n+m)$ -vector random differential equation into

$$R^n \times (S^1)^m:$$

$$\begin{aligned} dx(t) &= a(t) dt + F(t)x(t) dt + C(t)u(t) dt + Q^{1/2}(t) dw(t), \\ x(0) &= 0, \end{aligned}$$

and using J_{pq} to inject the following $p + q$ -vector random differential equation into $R^p \times (S^1)^q$:

$$\begin{aligned} dz(t) &= H(t)x(t) dt + R^{1/2}(t) dv(t), \\ z(0) &= 0, \end{aligned}$$

where the coefficient functions are of appropriate dimension and w and v are independent vector Brownian motions. Differentiating $X(t) = (J_{mn}(x))(t)$ and $Z(t) = (J_{pq}(z))(t)$ by the stochastic differentiation rule, we obtain a set of joint linear and bilinear stochastic differential equations. This calculation is straightforward and thus we will not display those equations. Let $X(t) = [x_1(t), \dots, x_n(t), X_{n+1}(t), \dots, X_{n+m}(t)]$, where $X_{n+i}(t) = (J(x_{n+i}))(t)$. A joint synchronization and orientation cost criterion can be written as follows: for $0 \leq l \leq m$,

$$\begin{aligned} n[u] &= \sum_{i=1}^{n+l} \int_0^{t_*} \gamma_i(s)(x_i(s) - \phi_i(s))^2 ds + \sum_{i=n+l+1}^{n+m} \int_0^{t_*} \gamma_i(s)(2 - \text{tr } X_i(s)\Phi'_i(s)) ds \\ &\quad + \int_0^{t_*} u'(s)V(s)u(s) ds, \end{aligned}$$

where γ_i are nonnegative functions over T , V is nonnegative definite over T , and $\phi_i(s)$ and $\Phi_i(s)$ are the desired total swept angles and the desired orientations at $t = s$. Because of the bijective property of J_{nm} and J_{pq} , it is clear that the optimal control analysis in the previous subsections can be easily generalized to this general Abelian case with little modification. The reader is referred to Lo and Willsky [9] for some examples, which illustrate the approach.

Acknowledgment. The authors wish to acknowledge with thanks Professor Roger W. Brockett of Harvard University, who provided the motivation and direction for this work.

REFERENCES

- [1] R. W. BROCKETT, *System theory on group manifolds and coset spaces*, this Journal, 10 (1972), pp. 265–284.
- [2] R. S. BUCY AND P. D. JOSEPH, *Filtering for Stochastic Processes with Applications to Guidance*, John Wiley, New York, 1968.
- [3] J. L. DOOB, *Stochastic Processes*, John Wiley, New York, 1953.
- [4] K. ITO AND H. P. MCKEAN, JR., *Diffusion Processes and Their Sample Paths*, Academic Press, New York, 1965.
- [5] V. JURDJEVIC AND H. J. SUSSMAN, *Control systems on Lie groups*, Differential Geometric Methods in Control, Tech. Rep. 628, Division of Engineering and Applied Physics, Harvard Univ., Cambridge, Mass., 1971.
- [6] H. J. KUSHNER, *Stochastic Stability and Control*, Academic Press, New York, 1967.
- [7] LEONDES, C. T., ed., *Guidance and Control of Aerospace Vehicles*, McGraw-Hill, New York, 1963.
- [8] J. T. LO AND A. S. WILLSKY, *Estimation for rotational processes with one degree of freedom*, Tech. Rep. 635, Division of Engineering and Applied Physics, Harvard Univ., Cambridge, Mass., 1972.

- [9] ———, *Stochastic control of rotational processes with one degree of freedom*, Mathematics Research Rep. 73-5, Division of Mathematics and Physics, University of Maryland, Baltimore County, Baltimore, Md., 1973.
- [10] F. W. WARNER, *Foundations of Differentiable Manifolds and Lie Groups*, Scott, Foresman, Glenview, Ill., 1971.
- [11] J. WEI AND E. NORMAN, *On global representations of the solutions of linear differential equations as a product of exponentials*, Proc. Amer. Math. Soc., 15 (1964), pp. 327–334.
- [12] ———, *Lie algebraic solution of linear differential equations*, J. Mathematical Physics, 4 (1963), pp. 575–581.
- [13] W. M. WONHAM, *Random differential equations in control theory*, Probabilistic Methods in Applied Mathematics, vol. 2, A. Bharucha-Reid, ed., Academic Press, New York, 1970, pp. 131–212.
- [14] W. WRIGLEY, W. HOLLISTER AND W. DENHARD, *Gyroscopic Theory, Design, and Instrumentation*, MIT Press, Cambridge, Mass., 1969.
- [15] Y. H. KU AND L. P. SHEPORAITIS, *Global properties of diffusion processes on cylindrical type phase space*, J. Franklin Inst., 289 (1970), pp. 1087–1103.
- [16] R. S. BUCY AND A. J. MALLINCKRODT, *An optical phase demodulator*, Stochastics, 1 (1973), pp. 3–23.
- [17] J. T. LO, *Signal detection of rotational processes and frequency demodulation*, Information and Control, 26 (1974), pp. 99–115.

AN EXAMPLE OF A CONTINUOUS JUNCTION FOR A SINGULAR CONTROL PROBLEM OF EVEN ORDER*

HELMUT MAURER†

Abstract. The junction theorem of McDanell and Powers [1] gives a characterization of the junction between optimal singular and nonsingular subarcs. The theorem states in particular that for a singular control problem of even order an optimal piecewise analytic control must be continuous. No example for such a continuous junction seems to be known in the literature. In this paper we give for every even order q an example for which optimality can be proved using standard sufficiency theorems.

1. Statement of the problem. We consider the following control problem with control appearing linearly: determine the scalar, real-valued and measurable control $u(t)$, $t \in [t_0, t_f]$, which minimizes the functional

$$(1) \quad J(u) = G(x(t_f)) + \int_{t_0}^{t_f} (L_0(t, x) + L_1(t, x)u) dt$$

subject to the constraints

$$(2) \quad \dot{x} = f_0(t, x) + f_1(t, x)u, \quad x(t_0) = x_0,$$

$$(3) \quad \psi(x(t_f)) = 0,$$

$$(4) \quad |u(t)| \leq K, \quad K > 0.$$

Here x is a real n -vector of state variables, the function ψ is a k -vector with $k \leq n$, the functions G , L_0 , L_1 are scalar, and the functions f_0 , f_1 are n -vectors. All functions are supposed to be analytic in a suitable domain.

The Hamiltonian is linear in the control u ,

$$(5) \quad H(t, x, \lambda, u) = \lambda^T f_0(t, x) + L_0(t, x) + \{\lambda^T f_1(t, x) + L_1(t, x)\}u,$$

where the superscript T denotes transposition. The equations for the multipliers $\lambda(t) \in \mathbb{R}^n$ are given by

$$(6) \quad \dot{\lambda} = -H_x(t, x, \lambda, u), \quad -\lambda(t_f) = G_x(x(t_f)) + v^T \psi_x(x(t_f)),$$

where $v \in \mathbb{R}^k$ and the subscript x means the derivative with respect to x . The coefficient of u in (5) is called the *switching function*

$$(7) \quad \phi(t) = \lambda^T(t) f_1(t, x(t)) + L_1(t, x(t)).$$

According to the minimum principle the optimal control $u(t)$ minimizes the Hamiltonian in (5) with respect to u . If $\phi(t) \equiv 0$ on $[t_1, t_2] \subset [t_0, t_f]$, then $u(t)$ is a *singular* control on $[t_1, t_2]$. If $\phi(t) = 0$ only at isolated values of $t \in [t_1, t_2]$, then

$$(8) \quad u(t) = -K \operatorname{sgn} \phi(t)$$

and $u(t)$ is a *nonsingular* control on $[t_1, t_2]$.

* Received by the editors November 2, 1973, and in revised form May 20, 1974.

† Mathematisches Institut der Universität Köln, Köln, West Germany.

2. The junction theorem. Let $u(t)$ be an optimal singular control on the interval $[t_1, t_2]$. The lowest order time derivative of the switching function $\phi(t)$ in (7) containing u with a coefficient not vanishing identically on $[t_1, t_2]$ is of even order $2q$ (see [2]), and has the form

$$(9) \quad \begin{aligned} \phi^{(2q)}(t) &= \frac{d^{2q}}{dt^{2q}} H_u(t, x(t), \lambda(t), u(t)) \\ &= \alpha(t, x(t), \lambda(t)) + \beta(t, x(t), \lambda(t))u(t), \end{aligned}$$

where H_u is the derivative of H with respect to u . The integer q is called the *order of the singular arc*. The well-known generalized Legendre–Clebsch condition states that for an optimal singular subarc of order q it is necessary that

$$(10) \quad (-1)^q \frac{\partial}{\partial u} \left[\frac{d^{2q}}{dt^{2q}} H_u \right] = (-1)^q \beta(t, x(t), \lambda(t)) \geq 0.$$

We cite the junction theorem of McDanell and Powers [1].

THEOREM 1. *Let t_c be a point at which singular and nonsingular subarcs of an optimal control u are joined, and let q be the order of the singular arc. Suppose that the strengthened generalized Legendre–Clebsch condition is satisfied at t_c , i.e., $(-1)^q \beta(t_c, x(t_c), \lambda(t_c)) > 0$, and assume that the control is piecewise analytic in a neighborhood of t_c . Let $u^{(r)}$, $r \geq 0$, be the lowest order derivative of $u(t)$ which is discontinuous at t_c . Then $q + r$ is an odd integer.*

It follows from this theorem that for q even an optimal piecewise analytic control is *continuous* at the junction. However, only examples for a *nonanalytic* junction seem to be known [3], [4]. Such a nonanalytic junction can be predicted by the following theorem of McDanell and Powers [1, Cor. 3].

THEOREM 2. *If q is even, $\alpha(t_c, x(t_c), \lambda(t_c)) = 0$, $\beta(t_c, x(t_c), \lambda(t_c)) \neq 0$, where t_c is a junction point between optimal singular and nonsingular subarcs, then the junction is nonanalytic.*

Thus in order to construct an example with a piecewise analytic control and a continuous junction for q even, we have to choose an example with $\alpha(t_c, x(t_c), \lambda(t_c)) \neq 0$ in (9).

3. The example.

$$(11) \quad \text{Minimize} \quad \frac{1}{2} \int_0^2 (x_1^2 + x_2^2) dt$$

subject to

$$(12) \quad \begin{aligned} \dot{x}_i &= x_{i+1}, \quad i = 1, \dots, q, \quad q \text{ even}, \quad q > 0, \\ \dot{x}_{q+1} &= u, \end{aligned}$$

$$(13) \quad |u| \leq 1,$$

and the boundary conditions

$$(14) \quad x_i(0) = \exp(-1), \quad x_i(2) = \sigma_i, \quad i = 1, \dots, q+1.$$

The Hamiltonian and the multiplier equations are given by

$$(15) \quad H = \frac{1}{2}(x_1^2 + x_2^2) + \sum_{i=1}^q \lambda_i x_{i+1} + \lambda_{q+1} u.$$

$$(16) \quad \dot{\lambda}_1 = -x_1, \quad \dot{\lambda}_2 = -x_2 - \lambda_1, \quad \dot{\lambda}_i = -\lambda_{i-1}, \quad i = 3, \dots, q+1.$$

For the switching function

$$(17) \quad \phi(t) = \lambda_{q+1}(t),$$

we obtain

$$(18) \quad \phi^{(2q)}(t) = (-1)^{q-1}(x_q(t) - u(t)),$$

so the order of a singular arc is q even. The singular control determined by $\phi^{(2q)}(t) \equiv 0$ is

$$(19) \quad u(t) = x_q(t).$$

The strengthened generalized Legendre-Clebsch condition (10) is obviously satisfied.

We choose now a candidate optimal trajectory composed by a singular subarc emanating from the initial state followed by a nonsingular subarc, where the junction point occurs at $t_c = 1$.

Singular subarc in $[0, 1]$. The initial conditions (14) and the singular control (19) yield

$$(20) \quad x_i(t) = \exp(t-1), \quad u(t) = \exp(t-1), \quad i = 1, \dots, q+1.$$

It can be seen that (16) is satisfied by

$$(21) \quad \lambda_1(t) = -\exp(t-1), \quad \lambda_i(t) \equiv 0, \quad i = 2, \dots, q+1.$$

Nonsingular subarc in $[1, 2]$ with $u = 1$. The initial conditions at $t_c = 1$ for this nonsingular subarc as obtained from (20), (21) are $x_i(1) = 1, i = 1, \dots, q+1, \lambda_1(1) = -1, \lambda_i(1) = 0, i = 2, \dots, q+1$. Integration of (12) and (16) with $u = 1$ then gives

$$(22) \quad x_i(t) = \sum_{k=0}^{q+2-i} \frac{1}{k!} (t-1)^k, \quad i = 1, \dots, q+1,$$

$$(23) \quad \lambda_1(t) = -\sum_{k=0}^{q+2} \frac{1}{k!} (t-1)^k, \quad \lambda_i(t) = (-1)^i \sum_{k=0}^1 \frac{1}{(q+i+k)!} (t-1)^{q+i+k},$$

$$i = 2, \dots, q+1.$$

The switching function is contained in (23) for $i = q+1$:

$$(24) \quad \phi(t) = \lambda_{q+1}(t) = (-1)^{q+1} \sum_{k=1}^2 \frac{1}{(2q+k)!} (t-1)^{2q+k}.$$

As we have chosen q even, it follows from (24) that

$$\phi(t) < 0 \quad \text{for } 1 < t \leq 2,$$

which is in accordance with condition (8) implied by the minimum principle.

Thus, if we specify in (14) because of (22) the final state

$$\sigma_i = x_i(2) = \sum_{k=0}^{q+2-i} \frac{1}{k!}, \quad i = 1, \dots, q+1,$$

we have found in

$$(25) \quad u(t) = \begin{cases} \exp(t-1), & 0 \leq t \leq 1, \\ 1, & 1 \leq t \leq 2, \end{cases}$$

an admissible control, which is *continuous* at $t_c = 1$ and satisfies all the necessary conditions of optimality.

As the cost functional (11) is *convex* and the differential equations (12) are *affine* in x and u , the control (25) is indeed *optimal*. This is a consequence of the sufficient conditions in [5]–[7].

Now we look for a possible imbedding of this optimal trajectory in the following way: consider for $0 < t_c < 1$ the singular control

$$(26) \quad u(t) = \begin{cases} \exp(t-1), & 0 \leq t \leq t_c, \\ 1, & t_c < t \leq 2, \end{cases}$$

producing certain final values $\sigma_i(t_c)$. The control is *discontinuous* at t_c and therefore *nonoptimal* according to Theorem 1. In fact, the nonoptimality can also be directly verified by evaluating the switching function for (26) which becomes

$$(27) \quad \phi(t) = \frac{1-c}{(2q)!} (t-t_c)^{2q} - \frac{c}{(2q+1)!} (t-t_c)^{2q+1} - \frac{1}{(2q+2)!} (t-t_c)^{2q+2},$$

$$t_c \leq t \leq 2,$$

where $c = \exp(t_c - 1) < 1$. But (27) implies because of $c < 1$ that

$$\phi(t) > 0 \quad \text{for } t_c < t < t_c + \varepsilon, \quad \varepsilon > 0 \text{ suitable,}$$

which violates the optimality condition (8). The existence of an *optimal* trajectory follows from [8, Chap. 3.5]. The structure of the optimal control is not at all obvious. The optimal trajectory may contain a singular arc to which a chattering nonsingular arc is joined; see [4]. We presume that the optimal control is nonsingular with a finite number of switches. As $t_c \rightarrow 1$ the number of switches for this nonsingular control tends to infinity.

Acknowledgment. I would like to thank Professor W. F. Powers who pointed out to me the sufficiency theorems fitting this example.

REFERENCES

- [1] J. P. McDANELL AND W. F. POWERS, *Necessary conditions for joining optimal singular and non-singular subarcs*, this Journal, 9 (1971), pp. 161–173.
- [2] H. J. KELLEY, R. E. KOPP AND H. G. MOYER, *Singular extremals*, Topics in Optimization, G. Leitmann, ed., Academic Press, New York, 1967, Chap. 3.
- [3] A. T. FULLER, *Study of an optimum nonlinear control system*, J. Electronics Control, 15 (1963), pp. 63–71.
- [4] C. MARCHAL, *Chattering arcs and chattering controls*, J. Optimization Theory Appl., 11 (1973), pp. 441–468.

- [5] J. E. FUNK AND E. G. GILBERT, *Some sufficient conditions for optimality in control problems with state space constraints*, this Journal, 8 (1970), pp. 498–504.
- [6] E. B. LEE, *A sufficient condition in the theory of optimal control*, this Journal, 1 (1963), pp. 241–245.
- [7] O. L. MANGASARIAN, *Sufficient conditions for the optimal control of nonlinear systems*, this Journal, 4 (1966), pp. 139–152.
- [8] E. G. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

OPTIMAL CONTROL OF STOCHASTIC LINEAR DISTRIBUTED PARAMETER SYSTEMS*

A. BENSOUSSAN† AND M. VIOT‡

Abstract. In this article we give necessary and sufficient conditions of optimality for linear stochastic distributed parameter systems, with convex differentiable payoffs and partial observation. They are obtained through variational methods, which can be applied only in the case of fixed information (i.e., not dependent on the control or the state). However, by a density argument it is proven that an optimal control adapted to the observation is also optimal for a space of controls adapted to some fixed information. Therefore we can get the necessary and sufficient conditions also in the case of feedback controls. We then prove, as a consequence, the separation principle for distributed parameter systems in the case of a quadratic payoff.

Introduction. In this article, one considers an optimal control problem for a stochastic linear infinite-dimensional system (with applications to distributed parameter systems) with partial observation. The payoff is not necessarily quadratic. The system is of the form

$$(i) \quad y(t) + \int_0^t A(\tau)y(\tau) d\tau = y_0 + \int_0^t f(\tau) d\tau + \int_0^t B(\tau) d\zeta(\tau) + \int_0^t D(\tau)u(\tau) d\tau,$$

$$(ii) \quad z(t) = \int_0^t C(\tau)y(\tau) d\tau + \eta(t)$$

and the payoff is

$$(iii) \quad J(u(\cdot)) = E\left(\int_0^T l(y(t), u(t), t) dt + \Lambda(y(T))\right).$$

As usual, one of the main difficulties, due to the stochastic character of the problem, is to define the control in feedback form. Direct approaches consist in considering *explicit* deterministic functions of the past observations. This approach has been extensively used for finite-dimensional systems (see W. H. Fleming [6], W. M. Wonham [13]), and for infinite-dimensional systems in A. Bensoussan [2] and H. J. Kushner [7].

A different approach is considered here; for $u(\cdot) \in L^2(0, T \times \Omega, dt \otimes d\mu; U)$, where U is the Hilbert space of controls, equation (i) well defines $y(\cdot)$ and thus $J(u(\cdot))$ has a meaning. Also z is defined for such $u(\cdot)$. In particular, $z(t; \omega)$ can be considered as a random variable with values in $C(0, T; F)$, where F is the Hilbert space of observations. Denoting by $z_t(\omega)$ the restriction of z to $C(0, t; F)$, which is a

* Received by the editors March 1, 1973, and in revised form November 9, 1973.

† University of Paris IX, Paris, France, and Institut de Recherche d'Informatique et d'Automatique, Rocquencourt 78150, Le Chesnay, France.

‡ Institut de Recherche d'Informatique et d'Automatique, Rocquencourt 78150, Le Chesnay, France.

random variable with values in $C(0, t; F)$, one can consider the subspace of $L^2(0, T \times \Omega, dt \otimes d\mu; U)$ of control functions $u(\cdot)$ such that

$$(iv) \quad \text{a.e. } t, \quad u(t) \text{ is } \nu_t\text{-measurable,}$$

where ν_t is the probability law of z_t .

The main problem concerns necessary conditions of optimality of controls $u(\cdot)$ satisfying (iv). Naturally, condition (iv) states that $u(t)$ must be adapted to the observation. In other words, what we do is to consider adapted (or feedback) controls as a *subset of a fixed Hilbert space*. Unfortunately, $L^2(0, T \times \Omega, dt \otimes d\mu; U)$ is to some extent too broad. But it is possible to define a *fixed sub-Hilbert space* \mathcal{W} containing the feedback controls and such that the set of feedback controls is a *dense* subset of \mathcal{W} . The density property is fundamental in the sense that if a feedback control is optimal, it is also optimal in \mathcal{W} . But necessary conditions of optimality are much easier to get in \mathcal{W} (which is a Hilbert space), because one can use variational methods as in the deterministic case, or as in the case of *fixed observation*.

Variational methods in stochastic control in the case of fixed observation (and in finite dimensions) have been extensively used by J. M. Bismut [4]. They have been also used by A. V. Balakrishnan [1] for partial observation (and still in finite dimensions) (see also R. A. Brooks [5]). The density argument has been introduced by M. Viot [12]. We get necessary conditions of optimality for the problem (iii). Specializing to the quadratic case, we get the well-known separation principle.

1. Formulation of the model.

1.1. Assumptions and notation. Let V and H be two Hilbert spaces such that

$$(1.1) \quad V \subset H, \quad V \text{ dense in } H \text{ with continuous injection.}$$

We identify H and its topological dual space. If V' denotes the dual of V , then according to (1.1), we have

$$(1.2) \quad V \subset H \subset V'.$$

Let $t \in]0, T[$ be the time. Let us consider a family $A(t)$ of linear bounded operators from $V \rightarrow V'$, such that

$$(1.3) \quad \langle A(t)z, z \rangle + \lambda|z|^2 \geq \alpha\|z\|^2, \quad \lambda \geq 0, \quad \alpha > 0,^1$$

$$(1.4) \quad t \rightarrow \langle A(t)z_1, z_2 \rangle \quad \text{is measurable} \quad \text{for all } z_1, z_2 \in V,$$

$$(1.5) \quad \|A(t)\|_{\mathcal{L}(V; V')} \leq M.$$

Let E be a Hilbert space and $B = B(t) \in L^\infty(0, T; \mathcal{L}(E; H))$. We now consider a topological probability space $(\Omega, \mathcal{A}, \mu)$, where Ω denotes the set of elementary events, \mathcal{A} a σ -algebra of Borel subsets of Ω , and μ a Radon probability measure on Ω (in the sense of L. Schwartz [11] and P. A. Meyer [10]). We consider a Wiener process with values in E , denoted by $\xi(t)$, i.e., a stochastic process with values in E

¹ We denote by $|\cdot|$ the norm in H , $\|\cdot\|$ the norm in V , (\cdot, \cdot) the scalar product in H , $((\cdot, \cdot))$ the scalar product in V and $\langle \cdot, \cdot \rangle$ the duality between V and V' .

such that for all t ,

$$(1.6) \quad (\xi(t), e) \text{ is a Gaussian real random variable,}$$

$$(1.7) \quad E\xi(t) = 0 \quad \text{for all } t,$$

$$(1.8) \quad E(\xi(t), e_1)(\xi(s), e_2) = \int_0^{\min(t, s)} (Q(\tau)e_1, e_2) d\tau \quad \text{for all } e_1, e_2 \in E,$$

where $Q \in L^\infty(0, T; \mathcal{L}(E; E))$ satisfies:

- (i) $Q(\tau)$ is self-adjoint nonnegative,
- (ii) $Q(\tau)$ has a finite trace, i.e., if e_1, \dots, e_n, \dots is an orthonormal basis of E , then

$$\sum_{i=1}^{\infty} (Q(\tau)e_i, e_i) < +\infty.$$

We set

$$\text{tr } Q(\tau) = \sum_{i=1}^{\infty} (Q(\tau)e_i, e_i),$$

This definition is independent of the particular basis e_1, \dots, e_i, \dots which is chosen. The family $Q(t)$ is called the *covariance operator* of $\xi(t)$. One can prove (see, for instance, Bensoussan [3]) that

$$(1.9) \quad \omega \rightarrow \xi(\cdot; \omega) \text{ is a random variable with values in } C(0, T; E).$$

Let $f \in L^2(0, T \times \Omega, d\mu \otimes dt; H)$ and $y_0 \in L^2(\Omega, d\mu; H)$ be such that

$$(1.10) \quad \begin{aligned} &\text{for all } t_1 \leq t_2, \text{ the random variable } \xi(t_2) - \xi(t_1) \text{ (with values in } \\ &E) \text{ is independent of } \{f|_{t_1}, y_0, \xi(\tau_1), \dots, \xi(\tau_n)\}, \text{ which is a ran-} \\ &\text{dom variable with values in } L^2(0, t_1; H) \times H \times E^n, \text{ for any } n \text{ and} \\ &0 \leq \tau_1 \leq \dots \leq \tau_n \leq t_1. \end{aligned}$$

In (1.10), $f|_{t_1}$ denotes the restriction of f to $]0, t_1[$.

We recall the following theorem (see A. Bensoussan [3]).

THEOREM 1.1. *Under the preceding assumptions, there exists a unique stochastic process $y(t)$ such that*

$$(1.11) \quad y \in L^2(0, T \times \Omega, dt \otimes d\mu; V) \cap C(0, T; L^2(\Omega, \mu; H)),$$

$$(1.12) \quad y(t) + \int_0^t A(\tau)y(\tau) d\tau = y_0 + \int_0^t f(s) ds + \int_0^t B(s) d\xi(s) \quad \text{for all } t, \quad \text{a.s. } \omega,$$

$$(1.13) \quad \begin{aligned} E|y(t)|^2 + 2E \int_0^t \langle A(s)y(s), y(s) \rangle ds &= E|y_0|^2 + 2E \int_0^t (f(s), y(s)) ds \\ &+ \int_0^t \text{tr } B(s)Q(s)B^*(s) ds \quad \text{for all } t. \end{aligned}$$

Equation (1.12) represents the dynamics of a stochastic system, the state of which is $y(t)$ (with values in H).

1.2. The observation process. Let F be a Hilbert space and

$$(1.14) \quad C \equiv C(t) \in L^\infty(0, T; \mathcal{L}(H; F)).$$

Let $\eta(t)$ be a stochastic process with values in F , such that

$$(1.15) \quad \eta \in L^2(0, T \times \Omega, dt \otimes d\mu; F) \cap \text{meas}(\Omega, \mu; C(0, T; F)).$$

We then define the observation process by setting

$$(1.16) \quad z(t) = \int_0^t C(s)y(s) ds + \eta(t).$$

In particular, we obtain

$$(1.17) \quad z \in L^2(0, T \times \Omega, dt \otimes d\mu; F) \cap \text{meas}(\Omega, \mu; C(0, T, F)).$$

2. Review on conditional expectations.

2.1. Definitions. Let Φ and Ψ be two Banach spaces, and let $\psi(\omega)$ be a random variable with values in Ψ . Let

$$(2.1) \quad \mathcal{U} = L^2(\Omega, \mu; \Phi)$$

and

$$(2.2) \quad \mathcal{U}^\psi = \{\varphi(\omega) \in \mathcal{U} | \varphi(\omega) = \lambda \circ \psi(\omega)\},$$

where λ is any mapping from Ψ into Φ which is measurable with respect to the image of μ on Ψ . Then \mathcal{U}^ψ is a sub-Hilbert space of \mathcal{U} . For any $\varphi \in \mathcal{U}$, one defines $E^\psi \varphi$, the conditional expectation of φ with respect to ψ , as the projection of φ on \mathcal{U}^ψ .

LEMMA 2.1. *Let Φ, Ψ, Ψ_1 be three Banach spaces, and $\psi(\omega), \psi_1(\omega)$ two random variables with values in Ψ and Ψ_1 , respectively. Let v and v_1 be the images of μ by ψ and ψ_1 . If there exists a v -measurable mapping Π , from Ψ into Ψ_1 , such that*

$$\psi_1 = \Pi \circ \psi,$$

then

$$\mathcal{U}^{\psi_1} \subset \mathcal{U}^\psi.$$

This is a direct consequence of the composition rule for measurable mappings (cf. L. Schwartz [11]).

Now let $\varphi(t)$ and $\psi(t)$ be two stochastic processes with values in Φ and Ψ , such that

$$(2.3) \quad \varphi \in L^2(0, T \times \Omega, dt \otimes d\mu; \Phi), \quad \psi \in \text{meas}(\Omega, \mu; C(0, T; \Psi)).$$

For any t , we define $\psi_t \equiv \psi_t(s; \omega) = \text{restriction of } \psi \text{ on } (0, t)$, which is a random variable with values in $C(0, t; \Psi)$. Furthermore we define \mathcal{U}^{ψ_t} as in (2.2). We consider the following Hilbertian sum

$$(2.4) \quad \int^\oplus \mathcal{U}^{\psi_t} dt = \{\varphi(t, \omega) \in L^2(0, T; \mathcal{U}) | \varphi(t) \in \mathcal{U}^{\psi_t} \text{ a.e. } t\}.$$

We consider also $E^{\psi_t} \varphi(t)$, called *the conditional expectation of the process $\varphi(t)$ with respect to the process $\psi(t)$* , as t varies.

2.2. Measurability property. Naturally $t \rightarrow E^{\psi_t} \varphi(t)$ is a stochastic process with values in Φ . We have the following result.

PROPOSITION 2.1. *Up to a modification on a set of measure 0 in $(0, T)$,*

$$(2.5) \quad E^{\psi_t} \varphi(t) \in \int^{\oplus} \mathcal{U}^{\psi_t} dt.$$

Proof. We shall prove that there exists an element of $\int^{\oplus} \mathcal{U}^{\psi_t} dt$, temporarily denoted by $\hat{\varphi}(t)$, such that

$$(2.6) \quad \hat{\varphi}(t) = E^{\psi_t} \varphi(t) \quad \text{a.e. } t.$$

Let us define $\hat{\varphi}$ by the projection of φ on $\int^{\oplus} \mathcal{U}^{\psi_t} dt$, which is a sub-Hilbert space of $L^2(0, T; \mathcal{U})$. By definition, we have

$$(2.7) \quad \int_0^T (\hat{\varphi}(t), g(t))_{\mathcal{U}} dt = \int_0^T (\varphi(t), g(t))_{\mathcal{U}} dt \quad \text{for all } g(\cdot) \in \int^{\oplus} \mathcal{U}^{\psi_t} dt.$$

Let us take $g \in \mathcal{U}^{\psi_{t_0}}$ for a given t_0 and consider in (2.7) the case of

$$(2.8) \quad g(t) = \begin{cases} 0 & \text{if } t \notin]t_0, t_0 + \varepsilon[, \\ g & \text{if } t \in]t_0, t_0 + \varepsilon[. \end{cases}$$

Since $\mathcal{U}^{\psi_{t_0}} \subset \mathcal{U}^{\psi_t}$ if $t \geq t_0$, the $g(\cdot)$ in (2.8) belongs to $\int^{\oplus} \mathcal{U}^{\psi_t} dt$. It follows from (2.7) that

$$(2.9) \quad \int_{t_0}^{t_0 + \varepsilon} (\hat{\varphi}(t), g)_{\mathcal{U}} dt = \int_{t_0}^{t_0 + \varepsilon} (\varphi(t), g)_{\mathcal{U}} dt.$$

If t_0 is a Lebesgue point, it follows from (2.9) that

$$(2.10) \quad (\hat{\varphi}(t_0), g)_{\mathcal{U}} = (\varphi(t_0), g)_{\mathcal{U}}.$$

Since the complement of the set of Lebesgue points is of measure 0, (2.10) holds a.e., so (2.10) means

$$\hat{\varphi}(t_0) = E^{\psi_{t_0}} \varphi(t_0).$$

Hence the proposition holds.

From now on, when we speak of $t \rightarrow E^{\psi_t}(t)$ we automatically imply the corresponding element of $\int^{\oplus} \mathcal{U}^{\psi_t} dt$.

Remark 2.1. We shall use the following notation. If $\varphi(\omega)$ and $\psi(\omega)$ are two random variables as in §2.1, then we shall indicate by \mathcal{U}^{ψ} (defined in (2.2)) the following:

$$(2.11) \quad \mathcal{U}^{\psi} = \tilde{L}^2(\Psi, dv; \Phi),$$

where v denotes the image probability of μ defined on Ψ by the random variable $\psi(\omega)$. Then as usual $L^2(\Psi, dv; \Phi)$ denotes the Hilbert space of functions $\psi \rightarrow \varphi(\psi)$, which are v -measurable and such that

$$\int \|\varphi(\psi)\|^2 dv(\psi) < +\infty.$$

The meaning of the right-hand side of (2.11) is the following. The mapping \mathcal{T} from $L^2(\Psi, dv; \Phi)$ into $L^2(\Omega, \mu; \Phi)$, defined by

$$\mathcal{T}\lambda(\cdot) = \lambda(\psi(\omega)),$$

is an *isometry*. Hence \mathcal{U}^ψ is isometric to $L^2(\Psi, dv; \Phi)$. This property is reflected by the notation $\tilde{L}^2(\Psi, dv; \Phi)$.

Remark 2.2. Let us consider $\varphi(\cdot) \in \int^\oplus \mathcal{U}^{\psi_t} dt$. Then

$$(2.12) \quad \varphi_t \in \tilde{L}^2(C(0, t; \Psi), dv_t; L^2(0, t; \Phi)),$$

where v_t denotes the probability defined on $C(0, t; \Psi)$ by the random variable ψ_t . Indeed, on $(0, t)$ we have

$$(2.13) \quad \varphi(s) \in \mathcal{U}^{\psi_t} \quad \text{a.e. } s \in (0, t).$$

Also from (2.4) it follows that

$$(2.14) \quad \varphi_t \in L^2(0, t; \tilde{L}^2(C(0, t; \Psi), dv_t; \Phi)),$$

while (2.12) follows from the equality

$$L^2(0, t; \tilde{L}^2(C(0, t; \Psi), dv_t; \Phi)) = \tilde{L}^2(C(0, t; \Psi), dv_t; L^2(0, t, \Phi)).$$

3. The optimal control problem.

3.1. Setting of the problem. We shall now add to (1.12) a control term. Let U be a Hilbert space (the space of controls) and $D \in L^\infty(0, T; \mathcal{L}(U; H))$. We consider instead of (1.12) the following equation:

$$(3.1) \quad y(t) + \int_0^t A(\tau)y(\tau) d\tau = y_0 + \int_0^t f(\tau) d\tau + \int_0^t B(\tau) d\xi(\tau) + \int_0^t D(\tau)u(\tau) d\tau.$$

For any $u(\cdot) \in L^2(0, T; \mathcal{U})$ ($\mathcal{U} = L^2(\Omega, \mu; U)$), if we apply Theorem 1.1 to equation (3.1), we get the existence of a stochastic process $y(t)$. We then define $z(u)(t) = z(t)$ by

$$(3.2) \quad z(t) = \int_0^t C(\tau)y(\tau) d\tau + \eta(t).$$

But the observation being given by (3.2), the control $u(\tau)$ at time τ , must be a function of the observation up to time τ .

This statement can be made precise by defining explicit feedback function as is done in Bensoussan [2]. We shall use here a different approach, which probably has a broader domain of application, relying on an idea introduced by Viot [12] (for finite-dimensional systems). We can say that the “natural” controls are those such that there exists a delay between the observation and the decision. More precisely, the control $u(\tau)$ at time τ is then a function of the observation up to time $\tau - \varepsilon$ ($\varepsilon > 0$). We prove that such controls are included in a fixed sub-Hilbert space of $L^2(0, T; \mathcal{U})$, and furthermore that the feedbacks obtained by these “natural” controls for $\varepsilon \downarrow 0$ form a dense subspace of this sub-Hilbert space (for more details see §3.3).

3.2. Set of admissible controls. Let us consider the process α and β defined as follows:

$$(3.3) \quad \begin{aligned} \alpha(t) + \int_0^t A(\tau)\alpha(\tau) d\tau &= y_0 + \int_0^t f(\tau) d\tau + \int_0^t B(\tau) d\xi(\tau), \\ \beta(t) &= \int_0^t C(\tau)\alpha(\tau) d\tau + \eta(t). \end{aligned}$$

Let us introduce \mathcal{U}^{β_t} , an increasing family of sub-Hilbert spaces of $\mathcal{U} = L^2(\Omega, \mu; U)$,

$$\mathcal{U}^{\beta_t} = \tilde{L}^2(C(0, t; F), dv_t; U),$$

v_t the probability law of β_t . In the same way, for the observation process $z(\cdot)$ given by (3.1) (3.2) and corresponding to any $u(\cdot) \in L^2(0, T; \mathcal{U})$, let

$$\mathcal{U}^{z_t} = \tilde{L}^2(C(0, t; F), dv_t(u); U),$$

$v_t(u)$ the probability law of z_t .

LEMMA 3.1. *Let $z(\cdot)$ be the observation process corresponding to any $u(\cdot) \in L^2(0, T; \mathcal{U})$. If $u(\cdot) \in \int^\oplus \mathcal{U}^{z_t} dt$, then for any $t \in [0, T]$, $\mathcal{U}^{\beta_t} \subset \mathcal{U}^{z_t}$.*

Proof. Let y_1 and z_1 be defined as follows:

$$(3.4) \quad \begin{aligned} \frac{dy_1}{dt} + A(t)y_1 &= D(t)u(t), \quad y_1(0) = 0, \\ \frac{dz_1}{dt} &= C(t)y_1, \quad z_1(0) = 0. \end{aligned}$$

Clearly we have

$$(3.5) \quad y(t) = y_1(t) + \alpha(t), \quad z(t) = z_1(t) + \beta(t).$$

Furthermore, from (3.4) it follows that the mapping

$$(3.6) \quad \chi_t: u_t \rightarrow z_{1t}$$

is linear and continuous from $L^2(0, t; U)$ into $C(0, t; F)$. Therefore we can write²

$$(3.7) \quad z_t(\omega) = \chi_t u_t(\omega) + \beta_t(\omega).$$

But by hypothesis and Remark 2.2, there exists

$$\lambda_t \in L^2(C(0, t; F), dv_t(u); L^2(0, t; U)),$$

such that

$$(3.8) \quad u_t(\omega) = \lambda_t(z_t(\omega)).$$

By virtue of (3.7), (3.8), we deduce

$$(3.9) \quad \beta_t(\omega) = z_t(\omega) - \chi_t \lambda_t(z_t(\omega)).$$

The result now follows from Lemma 2.1.

In the same way, we obtain the following lemma.

² We recall that the notation β_t means the restriction of $\beta(\cdot; \omega)$ on the interval $(0, t)$ and is a random variable with values in $C(0, t; F)$. The same thing applies for z_t, u_t, \dots .

LEMMA 3.2. If $u(\cdot) \in \int^{\oplus} \mathcal{U}^{\beta_t} dt$, then for any $t \in [0, T]$, $\mathcal{U}^{z_t} \subset \mathcal{U}^{\beta_t}$, where $z(\cdot)$ is the observation process corresponding to $u(\cdot)$.

COROLLARY 3.1. Let $z(\cdot)$ be the observation process corresponding to $u(\cdot) \in L^2(0, T; \mathcal{U})$. If $u(\cdot) \in \int^{\oplus} \mathcal{U}^{\beta_t} dt \cap \int^{\oplus} \mathcal{U}^{z_t} dt$, then for any $t \in [0, T]$, $\mathcal{U}^{\beta_t} = \mathcal{U}^{z_t}$.

PROPOSITION 3.1. Let $z(\cdot)$ be the observation process corresponding to $u(\cdot) \in L^2(0, T; \mathcal{U})$. Then for any $\varepsilon > 0$:

$$u(\cdot) \in \int^{\oplus} \mathcal{U}^{z_t - \varepsilon} dt \Leftrightarrow u(\cdot) \in \int^{\oplus} \mathcal{U}^{\beta_t - \varepsilon} dt,$$

where $\mathcal{U}^{z_t - \varepsilon} = \mathcal{U}^{\beta_t - \varepsilon} = U$ for $0 \leq t \leq \varepsilon$.

Proof. Let us suppose that $u(\cdot) \in \int^{\oplus} \mathcal{U}^{z_t - \varepsilon} dt$. Since we have

$$\int^{\oplus} \mathcal{U}^{z_t - \varepsilon} dt \subset \int^{\oplus} \mathcal{U}^{z_t} dt,$$

the inclusion $\mathcal{U}^{\beta_t} \subset \mathcal{U}^{z_t}$ is a consequence of Lemma 3.1. But for $t \in [0, \varepsilon]$, u_t belongs to $L^2(0, t; U)$, and (cf. (3.6))

$$z_t(\omega) = \chi_t u_t + \beta_t(\omega).$$

Therefore it is clear that

$$\mathcal{U}^{z_t} = \mathcal{U}^{\beta_t} \quad \text{for all } t \in (0, \varepsilon).$$

Let us prove now that, if for any $a \geq \varepsilon$,

$$\mathcal{U}^{z_t} = \mathcal{U}^{\beta_t} \quad \text{for all } 0 \leq t \leq a,$$

then

$$\mathcal{U}^{z_t} = \mathcal{U}^{\beta_t} \quad \text{for all } 0 \leq t \leq a + \varepsilon.$$

It is, of course, enough to prove that

$$\mathcal{U}^{z_t} \subset \mathcal{U}^{\beta_t} \quad \text{for all } a \leq t \leq a + \varepsilon.$$

But by assumption, we have

$$(3.10) \quad \begin{aligned} &\text{a.e. } s \in (a, a + \varepsilon), \quad u(s) \in \mathcal{U}^{z_{s-\varepsilon}}, \\ &\text{for all } s \in (a, a + \varepsilon), \quad \mathcal{U}^{z_{s-\varepsilon}} = \mathcal{U}^{\beta_{s-\varepsilon}} \subset \mathcal{U}^{\beta_s}. \end{aligned}$$

Therefore, for any fixed $t \in (a, a + \varepsilon)$, there exists Π_t belonging to $L^2(C(0, t; F), dv_t; L^2(0, t; U))$ such that

$$(3.11) \quad u_t(\omega) = \Pi_t(\beta(\omega))$$

and

$$(3.12) \quad z_t(\omega) = \chi_t \Pi_t(\beta_t(\omega)) + \beta_t(\omega).$$

Thus we obtain the inverse inclusion

$$\mathcal{U}^{z_t} \subset \mathcal{U}^{\beta_t} \quad \text{for all } t \in (a, a + \varepsilon).$$

It can be observed that, if we change the assumption $u(\cdot) \in \int^{\oplus} \mathcal{U}^{z_t - \varepsilon} dt$ into $u(\cdot) \in \int^{\oplus} \mathcal{U}^{\beta_t - \varepsilon}$, the same conclusion holds. Hence the result is proved.

PROPOSITION 3.2. Let $z(\cdot)$ be the observation process corresponding to any $u(\cdot) \in L^2(0, T; \mathcal{U})$, $\mathcal{U} = L^2(\Omega, \mu; U)$. The following statements are equivalent:

- (a) $u(\cdot) \in \int^\oplus \mathcal{U}^{\beta_t} dt$ and $\mathcal{U}^{\beta_t} = \mathcal{U}^{z_t}$ for all t .
- (b) $u(\cdot) \in \int^\oplus \mathcal{U}^{z_t} dt$ and there exists a sequence (u^n, ε^n) such that u^n converges toward u in $L^2(0, T; \mathcal{U})$, ε^n converges to 0 and $u^n(\cdot) \in \int^\oplus \mathcal{U}^{z_t - \varepsilon^n} dt$.

Proof. (i) At first, suppose that $\mathcal{U}^{\beta_t} = \mathcal{U}^{z_t}$ for all $t \in [0, T]$. Set

$$(3.13) \quad u^n(t) = \begin{cases} u(t - 1/n) & \text{if } t \in [1/n, T], \\ 0 & \text{if } t \in [0, 1/n]. \end{cases}$$

It is well known that $u^n \rightarrow u$ in $L^2(0, T; \mathcal{U})$. Moreover, we have

$$u^n(\cdot) \in \int^\oplus \mathcal{U}^{\beta_{t-1/n}} dt.$$

But from Proposition 3.1 and Corollary 3.1,

$$\mathcal{U}^{z^n} = \mathcal{U}^{\beta_t} \quad \text{for all } t \in (0, T),$$

where $z^n(\cdot)$ is the observation corresponding to $u^n(\cdot)$. Hence

$$(3.14) \quad u^n(\cdot) \in \int^\oplus \mathcal{U}^{z^n - 1/n} dt.$$

(ii) Suppose now that $u(\cdot) \in \int^\oplus \mathcal{U}^{z_t} dt$. Then from Lemma 3.1,

$$\mathcal{U}^{\beta_t} \subset \mathcal{U}^{z_t} \quad \text{for all } t \in (0, T).$$

The condition $u = \lim_n u^n$, with $u^n(\cdot) \in \int^\oplus \mathcal{U}^{z^n - \varepsilon^n} = \int^\oplus \mathcal{U}^{\beta_t - \varepsilon_n} dt$, implies

$$(3.15) \quad u(\cdot) \in \int^\oplus \mathcal{U}^{\beta_t} dt.$$

Hence the inverse inclusion $\mathcal{U}^{\beta_t} \supset \mathcal{U}^{z_t}$ holds for all $t \in [0, T]$.

Remark 3.1. The condition $u(\cdot) \in \int^\oplus \mathcal{U}^{z_t} dt$ is not generally sufficient to imply that

$$\mathcal{U}^{\beta_t} = \mathcal{U}^{z_t} \quad \text{for all } t,$$

as is shown by the following counterexample. Let

$$(3.16) \quad \begin{aligned} \frac{dy}{dt}(t) &= u(t) + f(t), & y(0) &= 0, \\ dz(t) &= y(t) dt + d\eta(t), & z(0) &= 0. \end{aligned}$$

Then $\alpha(t)$ is equal to $\int_0^t f(s) ds$ and we suppose the additional condition

$$(3.17) \quad \beta(t) = \int_0^t \alpha(s) ds + \eta(t) = 0 \quad \text{for all } t.$$

Let us define

$$(3.18) \quad u_0(t) = \xi, \quad 0 \leq t \leq T,$$

where ξ is some random variable with values in U . It is clear that the observation process $z_0(\cdot)$ corresponding to $u_0(\cdot)$ is given by

$$(3.19) \quad z_0(t) = t^2/2\xi.$$

Then we obtain

$$u_0(t) = \frac{2}{t^2} z_0(t),$$

which implies that $u_0(\cdot) \in \int^\oplus \mathcal{U}^{z_0, t} dt$. But according to (3.17), $\mathcal{U}^{\beta_t} \equiv U$ for all $t \in [0, T]$, so that

$$\mathcal{U}^{\beta_t} \neq \mathcal{U}^{z_0, t}.$$

The conclusion is that the unique condition $u(\cdot) \in \int^\oplus \mathcal{U}^{z_t} dt$ is not a correct definition of the admissible controls, because such controls can be dependent on events exterior to the system.

Now, let U_{ad} be a nonempty closed convex subset of U . We define

$$(3.20) \quad \mathcal{W} = \int^\oplus \mathcal{U}^{\beta_t} dt,$$

$$\mathcal{W}_{\text{ad}} = \{u(\cdot) \in \mathcal{W} \mid u(t) \in U_{\text{ad}} \text{ a.s.}\}.$$

Clearly, \mathcal{W}_{ad} is a closed convex subset of \mathcal{W} . Moreover, from Proposition 3.2 and Remark 3.1, the *admissible controls* can be defined by

$$(3.21) \quad \tilde{\mathcal{W}}_{\text{ad}} = \left\{ u(\cdot) \in \mathcal{W}_{\text{ad}} \mid u(\cdot) \in \int^\oplus \mathcal{U}^{z_t} dt \right\},$$

where $z(\cdot)$ is the observation corresponding to $u(\cdot)$. Notice that $\tilde{\mathcal{W}}_{\text{ad}}$ contains at least the open loop controls, i.e., the space $L^2(0, T; U_{\text{ad}})$. Therefore, we are sure that it is not an empty set. Furthermore we have the following *density theorem*.

THEOREM 3.1. $\tilde{\mathcal{W}}_{\text{ad}}$ is dense in \mathcal{W}_{ad} .

Proof. Let $u(\cdot) \in \mathcal{W}_{\text{ad}}$ and set

$$u^n(t) = \begin{cases} u(t - 1/n) & \text{if } t \in [1/n, T], \\ 0 & \text{if } t \in [0, 1/n]. \end{cases}$$

Then $u^n(\cdot)$ converges toward $u(\cdot)$ in $L^2(0, T; \mathcal{U})$ (hence in \mathcal{W}_{ad}), and by Proposition 3.1,

$$u^n(\cdot) \in \int^\oplus \mathcal{U}^{z^{t-1/n}} dt,$$

where z^n is the observation corresponding to u^n . Thus we obtain

$$u^n \in \tilde{\mathcal{W}}_{\text{ad}}.$$

Remark 3.2. A. V. Balakrishnan [1] has considered a finite-dimensional stochastic system, where disturbances are white noises, for which he proved that

$$\tilde{\mathcal{W}}_{\text{ad}} = \mathcal{W}_{\text{ad}}.$$

However, for more general noises, it is not difficult to get an example of the strict inclusion. Indeed, let us consider the system

$$dy(t) = u(t) dt + \xi dt + t dw(t), \quad 0 \leq t \leq T, \quad y(0) = 0,$$

where w is a Wiener process and ξ a random variable independent of w . We assume complete observation: $z(t) = y(t)$. Let us consider

$$u_0(t) = -\xi - \frac{1}{t} \int_0^t s dw(s).$$

Then in our notation,

$$u_0(t) = -\frac{\beta(t)}{t}, \quad \text{so that } u_0(\cdot) \in \int^\oplus \mathcal{U}^{t_0} dt.$$

But for this particular u_0 , the corresponding observation process is

$$z_0(t) = \int_0^t s dw(s) - \int_0^t \frac{ds}{s} \int_0^s \alpha dw(\alpha).$$

In view of the independence of ξ and w , we can conclude that

$$u_0(\cdot) \notin \int^\oplus \mathcal{U}^{z_0, t} dt, \quad \text{hence } u_0 \notin \tilde{\mathcal{W}}_{\text{ad}}.$$

3.3. Admissible controls and feedback functions. We must emphasize the fact that our controls are not directly based upon the notion of feedback. For us, the set of admissible controls is $\tilde{\mathcal{W}}_{\text{ad}}$, and for each $u(t; \omega)$ belonging to $\tilde{\mathcal{W}}_{\text{ad}}$ there exists a unique corresponding pair of processes $y(\cdot)$ and $z(\cdot)$ solving (3.1), (3.2). However, by the definition of $\tilde{\mathcal{W}}_{\text{ad}}$, for any $u \in \tilde{\mathcal{W}}_{\text{ad}}$ there exists a family of mappings $K(t; z_t)$ from

$$C(0, t; F) \rightarrow U_{\text{ad}}$$

such that

$$u(t; \omega) = K(t; z_t(\omega)) \quad \text{a.e. } t, \quad \text{a.s. } \omega,$$

where z is defined by (3.2).

It is very important to notice that if, conversely, we start with a family $K(t; z_t)$ and consider the system of equations

$$\begin{aligned} y(t) + \int_0^t A(s)y(s) ds &= y_0 + \int_0^t f(s) ds + \int_0^t B(s) d\xi(s) + \int_0^t D(s)K(s; z_s) ds, \\ (3.22) \quad z(t) &= \int_0^t C(s)y(s) ds + \eta(t), \end{aligned}$$

then we cannot guarantee existence nor uniqueness of the solution y, z of (3.22). Furthermore, if we have existence, we cannot guarantee even that

$$u(t; \omega) = K(t; z_t(\omega))$$

belongs to $\tilde{\mathcal{W}}_{\text{ad}}$ (for instance, take the counterexample of Remark 3.1, where there is existence and not uniqueness). The problem of uniqueness is discussed in A. Lindquist [8].

This difficulty is overcome by the introduction of a small delay. Let us assume that the family K satisfies

$$K(t; z_t(\omega)) \in L^2((0, T) \times \Omega, dt \otimes d\mu; U)$$

for any $z \in L^2((0, T) \times \Omega, dt \otimes d\mu; F) \cap \text{meas } \{\Omega, \mu; C(0, T; F)\}$. We then define

$$\begin{aligned} K_\varepsilon(t; z_t) &= K(t - \varepsilon; z_{t-\varepsilon}) \quad \text{for } t \geq \varepsilon, \\ K_\varepsilon(t) &\in U_{\text{ad}}, \quad 0 \leq t \leq \varepsilon \quad \text{and} \quad K_\varepsilon(\cdot) \in L^2(0, \varepsilon; U), \end{aligned}$$

and consider the equations (3.22), with K changed into K_ε . By splitting the interval $[0, T]$ into

$$0, \varepsilon, 2\varepsilon, \dots, N\varepsilon, T$$

and applying Theorem 1.1 on each of these subintervals successively, one easily checks that (3.22) define a unique pair $y_\varepsilon, z_\varepsilon$. Furthermore,

$$u_\varepsilon(t; \omega) = K_\varepsilon(t; z_\varepsilon(\omega))$$

belongs to $\tilde{\mathcal{W}}_{\text{ad}}$.

From a practical viewpoint, almost all feedbacks will have a small delay for physical reasons. This justifies the use of $\tilde{\mathcal{W}}_{\text{ad}}$. However, if we impose more conditions on feedbacks K , then we may get

$$u(t; \omega) = K(t; z_t(\omega)) \in \tilde{\mathcal{W}}_{\text{ad}}.$$

This will be proved in the next proposition.

PROPOSITION 3.3. *Let $u(\cdot) \in L^2(0, T; \mathcal{U})$ and suppose that there exists continuous mapping γ from $[0, T] \times C(0, T; F)$ into U_{ad} such that*

(a) *γ is nonanticipative:*

$$\gamma(t, f) = \gamma(t, g) \quad \text{if } f(s) = g(s) \quad \text{for all } s \leq t,$$

(b) *there exists $k > 0$ such that, for any $t \in [0, T]$,*

$$|\gamma(t, f) - \gamma(t, g)| \leq k \sup_{0 \leq s \leq t} |f(s) - g(s)|;$$

(c) *$u(t) = \gamma(t, z_t)$, where $z(\cdot)$ is the observation corresponding to $u(\cdot)$. Then u belongs to $\tilde{\mathcal{W}}_{\text{ad}}$.*

Proof. Clearly $u(\cdot)$ belongs to $\int^\oplus \mathcal{U}^{z_t} dt$. Moreover, according to Lemma 3.1,

$$\mathcal{U}^{B_t} \subset \mathcal{U}^{z_t} \quad \text{for all } t \in [0, T].$$

Now let $\phi(s, \tau)$ be the Green operator associated with $A(t)$. From (3.4), (3.5), it follows that

$$(3.23) \quad z(t) = \int_0^t ds C(s) \int_0^s \phi(s, \tau) D(\tau) \gamma(\tau, z_\tau) d\tau + \beta(t).$$

If we set

$$(3.24) \quad \begin{aligned} z^0(t) &= \beta(t), \\ z^n(t) &= \int_0^t ds \, C(s) \int_0^s \phi(s, \tau) D(\tau) \gamma(\tau, z_\tau^{n-1}) d\tau + \beta(t), \end{aligned}$$

we obtain the majorations

$$\sup_{0 \leq s \leq t} |z^{n+1}(s) - z^n(s)| \leq \text{const.} \int_0^t \sup_{0 \leq \tau \leq s} |z^n(\tau) - z^{n-1}(\tau)| ds \quad (\text{a.s.}),$$

while for some p sufficiently large,

$$(3.25) \quad \sup_{0 \leq s \leq t} |z^{(n+1)p}(s) - z^{np}(s)| \leq l \sup_{0 \leq s \leq t} |z^{np}(s) - z^{(n-1)p}(s)| \quad (\text{a.s.}),$$

with $0 < l < 1$. Then the sequence $\zeta_t^n = z_t^{np}$ converges toward z_t in $L^2(\Omega, \mu; C(0, t; F))$. But by construction

$$\zeta_t^n = \lambda_t^n(\beta_t)$$

for some λ_t^n belonging to $L^2(C(0, t; F), dv_t; C(0, t; F))$. Therefore $\zeta_t^n \in \tilde{L}^2(C(0, t; F), dv_t; C(0, t; F))$ and because this space is closed in $L^2(\Omega, \mu; C(0, t; F))$, the same holds for z_t . This completes the proof.

Notice that the same argument is valid if we suppose

$$u(t) = \gamma(t, \beta_t).$$

It is then enough to consider the sequence

$$(3.26) \quad \begin{aligned} \beta^0(t) &= z(t), \\ \beta^n(t) &= - \int_0^t ds \, C(s) \int_0^s \phi(s, \tau) D(\tau) \gamma(\tau, \beta_\tau^{n-1}) d\tau + z(t). \end{aligned}$$

Thus we have the following proposition.

PROPOSITION 3.4. *Let $u(\cdot)$ belong to $L^2(0, T; \mathcal{U})$ and suppose that there exists a continuous mapping γ from $[0, T] \times C(0, T; F)$ into U_{ad} , such that*

(a) γ is nonanticipative,

$$\gamma(t, f) = \gamma(t, g) \quad \text{if } f(s) = g(s) \quad \text{for all } s \leq t;$$

(b) there exists $k > 0$ such that for any $t \in [0, T]$,

$$|\gamma(t, f) - \gamma(t, g)| \leq k \sup_{0 \leq s \leq t} |f(s) - g(s)|;$$

(c) $u(t) = \gamma(t, \beta_t)$.

Then u belongs to $\tilde{\mathcal{W}}_{\text{ad}}$.

3.4. The payoff function. Let us consider a function $l(y, u, t)$ defined on $H \times U \times [0, T]$, which is continuous in y, u and measurable in t , such that

$$(3.27) \quad |l(y, u, t)| \leq C_1(|y|^2 + |u|^2 + 1),$$

(3.28) $y, u \rightarrow l(y, u, t)$ is convex and Gateaux-differentiable,

$$\left| \frac{\partial l}{\partial y}(y, u, t) \right| \leq C_2(|y| + |u| + 1),$$

$$\left| \frac{\partial l}{\partial u}(y, u, t) \right| \leq C_3(|y| + |u| + 1).$$

Also let Λ be a functional defined on H such that

$$(3.29) \quad |\Lambda(y)| \leq C_4(|y|^2 + 1),$$

(3.30) Λ is convex and Gateaux-differentiable with

$$\left| \frac{d\Lambda}{dy} \right| \leq C_5(|y| + 1).$$

Our aim now is to study the continuity and Gateaux-differentiability properties of the function

$$(3.31) \quad y(\cdot), u(\cdot) \rightarrow E \int_0^T l(y(t), u(t), t) dt + E\Lambda(y(T))$$

from $C(0, T; L^2(\Omega, \mu; H)) \times L^2(0, T, \mathcal{U})$ into R .

LEMMA 3.3. *Under assumptions (3.27)–(3.30), the function (3.31) is continuous and Gateaux-differentiable.*

Proof. (i) Let us consider for any $0 \leq \alpha \leq 1$, $0 < t < T$,

$$\begin{aligned} \Delta_t^\alpha(z, v) &= l(y(t) + \alpha z(t), u(t) + \alpha v(t), t) - l(y(t), u(t), t) \\ &\quad + \frac{1}{T} (\Lambda(y(T) + \alpha z(T)) - \Lambda(y(T))), \end{aligned}$$

where $y(\cdot), z(\cdot) \in C(0, T; L^2(\Omega, \mu; H))$, $u(\cdot), v(\cdot) \in L^2(0, T; \mathcal{U})$. By virtue of the convexity of l and Λ , we obtain the inequality

$$\begin{aligned} (3.32) \quad \Delta_t^\alpha(z, v) &\leq \alpha(l(y(t) + z(t), u(t) + v(t), t) - l(y(t), u(t), t)) \\ &\quad + \frac{\alpha}{T} (\Lambda(y(T) + z(T)) - \Lambda(y(T))) \quad (\text{a.s.}). \end{aligned}$$

Furthermore, from (3.27) and (3.29), we have

$$(3.33) \quad |\Delta_t^\alpha| \leq k\alpha(|y(t) + z(t)|^2 + |u(t) + v(t)|^2 + |y(T) + z(T)|^2 + 1) \quad (\text{a.s.}).$$

Now let (z^n, v^n) be a sequence in $C(0, T; L^2(\Omega, \mu; H)) \times L^2(0, T; \mathcal{U})$ converging toward 0. Since l and Λ are continuous, there exists a (sub)sequence such that

$$\Delta_t^\alpha(z^n, v^n) \rightarrow 0 \quad \text{a.e. } (t, \omega).$$

Moreover (3.33) shows that this sequence is uniformly integrable. Therefore $\Delta_t^\alpha(z^n, v^n)$ is convergent to 0 in $L^1(0, T \times \Omega, dt \otimes d\mu)$. Hence (3.31) is continuous.

(ii) Let us now set

$$\begin{aligned}
 \Gamma_t^\alpha(z, v) = & \frac{1}{\alpha} (l(y(t) + \alpha z(t), u(t) + \alpha v(t), t) - l(y(t), u(t), t)) \\
 (3.34) \quad & - \left(\frac{\partial l}{\partial y}(y(t), u(t), t), z(t) \right) - \left(\frac{\partial l}{\partial u}(y(t), u(t), t), v(t) \right) \\
 & + \frac{1}{\alpha T} (\Lambda(y(T) + \alpha z(T)) - \Lambda(y(T))) - \frac{1}{T} \left(\frac{\partial \Lambda}{\partial y}(y(T)), z(T) \right).
 \end{aligned}$$

The same arguments as before imply

$$|\Gamma_t^\alpha(z, v)| \leq k'(|y(t) + z(t)|^2 + |u(t) + v(t)|^2 + |y(T) + z(T)|^2 + 1),$$

and we can obtain the conclusion as in (i). This completes the proof of the lemma.

According to (3.34) we obtain

$$\begin{aligned}
 (3.35) \quad & \frac{d}{d\alpha} \left(E \int_0^T l(y(t) + \alpha z(t), u(t) + \alpha v(t), t) dt + E\Lambda(y(T) + \alpha z(T)) \right)_{\alpha=0} \\
 & = E \left(\int_0^T \left(\frac{\partial l}{\partial y}(y(t), u(t), t), z(t) \right) dt + \int_0^T \left(\frac{\partial l}{\partial u}(y(t), u(t), t), v(t) \right) dt \right. \\
 & \quad \left. + \left(\frac{d\Lambda}{dy}(y(T)), z(T) \right) \right)
 \end{aligned}$$

for all $y(\cdot), z(\cdot) \in C(0, T; L^2(\Omega, \mu; H))$, $u(\cdot), v(\cdot) \in L^2(0, T; \mathcal{U})$. For $u(\cdot) \in \mathcal{W}$ we set

$$(3.36) \quad J(u(\cdot)) = E \int_0^T l(y(t), u(t), t) dt + E\Lambda(y(T)),$$

where $y(\cdot)$ is a solution of (3.11).

From Theorem 3.1, we obtain the following corollary.

COROLLARY 3.2.

$$(3.37) \quad \inf_{u(\cdot) \in \mathcal{W}_{\text{ad}}} J(u(\cdot)) = \inf_{u(\cdot) \in \tilde{\mathcal{W}}_{\text{ad}}} J(u(\cdot)).$$

Proof. It follows from (3.1) and the energy equality (see Theorem 1.1) that the mapping $u(\cdot) \rightarrow y(\cdot)$ is continuous from $\mathcal{W}_{\text{ad}} \rightarrow C(0, T; L^2(\Omega, d\mu; H))$. This and Lemma 3.3 imply

$$(3.38) \quad u(\cdot) \rightarrow J(u(\cdot)) \text{ is continuous,}$$

and (3.37) is then a consequence of Theorem 3.1.

The optimal control problem is now stated as follows:

$$(3.39) \quad \text{Find } \hat{u}(\cdot) \in \tilde{\mathcal{W}}_{\text{ad}} \text{ such that } J(\hat{u}(\cdot)) = \inf_{u(\cdot) \in \tilde{\mathcal{W}}_{\text{ad}}} J(u(\cdot)).$$

It follows from Corollary 3.2 that $\hat{u}(\cdot)$ is also a solution of the optimal control problem:

$$(3.40) \quad \text{Find } \hat{u}(\cdot) \in \mathcal{W}_{\text{ad}} \text{ such that } J(\hat{u}(\cdot)) = \inf_{u(\cdot) \in \mathcal{W}_{\text{ad}}} J(u(\cdot)).$$

Property (3.40) is very important from the point of view of necessary and sufficient conditions (not for existence). Indeed, the necessary and sufficient conditions for optimality for problem (3.39) are the same as for problem (3.40). However, for problem (3.40) they can be found from the usual variational arguments. General results on necessary and sufficient conditions for optimality in problems similar to (3.40) (for finite-dimensional systems but not necessarily differentiable payoffs) can be found in J. M. Bismut [4].

4. Necessary and sufficient conditions of optimality.

4.1. Variational inequality. We have the following lemma.

LEMMA 4.1. *The function $J(u(\cdot))$ is convex and Gateaux-differentiable in $L^2(0, T; \mathcal{U})$.*

Proof. Convexity follows from the convexity of functions l and Λ . Let $u(\cdot)$ and $v(\cdot) \in L^2(0, T; \mathcal{U})$. We define $\tilde{y}(t; v(\cdot))$ as the solution of

$$(4.1) \quad \tilde{y}(t) + \int_0^t A(\tau) \tilde{y}(\tau) d\tau = \int_0^t D(\tau) v(\tau) d\tau.$$

For α real and $\alpha > 0$, we can write

$$(4.2) \quad \begin{aligned} J(u(\cdot) + \alpha v(\cdot)) &= E \int_0^T l(y(t) + \alpha \tilde{y}(t; v(\cdot)), u(t) + \alpha v(t), t) dt \\ &\quad + E \Lambda(y(T) + \alpha \tilde{y}(T; v(\cdot))). \end{aligned}$$

From (4.2) and Lemma 3.3, it follows that J is Gateaux-differentiable and that

$$(4.3) \quad \begin{aligned} (J'(u(\cdot)), v(\cdot)) &= E \int_0^T \left(\frac{\partial l}{\partial y}(y(t), u(t), t), \tilde{y}(t) \right) dt + E(\Lambda'(y(T)), \tilde{y}(T)) \\ &\quad + E \int_0^T \left(\frac{\partial l}{\partial u}(y(t), u(t), t), v(t) \right) dt, \end{aligned}$$

which completes the proof of Lemma 4.1.

Since \mathcal{W}_{ad} is a closed convex subset of $L^2(0, T; \mathcal{U})$, it follows from standard results that for $\hat{u}(\cdot)$ to be a solution of (3.40), it is necessary and sufficient that

$$(4.4) \quad (J'(\hat{u}(\cdot)), v(\cdot) - \hat{u}(\cdot)) \geq 0 \quad \text{for all } v(\cdot) \in \mathcal{W}_{\text{ad}}.$$

Remark 4.1. When one does not assume convexity of functions l and Λ , then (4.4) remains a necessary (but not sufficient) condition of optimality.

4.2. Stochastic maximum principle. Let us make explicit the relation (4.4). According to (4.3), we have

$$(4.5) \quad \begin{aligned} E \int_0^T \left(\frac{\partial l}{\partial y}(\hat{y}(t), \hat{u}(t), t), \tilde{y}(t; v(\cdot) - \hat{u}(\cdot)) \right) dt + E(\Lambda'(\hat{y}(T)), \tilde{y}(T; v(\cdot) \\ - \hat{u}(\cdot))) + E \int_0^T \left(\frac{\partial l}{\partial u}(\hat{y}(t), \hat{u}(t), t), v(t) - \hat{u}(t) \right) dt \geq 0. \end{aligned}$$

Let us now introduce \hat{p} as the solution of

$$(4.6) \quad -\frac{d\hat{p}}{dt} + A^*(t)\hat{p} = -\frac{\partial l}{\partial y}(\hat{y}(t), \hat{u}(t), t),$$

$$\hat{p}(T) = -\Lambda'(\hat{y}(T)).$$

Using (4.6) in (4.5) and integrating by parts (for fixed ω), we get

$$(4.7) \quad E \int_0^T \left(-D^*(t)\hat{p}(t) + \frac{\partial l}{\partial u}(\hat{y}(t), \hat{u}(t), t), v(t) - \hat{u}(t) \right) dt \geq 0 \quad \text{for all } v(\cdot) \in \mathcal{W}_{\text{ad}}.$$

Since \mathcal{W}_{ad} is a closed convex subset of $\mathcal{W} = \int^\oplus \mathcal{U}^{\beta_t} dt$, it follows from the definition of the projection on $\int^\oplus \mathcal{U}^{\beta_t} dt$ (as a sub-Hilbert space of $L^2(0, T; \mathcal{U})$) that (4.7) can be rewritten as

$$(4.8) \quad E \int_0^T \left(E^{\beta_t} \left(-D^*(t)\hat{p}(t) + \frac{\partial l}{\partial u}(\hat{y}(t), \hat{u}(t), t), v(t) - \hat{u}(t) \right) \right) dt \geq 0,$$

where E^{β_t} denotes the conditional expectation with respect to β_t (see §2.2).

Now let $v \in U_{\text{ad}}$, and let us introduce

$$(4.9) \quad \lambda(t, \omega) = E^{\beta_t} \left\{ \left(-D^*(t)\hat{p}(t) + \frac{\partial l}{\partial u}(\hat{y}(t), \hat{u}(t), t), v - \hat{u}(t) \right) \right\}.$$

We shall prove that

$$(4.10) \quad \lambda(t, \omega) \geq 0 \quad \text{a.e. } t, \quad \text{a.s. } \omega.$$

Indeed, let us set

$$(4.11) \quad A = \{t, \omega | \lambda(t, \omega) < 0\}$$

and take

$$(4.12) \quad v_0(t, \omega) = \begin{cases} v & \text{in } A, \\ \hat{u}(t, \omega) & \text{outside } A. \end{cases}$$

According to Proposition 2.1, $\lambda(t) \in \int^\oplus \mathcal{U}^{\beta_t} dt$. Thus, if we denote by $\chi_A(t)$ the characteristic function of A , then

$$\chi_A(\cdot) \in \int^\oplus \mathcal{U}^{\beta_t} dt.$$

From this we get

$$(4.13) \quad v_0(t) = \chi_A(t)v + (1 - \chi_A(t))\hat{u}(t) \in \int^\oplus \mathcal{U}^{\beta_t} dt,$$

which proves that $v_0(\cdot) \in \mathcal{W}_{\text{ad}}$.

Taking this particular choice of $v_0(\cdot)$ in (4.8), we get

$$(4.14) \quad \int_A \lambda(t, \omega) dt dP \geq 0,$$

contradicting (4.10) if A has measure different from 0. Hence the relation (4.10) holds, i.e.,

$$(4.15) \quad \left(-D^*(s)E^{\beta_s}\hat{p}(s) + E^{\beta_s}\frac{\partial l}{\partial u}(\hat{y}(s), \hat{u}(s), s), v - \hat{u}(s) \right) \geq 0 \quad \text{for all } v \in U_{\text{ad}}, \quad \text{a.e. } s, \quad \text{a.s. } \omega.$$

Now, (4.15) must hold if $\hat{u}(\cdot)$ is a solution of problem (3.40). If $\hat{u}(\cdot)$ solves problem (3.39), we know by the argument of Corollary 3.2 that $\hat{u}(\cdot)$ solves problem (3.40) and hence satisfies (4.15). But, there is then something more. According to (3.21) and Corollary 3.1, $\mathcal{U}^{z_t} = \mathcal{U}^{\beta_t}$, which implies $E^{\beta_t} = E^{z_t}$. We have thus proven the following.

THEOREM 4.1 (Stochastic maximum principle). *Assume the notations and assumptions of §§1.1 and 3.1. Then for $\hat{u}(\cdot)$ to be a solution of problem (3.39), it is necessary and sufficient that, denoting by $\hat{y}(\cdot)$ the corresponding trajectory, and defining $\hat{p}(\cdot)$ by (4.6), the following condition holds:*

$$(4.16) \quad \left(-D^*(t)E^{z_t}\hat{p}(t) + E^{z_t}\frac{\partial l}{\partial u}(\hat{y}(t), \hat{u}(t), t), v - \hat{u}(t) \right) \geq 0 \quad \text{a.e. } t, \quad \text{a.s. } \omega \quad \text{for all } v \in U_{\text{ad}}.$$

This is only a necessary condition, if one leaves out the convexity assumptions.

5. The separation principle.

5.1. Assumptions. We suppose now that the observation noise $\eta(t)$ is independent of the processes $f(t)$ and $\xi(t)$. Let us consider the quadratic case with

$$(5.1) \quad \begin{aligned} l(y, u, t) &= \frac{1}{2}(L(t)y, y) + \frac{1}{2}(N(t)u, u) + (g(t), y), \\ \Lambda(y) &= \frac{1}{2}(My, y), \\ U_{\text{ad}} &= U. \end{aligned}$$

In this case a unique optimal solution of (3.40) exists. Let (\hat{y}, \hat{u}) be such a solution. Then the adjoint system (4.6) now becomes

$$(5.2) \quad -\frac{d\hat{p}}{dt} + A^*(t)\hat{p} = -L(t)\hat{y}(t) - g(t), \quad \hat{p}(T) = -M\hat{y}(T)$$

and (4.16) becomes

$$(5.3) \quad (-D^*(t)E^{\beta_t}\hat{p}(t) + E^{\beta_t}N(t)\hat{u}(t), v - \hat{u}(t)) \geq 0 \quad \text{for all } v \in U, \quad \text{a.e. } t, \quad \text{a.s. } \omega.$$

It follows from (5.3) that

$$(5.4) \quad \hat{u}(t) = N^{-1}(t)D^*(t)E^{\beta_t}\hat{p}(t).$$

5.2. Decoupling. Let us set

$$(5.5) \quad \begin{aligned} \hat{y}_t(\tau) &= E^{\beta_t}\hat{y}(\tau), \quad \tau \geq t, \\ \hat{p}_t(\tau) &= E^{\beta_t}\hat{p}(\tau). \end{aligned}$$

According to (3.1),

$$(5.6) \quad \begin{aligned} \hat{y}(\tau) &= \phi(\tau, t)\hat{y}(t) + \int_t^\tau \phi(\tau, s)D(s)\hat{u}(s) ds \\ &+ \int_t^\tau \phi(\tau, s)f(s) ds + \int_t^\tau \phi(\tau, s)B(s) d\xi(s), \end{aligned}$$

where $\phi(t_1, t_2)$ is the Green operator associated with $A(t)$. Then

$$(5.7) \quad E^{\beta_t}\hat{y}(\tau) = \phi(\tau, t)E^{\beta_t}\hat{y}(t) + \int_t^\tau \phi(\tau, s)D(s)E^{\beta_t}\hat{u}(s) ds + \int_t^\tau \phi(\tau, s)E^{\beta_t}f(s) ds.$$

To get (5.7) from (5.6), we notice that if we write

$$\hat{u}(s) = E^{\beta_t}\hat{u}(s) + \tilde{u}(s),$$

then $\tilde{u}(s)$ is orthogonal to \mathcal{U}^{β_t} . Therefore for $\lambda \in L^2(C(0, t; F), dv_t; H)$, we have

$$\begin{aligned} E\left(\int_t^\tau \phi(\tau, s)D(s)\tilde{u}(s) ds, \lambda(\beta_t)\right) &= E\int_t^\tau (\tilde{u}(s), D^*(s)\phi^*(\tau, s)\lambda(\beta_t)) ds \\ &= \int_t^\tau E(\tilde{u}(s), D^*(s)\phi^*(\tau, s)\lambda(\beta_t)) ds = 0. \end{aligned}$$

Hence

$$E^{\beta_t} \int_t^\tau \phi(\tau, s)D(s)\tilde{u}(s) ds = 0.$$

A similar argument holds for the second integral in the right-hand side of (5.7).

Now

$$E^{\beta_t} \int_t^\tau \phi(\tau, s)B(s) d\xi(s) = 0,$$

which is an easy consequence of assumptions (1.10) and the independence of the processes ξ and η . Using (5.4) in (5.7) we get that $\hat{y}_t(\tau)$ solves

$$(5.8) \quad \begin{aligned} \frac{d\hat{y}_t(\tau)}{d\tau} + A(\tau)\hat{y}(\tau) &= D(\tau)N^{-1}(\tau)D^*(\tau)\hat{p}_t(\tau) + E^{\beta_t}f(\tau), \\ \hat{y}_t(t) &= \text{given}, \quad \tau > t. \end{aligned}$$

Similarly from (5.2) it follows that $\hat{p}_t(\tau)$ solves

$$(5.9) \quad \begin{aligned} -\frac{d\hat{p}_t(\tau)}{d\tau} + A^*(\tau)\hat{p}_t(\tau) &= -L(\tau)\hat{y}_t(\tau) - g(\tau), \\ \hat{p}_t(\tau) &= -M\hat{y}_t(T), \quad \tau > t. \end{aligned}$$

According to the well-known deterministic decoupling theory of Lions [9], introducing the Riccati equation

$$(5.10) \quad \frac{dP}{dt} - PA - A^*P - PDN^{-1}D^*P + L = 0, \quad P(T) = M$$

and the stochastic process r solving

$$(5.11) \quad \frac{dr}{dt} - A^*r - PDN^{-1}D^*r + Pf + g = 0, \quad r(T) = 0,$$

we have that the solutions $\hat{y}_t(\tau)$ and $\hat{p}_t(\tau)$ of (5.8) and (5.9) satisfy the following relationship:

$$(5.12) \quad \hat{p}_t(\tau) = -(P(\tau)\hat{y}_t(\tau) + E^{\beta_t}r(\tau)).$$

Therefore the process

$$(5.13) \quad \hat{u}(t) = -N^{-1}(t)D^*(t)(P(t)E^{\beta_t}\hat{y}(t) + E^{\beta_t}r(t))$$

is the optimal solution of the problem (3.40). For the problem (3.39), we now obtain the following result.

THEOREM 5.1. *Under the assumptions of §5.1, the sequence*

$$(5.14) \quad u_n(t) = -N^{-1}(t)D^*(t)(P(t)E^{\beta_t-1/n}y_n(t) + E^{\beta_t-1/n}r(t))$$

is a minimizing sequence for problem (3.39), which converges towards (5.13) in \mathcal{W}_{ad} . Moreover, if there exists an optimal control of (3.39), it is necessarily defined by the following feedback rule:

$$(5.15) \quad \hat{u}(t) = -N^{-1}(t)D^*(t)(P(t)E^{z_t}\hat{y}(t) + E^{z_t}r(t)),$$

where $z(\cdot)$ is the observation process corresponding to $\hat{u}(\cdot)$.

Proof. (i) Since the operator $E^{\beta_t-1/n}(\cdot)$ is a contraction, it is easy to prove that equation (3.1) has a unique solution with u_n given by (5.14). But according to Proposition 3.1, u_n belongs to $\tilde{\mathcal{W}}_{\text{ad}}$. And from (5.13), it is obvious that u_n is a minimizing sequence of problem (3.39).

(ii) If there exists an optimal control, it satisfies (5.13) (according to uniqueness and Corollary 3.2). The result follows now from the identity of the operators E^{β_t} , E^{z_t} for admissible controls.

Concerning the existence of an optimal control of problem (3.39), we can give a result under the additional properties:

$$(5.16) \quad y_0 \text{ is Gaussian, } Ey_0 = \bar{y}_0 \text{ and } E(y_0 - \bar{y}_0, h_1)(y_0 - \bar{y}_0, h_2) = (P_0h_1, h_2),$$

where P_0 is a trace operator in H ;

$$(5.17) \quad f(t) \text{ is a deterministic function;}$$

$$(5.18) \quad F \text{ is finite-dimensional space and } \eta(t) \text{ is a Wiener process with values in } F \text{ independent of the Wiener process } \xi(t). \text{ We denote by } R(t) \text{ its invertible covariance matrix.}$$

THEOREM 5.2. *If conditions (5.16) to (5.18) are satisfied, there exists an optimal control of problem (3.39).*

Proof. Since f is deterministic, then the solution r of (5.11) is also deterministic. Let us consider the Riccati equation

$$(5.19) \quad \frac{d\Pi}{dt} + A\Pi + \Pi A^* + \Pi C^*R^{-1}C\Pi = BQB^*, \quad \Pi(0) = P_0$$

and the system of stochastic operational differential equations of the type (1.12) (where the unknown functions are \hat{y} , \hat{s} , \hat{z}):

$$(5.20) \quad \begin{aligned} \hat{y}(t) + \int_0^t A(\tau)\hat{y}(\tau) d\tau &= y_0 + \int_0^t f(\tau) d\tau + \int_0^t B(\tau) d\zeta(\tau) \\ &\quad - \int_0^t D(\tau)N^{-1}(\tau)D^*(\tau)P(\tau)\hat{s}(\tau) d\tau - \int_0^t D(\tau)N^{-1}(\tau)D^*(\tau)r(\tau) d\tau, \end{aligned}$$

$$(5.21) \quad \begin{aligned} \hat{s}(t) + \int_0^t A(\tau)\hat{s}(\tau) d\tau + \int_0^t \Pi(\tau)C^*(\tau)R^{-1}(\tau)C(\tau)\hat{s}(\tau) d\tau \\ = \bar{y}_0 + \int_0^t f(\tau) d\tau - \int_0^t D(\tau)N^{-1}(\tau)D^*(\tau)P(\tau)\hat{s}(\tau) d\tau \\ - \int_0^t D(\tau)N^{-1}(\tau)D^*(\tau)r(\tau) d\tau + \int_0^t \Pi(\tau)C^*(\tau)R^{-1}(\tau)C(\tau)\hat{y}(\tau) d\tau \\ + \int_0^t \Pi(\tau)C^*(\tau)R^{-1}(\tau) d\eta(\tau), \end{aligned}$$

$$(5.22) \quad \hat{z}(t) = \int_0^t C(\tau)\hat{y}(\tau) d\tau + \eta(t).$$

Theorem 1.1 applies to the system of equations (5.20), (5.21) defining \hat{y} in $L^2(0, T \times \Omega, dt \otimes d\mu; H)$ (in particular). The interesting feature is the following:

$$(5.23) \quad \hat{s}(t) = E^{\hat{z}_t}\hat{y}(t).$$

For the proof of (5.23), see A. Bensoussan [3]. Therefore $\hat{u}(\cdot)$ defined by

$$(5.24) \quad \hat{u}(t) = -N^{-1}(t)D^*(t)(P(t)\hat{s}(t) + r(t))$$

is an admissible control for problem (3.39), and thus optimal according to Theorem 5.1.

Remark 5.1. Formula (5.24) expresses the separation principle, since the feedback rule is exactly the deterministic one, with $\hat{y}(t)$ changed into $\hat{s}(t) = E^{\hat{z}_t}\hat{y}(t)$.

6. Example. Let \mathcal{O} be an open subset of R^n . Let $a_{ij}(x, t)$ be functions belonging to $L^\infty(\mathcal{O} \times]0, T[)$, such that

$$(6.1) \quad \sum_{i,j=1}^n a_{ij}(x, t)\xi_i\xi_j \geq \alpha \left(\sum_{i=1}^n \xi_i^2 \right), \quad \alpha > 0, \quad \text{for all } \xi_i \in R, \quad \text{a.e. } x, t.$$

Let $H^1(\mathcal{O})$ be the Sobolev space defined as follows:

$$(6.2) \quad H^1(\mathcal{O}) = \left\{ z \in L^2(\mathcal{O}) \mid \frac{\partial z}{\partial x_i} \in L^2(\mathcal{O}) \right\}.$$

For $z_1, z_2 \in H^1(\mathcal{O})$, set

$$(6.3) \quad a(t; z_1, z_2) = \int_{\mathcal{O}} \sum_{i,j} a_{ij}(x, t) \frac{\partial z_1}{\partial x_j} \frac{\partial z_2}{\partial x_i} dx.$$

Take $H = L^2(\mathcal{O})$, $V = H_0^1(\mathcal{O}) = \text{closure of } \mathcal{D}(\mathcal{O}) \text{ in } H^1(\mathcal{O})$, where $\mathcal{D}(\mathcal{O})$ denotes the space of infinitely differentiable functions in \mathcal{O} with compact support. We define $A(t) \in \mathcal{L}(V; V')$ by

$$(6.4) \quad a(t; z_1, z_2) = \langle A(t)z_1, z_2 \rangle \quad \text{for all } z_1, z_2 \in V.$$

Take $U = H$ (distributed control) and $D(t) = I$.

Let us now define the stochastic features of the problem. Take $y_0 = \text{deterministic element}$; then $E = R$ and

$$B(t) = \sigma(x, t),$$

$$\xi(t) = \text{standard one-dimensional Wiener process.}$$

We also take $f = 0$. We can write equation (3.1) as

$$(6.5) \quad \begin{aligned} y(t, x) - \int_0^t \sum_{i,j=1}^n \frac{\partial}{\partial x_i} \left(a_{ij}(x, \tau) \frac{\partial y(\tau, x)}{\partial x_j} \right) d\tau \\ = y_0(x) + \int_0^t u(\tau, x) d\tau + \int_0^t \sigma(\tau, x) d\xi(\tau) \quad \text{for all } t, \text{ a.e. } x, \\ y(t, x) = 0, \quad t, x \in \partial(\mathcal{O} \times]0, T[). \end{aligned}$$

The observation process z is defined as follows. Take $F = R^p$ and

$$(6.6) \quad C(t)z = Cz = \int_{\mathcal{O}_1} z(x) dx, \dots, \int_{\mathcal{O}_p} z(x) dx,$$

where $\mathcal{O}_1, \dots, \mathcal{O}_p$ are nonoverlapping balls of \mathcal{O} . Then $\eta(t)$ is a Wiener process in R^p , with zero mean and covariance matrix $\sigma_1^2 I$. The payoff will be defined as

$$(6.7) \quad J = \int_0^T \int_{\mathcal{O}} E(y(x, t))^2 dx dt + \int_0^T \int_{\mathcal{O}} ENu^2(x, t) dx dt,$$

where N is a positive constant. The separation principle can be expressed as follows:

$$(6.8) \quad \hat{u}(x, t) = \int_{\mathcal{O}} K_0(x, \xi, t) \hat{\xi}(\xi, t) d\xi,$$

where K_0 is a solution of the following Riccati equation:

$$(6.9) \quad \begin{aligned} & - \frac{\partial K_0}{\partial t}(x, \xi, t) + (A_x^* + A_\xi^*)K_0(x, \xi, t) \\ & + \int_{\mathcal{O}} K_0(x, \xi_1, t) N^{-1} K_0(\xi_1, \xi, t) d\xi_1 = \delta(x - \xi), \\ & K_0(x, \xi, t) = K_0(\xi, x, t), \\ & K_0(x, \xi, t) = 0 \quad \text{if } x \in \Gamma, \xi \in \mathcal{O} \quad (\text{and thus } K_0(x, \xi, t) = 0 \text{ if } x \in \mathcal{O}, \xi \in \Gamma), \\ & K_0(x, \xi, T) = 0. \end{aligned}$$

In (6.8), $\hat{s}(t)$ is the best estimate of the optimal state $\hat{y}(t)$ and is given by the generalized Kalman filter

$$\begin{aligned}
 \hat{s}(t, x) - \int_0^t ds \left(\sum_{i,j=1}^n a_{ij}(x, s) \frac{\partial \hat{s}}{\partial x_j} \right) &= y_0(x) + \int_0^t ds \int_{\mathcal{O}} K_0(x, \xi, s) \hat{s}(s, \xi) d\xi \\
 (6.10) \quad &+ \frac{1}{\sigma_1^2} \sum_{i=1}^p \int_0^t \left(\int_{\mathcal{O}_i} \Pi(x, \xi, s) d\xi \right) \left(dz_i - ds \int_{\mathcal{O}_i} \hat{s}(\xi, s) d\xi \right), \\
 \hat{s}(t, x) &= 0 \quad \text{on } \partial(\Omega \times]0, T[).
 \end{aligned}$$

In (6.10), Π solves the Riccati equation:

$$\begin{aligned}
 \frac{\partial \Pi}{\partial t}(x, \xi, t) - \sum_{i,j=1}^n \frac{\partial}{\partial x_i} \left(a_{ij}(x, t) \frac{\partial \Pi}{\partial x_j}(x, \xi, t) \right) - \sum_{i,j=1}^n \frac{\partial}{\partial \xi_i} \left(a_{ij}(\xi, t) \frac{\partial \Pi}{\partial \xi_j}(x, \xi, t) \right) \\
 (6.11) \quad &+ \frac{1}{\sigma_1^2} \sum_{i=1}^p \int_{\mathcal{O}_i} \Pi(x, \eta, t) d\eta \int_{\mathcal{O}_i} \Pi(\eta, \xi, t) d\eta = \sigma(x, t) \sigma(\xi, t), \\
 \Pi(x, \xi, t) &= \Pi(\xi, x, t), \\
 \Pi(x, \xi, t) &= 0 \quad \text{if } x \in \Gamma, \xi \in \mathcal{O} \quad (\text{and thus } \Pi(x, \xi, t) = 0 \text{ if } x \in \mathcal{O}, \xi \in \Gamma), \\
 \Pi(x, \xi, 0) &= 0.
 \end{aligned}$$

REFERENCES

- [1] A. V. BALAKRISHNAN, *Stochastic control: A function space approach*, this Journal, 10 (1972), pp. 285–297.
- [2] A. BENSOUSSAN, *On the separation principle for distributed parameter systems*, IFAC Conference on Control for Distributed Parameter Systems, Banff, Canada, 1971.
- [3] ———, *Filtrage optimal des syst mes linéaires*, Dunod, Paris, 1971.
- [4] J. M. BISMUT, Thesis, Paris, 1971.
- [5] R. A. BROOKS, *Linear stochastic control: An extended separation principle*, J. Math. Anal. Appl., 38 (1972), pp. 569–587.
- [6] W. H. FLEMING, *Optimal continuous-parameter stochastic control*, SIAM Rev., 11 (1969), pp. 470–509.
- [7] H. J. KUSHNER, *On the optimal control of a system governed by a linear parabolic equation with white noise inputs*, this Journal, 6 (1968), pp. 596–614.
- [8] A. LINDQUIST, *On feedback control of linear stochastic systems*, this Journal, 11 (1973), pp. 323–343.
- [9] J. L. LIONS, *Contrôle optimal des systèmes gouvernés par des équations aux dérivées partielles*, Dunod, Paris, 1968.
- [10] P. A. MEYER, *Probabilités et potentiel*, Hermann, Paris, 1966.
- [11] L. SCHWARTZ, *Radon Measures on Arbitrary Topological Spaces*, Tata Institute of Fundamental Research, Bombay.
- [12] M. VIOT, *Théorème d'optimalité pour des systèmes stochastiques où la commande est adaptée à l'état*, Rev. Française Informat. Recherche Operationnelle, 10 (1968), pp. 115–128.
- [13] W. M. WONHAM, *On the separation theorem of stochastic control*, this Journal, 6 (1968), pp. 312–326.

ON THE MODELING OF SYSTEMS FOR IDENTIFICATION. PART I: ε -REPRESENTATIONS OF CLASSES OF SYSTEMS*

WILLIAM L. ROOT†

Abstract. The notion of a class of systems is formalized. The notion of an approximating class of systems is formalized with the definitions of ε -approximations and of ε -representations. Basic properties of and relationships between ε -representations are established. A fundamental existence theorem for ε -representations is proved, as well as a theorem giving ε -representations from ε -approximations. A construction of ε -approximations using Volterra polynomials is given.

Introduction. In this paper, Part I, and also in the succeeding Part II, subtitled *Time-varying systems*, there is presented an abstract theory of system models to be used in the identification of unknown systems. The basic ideas here are, (i) to formalize the obvious concept of a class of input-output systems, chosen so that the unknown system must be a member of the class, and (ii) to construct parametrized uniformly good approximate representations for a class of systems. Such a representation then gives a model to be used for (approximate) identification of a system in the class. These ideas are developed in §§1 and 2. The approximate representations are called ε -representations; the chief result is Proposition I.7, which is a constructive existence proof that “good” ε -representations exist under rather general conditions.

The theory is not restricted to linear or to time-invariant systems, nor are there very serious constraints on the generality of input and output spaces. Indeed, for the fundamental theory of §§1 and 2, the inputs and outputs do not have to be interpreted as functions of time, although that is what is intended for the applications. In §3 the classical integral polynomials of Volterra and Fréchet are introduced to provide examples.

The concept of an ε -representation was introduced in a conference paper [1], where some general results were stated but only the short proofs were included. The material here is an expansion of the basic portion of [1], with proofs.

Part II is a less general, much more detailed study of time-varying input-output systems. One of the chief reasons for the development in this paper is for application to such systems. However, the theory of Part II is formally independent of the results here except for an application of Proposition I.7. Some of the earlier results of Part II have been stated, again with only partial proofs, in the conference paper [2]. There is really no identification theory as such either here or in Part II; in particular, no statistical problems are mentioned. Identification of unknown systems with additive observation noise using the concepts of this paper is discussed in somewhat abridged form in [3], and touched on in [1].

* Received by the editors May 17, 1973.

† Department of Aerospace Engineering, College of Engineering, University of Michigan, Ann Arbor, Michigan 48105. This work was supported by the United States Air Force, Air Force Office of Scientific Research, Air Force Systems Command, under Grant 72-2328.

Basic definitions. A class of systems \mathcal{S} is defined to be $\mathcal{S} = (\mathcal{Y}, f, \mathcal{X}, \mathcal{U})$, where \mathcal{Y} is a Banach space, \mathcal{X} and \mathcal{U} are metric spaces, and f is a continuous mapping from the topological product of \mathcal{X} and \mathcal{U} , $\mathcal{X} \times \mathcal{U}$, into \mathcal{Y} . \mathcal{U} , \mathcal{Y} and \mathcal{X} are called, respectively, the input space, output space and system parameter space. The output $y \in \mathcal{Y}$ corresponding to input $u \in \mathcal{U}$ and system parameter $x \in \mathcal{X}$ is given by $y = f(x, u)$. The quantity x is to be interpreted as having the role of fixing one system from among all those in the class; thus, $s = (\mathcal{Y}, f, x, \mathcal{U})$ is called a system, for each $x \in \mathcal{X}$.

Further conditions will be imposed on the spaces \mathcal{Y} , \mathcal{X} and \mathcal{U} and the mapping f as needed, but some preliminary structure will be established at the level of generality of this definition. It is to be noted that the definition is formally symmetric in \mathcal{X} and \mathcal{U} . If one is considering a dynamical system, as the term is usually used, \mathcal{Y} and \mathcal{U} will be spaces of functions of a real variable (time), and the concept of state may be relevant. Since there is no explicit provision for state in the definition above, different initial states, in what would normally be described as one system, must be represented by different values of the "system parameter" x . Thus, what is normally called one system may be described here as a class of systems. This unconventional terminology comes about accidentally; the emphasis is to be on situations in which \mathcal{X} is interpretable as a parameter space in the ordinary sense, in which case the terminology is appropriate. The basic definition is further elaborated for use in situations in which there is any real consideration of state.

Let $\mathcal{S} = (\mathcal{Y}, f, \mathcal{X}, \mathcal{U})$ be a class of systems. A system $s \in \mathcal{S}$ with parameter x is *linear* if \mathcal{U} is a linear metric space and $f(x, \cdot)$ is a linear mapping from \mathcal{U} into \mathcal{Y} . (Note that this definition, although nearly inevitable in the context, agrees with the usual definition of a linear dynamical system only when the dynamical system is initially in a relaxed state). The class of systems \mathcal{S} is a *linear class* if \mathcal{X} is a linear metric space and $f(\cdot, u)$ is a linear mapping from \mathcal{X} into \mathcal{Y} for each $u \in \mathcal{U}$. We shall usually need to require that \mathcal{X} and \mathcal{U} be compact or at least bounded, so that they cannot be normed linear spaces. However, they may be subsets of linear spaces, so the following definitions are introduced. A system $s \in \mathcal{S}$ with parameter x is *prelinear* if \mathcal{U} is a subset of a linear metric space and $f(x, \alpha u_1 + \beta u_2) = \alpha f(x, u_1) + \beta f(x, u_2)$ for all $u_1, u_2 \in \mathcal{U}$ and all scalars α and β for which all three terms are defined. The class \mathcal{S} is a *prelinear class* if \mathcal{X} is a subset of a linear metric space, and $f(\alpha x_1 + \beta x_2, u) = \alpha f(x_1, u) + \beta f(x_2, u)$ for all $u \in \mathcal{U}$ whenever scalars α, β and parameters x_1, x_2 are such that all three terms are defined.

The class of systems \mathcal{S} is a *redundant class* if there are parameters $x_1, x_2 \in \mathcal{X}$, $x_1 \neq x_2$, such that $f(x_1, u) = f(x_2, u)$ for all $u \in \mathcal{U}$; if this does not happen, \mathcal{S} is *nonredundant*. If \mathcal{X}' is a subset of \mathcal{X} , then \mathcal{X}' is a metric space, and the restriction f' of f to $\mathcal{X}' \times \mathcal{U}$ is continuous. Hence, $\mathcal{S}' = (\mathcal{Y}, f', \mathcal{X}', \mathcal{U})$ is a class of systems; it is called a *subclass* of \mathcal{S} . Obviously if \mathcal{S} is a prelinear class, then \mathcal{S}' is also, and if \mathcal{S} is nonredundant, then \mathcal{S}' is also. If \mathcal{U}'' is a subset of \mathcal{U} , then $\mathcal{S}'' = (\mathcal{Y}, f'', \mathcal{X}, \mathcal{U}'')$ is a class of systems, where f'' is the restriction of f to $\mathcal{X} \times \mathcal{U}''$. \mathcal{S}'' is called a *restriction* of \mathcal{S} . If \mathcal{S} is prelinear, so is \mathcal{S}'' , and if \mathcal{S} is redundant, so is \mathcal{S}'' . Henceforth we shall not usually bother to use a different symbol for the function f when its domain is restricted either in \mathcal{X} or \mathcal{U} .

The classes $\mathcal{S}_1 = (\mathcal{Y}_1, f_1, \mathcal{X}_1, \mathcal{U}_1)$ and $\mathcal{S}_2 = (\mathcal{Y}_2, f_2, \mathcal{X}_2, \mathcal{U}_2)$ are said to be *equivalent* if there exist homeomorphisms of \mathcal{X}_1 onto \mathcal{X}_2 and \mathcal{U}_1 onto \mathcal{U}_2 and a

linear homeomorphism of \mathcal{Y}_1 on to \mathcal{Y}_2 such that if x_2, u_2 and y_2 are the images of x_1, u_1 and y_1 , then $y_2 = f_2(u_2, x_2)$ if and only if $y_1 = f_1(u_1, x_1)$. Redundancy or nonredundancy is preserved under equivalence, but nothing can be said about preservation of any of the linearity or prelinearity properties because there is no reference to any algebraic structure on $\mathcal{X}_1, \mathcal{U}_1, \mathcal{X}_2$ or \mathcal{U}_2 . If \mathcal{S}'_1 is a subclass of \mathcal{S}_1 , then there is a subclass \mathcal{S}'_2 of \mathcal{S}_2 that is equivalent to \mathcal{S}'_1 . An analogous statement can be made about restrictions.

Henceforth it is assumed that all classes \mathcal{S} consist only of *bounded* systems, that is, that for each $x \in \mathcal{X}$, $f(x, \cdot)$ is a bounded mapping.

1. Natural representations. Given an input space \mathcal{U} and output space \mathcal{Y} , one can consider the set of all bounded continuous mappings from \mathcal{U} into \mathcal{Y} as a sort of universal class of systems. One can then represent any particular class with the given \mathcal{U} and \mathcal{Y} as a subclass of this universal class. This is purely an artifice and introduces nothing new, but it provides a convenient structure within which one can discuss the approximation of one class by another.

Let $\mathcal{F} = \mathcal{F}(\mathcal{U}, \mathcal{Y})$ be the set of all bounded continuous mappings from \mathcal{U} into \mathcal{Y} . \mathcal{F} is made into a Banach space in the following standard way (see [4]):

- (i) Define $\alpha F_1 + \beta F_2$, where α, β are scalars and $F_1, F_2 \in \mathcal{F}$ by

$$(\alpha F_1 + \beta F_2)(u) = \alpha F_1(u) + \beta F_2(u), \quad u \in \mathcal{U}.$$

- (ii) Define $\|F\|$, $F \in \mathcal{F}$, by

$$\|F\| = \sup_{u \in \mathcal{U}} \|F(u)\|.$$

\mathcal{F} is then indeed a complete normed linear space, and henceforth the symbol \mathcal{F} , or $\mathcal{F}(\mathcal{U}, \mathcal{Y})$ when necessary, will always denote this space. Now, the equation $y = F(u)$ can be rewritten $y = g(F, u)$, where, in fact, g is defined by $g(F, u) = F(u)$, $F \in \mathcal{F}$, $u \in \mathcal{U}$. We then have the following proposition.

PROPOSITION I.1. $\mathcal{S}_{\mathcal{F}} = (\mathcal{Y}, g, \mathcal{F}, \mathcal{U})$ is a nonredundant linear class of systems.

Proof. All that needs to be proved is that g is continuous in $\mathcal{F} \times \mathcal{U}$ and linear in \mathcal{F} . Let $(F, u) \in \mathcal{F} \times \mathcal{U}$ and take $\varepsilon > 0$. Choose $\delta > 0$ so that $\|F(u) - F(u_1)\| \leq \varepsilon/2$ whenever $d(u, u_1) \leq \delta$.¹ Let (F', u') be any pair such that $d(u, u') \leq \delta$ and $\|F - F'\| \leq \varepsilon/2$. Then

$$\begin{aligned} \|g(F, u) - g(F', u')\| &\leq \|F(u) - F(u')\| + \|F(u') - F'(u')\| \\ &\leq \varepsilon/2 + \|F - F'\| \leq \varepsilon. \end{aligned}$$

The linearity follows immediately from the definitions:

$$\begin{aligned} g(\alpha F_1 + \beta F_2, u) &= (\alpha F_1 + \beta F_2)(u) = \alpha F_1(u) + \beta F_2(u) \\ &= \alpha g(F_1, u) + \beta g(F_2, u). \quad \square \end{aligned}$$

Let $\mathcal{S} = (\mathcal{Y}, f, \mathcal{X}, \mathcal{U})$ be any class of bounded systems. A subclass of $\mathcal{S}_{\mathcal{F}(\mathcal{U}, \mathcal{Y})}$ is assigned to it as follows. Since $f(x, \cdot)$ is a bounded continuous function, it can be written as $F(\cdot)$, $F \in \mathcal{F}$. Let ψ be the mapping from \mathcal{X} into \mathcal{F} defined by $\psi(x) = F$, where $F(u) = f(x, u)$ for all $u \in \mathcal{U}$. Let $\mathcal{H} = \psi(\mathcal{X})$. Then $\mathcal{S}_0 = (\mathcal{Y}, g, \mathcal{H}, \mathcal{U})$ is a subclass

¹ The symbol $d(u, u')$ means the distance from u to u' in the metric space \mathcal{U} .

of $\mathcal{S}_{\mathcal{F}}$, and hence is a nonredundant prelinear class of systems. We call \mathcal{S}_0 the natural representation of \mathcal{S} , and ψ its natural mapping. Additional conditions must be imposed to make ψ continuous.

PROPOSITION I.2. *Let \mathcal{S} be a class of systems. The natural mapping ψ for \mathcal{S} is continuous if and only if the functions of $x, f(\cdot, u)$, are equicontinuous for all $u \in \mathcal{U}$. In particular, ψ is continuous if \mathcal{U} is compact.*

Proof. Let $F' = \psi(x')$, $F = \psi(x)$. Then

$$\begin{aligned}\|F - F'\| &= \sup_{\mathcal{U}} \|(F - F')(u)\| = \sup_{\mathcal{U}} \|F(u) - F'(u)\| \\ &= \sup_{\mathcal{U}} \|f(x, u) - f(x', u)\|.\end{aligned}$$

Thus, given $\varepsilon > 0$, there will exist $\delta > 0$ such that $\|F - F'\| \leq \varepsilon$ whenever $d(x, x') \leq \delta$ if and only if the $f(\cdot, u)$ are equicontinuous functions of x .

Suppose \mathcal{U} is compact. Consider neighborhoods in $\mathcal{X} \times \mathcal{U}$ of the form $N_\varepsilon(x_1, u_1) = \{(x, y): d(x, x_1) < \delta_1, d(y, u_1) < \delta_2\}$, where δ_1 and δ_2 are chosen so that if $(x, u) \in N_\varepsilon(x_1, u_1)$, then $\|f(x, u) - f(x_1, u_1)\| \leq \varepsilon/2$. Then $\|f(x, u) - f(x', u)\| \leq \varepsilon$ if (x, u) and (x', u) belong to an N_ε . With x fixed, let there be given an $N_\varepsilon(x, u)$ for each $u \in \mathcal{U}$. The intersection of each of these neighborhoods with the compact subspace $\mathcal{U}_x = \{(x, u): u \in \mathcal{U}\}$ is an open set in \mathcal{U}_x ; these open sets form a covering and a finite subcovering can be extracted. Let this be given by the intersections of \mathcal{U}_x with neighborhoods $N_\varepsilon(x, u_i)$, $i = 1, 2, \dots, K$. If now $d(x, x') < \delta$, where δ is the minimum of the δ_1 's for the K neighborhoods $N_\varepsilon(x, u_i)$, the points (x', u) and (x, u) lie in one of the $N_\varepsilon(x, u_i)$ for each $u \in \mathcal{U}$. Hence,

$$\|F - F'\| = \sup_{\mathcal{U}} \|f(x, u) - f(x', u)\| \leq \varepsilon. \quad \square$$

A corollary of this result follows.

PROPOSITION I.3. *If \mathcal{S} has the properties (i) \mathcal{U} is compact, (ii) \mathcal{X} is compact, and (iii) \mathcal{S} is nonredundant, then \mathcal{S} is equivalent to its natural representation \mathcal{S}_0 .*

Proof. Since \mathcal{S} is nonredundant, ψ is one-to-one and onto $\psi(\mathcal{X})$. Then the fact that \mathcal{X} is compact implies that ψ is a homeomorphism. \square

The following is also to be noted.

PROPOSITION I.4. *If \mathcal{S} is prelinear, then ψ is a restriction of a linear mapping.*

Proof. Let α, β, x_1 and x_2 be such that x_1, x_2 and $\alpha x_1 + \beta x_2$ all belong to \mathcal{X} . Let $\psi(x_1) = F_1, \psi(x_2) = F_2$ and $\psi(\alpha x_1 + \beta x_2) = F$. Then,

$$\begin{aligned}F(\cdot) &= f(\alpha x_1 + \beta x_2, \cdot) = \alpha f(x_1, \cdot) + \beta f(x_2, \cdot) \\ &= \alpha F_1(\cdot) + \beta F_2(\cdot). \quad \square\end{aligned}$$

It will sometimes be convenient in special cases to deal with a functional representation, that is, a modification of the natural representation, as follows. Let $\mathcal{L} = \mathcal{L}(\mathcal{U}, \mathcal{Y})$ be a linear manifold, not necessarily closed, in $\mathcal{F}(\mathcal{U}, \mathcal{Y})$. Suppose \mathcal{L} is itself a Banach space with a norm to be denoted by $\|\cdot\|$, and that $\|H\| \geq c\|H\|$, $H \in \mathcal{L}$, c being a positive constant.

Then, with g defined as before, we have our next result.

PROPOSITION I.5. $\mathcal{S}_{\mathcal{L}} = (\mathcal{Y}, g, \mathcal{L}, \mathcal{U})$ is a nonredundant linear class of systems.

Proof. The proof is a trivial modification of that of Proposition I.1. \square

Let $\mathcal{S} = (\mathcal{Y}, f, \mathcal{X}, \mathcal{U})$ be a class of systems with the property that $f(x, \cdot)$ is an element of \mathcal{L} for each $x \in \mathcal{X}$. Let ψ be defined as before and put $\mathcal{H} = \psi(\mathcal{X})$, $\mathcal{H} \subset \mathcal{L}$. Then $\mathcal{S}_1 = (\mathcal{Y}, g, \mathcal{H}, \mathcal{U})$ is a nonredundant prelinear class of systems and provides what we shall call the functional representation of \mathcal{S} with respect to \mathcal{L} , where \mathcal{H} is understood to have the topology of \mathcal{L} . Note that Proposition I.4 still holds. However, the continuity of ψ remains to be investigated in each case, unless, of course, the norms $\|\cdot\|$ and $\|\cdot\|$ are equivalent.

2. ε -representations and ε -approximations. Let $\mathcal{S}_0 = (\mathcal{Y}, g, \mathcal{H}, \mathcal{U})$ be a subclass of \mathcal{S} , and let $\mathcal{S}_1 = (\mathcal{Y}, f_1, \mathcal{X}_1, \mathcal{U})$ be a class of systems. \mathcal{S}_1 is an ε -approximation to \mathcal{S}_0 , $\varepsilon > 0$, if for each $F \in \mathcal{H}$ there is an $x_1 \in \mathcal{X}_1$ such that $\|F - \psi_1(x_1)\| \leq \varepsilon$, where ψ_1 is the natural mapping for \mathcal{S}_1 . Note that it is not required that $\psi_1(x_1) \in \mathcal{H}$. Let ϕ_1 be a mapping from \mathcal{H} into \mathcal{X}_1 . Then (\mathcal{S}_1, ϕ_1) is an ε -representation of \mathcal{S}_0 if $\|F - \psi_1 \cdot \phi_1(F)\| \leq \varepsilon$ for all $F \in \mathcal{H}$. Again it is not required that $\psi_1 \cdot \phi_1(F) \in \mathcal{H}$. If \mathcal{S} is any class of systems, we say that \mathcal{S}_1 or (\mathcal{S}_1, ϕ_1) is an ε -approximation or, respectively, an ε -representation for \mathcal{S} if it is such for the natural representation \mathcal{S}_0 of \mathcal{S} . Note that the definitions imply that \mathcal{S} and \mathcal{S}_1 have the same input and output spaces \mathcal{U} and \mathcal{Y} .

An ε -approximation \mathcal{S}_1 is said to be *linear* if \mathcal{S}_1 is a prelinear class. This implies by Proposition I.4 that ψ_1 is a restriction of a linear mapping. It is said to be *finite-dimensional* if \mathcal{X}_1 is a subset of a *finite-dimensional* Euclidean space. It is *continuous* if ψ_1 is continuous. An ε -representation (\mathcal{S}_1, ϕ_1) is said to be *linear* if \mathcal{S}_1 is a prelinear class and in addition ϕ_1 is a restriction of a linear mapping. It is *finite-dimensional* if \mathcal{S}_1 is, and is *continuous* if both ϕ_1 and ψ_1 are continuous. An ε -representation is *determined* by \mathcal{U}_0 , $\mathcal{U}_0 \subset \mathcal{U}$, if the mapping ϕ depends only on the functions F , $F \in \mathcal{H}$, restricted to \mathcal{U}_0 .

The following observations are immediate from the definitions. If $\mathcal{S} = (\mathcal{Y}, f, \mathcal{X}, \mathcal{U})$ and $\mathcal{S}_1 = (\mathcal{Y}, f_1, \mathcal{X}_1, \mathcal{U})$ are two classes of systems, if $\psi_1(\mathcal{X}_1) \supset \psi(\mathcal{X})$, and if \mathcal{S}_1 is nonredundant, then $(\mathcal{S}_1, \psi_1^{-1})$ is an ε -representation of \mathcal{S} for any $\varepsilon > 0$. If \mathcal{S}_1 is equivalent to its natural representation, then $(\mathcal{S}_1, \psi_1^{-1})$ is a continuous ε -representation of \mathcal{S} for any $\varepsilon > 0$. If \mathcal{S} and \mathcal{S}_1 are equivalent nonredundant classes, then $(\mathcal{S}_1, \psi_1^{-1})$ is an ε -representation of \mathcal{S} for any ε . In each instance ψ_1 is one-to-one, so ψ_1^{-1} is defined.

PROPOSITION I.6. Let $\mathcal{S}_1 = (\mathcal{Y}, f_1, \mathcal{X}_1, \mathcal{U})$, $\mathcal{S}_2 = (\mathcal{Y}, f_2, \mathcal{X}_2, \mathcal{U})$ and $\mathcal{S}_3 = (\mathcal{Y}, f_3, \mathcal{X}_3, \mathcal{U})$ be classes of systems with natural representations \mathcal{S}_{10} , \mathcal{S}_{20} and \mathcal{S}_{30} , respectively. The parameter space of \mathcal{S}_{i0} is denoted \mathcal{H}_i , and the natural mapping for \mathcal{S}_i is ψ_i , $i = 1, 2, 3$. Then:

- (a) If (\mathcal{S}_2, ϕ_2) is an ε_2 -representation of \mathcal{S}_1 and (\mathcal{S}_3, ϕ_3) is an ε_3 -representation of \mathcal{S}_2 , then $(\mathcal{S}_3, \phi_3 \circ \psi_2 \circ \phi_2)$ is an $(\varepsilon_2 + \varepsilon_3)$ -representation of \mathcal{S}_1 . Further:
- (b) If (\mathcal{S}_2, ϕ_2) and (\mathcal{S}_3, ϕ_3) are continuous or linear, so is $(\mathcal{S}_3, \phi_3 \circ \psi_2 \circ \phi_2)$.
- (c) If (\mathcal{S}_3, ϕ_3) is finite-dimensional, so is $(\mathcal{S}_3, \phi_3 \circ \psi_2 \circ \phi_2)$.
- (d) If (\mathcal{S}_2, ϕ_2) is determined by a subset $\mathcal{U}_0 \subset \mathcal{U}$, so is $(\mathcal{S}_3, \phi_3 \circ \psi_2 \circ \phi_2)$.

Proof. The proof is obvious. \square

We now address the questions of the existence and construction of ε -representations with desirable properties for classes of systems for which there is little prior information. In different words, it is desired to obtain methods of constructing good ε -representations which do not require a knowledge of the structure of the

systems in the class, or indeed, do not require that they all have the same structure—admittedly, the term structure as used here is somewhat vague, but the idea should be clear.

Two different general methods for obtaining ε -representations are investigated. The first is to represent each system in the given class by a fixed interpolation formula which is based on a sufficiently large set of input–output pairs. The second is to obtain first an ε -approximation for the given class (e.g., by use of the Stone–Weierstrass theorem) and then construct a mapping ϕ so as to get an ε -representation. The first method leads to the following basic result.

PROPOSITION I.7. *Let \mathcal{Y} be a Banach space and let \mathcal{X} and \mathcal{U} be compact metric spaces. Let $\mathcal{S} = (\mathcal{Y}, f, \mathcal{X}, \mathcal{U})$ be a class of systems with \mathcal{X} , \mathcal{U} , \mathcal{Y} as prescribed. Then, given $\varepsilon > 0$, there exists an ε -representation (\mathcal{S}_1, ϕ_1) of \mathcal{S} that is continuous, finite-dimensional, and determined by a finite subset $\mathcal{U}_0 \subset \mathcal{U}$. \mathcal{S}_1 is a linear ε -approximation. If \mathcal{Y} is a Hilbert space, then there exists an ε -representation which, in addition to possessing the other properties listed, is linear; i.e., ϕ_1 is also a restriction of a linear mapping.*

The proof follows from construction of an ε -representation with the described properties. The construction yields what will be termed a *standard ε -representation*, described below by equations (5), (7) and (8). Given \mathcal{S} as prescribed and given $\varepsilon > 0$, there are many ε -representations of \mathcal{S} which are described by such equations; all of them are called standard. Even when \mathcal{Y} is not a Hilbert space, there may be a linear standard ε -representation, but the theorem does not guarantee linearity of ϕ_1 in general.

Proof. The proof proceeds in a series of steps. Let $\mathcal{S}_0 = (\mathcal{Y}, g, \mathcal{H}, \mathcal{U})$ be the natural representation of \mathcal{S} .

(a) \mathcal{H} and $\mathcal{H}(\mathcal{U})$ are compact.

Note that since \mathcal{U} is compact, ψ is continuous; then since \mathcal{X} is compact, $\mathcal{H} = \psi(\mathcal{X})$ is a compact subset of \mathcal{F} . Then by Ascoli's theorem, \mathcal{H} is equicontinuous and the set $\mathcal{H}(u)$ is compact in \mathcal{Y} for each $u \in \mathcal{U}$. The functions F in \mathcal{H} are uniformly continuous since \mathcal{U} is compact.

Consider an arbitrary infinite sequence in $\mathcal{H}(\mathcal{U})$; denote it by $\{y_n\}_n$, $y_n = F_n(u_n)$, $F_n \in \mathcal{H}$, $u_n \in \mathcal{U}$. Pick a convergent subsequence $\{u_{n_i}\}_i$ from $\{u_n\}$, and let $u_0 \in \mathcal{U}$ be its limit as $i \rightarrow \infty$. Then pick a convergent subsequence $\{F_{n_{ij}}\}_j$ from $\{F_{n_i}\}$, and let $F_0 \in \mathcal{H}$ be its limit as $j \rightarrow \infty$. Consider the sequence $\{F_{n_{ij}}(u_{n_{ij}})\}_j$ as $j \rightarrow \infty$, and rewrite it simply as $\{F_j(u_j)\}_j$, $F_0(u_0) \in \mathcal{H}(\mathcal{U})$, and

$$\begin{aligned} \|F_j(u_j) - F_0(u_0)\| &\leq \|F_j(u_j) - F_j(u_0)\| + \|F_j(u_0) - F_0(u_0)\| \\ &\leq \|F_j(u_j) - F_j(u_0)\| + \|F_j - F_0\|. \end{aligned}$$

The right side of this inequality approaches zero as $j \rightarrow \infty$ by the equicontinuity of the F_j and since $u_j \rightarrow u_0$ and $F_j \rightarrow F_0$.

(b) Construction of interpolation functionals on a compact set.

This construction is applied below, and is described here out of context. Let \mathcal{K} be a compact subset of a metric space \mathcal{M} , and let $\eta > 0$ be given. Let $\{\xi_i\}$, $i = 1, \dots, N$, be a maximal $3/4\eta$ -separated set in \mathcal{K} .² Such a finite set

² A set of $\{\xi_i\}$ is an ε -separated set if $d(\xi_i, \xi_j) > \varepsilon$ whenever $i \neq j$.

exists because \mathcal{K} is compact. Let

$$B_i = \{\xi \in \mathcal{M} : d(\xi, \xi_i) < \eta\}, \quad i = 1, \dots, N.$$

The open balls B_i provide a covering of \mathcal{K} , for if not the set $\{\xi_i\}$ would not be maximal. Let

$$A_i = \{\xi \in \mathcal{M} : d(\xi, \xi_i) \leq \eta/4\}, \quad i = 1, \dots, N.$$

The closed balls A_i are disjoint. Let

$$C_i = B_i - \bigcup_{j \neq i}^N A_j, \quad i = 1, \dots, N.$$

The C_i are open, $C_i \subset B_i$, $C_i \supset A_i$, and

$$\bigcup_{i=1}^N C_i = \bigcup_{i=1}^N B_i \supset \mathcal{K}.$$

Define

$$\begin{aligned} f_i(\xi) &= \frac{d(\xi, C_i^c)}{d(\xi, A_i) + d(\xi, C_i^c)} \\ (1) \quad &= 1 - \frac{d(\xi, A_i)}{d(\xi, A_i) + d(\xi, C_i^c)}, \quad i = 1, \dots, N. \end{aligned}$$

Since the distance from a point ξ to a closed set in a metric space is a continuous function of ξ , and since $d(\xi, A_i) + d(\xi, C_i^c) \geq \eta/4$, each $f_i(\xi)$ is continuous on \mathcal{M} . Furthermore, $f_i(\xi) = 1$ for $\xi \in A_i$; $f_i(\xi) = 0$ for $\xi \in C_i^c$; $0 \leq f_i(\xi) \leq 1$, and $f_i(\xi) > 0$ for $\xi \in C_i$.

Now define, for $i = 1, \dots, N$,

$$\begin{aligned} (2) \quad \gamma_i(\xi) &= \frac{f_i(\xi)}{\sum_{j=1}^N f_j(\xi)} \quad \left(\text{for } \xi \in \mathcal{K} \subset \bigcup_{i=1}^N C_i \right) \\ &= \inf_{v \in \mathcal{K}} \left[\frac{(\gamma_i(v) + 1) d(\xi, v)}{d(\xi, \mathcal{K})} \right] - 1 \quad (\text{for } \xi \in \mathcal{M} - \mathcal{K}). \end{aligned}$$

The definition is meaningful for $\xi \in \mathcal{K}$ because $\sum_{j=1}^N f_j(\xi) > 0$ for $\xi \in \mathcal{K}$. Furthermore, γ_i in \mathcal{K} is continuous and takes values between 0 and 1. Thus the Tietze–Urysohn extension theorem applies, and in fact one continuous extension to \mathcal{M} taking values between 0 and 1 is given by the second part of the definition (2) (see [4]). We have then: (i) each γ_i is continuous on \mathcal{M} ; (ii) $\gamma_i(\xi) = 1$ for $\xi \in A_i$ (since $f_j(\xi) = 0$ for $\xi \in A_i$, $i \neq j$); (iii) $\gamma_i(\xi) = 0$ for $\xi \in \mathcal{K} - C_i$; and (iv) $\sum_{i=1}^N \gamma_i(\xi) = 1$ for $\xi \in \mathcal{K}$.

The $\gamma_i(\xi)$ are called the interpolation functionals for \mathcal{K} with respect to $\{\xi_1, \dots, \xi_N\}$, with mesh η . Given any compact $\mathcal{K} \subset \mathcal{M}$, and any maximal $3/4\eta$ -separated set in \mathcal{K} , a unique set of interpolation functionals is defined by equations (1) and (2). The positive integer N may depend on the particular maximal $3/4\eta$ -separated set.

(c) Specification of an approximating subspace \mathcal{Y}' for $\mathcal{H}(\mathcal{U})$ in \mathcal{Y} and a mapping π from \mathcal{Y} into \mathcal{Y}' .

Let $\eta = \varepsilon/6$. Since $\mathcal{H}(\mathcal{U})$ is a compact subset of \mathcal{Y} , we can choose $\{y_i\}$, $i = 1, \dots, K$, to be a maximal $3/4\eta$ -separated subset of $\mathcal{H}(\mathcal{U})$ and construct the interpolation functionals for $\mathcal{H}(\mathcal{U})$ with respect to $\{y_1, \dots, y_K\}$. Denote these by $\alpha_i(y)$, $i = 1, \dots, K$.

Let \mathcal{Y}' be the (finite-dimensional) linear subspace of \mathcal{Y} spanned by $\{y_1, \dots, y_K\}$. \mathcal{Y}' has dimension $M \leq K$. \mathcal{Y}' is to be the subspace in which we approximate $\mathcal{H}(\mathcal{U})$ by means of a continuous mapping π . Let π be defined by $\pi(y) = y'$, where $y' = \sum_{i=1}^K \alpha_i(y)y_i$; π is a bounded continuous map from \mathcal{Y} into \mathcal{Y}' . We have for $y \in \mathcal{H}(\mathcal{U})$

$$\begin{aligned} \|y - y'\| &= \left\| y - \sum_{i=1}^K \alpha_i(y)y_i \right\| \\ &= \left\| \sum_{i=1}^K \alpha_i(y)y - \sum_{i=1}^K \alpha_i(y)y_i \right\| \leq \sum_{i=1}^K \|\alpha_i(y)(y - y_i)\| \\ &= \sum_{i=1}^p \alpha_{n_i}(y) \|y - y_{n_i}\| + \sum_{i=1}^q \alpha_{m_i}(y) \|y - y_{m_i}\|, \end{aligned}$$

where y_{n_1}, \dots, y_{n_p} are those y_i for which $y \in \mathcal{H}(\mathcal{U}) \cap C_i$ (defined as in (b)), and y_{m_1}, \dots, y_{m_q} are those y_i for which $y \in \mathcal{H}(\mathcal{U}) - C_i$. By (b, iii) the second sum is zero. Since $\|y - y_i\| \leq \eta$ for all i such that $y \in C_i$,

$$(3) \quad \|y - \pi(y)\| \leq \sum_{i=1}^p \alpha_{n_i}(y) \|y - y_{n_i}\| \leq \eta, \quad y \in \mathcal{H}(\mathcal{U}).$$

Now select a maximal linearly independent subset of $\{y_1, \dots, y_K\}$. By re-numbering if necessary, denote these $\{y_1, \dots, y_M\}$, $M \leq K$. Let

$$y_j = \sum_{i=1}^M \beta_{ji}y_i, \quad j = M+1, \dots, K.$$

Then,

$$(4) \quad \pi(y) = y' = \sum_{i=1}^M \left[\alpha_i(y) + \sum_{k=M+1}^K \alpha_k(y)\beta_{ki} \right] y_i.$$

We shall also write

$$(5) \quad y' = \sum_{i=1}^M e_i^*(y')e_i \quad \text{for all } y' \in \mathcal{Y}',$$

where e_i is an arbitrary basis in \mathcal{Y}' and the e_i^* are continuous linear functionals. Of course, if $e_i = y_i$, then the $e_i^*(\pi(y))$ are given by the bracketed expressions in equation (4).

(d) Specification of \mathcal{X}_1 , ϕ_1 and f_1 .

Since \mathcal{U} is compact the $F \in \mathcal{H}$ are uniformly continuous on \mathcal{U} , and since \mathcal{H} is compact they are equicontinuous. Hence, given $\varepsilon > 0$ there is $\delta = \delta(\varepsilon)$ such that $\|F(u) - F(u')\| \leq \varepsilon/6$ whenever $d(u, u') \leq \delta$ for all $F \in \mathcal{H}$. With such a δ

fixed, choose a maximal $3/4\delta$ -separated set $\{u_1, \dots, u_N\}$ in \mathcal{U} and form the interpolation functionals for \mathcal{U} with respect to $\{u_1, \dots, u_N\}$. Denote these (as in (b)) by $\gamma_i(u)$, $i = 1, \dots, N$.

We take \mathcal{X}_1 to be a subset of an NM -dimensional Euclidean space \mathcal{E} with an arbitrary fixed orthogonal basis, and represent points x in \mathcal{E} by an NM -vector with real components indexed as shown:

$$(6) \quad x = (a_{11}, a_{21}, \dots, a_{M1}; a_{12}, \dots, a_{M2}; \dots; a_{1N}, \dots, a_{MN}).$$

The mapping ϕ_1 from \mathcal{H} into \mathcal{E} is then defined by $\phi_1(F) = x$, $F \in \mathcal{H}$, where the components a_{kj} of x corresponding to F are given by

$$(7) \quad a_{kj}(F) = e_k^* \circ \pi \circ F(u_j), \quad k = 1, \dots, M, \quad j = 1, \dots, N.$$

We put $\mathcal{X}_1 = \phi_1(\mathcal{H}) \subset \mathcal{E}$. The functionals a_{kj} are continuous on \mathcal{H} ; in fact, $F(u_j)$ regarded as a function of F for fixed u_j is continuous since $\|F(u_j) - F'(u_j)\| = \|(F - F')(u_j)\| \leq \|F - F'\|$, and π and e_k^* are continuous. Hence ϕ_1 is continuous, and \mathcal{X}_1 is compact.

The function f is defined by

$$(8) \quad f(x, u) = \sum_{k=1}^M \left[\sum_{j=1}^N \gamma_j(u) a_{kj} \right] e_k, \quad u \in \mathcal{U},$$

where the a_{kj} are the components of x as indicated in equation (7). Obviously f is actually defined on all of \mathcal{E} by equation (8), but unless there is a specific statement to the contrary, we consider f always as a mapping only from $\mathcal{X}_1 \times \mathcal{U}$ into $\mathcal{Y}' \subset \mathcal{Y}$. Since the γ_i are continuous functionals on \mathcal{U} and the a_{kj} are continuous functionals on \mathcal{X}_1 , it is clear that f is continuous and bounded on $\mathcal{X}_1 \times \mathcal{U}$. It follows that $\mathcal{S}_1 = (\mathcal{Y}, f_1, \mathcal{X}_1, \mathcal{U})$ is a class of systems.

(e) (\mathcal{S}_1, ϕ_1) is an ε -representation of \mathcal{S}_0 .

It is necessary to prove that if $F \in \mathcal{H}$, then $\|F - \psi_1 \circ \phi_1(F)\| \leq \varepsilon$. Or, since

$$\begin{aligned} \|F - \psi_1 \circ \phi_1(F)\| &= \sup_{\mathcal{U}} \|F(u) - (\psi_1 \circ \phi_1 \circ F)(u)\| \\ &= \sup_{\mathcal{U}} \|F(u) - f(\phi_1(F), u)\|, \end{aligned}$$

it is necessary to prove that $\|F(u) - f(\phi_1(F), u)\| \leq \varepsilon$ for all $u \in \mathcal{U}$. Note that

$$\begin{aligned} f(\phi_1(F), u) &= \sum_{j=1}^N \gamma_j(u) \sum_{k=1}^M a_{kj}(F) e_k \\ &= \sum_{j=1}^N \gamma_j(u) \sum_{k=1}^M [e_k^* \circ \pi \circ F(u_j)] e_k \\ &= \sum_{j=1}^N \gamma_j(u) [\pi \circ F(u_j)]. \end{aligned}$$

Temporarily fix u and let u_n be any one of $\{u_1, \dots, u_N\}$ such that $d(u, u_n) \leq \delta$. Then,

$$(9) \quad \begin{aligned} \|F(u) - f(\phi_1(F), u)\| &\leq \|F(u) - F(u_n)\| \\ &\quad + \|F(u_n) - f(\phi_1(F), u_n)\| + \|f(\phi_1(F), u_n) - f(\phi_1(F), u)\|. \end{aligned}$$

Since $d(u, u_n) \leq \delta$, the first term on the right side of the inequality (9) is less than or equal to $\varepsilon/6$, by the choice of δ . Since $\gamma_i(u_n) = \delta_{in}$, $f(\phi_k(F), u_n) = \pi \circ F(u_n)$, so we have that the second term is

$$\|F(u_n) - \pi \circ F(u_n)\| \leq \varepsilon/6$$

by (3). Also the final term on the right side of (9) can be written

$$\left\| \pi \circ F(u_n) - \sum_{j=1}^N \gamma_j(u) [\pi \circ F(u_j)] \right\|.$$

Let $\{u_{j_1}, \dots, u_{j_P}\}$, $P \leq N$, be the set of u_i 's which are within distance δ of u . Since $\gamma_i(u) = 0$ if i is not one of j_1, \dots, j_P , and since $d(u_{j_n}, u_{j_m}) \leq 2\delta$ for $n, m = 1, \dots, P$, this final term becomes

$$\begin{aligned} & \left\| \pi \circ F(u_n) - \sum_{p=1}^P \gamma_{j_p}(u) [\pi \circ F(u_{j_p})] \right\| \\ &= \left\| \sum_{p=1}^P \gamma_{j_p}(u) [\pi \circ F(u_n) - \pi \circ F(u_{j_p})] \right\| \\ &\leq \sum_{p=1}^P \gamma_{j_p}(u) \|\pi \circ F(u_n) - \pi \circ F(u_{j_p})\|. \end{aligned}$$

But

$$\|F(u_n) - F(u_{j_p})\| \leq \|F(u_n) - F(u)\| + \|F(u) - F(u_{j_p})\| \leq 2\varepsilon/6$$

for each $p = 1, \dots, P$. Consequently

$$\|\pi \circ F(u_n) - \pi \circ F(u_{j_p})\| \leq 4\varepsilon/6$$

for each p , and the last term in (9) is less than or equal to $\sum_{p=1}^P \gamma_{j_p}(u) \cdot 4\varepsilon/6 = 4\varepsilon/6$. Hence we have that $\|F(u) - f(\phi_1(F), u)\| \leq \varepsilon$ for all $u \in \mathcal{U}$, which proves the assertion.

(f) Properties of the ε -representation (\mathcal{S}_1, ϕ_1) .

It has been shown in (d) that ϕ_1 is continuous. Since \mathcal{U} is compact, ψ_1 is continuous, and the ε -representation (\mathcal{S}_1, ϕ_1) is continuous. From the formula (8) for $f(x, u)$ it is clear that f is linear in x , hence \mathcal{S}_1 is prelinear. Obviously (\mathcal{S}_1, ϕ_1) is finite-dimensional, and equally obvious ϕ_1 is determined by the functions $F \in \mathcal{H}$ acting only on $\{u_1, \dots, u_N\}$.

The construction given provides a ϕ_1 which is not a restriction of a linear map, because π is not linear. However, in any special case where a continuous linear mapping π' can be found that yields a uniformly good approximation to $\mathcal{H}(\mathcal{U})$ in a finite-dimensional subspace of \mathcal{Y} , then π' can replace π , ϕ_1 becomes a restriction of a linear mapping and the ε -representation is linear. This can always be done when \mathcal{Y} is a Hilbert space, by taking π' to be the orthogonal projection on a finite-dimensional subspace that approximates the compact set $\mathcal{H}(\mathcal{U})$ uniformly to within the desired tolerance. \square

The ε -representation constructed in the proof of the theorem is not unique since it depends on the choice of the u_i 's and the y_i 's. These are not unique and

even the integers N and M , which fix the "size" of the ε -representation, are not unique. However, since the number of elements in an η -separated subset of a compact metric space is bounded, there are positive integers N_ε and M_ε such that $N \leq N_\varepsilon$ and $M \leq M_\varepsilon$. The question of how big M and N are is obviously of interest, but it is not considered here. As far as achieving the desired degree of approximation is concerned, there is no reason why the u_i 's and y_i 's must have the separation property; it is only necessary that they provide a sufficiently fine mesh in \mathcal{U} and $\mathcal{H}(\mathcal{U})$, respectively. However, requiring the separation property guarantees that N and M are not unnecessarily large. If they do not satisfy the separation property, the construction of the interpolation functionals needs to be modified slightly (as is done in the proof of Proposition I.8).

Also, the ε -representation can be modified in other ways without changing its essential characteristics. Any continuous mapping π' that satisfies $\|y - \pi'(y)\| \leq \eta$, $y \in \mathcal{H}(\mathcal{U})$, can be used instead of π , and it has already been pointed out that in some instances at least it is advantageous to replace π . Further, any interpolation functionals that satisfy the conditions of step (b) can be used exactly as are the γ_i 's. Any changes of the types indicated can be made without invalidating equations (5), (7) and (8), which finally describe the ε -representation in terms of π , the γ_i 's, the u_i 's and the y_i 's.

The other general method mentioned for getting continuous ε -representations is to break the problem into two parts: first find a convenient class of functions from \mathcal{U} to \mathcal{Y} to approximate uniformly the functions in the particular set \mathcal{H} in question, and second, construct a continuous ϕ carrying \mathcal{H} into the approximating class. The first step yields an ε_1 -approximation, the second makes it into an ε_2 -representation, where most likely ε_2 must exceed ε_1 . The following result guarantees that this second step can sometimes be accomplished.

PROPOSITION I.8. *Let $\mathcal{S}_0 = (\mathcal{Y}, g, \mathcal{H}, \mathcal{U})$, and let $\mathcal{S}_1 = (\mathcal{Y}, f_1, \mathcal{X}_1, \mathcal{U})$ be a linear ε -approximation to \mathcal{S}_0 . In addition, require that: (i) \mathcal{U} be compact; (ii) \mathcal{X}_1 be compact and convex; and (iii) \mathcal{S}_1 be nonredundant. Then, given $\eta > \varepsilon$, there exists a mapping $\phi_1: \mathcal{H} \rightarrow \mathcal{X}_1$, such that (\mathcal{S}_1, ϕ_1) is a continuous η -representation of \mathcal{S}_0 determined by a finite set $\{u_1, \dots, u_N\}$.*

Proof. Let $\alpha = (\eta - \varepsilon)/4$. By Proposition I.3, the natural mapping ψ_1 is uniformly continuous on \mathcal{X}_1 . Hence there exists $\delta = \delta(\alpha) > 0$ such that $\|\psi(x) - \psi(x')\| \leq \alpha$ for any pair x, x' belonging to \mathcal{X}_1 that satisfies $d(x, x') \leq \delta$. Let $\{x_i\}$, $i = 1, \dots, K$, $x_i \in \mathcal{X}_1$, be chosen so that the open balls of radius δ centered at the x_i cover \mathcal{X}_1 . Put $F_i = \psi_1(x_i)$, $i = 1, \dots, K$. Then every $F \in \psi_1(\mathcal{X}_1)$ satisfies $\|F - F_i\| \leq \alpha$ for some i .

By hypothesis, ψ_1 is one-to-one from \mathcal{X}_1 onto $\psi_1(\mathcal{X}_1)$, so a mapping ϕ can be defined on the F_i by $\tilde{\phi}(F_i) = x_i$, $i = 1, \dots, K$; $\tilde{\phi}$ can then be extended to \mathcal{H} by an interpolation that is essentially the same as that described in the proof of Proposition I.7. In fact, since \mathcal{S}_1 is an ε -approximation to \mathcal{S}_0 , it follows that for any $F \in \mathcal{H}$ there is an F_i such that $\|F - F_i\| \leq \varepsilon + \alpha$. Let the B_i , $i = 1, \dots, K$, be open balls of radius $\varepsilon + 3\alpha$ centered at the F_i . Then $\bigcup_{i=1}^K B_i \supset \mathcal{H}$. At each F_i , let d_i be the minimum of the distances to the other F_j . Let A_i , $i = 1, \dots, K$, be closed balls of radius $d_i/3$, respectively, centered at the F_i . Put $C_i = B_i - \bigcup_{j \neq i}^K A_j$. Using the sets C_i and A_i , define interpolation functionals γ_i on $\bigcup_{i=1}^K C_i = \bigcup_{i=1}^K B_i$ exactly as in step (b) of the previous proof; the properties of the γ_i stated there all hold.

Then, define $\tilde{\phi}$ on \mathcal{H} by

$$(10) \quad \tilde{\phi}(F) = \sum_{i=1}^K \gamma_i(F) \tilde{\phi}(F_i) = \sum_{i=1}^K \gamma_i(F) x_i.$$

The mapping given by (10) is continuous, and it carries \mathcal{H} into the convex hull of the x_i and hence into \mathcal{X}_1 . Finally, by the linearity of ψ_1 and by arguments employed previously,

$$(11) \quad \begin{aligned} \|F - \psi_1 \circ \tilde{\phi}(F)\| &= \left\| F - \psi_1 \left[\sum_{i=1}^K \gamma_i(F) x_i \right] \right\| \\ &= \left\| F - \sum_{i=1}^K \gamma_i(F) F_i \right\| \\ &\leq \sum_{i=1}^K \gamma_i(F) \cdot (\varepsilon + 3\alpha) = \varepsilon + 3\alpha. \end{aligned}$$

Hence $(\mathcal{S}_1, \tilde{\phi})$ is a continuous $(\varepsilon + 3\alpha)$ -representation of \mathcal{S}_0 . However, it does not have the property that it is determined by a finite subset of \mathcal{U} . To get this further property, we proceed as follows.

Let $(\mathcal{S}', \tilde{\phi}')$ $\mathcal{S}' = (\mathcal{Y}, f', \mathcal{S}', \mathcal{U})$ be a standard α -representation of \mathcal{S}_0 ; such is known to exist by Proposition I.7. Let \mathcal{S}'_0 be the natural representation of \mathcal{S}' , with natural mapping ψ' , and put $\mathcal{H}' = \psi'(\mathcal{X}')$. Each $F' \in \mathcal{H}'$ is within a distance α of some $F \in \mathcal{H}$, and hence is within a distance $\varepsilon + 2\alpha$ of an F_i . Thus the set $\bigcup_{i=1}^K B_i$ on which the interpolation functionals are defined contains \mathcal{H}' , and the inequality (11) holds for all $F' \in \mathcal{H}'$. We now put $\phi_1 = \tilde{\phi} \circ \psi' \circ \phi'$. Since ϕ', ψ' are continuous, ϕ_1 is continuous; since ϕ' is determined by a finite subset of \mathcal{U} , say $\{u_1, \dots, u_N\}$, so is ϕ_1 . Finally,

$$\|F - \psi_1 \circ \phi_1(F)\| \leq \varepsilon + 4\alpha = \eta \quad \text{for all } F \in \mathcal{H}. \quad \square$$

The condition that \mathcal{X}_1 be convex is really of no consequence, by the following observation. With the terminology of Proposition I.8, suppose \mathcal{X}_1 is a compact nonconvex subset of a normed linear space. Then the convex closure of the finite set $\{x_1, \dots, x_k\}$ is both convex and compact in that normed linear space, and can be used to replace \mathcal{X}_1 with no inconvenience.

It will be observed, however, that the mapping ϕ provided by the theorem is awkward. It appears that this awkwardness is inherent in the problem of constructing a mapping ϕ to convert an ε -approximation to an ε -representation when the original class of systems does not have a structural characterization. In an actual identification one may be forced to pretend that there is a structural characterization known for the class of unknown systems, even when clearly there is not, so as to be able to construct a usable “ ϕ mapping”. An unknown error is then introduced into the identification, which it may be hoped is washed out by statistical estimation procedures, but which is really a modeling error, and not an error due to randomness of observations. This point is discussed very briefly in [1] and somewhat more fully in [3].

3. ε -approximations with Volterra polynomials. By a Volterra polynomial is meant a function of the form given by the right side of the equation

$$(12) \quad y(t) = \sum_{n=0}^N \int_A \cdots \int_A x_n(t, s_1, \dots, s_n) u(s_1) \cdots u(s_n) ds_1 \cdots ds_n, \quad t \in T,$$

where A and T are subsets of \mathbb{R}^1 , usually intervals, and u and y are real-valued or vector-valued functions of a real variable (further details are given below). The term is intended to include the special case where the kernels x_n are not functions of t , so that the polynomial gives a mapping into \mathbb{R}^1 .

Volterra polynomials obviously provide very general classes of mappings, and they are relatively easy to deal with in principle. For these reasons they have been fairly widely used in theoretical treatments for some time to give models for nonlinear system synthesis and identification (see, e.g., [5], [6], [7] and [8]). The chief drawback to their more extensive use has been that their application has not usually been very practical, often leading to an exorbitant amount of actual numerical calculation. It is the author's guess that to a certain extent this is unavoidable, that any model that provides generality comparable to that of the Volterra polynomials will lead to great numerical complexity. The objective, of course, in any given problem is to find a model that is not only convenient to handle but that also has just the right amount of generality; the Volterra polynomials are often too general for a particular application. Nevertheless, they are theoretically important in identification because, in the language of this paper, they give *linear* finite-dimensional ε -approximations for almost any class of systems for which there are ε -approximations.

The treatment here is limited pretty much to statements and proofs of the formations of ε -approximations by Volterra polynomials in two common situations. This material is included for convenience and completeness; it is essentially known, is mathematically straightforward, and, in fact, a proof of Proposition I.10 does appear in the conference paper on which [9] is based (which may not, however, be readily available). There is no discussion of the details of using Volterra polynomials in identification. To some extent these details are discussed in the spirit and language of this paper in [1], and closely related material is given in [9].

Let

$$(13) \quad y^{(i)}(t) = \sum_{n=0}^N \int_A \cdots \int_A \sum_{i_1, \dots, i_n=1}^{\mu} x_n^{(i, i_1, \dots, i_n)}(t, s_1, \dots, s_n) \cdot u_{i_1}(s_1) \cdots u_{i_n}(s_n) ds_1 \cdots ds_n, \quad i = 1, \dots, v, \quad t \in T,$$

where μ , v and N are positive integers, A and T are measurable subsets of \mathbb{R}^1 , and all the u_j , $x_n^{(i, i_1, \dots, i_n)}$ and y_i are real-valued functions. The $n = 0$ terms are by convention $x_0^{(i)}(t)$. We let $y(t)$ be the v -vector with components $y^{(1)}(t), \dots, y^{(v)}(t)$, and $u(t)$ the μ -vector with components $u_1(s), \dots, u_{\mu}(s)$, and $x_n(t, s_1, \dots, s_n)$ the element of $\mathbb{R}^{v \times \mu^n}$ with components $x_n^{(i, i_1, \dots, i_n)}(t, s_1, \dots, s_n)$. Then the set of equations (13) can be rewritten in the form

$$(14a) \quad y(t) = [H(u)](t) = \sum_{n=0}^N y_n(t),$$

$$\begin{aligned}
 y_n(t) &= [H_n(u)](t) \\
 (14b) \quad &= \int_A \cdots \int_A x_n(t, s_1, \dots, s_n) u(s_1) \cdots u(s_n) ds_1 \cdots ds_n, \quad t \in T,
 \end{aligned}$$

where the multiplication convention, and the definitions of y_n and of the functions H_n and H are obvious from comparison of (13) and (14).

The Euclidean norm of $u(s)$ in \mathbb{R}^μ is written $|u(s)|$, and, in general, Euclidean norms of elements of finite-dimensional Euclidean spaces are denoted similarly. Since μ and ν are fixed in context, we can denote the L_2 -space of Lebesgue square-integrable μ -vector-valued functions u on A with real components simply by $L_2(A)$, and, similarly, the L_2 -space of square-integrable ν -vector-valued functions y on T with real components by $L_2(T)$. The L_2 -space of functions x_n with the t variable ranging over T and the s variables over A is correspondingly denoted by $L_2(T \times A^n)$. Norms in all these spaces are written $\|\cdot\|$. Thus, for example,

$$\begin{aligned}
 \|x_n\|^2 &= \int_T \int_A \cdots \int_A \sum_{i=1}^\nu \cdots \sum_{i_1 \cdots i_n=1}^\mu \\
 &\quad |x_n^{(i, i_1, \dots, i_n)}(t, s_1, \dots, s_n)|^2 ds_1 \cdots ds_n dt \\
 &= \int_T \int_A \cdots \int_A |x_n(t, s_1, \dots, s_n)|^2 ds_1 \cdots ds_n dt.
 \end{aligned}$$

Also, we let $x_n(t)$ be the function of $t \in T$, taking on values in $L_2(A^n)$, given by the function $x_n(t, s_1, \dots, s_n)$ whenever the following integral exists:

$$\|x_n(t)\|^2 \stackrel{\text{def}}{=} \int_A \cdots \int_A |x_n(t, s_1, \dots, s_n)|^2 ds_1 \cdots ds_n, \quad t \in T.$$

Finally, the norm of a bounded ν -vector-valued function y on T given by $\sup_T |y(t)|$ is written $\|y\|_b$.

We now list certain basic inequalities applying to (13) or (14).

(a) Let $u \in L_2(A)$ and $x_n \in L_2(T \times A^n)$. Then

$$(15) \quad \|y_n\| \leq \|x_n\| \cdot \|u\|^n$$

and

$$(16) \quad \|y\| \leq \sum_{n=0}^N \|x_n\| \cdot \|u\|^n.$$

Also,

$$(17) \quad |y_n(t)|^2 \leq \|u\|^{2n} \cdot \int_A \cdots \int_A |x_n(t, s_1, \dots, s_n)|^2 ds_1 \cdots ds_n$$

and

$$(18) \quad \|y\|_b \leq \sup_T \sum_{n=0}^N \|x_n(t)\| \cdot \|u\|^n$$

whenever the right-hand sides exist.

(b) Let $y = H(u)$ and $\eta = H(\xi)$. Then if $u, \xi \in L_2(A)$ and $x_n \in L_2(T \times A^n)$,

$$(19) \quad \|y_n - \eta_n\|^2 \leq n \|x_n\|^2 \cdot \|u - \xi\|^2 \\ \cdot \{ \|u\|^{2(n-1)} + \|u\|^{2(n-2)} \|\xi\|^2 + \dots + \|\xi\|^{2(n-1)} \}$$

and

$$(20) \quad \|y - \eta\| \leq \|u - \xi\| \left[\sum_{n=1}^N n \|x_n\| [\max(\|u\|, \|\xi\|)]^{n-1} \right].$$

Also,

$$(21) \quad |y_n(t) - \eta_n(t)|^2 \leq n \|x_n(t)\|^2 \cdot \|u - \xi\|^2 \\ \cdot \{ \|u\|^{2(n-1)} + \|u\|^{2(n-2)} \|\xi\|^2 + \dots + \|\xi\|^{2(n-1)} \}$$

and

$$(22) \quad \|y - \eta\|_b \leq \|u - \xi\| \sup_T \left[\sum_{n=1}^N n \|x_n(t)\| [\max(\|u\|, \|\xi\|)]^{n-1} \right]$$

whenever the right-hand sides exist.

(c)

$$(23) \quad |y_n(t) - y_n(t')| \leq \|u\|^n \|x_n(t) - x_n(t')\|.$$

The inequalities of (a) and (c) result immediately from applications of the Schwarz and triangle inequalities. The inequalities of (b) also result from the Schwarz inequality, the inequality

$$\left(\sum_{i=1}^n a_i \right)^2 \leq n \sum_{i=1}^n a_i^2, \quad n = 1, 2, \dots,$$

where the a_i are real numbers, and the identity

$$u_1 u_2 \cdots u_n - \xi_1 \xi_2 \cdots \xi_n = u_1 u_2 \cdots u_{n-1} (u_n - \xi_n) + u_1 \cdots u_{n-2} \\ \cdot \xi_n (u_{n-1} - \xi_{n-1}) + \cdots + \xi_2 \xi_3 \cdots \xi_n (u_1 - \xi_1).$$

We let x represent the $(n+1)$ -tuple of functions (x_0, x_1, \dots, x_N) . Then the equations (14) define a function f_N , $y = f_N(x, u)$, $u \in \mathcal{U}$, $x \in \mathcal{X}$, and $y \in \mathcal{Y}$, where \mathcal{U} , \mathcal{X} and \mathcal{Y} must be specified appropriately. Two cases are considered.

First, let \mathcal{U} be a bounded subset of $L_2(A)$. Consider those $x = (x_0, x_1, \dots, x_N)$ which can be regarded as elements in the Hilbert space $\mathcal{M}_N = L_2(T) \oplus L_2(T \times A) \oplus \cdots \oplus L_2(T \times A^N)$, so that $\|x\|^2 = \sum_{n=0}^N \|x_n\|^2$, and take \mathcal{X} to be a bounded subset of \mathcal{M}_N . Let $\mathcal{Y} = L_2(T)$.

PROPOSITION I.9. *With \mathcal{X} , \mathcal{Y} and \mathcal{U} as just defined, $\mathcal{S} = \{\mathcal{Y}, f_N, \mathcal{X}, \mathcal{U}\}$ is a prelinear class of bounded systems.*

Proof. By the inequality (16), f is a bounded function from $\mathcal{X} \times \mathcal{U}$ into \mathcal{Y} . Continuity in $\mathcal{X} \times \mathcal{U}$ follows easily from

$$\|f(x, u) - f(x', u')\| \leq \|f(x, u) - f(x', u)\| + \|f(x', u) - f(x', u')\|,$$

and the inequalities (20) and (15). The linearity condition is obvious. \square

Second, let \mathcal{U} again be a bounded subset of $L_2(A)$. Let $x_n(t)$ be as defined above and take $x(t) = (x_0(t), x_1(t), \dots, x_N(t))$. Consider those x such that for each t , $x(t)$ can be regarded as an element in the Hilbert space $\mathcal{N}_N = \mathbb{R}^1 \oplus L_2(A) \oplus \dots \oplus L_2(A^N)$ and such that x is a bounded continuous \mathcal{N} -valued function on T . Let the Banach space of bounded continuous \mathcal{N} -valued functions on T with the usual sup norm be denoted \mathcal{C} , and take \mathcal{X}_1 to be a bounded subset of \mathcal{C} . Let \mathcal{Y}_1 be the Banach space $C(T)$ of bounded continuous functions from T to \mathbb{R}^L with norm $\|y\|_b$.

PROPOSITION I.10. *With \mathcal{X}_1 , \mathcal{Y}_1 and \mathcal{U} as just defined, $\mathcal{S}_1 = \{\mathcal{Y}_1, f_N, \mathcal{X}_1, \mathcal{U}\}$ is a prelinear class of systems.*

Proof. The proof is analogous to that of Proposition I.9. \square

We now have basic approximation theorems using these classes of Volterra polynomials.

PROPOSITION I.11. *Let \mathcal{U} be a compact subset of $L_2(A)$, and \mathcal{Y} be $L_2(T)$. Let \mathcal{K} be a compact subset of $\mathcal{F}(\mathcal{U}, \mathcal{Y})$. Then, given $\varepsilon > 0$, the class of systems (in the natural representation form) $\mathcal{S} = (\mathcal{Y}, g, \mathcal{K}, \mathcal{U})$ has a linear ε -approximation $\mathcal{S}_N = (\mathcal{Y}, f_N, \mathcal{X}_N, \mathcal{U})$ for some N , where f_N is as defined above, and \mathcal{X}_N is a finite subset of \mathcal{M}_N . In other words, \mathcal{S} has an ε -approximation in terms of Volterra polynomials of fixed degree with L_2 kernels.*

Proof. Since $\mathcal{K}(\mathcal{U})$ is a compact subset of $\mathcal{Y} = L_2(T)$, given $\alpha > 0$, there is a finite-dimensional subspace of \mathcal{Y} , say of dimension K , with orthonormal projection π such that $\|\pi y - y\| \leq \varepsilon/4$ for all $y \in \mathcal{K}_1(\mathcal{U})$. Let $\{\psi_1, \dots, \psi_K\}$ be an orthonormal basis in $L_2(T)$ for the range of π .

Let \mathcal{K} be covered by finitely many (say B) balls of diameter $\varepsilon/2$, and let F_1, \dots, F_B be arbitrary elements of \mathcal{K} such that F_p is contained in the p th ball. We seek mappings $H_1, \dots, H_B \in \mathcal{K}_N$, the image of \mathcal{K}_N under the natural mapping ψ , such that $\|F_p - H_p\| \leq \varepsilon/2$. With such H_p , $\|F - H_p\| \leq \|F - F_p\| + \|F_p - H_p\| \leq \varepsilon$ for some p , and the proof would be completed.

To find the H_p , first consider the class of all real-valued functionals of the form

$$\begin{aligned} k(x) &= \sum_{n=0}^N \int_A \dots \int_A \sum_{i_1, \dots, i_n} k^{(i_1, \dots, i_n)}(s_1, \dots, s_n) \\ &\quad \cdot u_{i_1}(s_1) \dots u_{i_n}(s_n) ds_1 \dots ds_n \\ (24) \quad &= \sum_{n=0}^N \int_A \dots \int_A k(s_1, \dots, s_n) u(s_1) \dots u(s_n) ds_1 \dots ds_n \end{aligned}$$

(using the multiplication convention), where

$$\int_A \dots \int_A |k(s_1, \dots, s_n)|^2 ds_1 \dots ds_n < \infty,$$

and where by convention the $n = 0$ term is a constant.

This class of functionals is an algebra and it separates points in \mathcal{U} , since if $u_1 \neq u_2$, there is a $k(s)$ such that

$$\int_A k(s) u_1(s) ds \neq \int_A k(s) u_2(s) ds.$$

Hence by the Stone–Weierstrass theorem, any continuous real-valued functional on \mathcal{U} can be approximated to within η in the sup norm by an element of the class, η any positive number.

Now each $\langle \psi_i, F_p(u) \rangle$, $i = 1, \dots, k$; $p = 1, \dots, B$, is a continuous real-valued functional on \mathcal{U} . Hence there exists a positive integer N_{ip} and a $k^{(i,p)}$,

$$k^{(i,p)}(u) = \sum_{n=0}^{N_{ip}} \int_A \cdots \int_A k_n^{(i,p)}(s_1, \dots, s_n) \cdot u(s_1) \cdots u(s_n) ds_1 \cdots ds_n,$$

such that

$$|\langle \psi_i, F_p(u) \rangle - k^{(i,p)}(u)| \leq \varepsilon/4K \quad \text{for all } u \in \mathcal{U}.$$

Put $N = \max_{i,p} (N_{ip})$ and put $k_n^{(i,p)} = 0$ whenever $N_{ip} < n \leq N$. Define H_p by

$$(25) \quad H_p(u) = \sum_{n=0}^N \int_A \cdots \int_A \left[\sum_{k=1}^K k_n^{(k,p)}(s_1, \dots, s_n) \psi_k(t) \right] \cdot u(s_1) \cdots u(s_n) ds_1 \cdots ds_n, \quad t \in T.$$

Then,

$$\langle \psi_i, H_p(u) \rangle = k^{(i,p)}(u),$$

$$|\langle \psi_i, F_p(u) \rangle - \langle \psi_i, H_p(u) \rangle| \leq \varepsilon/4K$$

so that

$$\|\pi F_p(u) - \pi H_p(u)\| \leq K \cdot \varepsilon/4K = \varepsilon/4 \quad \text{for all } u \in \mathcal{U}.$$

Since

$$\|\pi^\perp F_p(u)\| \leq \varepsilon/4 \quad \text{and} \quad \|\pi^\perp H_p(u)\| = 0,$$

we have

$$\|F_p(u) - H_p(u)\| \leq \varepsilon/2 \quad \text{for all } u \in \mathcal{U}. \quad \square$$

PROPOSITION I.12. *Let \mathcal{U} be a compact subset of $L_2(A)$, and \mathcal{Y}' be $C(T)$, where T is compact in \mathbb{R}_1 . Let \mathcal{K}' be a compact subset of $\mathcal{F}(\mathcal{U}, \mathcal{Y}')$. Then, given $\varepsilon > 0$, the class of systems $\mathcal{S} = (\mathcal{Y}', g, \mathcal{K}', \mathcal{U})$ has a linear ε -approximation $\mathcal{S}'_N = (\mathcal{Y}', f_N, \mathcal{X}'_N, \mathcal{U})$ for some N , where f_N is as defined above, and \mathcal{X}'_N is a finite subset of \mathcal{C} .*

Proof. The proof is very close to that of the preceding proposition. The difference is that the inner products with the orthogonal functions $\psi_k(t)$ appearing there are replaced by the coordinate functionals on \mathcal{Y}' provided by $y(t_i)$, $t_i \in T$, $i = 1, \dots, K$ (say), where the mesh of $\{t_i\}$ is chosen sufficiently fine to give the desired approximation to the equi- and uniformly continuous functions of $\mathcal{K}'(\mathcal{U})$. The kernels h_n for the Volterra polynomials are constructed by interpolating linearly between the kernels k_n of the functionals. \square

The ε -approximations given by Propositions I.11 and I.12 need never be redundant, because \mathcal{X}_N and \mathcal{X}'_N are finite, and any redundant x 's can simply be dropped. Hence, by Proposition I.8 and the comment following it, continuous ε -representations can always be formed from such ε -approximations.

Remarks. It should be fairly clear, roughly at least, how ε -representations are to be used as system models for identification. The reasons for the emphasis on the special properties of ε -representations of continuity, linearity, finite-dimensionality, and determination by a finite input set are as follows. Continuity implies that the model is stable. Linearity implies that the identification is "linear-in-the parameters" in the terminology of identification; linear methods are applicable. Finite-dimensionality implies that only finitely many parameters are to be determined in the identification. Determination by a finite input set implies that the identification can be made with only finitely many measurements. All of these properties except linearity are essential, even in noise-free situations, and linearity is very useful, especially when statistical aspects of identification are considered. The use of ε -representations in identification is discussed in [3] for the situation that there is additive output noise.

Algebras of functionals other than Volterra polynomials can be used to construct ε -approximations for rather general classes of systems; see [1]. However, only the Volterra polynomials seem to have been considered, and it appears there is good reason for this. At least it is difficult to envision other algebras which have all the desirable properties of the Volterra polynomials.

REFERENCES

- [1] W. L. ROOT, *Some general structure theory of systems to be used in identification and measurement*, Proc. Fifth Annual Princeton Conf. on Information Sciences and Systems, 1971, pp. 13–19.
- [2] ———, *Approximate representations of causal systems with bounded memory*, Techniques of Optimization, A. V. Balakrishnan, ed., Academic Press, New York, 1972, pp. 51–64. (Proc. Fourth IFIP Colloquium on Optimization Techniques, Los Angeles, 1971.)
- [3] ———, *On the modelling and estimation of communication channels*, Multivariate Analysis, III, P. R. Krishnaiah, ed., Academic Press, New York, 1973, pp. 61–78. (Proc. Third International Symposium on Multivariate Analysis, Wright State Univ., 1972.)
- [4] J. DIEUDONNÉ, *Foundations of Modern Analysis*, Academic Press, New York, 1960.
- [5] N. WIENER, *Nonlinear Problems in Random Theory*, Technology Press of MIT and John Wiley, Cambridge and New York, 1958.
- [6] A. V. BALAKRISHNAN, *Determination of nonlinear systems from input–output data*, Proc. Princeton Univ. Conference on Identification Problems in Communications and Control Systems, 1963, pp. 31–49.
- [7] A. V. BALAKRISHNAN AND P. PETERKA, *Identification in automatic control systems*, Automatica, 5 (1969), no. 6, pp. 817–829.
- [8] W. L. ROOT, *On system measurement and identification theory*, Proc. Symposium on System Theory, Brooklyn Polytechnic Institute, 1965, pp. 133–157.
- [9] ———, *On the structure of a class of system identification problems*, Automatica, 7 (1970), no. 2, pp. 219–231.

ON THE MODELING OF SYSTEMS FOR IDENTIFICATION.

PART II: TIME-VARYING SYSTEMS*

WILLIAM L. ROOT†

Abstract. Certain Banach spaces, denoted \mathcal{L}_T^p , of equivalence classes of functions of a real variable are introduced and investigated. These are to be used as system input and output spaces for systems operating for all time. Some basic properties of causal systems with finite memory are established. The concept of forming time-interval truncations of a time-varying system is formalized and investigated. Trajectories of such truncations are studied. It is proved that under certain reasonable conditions, the trajectories of a class of systems are generated by a strongly continuous semigroup of linear operators. It is also shown that ε -representations of the truncations can be generated by an induced semigroup. Thus, the evolution in time of classes of, in general, nonlinear, time-varying systems, is described in the framework of a linear dynamical theory.

Introduction. A *system* as defined in Part I [2] is simply an input space, an output space, and a mapping carrying inputs into outputs. In Part I some abstract structure and representation theory is established for classes of systems for which the system mappings are bounded and continuous, and for which certain conditions are satisfied by the input and output spaces and by the class of mappings itself. In Part II the interest is in systems and classes of systems where the inputs and outputs are functions of time. Again there is no restriction to linear or to time-invariant systems. The chief emphasis is on causal systems, and indeed on causal systems with bounded memory.

The emphasis on causal systems needs no justification, but there might be a question raised as to why one should consider systems with bounded memory. The primary answer is: the bounded memory condition turns out to fit very conveniently in the mathematical structure used, and since almost any system of interest has a decaying memory it can be approximated as well as desired by a system with finite memory. The goal in this work is to set up approximate models or representations of classes of systems to be used in identification, so approximation is permissible. It is certainly realistic to stipulate that the observation periods for both inputs and outputs be of finite duration, and this requirement influences the mathematical structure chosen.

The first section after this Introduction is devoted to setting up and investigating certain function spaces which are appropriate for modeling input and output spaces for systems; the second to establishing basic facts about causal and bounded memory systems. In the third section the concept of trajectories of time-limited truncations of systems is developed. The time-limited truncations are, roughly speaking, observable portions of a system which is operating for all time. In the fourth section families of trajectories associated with a class of systems

* Received by the editors May 17, 1973.

† Department of Aerospace Engineering, College of Engineering, University of Michigan, Ann Arbor, Michigan 48105. This work was supported by the United States Air Force, Air Force Office of Scientific Research, Air Force Systems Command, under Grant 72-2328.

are considered. Under certain circumstances, these trajectories can be generated by semigroups of linear operators.

The results of Part I are not used explicitly in Part II till near the end of the fourth section. However, the work of Part I influences what is done in Part II throughout. Some of the material of § § 1 and 3 appeared, with only partial proofs, in the conference paper [1].

1. Some function spaces for inputs and outputs. We want to treat systems for which the inputs and outputs are functions of time, real or vector-valued with finitely many components, and extending for infinite time. We do not want it to be required that the inputs and outputs must always die out in some sense in the infinite future or infinite past. Hence, function spaces, such as the L_p -spaces with $1 \leq p < \infty$, which have the property that their constituent functions all get arbitrarily small (in some sense) as $t \rightarrow \pm \infty$, will usually not be suitable for modeling admissible collections of inputs or outputs. The spaces of bounded or essentially bounded functions are satisfactory on this score, and sometimes we shall use the space of bounded continuous functions on \mathbb{R}^1 with the sup norm. However, there are certain nonstandard function spaces that are suitable and especially convenient, and which will be used customarily. These are spaces of functions that are uniformly local-time L_2 provided with one of a family of norms to be given in the definition below. These spaces are only Banach spaces, but the local L_2 character is advantageous. Since it is quite as easy to define such spaces more generally using a local-time L_p property, $1 \leq p < \infty$, we do so, even though the local L_2 -spaces are the ones chiefly desired.

Let y be either a real-valued function of a real variable, or a vector-valued function that has finitely many real-valued components. In the second case, $|y(t)|$ will denote the Euclidean norm of the vector $y(t)$. Define the operator P_t by

$$(1) \quad [P_t y](s) = \begin{cases} y(s), & s \leq t, \\ 0, & s > t. \end{cases}$$

As usual, let $L_p(A)$ denote the Lebesgue space of p -integrable N -vector-valued functions on the measurable set $A \subset \mathbb{R}^1$. It will always be assumed that $1 \leq p < \infty$. The L_p -norm is written $\|\cdot\|_p$. Let $\mathcal{L}_0^{(p)}$ be the space of all functions y that satisfy the following condition: $y \in \mathcal{L}_0^{(p)}$ iff for any T , $0 < T < \infty$, there is a positive number $K = K(T, y)$ such that $\|(P_{t+T} - P_t)y\|_p \leq K$ for all $t \in \mathbb{R}^1$. Obviously $\mathcal{L}_0^{(p)}$ is a linear space over the real numbers under the usual addition and scalar multiplication of functions. It is made into a normed linear space by the assignment of a norm:

$$(2) \quad \|y\|_T^{(p)} = \sup_t (\|(P_{t+T} - P_t)y\|_p),$$

where T is an arbitrary fixed number, $0 < T < \infty$. We call the resulting normed linear space $\mathcal{L}_T^{(p)}$.

PROPOSITION II.1. $\mathcal{L}_T^{(p)}$ is a Banach space if its elements are interpreted to be the equivalence classes of functions in $\mathcal{L}_0^{(p)}$ that are equal a.e. Lebesgue.

Proof. It is immediately verifiable that $\mathcal{L}_T^{(p)}$ is indeed a normed linear space, so it remains only to show it is complete. Consider a particular set of half-open, half-closed intervals $(kT, (k+1)T]$, where k is any integer, $-\infty < k < \infty$. Since

for any t , $(t, t + T] \subset (kT, (k + 1)T] \cup ((k + 1)T, (k + 2)T]$ for some k , we have

$$\begin{aligned} \sup_t \|(P_{t+T} - P_t)y\|_p &\leq \sup_k \|(P_{(k+2)T} - P_{kT})y\|_p \\ &\leq 2 \sup_k \|(P_{(k+1)T} - P_{kT})y\|_p \end{aligned}$$

or $\|y\|_T^{(p)} \leq 2 \sup_k \|(P_{(k+1)T} - P_{kT})y\|_p \leq 2\|y\|_T^{(p)}$. Now let $\{y_n\}$ be a sequence such that for any $\varepsilon > 0$, $\|y_m - y_n\|_T^{(p)} \leq \varepsilon$ whenever $m, n \geq n_0(\varepsilon)$. Then, for any integer k ,

$$\|(P_{(k+1)T} - P_{kT})(y_n - y_m)\|_p \leq \varepsilon \quad \text{for } m, n \geq n_0(\varepsilon).$$

Since $L_p(kT, (k + 1)T]$ is complete, there is a $y^{(k)} \in L_p(kT, (k + 1)T]$ which is the limit of the sequence $\{(P_{(k+1)T} - P_{kT})(y_n)\}$, regarded as a sequence of functions on $(kT, (k + 1)T]$. Let y be the equivalence class of functions on \mathbb{R}^1 that are equal a.e. on each interval $(kT, (k + 1)T]$ to any function representing $y^{(k)}$. Then $y \in \mathcal{L}_T^{(p)}$, and

$$\|y - y_n\|_T^{(p)} \leq 2 \sup_k \|(P_{(k+1)T} - P_{kT})(y - y_n)\|_p.$$

Since

$$\|(P_{(k+1)T} - P_{kT})(y - y_n)\|_p \leq 2\varepsilon \quad \text{for } n \geq n_0(\varepsilon)$$

and all k , $\|y - y_n\|_T^{(p)} \leq 4\varepsilon$ for $n \geq n_0(\varepsilon)$. Thus y is the limit of y_n . \square

Henceforth, unless there is some particular reason to be precise, we shall refer to the “functions in $\mathcal{L}_T^{(p)}$ ” or the “functions in $L_p(A)$ ” instead of to the elements, which are properly equivalence classes of functions. Some elementary properties of the spaces $\mathcal{L}_T^{(p)}$ are noted in the next proposition and succeeding remarks.

PROPOSITION II.2. *Let N , the number of (real) components of the vector-valued functions under consideration, be fixed. Let T_1, T_2 be any positive numbers. Then*

(a) $\mathcal{L}_{T_1}^{(p)}$ and $\mathcal{L}_{T_2}^{(p)}$ are comprised of the same elements, and the norms on these two spaces are equivalent.

(b) If $p > q$, then any function belonging to $\mathcal{L}_T^{(p)}$ belongs also to $\mathcal{L}_T^{(q)}$. Also, convergence in $\mathcal{L}_T^{(p)}$ implies convergence in $\mathcal{L}_T^{(q)}$.

Proof. The elements of $\mathcal{L}_{T_1}^{(p)}$ and $\mathcal{L}_{T_2}^{(p)}$ are the elements of $\mathcal{L}_0^{(p)}$ (for the fixed N in question). Suppose $T_1 < T_2$ and m is an integer such that $mT_1 > T_2$, then for $y \in \mathcal{L}_0^{(p)}$, $\|y\|_{T_1} \leq \|y\|_{T_2} \leq m\|y\|_{T_1}$.

Part (b) follows from Hölder's inequality. In fact,

$$\begin{aligned} \|y\|_T^{(q)} &= \sup_t \|(P_{t+T} - P_t)y\|_q \\ &\leq \sup_t \|(P_{t+T} - P_t)y\|_p \cdot T^{(p-q)/pq} \\ &= T^{(p-q)/pq} \|y\|_T^{(p)}. \quad \square \end{aligned}$$

Let $\mathcal{M}(\Delta)$ denote the set of functions in $\mathcal{L}_0^{(p)}$ which vanish a.e. outside the interval $\Delta = (a, b)$, where $-\infty \leq a < b \leq +\infty$. Then clearly $\mathcal{M}(\Delta)$ is a closed linear subspace of $\mathcal{L}_T^{(p)}$ for any $0 < T < \infty$. If a and b are finite, $\mathcal{M}(\Delta)$ may be identified with either $L_p(\Delta)$ or with a closed linear subspace of $L_p(\mathbb{R}^1)$, and the

one-to-one correspondence in either case is a linear homeomorphism. If $T = b - a$, the correspondence is isometric.

We denote the operations of translation by c to the left or right, respectively, by L_c and R_c ; i.e., $(L_c u)(t) = u(t + c)$. L_c and R_c are linear operations on $\mathcal{L}_0^{(p)}$ which preserve norm in any $\mathcal{L}_T^{(p)}$, and $L_c = R_{-c} = R_c^{-1}$. The following identities hold for any function u defined on \mathbb{R}^1 and any real numbers a, b and c :

$$(3) \quad \begin{aligned} L_c(P_b - P_a)u &= (P_{b-c} - P_{a-c})L_c u, \\ R_c(P_{b-c} - P_{a-c})u &= (P_b - P_a)R_c u. \end{aligned}$$

It will sometimes be convenient in order to avoid an awkward locution to apply L_c or R_c to elements of an $L_p(\Delta)$, for some finite interval Δ . When this is done it will always be intended that $L_p(\Delta)$ be identified with $\mathcal{M}(\Delta)$, as above, so that the operation is defined. Care must be taken of course to ensure that this operation is meaningful.

Compactness of input spaces is required for much of the structure described in Part I. Because that structure is to be applied to what follows, compactness in some form will again often appear as a requirement, but usually not as the condition that an input space be a compact subset of an $\mathcal{L}_T^{(p)}$ -space. A weaker condition is appropriate, one which says that ordinary compactness only hold locally in time (not local compactness). This notion is formalized, and it is proved below that there is an abundance of subsets of $\mathcal{L}_T^{(p)}$ which have this property along with certain other desirable properties.

A subset \mathcal{A} of $\mathcal{L}_0^{(p)}$ is T -compact if $(P_{t+T} - P_t)\mathcal{A}$ regarded as a subset of $L_p(t, t + T]$ is compact for every t . Relative T -compactness is defined correspondingly.

PROPOSITION II.3. *If \mathcal{A} is a compact subset of $\mathcal{L}_T^{(p)}$, it is T -compact, but the converse is not necessarily true. Also, if \mathcal{A} is T -compact for a positive number T , it is T_1 -compact for any other positive number T_1 .*

Proof. The proof is obvious. \square

Another property of input sets that will be essential is the following: a subset \mathcal{U} of $\mathcal{L}_0^{(p)}$ will be said to have the *projection property*, denoted (P), if $u \in \mathcal{U}$ implies that $P_t u$, $(I - P_t)u$ and $(P_t - P_s)u$ belong to \mathcal{U} for any real numbers s and t . Of course, if $s = t$, $(P_t - P_s)u = 0$, so in particular the zero function must belong to \mathcal{U} .

PROPOSITION II.4. *Let \mathcal{U}_0 be any compact subset of $L_p(0, T]$. Then there exists a set $\mathcal{U} \subset \mathcal{L}_T^{(p)}$ with the following properties:*

- (i) $(P_T - P_0)\mathcal{U} \supset \mathcal{U}_0$ (using the identification explained previously between $L_p(0, T]$ and $\mathcal{M}(0, T]$);
- (ii) \mathcal{U} has property (P);
- (iii) \mathcal{U} is T -compact;
- (iv) \mathcal{U} is invariant under time shift; thus if $u \in (P_{t+T} - P_t)\mathcal{U}$, then

$$L_t u \in (P_T - P_0)\mathcal{U},$$

and vice versa.

Proof. We first enlarge \mathcal{U}_0 so as to have a set that is closed under the projections $(P_t - P_s)$, $0 \leq s, t \leq T$. Let \mathcal{U}'_0 be the subset of $L_p(0, T]$ consisting of all functions u' satisfying

$$u' = (P_t - P_s)u, \quad u \in \mathcal{U}_0, \quad 0 \leq s, t \leq T,$$

a.s. \mathcal{U}'_0 is compact in $L_p(0, T]$. In fact, let $\{u'_n\}$ be an infinite sequence of elements of \mathcal{U}'_0 ; $u'_n = (P_{t_n} - P_{s_n})u_n$. Form a subsequence of the positive integers, $\{n_i\}_i$, such that $t_{n_i} \rightarrow t_0$, $s_{n_i} \rightarrow s_0$ and $\|u_{n_i} - u_0\|_p \rightarrow 0$ as $i \rightarrow \infty$. Put $u'_0 = (P_{t_0} - P_{s_0})u_0$. Then

$$\begin{aligned} \|u'_{n_i} - u'_0\|_p &\leq \|(P_{t_{n_i}} - P_{s_{n_i}})u_{n_i} - (P_{t_{n_i}} - P_{s_{n_i}})u_0\|_p \\ &\quad + \|(P_{t_{n_i}} - P_{s_{n_i}})u_0 - (P_{t_0} - P_{s_0})u_0\|_p \\ &\leq \|u_{n_i} - u_0\|_p + \|(P_{t_{n_i}} - P_{t_0})u_0\|_p + \|(P_{s_0} - P_{s_{n_i}})u_0\|_p, \end{aligned}$$

which is arbitrarily small for sufficiently large i . \mathcal{U}'_0 is closed under the projections $(P_t - P_s)$ since $(P_t - P_s)u$, $u \in \mathcal{U}'_0$, can always be written $(P_{t_1} - P_{s_1})u$, $u \in \mathcal{U}_0$, for some t_1 and s_1 .

Now construct a subset $\tilde{\mathcal{U}}$ of $\mathcal{L}_T^{(p)}$ as follows: let the elements $u \in \tilde{\mathcal{U}}$ be defined by

$$u(t) = \begin{cases} u_0(t), & 0 < t \leq T, \\ u_{-1}(t + T), & -T < t \leq 0, \\ u_1(t - T), & T < t \leq 2T, \\ \dots & \\ u_k(t - kT), & kT < t \leq (k + 1)T, \\ \dots & \end{cases}$$

where the u_k are any sequence of elements from \mathcal{U}'_0 . Put $\mathcal{U} = \bigcup_{0 \leq \eta \leq T} L_\eta \tilde{\mathcal{U}}$. Obviously $\mathcal{U} \supset \mathcal{U}_0$ and $\mathcal{U} \subset \mathcal{U}_T^{(p)}$.

\mathcal{U} has property (P) since $\tilde{\mathcal{U}}$ does and translation does not affect the property.

\mathcal{U} is invariant under time shift. By the way it is constructed, \mathcal{U} is invariant under shifts which are integer multiples of T . Since any shift can be decomposed into a shift by NT for some integer N , and a shift L_η , $0 \leq \eta < T$, \mathcal{U} is invariant for any shift.

\mathcal{U} is T -compact. It is sufficient to prove that $(P_T - P_0)\mathcal{U}$ is compact. Let $\{z_n\}$ be an infinite sequence contained in $(P_T - P_0)\mathcal{U}$. By the construction given, each z_n must be of the form

$$z_n = (I - P_0)L_{\eta_n}u_n + P_TR_{T-\eta_n}u'_n,$$

where $u_n, u'_n \in \mathcal{U}'_0$, and $0 \leq \eta_n \leq T$. Let n_i be a subsequence such that $\|u_{n_i} - u_0\|_p \rightarrow 0$, $\|u'_{n_i} - u'_0\|_p \rightarrow 0$ and $\eta_{n_i} \rightarrow \eta$, where u_0 and u'_0 belong to \mathcal{U}'_0 and $0 \leq \eta \leq T$. There is such a subsequence because of the compactness of \mathcal{U}'_0 (and of the interval $[0, T]$). Then

$$\begin{aligned} \lim_{i \rightarrow \infty} z_{n_i} &= (I - P_0)L_\eta u_0 + P_TR_{T-\eta}u'_0 \\ &= L_\eta(I - P_\eta)u_0 + R_{T-\eta}P_\eta u'_0. \end{aligned}$$

In fact, since $\|L_\alpha u - u\|_p \rightarrow 0$ as $\alpha \rightarrow 0$, it follows that $L_{\eta_{n_i}}u_{n_i} \rightarrow L_\eta u_0$ and $R_{T-\eta_{n_i}}u'_{n_i} \rightarrow R_{T-\eta}u'_0$ by the triangle inequality. The limit element belongs to $(P_T - P_0)\mathcal{U}$ by the definition of \mathcal{U} . \square

T -compactness cannot be replaced by compactness here. In fact, it is trivially verifiable that if \mathcal{U}_0 has even two distinct elements, then any \mathcal{U} satisfying (i), (ii)

and (iv) in the theorem is not compact. Indeed, \mathcal{U} need only be invariant under shifts by integer multiples of T to make compactness impossible: with no loss of generality let \mathcal{U}_0 consist of the functions $f_0(t) = 0$, $0 \leq t < T$, and $f_1(t) = 1$, $0 \leq t < T$. Then it is sufficient to observe that the sequence $\{u_n\}$, u_n defined by $u_n(t) = 1$, $nT \leq t < (n+1)T$, $u_n(t) = 0$ for all other $t \in \mathbb{R}^1$, has no limit point.

The construction given in the proof of Proposition II.4 depends on T and happens to give a class \mathcal{U} that includes all the periodic functions of period T generated by \mathcal{U}'_0 . However, we remark that if \mathcal{U} is a subset of $\mathcal{L}_0^{(p)}$ which is T_1 -compact, shift-invariant and has property (P), then it is T_2 -compact and, of course, still shift-invariant with property (P). Thus, in modeling a system, an input space \mathcal{U} can be chosen with the desirable properties listed without any consideration being given to the value of T to be used.

Clearly the bounded continuous functions on \mathbb{R}^1 (denoted¹ \mathcal{B}_c) are contained in all $\mathcal{L}_0^{(p)}$, and convergence in the uniform norm of \mathcal{B}_c implies convergence in $\mathcal{L}_T^{(p)}$ for any p and any T . It is easy to see also that the functions of \mathcal{B}_c are not dense in any $\mathcal{L}_T^{(p)}$. We give a characterization of the closed subspace of $\mathcal{L}_T^{(p)}$ which is generated by \mathcal{B}_c .

PROPOSITION II.5. *For any $y \in \mathcal{L}_0^{(p)}$, let $y^{(k)} \in L_p[0, T]$ be defined by $y^{(k)}(t) = y(t + kT)$, $0 \leq t < T$, for all integers k . Then, a necessary and sufficient condition that y can be approximated in the $\mathcal{L}_T^{(p)}$ -norm by functions from \mathcal{B}_c is that the functions $[y^{(k)}(t)]^p$ be uniformly integrable.*

Proof. Sufficiency. Let $B_k(b) = \{t \in [0, T] : |y^{(k)}(t)|^p > b\}$. The condition that the $[y^{(k)}]^p$ are uniformly integrable is that, given any $\eta > 0$, there exists $b > 0$ such that $\int_{B_k(b)} |y^{(k)}(t)|^p dt \leq \eta$ for all integers k . Let $\varepsilon > 0$ be given, and put $\eta = (\varepsilon/2)^p$. Let $y_b^{(k)}(t) = y^{(k)}(t)$ whenever $|y^{(k)}(t)| \leq b^{1/p}$ and equal to zero otherwise. For each k , there is a function $f^{(k)}$ defined on $[0, T]$ which is continuous on $[0, T]$, and satisfies $f^{(k)}(0) = f^{(k)}(T) = 0$, $|f^{(k)}(t)| \leq b^{1/p}$ and $\|y_b^{(k)} - f^{(k)}\|_p \leq \eta$. Then

$$\begin{aligned} \|y^{(k)} - f^{(k)}\|_p &\leq \|y^{(k)} - y_b^{(k)}\|_p + \|y_b^{(k)} - f^{(k)}\|_p \\ &\leq \left[\int_{B_k(b)} |y^{(k)}(t)|^p dt \right]^{1/p} + \eta \leq \eta^{1/p} + \eta \leq \varepsilon \end{aligned}$$

when $\varepsilon < 2$. Now if $f \in \mathcal{B}_c$ is the function formed by piecing together the $f^{(k)}$,

$$\begin{aligned} \|y - f\|_T^{(p)} &= \sup_t \|(P_{t+T} - P_t)(y - f)\|_p \\ &\leq \sup_k \|(P_{(k+2)T} - P_{kT})(y - f)\|_p \\ &\leq 2\varepsilon. \end{aligned}$$

Necessity. Suppose that $\|y - f_n\|_T^{(p)} \rightarrow 0$ as $n \rightarrow \infty$, where the $f_n \in \mathcal{B}_c$. It follows that $\|y^{(k)} - f_n^{(k)}\|_p \rightarrow 0$ uniformly in k . Suppose further that the $[y^{(k)}]^p$ are not uniformly integrable; we shall obtain a contradiction. Then, for some $\varepsilon > 0$, $\varepsilon < \frac{1}{3}$,

¹ To be consistent, what is here denoted \mathcal{B}_c should be $\mathcal{S}(\mathbb{R}^1, \mathbb{R}^n)$, but it seems less confusing to introduce a new symbol for this special case.

and for every real number $b > 0$, no matter how large, there is an integer $k' = k'(b, \varepsilon)$ such that

$$\int_{B_{k'(b)}} |y^{(k')}(t)|^p dt > 3\varepsilon.$$

For the same ε , let n be a fixed integer so large that $\|y^{(k)} - f_n^{(k)}\|_p \leq \varepsilon$ for all k . We have

$$\|y^{(k)} - f_n^{(k)}\|_p \geq \left[\int_{B_k(b)} |y^{(k)}(t) - f_n^{(k)}(t)|^p dt \right]^{1/p}$$

for any b , for all k . Put $K_n = \sup_t |f_n(t)|$ and $b = 2K_n$. Then, for $k = k'(b, \varepsilon)$,

$$\begin{aligned} |y^{(k')}(t) - f_n^{(k')}(t)| &\geq ||y^{(k')}(t)| - |f_n^{(k')}(t)|| \\ &= |y^{(k')}(t)| - |f_n^{(k')}(t)| \geq |y^{(k')}(t)| - K_n \quad \text{for } t \in B_{k'}(b). \end{aligned}$$

Hence,

$$\begin{aligned} \|y^{(k')} - f_n^{(k')}\|_p &\geq \left[\int_{B_{k'}(b)} (|y^{(k')}(t)| - K_n)^p dt \right]^{1/p} \\ &\geq \left[\left[\int_{B_{k'}(b)} |y^{(k')}(t)|^p dt \right]^{1/p} - \left[\int_{B_{k'}(b)} K_n^p dt \right]^{1/p} \right] \\ &\geq (3\varepsilon)^{1/p} - K_n [\mu(B_{k'}(b))]^{1/p}, \end{aligned}$$

where $\mu(B)$ is the Lebesgue measure of B . But since

$$\varepsilon \geq \|y^{(k')} - f_n^{(k')}\|_p \geq \left[\int_{B_{k'}(b)} (2K_n - K_n)^p dt \right]^{1/p} = K_n [\mu(B_{k'}(b))]^{1/p}$$

and since $3\varepsilon \geq (3\varepsilon)^p$, we have

$$\|y^{(k')} - f_n^{(k')}\|_p \geq 3\varepsilon - \varepsilon = 2\varepsilon$$

which yields a contradiction. \square

Clearly, if $y \in \mathcal{L}_0^{(p)}$ satisfies the condition of Proposition II.5 for $T > 0$ it also satisfies the condition for any other $T' > 0$. Since the value of T is immaterial, we can denote the class of all $y \in \mathcal{L}_0^{(p)}$ which satisfy the condition by $\mathcal{L}_{0u}^{(p)}$. The functions belonging to $\mathcal{L}_{0u}^{(p)}$, or more properly the usual equivalence classes of such functions, belong to $\mathcal{L}_T^{(p)}$ for any $T > 0$, and as a subset of $\mathcal{L}_T^{(p)}$ this class is denoted $\mathcal{L}_{Tu}^{(p)}$. An immediate corollary of Proposition II.5 is the following.

PROPOSITION II.6. $\mathcal{L}_{Tu}^{(p)}$ is a closed linear subspace of $\mathcal{L}_T^{(p)}$ and is the smallest closed linear subspace containing \mathcal{B}_c .

Proof. The proof is obvious. \square

With reference to Proposition II.4 it may be noted that since the set \mathcal{U}_0 is a compact subset of $L_p[0, T]$ the functions belonging to \mathcal{U}_0 are uniformly integrable. Then the construction given for \mathcal{U} guarantees that the p th powers of functions belonging to \mathcal{U} satisfy the hypothesis of Proposition II.5. Hence $\mathcal{U} \subset \mathcal{L}_{Tu}^{(p)}$.

2. Preliminaries on causal and bounded memory transformations. Let \mathcal{U} be a metric space whose elements are either functions of a real variable t (time) or are equivalence classes of such functions that are equal a.e. Lebesgue. Correspondingly, let \mathcal{Y} be a Banach space of functions or equivalence classes of functions of t . If both \mathcal{U} and \mathcal{Y} have property (P), the properties of causality and bounded memory for a mapping F from \mathcal{U} into \mathcal{Y} can be defined, and in the usual way: F is *causal* if $P_t F(u) = P_t F P_t(u)$ for all t and all $u \in \mathcal{U}$; F has *bounded memory* (d) if $(I - P_t)F(u) = (I - P_t)F(I - P_{t-d})(u)$ for all t and all $u \in \mathcal{U}$. Note that the same symbol, P_t , is being used to denote the linear projection on the past in both \mathcal{U} and \mathcal{Y} , but this should cause no confusion.

PROPOSITION II.7. *If F is a mapping from \mathcal{U} into \mathcal{Y} that is causal and has bounded memory (d), then for every $T > 0$,*

$$(4) \quad (P_{t+T} - P_t)F(u) = (P_{t+T} - P_t)F(P_{t+T} - P_{t-d})(u)$$

for all t and $u \in \mathcal{U}$.

Conversely, if (4) is satisfied for some $T > 0$ and all t and all $u \in \mathcal{U}$, then F is causal and has bounded memory (d).

Proof. The assertions appear to be obvious. However a proof is given in Appendix A, where the algebraic properties of the P_t are isolated and are used precisely. \square

In the class of bounded continuous mappings $\mathcal{F} = \mathcal{F}(\mathcal{U}, \mathcal{Y})$, let \mathcal{F}° denote the subclass of causal mappings, and let \mathcal{F}_d° denote the subclass of causal mappings with bounded memory (d). Henceforth we only consider metric function spaces \mathcal{U} that have property (P), and \mathcal{Y} can always be chosen to be either one of the \mathcal{L}_T^p or \mathcal{B} , both of which have property (P). Hence \mathcal{F}° and \mathcal{F}_d° are defined. In some instances, however, where \mathcal{B} could be used for \mathcal{Y} it may be convenient to take \mathcal{Y} to be a subspace of \mathcal{B} that does not possess property (P), e.g., the subspace of bounded continuous functions \mathcal{B}_c . This is all right, because in this situation where the elements of \mathcal{Y} are functions (not equivalence classes of functions) the definition of causality may be replaced by: F is causal if, for all t and all u ,

$$[Fu](s) = [FP_t u](s) \quad \text{for all } s \leq t.$$

An equivalent condition is the apparently weaker statement:

$$[Fu](s) = [FP_s u](s) \quad \text{for all } s \text{ and all } u.$$

In fact, suppose the second condition holds. Since \mathcal{U} has property (P), $P_t u \in \mathcal{U}$ for all $u \in \mathcal{U}$. Take $t > s$. Then

$$[F(P_t u)](s) = [FP_s(P_t u)](s) = [FP_s u](s)$$

and

$$[Fu](s) = [FP_s u](s).$$

Hence,

$$[Fu](s) = [FP_t u](s) \quad \text{for all } s \leq t.$$

Analogous statements hold for the case of bounded memory.

PROPOSITION II.8. \mathcal{F} and \mathcal{F}_d° are closed linear subspaces of \mathcal{F} .

Proof. \mathcal{F}_d° is obviously linear; we need to prove it is closed. First, let us note the following. If $y \in \mathcal{Y} = \mathcal{L}_T^{(p)}$, then by definition,

$$\|y\| = \sup_t \|(P_{t+T} - P_t)y\|_p, \quad y \in \mathcal{Y}.$$

On the other hand if $y \in \mathcal{Y} = \mathcal{B}$, then

$$\|y\| = \sup_t |y(t)| = \sup_t \|(P_{t+T} - P_t)y\|_{\mathcal{B}},$$

where

$$\|(P_{t+T} - P_t)y\|_{\mathcal{B}} = \sup_{t \leq s \leq t+T} |y(s)|$$

is the norm in \mathcal{B} of the truncation of y to $[t, t+T]$. Thus in either case, $\|y\| = \sup_t \|(P_{t+T} - P_t)y\|$, where the norm on the right side is appropriately interpreted.

Now suppose that $F_n \in \mathcal{F}_d^\circ$ and $\lim_n F_n = F$, where $F \notin \mathcal{F}_d^\circ$. Put $\Delta_t = P_{t+T} - P_t$ and $\Delta'_t = P_{t+T} - P_{t-d}$. Then

$$\begin{aligned} \|F_n - F\| &= \sup_{u \in \mathcal{U}} \sup_t \|\Delta_t(F_n u - F u)\| \\ &= \sup_u \sup_t \|\Delta_t F_n \Delta'_t u - \Delta_t F u\| \\ &\leq \sup_u \sup_t \|\Delta_t F_n \Delta'_t u - \Delta_t F \Delta'_t u\| + \|\Delta_t F \Delta'_t u - \Delta_t F u\|. \end{aligned}$$

For some t_0 and u_0 , $\|\Delta_{t_0} F u_0 - \Delta_{t_0} F \Delta'_{t_0} u_0\| \geq \alpha$, $\alpha > 0$, whereas for sufficiently large n_0 ,

$$\|\Delta_{t_0} F_n(\Delta'_{t_0} u) - \Delta_{t_0} F(\Delta'_{t_0} u)\| \leq \|F_n(\Delta'_{t_0} u) - F(\Delta'_{t_0} u)\| \leq \alpha/2, \quad n \geq n_0.$$

Hence $\|F - F_n\| \geq \alpha/2$, $n \geq n_0$, which is a contradiction. The proof for \mathcal{F}° is similar. \square

If F is any mapping from \mathcal{U} into \mathcal{B} , then one can reasonably define the causal part of F , denoted F° , and the causal and bounded memory (d) part of F , denoted F_d° , by

$$[F^\circ u](t) = [F P_t u](t) \quad \text{for all } t, \quad \text{all } u \in \mathcal{U},$$

$$[F_d^\circ u](t) = [F(P_t - P_{t-d})u](t) \quad \text{for all } t, \quad \text{all } u \in \mathcal{U}.$$

For the rest of this section, we assume $\mathcal{Y} = \mathcal{B}$.

PROPOSITION II.9. Let \mathcal{U} have the property that for any $u, u' \in \mathcal{U}$ and any s, t , $d[(P_t - P_s)u, (P_t - P_s)u'] \leq d[u, u']$. Let $F \in \mathcal{F}$. Then a sufficient condition that $F^\circ \in \mathcal{F}^\circ$ and $F_d^\circ \in \mathcal{F}_d^\circ$ is that F be uniformly continuous on \mathcal{U} .

Proof. That F_d° is causal and has bounded memory (d) is shown by a simple verification. F_d° is a mapping into \mathcal{B} and is bounded; in fact, for any u and t ,

$$|[F_d^\circ u](t)| = |[F(P_t - P_{t-d})u](t)| \leq \|F[(P_t - P_{t-d})u]\| \leq \|F\|.$$

To show that F_d° is continuous, choose $\varepsilon > 0$ arbitrarily. Let $\delta > 0$ be small enough that if u', u satisfy $d[u', u] < \delta$, then $\|F u' - F u\| \leq \varepsilon/2$. Take any such pair u', u .

Then there is a t_0 such that

$$\begin{aligned}
 \|F_d^\circ u' - F_d^\circ u\| &\leq |[F_d^\circ u'](t_0) - [F_d^\circ u](t_0)| + \varepsilon/2 \\
 &= |[F(P_{t_0} - P_{t_0-d})u'](t_0) - [F(P_{t_0} - P_{t_0-d})u](t_0)| + \varepsilon/2 \\
 &\leq \|F[(P_{t_0} - P_{t_0-d})u'] - F[(P_{t_0} - P_{t_0-d})u]\| + \varepsilon/2 \\
 &\leq \varepsilon/2 + \varepsilon/2 = \varepsilon
 \end{aligned}$$

by the uniform continuity. Hence F_d° is continuous, and indeed uniformly continuous. The same sort of argument shows that $F^\circ \in \mathcal{F}$. \square

It is to be noted that $F \in \mathcal{F}$ does not by itself necessarily imply that $F^\circ \in \mathcal{F}^\circ$, or $F_d^\circ \in \mathcal{F}_d^\circ$; i.e., a continuous mapping from \mathcal{U} into \mathcal{Y} , where \mathcal{U} and \mathcal{Y} satisfy the conditions of Proposition II.9, does not necessarily have a continuous causal part, nor a continuous causal and bounded memory part. The condition of uniform continuity is perhaps the most obvious condition that guarantees the continuity of F° and F_d° . The following is an example where F_d° is not continuous, even though $F \in \mathcal{F}$.

Consider the set \mathcal{E} of real-valued functions on \mathbb{R}^1 described as follows. Each $u \in \mathcal{E}$ is of the form, for some τ_1, τ_2 , $-\infty \leq \tau_1 < \tau_2 \leq \infty$,

$$u(t) = \begin{cases} 0, & t \leq \tau_1, \\ u(\tau_2), & \tau_1 < t \leq \tau_2, \\ 0, & \tau_2 < t, \end{cases}$$

where $-1 \leq u(\tau_2) \leq 1$, and where the convention is made that if $\tau_1 = -\infty$, $u(t) = u(\tau_2)$, $t \leq \tau_2$; and correspondingly, if $\tau_2 = +\infty$, $u(t)$ has a constant value for all $t > \tau_1$. Thus the constant functions and the functions that are constant except for a single step up from zero or down to zero are included. Obviously $\mathcal{E} \subset \mathcal{B}$ has property (P).

Let F , a mapping from \mathcal{E} into \mathcal{B} , be defined as follows:

$$[Fu](t) = \begin{cases} 0, & t \leq \tau_1, \\ \phi(u(\tau_2), \tau_2), & \tau_1 < t \leq \tau_2, \\ 0, & \tau_2 < t, \end{cases}$$

where

$$\phi(a, \tau) = \begin{cases} |a| & \text{if } |\tau| \leq 1, \\ |\tau| \cdot |a| + 1 - |\tau| & \text{if } 1 < |\tau| < 1/(1 - |a|), \\ 0 & \text{if } |\tau| \geq 1/(1 - |a|). \end{cases}$$

It will be noted that F carries all the constant functions in \mathcal{E} , and in fact all the functions in \mathcal{E} with $\tau_2 = +\infty$, into zero. It is readily verified that F is a bounded continuous mapping from \mathcal{E} (regarded as a metric subspace of \mathcal{B}) into \mathcal{B} . Actually, the range of F is contained in \mathcal{E} .

Now, let $u(t) \equiv 1$ and $u_n(t) \equiv 1 - 1/n$, $n = 1, 2, \dots$. The functions u and $u_n \in \mathcal{E}$, and $u_n \rightarrow u$ in \mathcal{E} . Consider $F^\circ u$,

$$\begin{aligned} [F^\circ u](t) &= [FP_t u](t) \\ &= u(t) = 1 \quad (|t| < 1) \\ &= |t|u(t) + 1 - |t| = 1 \quad (|t| < 1/(1 - 1) = \infty), \end{aligned}$$

i.e., $[F^\circ u](t) \equiv 1$. On the other hand,

$$[F^\circ u_n](t) = [FP_t u_n](t) = \begin{cases} 1 - 1/n, & |t| \leq 1, \\ |t|(1 - 1/n) + 1 - |t| = 1 - |t|/n, & 1 < |t| < n, \\ 0, & |t| > n. \end{cases}$$

Thus $F^\circ u_n$ does not converge to $F_0 u$ in \mathcal{B} , although of course $[F^\circ u_n](t) \rightarrow [F^\circ u](t)$ for each t . Hence F° is not a continuous mapping. For these particular u and u_n , $F_1^\circ u = F^\circ u$, and $F_1^\circ u_n = F^\circ u_n$, so it follows that F_1° is not continuous either.

The following very simple result gives some justification for introducing the concepts of causal part and of bounded memory causal parts of mappings, at least when the intended use of these mappings is for approximation.

PROPOSITION II.10. *Let F and $F_d^\circ \in \mathcal{F}$. If for some $\alpha > 0$ there is a $G \in \mathcal{F}_d^\circ$ such that $\|F - G\| \leq \alpha$, then $\|F - F_d^\circ\| \leq 2\alpha$. The corresponding statement is true for F° and $G \in \mathcal{F}^\circ$.*

Proof. For any $\varepsilon > 0$ there is a $u \in \mathcal{U}$ and a t_0 such that

$$\begin{aligned} \|G - F_d^\circ\| &\leq |[Gu](t_0) - [F(P_{t_0} - P_{t_0-d})u](t_0)| + \varepsilon/2 \\ &= |[G(P_{t_0} - P_{t_0-d})u](t_0) - [F(P_{t_0} - P_{t_0-d})u](t_0)| + \varepsilon/2 \\ &\leq \|G - F\| + \varepsilon/2. \end{aligned}$$

Hence $\|G - F_d^\circ\| \leq \|G - F\|$ and $\|F - F_d^\circ\| \leq 2\alpha$. \square

3. Finite-time-interval projections of systems and their trajectories. The general situation to be discussed next is the following. The kind of system in question consists of an input space \mathcal{U} of functions of time, an output space \mathcal{Y} of functions of time, and a continuous bounded mapping F from \mathcal{U} into \mathcal{Y} . The mapping F may or may not be causal and of finite memory, but there is some emphasis on the case where it is. Such a system operates for infinite time. We want to look at pieces of the system corresponding to finite observation intervals for both input and output, and at the relations among such pieces and between them and the entire system. Each real number t can be taken to be the epoch of an observation interval. If the observation intervals are of fixed duration, then as t changes, a trajectory of comparable finite-time systems is generated by the original system. The elementary properties of these trajectories are investigated in this section.

It is assumed for the remainder of the paper that \mathcal{U} is a subset of $\mathcal{L}_0^{(p)}$ for some fixed p , $1 \leq p < \infty$, and that it is T -compact, shift-invariant and has property (P). \mathcal{U} is to be regarded as a metric subspace of $\mathcal{L}_{T+d}^{(p)}$ for some T and $d > 0$ as

given. Since all $\mathcal{L}_T^{(p)}$ -spaces with the same p are topologically equivalent, changing T and d changes only the metric on \mathcal{U} ; it does not affect the T -compactness. We have then always

$$\|u\| = \sup_t \|(P_{t+T} - P_{t+T-d})u\|_p, \quad u \in \mathcal{U}.$$

\mathcal{Y} is always either an $\mathcal{L}_T^{(p)}$ -space or \mathcal{B} , the Banach space of bounded functions on \mathbb{R}^1 with the uniform norm, or a closed linear subspace of one of these. In the propositions of this section, whenever \mathcal{Y} is to be one of some particular class of spaces that fact is stated; otherwise it may be any of the spaces just indicated. One slight technical annoyance is that sometimes it is desirable to take \mathcal{Y} to be \mathcal{B}_c , the bounded continuous functions regarded as a subspace of \mathcal{B} , but this space quite obviously does not have property (P), which is usually needed. It is not always satisfactory just to replace \mathcal{B}_c with \mathcal{B} in every statement, but it will be clear that when necessary, \mathcal{B}_c can be imbedded in \mathcal{B} in order to make the calculations meaningful. $\mathcal{F} = \mathcal{F}(\mathcal{U}, \mathcal{Y})$ is the family of bounded continuous mappings from \mathcal{U} into \mathcal{Y} made into a Banach space with the sup norm, as before. Thus,

$$\|F\| = \sup_{u \in \mathcal{U}} \sup_t \|(P_{t+T} - P_t)F(u)\|_{\mathcal{Y}}, \quad F \in \mathcal{F},$$

in all cases, where the norm on the right is the L_p -norm or the uniform norm as appropriate.

We now introduce notations for the finite-time pieces of a system. Let $T > 0$ and $d > 0$ be given. Put

$$(5) \quad \mathcal{U}'_{t,T} \stackrel{\text{def}}{=} (P_{t+T} - P_{t-d})\mathcal{U},$$

$$(6) \quad F'_{t,T}u' \stackrel{\text{def}}{=} (P_{t+T} - P_t)F u', \quad u' \in \mathcal{U}'_{t,T}.$$

Equation (6) does define a mapping on $\mathcal{U}'_{t,T}$ since u' belongs to the domain of F by property (P). Further, because of shift-invariance we can write $\mathcal{U}_T \stackrel{\text{def}}{=} L_t \mathcal{U}'_{t,T} = \mathcal{U}'_{0,T}$ for all t . Define $F_{t,T}$ by

$$(7) \quad \begin{aligned} F_{t,T}z &\stackrel{\text{def}}{=} L_t F'_{t,T} R_t z = L_t (P_{t+T} - P_t) F R_t z \\ &= (P_T - P_0) L_t F R_t z, \quad z \in \mathcal{U}_T. \end{aligned}$$

If \mathcal{Y} has property (P), then $F_{t,T}$ is a mapping from \mathcal{U}_T into \mathcal{Y} , and clearly it is bounded and continuous. But $F_{t,T}$ can also always be regarded as a mapping into a smaller space, denoted \mathcal{Y}_T . If $\mathcal{Y} = \mathcal{L}_T^{(p)}$, then $F_{t,T}$ is a bounded continuous mapping into $\mathcal{Y}_T = L_p(0, T]$, and with the same norm as if its range space is taken to be \mathcal{Y} . Similarly, if $\mathcal{Y} = \mathcal{B}$, $F_{t,T}$ is a bounded continuous mapping into $\mathcal{Y}_T = \mathcal{B}(0, T]$ with the same norm; even if $\mathcal{Y} = \mathcal{B}_c$, $F_{t,T}$ is a bounded continuous mapping into $\mathcal{Y}_T = \mathcal{B}_c(0, T]$, although it is not a mapping into \mathcal{Y} .

If T is fixed throughout a calculation, we write simply F_t for $F_{t,T}$. It often avoids confusion to write

$$F_t = (P_T - P_0) L_t F R_t (P_t - P_{-d})$$

even though the projection on the right is redundant. When we are dealing with mappings F with finite memory, d is usually chosen to be the duration of the

memory; however the above definitions are to be applied in the general case, whether F is causal with finite memory or not. Causality and finite memory are not to be assumed in what follows unless explicitly stipulated.

Let $\pi_t: \mathcal{F}(\mathcal{U}, \mathcal{Y}) \rightarrow \mathcal{F}(\mathcal{U}_T, \mathcal{Y}_T)$ be the mapping that carries F into F_t according to equation (7).

PROPOSITION II.11. *The mapping π_t is linear and continuous; in fact, $\|\pi_t F\| \leq \|F\|$.*

Proof. The linearity is obvious. Also

$$\begin{aligned} \|\pi_t F\| &= \|F_t\| = \sup_{u \in \mathcal{U}_T} \|(P_T - P_0)L_t F R_t u\| \\ &\leq \sup_{u \in \mathcal{U}_T} \|F R_t u\| \leq \sup_{\mathcal{U}} \|F z\|_{\mathcal{Y}} = \|F\|, \end{aligned}$$

where the norm on the left is for the space $\mathcal{F}(\mathcal{U}_T, \mathcal{Y}_T)$. \square

For each t , $F_t = \pi_t F$ is an element of $\mathcal{F}(\mathcal{U}_T, \mathcal{Y}_T)$, so as t runs through \mathbb{R}^1 a trajectory is generated in $\mathcal{F}(\mathcal{U}_T, \mathcal{Y}_T)$ corresponding to F . If F is a time-invariant mapping, this "trajectory" reduces to a single point, of course, but we are interested in time-invariant systems only as a special case. Since in general for time-varying systems these trajectories describe the evolution of the systems, we wish to investigate their properties. Note that the trajectories depend on T ; however, for now, we keep T fixed arbitrarily.

PROPOSITION II.12. *Let $\mathcal{U} \subset \mathcal{L}_T^{(q)}$, $1 \leq q < \infty$, and \mathcal{Y} be $\mathcal{L}_T^{(p)}$, $1 \leq p < \infty$, or \mathcal{B}_c . Then, for any $F \in \mathcal{F}(\mathcal{U}, \mathcal{Y})$ the trajectories $F_t = \pi_t F$ with values in $\mathcal{F}(\mathcal{U}_T, \mathcal{Y}_T)$ are continuous in t . Furthermore, if \mathcal{H} is a compact subset of $\mathcal{F}(\mathcal{U}, \mathcal{Y})$, then the trajectories $F_t = \pi_t F$, $F \in \mathcal{H}$, are equicontinuous functions of t .*

Proof. Suppose $\mathcal{Y} = \mathcal{L}_T^{(p)}$. Then

$$\begin{aligned} \|F_t - F_{t+h}\| &= \sup_{u \in \mathcal{U}_T} \|(F_t - F_{t+h})u\|_p \\ &= \sup_{\mathcal{U}_T} \|(P_T - P_0)L_t F R_t u - (P_T - P_0)L_{t+h} F R_{t+h} u\|_p \\ &\leq \sup_{\mathcal{U}_T} \|(P_T - P_0)L_t F R_t u - (P_T - P_0)L_{t+h} F R_t u\|_p \\ &\quad + \sup_{\mathcal{U}_T} \|(P_T - P_0)L_{t+h} F R_t u - (P_T - P_0)L_{t+h} F R_{t+h} u\|_p. \end{aligned}$$

Denote the first term on the right-hand side of the inequality by I, the second by II. Then,

$$\begin{aligned} \text{I} &\leq \sup_{\mathcal{U}_T} \|L_t(P_{T+t} - P_t)F R_t u - L_{t+h}(P_{T+t} - P_t)F R_t u\|_p \\ &\quad + \sup_{\mathcal{U}_T} \|L_{t+h}(P_{T+t} - P_t)F R_t u - L_{t+h}(P_{T+t+h} - P_{t+h})F R_t u\|_p. \end{aligned}$$

Denote the two terms on the right-hand side of this inequality by I_a , I_b , respectively. Then

$$\begin{aligned} I_b &= \sup_{\mathcal{U}_T} \|L_{t+h}[P_{T+t} - P_{T+t+h} + P_{t+h} - P_t]F R_t u\|_p \\ &\leq \sup_{\mathcal{U}_T} \|(P_{T+t} - P_{T+t+h})F R_t u\|_p + \sup_{\mathcal{U}_T} \|(P_{t+h} - P_t)F R_t u\|_p. \end{aligned}$$

Now, $F(R_t \mathcal{U}_T)$ is a compact subset of \mathcal{Y} since \mathcal{U}_T is compact in L_q (by the T -compactness of \mathcal{U}) and FR_t is continuous. Let $y_i, i = 1, \dots, N$, be a set of points in \mathcal{Y} such that the balls of radius ε about the y_i cover $F(R_t \mathcal{U}_T)$. Then the first term in the expression dominating I_b is in turn dominated by

$$\sup_{\mathcal{U}_T} \min_{i=1, \dots, N} \{ \|(P_{T+t} - P_{T+t+h})(FR_t u - y_i)\|_p + \|(P_{T+t} - P_{T+t+h})y_i\|_p \}.$$

Let h be sufficiently small that

$$\|(P_{T+t} - P_{T+t+h})y_i\|_p \leq \varepsilon \quad \text{for all } i = 1, \dots, N.$$

Then for any such h the above expression has a value $\leq 2\varepsilon$. The second term in the expression dominating I_b can be treated in the same way, so we have that for some $h_1 > 0$, $I_b \leq 4\varepsilon$ whenever $|h| < h_1$.

The term I_a can be written as

$$\sup_{\mathcal{U}_T} \|(L_t - L_{t+h})(P_{t+T} - P_t)FR_t u\|_p.$$

Since $(P_{t+T} - P_t)FR_t \mathcal{U}_T$ is a compact subset of L_p one can choose a finite set $\{z_i\}, i = 1, \dots, M$, of elements of L_p such that the balls of radius ε about the z_i cover $(P_{t+T} - P_t)FR_t \mathcal{U}_T$. Then, since for h sufficiently small $\|(L_t - L_{t+h})z_i\|_p \leq \varepsilon$ for all $i = 1, \dots, M$, we have, very much as above, that for some $h_2 > 0$, $I_a \leq 2\varepsilon$ whenever $|h| \leq h_2$.

To bound Π , we have

$$\begin{aligned} \Pi &= \sup_{\mathcal{U}_T} \|L_{t+h}(P_{t+T+h} - P_{t+h})[FR_t u - FR_{t+h} u]\|_p \\ &\leq \sup_{\mathcal{U}_T} \|(P_{2T+t} - P_{t-T})[FR_t u - FR_{t+h} u]\|_p \end{aligned}$$

when $|h| \leq T$. Now $\mathcal{K} = \bigcup_{|h| \leq T} R_h \mathcal{U}_T$ is a compact subset of L_q (as in Proposition II.4) and $(P_{2T+t} - P_{t-T})FR_t$ is a uniformly continuous mapping from \mathcal{K} into L_p . Hence, there is an $\eta > 0$ so that

$$\|(P_{t+2T} - P_{t-T})FR_t u' - (P_{t+2T} - P_{t-T})FR_t u''\|_p \leq \varepsilon$$

whenever $\|u' - u''\|_q \leq \eta$. Let $\{w_i\}, i = 1, \dots, K$, be the centers of balls of radius η that cover \mathcal{K} . Then

$$\begin{aligned} \Pi &\leq \sup_{\mathcal{U}_T} \{ \|(P_{t+2T} - P_{t-T})FR_t u - (P_{t+2T} - P_{t-T})FR_t w_i\|_p \\ &\quad + \|(P_{t+2T} - P_{t-T})FR_t w_i - (P_{t+2T} - P_{t-T})FR_t R_h u\|_p \}. \end{aligned}$$

Let $h_3 > 0$ be small enough that $\|R_h w_i - w_i\|_q \leq \eta/2$ for all $i = 1, \dots, K$ whenever $|h| < h_3$, and temporarily fix such an h . With this fixed value of h , there is u_0 so that the supremum in the inequality above is realized to within ε by $u = u_0$. This gives

$$\begin{aligned} \Pi &\leq \|(P_{t+2T} - P_{t-T})FR_t u_0 - (P_{t+2T} - P_{t-T})FR_t w_i\|_p \\ &\quad + \|(P_{t+2T} - P_{t-T})FR_t w_i - (P_{t+2T} - P_{t-T})FR_t R_h u_0\|_p \\ &\quad + \varepsilon \quad \text{for all } i = 1, \dots, K. \end{aligned}$$

There is at least one w_i so that $\|u_0 - w_i\|_q \leq \eta/2$; choose such a w_i . Then the first term on the right is $\leq \varepsilon$. With this particular w_i ,

$$\begin{aligned} \|R_h u_0 - w_i\|_q &\leq \|R_h u_0 - R_h w_i\|_q + \|R_h w_i - w_i\|_q \\ &= \|u_0 - w_i\|_q + \|R_h w_i - w_i\|_q \\ &\leq \eta/2 + \eta/2 = \eta \end{aligned}$$

so the second term is also $\leq \varepsilon$. Thus for $|h| < h_3$, $\Pi \leq 3\varepsilon$. Combining these estimates gives the result that if $|h| \leq \max(h_1, h_2, h_3)$, then

$$\|F_t - F_{t+h}\| \leq I_a + I_b + \Pi \leq 2\varepsilon + 4\varepsilon + 3\varepsilon = 9\varepsilon.$$

This proves the assertion for a single trajectory with $\mathcal{Y} = \mathcal{L}_T^{(p)}$. An inspection of the proof will show that if $F \in \mathcal{H}$, \mathcal{H} a compact subset of $\mathcal{F}(\mathcal{U}, \mathcal{Y})$, then the compact sets chosen above can each be replaced by compact sets chosen independently of F in \mathcal{H} . For example, the compact set $F(R_t \mathcal{U}_T)$ is replaced by $\mathcal{H}(R_t \mathcal{U}_T)$, which is a compact subset of \mathcal{Y} since \mathcal{H} restricted to $R_t \mathcal{U}_T$ is a compact set of bounded continuous mappings, and $R_t \mathcal{U}_T$ is a compact subset of L_q . Also the mappings $(P_{2T+t} - P_{t-T})FR_t$ restricted to \mathcal{H} , $F \in \mathcal{H}$, are equicontinuous by Ascoli's theorem. These facts yield the assertion that the F_t are equicontinuous.

The proof of the assertions for the case $\mathcal{Y} = \mathcal{B}_c$ is similar, although obviously some modifications are required. The details are not given. \square

Two consistency relations are introduced for the trajectories F_t . The second of these will also be used as an interpolation formula. Conditions under which they hold are given in the proposition to follow.

$$\begin{aligned} (8) \quad & (P_{T-\eta} - P_0)L_\eta F_t R_\eta (P_{T-\eta} - P_{-d}) \\ & = (P_{T-\eta} - P_0)F_{t+\eta}(P_{T-\eta} - P_{-d}), \quad 0 \leq \eta \leq T; \end{aligned}$$

$$\begin{aligned} (9) \quad & F_{t+\eta} = (P_{T-\eta} - P_0)L_\eta F_t R_\eta (P_{T-\eta} - P_{-d}) \\ & + (P_T - P_{T-\eta})L_{\eta-T}F_{t+T}R_{\eta-T}(P_T - P_{T-\eta-d}), \quad 0 \leq \eta \leq T. \end{aligned}$$

PROPOSITION II.13. (i) If $F_t = \pi_t F$, $F \in \mathcal{F}(\mathcal{U}, \mathcal{Y})$, then F_t satisfies (8) for all t . (ii) If $F \in \mathcal{F}_d^\circ(\mathcal{U}, \mathcal{Y})$, then F_t satisfies (9) for all t . (iii) If $H_t, H_{t+T}, H_{t+\eta}$ are any mappings from \mathcal{U}_T into \mathcal{Y}_T that satisfy (9), then they satisfy (8).

Proof. The proof of (i) is given by the calculation

$$\begin{aligned} & (P_{T-\eta} - P_0)L_\eta[(P_T - P_0)L_t F R_t (P_T - P_{-d})]R_\eta(P_{T-\eta} - P_{-d}) \\ & = (P_{T-\eta} - P_0)(P_{T-\eta} - P_{-\eta})L_{\eta+t}F R_{\eta+t}(P_{T-\eta} - P_{-d-\eta})(P_{T-\eta} - P_{-d}) \\ & = (P_{T-\eta} - P_0)L_{\eta+t}F R_{\eta+t}(P_{T-\eta} - P_{-d}) \\ & = (P_{T-\eta} - P_0)[(P_T - P_0)L_{t+\eta}F R_{t+\eta}(P_T - P_{-d})](P_{T-\eta} - P_{-d}) \\ & = (P_{T-\eta} - P_0)F_{t+\eta}(P_{T-\eta} - P_{-d}). \end{aligned}$$

To prove (ii) we use (i) for the first term on the right side of equation (9) and make an analogous calculation for the second term. Then the right-hand side of (9) becomes

$$(10) \quad (P_{T-\eta} - P_0)F_{t+\eta}(P_{T-\eta} - P_{-d}) + (P_T - P_{T-\eta})F_{t+\eta}(P_T - P_{T-\eta-d}).$$

If F is causal with bounded memory (d), then so are all the F_t , and this expression reduces to

$$(P_{T-\eta} - P_0)F_{t+\eta} + (P_T - P_{T-\eta})F_{t+\eta} = F_{t+\eta},$$

which proves (ii). Part (iii) can be verified immediately by substituting $F_{t+\eta}$ from (9) into the right-hand side of (8). \square

The consistency condition (9), if required to hold for all t and all η , $0 \leq \eta \leq T$, is not quite enough to guarantee that F is causal with bounded memory (d). It does guarantee something a little weaker, and to state this we use the definition: F is weakly causal and of bounded memory (d) if for every $A > 0$, and for all t ,

$$(11) \quad (P_{t+A} - P_t)F(P_{t+A} - P_{t-d}) = (P_{t+A} - P_t)F(P_a - P_{t-d})$$

whenever $t + A \leq a$, and

$$(12) \quad (P_{t+A} - P_t)F(P_{t+A} - P_{t-d}) = (P_{t+T} - P_t)F(P_{t+A} - P_b)$$

whenever $b \leq t - d$.

This definition rules out noncausality and non-bounded-memory (d) that depend on interactions between past and future.

PROPOSITION II.14. *If $F \in \mathcal{F}(\mathcal{U}, \mathcal{Y})$, then (9) is satisfied for all $T > 0$, all t , and all η , $0 \leq \eta \leq T$, if and only if F is weakly causal and of bounded memory (d).*

Proof. The right-hand side of (9) is given in different form in (10); consider the first term of (10). Since (9) is satisfied, we must have that

$$(P_{T-\eta} - P_0)F_{t+\eta}(P_{T-\eta} - P_{-d}) = (P_{T-\eta} - P_0)F_{t+\eta}(P_T - P_{-d}).$$

This may be rewritten

$$\begin{aligned} L_{t+\eta}(P_{T+t} - P_{t+\eta})F(P_{T+t} - P_{t+\eta-d})R_{t+\eta} \\ = L_{t+\eta}(P_{T+t} - P_{t+\eta})F(P_{T+t+\eta} - P_{t+\eta-d})R_{t+\eta}, \end{aligned}$$

which is equivalent to

$$(P_{T+t} - P_{t+\eta})F(P_{T+t} - P_{t+\eta-d}) = (P_{T+t} - P_{t+\eta})F(P_{T+t+\eta} - P_{t+\eta-d}).$$

Put $a = T + t + \eta$, $s = t + \eta$ and $A = T - \eta$. Then this is in the form of condition (11) for weak causality and bounded memory (d), and a , s and A can be given arbitrarily by choosing $\eta > 0$, t and $T > \eta$. An analogous argument applied to the second term of (10) yields condition (12). The converse follows immediately from equation (10) and the definition of $H_{t+\eta}$. \square

From a family of mappings carrying \mathcal{U}_T into \mathcal{Y}_T it is possible under certain circumstances to synthesize a mapping from \mathcal{U} into \mathcal{Y} . We want to be able to do this, because we want to be able to go from trajectories $\{F_t\}$ back to an overall system mapping F . The transformations ρ_s to be defined below accomplish this. If π denotes the transformation carrying F into a trajectory $\{F_t\}$, then the ρ_s are roughly inverse to π . However the situation is a little complicated in general, and ρ_s and π are inverse to each other only when F is causal with bounded memory of sufficiently short duration. These comments are made precise in what follows.

We use the notations

$$\begin{aligned}\Delta_{n,t} &= (P_{t-(n-1)T} - P_{t-nT}), \\ \Delta'_{n,t} &= (P_{t-(n-1)T} - P_{t-nT-d}),\end{aligned}$$

where $T > 0$, $d > 0$ are fixed. Let \mathcal{G} be a bounded subset of $\mathcal{F}(\mathcal{U}_T, \mathcal{Y}_T)$ with the additional property that the $G \in \mathcal{G}$ are equicontinuous. Let $\{G_n\}$ be a sequence from \mathcal{G} . For any real number t , define

$$(13) \quad G^t = \sum_{-\infty}^{\infty} \Delta_{n,t} R_{t-nT} G_n L_{t-nT} \Delta'_{n,t}.$$

It is clear that G^t is a mapping from \mathcal{U} into \mathcal{Y} if $\mathcal{Y} = \mathcal{L}_T^{(p)}$ or \mathcal{B} ; however, see Appendix A for a formal justification of the infinite sum.

PROPOSITION II.15. *G^t as defined by (13) is an element of $\mathcal{F}(\mathcal{U}, \mathcal{Y})$, where \mathcal{Y} is either $\mathcal{L}_T^{(p)}$ or \mathcal{B} .*

Proof. Take $\mathcal{Y} = \mathcal{L}_T^{(p)}$. Given $\varepsilon > 0$, let $\delta > 0$ be such that $\|z_1 - z_2\| \leq \delta$, $z_1, z_2 \in \mathcal{U}_T$, implies $\|G_n(z_1) - G_n(z_2)\| \leq \varepsilon$ for all $n = 1, 2, \dots$, as is possible from the hypothesis on \mathcal{G} . For $u_1, u_2 \in \mathcal{U}$, $\|u_1 - u_2\| \leq \delta$, one has

$$\begin{aligned}\|G^t(u_1) - G^t(u_2)\| &= \sup_s \|(P_{s+T} - P_s) [G^t(u_1) - G^t(u_2)]\|_p \\ &\leq \sup_k \|\Delta_{k,t} R_{t-kT} G_k L_{t-kT} \Delta'_{k,t}(u_1) \\ &\quad + \Delta_{k+1,t} R_{t-(k+1)T} G_{k+1} L_{t-(k+1)T} \Delta'_{k+1,t}(u_1) \\ &\quad - \Delta_{k,t} R_{t-kT} G_k L_{t-kT} \Delta'_{k,t}(u_2) \\ &\quad - \Delta_{k+1,t} R_{t-(k+1)T} G_{k+1} L_{t-(k+1)T} \Delta'_{k+1,t}(u_2)\|_p \\ &\leq 2\|\Delta_{n,t} R_{t-nT} G_n L_{t-nT} \Delta'_{n,t}(u_1) \\ &\quad - \Delta_{n,t} R_{t-nT} G_n L_{t-nT} \Delta'_{n,t}(u_2)\|_p + \varepsilon \\ &= 2\|(P_T - P_0)[G_n L_{t-nT} \Delta'_{n,t}(u_1) - G_n L_{t-nT} \Delta'_{n,t}(u_2)]\|_p + \varepsilon\end{aligned}$$

for some n . But

$$\|L_{t-nT} \Delta'_{n,t} u_1 - L_{t-nT} \Delta'_{n,t} u_2\| \leq \|u_1 - u_2\| \leq \delta,$$

and $L_{t-nT} \Delta'_{n,t} u \in \mathcal{U}_T$, hence

$$\|G^t(u_1) - G^t(u_2)\| \leq 2\varepsilon + \varepsilon = 3\varepsilon.$$

The boundedness of G^t follows similarly. The same proof holds for $\mathcal{Y} = \mathcal{B}$ if the L_p -norms are changed to uniform norms. \square

The transformation that carries the sequence $\{G_n\}$ of equicontinuous mappings into G^t is denoted by ρ_t . Note that the equicontinuity condition is natural, since when we go the other way we have that the $\pi_t F$, $F \in \mathcal{F}(\mathcal{U}, \mathcal{Y})$, are equicontinuous with respect to t .

PROPOSITION II.16. *The transformation ρ_t is continuous in the following sense: if there are two sequences of equicontinuous mappings from \mathcal{U}_T to \mathcal{Y}_T , $\{G_n\}$ and $\{\tilde{G}_n\}$, and $\|G_n - \tilde{G}_n\| \leq \delta$ for all integers n for some $\delta = \delta(\varepsilon)$, then $\|\rho_t(\{G_n\}) - \rho_t(\{\tilde{G}_n\})\| \leq \varepsilon$.*

Proof. Again take $\mathcal{Y} = \mathcal{L}_T^{(p)}$. Then

$$\begin{aligned} & \|\rho_t(\{G_n\}) - \rho_t(\{\tilde{G}_n\})\| \\ & \leq 2 \sup_{\mathcal{U}} \|\Delta_{n,t} R_{t-nT} [G_n L_{t-nT} \Delta'_{n,t}(u) - \tilde{G}_n L_{t-nT} \Delta'_{n,t}(u)]\|_p + \varepsilon \end{aligned}$$

for some n by a calculation very similar to that in the previous proof. But the right side of this inequality can be rewritten as

$$2 \sup_{\mathcal{U}} \|(G_n - \tilde{G}_n)(L_{t-nT} \Delta'_{n,t} u)\|_p + \varepsilon = 2 \sup_{z \in \mathcal{U}_T} \|(G_n - \tilde{G}_n)z\|_p + \varepsilon$$

since, by the shift invariance of \mathcal{U} , any $z \in \mathcal{U}_T$ can be obtained by truncating some u by $(P_{t-(n-1)T} - P_{t-nT-\delta})$ for arbitrary t, n . Hence, if $\|G_n - \tilde{G}_n\|$ is sufficiently small for all n ,

$$\|\rho_t(\{G_n\}) - \rho_t(\{\tilde{G}_n\})\| \leq 2\varepsilon + \varepsilon = 3\varepsilon.$$

Again, the same proof holds for $\mathcal{Y} = \mathcal{B}$ if the L_p -norms are changed to sup norms. \square

PROPOSITION II.17. *If $F \in \mathcal{F}_d^{\circ}(\mathcal{U}, \mathcal{Y})$, then for any t , $\{\pi_{t-nT}F\}$ is a family of equicontinuous causal mappings from \mathcal{U}_T to \mathcal{Y}_T with bounded memory (d), and $F = \rho_t(\{\pi_{t-nT}F\})$. Conversely, if $\{G_n\}$ is a sequence of equicontinuous causal mappings from \mathcal{U}_T to \mathcal{Y}_T with bounded memory (d), then $\rho_t(\{G_n\}) \in \mathcal{F}_d^{\circ}(\mathcal{U}, \mathcal{Y})$ and $G_k = \pi_{t-kT} \circ \rho_t(\{G_n\})$.*

Proof. The assertion that the $\pi_{t-nT}F$ are causal with bounded memory (d) is obvious; indeed all the $\pi_s F$ are causal with bounded memory (d). Further,

$$\begin{aligned} \rho_t(\{\pi_{t-nT}F\}) &= \sum_{-\infty}^{\infty} \Delta_{n,t} R_{t-nT} [(P_T - P_0) L_{t-nT} F R_{t-nT} (P_T - P_{-d})] L_{t-nT} \Delta'_{n,t} \\ &= \sum_{-\infty}^{\infty} \Delta_{n,t} F \Delta'_{n,t} = F. \end{aligned}$$

It is also obvious that $\rho_t(\{G_n\})$ is causal with bounded memory (d). The second inversion identity is given by the calculation

$$\begin{aligned} \pi_{t-kT} \circ \rho_t(\{G_n\}) &= (P_T - P_0) L_{t-kT} \left[\sum_{-\infty}^{\infty} \Delta_{n,t} R_{t-nT} G_n L_{t-nT} \Delta'_{n,t} \right] \cdot R_{t-kT} (P_T - P_{-d}) \\ &= L_{t-kT} \Delta_{k,t} R_{t-kT} G_k L_{t-kT} \Delta'_{k,t} R_{t-kT} (P_T - P_{-d}) \\ &= (P_T - P_0) G_k (P_T - P_{-d}) = G_k. \quad \square \end{aligned}$$

When F is not causal with bounded memory, the operations π and ρ_t obviously cannot be inverse to each other because some information about F is lost in the truncations given by the π_t which cannot be restored. The sense in which they are approximately inverse to each other is given in the next proposition.

PROPOSITION II.18. *Let $\{F_t\}$, $-\infty < t < \infty$, be a family of mappings in $\mathcal{F}(\mathcal{U}_T, \mathcal{Y}_T)$ which is bounded and in which the F_k are equicontinuous. For any fixed s consider $\{F_{s-nT}\}$, $n = \dots, -2, -1, 0, 1, 2, \dots$. Put*

$$\begin{aligned} H_{s-nT+\eta} &\stackrel{\text{def}}{=} (P_{T-\eta} - P_0) L_{\eta} F_{s-nT} R_{\eta} (P_{T-\eta} - P_{-d}) \\ (14) \quad &+ (P_T - P_{T-\eta}) L_{\eta-T} F_{s-(n-1)T} R_{\eta-T} (P_T - P_{T-\eta-D}), \quad 0 \leq \eta < T. \end{aligned}$$

This defines H_t for all t , and $H_{s-nT} = F_{s-nT}$. Further, define

$$\begin{aligned} H^{(1)} &\stackrel{\text{def}}{=} \rho_s(\{H_{s-nT}\}), \\ H_t^{(1)} &\stackrel{\text{def}}{=} \pi_t \circ \rho_s(\{H_{s-nT}\}) = \pi_t H^{(1)}, \\ H^{(2)} &\stackrel{\text{def}}{=} \rho_s(\{\pi_{s-nT} H^{(1)}\}) = \rho_s(\{H_{s-nT}^{(1)}\}). \end{aligned}$$

Then,

- (i) $H_t^{(1)} = H_t$ for all t ;
- (ii) $H^{(2)} = H^{(1)}$;
- (iii) if F_t satisfies (9), then $H_t = F_t$ and $H_t^{(1)} = F_t$.

Proof. By the definitions, $H_t^{(1)}$ is given by

$$H_t^{(1)} = (P_T - P_0)L_t \left[\sum_{-\infty}^{\infty} \Delta_{n,s} R_{s-nT} H_{s-nT} L_{s-nT} \Delta'_{n,s} \right] R_t (P_T - P_{-d}).$$

At most two terms from the infinite sum can contribute anything, by virtue of the projection $(P_T - P_0)$. Let k be that integer such that

$$s - kT \leq t < s - (k-1)T,$$

and let

$$\eta = t - (s - kT).$$

Then, since $F_{s-nT} = H_{s-nT}$, the expression for $H_t^{(1)}$ above reduces to the expression for $H_t = H_{s-nT+\eta}$ given by (14). Thus $H_t^{(1)} = H_t$ for all t , and $H^{(2)} = H^{(1)}$ follows from this equality and from the definitions. The assertion (iii) is obvious, since (14) becomes (9) if H is replaced by F . \square

We conclude this section with a simple error bound on the interpolated $H_{t+\eta}$ as given by (9) when H_t and H_{t+T} are in error.

PROPOSITION II.19. *Let $\|\tilde{F}_t - F_t\| \leq \varepsilon$ and $\|\tilde{F}_{t+T} - F_{t+T}\| \leq \varepsilon$. Then, if $\tilde{F}_{t+\eta}$ and $F_{t+\eta}$ are each given by (9) in terms of $\tilde{F}_t, \tilde{F}_{t+T}$ and F_t, F_{t+T} , respectively, $\|\tilde{F}_{t+\eta} - F_{t+\eta}\| \leq 2\varepsilon$.*

Proof. Expressing $F_{t+\eta}, \tilde{F}_{t+\eta}$ in terms of F_t, F_{t+T} and $\tilde{F}_t, \tilde{F}_{t+T}$ from (9) yields

$$\begin{aligned} \|\tilde{F}_{t+\eta} - F_{t+\eta}\| &\leq \sup_{\mathcal{U}_T} \|\tilde{F}_t R_\eta (P_{T-\eta} - P_{-d})u - F_t R_\eta (P_{T-\eta} - P_{-d})u\| \\ (15) \quad &+ \sup_{\mathcal{U}_T} \|\tilde{F}_{t+T} R_{\eta-T} (P_T - P_{T-\eta-d})u - F_{t+T} R_{\eta-T} (P_T - P_{T-\eta-d})u\|. \end{aligned}$$

Now, by the properties of \mathcal{U} ,

$$\bigcup_{0 \leq \eta \leq T} R_\eta (P_{T-\eta} - P_{-d}) \mathcal{U}_T \subset \mathcal{U}_T.$$

Hence,

$$\sup_{\mathcal{U}_T} \|\tilde{F}_t R_\eta (P_{T-\eta} - P_{-d})u - F_t R_\eta (P_{T-\eta} - P_{-d})u\| \leq \sup_{\mathcal{U}_T} \|\tilde{F}_t u - F_t u\| \leq \varepsilon.$$

The second term in the inequality (15) is also dominated by ε , by essentially the same argument. \square

4. Trajectories of the finite-time projections for classes of systems. We now consider a class of bounded systems $\mathcal{S} = (\mathcal{Y}, f, \mathcal{X}, \mathcal{U})$ in its natural representation form $\mathcal{S}_0 = (\mathcal{Y}, g, \mathcal{H}, \mathcal{U})$, where \mathcal{U} is a shift-invariant, T -compact subset of $\mathcal{L}_{T+d}^{(p)}$ with property (P), and where \mathcal{Y} is $\mathcal{L}_T^{(p)}$ or \mathcal{B} . \mathcal{H} is, of course, a subset of $\mathcal{F}(\mathcal{U}, \mathcal{Y})$; further hypotheses on \mathcal{H} will be made as needed. Each $F \in \mathcal{H}$ will generate a trajectory $\{F_t\} \in \mathcal{H}_T$, whether F is causal with bounded memory less than or equal to d , or not. If $\mathcal{H} \subset \mathcal{F}_d^0(\mathcal{U}, \mathcal{Y})$, then each of these trajectories will yield the corresponding F through the mapping ρ . We investigate some basic properties of these families of trajectories.

Temporarily take $T > 0$ to be fixed. Let \mathcal{M} be the closed linear subspace of the Banach space $\mathcal{F}(\mathcal{U}, \mathcal{Y})$ generated by \mathcal{H} , and let $\mathcal{M}_t = \pi_t \mathcal{M}$. \mathcal{M}_t is a linear subset of $\mathcal{F}(\mathcal{U}_T, \mathcal{Y}_T)$; its closure, $\overline{\mathcal{M}}_t$, is the closed linear subspace of $\mathcal{F}(\mathcal{U}_T, \mathcal{Y}_T)$ generated by $\pi_t \mathcal{H}$. We define \mathcal{S}_0 (or \mathcal{S}) to be a *linearly predictable class of systems with respect to T* if each mapping π_t is one-to-one from \mathcal{M} onto \mathcal{M}_t , $t \in \mathbb{R}^1$. When \mathcal{S}_0 is a linearly predictable class a prediction mapping $\theta(t, s)$ carrying H_t into H_s , $t \leq s$, can be defined by

$$\theta(t, s) = \pi_s \circ \pi_t^{-1}, \quad -\infty < t, s < \infty.$$

For each t, s , $\theta(t, s)$ is obviously a linear transformation with domain \mathcal{M}_t and range \mathcal{M}_s .

The intuitive meaning of \mathcal{S}_0 being a linearly predictable class is that no two trajectories associated with the $F \in \mathcal{M}$ corresponding to \mathcal{S}_0 can cross or touch and be at the common point at the same time. Two trajectories can cross or touch provided the time of arrival at the common point is different for the two. A class of systems consisting of a single system (\mathcal{H} has only one element) is always predictable in the sense of this definition.

We further define a *stationarily predictable class of systems with respect to T* to be a class \mathcal{S}_0 with the property that whenever $\pi_t F = \pi_s G$, F and $G \in \mathcal{M}$, then $\pi_{t+a} F = \pi_{s+a} G$ for all real numbers a . Intuitively, this implies that the systems F and G have trajectories which as geometrical entities are identical. Furthermore, no individual trajectory can cross itself. If the definition is weakened to read: $\pi_t F = \pi_s G$, F and $G \in \mathcal{M}$, implies $\pi_{t+a} F = \pi_{s+a} G$ for all $a \geq 0$, we call the class \mathcal{S}_0 a *future-time (f.t.) stationarily predictable class with respect to T* .

If either of F or G is not causal with bounded memory (d), it is obviously possible that $\pi_a F = \pi_a G$ for all a without F and G being the same. In this case, \mathcal{S}_0 can be stationarily predictable without being linearly predictable. A fortiori, \mathcal{S}_0 can be f.t. stationarily predictable without being linearly predictable. However, if the \mathcal{H} associated with \mathcal{S}_0 is a subset of $\mathcal{F}_d^0(\mathcal{U}, \mathcal{Y})$, so is \mathcal{M} . Then if for some t , $\pi_t F = \pi_t G$, it follows from stationary predictability that $\pi_a F = \pi_a G$ for all a , and hence by Proposition II.17 that $F = G$. Thus, in this situation stationary predictability implies linear predictability. Under the same condition that $\mathcal{H} \subset \mathcal{F}_d^0(\mathcal{U}, \mathcal{Y})$, if \mathcal{S}_0 is only f.t. stationarily predictable, the situation is complicated a little, but can be interpreted in much the same way as will be seen below.

In case \mathcal{S}_0 is linearly and stationarily predictable the prediction mapping $\theta(t, s)$ can be written as a function of the difference $s - t$ only, once the domain has been defined properly. In fact, suppose to start with that $F' \in \mathcal{M}_t$ and also $F' \in \mathcal{M}_{t+a}$. Then $F' = \pi_t F$ for some $F \in \mathcal{M}$, and also $F' \in \pi_{t+a} G$ for some $G \in \mathcal{M}$.

Thus,

$$\theta(t, s)F' = \pi_s \circ \pi_t^{-1}(\pi_t F) = \pi_s F$$

and

$$\theta(t + a, s + a)F' = \pi_{s+a} \circ \pi_{t+a}^{-1}(\pi_{t+a} G) = \pi_{s+a} G.$$

By the definition of a stationarily predictable class, $\pi_s F = \pi_{s+a} G$; hence $\theta(t, s)F' = \theta(t + a, s + a)F'$. Now (with a slight abuse of notation) let $\theta(\tau)F' = \theta(t, s)F'$, $s = t + \tau$, for all F' such that for some t , $F' \in \mathcal{M}_t$.

This definition is meaningful, because if more than one pair (t, s) satisfy the conditions they all yield the same $\theta(t, s)F'$. The domain of $\theta(\tau)$, for any τ , will now include $\bigcup_{t \in \mathbb{R}^1} \mathcal{M}_t$; extend this by linearity to $\mathcal{N} = \text{linear span } \{\bigcup_{t \in \mathbb{R}^1} \mathcal{M}_t\}$. The family $\{\theta(\tau)\}$, $\tau \in \mathbb{R}^1$, is now a one-parameter group of linear transformations on \mathcal{N} . We note that $\mathcal{N} \subset \mathcal{F}(\mathcal{U}_T, \mathcal{Y}_T)$. In fact, the elements of \mathcal{N} are of the form

$$\begin{aligned} F' &= \sum_{n=1}^N \alpha_n (P_T - P_0) L_{t_n} F_n P_{t_n} (P_T - P_{-d}) \\ &= (P_T - P_0) \left(\sum_{n=1}^N \alpha_n L_{t_n} F_n R_{t_n} \right) (P_T - P_{-d}), \end{aligned}$$

where $\{t_1, \dots, t_N\}$ is an arbitrary finite set of real numbers, as is also $\{\alpha_1, \dots, \alpha_N\}$, and F_1, \dots, F_N are each elements of $\mathcal{F}(\mathcal{U}, \mathcal{Y})$. Since $\sum_{n=1}^N \alpha_n L_{t_n} F_n R_{t_n}$ is also a bounded, continuous mapping, we can denote it by $F \in \mathcal{F}(\mathcal{U}, \mathcal{Y})$. Then

$$F' = (P_T - P_0)F(P_T - P_{-d}) \in \mathcal{F}(\mathcal{U}_T, \mathcal{Y}_T).$$

In case \mathcal{S}_0 is a linearly and f.t. stationarily predictable class we can, similarly, for any $\tau \geq 0$, put $\theta(\tau)F_1 = \theta(t, s)F_1$ for all F_1 such that for some pair (t, s) with $t \geq 0$, $s - t = \tau$, it holds that $F_1 \in \mathcal{M}_t$. The domain of $\theta(\tau)$, $\tau \geq 0$, can now be extended by linearity to $\mathcal{N}_+ = \text{linear span } \{\bigcup_{t \geq 0} \mathcal{M}_t\}$. The family $\{\theta(\tau)\}$, $\tau \geq 0$, is now a one-parameter semigroup of linear transformations on \mathcal{N}_+ , which is also contained in $\mathcal{F}(\mathcal{U}_T, \mathcal{Y}_T)$.

If \mathcal{S}_0 is f.t. stationarily (but not linearly) predictable and $\mathcal{H} \subset \mathcal{F}_d^\circ(\mathcal{U}, \mathcal{Y})$, a semigroup can be established in essentially the same way. Suppose $F' = \pi_t F = \pi_t G$. Then the fact that $\pi_{t+\tau} F = \pi_{t+\tau} G$, $\tau \geq 0$, implies that F and G restricted as desired are the same mapping. We now redefine π_t^{-1} as the set function: $\pi_t^{-1}(F') = \{F: \pi_t F = F'\}$. Then $\theta(t, s)$ can again be defined as $\pi_s \circ \pi_t^{-1}$, but only, of course, for $t \leq s$. $\theta(t, s)$ is again linear on \mathcal{M}_t , and the development that follows for the semigroup case can be repeated exactly. In what follows we restrict attention to the semigroups of linear transformations, as being of more immediate interest than groups in modeling for system identification.

The usual linear operator norm, when it exists, of the linear transformation $\theta(\tau)$ is given by

$$|\theta(\tau)| = \sup_{F' \in \mathcal{N}_+} \left\{ \frac{1}{\|F'\|_{\mathcal{Y}_T}} \|\theta(\tau)F'\|_{\mathcal{Y}_T} \right\},$$

where the symbol $|\cdot|$ has been used to provide a reminder that this is a different kind of norm than has been used for the other mappings that have appeared.

From the definition of \mathcal{N}_+ it follows that $\theta(\tau)$ is a bounded operator if and only if there is a number $B > 0$ such that

$$(16) \quad \sup_{u \in \mathcal{U}} \left\| (P_T - P_0) \left(\sum_{n=1}^N \alpha_n L_{t_n + \tau} F_n R_{t_n + \tau} \right) (P_T - P_{-d}) u \right\|_{\mathcal{Y}_T} \\ \leq B \sup_{u \in \mathcal{U}} \left\| (P_T - P_0) \left(\sum_{n=1}^N \alpha_n L_{t_n} F_n R_{t_n} \right) (P_T - P_{-d}) u \right\|_{\mathcal{Y}_T}$$

for any positive integer N , any set of points t_1, \dots, t_N all greater than or equal to zero, any set of scalars $\alpha_1, \dots, \alpha_N$ and any F_1, \dots, F_N belonging to \mathcal{M} .

This is a regularity condition on the time behavior of the mappings F . Note that, unfortunately, it is not sufficient to consider just those $F \in \mathcal{H}$, but rather all finite linear combinations of these and of their translations. If \mathcal{H} is itself a subset of $\mathcal{F}(\mathcal{U}, \mathcal{Y})$ that is invariant under time shift, then all the \mathcal{M}_t are the same and the sums in condition (16) collapse to single terms.

Using the definitions established, we can now state a basic fact, which is really a corollary to Proposition II.12.

PROPOSITION II.20. *Let \mathcal{S}_0 be such that $\mathcal{H} \subset \mathcal{F}_d^\circ(\mathcal{U}, \mathcal{Y})$, let it be f.t. stationarily predictable with respect to T , and let the $\theta(\tau)$, $\tau \geq 0$, be bounded operators. Then $\{\theta(\tau)\}$, $\tau \geq 0$, is a strongly continuous semigroup of bounded linear operators on the Banach space $\bar{\mathcal{N}}_+$, the closure of \mathcal{N}_+ in $\mathcal{F}(\mathcal{U}_T, \mathcal{Y}_T)$.*

Proof. It is supposed of course that the $\theta(\tau)$ are extended by continuity to $\bar{\mathcal{N}}_+$. All that has to be shown is that $\|\theta(\tau)F' - F'\| \rightarrow 0$ as $\tau \rightarrow 0$, for any $F' \in \bar{\mathcal{N}}_+$. Since $\bar{\mathcal{N}}_+ \subset \mathcal{F}(\mathcal{U}_T, \mathcal{Y}_T)$, and since any $F' \in \mathcal{F}(\mathcal{U}_T, \mathcal{Y}_T)$ can be written $F' = (P_T - P_0)F'(P_T - P_{-d})$, it follows that F' is the image under π_0 of itself, regarded as an element of $\mathcal{F}(\mathcal{U}, \mathcal{Y})$. We write $F' \times \pi_0 F' \times F_0$. Then

$$\|\theta(\tau)F' \times F' - \|F_\tau - F_0\| \rightarrow 0 \quad \text{as } \tau \rightarrow 0$$

by Proposition II.12. \square

Clearly the hypothesis that $\mathcal{H} \subset \mathcal{F}_d^\circ(\mathcal{U}, \mathcal{Y})$ can be replaced by the hypothesis that \mathcal{S}_0 is linearly predictable, and then with the other hypotheses in force the conclusion still follows. For convenience we shall refer to an \mathcal{S}_0 that satisfies either the conditions of Proposition II.20 or the modified conditions just given as a linear dynamical class of systems with respect to T . This terminology is introduced with some apology since dynamical is such a widely used term; however, it seems reasonably appropriate. There is no inference, of course, that the individual systems in the linear dynamical class are linear.

Thus far, T , the length of the interval of observation of the output, has remained fixed. We now look at how the properties of the special classes of systems introduced in this section are affected by changes in T . When T is changed, so is the norm on the input space, which is always assumed to be a subset of $\mathcal{L}_{T+d}^{(p)}$. In fact, $\|u\|_T^{(p)} \leq \|u\|_{T'}^{(p)}$ when $T' \leq T$. However, as has been pointed out earlier, the membership of \mathcal{U} does not depend on the value of T , nor does the topology on \mathcal{U} , nor do the properties of T -compactness and shift-invariance. Similar statements can be made for \mathcal{Y} if $\mathcal{Y} = \mathcal{L}_T^{(p)}$. If $\mathcal{Y} = \mathcal{B}$, then not even the norm on \mathcal{Y} is changed. In any event, the class of mappings $\mathcal{F}(\mathcal{U}, \mathcal{Y})$ is not affected.

PROPOSITION II.21. *If \mathcal{S}_0 is a linearly predictable class of systems with respect to T' , then it is also linearly predictable with respect to any $T \geq T'$.*

Proof. Suppose the linear mapping $\pi_t(T)$ given by

$$\pi_t(T)F = (P_T - P_0)L_tFR_t(P_T - P_{-d})$$

is singular. Then for some $F \neq 0$, $\pi_t(T)F = 0$; and for $T' \leq T$,

$$(P_{T'} - P_0)[(P_T - P_0)L_tFR_t(P_T - P_{-d})u] = 0$$

for all $u \in \mathcal{U}$. Since $(P_{T'} - P_{-d})u \in \mathcal{U}$ for all $u \in \mathcal{U}$,

$$(P_{T'} - P_0)L_tFR_t(P_{T'} - P_{-d})u = 0$$

for all $u \in \mathcal{U}$. Hence $\pi_t(T')$ is singular, and the assertion is proved by contradiction. \square

PROPOSITION II.22. *If \mathcal{S}_0 is a class of systems with the property that $\mathcal{H} \subset \mathcal{F}_d^\circ(\mathcal{U}, \mathcal{Y})$, and if \mathcal{S}_0 is stationarily predictable with respect to T' , then it is stationarily predictable with respect to any $T \geq T'$. Stationary predictability can be replaced simultaneously in hypothesis and conclusion by future-time stationary predictability.*

Proof. Suppose to start with that $T' \leq T \leq 2T'$. We need to show that the condition $\pi_t(T)F = \pi_s(T)G$, where $F, G \in \mathcal{M} \subset \mathcal{F}_d^\circ(\mathcal{U}, \mathcal{Y})$ implies that $\pi_{t+a}(T)F = \pi_{s+a}(T)G$ for all a . We note that the condition $\pi_t(T)F = \pi_s(T)G$ can be written

$$(P_T - P_0)(L_tFR_t - L_sGR_s)(P_T - P_{-d})u = 0$$

for all $u \in \mathcal{U}$. Since $(P_{T'} - P_{-d})u \in \mathcal{U}$ for all $u \in \mathcal{U}$, it follows that

$$(P_{T'} - P_0)(L_tFR_t - L_sGR_s)(P_{T'} - P_{-d})u = 0$$

for all $u \in \mathcal{U}$; i.e., $\pi_t(T')F = \pi_s(T')G$. By hypothesis, it follows that $\pi_{t+a}(T')F = \pi_{s+a}(T')G$, or

$$(17) \quad (P_{T'} - P_0)L_a(L_tFR_t - L_sGR_s)R_a(P_{T'} - P_{-d})u = 0$$

for all $u \in \mathcal{U}$, and any real number a .

Now

$$\begin{aligned} & (P_{2T'} - P_{T'})L_a(L_tFR_t - L_sGR_s)R_a(P_{2T'} - P_{T'-d})u \\ &= L_{-T'}(P_{T'} - P_0)L_{a+T'}(L_tFR_t - L_sGR_s)R_{a+T'}(P_{T'} - P_{-d})R_{-T'}u = 0 \end{aligned}$$

for all $u \in \mathcal{U}$, since $R_{-T'}(u) \in \mathcal{U}$ for all $u \in \mathcal{U}$, and we can replace the a of (15) by $a + T'$. Since the mappings F and G are causal with bounded memory (d),

$$\begin{aligned} & (P_{2T'} - P_0)L_a(L_tFR_t - L_sGR_s)R_a(P_{2T'} - P_{-d})u \\ &= (P_{2T'} - P_{T'})L_a(L_tFR_t - L_sGR_s)R_a(P_{2T'} - P_{T'-d})u \\ &+ (P_{T'} - P_0)L_a(L_tFR_t - L_sGR_s)R_a(P_{T'} - P_{-d})u, \end{aligned}$$

which equals zero by the calculations above. It follows then by a now familiar argument that

$$(P_T - P_0)L_a(L_tFR_t - L_sGR_s)R_a(P_T - P_{-d})u = 0$$

for all $u \in \mathcal{U}$, which is what needs to be shown. The extension to arbitrary $T \geq T'$ follows by induction. The proof for future-time stationary predictability is the same with a restricted to be ≥ 0 . \square

PROPOSITION II.23. *If \mathcal{S}_0 is a linear dynamical class of systems with respect to T' , and if it further has the property that $\mathcal{H} \subset \mathcal{F}_d^0(\mathcal{U}, \mathcal{Y})$, then \mathcal{S}_0 is a dynamical class with respect to any $T \geq T'$.*

Proof. In view of the preceding proposition, all that needs to be proved is that if $\theta_{T'}(\tau)$ is a bounded operator for all $\tau \geq 0$, then $\theta_T(\tau)$ is a bounded operator for all $\tau \geq 0$ whenever $T \geq T'$. The meaning of the subscripts T and T' on $\theta(\tau)$ is obvious. In what follows it is necessary to go back and forth between norms in \mathcal{Y}_T and in $\mathcal{Y}_{T'}$, so a subscript T or T' is used. The facts that, by an obvious identification of elements, $\mathcal{Y}_{T'}$ can be thought of as a subset of \mathcal{Y}_T , $T' \leq T$, and that then $\|y\|_{T'} = \|y\|_T$ when $y \in \mathcal{Y}_{T'}$ are used without comment.

Again assume to start with that $T \leq 2T'$. We have

$$\|\theta_T(\tau)F\|_T = \sup_{\mathcal{U}} \left\| (P_T - P_0) \left(\sum_{n=1}^N \alpha_n L_{t_n+\tau} F_n R_{t_n+\tau} \right) \cdot (P_T - P_{-d})u \right\|_T, \\ \tau \geq 0, \quad F \in \mathcal{N}_+,$$

for some $F_n \in \mathcal{M} \subset \mathcal{F}_d^0(\mathcal{U}, \mathcal{Y})$, and some scalars α_n . Now

$$\begin{aligned} \|\theta_T(\tau)F\|_T &\leq \sup_{\mathcal{U}} \left\| (P_{T'} - P_0) \left(\sum_{n=1}^N \alpha_n L_{t_n+\tau} F_n R_{t_n+\tau} \right) \cdot (P_T - P_{-d})u \right\|_T \\ (18) \quad &+ \sup_{\mathcal{U}} \left\| (P_T - P_{T'}) \left(\sum_{n=1}^N \alpha_n L_{t_n+\tau} F_n R_{t_n+\tau} \right) \cdot (P_T - P_{-d})u \right\|_T \\ &= \sup \|A(u)\|_T + \sup \|B(u)\|_T, \end{aligned}$$

where the A and B are defined implicitly.

Because the $F_n \in \mathcal{F}_d^0(\mathcal{U}, \mathcal{Y})$,

$$A = (P_{T'} - P_0) \left(\sum_{n=1}^N \alpha_n L_{t_n+\tau} F_n R_{t_n+\tau} \right).$$

Since $A(u)$ is different from zero only on $[0, T']$, and since $\theta_{T'}(\tau)$ is bounded,

$$\begin{aligned} \sup_{\mathcal{U}} \|A(u)\|_T &= \sup_{\mathcal{U}} \|A(u)\|_{T'} = \|\theta_{T'}(\tau)F\|_{T'} \leq |\theta_{T'}(\tau)| \cdot \|F\|_{T'} \\ &\leq |\theta_{T'}(\tau)| \cdot \sup_{\mathcal{U}} \left\| (P_{T'} - P_0) \left(\sum_{n=1}^N \alpha_n L_{t_n} F_n R_{t_n} \right) u \right\|_{T'} \\ (19) \quad &\leq |\theta_{T'}(\tau)| \cdot \sup_{\mathcal{U}} \left\| (P_T - P_0) \left(\sum_{n=1}^N \alpha_n L_{t_n} F_n R_{t_n} \right) u \right\|_T \\ &= |\theta_{T'}(\tau)| \cdot \|F\|_T. \end{aligned}$$

Using the fact that $T - T' \leq T'$, and also using again the fact that the F_n are causal with bounded memory (d) yields

$$\begin{aligned} \|B(u)\|_T &\leq \left\| (P_T - P_{T-T'}) \left(\sum_{n=1}^N \alpha_n L_{t_n+\tau} F_n R_{t_n+\tau} \right) (P_T - P_{T-T'-d})u \right\|_T \\ &= \|C(u)\|_T, \end{aligned}$$

where C is defined implicitly. Now,

$$L_{T-T'}CR_{T-T'} = (P_{T'} - P_0) \left(\sum_{n=1}^N \alpha_n L_{T-T'+t_n+\tau} F_n R_{T-T'+t_n+\tau} \right) (P_{T'} - P_{-d})$$

so, by the fact that $\theta_{T'}(T - T' + \tau)$ is a bounded operator,

$$\begin{aligned} \sup_{\mathcal{U}} \|L_{T-T'}CR_{T-T'}(u)\|_{T'} &\leq |\theta_{T'}(T - T' + \tau)| \sup_{\mathcal{U}} \left\| (P_{T'} - P_0) \left(\sum_{n=1}^N \alpha_n L_{t_n} F_n R_{t_n} \right) (P_{T'} - P_{-d}) \right\|_{T'} \\ &= |\theta_{T'}(T - T' + \tau)| \cdot \|F\|_{T'}. \end{aligned}$$

But, $\sup_{\mathcal{U}} \|L_{T-T'}CR_{T-T'}(u)\|_{T'} = \sup_{\mathcal{U}} \|C(u)\|_{T'}$. Thus

$$\begin{aligned} \sup_{\mathcal{U}} \|B(u)\|_T &\leq \sup_{\mathcal{U}} \|C(u)\|_{T'} \leq |\theta_{T'}(T - T' + \tau)| \cdot \|F\|_{T'} \\ (20) \qquad \qquad &\leq |\theta_{T'}(T - T' + \tau)| \cdot \|F\|_T. \end{aligned}$$

Combining the inequalities (18), (19) and (20) yields

$$\|\theta_T(\tau)F\|_T \leq (|\theta_T(\tau)| + |\theta_{T'}(T - T' + \tau)|) \cdot \|F\|_T$$

for all $F \in \mathcal{N}_+$, which establishes the result when $T \leq 2T'$. This can be extended to all $T \geq T'$ by induction. \square

If now \mathcal{S}_0 is a class of systems with $\mathcal{H} \subset \mathcal{F}_d^0(\mathcal{U}, \mathcal{Y})$ and is dynamical with respect to some $T' > 0$, one can put T_0 equal to the infimum of all such T' and know that \mathcal{S}_0 is dynamical with respect to any $T > T_0$. It is to be noted that the hypothesis that $\mathcal{H} \subset \mathcal{F}_d^0(\mathcal{U}, \mathcal{Y})$ cannot be dropped in this assertion. In fact, it is not very difficult to give an example where Proposition II.22 is violated if the mappings F are not causal with bounded memory (d); thus the semigroup property is not preserved.

If \mathcal{S}_0 is a linear dynamical class with respect to T and with $\mathcal{H} \subset \mathcal{F}_d^0(\mathcal{U}, \mathcal{Y})$, then it is clearly possible to deal with the discrete parameter semigroup $\{\theta^n = \theta(nT)\}$, $n = 0, 1, 2, \dots$, and still completely describe the future of the system by virtue of the interpolation formula (9). Under certain conditions when $\mathcal{S}_0 = (\mathcal{Y}, g, \mathcal{H}, \mathcal{U})$ is a linear dynamical class, the discrete parameter semigroup $\{\theta^n\}$ can be used to induce a "corresponding" semigroup $\{\tilde{\theta}^n\}$ of linear operators on the linear space spanned by the system parameter space \mathcal{X}_1 of an ε -representation of \mathcal{S}_0 . We describe a situation in which this can be done and construct the $\tilde{\theta}^n$. The construction is not unique, as will be seen, but any $\{\tilde{\theta}^n\}$ so devised approximates $\{\theta^n\}$ in the sense to be indicated.

Let it be assumed that \mathcal{Y} is $\mathcal{L}^{(2)}$. Write $\mathcal{S}_n = (\mathcal{Y}_T, g, \mathcal{H}_{nT}, \mathcal{U}_T)$, $n = 0, 1, 2, \dots$, for the classes of truncated systems, where \mathcal{H}_{nT} is the set of all $\pi_{nT}F$, $F \in \mathcal{H}$. By the assumption on \mathcal{Y} , $\mathcal{Y}_T = L_2$. Since \mathcal{S}_0 is a linear dynamical class, $\mathcal{H}_{nT} = \theta(nT)\mathcal{H}_0 = \theta^n\mathcal{H}_0$. Let it further be required that $\bigcup_{n=0}^{\infty} \mathcal{H}_{nT}$ is a compact subset of $\mathcal{F}(\mathcal{U}_T, \mathcal{Y}_T)$, and for convenience denote $\bigcup_{n=0}^{\infty} \mathcal{H}_{nT}$ by \mathcal{G}_T . Then each \mathcal{S}_n is a subclass of $\mathcal{S} = (\mathcal{Y}_T, g, \mathcal{G}_T, \mathcal{U}_T)$. Since \mathcal{U}_T and \mathcal{G}_T are compact and $\mathcal{Y}_T = L_2$, \mathcal{S} has a standard ε -representation (\mathcal{S}_1, ϕ_1) , $\mathcal{S}_1 = (\mathcal{Y}_T, f_1, \mathcal{X}_1, \mathcal{U}_T)$ as given by Proposition I.7 of Part I, and ϕ_1 is linear. $\mathcal{X}_1 = \phi_1\mathcal{G}_T$ is a subset of a finite-dimensional Euclidean space; let \mathbb{R}^K be the Euclidean space generated by \mathcal{X}_1 . The representation mapping

ϕ_1 as given by Proposition I.7 is actually defined as a continuous linear map from the closed linear span of \mathcal{G}_T onto \mathbb{R}^K . Obviously the closed linear span $\tilde{\mathcal{G}}_T$ of \mathcal{G}_T is contained in \mathcal{N}_+ , the domain of the $\theta(\tau)$. Let $\{b_1, \dots, b_K\}$ be elements of \mathcal{X}_1 which form a basis for \mathbb{R}^K , and denote the coordinate functionals $\{b_1^*, \dots, b_K^*\}$, so that any element $x \in \mathbb{R}^K$ can be written

$$x = \sum_{i=1}^K b_i^*(x) b_i.$$

The idea of the construction of $\tilde{\theta}$ is that $\tilde{\theta}$ should be the composition of the mappings, ψ, θ, ϕ in that order. However, this will not quite do, because $\psi(x)$, $x \in \mathcal{X}_1$, is not necessarily contained in $\tilde{\mathcal{G}}_T$, and hence is not necessarily in \mathcal{N}_+ , the domain of $\theta(\tau)$. To correct this, we construct a linear mapping $\tilde{\psi}$ which does satisfy the condition $\psi(x) \in \tilde{\mathcal{G}}_T$, $x \in \mathcal{X}_1$, and which is close to ψ . Consider the continuous linear functionals on the closed linear span of \mathcal{G}_T given by $b_i^* \circ \phi$, $i = 1, \dots, K$. Let \mathcal{E}_i be the null space of $b_i^* \circ \phi$. First choose an element H_1 belonging to \mathcal{G}_T that does not belong to \mathcal{E}_1 ; this is possible by the definitions of \mathcal{X}_1 and b_1 . Then $b_1^* \circ \phi(H_1) = \alpha_1 \neq 0$. Next, choose $H_2 \in \mathcal{G}_T$, not in \mathcal{E}_2 , and linearly independent of H_1 . This can be done by virtue of the linear independence of the b_i , and yields $b_2^* \circ \phi(H_2) = \alpha_2 \neq 0$. Continue this procedure to obtain a linearly independent set $\{H_1, \dots, H_K\}$, $H_i \in \mathcal{G}_T$, satisfying $b_i^* \circ \phi(H_i) = \alpha_i \neq 0$. Define another basis for \mathbb{R}^K with elements in \mathcal{X}_1 by

$$c_j = \sum_{i=1}^K [b_i^* \circ \phi(H_j)] b_i = \phi(H_j);$$

since each $\phi(H_j) \in \mathcal{X}_1$, it is clear that the c_j do belong to \mathcal{X}_1 . Define $\tilde{\phi}$, a linear mapping from the linear span of $\{H_1, \dots, H_K\}$ onto \mathbb{R}^K , by $\tilde{\phi}(H_i) = c_i$, $i = 1, \dots, K$, and extending linearly. $\tilde{\phi}$ is one-to-one, so we can define $\tilde{\psi} = \tilde{\phi}^{-1}$, a linear mapping from \mathbb{R}^K onto the linear span of $\{H_1, \dots, H_K\}$. If \tilde{H} belongs to the linear span of $\{H_1, \dots, H_K\}$ then we have

$$\phi(\tilde{H}) = \phi\left(\sum_{i=1}^K \gamma_i H_i\right) = \tilde{\phi}(\tilde{H}).$$

Thus $\tilde{\psi}$ carries any element in \mathcal{X}_1 into $\tilde{\mathcal{G}}_T$. As was already mentioned, $\tilde{\psi}$ is not uniquely defined, except in certain cases of finite-dimensional \mathcal{G}_T , since the choice of H_1, \dots, H_K is not unique and the resulting linear space spanned by them is not unique.

It now follows that if $H \in \mathcal{G}_T$, then

$$\|H - \tilde{\psi} \circ \phi_1(H)\| \leq \varepsilon \left[1 + \sum_{k=1}^K |a_k| \right],$$

where the a_k are defined by

$$\phi_1(H) = \sum_{k=1}^K a_k c_k = \sum_{k=1}^K a_k \phi(H_k).$$

In fact, $\|H - \psi_1 \circ \phi_1(H)\| \leq \varepsilon$, and a routine calculation shows that

$$\|\psi \circ \phi_1(H) - \tilde{\psi} \circ \phi_1(H)\| \leq \sum_{k=1}^K |a_k| \|\psi \circ \phi(H_k) - H_k\| \leq \varepsilon \sum_{k=1}^K |a_k|.$$

The set $\phi_1(\mathcal{G}_T) = \mathcal{X}_1$, being compact, is bounded so

$$b \stackrel{\text{d}}{=} \sup_{\mathcal{G}_T} \left(\sum_{k=1}^K |a_k| \right)$$

is a finite positive number. The approximating semigroup can now be defined and a formula given for the n -step error in the approximation.

PROPOSITION II.24. *The mapping $\tilde{\theta}$ from \mathbb{R}^K into \mathbb{R}^K given by $\tilde{\theta} = \phi_1 \circ \theta \circ \tilde{\psi}$ is well-defined and linear. If $H_0 \in \mathcal{G}_T$, then with b as defined above,*

$$(21) \quad \|\theta^n H_0 - \tilde{\psi} \circ \tilde{\theta}^n \circ \phi_1(H_0)\| \leq \varepsilon[1 + |\theta| + \cdots + |\theta|^n][1 + b],$$

where $|\theta|$ denotes the norm of $\theta = \theta(T)$.

Proof. It has already been ascertained that the range of $\tilde{\psi}$ is contained in the linear span of \mathcal{G}_T , which in turn is contained in \mathcal{N}_+ . So $\theta \circ \tilde{\psi}$ is defined. By definition, \mathcal{G}_T is invariant with respect to θ ; since θ is linear, the linear span of \mathcal{G}_T is carried into itself by θ . Thus the range of $\theta \circ \tilde{\psi}$ is contained in the domain of ϕ_1 , and $\phi_1 \circ \theta \circ \tilde{\psi}$ is defined as a linear transformation from \mathbb{R}^K into itself.

If $H_0 \in \mathcal{G}_T$, $\tilde{H}_0 = \tilde{\psi} \circ \phi_1(H_0) \in \tilde{\mathcal{G}}_T$ and $\|H_0 - \tilde{H}_0\| \leq (1 + b)\varepsilon$, as already shown. Then

$$\|\theta \tilde{H}_0 - \theta H_0\| \leq |\theta| \|H_0 - \tilde{H}_0\| \leq \varepsilon |\theta| (1 + b)$$

and

$$\|\theta H_0 - \tilde{\psi} \circ \phi_1 \circ \theta \circ \tilde{\psi} \circ \phi_1(H_0)\| \leq \varepsilon(1 + b)(1 + |\theta|).$$

The inequality (21) follows by induction. \square

Only linear predictability and associated ideas have been considered in this section. However, it probably should be noted, although the fact is obvious, that a class of systems could be described as predictable in a wider sense. Indeed, if $\{T_n\}$, $n = 1, 2, \dots$, is any sequence of mappings from $\mathcal{F}_d^o(\mathcal{U}_T, \mathcal{Y}_T)$ into $\mathcal{F}_d^o(\mathcal{U}_T, \mathcal{Y}_T)$ so that the images under these mappings satisfy the conditions of Proposition II.15, then the class is “predictable” in an obvious sense.

5. Remarks. It will be noticed that, for what has been labeled a linear dynamical class of systems, a structure has been described that is analogous to the usual state-variable formulation of a linear system. In fact, we can write either

$$F_t = \theta(t)F_0, \quad y_t = F_t u_t$$

or

$$F_{nt} = \theta(T)F_{(n-1)T}, \quad y_{nT} = F_{nT}u_{nT},$$

where $u_t = (P_{t+T} - P_{t-d})u$, $y_t = (P_{t+T} - P_t)y$. The first equation in either case corresponds to the state equation for a linear, time-invariant unforced system, and the second to a time-varying observation equation—actually a linear observation equation, since $F_t(u_t)$ for fixed u_t defines a linear mapping from $\mathcal{F}(\mathcal{U}_T, \mathcal{Y}_T)$ into \mathcal{Y}_T . It follows that the identification problem, when there is noise added,

is thereby analogous to the problem of estimating state in a linear system when there is additive noise. A study of identification of $F \in \mathcal{F}$ along the lines of this analogy will be made in a future report. A practical difficulty is, of course, that in modeling many real problems involving rapid time variation the transformations $\theta(\tau)$ cannot be known; but this is simply to say that a rapidly time-varying system is not identifiable if there is no information about the future time variation.

The characterization of system trajectories in terms of strongly continuous semigroups of linear operators obviously suggests the application of some of the elaborate theory of such semigroups to further study of the structure of these classes of systems, but this is a matter for future work.

Appendix. Projections on past and future. The projections P_t used in this paper are defined by

$$(A.1) \quad [P_t f](s) = \begin{cases} f(s), & s \leq t, \\ 0, & s > t, \end{cases}$$

where f is a function on \mathbb{R}^1 . This definition is still meaningful if f is an element of a space for which the elements are equivalence classes of functions equal a.e. Lebesgue, for then it is applied to each representative of the equivalence class. Most of the operations involving these projections are intuitively clear from the definition. Here and there, however, one may want a formal proof of an identity involving these projections. If one is going to the trouble to provide such proofs, it seems as if the properties that are used might as well be axiomatized, particularly since this does not involve much effort. Then generalizations are at least possible. There is nothing new in thus generalizing the notions of past and future, of course; see, e.g., [3], [4] and [5]. However, it is not the intent in this paper really to pursue any notion of generalized time; so we do not build on theory established in the references cited, but merely develop some simple results ad hoc. These results are more than sufficient for what is needed here.

For the remainder of this Appendix, the operators P_t are not to be taken as defined in §1 unless such an interpretation is specifically indicated, but are to be considered abstractly as operators belonging to a family according to the following definition.

DEFINITION A.1. Let \mathcal{Z} be a linear space. Let $\{P_t\}$, $-\infty \leq t \leq \infty$, be a parametrized family of operators on \mathcal{Z} (that is, mappings from \mathcal{Z} into \mathcal{Z}) such that the following conditions are satisfied:

- (i) $P_{-\infty} = 0$ (the zero operator); $P_{+\infty} = I$ (the identity operator).
- (ii) $P_t P_s = P_s P_t$ for all t, s .
- (iii) If $t \leq s$, $P_t P_s = P_t$.
- (iv) P_t is linear on \mathcal{Z} .
- (v) If $(P_b - P_a)y = (P_b - P_a)z$ for arbitrarily large positive numbers b and arbitrarily large negative numbers a where y and z are elements of \mathcal{Z} , then $y = z$.

Then $\{P_t\}$ will be called a family of *generalized time projections* on \mathcal{Z} (g.t. projections).

PROPOSITION A.1. Let \mathcal{Z} be any \mathcal{L}^p -space, or any $L^p(\mathbb{R}^1)$ -space, or \mathcal{B} . Let $\{P_t\}$ be the family of projection operators defined by equation (1), or the extension

of (1) to equivalence classes of functions. Then $\{P_t\}$ is a family of g.t. projections on the space in question.

Proof. The proof is obvious. \square

The projection property (P) as defined in §1 is still a meaningful concept when applied to g.t. projections on a subset of \mathcal{L} . Let \mathcal{L}_1 and \mathcal{L}_2 be linear spaces with families of g.t. projections $\{P_t\}$ and $\{Q_t\}$, respectively. Let \mathcal{U} be a subset of \mathcal{L}_1 with property (P), and let F be a mapping from \mathcal{U} into \mathcal{L}_2 . As in the special case, F is said to be *causal* if $Q_t F(u) = Q_t F P_t(u)$ for all t and all $u \in \mathcal{U}$; F has *bounded memory (d)* if $(Q_\infty - Q_t)F(u) = (Q_\infty - Q_t)F(P_{\infty} - P_{t-d})(u)$ for all t and all $u \in \mathcal{U}$.

PROPOSITION A.2 (Proposition II.7). *If F is a mapping from \mathcal{U} into \mathcal{L}_2 that is causal and has bounded memory (d), then for every $T > 0$,*

$$(A.2) \quad (Q_{t+T} - Q_t)F(u) = (Q_{t+T} - Q_t)F(P_{t+T} - P_{t-d})(u)$$

for all t and all $u \in \mathcal{U}$.

Conversely, if equation (A.2) is satisfied for some $T > 0$, all t and all $u \in \mathcal{U}$, then F is causal and has bounded memory (d).

Proof. We prove first that causality and bounded memory (d) imply the property (A.2). For any $u \in \mathcal{U}$, any real number t and any $T > 0$,

$$\begin{aligned} (Q_{t+T} - Q_t)Fu &= (Q_{t+T} - Q_t)Q_{t+T}Fu = (Q_{t+T} - Q_t)Q_{t+T}FP_{t+T}u \\ &= (Q_{t+T} - Q_t)FP_{t+T}u = Q_{t+T}(Q_\infty - Q_t)F(P_{t+T}u) \\ &= Q_{t+T}(Q_\infty - Q_t)F(P_\infty - P_{t-d})(P_{t+T}u) \\ &= (Q_{t+T} - Q_t)F(P_{t+T} - P_{t-d})u. \end{aligned}$$

Only conditions (i)–(iv) of Definition A.1 and the properties of causality and bounded memory have been used.

Now suppose that (A.2) is satisfied. We prove causality. Let $b > t$ be positive and $a < t$ be negative. Then,

$$(Q_b - Q_a)Q_t Fu = (Q_b - Q_a) \sum_{k=0}^K (Q_{t-kT} - Q_{t-(k+1)T})Fu$$

for any K such that $t = (K+1)T < a$. By (A.2) this is equal to

$$\begin{aligned} (Q_b - Q_a) &\left[\sum_{k=0}^K (Q_{t-kT} - Q_{t-(k+1)T})F(P_{t-kT} - P_{t-(k+1)T-d}) \right] u \\ &= (Q_b - Q_a) \left[\sum_{k=0}^K (Q_{t-kT} - Q_{t-(k+1)T})F(P_{t-kT} - P_{t-(k+1)T-d})P_t \right] u \\ &= (Q_b - Q_a) \left[\sum_{k=0}^K (Q_{t-kT} - Q_{t-(k+1)T})F \right] P_t u \\ &= (Q_b - Q_a)Q_t F P_t u. \end{aligned}$$

Hence, by condition (v) of Definition A.1, $Q_t Fu = Q_t F P_t$.

The proof that F has bounded memory is completely analogous. \square

Let $\{z_k\}$ be an arbitrary sequence of elements belonging to \mathcal{L} , and let $\{\Delta_k = P_{t_k} - P_{t_{k-1}}\}$ be a sequence of differences of g.t. projections, where the $\{t_k\}$, \dots , -2 , -1 , 0 , 1 , 2 , \dots , satisfy $t_k < t_{k+1}$ and $\lim_{k \rightarrow \infty} t_k = \infty$, $\lim_{k \rightarrow -\infty} t_k$

$= -\infty$. In §3 infinite sums of the form $\sum_{k=1}^{\infty} \Delta_k z_k$ are used. These sums have no meaning as far as the structure given by Definition A.1 is concerned, and some further condition is necessary. It is sufficient to require:

Corresponding to $\{z_k\}$, $z_k \in \mathcal{Z}$, and $\{\Delta_k\}$, $k = 1, 2, \dots$, where the Δ^k are as defined above, there exists a $z \in \mathcal{Z}$ with the property

$$(P_b - P_a)z = (P_b - P_a) \sum_{k=-K_1(a)}^{K_2(b)} \Delta_k z_k$$

for all $b \geq a$, where K_1 and K_2 are any integers large enough that the interval $(a, b]$ is contained in the interval $(-t_{K_1}, t_{K_2}]$. Then the sum is defined to be z , in agreement with (v).

It is clear that if \mathcal{Z} is any $\mathcal{L}_T^{(p)}$ -space, or \mathcal{B} (but not, of course, L^p), and the $\{P_t\}$ are ordinary time projections, then the condition holds whenever the z_k are bounded.

REFERENCES

- [1] W. L. ROOT, *Approximate representations of causal systems with bounded memory*, Techniques of Optimization, A. V. Balakrishnan, ed., Academic Press, New York, 1972, pp. 51–64. (Proc. Fourth IFIP Colloquium on Optimization Techniques, Los Angeles, 1971.)
- [2] ———, *On the modeling of systems for identification. Part I: ε -representations of classes of systems*, this Journal, 13 (1975), pp. 927–944.
- [3] W. A. PORTER, *Some circuit theory concepts revisited*, Internat. J. Control, 12 (1970), pp. 433–488.
- [4] R. M. DESANTIS AND W. A. PORTER, *On time-related properties of nonlinear systems*, SIAM J. Appl. Math, 24 (1972), pp. 188–206.
- [5] R. SAEKS, *Resolution Space. Operators and Systems*, Lecture Notes in Economics and Mathematical Systems, No. 82, Springer-Verlag, Berlin, 1973.

ERRATA: ON THE APPROXIMATION OF ITO INTEGRALS USING BAND-LIMITED PROCESSES*

A. V. BALAKRISHNAN†

The second term and the third term in formula (3.8) on p. 249 should be corrected as follows.

In the second term:

$$\frac{p!}{2^v v!} \text{ should read } (p - 2v)! \left(\frac{p!}{(p - 2v)! 2^v v!} \right)^2.$$

In the third term:

$$- \frac{p!}{2^v v!} \text{ should read } + (p - 2v)! \left(\frac{p!}{(p - 2v)! 2^v v!} \right)^2.$$

* This Journal, 12 (1974), pp. 237–251. Received by the editors September 15, 1974.

† Department of Systems Science, University of California at Los Angeles, Los Angeles, California 90024.

EXISTENCE OF SADDLE POINTS AND NASH EQUILIBRIUM POINTS FOR DIFFERENTIAL GAMES*

T. PARTHASARATHY AND T. E. S. RAGHAVAN†

Abstract. In this paper we give sufficient conditions for the existence of saddle points and Nash equilibrium points in the class of classical and relaxed controls for differential games.

1. Introduction. Consider a two person differential game with

$$(1) \quad \frac{dx}{dt} = A(t)x + B(t, u) + C(t, v)$$

as the dynamical system with starting point $X(0) = X_0$ and with payoffs for the two players

$$(2) \quad P_1(u, v) = \mu_1(x) + \int_0^1 F_1(u, t) dt + \int_0^1 G_1(v, t) dt,$$

$$(3) \quad P_2(u, v) = \mu_2(x) + \int_0^1 F_2(u, t) dt + \int_0^1 G_2(v, t) dt.$$

Here, the game is played as follows. A measurable control function $u(t)$ is chosen once for all by player I and a measurable control function $v(t)$ is chosen once for all by player II with $u(t) \in U$, $v(t) \in V$ for all $0 \leq t \leq 1$. Here U , V are compact subsets of real Euclidean spaces. The pair (u, v) determines a unique trajectory $x(t)$ satisfying the dynamical system almost everywhere with the given initial point. Here $x(t)$ is a vector in n -space, $A(t)$ an $n \times n$ matrix, with continuous elements. $B(t, u)$ and $C(t, v)$ are appropriate vector-valued continuous functions on $[0, 1] \times U$, $[0, 1] \times V$ respectively. Further we assume F_1 , G_1 , F_2 , G_2 to be continuous and μ_1 , μ_2 to be continuous linear functionals on $C[0, 1]$, the space of continuous functions on $[0, 1]$.

A relaxed control for the first player is a measurable function σ from $[0, 1]$ to P_U , where P_U is the space of probability distributions on U . A relaxed control τ is similarly defined.

We know from a theorem of Warga [4] that for every (σ, τ) there is a unique absolutely continuous solution for the differential equation

$$(4) \quad \frac{dx}{dt} = \int_U \int_V [A(t)x + B(t, u) + C(t, v)] d\sigma(u, t) d\tau(v, t)$$

with the initial condition $x(0) = x_0$. Using that trajectory and σ, τ we define

$$P_1(\sigma, \tau) = \mu_1(x) + \int_0^1 \left(\int_U F_1(u, t) d\sigma(u, t) \right) dt + \int_0^1 \left(\int_V G_1(v, t) d\tau(v, t) \right) dt.$$

Similarly $P_2(\sigma, \tau)$ is defined.

* Received by the editors June 20, 1974, and in revised form August 20, 1974.

† Department of Mathematics, University of Illinois at Chicago Circle, Chicago, Illinois 60680.
The first author is on leave from the Indian Statistical Institute, Calcutta, India.

We will prove the following theorems.

THEOREM 1. *Consider the differential game with dynamics and payoffs given by (1), (2) and (3). Then there exists a pair of relaxed controls (σ^*, τ^*) forming a Nash equilibrium over the set of all relaxed controls for the two players. That is,*

$$P_1(\sigma^*, \tau^*) \geq P_1(\sigma, \tau^*) \text{ for all relaxed control } \sigma \text{ for player I.}$$

$$P_2(\sigma^*, \tau^*) \geq P_2(\sigma^*, \tau) \text{ for all relaxed control } \tau \text{ for player II.}$$

THEOREM 2. *Consider the differential game with dynamics and payoffs given by (1), (2) and (3). Let U, V be compact convex and further assume that for fixed t , B , F_1 and F_2 are linear in u and C , G_1 and G_2 are linear in V . Then there exists a Nash equilibrium point in the class of classical controls.*

The next theorem is stated for zero-sum games. That is, $P_1(u, v) = -P_2(u, v)$ for all u, v .

THEOREM 3. *Let the control sets U, V be compact convex. Consider the dynamical system $dx/dt = A(t)x + f(t, u, v)$ with initial condition $x(0) = x_0$ and the payoff*

$$P_1(u, v) = \mu(x) + \int_0^1 h(t, u, v) dt,$$

where f, h are continuous in (t, u, v) and convex in v and concave in u . μ is a nonnegative continuous linear function. Then there is a saddle point (u^*, v^*) in the class of classical controls.

Remark. Proofs of Theorems 1 and 2 are similar to Theorems 4.1 and 4.2 in [1].

2. The set of all relaxed controls for player I can be viewed as a compact convex subset of the unit ball B^* in w^* topology, where B is the Banach space $L'_{[0,1]}[C(U)]$ with the norm $\|\varphi\| = \int_0^1 \sup_{u \in U} |\varphi(u, t)|$. For a proof of this see [4]. Also in our setup the space of relaxed trajectories satisfying (4) is compact in the uniform topology [4]. Since (1), (2), (3) are separated in the variables u, v , the payoffs $P_1(\sigma, \tau)$ and $P_2(\sigma, \tau)$ are jointly continuous on the w^* topologies. Also they are bilinear. For the proof of Theorem 1 we need the following proposition of Glicksberg [3].

PROPOSITION. *Let S be a nonempty compact convex subset of a Hausdorff topological vector space X . Let $\Phi: S \rightarrow S$ be a point-to-set mapping satisfying the following conditions:*

- (i) $\Phi(x)$ is nonempty closed convex for each $x \in S$.
- (ii) $\Phi(x)$ is upper-semicontinuous.

Then there exists a fixed point $x_0 \in \Phi(x_0)$.

Proof of Theorem 1. Consider the following set-valued mapping:

$$\Phi: (\sigma^0, \tau^0) \rightarrow \{(\sigma', \tau'): \max_{\sigma} P_1(\sigma, \tau^0) = P_1(\sigma', \tau^0) \text{ and} \\ \max_{\tau} P_2(\sigma^0, \tau) = P_2(\sigma^0, \tau')\}.$$

Clearly this mapping satisfies the conditions of the above proposition. Hence we have a $(\sigma^*, \tau^*) \in \Phi(\sigma^*, \tau^*)$ and this (σ^*, τ^*) is the required Nash equilibrium pair. This completes the proof of Theorem 1.

Proof of Theorem 2. Let (σ^*, τ^*) be an equilibrium pair in accord with Theorem 1. Consider the measurable vector function

$$\mathcal{S}(t) = \begin{cases} \int_U B(u, t) d\sigma^*(u, t), \\ \int_U F_1(u, t) d\sigma^*(u, t), \\ \int_U F_2(u, t) d\sigma^*(u, t). \end{cases}$$

Since B, F_1, F_2 are linear in U , $\mathcal{S}(t)$ is an element of the compact convex set L_t , where

$$L_t = \left\{ \begin{bmatrix} B(u, t) \\ F_1(u, t) \\ F_2(u, t) \end{bmatrix} ; u \in U \right\}.$$

By the Fillipov lemma [2] there is a measurable function u_0 from $[0, 1]$ to U such that

$$\begin{bmatrix} B(u_0(t), t) \\ F_1(u_0(t), t) \\ F_2(u_0(t), t) \end{bmatrix} = \mathcal{S}(t).$$

Similarly there is a measurable $v_0(t)$ from $[0, 1]$ to V such that

$$\begin{bmatrix} C(v_0(t), t) \\ G_1(v_0(t), t) \\ G_2(v_0(t), t) \end{bmatrix} = \int_V \begin{bmatrix} C(v, t) \\ G_1(V, t) \\ G_2(V, t) \end{bmatrix} d\tau^*(v, t).$$

Also in our case (u_0, v_0) does the same job as (σ^*, τ^*) for the payoff. This completes the proof.

Proof of Theorem 3. From Theorem 4.1 in [1] there exists a saddle point in relaxed controls. Namely $P_1(\sigma^*, \tau) \geq P_1(\sigma^*, \tau^*) \geq P_1(\sigma^*, \tau^*)$ for all (σ, τ) , and in particular $P_1(\sigma^*, \tau^*) \leq P_1(\sigma^*, v)$, where $P_1(\sigma^*, v)$ denotes the payoff to I when II uses the classical control v , and I uses the relaxed control σ^* . The sets U, V being convex, the classical control sets are convex. We will show that $P_1(\sigma^*, v)$ is convex in v .

Let v_1, v_2 be any two classical controls for II. Let x_1, x_2, x_3 be the associated trajectories corresponding to $(\sigma^*, v_1), (\sigma^*, v_2)$ and $(\sigma^*, \lambda v_1 + (1 - \lambda)v_2)$ respectively, for any $0 < \lambda < 1$. We have

$$\begin{aligned} \frac{dx_3}{dt} &= A(t)x_3 + \int f(t, u, \lambda v_1 + (1 - \lambda)v_2) d\sigma^*(u, t) \\ &\leq A(t)x_3 + \lambda \int f(t, u, v_1) d\sigma^*(u, t) + (1 - \lambda) \int f(t, u, v_2) d\sigma^*(u, t). \end{aligned}$$

Here the inequalities are meant componentwise.

$$\frac{dx_1}{dt} = A(t)x_1 + \int f(t, u, v_1) d\sigma^*(u, t),$$

$$\frac{dx_2}{dt} = A(t)x_2 + \int f(t, u, v_2) d\sigma^*(u, t),$$

$$\frac{d}{dt}(\lambda x_1 + (1 - \lambda)x_2) - \frac{dx_3}{dt} \geq A(t)(\lambda x_1 + (1 - \lambda)x_2) - A(t)x_3.$$

Let $z = \lambda x_1 + (1 - \lambda)x_2 - x_3$, where $dz/dt \geq A(t) \cdot z$. Further $z(0) = 0$. Since $A(t)$ is nonnegative it follows that $z(t)$ is a nonnegative vector function. Further μ_1 is a nonnegative linear functional and hence $\mu_1(z) \geq 0$.

Now it follows that $P_1(\sigma^*, \lambda v_1 + (1 - \lambda)v_2) \leq \lambda P_1(\sigma^*, v_1) + (1 - \lambda)P_1(\sigma^*, v_2)$. Similarly we can show that $P_1(u, \tau^*)$ is concave in u . Define

$$v_0(t) = \int_V v d\tau^*(v, t).$$

$v_0(t)$ is measurable and since V is compact convex, $v_0(t)$ is itself in V for each t . Let x^* be the trajectory for (σ^*, τ^*) . Since $h(t, u, v)$ is convex in v for each fixed t it follows from Jensen's inequality that

$$\int h(t, u, v) d\tau^*(v, t) \geq h(t, u, v_0(t)).$$

This implies that

$$P_1(\sigma^*, \tau^*) \geq \mu_1(x^*) + \iint h(t, u, v_0(t)) d\sigma^*(u, t) dt.$$

Let x^0 be the trajectory for (σ^*, v_0) . Again using Jensen's inequality for the vector convex function $f(t, u, v)$ in v we can show that $x^* \geq x^0$, and that $P_1(\sigma^*, \tau^*) \geq P_1(\sigma^*, v_0)$. Of course $P_1(\sigma, \tau^*) \geq P(\sigma, v_0)$ for any general σ . Since $P_1(\sigma^*, \tau^*) \leq P_1(\sigma^*, v_0)$ is satisfied by the optimality of (σ^*, τ^*) , we get $P_1(\sigma^*, \tau^*) = P_1(\sigma^*, v_0)$. Also $P_1(\sigma, v_0) \leq P_1(\sigma, \tau^*) \leq P_1(\sigma^*, \tau^*)$ for all σ . This proves that v_0 is optimal for II. Similarly one can prove that the function u_0 defined by

$$u_0(t) = \int_U u d\sigma^*(u, t)$$

is optimal for I. This completes the proof of the theorem.

REFERENCES

- [1] R. J. ELLIOTT, N. J. KALTON AND L. MARKUS, *Saddle points for linear differential games*, this Journal, 11 (1973), pp. 100-112.
- [2] A. FRIEDMAN, *Differential Games*, John Wiley, Interscience, New York, 1971.
- [3] I. L. GLICKSBERG, *A further generalization of the Kakutani fixed point theorem, with application to Nash Equilibrium points*, Proc. Amer. Math. Soc., 3 (1952), pp. 170-174.
- [4] J. WARGA, *Functions of relaxed controls*, this Journal, 5 (1967), pp. 628-641.

NECESSARY CONDITIONS FOR OPTIMALITY OF CAUCHY PROBLEMS FOR PARABOLIC PARTIAL DIFFERENTIAL SYSTEMS*

N. U. AHMED† AND K. L. TEO‡

Abstract. Recently Zolezzi [4] presented a necessary condition for optimality for the problem of optimal control of systems governed by parabolic partial differential equations with a first boundary condition. In this paper, we have developed (Theorem 4.3 and Corollary 4.5) a similar necessary condition for Cauchy problems (unbounded domains) that arise naturally in the optimal control problems of stochastic Ito differential systems with fixed terminal time. For the proof of our results we use several theorems of Aronson [1] which have been slightly modified to suit our requirements.

1. Introduction. In this paper, we consider the problem of optimal control of systems described by parabolic partial differential equations with a Cauchy condition. For this problem, a necessary condition for optimality in integral form and its pointwise version are given in Theorem 4.3 and Corollary 4.5, respectively. Note that this problem arises naturally in the optimal control problems of stochastic Ito differential systems with fixed terminal time. The proofs of our results are based on several theorems of Aronson [1], as modified in this paper.

Consider the system described by the following Cauchy problem:

$$\begin{aligned} S \quad & L(u)\phi(u)(x, t) = f(x, t, u(x, t)), \quad (x, t) \in R^n \times I \triangleq G, \\ & \phi(u)(x, 0) = \psi_0(x), \quad x \in R^n, \end{aligned}$$

for each $u \in D$, where $I \triangleq [0, T]$; T is a positive number; R^n is the n -dimensional Euclidean space; the set D , to be defined later, is the class of admissible controls, and the (parabolic) operator $L(u)$, dependent on the control $u \in D$, is given by

$$\begin{aligned} L(u) \cdot \psi \triangleq & \psi_t - \{a_{ij}(x, t, u(x, t)) \cdot \psi_{x_i} + a_j(x, t, u(x, t)) \cdot \psi_{x_j} \\ & - b_j(x, t, u(x, t)) \cdot \psi_{x_j} - c(x, t, u(x, t)) \cdot \psi \} \end{aligned}$$

with

$$\psi_t \triangleq \frac{\partial \psi}{\partial t} \quad \text{and} \quad \psi_{x_i} \triangleq \frac{\partial \psi}{\partial x_i}.$$

For convenience, the variable (x, t) will be suppressed, and $\alpha(u)$ will be used to denote the function $\alpha(x, t, u)$, where $\alpha(u)$ stands for any of the coefficients or the solution of the system S .

Note that we make use of the standard convention throughout the paper of taking summation up to n over repeated indices.

* Received by the editors February 11, 1974, and in revised form August 19, 1974.

† Department of Electrical Engineering, University of Ottawa, Ottawa, Canada.

‡ Department of Applied Mathematics, University of New South Wales, New South Wales, Australia.

2. Basic assumptions, definitions and problem statement. For each $u \in D$, the coefficients of the operator $L(u)$ are assumed to be defined and measurable on G . Before describing the remaining assumptions on L , we introduce some useful notation.

$|B|$ denotes the Lebesgue measure of a measurable set B of any finite-dimensional Euclidean space. $\partial\Omega$ denotes the boundary of the set $\Omega \subset R^n$ and $\bar{\Omega}$ its closure. Let K be any connected subset of an n -dimensional Euclidean space, and denote by $C^l(K)$, $1 \leq l \leq \infty$, the class of l -times differentiable real-valued functions on K . $C_0^l(K)$ denotes the class of all functions in $C^l(K)$ with compact support on K . $W^1(K)$ is the completion of $C_0^\infty(K)$ in the norm

$$\|z\| \triangleq \|z\|_2 + \|z_x\|_2,$$

where

$$\|z\|_2 \triangleq \left(\int_K |z|^2 dx \right)^{1/2} \quad \text{and} \quad \|z_x\|_2 \triangleq \left(\int_K \sum_{i=1}^n |z_{x_i}|^2 dx \right)^{1/2}.$$

$Y(I, X)$ denotes a normed space of functions defined on the interval I with values in a normed space X , for example, $L^p(I, L^q(R^n))$, $p, q \geq 1$.

For any nonintegral positive number λ and for any measurable subset $G_k \subset G$, $H^{\lambda, \lambda/2}(G_k)$ denotes the Banach space of functions z that are continuous on G_k and have derivatives of the form

$$D_t^\alpha \cdot D_x^\beta z \triangleq \frac{\partial \alpha}{\partial t^\alpha} \cdot \left(\frac{\partial^{\beta_1 + \dots + \beta_n}}{\partial x_1^{\beta_1} \dots \partial x_n^{\beta_n}} z \right), \quad \sum_{i=1}^n \beta_i = \beta,$$

β_i a nonnegative integer, $2\alpha + \beta < \lambda$, and have a finite norm

$$|z|_{G_k}^{(\lambda)} \triangleq \|z\|_{G_k}^{(\lambda)} + \sum_{j=0}^{[\lambda]} \|z\|_{G_k}^{(j)}.$$

Note that $[\lambda]$ denotes the largest integer of λ and

$$\begin{aligned} \|z\|_{G_k}^{(0)} &\triangleq |z|_{G_k}^{(0)} = \max_{G_k} |z|, \\ \|z\|_{G_k}^{(j)} &\triangleq \sum_{(2\alpha + \beta = j)} |D_t^\alpha \cdot D_x^\beta \cdot z|_{G_k}^{(0)}, \\ \|z\|_{G_k}^{(\lambda)} &\triangleq \|z\|_{x, G_k}^{(\lambda)} + \|z\|_{t, G_k}^{(\lambda/2)}, \\ \|z\|_{x, G_k}^{(\lambda)} &\triangleq \sum_{(2\alpha + \beta = [\lambda])} \|D_t^\alpha \cdot D_x^\beta \cdot z\|_{x, G_k}^{(\lambda - [\lambda])}, \\ \|z\|_{t, G_k}^{(\lambda/2)} &\triangleq \sum_{0 < \lambda - 2\alpha - \beta < 2} \|D_t^\alpha \cdot D_x^\beta \cdot z\|_{t, G_k}^{(\lambda - 2\alpha - \beta)/2}, \\ \|z\|_{x, G_k}^{(\gamma)} &\triangleq \sup_{(x, t), (x', t) \in \bar{G}_k} \frac{|z(x, t) - z(x', t)|}{|x - x'|^\gamma}, \quad 0 < \gamma < 1, \\ \|z\|_{t, G_k}^{(\gamma)} &\triangleq \sup_{(x, t), (x, t') \in \bar{G}_k} \frac{|z(x, t) - z(x, t')|}{|t - t'|^\gamma}, \quad 0 < \gamma < 1. \end{aligned}$$

For the class of admissible controls D , we introduce the following.

DEFINITION 2.1. The class of admissible controls on G is given by the set

$$D \triangleq \{u: u \text{ measurable on } G, u(x, t) \in U \text{ for a.a. } (x, t) \in G\},$$

where U is a bounded convex subset of R^m .

For the weak solution of the system S , we use the definition of Aronson [1, p. 638].

DEFINITION 2.2. For each $u \in D$, a function $\phi(u)$ is said to be a *weak solution of the problem S* , if

$$\phi(u) \in L^\infty(I, L^2(R^n)) \cap L^2(I, W^1(R^n))$$

and

$$(2.1) \quad \int_G [-\phi(u) \cdot \varphi_t + a_{ij}(u) \cdot \phi_{x_i}(u) \cdot \varphi_{x_j} + a_j(u) \cdot \phi(u) \cdot \varphi_{x_j} - b_j(u) \cdot \phi_{x_j}(u) \cdot \varphi - c(u) \cdot \phi(u) \cdot \varphi - f(u) \cdot \varphi] dx dt = 0$$

for every $\varphi \in C_0^1(G^0)$, G^0 interior of the set G , and if

$$(2.2) \quad \lim_{t \downarrow 0} \int_{R^n} \phi(u)(x, t) \cdot z(x) dx = \int_{R^n} \psi_0(x) \cdot z(x) dx$$

for every $z \in C_0^1(R^n)$.

With these preparations, we may state our optimal control problem "P" as: Given the system S , find a control $u^0 \in D$ that minimizes the cost functional

$$J(u) = \int_{R^n} H(x, \phi(u)(x, T)) dx,$$

where $\phi(u)$ is the weak solution of the system S corresponding to the control $u \in D$ and H is a Carathéodory function in $R^n \times R^1$ (i.e., measurable in $x \in R^n$ for each $\phi \in R^1$ and continuous in $\phi \in R^1$ for almost all $x \in R^n$). The function H is assumed to satisfy the following properties:

H(i). There exists a Carathéodory function h so that $H(x, z_1) \leq H(x, z_2) + h(x, z_2)(z_1 - z_2)$ for almost every $x \in R^n$, every $z_1, z_2 \in R^1$, and $|h(x, z)| \leq h_1(x) + h_2 \cdot |z|$ for some $h_1 \in L^2(R^n)$ and $h_2 \in R^1$;

H(ii). $H(\cdot, \bar{z}(\cdot)) \in L^1(R^n)$ for some $\bar{z} \in L^2(R^n)$.

Throughout the paper, we need the following assumptions which will be referred to collectively as (A): $a_{ij}, a_j, b_j, i, j = 1, \dots, n, c$ are bounded Carathéodory functions in $G \times U$; there exist constants $\alpha_t, \alpha_u > 0$ such that $\alpha_t \cdot |y|^2 \leq a_{ij}(x, t, v) \cdot y_i \cdot y_j \leq \alpha_u \cdot |y|^2$ almost everywhere on G for every $v \in R^n$; and $f(u) \in L^p(I, L^2(R^n))$ for every $u \in D$, where $p \in (1, 2]$.

Note that it was shown [2] that every weak solution ϕ of the Cauchy problem S has a continuous representation in G . Thus, ϕ will always be assumed to denote the continuous representative of a given weak solution. Hence there is no difficulty in talking about the value of ϕ at any point of its domain of definition.

3. Certain preparatory results. Let $u^0 \in D$ be the control that solves the problem P (called the optimal control) and consider S^* the system adjoint to the system S corresponding to the optimal control.

$$\begin{aligned} S^* \quad & L^*(u^0) \cdot q(x, t) = 0, & (x, t) \in G, \\ & q(x, T) = h(x, \phi^0(x, T)) \triangleq h_0(x), & x \in R^n, \end{aligned}$$

where

$$L^*(u^0)\psi \triangleq -\psi_t - \{c(u^0) \cdot \psi_{x_j} - b_i(u^0) \cdot \psi\}_{x_i} + a_i(u^0) \cdot \psi_{x_i} - c(u^0) \cdot \psi;$$

h is as defined in H(i) (§ 4). ... ϕ^0 is the weak solution of the system S corresponding to the optimal control $u^0 \in D$.

For brevity, it is noted that the statement " C depends on the structure of differential equation of the system S " means that C is determined by the quantities α_i, α_u, p and the bounds of the coefficients $a_j, b_j, = 1, \dots, n$, and c .

THEOREM 3.1. *Consider the system S . Suppose that the assumption (A) is satisfied and that $\psi_0(\cdot) \in L^2(R^n)$. Then, for any $u \in D$ there exists a unique weak solution $\phi(u) \in L^\infty(I, L^2(R^n)) \cap L^2(I, W^1(R^n))$ of the Cauchy problem S and*

$$\|\phi(u)\|_{2,\infty}^2 + \|\phi_x(u)\|_{2,2}^2 \leq C\{\|\psi_0\|_2^2 + \|f(u)\|_{2,p}^2\},$$

where

$$\|\cdot\|_{2,p} \triangleq \left\{ \int_I \left(\int_{R^n} |\cdot|^2 dx \right)^{p/2} dt \right\}^{1/p}; \quad \|\cdot\|_2 \triangleq \left\{ \int_{R^n} |\cdot|^2 dx \right\}^{1/2};$$

C is a positive constant depending only on T and the structure of the equation of the system S ; and p is as defined in (A).

Proof. The proof is analogous to that given for Theorem 3 in [1, pp. 640–642] with the following minor modifications:

(i) use $\mu = \infty$ in the application of Lemma 1 in [1, pp. 623–624] to obtain the estimate

$$\|\phi^k(u)\|_{2,\infty,Q_k}^2 + \|\phi^k(u)\|_{2,2,Q_k}^2 \leq e^{\beta T} \cdot C_1 \cdot \{\|\psi_0\|_{2,\Sigma_k}^2 + \|f(u)\|_{2,p,Q_k}^2\}$$

and use this estimate instead of the estimate (4.6) in [1, p. 640] in the proof,

(ii) use the norm of $f(u)$ as given in the statement of the theorem throughout the proof.

Remark 3.2. Theorem 3.1 remains valid for the adjoint system S^* .

For the proof of necessary conditions for optimality, it is required to consider the systems S and S^* with their coefficients and data replaced by their corresponding integral averages. For this, let $\omega(\eta; s)$ be a sufficiently smooth nonnegative function defined on R^m for each positive integer s so that $\omega(\eta; s) = 0$ for $|\eta| \geq 1/s$ and $K_m(s) \int_{|\eta| < 1/s} \omega(\eta; s) d\eta = 1$ for all positive integers s , where $K_m(s)$ is the volume of the hypersphere $\{\eta: |\eta| \leq 1/s\}$. For any real-valued measurable function ξ on R^m and for any positive integer s , let us define on R^m the function ξ^s , called the *integral average of ξ* , by

$$\xi^s(\eta) \triangleq K_m(s) \cdot \int_{R^m} \omega(\eta - \eta'; s) \cdot \xi(\eta') d\eta'.$$

For every $u \in D$ and for every positive integer s , let $a_{ij}^s(u)$, $a_j^s(u)$, $b_j^s(u)$, $i, j = 1, \dots, n$, $c^s(u)$ and $f^s(u)$ denote, respectively, the integral averages in (x, t) -space of the functions $a_{ij}(u)$, $a_j(u)$, $b_j(u)$, $i, j = 1, \dots, n$, $c(u)$ and $f(u)$. Similarly, let

ψ_0^s and h_0^s be the integral averages of ψ_0 and h_0 respectively on R^n .

The integral averages are known to have derivatives of arbitrary order [8, p. 14].

With these preparations, we now consider the sequence of Cauchy problems

$$(3.1) \quad \begin{aligned} L^s(u)\phi &= f^s(u), & (x, t) \in G, \\ \phi(x, 0) &= \psi_0^s(x), & x \in R^n, \end{aligned}$$

where, for each $u \in D$ and for each positive integer s , the operator $L^s(u)$ is defined by

$$L^s(u)\psi \triangleq \psi_t - \{a_{ij}^s(u) \cdot \psi_{x_i} + a_j^s(u) \cdot \psi\}_{x_j} - b_j^s(u) \cdot \psi_{x_j} - c^s(u) \cdot \psi.$$

The solution to the Cauchy problem (3.1), corresponding to any $u \in D$, will be denoted by $\phi^s(u)$.

Suppose an optimal control u^0 exists. Corresponding to this control $u^0 \in D$ and for each positive integer s , let q^s be the weak solution of the system (3.2) adjoint to (3.1).

$$(3.2) \quad \begin{aligned} L^{*s}(u^0)q &= 0, & (x, t) \in G, \\ q(x, T) &= h_0^s(x), & x \in R^n. \end{aligned}$$

Since the coefficients and data of the integral averaged problems obviously satisfy the assumptions of Theorem 3.1, the existence and uniqueness of the weak solution of these problems follow from that theorem.

In the sequel, we need the following result.

LEMMA 3.3. *Consider the systems S. Suppose that the assumption (A) is satisfied and that $\psi_0(\cdot) \in L^2(R^n)$. Then for each $u \in D$ the weak solution $\phi(u)$ of the Cauchy problem S is the weak limit in $L^2(I, W^1(R^n))$ of the sequence $\{\phi^s(u)\}_{s=1}^\infty$ of the weak solutions of the Cauchy problems (3.1). Moreover, $\{\phi^s(u)\}_{s=1}^\infty$ converges uniformly to $\phi(u)$ in any compact subset of G .*

Proof. From the estimate of Theorem 3.1 and the fact that integral averaging on G does not increase the norm, we have

$$(3.3) \quad \|\phi^s(u)\|_{2,\infty} + \|\phi_x^s(u)\|_{2,2} \leq C\{\|\psi_0\|_2 + \|f(u)\|_{2,p}\},$$

where C and p are as defined in Theorem 3.1. Thus, using this estimate instead of the estimate (3.3) in [1, p. 635] and replacing Q^m, Ω^m, Q, p', q' and $H_0^{1,2}(\Omega)$ in the rest of the proof for the first two assertions of Theorem 1 in [1, pp. 635–637] by $G, R^n, G, 1, 1$ and $W^1(R^n)$ respectively, we obtain the proof of convergence of $\{\phi^s(u)\}_{s=1}^\infty$ in the weak sense to the weak solution $\phi(u)$.

To prove the second part of the lemma, let K be any compact subset of G . By the estimate (3.3) and Theorem B in [1, p. 616], the sequence $\{\phi^s(u)\}_{s=1}^\infty$ is uniformly bounded on K for any given $u \in D$. Obviously, for each positive integer $s \geq 1$ and for any $u \in D$, $\phi^s(u)$ satisfies the condition (2.1) with the set G replaced by K . This implies that $\phi^s(u)$ is also a weak solution of the differential equation of the system (3.1) on K . Thus it follows from Theorem C in [1, p. 616] and the uniform boundedness of the family $\{\phi^s(u)\}_{s=1}^\infty$ that this family is also equicontinuous in K for any $u \in D$. Therefore it follows from Arzela–Ascoli's theorem

that there is a subsequence which converges uniformly in K . However, $\phi(u)$ is unique. Thus the sequence $\{\phi^s(u)\}_{s=1}^\infty$ converges to $\phi(u)$ uniformly in K . This completes the proof.

Remark 3.4. Lemma 3.3 remains valid for the adjoint systems S^* and (3.2).

For the proof of the necessary condition for optimality of the problem P , it would be required to construct a sequence of first boundary value problems for the system (3.1) and its adjoint system (3.2). Further, we note that for each $u \in D$ it will be convenient to regard the operator $L(u)$ and the forcing function $f(u)$ of the first boundary value problems as being defined throughout the $(n+1)$ -dimensional (x, t) -space. Thus for any first boundary value problem to be considered in this paper, we will adopt the convention that $L(u) \psi \triangleq \psi_t - \Delta \psi$ and $f(x, t, u) \equiv 0$ for all (x, t) outside the domain of definition of the first boundary value problem, where Δ is the Laplacian operator.

Let $G_k = \Sigma_k \times (0, T]$, where $\Sigma_k = \{x: |x| < k\}$ for integers $k \geq 1$. For any $u \in D$ let us consider the first boundary value problems

$$\begin{aligned} (3.4) \quad & L^s(u)\phi(x, t) = f^s(u), & (x, t) \in G_k, \\ & \phi(x, 0) = \psi_0^s(x) \cdot g_k(x), & x \in \Sigma_k, \\ & \phi(x, t) = 0, & (x, t) \in \partial \Sigma_k \times [0, T], \end{aligned}$$

where $L^s(u)$, $f^s(u)$ and ψ_0^s are as defined for the system (3.1) and for each integer $k \geq 1$ the function g_k belongs to $C_0^\infty(\Sigma_k)$ so that $g_k = 1$ on $\bar{\Sigma}_{k-1}$ and $0 \leq g_k(x) \leq 1$ for $x \in \Sigma_k \setminus \Sigma_{k-1}$.

In the sequel, we need the following definition.

DEFINITION 3.5. For each pair of positive integers s, k and for any $u \in D$, a function ϕ is said to be a *weak solution of the first boundary value problem* (3.4) if $\phi \in L^\infty(I, L^2(\Sigma_k)) \cap L^2(I, W^1(\Sigma_k))$ and

$$\begin{aligned} (3.5) \quad & \int_{G_k} [-\phi(u) \cdot \varphi_t + a_{ij}^s(u) \cdot \phi_{x_i}(u) \cdot \varphi_{x_j} + a_j^s(u) \cdot \phi(u) \cdot \varphi_{x_j} \\ & - b_j^s(u) \cdot \phi_{x_j}(u) \cdot \varphi - c^s(u) \cdot \phi(u) \cdot \varphi - f^s(u) \cdot \varphi] dx \cdot dt = 0 \end{aligned}$$

for every $\varphi \in C_0^1(G_k^0)$, G_k^0 interior of the set G_k , and if

$$(3.6) \quad \lim_{t \downarrow 0} \int_{\Sigma_k} \phi(u)(x, t) \cdot z(x) dx = \int_{\Sigma_k} \psi_0(x) \cdot z(x) dx$$

for every $z \in C_0^1(\Sigma_k)$.

Corresponding to the adjoint system (3.2), we consider the sequence of first boundary value problems

$$\begin{aligned} (3.7) \quad & L^{*s}(u^0)q(x, t) = 0, & (x, t) \in G_k, \\ & q(x, T) = h_0^s(x) \cdot g_k(x), & x \in \Sigma_k, \\ & q(x, t) = 0, & (x, t) \in \partial \Sigma_k \times [0, T]. \end{aligned}$$

Since the integral averages obviously satisfy the assumptions of Theorem 5.2 in [3, p. 320], it follows from that theorem that for each pair of positive integers

s and k the system (3.4) [(3.7)] has a classical solution $\phi_k^s(u)[q_k^s]$ belonging to $H^{\lambda, \lambda/2}(G_k)$ with λ any positive nonintegral number (§ 2). In particular, $\phi_k^s(u)[q_k^s]$ is also a weak solution of the system (3.4) [(3.7)]. Further, it follows from [1, Thm. 1, p. 634] that it is the only weak solution.

With these preparations, we may state the following result.

LEMMA 3.6. *Consider the system (3.1). Suppose that all the hypotheses given in Lemma 3.3 are satisfied. Then for each integer $s \geq 1$ and for any $u \in D$, the weak solution $\phi^s(u)$ of the Cauchy problem (3.1) is the weak limit in $L^2(I, W^1(R^n))$ of the sequence $\{\phi_k^s(u)\}_{k=1}^\infty$ of the solutions of the first boundary value problems (3.4). Moreover, $\{\phi_k^s(u)\}_{k=1}^\infty$ converges uniformly to $\phi^s(u)$ in any compact subset of G .*

Proof. By defining $L(u)\psi \triangleq \psi_t - \Delta\psi$ and setting $f \equiv 0$ for all $(x, t) \notin G_k$, the first boundary value problem (3.4) is converted into an equivalent Cauchy problem on G . If we set $\phi_k^s(u)(x, t) \equiv 0$ for all $(x, t) \notin G_k$, then it is clear that $\phi_k^s(u)$ is also the unique weak solution of the Cauchy problem constructed above. Thus, from the estimate of Theorem 3.1 and the fact that $f^s(x, t, u) \equiv 0$ for all $(x, t) \notin G_k$, we have

$$(3.8) \quad \begin{aligned} \|\phi_k^s(u)\|_{2, \infty} + \|(\phi_k^s)_x(u)\|_{2, 2} &\leq C\{\|\psi_0\|_2 + \|f(u)\|_{2, p}\} \\ &\triangleq C_1, \end{aligned}$$

where C and p are as defined in Theorem 3.1.

Therefore, replacing the estimate (4.7) in [1, p. 640] by the estimate (3.8), we obtain the proof of convergence of $\{\phi_k^s(u)\}_{k=1}^\infty$ in the weak sense to $\phi^s(u)$ for each positive integer s and for any $u \in D$ similar to that as given in the proof of Theorem 3 in [1, pp. 640–642]. The second part of the lemma is analogous to that as given in Lemma 3.3. This completes the proof.

Remark 3.7. Lemma 3.6 remains valid for the adjoint systems (3.2) and (3.7).

Remark 3.8. Lemma 3.6 remains valid both for the weak solutions $\phi_k(u)$ and q_k of the first boundary value problems corresponding to the original Cauchy problem S and the adjoint system S^* respectively.

In the sequel, we need the following.

LEMMA 3.9. *Consider the Cauchy problem*

$$(3.9) \quad \begin{aligned} \frac{\partial \phi}{\partial t} &= (a_{ij}^k(x, t) \cdot \phi_{x_i} + a_j^k(x, t) \cdot \phi)_{x_j} + b_j^k(x, t) \cdot \phi_{x_j} \\ &\quad + c^k(x, t) \cdot \phi + f^k(x, t), \quad (x, t) \in G, \\ \phi(x, 0) &= \psi_0(x), \quad x \in R^n. \end{aligned}$$

Suppose that $a_{ij}^k, i, j = 1, \dots, n$, satisfy the inequality in assumption (A) independent of k and that $a_j^k, b_j^k, j = 1, \dots, n, c^k$ are bounded on G uniformly with respect to k and that $\|f^k\|_{2, p}$ is bounded independently of k , where p is as defined in (A). Further, it is assumed that $\psi_0 \in L^2(R^n)$ and that $a_{ij}^k, a_j^k, b_j^k, c^k$ and f^k converge, respectively, to a_{ij}, a_j, b, c and f almost everywhere in G . Then the sequence of the weak solutions $\{\phi^k\}_{k=1}^\infty$ of the Cauchy problems (3.9) converges to ϕ weakly in $L^2(I, W^1(R^n))$, where ϕ is the weak solution of the system (3.9) with its coefficients replaced by their corresponding limits.

Proof. From the estimate of Theorem 3.1 and the hypotheses given for the coefficients of the Cauchy problems (3.9), we have

$$\begin{aligned}\|\phi^k\|_{2,\infty} + \|\phi_x^k\|_{2,2} &\leq C\{\|\psi_0\|_2 + \|f^k\|_{2,p}\} \\ &\leq C_1,\end{aligned}$$

where C and p are as defined in Theorem 3.1 and

$$C_1 = C \cdot \{\|\psi_0\|_2 + \sup_k \|f^k\|_{2,p}\}.$$

Thus, using the above estimate instead of the estimate (3.8), we obtain the proof of convergence of $\{\phi^k\}_{k=1}^\infty$ in the weak sense to ϕ by the argument similar to that given in Lemma 3.6. This completes the proof.

4. Necessary conditions for optimality. For the sake of brevity, let $\int_G f(x, t)g(x, t) dx dt$ be denoted by $\langle f, g \rangle$.

For the proof of the desired necessary condition for optimality for the problem P, we need the following lemmas.

LEMMA 4.1. *Consider the problem P. Suppose that $u^0 \in D$ is an optimal control (whose existence is assumed) and that*

- (i) *the assumption (A) holds;*
- (ii) *$\psi_0 \in L^2(R^n)$.*

Then there exists sequence of solutions q_k^s of the system (3.7) such that

$$(4.1) \quad \lim_{s \rightarrow \infty} \lim_{k \rightarrow \infty} \langle L(u^0)(\phi(u) - \phi(u^0)), q_k^s \rangle = 0 \quad \text{for all } u \in D$$

independently of the order of taking the limit.

Proof. For any pair of integers $s, k \geq 1$ and for any $u \in D$, we have

$$\begin{aligned}(4.2) \quad &\langle L(u^0)(\phi(u^0) - \phi(u)), q_k^s \rangle \\ &= \langle L(u^0)(\phi(u^0) - \phi(u) - \phi^r(u^0) + \phi^r(u)), q_k^s \rangle \\ &\quad + \langle L(u^0)(\phi^r(u^0) - \phi^r(u) - \phi_l^r(u^0) + \phi_l^r(u)), q_k^s \rangle \\ &\quad + \langle L(u^0)(\phi_l^r(u^0) - \phi_l^r(u)), q_k^s \rangle,\end{aligned}$$

where l is any positive integer; $\phi^r(u)$ and $\phi_l^r(u)$ are, respectively, the weak solution and the solution of the systems (3.1) and (3.4) both corresponding to the control $u \in D$; and q_k^s is the solution of the system (3.7). By defining $L(u)\psi = \psi_t - \Delta\psi$ and $f(x, t, u) \equiv 0$ for all $(x, t) \notin G_l$, the first boundary value problem (3.4) with $s = r$ and $k = l$ is converted into an equivalent Cauchy problem on G . If, for the first boundary value problem, we set $\phi_l^r(u) \equiv 0$ for $(x, t) \notin G_l$, then it is clear that the solution $\phi_l^r(u)$ of this problem is also the weak solution of the (extended) Cauchy problem constructed above. Similarly, corresponding to the adjoint system (3.7) if we set its solution $q_k^s(x, t) \equiv 0$ for all $(x, t) \notin G_k$, then this solution is also necessarily the weak solution of the (extended) Cauchy problem constructed by adjoining $L^*(u^0)\psi = -\psi_t - \Delta\psi$ for all (x, t) outside the set G_k to the adjoint system (3.7).

Thus, by integrating by parts the last term in the right-hand side of (4.2) and noting that $\phi_l^r(u^0)(\cdot, 0) = \phi_l^r(u)(\cdot, 0)$, one obtains

$$\begin{aligned}
 (4.3) \quad & \langle L(u^0)(\phi_l^r(u^0) - \phi_l^r(u)), q_k^s \rangle \\
 &= \langle \phi_l^r(u^0) - \phi_l^r(u), L^*(u^0)q_k^s \rangle \\
 &+ \int_{R^n} [\phi_l^r(u^0)(x, T) - \phi_l^r(u)(x, T)] \cdot h_0^s(x) \cdot g_k(x) dx.
 \end{aligned}$$

Writing

$$\begin{aligned}
 (4.4) \quad & \langle \phi_l^r(u^0) - \phi_l^r(u), L^*(u^0)q_k^s \rangle \\
 &= \langle \phi_l^r(u^0) - \phi_l^r(u), (L^*(u^0) - L^{*s}(u^0))q_k^s \rangle \\
 &+ \langle \phi_l^r(u^0) - \phi_l^r(u), L^{*s}(u^0)q_k^s \rangle,
 \end{aligned}$$

it follows from the relation (1.4) in [1, p. 620] that the last term of the above expression reduces to

$$\begin{aligned}
 (4.5) \quad & \langle \phi^r(u^0) - \phi_l^r(u), L^{*s}(u^0)q_k^s \rangle \\
 &= - \int_{R^n} [\phi_l^r(u^0)(x, T) - \phi_l^r(u)(x, T)] \cdot h_0^s(x) \cdot g_k(x) dx.
 \end{aligned}$$

Combining (4.2), (4.3), (4.4) and (4.5), we have

$$\begin{aligned}
 (4.6) \quad & \langle L(u^0)(\phi(u^0) - \phi(u)), q_k^s \rangle \\
 &= \langle L(u^0)(\phi(u^0) - \phi(u) - \phi^r(u^0) + \phi^r(u)), q_k^s \rangle \\
 &+ \langle L(u^0)(\phi^r(u^0) - \phi^r(u) - \phi_l^r(u^0) + \phi_l^r(u)), q_k^s \rangle \\
 &+ \langle \phi_l^r(u^0) - \phi_l^r(u), (L^*(u^0) - L^{*s}(u^0))q_k^s \rangle.
 \end{aligned}$$

Note that $q_k^s \in H^{\lambda, \lambda/2}(G_k)$ (Remark 3.4) and $q_k^s(x, t) \equiv 0$ for $(x, t) \notin G_k$. Integrating by parts the terms containing t -differentials and those with coefficients appearing under the x -differentials and then taking limits with respect to l and r in the same order, it follows from Lemma 3.6 that the second term of (4.6) vanishes and from Lemma 3.3 that the first term vanishes. The remaining term yields

$$\begin{aligned}
 (4.7) \quad & \langle L(u^0)(\phi(u^0) - \phi(u)), q_k^s \rangle \\
 &= \langle (a_{ij}(u^0) - a_{ij}^s(u^0)) \cdot (q_k^s)_{x_j}, \phi_{x_i}(u^0) - \phi_{x_i}(u) \rangle \\
 &- \langle (b_i(u^0) - b_i^s(u^0)) \cdot q_k^s, \phi_{x_i}(u^0) - \phi_{x_i}(u) \rangle \\
 &+ \langle (a_i(u^0) - a_i^s(u^0)) \cdot (q_k^s)_{x_i}, \phi(u^0) - \phi(u) \rangle \\
 &- \langle (c(u^0) - c^s(u^0)) \cdot q_k^s, \phi(u^0) - \phi(u) \rangle.
 \end{aligned}$$

(Note that the order of taking limits with respect to r or l is immaterial due to Lemma 3.3, Theorem 1 in [1, p. 634] and Remark 3.8). Letting $k \rightarrow \infty$ in (4.7), we have from Remark 3.7,

$$\begin{aligned}
 (4.8) \quad & \lim_{k \rightarrow \infty} \langle L(u^0)(\phi(u^0) - \phi(u)), q_k^s \rangle \\
 &= \langle (a_{ij}(u^0) - a_{ij}^s(u^0)) \cdot q_{x_j}^s, \phi_{x_i}(u^0) - \phi_{x_i}(u) \rangle \\
 &- \langle (b_i(u^0) - b_i^s(u^0)) \cdot q^s, \phi_{x_i}(u^0) - \phi_{x_i}(u) \rangle \\
 &+ \langle (a_i(u^0) - a_i^s(u^0)) \cdot q_{x_i}^s, \phi(u^0) - \phi(u) \rangle \\
 &- \langle (c(u^0) - c^s(u^0)) \cdot q^s, \phi(u^0) - \phi(u) \rangle.
 \end{aligned}$$

Since the integral averages converge almost everywhere on G and the coefficients of the Cauchy problem S are bounded on G , it can be shown, from the Lebesgue dominated convergence theorem, that (4.8) in the limit with respect to s converges to zero.

On the other hand, using the properties of the integral averages and the coefficients of the Cauchy problem S, it follows from the Lebesgue dominated convergence theorem that (4.7) in the limit with respect to s reduces to

$$(4.9) \quad \lim_{s \rightarrow \infty} \langle L(u^0)(\phi(u^0) - \phi(u)), q_k^s \rangle = 0.$$

Therefore, we obtain the expression (4.1) independently of the order of taking the limits. This completes the proof.

LEMMA 4.2. *Consider the problem P. Suppose that $u^0 \in D$ is the optimal control (whose existence is assumed) and that*

- (i) *the assumption (A) holds;*
- (ii) *$\psi_0 \in L^2(R^n)$.*

Then there exists a weak solution q of the adjoint system S^ such that*

$$(4.10) \quad \begin{aligned} & \langle (a_{ij}(u^0) - a_{ij}(u)) \cdot \phi_{x_i}(u), q_{x_j} \rangle + \langle (a_f(u^0) - a_f(u)) \cdot \phi(u), q_{x_j} \rangle \\ & - \langle (b_f(u^0) - b_f(u)) \cdot \phi_{x_j}(u), q \rangle - \langle (c(u^0) - c(u)) \cdot \phi(u), q \rangle \\ & \geq \langle f(u^0) - f(u), q \rangle \end{aligned}$$

for all $u \in D$.

Proof. Clearly, for any pair of integers $s, k \geq 1$ and for any $u \in D$,

$$(4.11) \quad \begin{aligned} & \langle L(u^0)\phi(u^0), q_k^s \rangle - \langle L(u)\phi(u), q_k^s \rangle \\ & = \langle (L(u^0) - L(u))\phi(u), q_k^s \rangle + \langle L(u^0)(\phi(u^0) - \phi(u)), q_k^s \rangle. \end{aligned}$$

where $\phi(u)$ is the weak solution of the system S corresponding to the control $u \in D$ and q_k^s is the solution of the system (3.7) corresponding to the optimal control $u^0 \in D$.

Since $q_k^s(x, t) \equiv 0$ for all $(x, t) \notin G_k$, it follows from the relation (1.4) in [1, p. 620] that the left-hand side of the expression (4.11) reduces to

$$(4.12) \quad \begin{aligned} & \langle L(u^0)\phi(u^0), q_k^s \rangle - \langle L(u)\phi(u), q_k^s \rangle \\ & = \langle (f(u^0) - f(u)), q_k^s \rangle + \int_{R^n} [\phi(u)(x, T) - \phi(u^0)(x, T)] g_k(x) \cdot h_0^s(x) dx. \end{aligned}$$

Therefore

$$(4.13) \quad \begin{aligned} & \langle (L(u^0) - L(u))\phi(u), q_k^s \rangle + \langle L(u^0)(\phi(u^0) - \phi(u)), q_k^s \rangle \\ & = \langle (f(u^0) - f(u)), q_k^s \rangle + \int_{R^n} [\phi(u)(x, T) - \phi(u^0)(x, T)] g_k(x) h_0^s(x) dx. \end{aligned}$$

By the construction of the function g_k , we note that $\lim_{k \rightarrow \infty} g_k(x) = 1$ for every $x \in R^n$ and $|g_k(x)| \leq 1$ on R^n for all integers $k \geq 1$. Since the weak solution $\phi(u) \in L^\infty(I, L^2(R^n)) \cap L^2(I, W^1(R^n))$ and is continuous in G , $\phi(u)(x, T) \in L^2(R^n)$ for every $u \in D$. Further it follows from the properties of the function $h_0(x)$

($\triangleq h(x, \phi^0(x, T))$) given in H(i) (§ 2) that $h_0 \in L^2(R^n)$. Thus it is easily verified with the help of the Lebesgue dominated convergence theorem and the fact that $h_0^s \rightarrow h_0$ almost everywhere in R^n that

$$(4.14) \quad \begin{aligned} & \lim_{s \rightarrow \infty} \cdot \lim_{k \rightarrow \infty} \cdot \int_{R^n} [\phi(u)(x, T) - \phi(u^0)(x, T)] \cdot g_k(x) \cdot h_0^s(x) dx \\ &= \int_{R^n} [\phi(u)(x, T) - \phi(u^0)(x, T)] h_0(x) dx \end{aligned}$$

independently of the order of taking limits.

Therefore it follows from Lemma 4.1 that (4.13) in the limit with respect to k and s reduces to

$$(4.15) \quad \begin{aligned} & \langle (a_{ij}(u^0) - a_{ij}(u)) \cdot \phi_{x_i}(u), q_{x_j} \rangle + \langle (a_j(u^0) - a_j(u)) \cdot \phi(u), q_{x_j} \rangle \\ & - \langle (b_j(u^0) - b_j(u)) \cdot \phi_{x_j}(u), q \rangle - \langle (c(u^0) - c(u)) \cdot \phi(u), q \rangle \\ &= \langle (f(u^0) - f(u)), q \rangle + \int_{R^n} [\phi(u)(x, T) - \phi(u^0)(x, T)] \cdot h_0(x) dx. \end{aligned}$$

Using the property H(i) and the definition of the function h_0 , we note that

$$(4.16) \quad \int_{R^n} [\phi(u)(x, T) - \phi(u^0)(x, T)] \cdot h_0(x) \geq 0.$$

Thus, the condition (4.10) follows from the expressions (4.15) and (4.16). This completes the proof of the lemma.

Based on the above results, we present in the theorem below the desired necessary conditions for optimality.

THEOREM 4.3. *Consider the problem P. Suppose that the assumption (A) holds and that $a_{ij}(x, t, \cdot)$, $a_j(x, t, \cdot)$, $b_j(x, t, \cdot)$, $i, j = 1, \dots, n$, $c(x, t, \cdot)$ and $f(x, t, \cdot)$ belong to $C^1(U)$ almost everywhere in G with the gradients bounded in R^m for almost all $(x, t) \in G$ and for every $v \in U$. Then, if $u^0 \in D$ is an optimal control (whose existence is assumed) it is necessary that there exists a weak solution q of the adjoint system S^* so that*

$$(4.17) \quad \begin{aligned} & \sum_{k=1}^m \langle (a_{ij,k}(u^0) \cdot \phi_{x_i}(u^0) + a_{j,k}(u^0) \cdot \phi(u^0)_{x_j} - b_{j,k}(u^0) \cdot \phi_{x_j}(u^0) \cdot q \\ & - c_k(u^0) \cdot \phi(u^0) \cdot q - f_k(u^0) \cdot q), u_k^0 - u_k \rangle \geq 0 \end{aligned}$$

for all $u \in D$, where

$$\alpha_k(x, t, u^0(x, t)) \triangleq \frac{\partial \alpha(x, t, u_1^0(x, t), \dots, u_k, \dots, u_r^0(x, t))}{\partial u_k} \Big|_{u_k = u_k^0(x, t)}$$

and

$$\alpha_k \triangleq a_{ij,k}, a_{j,k}, b_{j,k}, (i, j = 1, \dots, n) c_k \text{ or } f_k.$$

Proof. For any $u \in D$, let $\varepsilon \in [0, 1]$ and let $u - u^0 \triangleq u^1$. Since D is convex, $u^0 + \varepsilon u^1 \in D$. Thus, dividing the inequality (4.10) by ε and replacing u by $u^0 + \varepsilon u^1$, we obtain

$$\begin{aligned}
 (1/\varepsilon) \{ & \langle (a_{ij}(u^0) - a_{ij}(u^0 + \varepsilon u^1)) \cdot \phi_{x_i}(u^0 + \varepsilon u^1), q_{x_j} \rangle \\
 & + \langle (a_j(u^0) - a_j(u^0 + \varepsilon u^1)) \cdot \phi(u^0 + \varepsilon u^1), q_{x_j} \rangle \\
 & - \langle (b_j(u^0) - b_j(u^0 + \varepsilon u^1)) \cdot \phi_{x_j}(u^0 + \varepsilon u^1), q \rangle \\
 & - \langle (c(u^0) - c(u^0 + \varepsilon u^1)) \cdot \phi(u^0 + \varepsilon u^1), q \rangle \} \\
 & \geq (1/\varepsilon) \{ \langle f(u^0) - f(u^0 + \varepsilon u^1), q \rangle \}.
 \end{aligned}
 \tag{4.18}$$

Since, by hypothesis, the coefficients of the system S belong to $C^1(U)$ almost everywhere in G with bounded gradients for almost every $(x, t) \in G$ and every $v \in U$, it follows that their incremental ratios converge to the corresponding Gateaux differentials almost everywhere in G as $\varepsilon \rightarrow 0$. Denoting by α any of the coefficients a_{ij}, a_j, b_j, c, f and noting that $\alpha(u^0 + \varepsilon u^1)(x, t) \rightarrow \alpha(u^0)(x, t)$ a.e. on G as $\varepsilon \rightarrow 0$, it follows from Lemma 3.9 that $\phi(u^0 + \varepsilon u^1) \rightarrow \phi(u^0)$ weakly in $L^2(I, W^1(R^n))$. Thus, by taking the limit with respect to ε in the expression (4.18) and using the facts just mentioned, we obtain the condition (4.17). This completes the proof of the theorem.

In order to obtain the pointwise necessary condition for optimality for the problem P , we need the following well-known result which is presented in the form of a lemma.

LEMMA 4.4. *Let γ be a Lebesgue integrable function defined on G , y a regular point in G and let $E \subset G$ be any measurable set containing y and contracting to the one point set $\{y\}$. Then*

$$\lim_{|E| \rightarrow 0} \left\{ \frac{1}{|E|} \int_E \gamma(\theta) d\theta \right\} = \gamma(y).$$

COROLLARY 4.5. *Consider the problem P . Suppose that all the hypotheses of Theorem 4.3 are satisfied. Then*

$$\begin{aligned}
 \sum_{k=1}^m [& \{ (a_{ij,k}(x, t, u^0(x, t)) \cdot \phi_{x_i}(u^0)(x, t) + a_{j,k}(x, t, u^0(x, t)) \cdot \phi(u^0)(x, t) \\
 & \cdot q_{x_j}(x, t) - b_{j,k}(x, t, u^0(x, t)) \cdot \phi_{x_j}(u^0)(x, t) \cdot q(x, t) - c_k(x, t, u^0(x, t)) \\
 & \cdot \phi(u^0)(x, t) \cdot q(x, t) - f_k(x, t, u^0(x, t)) \cdot q(x, t) \} \cdot \{u_k^0 - v_k\}] \geq 0
 \end{aligned}
 \tag{4.19}$$

for almost all $(x, t) \in G$ and every $v \in U$.

Proof. Let (x^0, t^0) be a regular point contained in the interior of G and suppose that E is a measurable subset of G containing $\{(x^0, t^0)\}$ and contracting to the point $\{(x^0, t^0)\}$ as $|E| \rightarrow 0$. Then, dividing the expression (4.17) by $|E|$ and replacing the control u by the one defined below,

$$u(x, t) = \begin{cases} v & \text{for } (x, t) \in E, \\ u^0(x, t) & \text{elsewhere,} \end{cases}$$

we have

$$\begin{aligned}
 (4.20) \quad & \sum_{k=1}^m \left[\frac{1}{|E|} \cdot \int_E \{ (a_{ij,k}(x, t, u^0(x, t)) \cdot \phi_{x_i}(u^0)(x, t) + a_{j,k}(x, t, u^0(x, t)) \right. \\
 & \qquad \qquad \qquad \cdot \phi(u^0)(x, t)) \cdot q_{x_j}(x, t) \\
 & \qquad \qquad \qquad - b_{j,k}(x, t, u^0(x, t)) \cdot \phi_{x_j}(u^0)(x, t) \cdot q(x, t) - c_k(x, t, u^0(x, t)) \cdot \phi(u^0)(x, t) \\
 & \qquad \qquad \qquad \cdot q(x, t) \\
 & \qquad \qquad \qquad \left. - f_k(x, t, u^0(x, t)) \cdot q(x, t) \} \cdot \{ u_k^0(x, t) - v_k \} dx \cdot dt \right] \geq 0.
 \end{aligned}$$

Letting $|E| \rightarrow 0$ and noting that almost all $(x, t) \in G$ are regular points, we obtain from Lemma 4.4 the condition (4.19). This completes the proof.

Remark 4.6. Existence results on optimal control for systems governed by parabolic partial differential equations with first boundary conditions arising naturally from stochastic optimal control problems has been recently reported in [7, Thm. 3, p. 205] and [10, Thm. 1]. To the knowledge of the authors, similar results do not exist for the Cauchy problem as considered in this paper. This is an open problem.

Acknowledgment. The authors would like to thank the reviewers of the paper for their valuable comments and suggestions.

REFERENCES

- [1] D. G. ARONSON, *Non-negative solutions of linear parabolic equations*, Ann. Scuola Norm. Sup. Pisa, 22 (1968), pp. 607–694.
- [2] D. G. ARONSON AND J. SERRIN, *Local behavior of solutions of quasi-linear parabolic equations*, Arch. Rational Mech. Anal., 25 (1967), pp. 81–122.
- [3] O. A. LADYZHENSKAYA, V. A. SOLONNIKOV AND N. N. URAL'CEVA, *Linear and quasilinear equations of parabolic type*, Translations of Mathematical Monographs, American Mathematical Society, Providence, R.I., 1968.
- [4] T. ZOLEZZI, *Necessary conditions for optimal conditions of elliptic or parabolic problems*, this Journal, 10 (1972), pp. 594–607.
- [5] W. H. FLEMING, *The Cauchy problem for degenerate parabolic equations*, J. Math. Mech., 13 (1964), pp. 987–1008.
- [6] ———, *Optimal continuous-parameter stochastic control*, SIAM Rev., 11 (1969), pp. 470–508.
- [7] ———, *Optimal control of partially observable diffusions*, this Journal, 6 (1968), pp. 194–214.
- [8] S. L. SOBOLEV, *Some applications of functional analysis to mathematical physics*, Translations of Mathematical Monographs, American Mathematical Society, Providence, R.I., 1963.
- [9] J. L. LIONS, *Optimal control of systems governed by partial differential equations*, Springer-Verlag, Berlin, 1971.
- [10] N. U. AHMED AND K. L. TEO, *An existence theorem on optimal control of partially observable diffusions*, this Journal, 12 (1974), pp. 351–355.

MEASURABILITY THEOREMS FOR STOCHASTIC EXTREMALS*

P. KALL† AND W. OETTLI‡

Abstract. Measurability of the optimal value is proved for a rather general class of parametric optimization problems. The class considered includes in particular the stochastic convex programs.

In [1] a direct and elementary proof was given for the measurability of the optimal value of a stochastic linear program. It turns out that the same technique yields measurability statements for very general nonlinear optimization problems, too.

1. Let Ω be some measurable space, and let X be some subset of \mathbb{R}^n . X contains then a countable dense subset which we denote by $\Xi = \{\xi_i\}_{i \in \mathbb{N}}$. Let the functions $F: X \times \Omega \rightarrow \mathbb{R}$ and $f: X \times \Omega \rightarrow \mathbb{R}$ be measurable on Ω for every $x \in X$. We are interested in the measurability of the optimal value

$$\Phi(\omega) = \begin{cases} \inf_x \{F(x, \omega) | x \in X, f(x, \omega) \leq 0\} & \text{if } \{x | x \in X, f(x, \omega) \leq 0\} \neq \emptyset, \\ +\infty & \text{otherwise.} \end{cases}$$

Let us define in addition for $n \in \mathbb{N}$,

$$\tau_n(\omega) = \begin{cases} \inf_x \{F(x, \omega) | x \in X, f(x, \omega) \leq 1/n\} & \text{if } \{x | x \in X, f(x, \omega) \leq 1/n\} \neq \emptyset, \\ +\infty & \text{otherwise,} \end{cases}$$

and for all $n \in \mathbb{N}$, $i \in \mathbb{N}$,

$$\Phi_{in}(\omega) = \begin{cases} F(\xi_i, \omega) & \text{if } f(\xi_i, \omega) \leq 1/n, \\ +\infty & \text{otherwise.} \end{cases}$$

According to our measurability assumptions, $\Phi_{in}(\omega)$ is an extended real-valued measurable function for every $n \in \mathbb{N}$ and $i \in \mathbb{N}$.

LEMMA. Let F and f be upper semicontinuous on X for every $\omega \in \Omega$, and suppose that $\sup_n \tau_n(\omega) \geq \Phi(\omega)$ for all $\omega \in \Omega$. Then $\Phi(\omega)$ is measurable.

Proof. For all $n \in \mathbb{N}$ and $i \in \mathbb{N}$ we have $\tau_n(\omega) \leq \Phi_{in}(\omega)$, implying $\tau_n(\omega) \leq \inf_i \Phi_{in}(\omega)$, and hence $\sup_n \tau_n(\omega) \leq \sup_n \inf_i \Phi_{in}(\omega)$. By hypothesis, then,

$$(1) \quad \Phi(\omega) \leq \sup_n \inf_i \Phi_{in}(\omega).$$

To show the converse inequality we suppose first that $\Phi(\omega) < +\infty$. Then there exist points $x \in X$ satisfying $f(x, \omega) \leq 0$, and—due to the upper semicontinuity of F and f —for every such x and for all $\varepsilon > 0$, $n \in \mathbb{N}$, there exists a $\xi_i \in \Xi$ such that

$$f(\xi_i, \omega) \leq 1/n, \quad F(\xi_i, \omega) \leq F(x, \omega) + \varepsilon.$$

Hence for every $n \in \mathbb{N}$ we have $\inf_i \Phi_{in}(\omega) \leq F(x, \omega) + \varepsilon$, and therefore $\sup_n \inf_i \Phi_{in}(\omega) \leq F(x, \omega) + \varepsilon$. Since this inequality is true for all $x \in \{x | x \in X, f(x, \omega) \leq 0\}$

* Received by the editors May 14, 1974, and in revised form August 30, 1974.

† Institut für Operations Research und Mathematische Methoden der Wirtschaftswissenschaften, Universität Zürich, CH-8006 Zürich, Switzerland.

‡ Fakultät für Mathematik und Informatik, Universität Mannheim, D-6800 Mannheim, Germany.

and for every $\varepsilon > 0$, we have

$$(2) \quad \sup_n \inf_i \Phi_{in}(\omega) \leq \Phi(\omega).$$

Inequality (2) is trivially satisfied if $\Phi(\omega) = +\infty$. From (1) and (2) we obtain

$$(3) \quad \Phi(\omega) = \sup_n \inf_i \Phi_{in}(\omega).$$

Since the infimum and supremum of countably many measurable functions is again measurable, the Lemma follows. Q.E.D.

The assumption $\sup_n \tau_n(\omega) \geq \Phi(\omega)$ may be replaced by the assumption that the Kuhn–Tucker condition holds for all ω with $\Phi(\omega) < +\infty$. More precisely we have the following.

THEOREM 1. *Let F and f be upper semicontinuous on X for every $\omega \in \Omega$. Suppose that for every $\omega \in \{\omega | \Phi(\omega) < +\infty\}$ there exists a real number $u(\omega) \geq 0$ such that*

$$\Phi(\omega) \leq F(x, \omega) + u(\omega) \cdot f(x, \omega) \quad \forall x \in X \quad (\text{K.-T. condition}),$$

and suppose that for every $\omega \in \{\omega | \Phi(\omega) = +\infty\}$ we have $\sup_n \tau_n(\omega) = +\infty$. Then $\Phi(\omega)$ is measurable.

Proof. We have to show that $\sup_n \tau_n(\omega) \geq \Phi(\omega)$ for all ω satisfying $\Phi(\omega) < +\infty$. Then the result follows from the Lemma. According to the Kuhn–Tucker condition assumed, $F(x, \omega) \geq \Phi(\omega) - u(\omega) \cdot (1/n)$ for all $x \in X$ such that $f(x, \omega) \leq 1/n$. Hence $\tau_n(\omega) \geq \Phi(\omega) - u(\omega) \cdot (1/n)$, which implies

$$\sup_n \tau_n(\omega) \geq \Phi(\omega). \quad \text{Q.E.D.}$$

COROLLARY 1. *If $X = \mathbb{R}^n$, if F is convex in x for every $\omega \in \Omega$, and if $f(x, \omega) = \max_{1 \leq j \leq m} l_j(x, \omega)$, where the functions l_j are linear-affine in x , then $\Phi(\omega)$ is measurable.*

Proof. F and f are continuous in x , since they are convex over all of \mathbb{R}^n . The Kuhn–Tucker condition is satisfied, since it always holds for convex programs with only linear constraints. If the linear system $l_j(x, \omega) \leq 0$ (with $j = 1, \dots, m$) has no solution, then it is a standard result of linear programming that the system $l_j(x, \omega) \leq (1/n)$, $j = 1, \dots, m$, also has no solution for all sufficiently large $n \in \mathbb{N}$. Thus $\Phi(\omega) = +\infty$ implies $\sup_n \tau_n(\omega) = +\infty$. The assumptions of Theorem 1 are therefore satisfied. Q.E.D.

Corollary 1 implies in particular that the optimal value of a stochastic linear program is measurable.

2. The assumption made in Theorem 1 that the Kuhn–Tucker condition be satisfied for all ω with $\Phi(\omega) < +\infty$ is very restrictive, since even for convex programs the Kuhn–Tucker condition generally holds only if $\inf_{x \in X} f(x, \omega) < 0$. It is for this reason that we introduce a modified optimal value, $\Psi(\omega)$, defined as

$$\Psi(\omega) = \begin{cases} \inf_x \{F(x, \omega) | x \in X, f(x, \omega) \leq 0\} & \text{if } \inf_{x \in X} f(x, \omega) < 0, \\ \sup_n \tau_n(\omega) & \text{if } \inf_{x \in X} f(x, \omega) = 0, \\ +\infty & \text{if } \inf_{x \in X} f(x, \omega) > 0. \end{cases}$$

THEOREM 2. Let F and f be upper semicontinuous on X for every $\omega \in \Omega$. Suppose that for all $\omega \in \{\omega | \inf_X f(x, \omega) < 0\}$ there exists a real number $u(\omega) \geq 0$ such that

$$\Psi(\omega) \leq F(x, \omega) + u(\omega) \cdot f(x, \omega) \quad \forall x \in X \quad (\text{K.-T. condition}).$$

Then $\Psi(\omega)$ is measurable.

Proof. As in the proof of the Lemma we have for all $\omega \in \Omega$ that

$$\sup_n \tau_n(\omega) \leq \sup_n \inf_i \Phi_{in}(\omega).$$

If $\inf_X f(x, \omega) < 0$ we conclude from the Kuhn–Tucker condition, as in the proof of Theorem 1, that

$$\Psi(\omega) \leq \sup_n \tau_n(\omega).$$

This is also true if $\inf_X f(x, \omega) = 0$, from the definition of Ψ . If $\inf_X f(x, \omega) > 0$, then there is a real number $\rho(\omega) > 0$ such that $f(x, \omega) \geq \rho(\omega)$ for all $x \in X$, implying $\tau_n(\omega) = +\infty$ for all $n > 1/\rho(\omega)$, and thereby $\Psi(\omega) = \sup_n \tau_n(\omega) = +\infty$. Hence we have for all $\omega \in \Omega$,

$$(4) \quad \Psi(\omega) \leq \sup_n \inf_i \Phi_{in}(\omega).$$

On the other hand, for all ω satisfying $\inf_X f(x, \omega) \neq 0$, the relation

$$(5) \quad \sup_n \inf_i \Phi_{in}(\omega) \leq \Psi(\omega)$$

follows from the upper semicontinuity of F and f , as in the proof of the Lemma. Let now $\inf_X f(x, \omega) = 0$. Choose $n \in \mathbb{N}$ and $\varepsilon > 0$ arbitrarily. Then for every $x \in X$ satisfying $f(x, \omega) \leq 1/2n$ there exists, according to the upper semicontinuity of F and f , an element $\xi_i \in \Xi$ such that

$$f(\xi_i, \omega) \leq f(x, \omega) + \frac{1}{2n} \leq \frac{1}{n}, \quad F(\xi_i, \omega) \leq F(x, \omega) + \varepsilon.$$

Hence $\Phi_{in}(\omega) \leq F(x, \omega) + \varepsilon$ and $\inf_i \Phi_{in}(\omega) \leq \tau_{2n} + \varepsilon$. Since ε was arbitrary we get

$$\sup_n \inf_i \Phi_{in}(\omega) \leq \sup_n \tau_{2n}(\omega) \leq \sup_n \tau_n(\omega).$$

Since $\sup_n \tau_n(\omega) = \Psi(\omega)$ in the case under consideration, (5) again holds. In conclusion, we have from (4), (5),

$$\Psi(\omega) = \sup_n \inf_i \Phi_{in}(\omega),$$

which proves the measurability of $\Psi(\omega)$. Q.E.D.

COROLLARY 2. Let X be a convex set, and let F and f be convex functions in x for every $\omega \in \Omega$. Then $\Psi(\omega)$ is measurable.

Proof. The Kuhn–Tucker condition, as required in Theorem 2, is satisfied, since $\inf_X f(x, \omega) < 0$ is the well-known Slater-condition, the latter implying in the convex case the validity of the Kuhn–Tucker condition. The requirement of upper semicontinuity may be dropped in the convex case. Indeed, the upper semicontinuity of F (resp. f) was used only to conclude that for every $x \in X$ and

$\varepsilon > 0$ there exists $\xi_i \in \Xi$ satisfying

$$(6) \quad F(\xi_i, \omega) \leq F(x, \omega) + \varepsilon.$$

In the convex case, the same conclusion may be reached as follows. Let z be an arbitrary point in the relative interior of X . Then $x_\lambda = x + \lambda(z - x)$ with $0 < \lambda \leq 1$ is also in the relative interior of X . Since F is convex in x we have

$$F(x_\lambda, \omega) \leq F(x, \omega) + \lambda(F(z, \omega) - F(x, \omega)).$$

Choose $\lambda > 0$ so small that

$$(7) \quad F(x_\lambda, \omega) \leq F(x, \omega) + \varepsilon/2.$$

Since x_λ is in the relative interior of X , since F —as a convex function—is continuous in the relative interior of its domain X , and since Ξ is dense in X , there exists $\xi_i \in \Xi$ such that

$$(8) \quad |F(\xi_i, \omega) - F(x_\lambda, \omega)| \leq \varepsilon/2.$$

From (7) and (8) follows (6). Q.E.D.

3. We would like to point out that Corollary 1 could also be derived from Theorem 2 instead of from Theorem 1, since it may be shown under the assumptions of Corollary 1 that Φ and Ψ coincide. Under the weaker assumptions of Corollary 2, however, Φ and Ψ may differ, as the following examples show. Choose

$$X = \{(x_1, x_2) | x_1 \geq 1\} \subset \mathbb{R}^2, \quad F(x_1, x_2, \omega) = x_2, \quad f(x_1, x_2, \omega) = (x_2)^2/x_1.$$

Then $\Phi(\omega) = 0$, but $\Psi(\omega) = \sup_n \tau_n(\omega) = -\infty$. To take another example, let

$$X = [0, 1] \subset \mathbb{R}^1, \quad F(x, \omega) \equiv 0, \quad f(0, \omega) = 1, \quad f(x, \omega) = x^2 \quad \text{for } x > 0.$$

Then $\Phi(\omega) = +\infty$, but $\Psi(\omega) = 0$. A further comparison of Φ and Ψ therefore seems appropriate.

THEOREM 3. *If X is compact, and F and f are lower semicontinuous in x for every $\omega \in \Omega$, then $\Phi(\omega) = \Psi(\omega)$.*

Proof. We have to show that if $\inf_X f(x, \omega) = 0$, then $\sup_n \tau_n(\omega) = \Phi(\omega)$. Obviously $\sup_n \tau_n(\omega) \leq \Phi(\omega)$. On the other hand, by lower semicontinuity and compactness, for every $n \in \mathbb{N}$ there exists $x_n \in X$ satisfying

$$f(x_n, \omega) \leq 1/n, \quad F(x_n, \omega) = \tau_n(\omega).$$

Let $\{x_{n_j}\}$ be a subsequence converging to some $x_0 \in X$. Then, by lower semicontinuity and by the monotonicity of $\tau_n(\omega)$,

$$f(x_0, \omega) \leq \liminf_{j \rightarrow \infty} f(x_{n_j}, \omega) \leq 0, \quad F(x_0, \omega) \leq \liminf_{j \rightarrow \infty} F(x_{n_j}, \omega) \leq \sup_n \tau_n(\omega),$$

implying $\sup_n \tau_n(\omega) \geq \Phi(\omega)$. Q.E.D.

Combining Theorems 2 and 3 one can derive measurability statements for $\Phi(\omega)$. In particular, we obtain very easily the following result which is contained in [2, Cor. 4.3].

COROLLARY 3. *Let X be a closed convex set, and let F and f be lower semicontinuous convex functions on X for every $\omega \in \Omega$. Then $\Phi(\omega)$ is measurable.*

Proof. For all $k \in \mathbb{N}$ denote by $\Phi_k(\omega)$ [resp. $\Psi_k(\omega)$] the functions which are obtained if in the definition of Φ [resp. Ψ] we replace X by the compact subset $X_k \equiv \{x | x \in X, \|x\| \leq k\}$. By Corollary 2, $\Psi_k(\omega)$ is measurable. By Theorem 3, Φ_k equals Ψ_k , hence is measurable. The measurability of Φ follows since obviously

$$\Phi(\omega) = \inf_k \Phi_k(\omega). \quad \text{Q.E.D.}$$

Acknowledgment. The authors are indebted to R. T. Rockafellar for some clarifying remarks.

REFERENCES

- [1] P. KALL, *Some remarks on the distribution problem of stochastic linear programming*, Methods of Operations Research, 16 (1973), pp. 189–196.
- [2] R. T. ROCKAFELLAR, *Measurable dependence of convex sets and functions on parameters*, J. Math. Anal. Appl., 28 (1969), pp. 4–25.

MARTINGALES ON JUMP PROCESSES. I: REPRESENTATION RESULTS*

R. BOEL, P. VARAIYA AND E. WONG†

Abstract. The paper is a contribution to the theory of martingales of processes whose sample paths are piecewise constant and have finitely many discontinuities in a finite time interval. The assumption is made that the jump times of the underlying process are totally inaccessible and necessary and sufficient conditions are given for this to be true. It turns out that all martingales are then discontinuous, and can be represented as stochastic integrals of certain basic martingales. This representation theorem is used in a companion paper to study various practical problems in communication and control. The results in the two papers constitute a sweeping generalization of recent work on Poisson processes.

1. Introduction and summary. The theory of martingales has proved to be successful as a framework for formulating and analyzing many issues in stochastic control, and in detection and filtering problems [2], [4], [5], [10], [11], [12], [19], [32], [33], [34]. Three sets of results in the abstract or general theory of martingales seem to be the most useful ones in these applications. The first set consists of the optional sampling theorem and the classical martingale inequalities [17]. The second set consists of the locus of results culminating in the decomposition theorem for supermartingales [24]. The third set includes the calculus of stochastic integrals [16], [22], the differentiation formula and its application to the so-called “exponentiation formula” [15].

In applications one is concerned with martingales which are functionals of a basic underlying process such as a Wiener or Poisson process, and in order to use the abstract theory one needs to know how to represent these martingales usefully and explicitly in terms of the underlying process. Thus the “martingale representation theorems” serve as a bridge linking the abstract theory and the concrete applications. Their role is quite analogous to that of matrix representations of linear operators which serve as the instrument with which one can apply the abstract theory of linear algebra.

The most familiar of all the basic processes which can arise in practice is the Wiener process. It is known that every martingale of a Wiener process can be represented as a stochastic integral of the Wiener process [6], [22]. This fundamental representation theorem, together with the exponentiation formula, has been used to derive solutions of stochastic differential equations [2], [19], [20], to obtain recursive equations for filters [5], [21], [30], [31] and the likelihood ratios for some detection problems [10], [18], to mention just a few applications. These very results combined with the decomposition theorem for supermartingales form the foundation of an approach to one family of stochastic optimal control

* Received by the editors January 12, 1973, and in revised form August 13, 1974.

† Department of Electrical Engineering and Computer Sciences and the Electronics Research Laboratory, University of California, Berkeley, California 94720. This research was sponsored by the U.S. Army Research Office—Durham, under Contract DAHC 04-67-C-0046 and the National Science Foundation under Grant GK-10656X3.

problems [12]. It turns out that every martingale of a Wiener process has continuous sample paths. This is fortunate because it implies that the martingale is locally square integrable, and hence most of the questions about martingales can be posed within the Hilbert space structure of the space of square integrable random variables.

However, for many processes, e.g., Poisson process, one can have martingales which are not locally square integrable. As Meyer and his co-workers have pointed out [16], [26] the L^2 structure is no longer appropriate and one needs to be more careful in defining stochastic integrals and in obtaining the differentiation formula. Indeed the current theory of stochastic integration with respect to such martingales is still not completely satisfactory.

This paper is a contribution to the abstract theory and to its applications for the relatively simple case where the sample functions of the underlying process are step functions which have only a finite number of jumps in every finite time interval. In a sense this is the polar opposite of the Wiener process case since all the martingales are discontinuous, that is, all the continuous martingales have constant sample paths. The most important special cases covered by this paper include the Poisson process, Markov chains and extensions of these, such as processes arising in queueing theory. To some extent the results for some of these special cases are also covered in [4], [5], [10], [11], [29], [30], [31].

The next section gives a precise definition of the underlying process and exhibits some of the important properties of the generated σ -fields. Conditions are derived which guarantee that the jump times of the process are totally inaccessible stopping times. These preliminary results are used in § 3 to show first that there are no nonconstant continuous martingales and then to obtain an integral representation of all martingales. A particular example, which includes most of the special cases mentioned above, is presented in § 4. Applications of the results are given in the companion paper [3].

2. The basic process and its stopping times. Let (Z, \mathcal{Z}) be a Blackwell space, that is a measurable space such that \mathcal{Z} is a separable σ -field and every measurable function $f: Z \rightarrow R$ maps Z onto an analytic subset of R (see [24, p. 61]). Let Ω be a family of functions on $R^+ = [0, \infty)$ with values in Z , such that each $\omega \in \Omega$ is a step function with only a finite number of jumps in every finite interval, and such that for all $\omega \in \Omega$, $t \in R^+$, $\omega(t) = \omega(t + \varepsilon)$ for all ε less than some sufficiently small $\varepsilon_0 > 0$. If Z is also a topological space, then each function ω is right-continuous and has left-hand limits. Let x_t be the evaluation process on Ω , i.e., $x_t(\omega) = \omega(t)$, $t \in R^+$. Let \mathcal{F}_t be the σ -field on Ω generated by sets of the form $\{x_s \in B\}$, $B \in \mathcal{Z}$, $s \leq t$. Let $\mathcal{F} = \bigvee_{t \in R^+} \mathcal{F}_t$.¹

Because the positive rationals are dense in R^+ , it is clear that \mathcal{F} can also be written as $V_r \sigma(x_{r_n})$, where $\sigma(x_{r_n})$ is the σ -field generated by the function x_{r_n} and r_n is rational. Hence the separability of \mathcal{Z} implies the separability of \mathcal{F} . Moreover, as will be shown, every real-valued \mathcal{F} -measurable function on Ω will map Ω onto an analytic subset, hence (Ω, \mathcal{F}) is a Blackwell space. The assertion follows from considering approximations for any measurable $f: \Omega \rightarrow R$ of the form

¹ If A_x is a family of subsets then $V_x A_x$ denotes the smallest σ -field containing all the A_x .

$f^n = g^n \cdot h^n \cdot i$, where $i: (\Omega, \mathcal{F}) \rightarrow (Z^{\mathbb{N}}, \mathcal{Z}^{\mathbb{N}})$ is the natural isomorphism (\mathbb{N} is the set of natural numbers), and $h^n: (Z^{\mathbb{N}}, \mathcal{Z}^{\mathbb{N}}) \rightarrow (R^{\mathbb{N}}, \mathcal{B}^{\mathbb{N}})$ (\mathcal{B} is the Borel field on R) consists of measurable components h_1^n, h_2^n, \dots with $h^n(z_1, z_2, \dots) = (h_1^n(z_1), h_2^n(z_2), \dots)$, and finally g^n is a measurable mapping from $(R^{\mathbb{N}}, \mathcal{B}^{\mathbb{N}})$ into (R, \mathcal{B}) . Since the Cartesian product of analytic sets is analytic (see [1]), the image of $Z^{\mathbb{N}}$ in $R^{\mathbb{N}}$ under h^n is an analytic set which is in turn mapped into an analytic subset of R by g^n . Since analytic sets form a class closed under countable unions and intersections, this limiting procedure shows that every measurable function $f: \Omega \rightarrow R$ maps Ω onto an analytic set. Since (Ω, \mathcal{F}) is a Blackwell space it follows from [24, § II, Thm. 16] that (Ω, \mathcal{F}) is isomorphic to $(A, \mathcal{B}(A))$ where A is an analytic subset of R . Hence the results of [28] can be applied without assuming a topological structure on Z itself.

A Z -valued or $R \cup \{\infty\}$ -valued function f on Ω is a *random variable* (r.v.) if $f^{-1}(B) \in \mathcal{F}$ whenever $B \in \mathcal{Z}$ or whenever B is a Borel subset of $R \cup \{\infty\}$. Unless otherwise stated a r.v. is $R \cup \{\infty\}$ -valued. A nonnegative r.v. T is said to be a *stopping time* (s.t.) if for every $t \in R^+$, $\{T \leq t\} \in \mathcal{F}_t$. If T is a s.t., then \mathcal{F}_T consists of those sets $A \in \mathcal{F}$ for which $A \cap \{T \leq t\} \in \mathcal{F}_t$ for each $t \in R_+$, whereas \mathcal{F}_T^- is the σ -field generated by \mathcal{F}_0 and sets of the form $A \cap \{t < T\}$, where $A \in \mathcal{F}_t$, and finally $\mathcal{F}_{T^+} = \bigcap_{n>0} \mathcal{F}_{T+1/n}$.

Define inductively the functions T_n :

$$T_0 \equiv 0, \quad T_{n+1}(\omega) = \inf \{t | t \geq T_n(\omega) \text{ and } x_t(\omega) \neq x_{T_n(\omega)}(\omega)\},$$

where the infimum over an empty set is taken to be $+\infty$. The next few results characterize the σ -field \mathcal{F}_t and demonstrate that the T_n are indeed s.t.s. The key results, Corollary 2.2 and Proposition 2.3, which are the only ones used subsequently, can in fact be proved from first principles assuming only the separability of \mathcal{Z} , but it is much more intuitive and easier to rely on the results of [7] and [28].

Let $H: \Omega \rightarrow [0, \infty]$ be any function. Then H defines three equivalence relations on Ω as follows:

$$\omega \stackrel{H}{\sim} \omega' \Leftrightarrow H(\omega) = H(\omega') \text{ and } x_t(\omega) = x_t(\omega') \text{ for } t \leq H(\omega),$$

$$\omega \stackrel{H^+}{\sim} \omega' \Leftrightarrow H(\omega) = H(\omega') \text{ and there is } \varepsilon > 0 \text{ such that } x_t(\omega) = x_t(\omega')$$

$$\text{for } t \leq H(\omega) + \varepsilon,$$

$$\omega \stackrel{H^-}{\sim} \omega' \Leftrightarrow H(\omega) = H(\omega') \text{ and } x_t(\omega) = x_t(\omega') \text{ for } t < H(\omega).$$

A set $A \subset \Omega$ is said to be *saturated* for H , respectively H^+ , H^- , if $\omega \in A$, and $\omega \stackrel{H}{\sim} \omega'$, respectively $\omega \stackrel{H^+}{\sim} \omega'$, $\omega \stackrel{H^-}{\sim} \omega'$, implies $\omega' \in A$. Let \mathcal{S}_H , \mathcal{S}_{H^+} , \mathcal{S}_{H^-} denote the family of subsets of Ω which are saturated for H , H^+ , H^- respectively.

PROPOSITION 2.1. $\mathcal{F}_t = \mathcal{S}_t \cap \mathcal{F}$, where $\mathcal{S}_t = \mathcal{S}_H$ for $H \equiv t$.

Proof. The proof follows from [28, Prop. 1]. \square

COROLLARY 2.1. A nonnegative r.v. T is a s.t. if and only if $\{T \leq t\} \in \mathcal{S}_t$ for all $t \in R_+$.

COROLLARY 2.2. T_n is a s.t. for all n .

Proof. T_n is obviously a nonnegative r.v. and $\{T_n \leq t\} \in \mathcal{S}_t$ by definition. \square

PROPOSITION 2.2. *Let T be a s.t. Then*

$$\mathcal{F}_T = \mathcal{S}_T \cap \mathcal{F}, \quad \mathcal{F}_{T^+} = \mathcal{S}_{T^+} \cap \mathcal{F}, \quad \mathcal{F}_{T^-} = \mathcal{S}_{T^-} \cap \mathcal{F}.$$

Proof. This follows from [28, Props. 1, 2]. \square

For a s.t. T , $\mathcal{F}_\infty(x_{t \wedge T})$ denotes the σ -field generated by the Z -valued r.v.s $X_{t \wedge T}$, $t \in R_+$. (If S, T are r.v.s, then $S \wedge T = \{\min S, T\}$.)

PROPOSITION 2.3. *Let T be a s.t. Then $\mathcal{F}_T = \mathcal{F}_\infty(x_{t \wedge T})$.*

Proof. First of all since every measurable set generated by $x_{t \wedge T}$ clearly belongs to \mathcal{S}_T , it follows that $\mathcal{F}_\infty(x_{t \wedge T}) \subset \mathcal{S}_T \cap \mathcal{F} = \mathcal{F}_T$. To prove the reverse inclusion, we begin by noting that $\mathcal{F}_\infty(x_{t \wedge T})$ is separable by the same argument which was used to show that $\mathcal{F} = \mathcal{F}_\infty(x_t)$ is separable. By [24, § III, Thm. 17] it follows that $\mathcal{F}_T \subset \mathcal{F}_\infty(x_{t \wedge T})$ since the two families have the same atoms, namely, $\bigcap_n \{x_{r_n \wedge T} \in B_n\}$, where r_n is rational and B_n is an atom of \mathcal{Z} . \square

COROLLARY 2.3. $\mathcal{F}_{T_n} = \sigma(x_{T_i}, T_i; 0 \leq i \leq n)$.

Proof. The proof follows from Proposition 2.3 since

$$\begin{aligned} \mathcal{F}_\infty(x_{t \wedge T_n}) &= \sigma(x_{T_i \wedge T_n}, T_i \wedge T_n; 0 \leq i \leq \infty) \\ &= \sigma(x_{T_i}, T_i; 0 \leq i \leq n). \end{aligned}$$

COROLLARY 2.4. $\mathcal{F}_t = \sigma(x_{T_i \wedge t}, T_i \wedge t, 0 \leq i < \infty)$.

COROLLARY 2.5. *Let T be a s.t. Then $\mathcal{F}_{T^+} = \mathcal{F}_T$.*

Proof. Since the sample functions are piecewise constant and $\omega(t) = \omega(t^+)$, it follows that $\mathcal{S}_T = \mathcal{S}_{T^+}$ and then the result follows from Proposition 2.2. \square

PROPOSITION 2.4. $\mathcal{F}_{T_{n-}} = \sigma(x_{T_i}, T_{i+1}, 0 \leq i \leq n-1)$.

Proof. This proof is similar to the proof of Proposition 2.3, with both σ -fields having the atoms $\{x_{T_i} \in A_i, T_{i+1} \in B_i; 0 \leq i \leq n-1\}$, where A_i is an atom of Z and B_i is an atom of R .

PROPOSITION 2.5. *Let $n \geq 1$, and $\delta > 0$. Let $T = (T_{n-1} + \delta) \wedge T_n$, and let $A \in \mathcal{F}_T$. Then there exists $A^0 \in \mathcal{F}_{T_{n-1}}$ such that $A \cap \{T < T_n\} = A^0 \cap \{T < T_n\}$.*

Proof. By Proposition 2.3, $\mathcal{F}_T = \mathcal{F}_\infty(x_{t \wedge T})$ and it is easy to see that the latter coincides with the σ -field generated by the r.v.s $\{x_{T_i \wedge T}, T_i \wedge T; i = 0, 1, 2, \dots\}$. Hence there exists a function g measurable in its arguments such that

$$\begin{aligned} I_A(\omega) &= g(x_{T_0 \wedge T}(\omega), T_0 \wedge T(\omega), \dots, x_{T_{n-1} \wedge T}(\omega), T_{n-1} \wedge T(\omega), x_{T_n \wedge T}(\omega), \\ &\quad T_n \wedge T(\omega), \dots) \\ &= g(x_{T_0}(\omega), T_0(\omega), \dots, x_{T_{n-1}}(\omega), T_{n-1}(\omega), x_{T_n \wedge T}(\omega), T_n \wedge T(\omega), \dots). \end{aligned}$$

Define the measurable function g^0 by

$$\begin{aligned} g^0(x_0, t_0, \dots, x_{n-1}, t_{n-1}) &= g(x_0, t_0, \dots, x_{n-1}, t_{n-1}, x_{n-1}, t_{n-1} + \delta, x_{n-1}, \\ &\quad t_{n-1} + \delta, \dots). \end{aligned}$$

Now if $T_{n-1} \leq T < T_n$, then $x_{T_{n+k} \wedge T}(\omega) = x_{T_{n-1}}(\omega)$ and $T_{n+k} \wedge T(\omega) = T_{n-1}(\omega) + \delta$ for all $k \geq 0$. Therefore,

$$I_A(\omega) I_{\{T < T_n\}}(\omega) = g^0(x_{T_0}(\omega), T_0(\omega), \dots, x_{T_{n-1}}(\omega), T_{n-1}(\omega)) (I_{\{T < T_n\}}(\omega)),$$

so that the set $A^0 = \{\omega | g^0(x_{T_0}(\omega), \dots, T_{n-1}(\omega)) = 1\}$ satisfies the assertion. \square

LEMMA 2.1. *Let $n \geq 1$, and let S be a s.t. Then there exists a r.v.f, measurable with respect to $\mathcal{F}_{T_{n-1}}$ such that $SI_{\{S < T_n\}} = fI_{\{S < T_n\}}$.*

Proof. $SI_{\{S < T_n\}} = SI_{\{S < T_{n-1}\}} + SI_{\{T_{n-1} \leq S < T_n\}}$, and $SI_{\{S < T_{n-1}\}}$, $I_{\{S < T_{n-1}\}}$ are $\mathcal{F}_{T_{n-1}}$ -measurable so that by replacing S by $S \vee T_{n-1}$ if necessary, one can assume that $S \geq T_{n-1}$. Let $\Gamma = \{S < T_n\}$. Then $\Gamma = \bigcup_m \Gamma_m$, where

$$\Gamma_m = \bigcup_k \{S \leq T_{n-1} + k2^{-m}\} \cap \{T_{n-1} + k2^{-m} < T_n\}.$$

Fix $\delta = 2^{-m}$. By Proposition 2.5 there exist sets $A_k \in \mathcal{F}_{T_{n-1}}$ such that

$$\{S \leq T_{n-1} + k\delta\} \cap \{T_{n-1} + k\delta < T_n\} = A_k \cap \{T_{n-1} + k\delta < T_n\}, \quad k \geq 1.$$

Define sets B_k by

$$B_1 = A_1 \quad \text{and} \quad B_k = \{\omega \in A_k | \omega \notin A_i \text{ for } i < k\} \quad \text{for } k > 1,$$

and then define the function $f_m: \Omega \rightarrow [0, \infty]$ by

$$f_m(\omega) = T_{n-1} + k\delta \quad \text{if } \omega \in B_k \quad \text{and} \quad f_m(\omega) = T_{n-1}(\omega) \quad \text{if } \omega \notin \bigcup_k B_k.$$

Certainly f_m is $\mathcal{F}_{T_{n-1}}$ -measurable. Also

$$(2.1) \quad f_m(\omega) - \delta \leq S(\omega) \leq f_m(\omega) < T_n(\omega) \quad \text{for } \omega \in \Gamma_m.$$

To see this note first that if $\omega \in A_1 \cup \{T_{n-1} + \delta < T_n\}$, then clearly $T_{n-1}(\omega) = f_m(\omega) - \delta \leq S(\omega) < f_m(\omega) < T_n(\omega)$. Next, as an induction hypothesis, suppose that the inequalities in (2.1) hold for

$$\omega \in \bigcup_{k=1}^N A_k \cap \{T_{n-1} + k\delta < T_n\},$$

and let

$$(2.2) \quad \omega \in A_{N+1} \cap \{T_{n-1} + (N+1)\delta < T_n\}, \quad \omega \notin \bigcup_{k=1}^N A_k \cap \{T_{n-1} + k\delta < T_n\}.$$

Let $k \leq N+1$ be the smallest integer such that $\omega \in B_k$. Suppose $k \leq N$. Then, since $B_k \subset A_k$, and since from (2.2), $T_n > T_{n-1} + k\delta$, it follows that

$$\omega \in A_k \cap \{T_{n-1} + k\delta < T_n\}$$

which contradicts the second condition of (2.2). Hence $\omega \in B_{N+1}$ and so $T_{n-1}(\omega) + N\delta \leq S(\omega) \leq T_{n-1}(\omega) + (N+1)\delta = f_m(\omega) < T_n(\omega)$. Therefore (2.1) holds by induction. Finally, define the $\mathcal{F}_{T_{n-1}}$ -measurable function f by $f(\omega) = \lim_m \inf f_m(\omega)$. The obvious inclusion $\Gamma_m \subset \Gamma_{m+1}$ implies that if $\omega \in \Gamma_m$, then $f_{m+k}(\omega) - 2^{-(m+k)} \leq S(\omega) \leq f_{m+k}(\omega)$ for all $k \geq 0$. Hence $f(\omega) = S(\omega)$ and the assertion is proved. \square

To prove further it is convenient to introduce a probability measure on $(\Omega, \mathcal{F})^2$. Throughout this paper let P denote a fixed probability measure on (Ω, \mathcal{F}) . Recall the following important classification of stopping times [25].

² It may be of interest to note that Lemmas 2.2, 2.3 and 2.4 below can be proved without imposing a probability measure P by using the algebraic definition of a predictable s.t. of [28]. Then a predictable s.t. in the sense used here is simply a nonnegative r.v. which is a.s. P equal to a predictable s.t. in the sense of [28].

Let T be a s.t. T is said to be *totally inaccessible* if $T > 0$ a.s. and if for every increasing sequence of s.t.s $S_1 \leq S_2 \leq \dots$,

$$P\{S_k(\omega) < T(\omega) \text{ for all } k \text{ and } \lim_{k \rightarrow \infty} S_k(\omega) = T(\omega) < \infty\} = 0;$$

whereas T is said to be *predictable* if there exists an increasing sequence of s.t.s $S_1 \leq S_2 \leq \dots$ such that

$$P\{T = 0, \text{ or } S_k < T \text{ for all } k \text{ and } \lim_{k \rightarrow \infty} S_k = T\} = 1.$$

The next three lemmas relate this classification to the properties of the jump times T_n of the process x .

LEMMA 2.2. *Let T be a totally inaccessible s.t. Then*

$$TI_{\{T < \infty\}} = \left[\sum_{n=1}^{\infty} T_n I_{\{T = T_n\}} \right] I_{\{T < \infty\}} \quad a.s.$$

Proof. The equality above holds if and only if $P\{T_{n-1} < T < T_n\} = 0$ for each $n \geq 1$. Let n be fixed. By Lemma 2.1 there exists a $\mathcal{F}_{T_{n-1}}$ -measurable function f such that $f(\omega) = T(\omega)$ for $\omega \in \{T_{n-1} < T < T_n\}$. Let $S_k = T_{n-1} \vee (f - 1/k)$. Then $S_k \geq T_{n-1}$ and S_k is $\mathcal{F}_{T_{n-1}}$ -measurable so that it is a s.t. Also S_k is increasing and clearly

$$\{T_{n-1} < T < T_n\} \subset \left\{ S_k < T \text{ for all } k \text{ and } \lim_{k \rightarrow \infty} S_k = T < \infty \right\}.$$

Since T is totally inaccessible, the set on the right has probability measure zero. The assertion is proved. \square

LEMMA 2.3. *Let T be a s.t. such that for all $n \geq 1$, $P\{T = T_n < \infty\} = 0$. Then T is predictable.*

Proof. Let h be a function measurable in its arguments and taking values in the set $\{0, 1\}$ such that the process $I_{T \leq t}$ has the representation

$$I_{T \leq t} = h(t, x_{T_0 \wedge t}, T_0 \wedge t, \dots, x_{T_n \wedge t}, T_n \wedge t, \dots).$$

Since $I_{T \leq t} = \max_{s \leq t} I_{T \leq s}$, by modifying h if necessary it can be assumed that

$$h(t, \xi) = \max_{s \leq t} h(s, \xi).$$

The r.v. $(h(t + \varepsilon), x_{T_0 \wedge t}, T_0 \wedge t, \dots)$ is \mathcal{F}_t -measurable and so the r.v.

$$T_\varepsilon(\omega) = \inf \{t | h(t + \varepsilon, x_{T_0 \wedge t}, T_0 \wedge t, \dots) = 1\}$$

is a s.t., and it is immediate that for $\varepsilon > 0$,

$$T_\varepsilon(\omega) < T(\omega) \quad \text{for } \omega \in \{0 < T < \infty\}.$$

Furthermore $T_\varepsilon \leq T_{\varepsilon'}$ if $\varepsilon' \leq \varepsilon$. Define the s.t.s S_k by $S_k = T_{1/k} \wedge k$. It will now be shown that

$$\lim_{k \rightarrow \infty} S_k(\omega) = T(\omega) \quad \text{for } \omega \in \bigcup_{n=1}^{\infty} \{T_{n-1} < T < T_n\}.$$

Let $\omega \in \{T_{n-1} < T < T_n\}$. Then

$$\begin{aligned} & h(t, x_{T_0 \wedge t}(\omega), T_0 \wedge t(\omega), \dots, x_{T_n \wedge t}(\omega), T_n \wedge t(\omega) \dots) \\ &= \begin{cases} 0 & \text{for } T_{n-1}(\omega) < t < T(\omega), \\ 1 & \text{for } t \geq T(\omega), \end{cases} \end{aligned}$$

so that

$$\begin{aligned} & h\left(t + \frac{1}{k}, x_{T_0 \wedge t}(\omega), T_0 \wedge t(\omega), \dots\right) \\ &= \begin{cases} 0 & \text{for } T_{n-1}(\omega) < t + (1/k) < T(\omega) \text{ or } T_{n-1}(\omega) < t < T(\omega), \\ 1 & \text{for } t \geq T(\omega). \end{cases} \end{aligned}$$

Hence $T_{1/k}(\omega) = T(\omega) - 1/k$ for $1/k < T(\omega) - T_{n-1}(\omega)$. It follows that $S_k(\omega)$ converges to $T(\omega)$ and the assertion follows. \square

LEMMA 2.4. T_n is totally inaccessible if and only if for every $\mathcal{F}_{T_{n-1}}$ -measurable function f , $P\{T_n = f < \infty\} = 0$.

Proof. Suppose $P\{T_n = f < \infty\} > 0$. Let $S_k = T_{n-1} \vee (f - 1/k)$. Then S_k is an increasing sequence of s.t.s and

$$\{T_n = f < \infty\} \subset \left\{ S_k < T_n \text{ for all } k \text{ and } \lim_{k \rightarrow \infty} S_k = T_n < \infty \right\}$$

so that T_n cannot be totally inaccessible thereby proving necessity. To prove sufficiency suppose that T_n is not totally inaccessible so that there is an increasing sequence of s.t.s S_k such that

$$(2.3) \quad P\{\Gamma\} = P\left\{ S_k < T_n \text{ for all } k \text{ and } \lim_{k \rightarrow \infty} S_k = T_n < \infty \right\} > 0.$$

By Lemma 2.1 there exist functions f_k , measurable with respect to $\mathcal{F}_{T_{n-1}}$, such that $S_k(\omega) = f_k(\omega)$ for $\omega \in \{S_k < T_n\}$. Let $f = \liminf f_k$. Then from (2.3) it follows that $f(\omega) = T_n(\omega)$ for $\omega \in \Gamma$ so that $P\{f = T_n < \infty\} > 0$ and sufficiency is proved. \square

From the lemma above the following intuitive sufficient condition follows immediately.

THEOREM 2.1. Let $F(t_n|x_0, t_0, \dots, x_{n-1}, t_{n-1})$ be the conditional probability distribution of T_n given $x_{T_0}, T_0, \dots, x_{T_{n-1}}, T_{n-1}$. Suppose that F is continuous in t_n for all values of $(x_0, t_0, \dots, x_{n-1}, t_{n-1})$. Then T_n is totally inaccessible.³

As an application of Theorem 2.1 note that if x_t is a Poisson process, then $F(t_n|x_0, t_0, \dots, x_{n-1}, t_{n-1}) = (1 - \exp - (t_n - t_{n-1}))I_{t_n \geq t_{n-1}}$ is continuous. Hence the jump times of a Poisson process are totally inaccessible.

3. The martingale representation theorem. It will be necessary from now on to complete the σ -fields \mathcal{F}_t and \mathcal{F} with respect to the measure P . An additional condition is also imposed.

Assumptions. (i) The σ -fields $\mathcal{F}_t, \mathcal{F}$ are augmented so as to be complete with respect to P . (ii) The stopping times T_n are totally inaccessible for $n \geq 1$.

³ If Z is a Borel subset of \mathcal{R}^p and \mathcal{Z} contains all Borel subsets of Z , then the conditional probability F exists by [23, p. 361].

Note that after completion of the space (Ω, \mathcal{F}) it ceases to be a Blackwell space. But, of course, the results of § 2 continue to hold if the relevant equalities are interpreted as being true almost surely P .

The family \mathcal{F}_t is said to be *free of times of discontinuity* if for every increasing sequence of s.t.s S_k , $\mathcal{F}_{\lim S_k} = \bigvee_k \mathcal{F}_{S_k}$.

PROPOSITION 3.1. *The family \mathcal{F}_t is free of times of discontinuity.*

Proof. By Lemma 2.2 and Assumption (ii) a s.t. T is totally inaccessible if and only if its graph⁴ $[T]$ is contained in the union $\bigcup_n [T_n]$ of the graphs of T_n , whereas by Lemma 2.3, T is predictable if $[T] \cap \bigcup_n [T_n] = \emptyset$. The assertion follows from [14, § III, Thm. 51, p. 62]. \square

It will be useful to recall some definitions at this time. This will be followed by some remarks and a reproduction of some known results which will be used in the discussion to follow.

A process y_t is said to be *adapted* (to the family \mathcal{F}_t) if y_t is \mathcal{F}_t -measurable for all t . Two processes y_t and y'_t are said to be *indistinguishable*, and are written $y_t \equiv y'_t$, if for almost all ω , $y_t(\omega) = y'_t(\omega)$ for all $t \in R_+$.

Let m_t be a martingale with respect to $(\Omega, \mathcal{F}_t, P)$. It is said to be *uniformly integrable* (u.i.), and one writes $m_t \in \mathcal{M}^1$, if $\{m_t | t \in R_+\}$ is a u.i. set of r.v.s. It is said to be *square integrable* (s.i.), and one writes $m_t \in \mathcal{M}^2$, if $\sup \{Em_t^2 | t \in R_+\} < \infty$.

Let m_t be a process. It is said to be a *locally integrable martingale* [locally square integrable martingale], and one writes $m_t \in \mathcal{M}_{\text{loc}}^1$ [$m_t \in \mathcal{M}_{\text{loc}}^2$], if there is an increasing sequence of s.t.s S_k with $S_k \rightarrow \infty$ a.s. such that for each k ,

$$m_t \wedge S_k I_{\{S_k > 0\}} \in \mathcal{M}^1 [m_t \wedge S_k I_{\{S_k > 0\}} \in \mathcal{M}^2].$$

An adapted process a_t is said to be an *increasing process* if $a_0 = 0$ and if its sample paths are nondecreasing and right-continuous. It is said to be *integrable*, and one writes $a_t \in \mathcal{A}^+$ if $\sup \{Ea_t | t \in R_+\} < \infty$. $\mathcal{A}_{\text{loc}}^+$ is defined in a manner analogous to the previous definition. Finally let $\mathcal{A} = \mathcal{A}^+ - \mathcal{A}^+ = \{a_t - a'_t | a_t \in \mathcal{A}^+\}$ and $\mathcal{A}_{\text{loc}} = \mathcal{A}_{\text{loc}}^+ - \mathcal{A}_{\text{loc}}^+$.

It will be assumed throughout that all the local martingales have sample paths which are right-continuous and have left-hand limits. It is known that since the σ -fields \mathcal{F}_t are complete and since by Corollary 2.5, $\mathcal{F}_{t+} = \mathcal{F}_t$ for all $t \in R_+$, therefore one can always choose a modification of a local martingale so that its sample paths have the above mentioned property (see [24, § VI, Thm. 4]). Two modifications with this property are indistinguishable.

It can be immediately verified that $\mathcal{M}^2 \subset \mathcal{M}^1$ and so $\mathcal{M}_{\text{loc}}^2 \subset \mathcal{M}_{\text{loc}}^1$, and if $m_t \in \mathcal{M}^1$ has continuous sample paths, then $m_t \in \mathcal{M}_{\text{loc}}^2$. However if the sample paths of $m_t \in \mathcal{M}^1$ are not continuous, then m_t may not belong to $\mathcal{M}_{\text{loc}}^2$. Thus in dealing with discontinuous martingales one may be unable to use the Hilbert space structure of square integrable r.v.s.

The next result follows from Proposition 3.1 and [22, Thm. 1.1].

THEOREM 3.1. *Let m_t and m'_t be in $\mathcal{M}_{\text{loc}}^2$. Then there exists a unique,⁵ continuous process $\langle m, m' \rangle_t \in \mathcal{A}$ such that $m_t m'_t - \langle m, m' \rangle_t \in \mathcal{M}_{\text{loc}}^1$.*

⁴ $[T] = \{(\omega, T(\omega)) | \omega \in \Omega\} \subset \Omega \times [0, \infty]$.

⁵ Throughout "unique" means unique up to indistinguishability.

DEFINITION 3.1. Let $B \in \mathcal{Z}$. Let

$$P(B, t) = \sum_{s \leq t} I_{\{x_s \neq x_{s-}\}} I_{\{x_s \in B\}}$$

be the number of jumps of x which occur prior to t and which end in the set B .

PROPOSITION 3.2. *There is a unique continuous process $\tilde{P}(B, t) \in \mathcal{M}_{\text{loc}}^+$ such that the process $Q(B, t) = P(B, t) - \tilde{P}(B, t)$ is in $\mathcal{M}_{\text{loc}}^2$.*

Proof. Let $P_n(B, t) = P(B, t \wedge T_n)$. Then $P_n(B, t) \leq n$ so that it is square integrable. Furthermore the jumps of $P_n(B, t)$ occur at the s.t.s T_n , $1 \leq i \leq n$, and these s.t.s are totally inaccessible by assumption. It follows from [24, § VIII, Thm. 31] that there is a unique, continuous, integrable, increasing process $\tilde{P}_n(B, t)$ such that $Q_n(B, t) = P_n(B, t) - \tilde{P}_n(B, t) \in \mathcal{M}^2$. From this last relation and the uniqueness of \tilde{P}_n one can conclude that $\tilde{P}_{n+1}(B, t \wedge T_n) \equiv \tilde{P}_n(B, t)$, $Q_{n+1}(B, t \wedge T_n) \equiv Q_n(B, t)$. Hence the processes \tilde{P}, Q defined by

$$\tilde{P}(B, t \wedge T_n) \equiv \tilde{P}_n(B, t), \quad Q(B, t \wedge T_n) \equiv Q_n(B, t)$$

satisfy the assertion. \square

Remark. If the conditional distribution of the jump times T_{n+1} and the jumps $x_{T_{n+1}}$ given \mathcal{F}_{T_n} is available, then, following the results of [35], [36] and using Lemma 2.1, an explicit characterization of the processes $\tilde{P}(B, t)$ can be obtained. Specifically, for each $B \in \mathcal{Z}$ and integer n let

$$F_n(B, t) = P(T_{n+1} - T_n \leq t, x_{T_{n+1}} \in B | \mathcal{F}_{T_n}).$$

Then

$$\tilde{P}(B, t) = \sum_{T_i \leq t} \left[\int_0^{T_i - T_{i-1}} \frac{F_i(B, ds)}{1 - F_i(Z, s)} \right] + \int_0^{t - T_n} \frac{F_n(B, ds)}{1 - F_n(Z, s^-)}.$$

It follows from this result that $\tilde{P}(B, t)$ is continuous (absolutely continuous) in t if for each n , $F_n(B, t)$ is continuous (absolutely continuous) in t . (Compare Theorem 2.1.)

Two processes m_t, m'_t in $\mathcal{M}_{\text{loc}}^2$ are said to be *orthogonal* if $m_t m'_t \in \mathcal{M}_{\text{loc}}^2$ or equivalently if $\langle m, m' \rangle_t \equiv 0$.

LEMMA 3.1. *Let $B_i \in \mathcal{Z}$, $i = 1, 2$. Then $Q(B_1, t)Q(B_2, t) - \tilde{P}(B_1 \cap B_2, t) \in \mathcal{M}_{\text{loc}}^1$, i.e., $\langle Q(B_1, \cdot), Q(B_2, \cdot) \rangle_t \equiv \tilde{P}(B_1 \cap B_2, t)$. In particular, $Q(B_1, t)$ and $Q(B_2, t)$ are orthogonal if $B_1 \cap B_2 = \emptyset$.*

Proof.

$$Q(B_1, t \wedge T_n) = Q(B_1 \cap B_2, t \wedge T_n) + Q(B_1 - B_2, t \wedge T_n)$$

and

$$Q(B_2, t \wedge T_n) = Q(B_1 \cap B_2, t \wedge T_n) + Q(B_2 - B_1, t \wedge T_n),$$

where $B - B' = \{z | z \in B, z \notin B'\}$. The s.i. martingales

$$Q(B_1 \cap B_2, t \wedge T_n), \quad Q(B_1 - B_2, t \wedge T_n) \quad \text{and} \quad Q(B_2 - B_1, t \wedge T_n)$$

have no discontinuities in common so that they are pairwise orthogonal by

[24, § VIII, Thm. 31]. The assertion follows then if one can show that for any $B \in \mathcal{L}$,

$$(3.1) \quad Q^2(B, t \wedge T_n) - \tilde{P}(B, t \wedge T_n) \in \mathcal{M}^1.$$

Let $Q(t) = Q(B, t \wedge T_n)$, $P(t) = P(B, t \wedge T_n)$ and $\tilde{P}(t) = \tilde{P}(B, t \wedge T_n)$. Let $\varepsilon > 0$ and $s < t$ be arbitrary. Let $S_0 \leq S_1 \leq S_2 \leq \dots$ be a sequence of s.t.s such that $S_0 \equiv s$, $\lim_{k \rightarrow \infty} S_k = t$ a.s. and such that $0 \leq \tilde{P}(S_k) - \tilde{P}(S_{k-1}) \leq \varepsilon$ and $0 \leq P(S_k) - P(S_{k-1}) \leq 1$ a.s. Such a sequence exists since \tilde{P} is continuous. Then

$$\begin{aligned} \sum_{k=1}^{\infty} (Q(S_k) - Q(S_{k-1}))^2 &= \sum_{k=1}^{\infty} (P(S_k) - P(S_{k-1}) - \tilde{P}(S_k) + \tilde{P}(S_{k-1}))^2 \\ &= \sum_{k=1}^{\infty} (P(S_k) - P(S_{k-1}))^2 \\ &\quad - 2 \sum_{k=1}^{\infty} (P(S_k) - P(S_{k-1}))(\tilde{P}(S_k) - \tilde{P}(S_{k-1})) \\ &\quad + \sum_{k=1}^{\infty} (\tilde{P}(S_k) - \tilde{P}(S_{k-1}))^2. \end{aligned}$$

The first term in the last expression is equal to $P(t) - P(s)$ since $P(S_k) - P(S_{k-1})$ is zero or one. Hence

$$\begin{aligned} &\left| E \left\{ \sum_{k=1}^{\infty} (Q(S_k) - Q(S_{k-1}))^2 - (P(t) - P(s)) \middle| \mathcal{F}_s \right\} \right| \\ &\leq 2\varepsilon E\{P(t) - P(s) | \mathcal{F}_s\} + \varepsilon E\{\tilde{P}(t) - \tilde{P}(s) | \mathcal{F}_s\}. \end{aligned}$$

Since $\varepsilon > 0$ is arbitrary it follows that

$$(3.2) \quad E \sum_{k=1}^{\infty} \left\{ (Q(S_k) - Q(S_{k-1}))^2 - (P(t) - P(s)) \middle| \mathcal{F}_s \right\} = 0.$$

Now $Q_t \in \mathcal{M}^2$ so that $E\{(Q(S_k) - Q(S_{k-1}))^2 | \mathcal{F}_s\} = E\{Q^2(S_k) - Q^2(S_{k-1}) | \mathcal{F}_s\}$. Also $P_t - \tilde{P}_t \in \mathcal{M}^1$ so that $E\{P(t) - P(s) | \mathcal{F}_s\} = E\{\tilde{P}(t) - \tilde{P}(s) | \mathcal{F}_s\}$. Substituting these relations in (3.2) one obtains

$$\begin{aligned} E \left\{ \sum_{k=1}^{\infty} (Q^2(S_k) - Q^2(S_{k-1})) - (\tilde{P}(t) - \tilde{P}(s)) \middle| \mathcal{F}_s \right\} &= E\{Q^2(t) - Q^2(s) \\ &\quad - (\tilde{P}(t) - \tilde{P}(s)) | \mathcal{F}_s\} = 0, \end{aligned}$$

which is the same as (3.1). \square

For fixed t , $Q(B, t)$, $P(B, t)$ and $\tilde{P}(B, t)$ can be regarded as set functions on \mathcal{L} . In order to define stochastic integrals and Lebesgue–Stieltjes integrals with respect to these set functions it is necessary to show that they are countably additive.

LEMMA 3.2. *Let $B_k, k \geq 1$, be a decreasing sequence in \mathcal{L} such that $\bigcap B_k = \emptyset$. Then for almost all $\omega \in \Omega$, $Q(B_k, t) \rightarrow 0$, $P(B_k, t) \rightarrow 0$, $\tilde{P}(B_k, t) \rightarrow 0$ for all $t \in R_+$ as $k \rightarrow \infty$. Furthermore for all $t \in R_+$ and $n \geq 0$, $EQ^2(B_k, t \wedge T_n) \rightarrow 0$ as $k \rightarrow \infty$.*

Proof. Fix $t \in R_+$. The nonnegative r.v.s $P(B_k, t)$ and $\tilde{P}(B_k, t)$ decrease as k increases so that they converge to some r.v.s $P(t)$ and $\tilde{P}(t)$ respectively. Hence $Q(B_k, t) = P(B_k, t) - \tilde{P}(B_k, t)$ converges to $Q(t) = P(t) - \tilde{P}(t)$. From the definition of $P(B_k, t)$ it is clear that $P(t) = 0$ a.s. and from Lemma 3.1 it follows that $Q_t \in \mathcal{M}_{\text{loc}}^2$. Thus $Q(t) = -\tilde{P}(t) \in \mathcal{M}_{\text{loc}}^2$. But $\tilde{P}(t)$ is an increasing process and $\tilde{P}(0) = 0$ so that this is possible only if $Q(t) = -\tilde{P}(t) = 0$ a.s. Thus $P(t) = \tilde{P}(t) = Q(t) = 0$ for ω not belonging to a null set $N \in \mathcal{F}$. The monotonicity of the sample functions of P, \tilde{P} implies that $P(s) = \tilde{P}(s) = 0$, hence $Q(s) = 0$ for $\omega \notin N$ and $s \leq t$. To prove the remaining assertion it is enough to note that by Lemma 3.1 and by what has just been shown,

$$EQ^2(B_k, t \wedge T_n) = E\tilde{P}(B_k, t \wedge T_n) \rightarrow 0 \quad \text{as } k \rightarrow \infty. \quad \square$$

The following definition relates to the different classes of integrands for which a satisfactory theory of integration is available.

Let \mathcal{H} denote the set of all processes $h(t) = h(\omega, t)$ of the form

$$h(t) = h_0 I_{(t_0, t_1]} + h_1 I_{(t_1, t_2]} + \cdots + h_k I_{(t_k, t_{k+1}]},$$

where h_i is a bounded r.v. measurable with respect to \mathcal{F}_{t_i} and $0 \leq t_0 \leq \cdots \leq t_{k+1} < \infty$. Let \mathcal{P}_0 denote the set of all functions $f(z, t) = f(z, \omega, t)$ of the form

$$f(z, \omega, t) = \sum_{i=0}^k \phi_i(z) h_i(\omega, t),$$

where ϕ_i is a bounded function measurable with respect to \mathcal{Z} and $h_i \in \mathcal{H}$.

DEFINITION 3.2. A function $f(z, t) = f(z, \omega, t)$ is said to be *predictable* if there exists a sequence f_k in \mathcal{P}_0 such that

$$\lim_{k \rightarrow \infty} f_k(z, \omega, t) = f(z, \omega, t) \quad \text{for all } (z, \omega, t) \in \mathcal{Z} \times \Omega \times R_+.$$

Let \mathcal{P} denote the set of all predictable functions and let \mathcal{F}^P be the sub- σ -field of $\mathcal{Z} \otimes \mathcal{F} \otimes \mathcal{B}$ generated by \mathcal{P} .

If $f(z, t) = f(z, \omega, t)$ is measurable with respect to $\mathcal{Z} \otimes \mathcal{F} \otimes \mathcal{B}$ and if for all fixed (z, ω) , $f(z, \omega, t)$ is left-continuous in t , then $f \in \mathcal{P}$.

DEFINITION 3.3.

$$L^2(\tilde{P}) = \left\{ f \in \mathcal{P} \mid (\|f\|_{\tilde{z}})^2 = E \int_Z \int_{R^+} f^2(z, t) \tilde{P}(dz, dt) < \infty \right\}.$$

$$L^1(\tilde{P}) = \left\{ f \in \mathcal{P} \mid \|f\|_{\tilde{1}} = E \int_Z \int_{R^+} |f(z, t)| \tilde{P}(dz, dt) < \infty \right\}.$$

Similarly

$$L^1(P) = \left\{ f \in \mathcal{P} \mid \|f\|_1 = E \int_Z \int_{R^+} |f(z, t)| P(dz, dt) < \infty \right\}.$$

$L_{\text{loc}}^2(\tilde{P})$ is the set of all $f \in \mathcal{P}$ for which there exists a sequence of s.t.s $S_k \uparrow \infty$ a.s. such that $f I_{t \leq S_k} \in L^2(\tilde{P})$ for all k . $L_{\text{loc}}^1(\tilde{P})$ and $L_{\text{loc}}^1(P)$ are defined in an analogous manner. The integrals in this definition are to be interpreted as Lebesgue–Stieltjes

integrals. Finally let $L^1(Q) = L^1(P) \cap L^1(\tilde{P})$, $L^1_{\text{loc}}(Q) = L^1_{\text{loc}}(P) \cap L^1_{\text{loc}}(\tilde{P})$. If $f(z, t) \in L^1(Q)$, then the integral

$$\int_Z \int_{R^+} f(z, t) P(dz, dt) - \int_Z \int_{R^+} f(z, t) \tilde{P}(dz, dt)$$

is denoted

$$\int_Z \int_{R^+} f(z, t) Q(dz, dt).$$

LEMMA 3.3. To each $f \in L^2(\tilde{P})$ there corresponds a unique process $(f \circ Q)_t \in \mathcal{M}^2$, called the “stochastic integral of f with respect to Q ” with the following properties:

(i) if $f(z, \omega, t) = I_B(z) I_A(\omega) I_{(t_0, t_1]}(t) \in L^2(\tilde{P})$, where $B \in \mathcal{Z}$ and $A \in \mathcal{F}_{t_0}$, then

$$(f \circ Q)_t = \begin{cases} I_A(\omega) [Q(B, t \wedge t_1) - Q(B, t \wedge t_0)] & \text{for } t > t_0, \\ 0 & \text{for } t \leq t_0. \end{cases}$$

(ii) if f, g are in $L^2(\tilde{P})$ and α, β are in R , then

$$(\alpha f + \beta g) \circ Q \equiv \alpha(f \circ Q) + \beta(g \circ Q).$$

Furthermore the stochastic integral satisfies the following relations:

$$(3.3) \quad \langle f \circ Q, g \circ Q \rangle_t = \int_Z \int_{R^+} f(z, s) g(z, s) I_{(0, t]}(s) \tilde{P}(dz, ds),$$

and in particular,

$$(3.4) \quad E(f \circ Q)_\infty^2 = (\|f\|_2^2)^2.$$

Proof. The proof follows quite closely that of [22, Prop. 5.1]. Let

$$f^j = \sum_{i=0}^k \alpha_i^j I_{B_i^j}(z) I_{A_i^j}(\omega) I_{(t_i, t_{i+1}]}(t), \quad j = 1, 2,$$

be simple functions in $L^2(\tilde{P})$ with $B_i^j \in \mathcal{Z}$, $A_i^j \in \mathcal{F}_{t_i}$ and $0 = t_0 < t_1 < \cdots < t_{k+1} < \infty$. Then from (i), (ii) and Lemma 3.1, it can be verified directly that

$$(f^1 \circ Q)_t (f^2 \circ Q)_t - \int_Z \int_{R^+} f^1(z, s) f^2(z, s) I_{(0, t]}(s) \tilde{P}(dz, ds) \in \mathcal{M}^1$$

so that (3.3) and (3.4) hold for all simple functions in $L^2(\tilde{P})$. Since such simple functions are dense in $L^2(\tilde{P})$, (3.4) implies that there is a unique extension of the map $f \rightarrow (f \circ Q)$ to all of $L^2(\tilde{P})$. Evidently (3.3) and (3.4) will hold for the extension. \square

LEMMA 3.4. Let $m_t \in \mathcal{M}^2$ have continuous sample paths. Then $m_t \equiv m_0$.

Proof. By replacing the martingale m_t by $m_t - m_0$ it can be assumed that $m_0 = 0$. It will be shown that $m_t \equiv 0$. Suppose $m_{T_{n-1}} = 0$ for some $n \geq 1$ so that in fact $m_{t \wedge T_{n-1}} = E\{m_{T_{n-1}} | \mathcal{F}_{t \wedge T_{n-1}}\} = 0$ for all t , and consider the continuous martingale $\mu_t = m_{t \wedge T_n}$. By Corollary 2.2 there exists a function h , measurable in its arguments, such that $\mu_t \equiv h(t, x_{T_0 \wedge t}, T_0 \wedge t, \cdots, x_{T_n \wedge t}, T_n \wedge t)$. The process $\mu'_t = h(t, x_{T_0 \wedge t}, \cdots, x_{T_{n-1} \wedge t}, T_{n-1} \wedge t, x_{T_{n-1} \wedge t}, t)$ is then measurable with respect

to \mathcal{F}_{T_n-1} . Since for $t < T_n$, $x_{T_n \wedge t} = x_{T_n-1 \wedge t}$ and $t = T_n \wedge t$, it follows that $\mu_t = \mu'_t$ for $t < T_n$, and so by continuity of μ_t , $\mu_t = \mu'_t$ for $t \leq T_n$. For $\alpha \in R_+$ define S_α by

$$S_\alpha(\omega) = \sup \{s \leq \alpha \mid \mu'_s(\omega) \geq 0\}.$$

Then since $\mu'_s = \mu_s = 0$ for $s \leq T_n-1$ it follows that $S_\alpha \geq T_n-1$, and since S_α is measurable with respect to \mathcal{F}_{T_n-1} , therefore S_α is a s.t. for every α . Now let

$$T_\alpha(\omega) = \sup \{s \leq \alpha \wedge T_n(\omega) \mid \mu'_s(\omega) \geq 0\}.$$

It will be shown that T_α is a s.t. Fix t . If $\alpha \leq t$, then $\{T_\alpha \leq t\} = \Omega \in \mathcal{F}_t$ since $T_\alpha \leq \alpha$. Suppose then that $\alpha > t$. Now

$$(3.5) \quad \{T_\alpha \leq t\} = (\{T_\alpha \leq t\} \cap \{T_n \leq t\}) \cup (\{T_\alpha \leq t\} \cap \{T_n > t\}).$$

Since $T_\alpha \leq T_n$, therefore $\{T_n \leq t\} \subset \{T_\alpha \leq t\}$, so that the first set on the right in (3.5) is equal to $\{T_n \leq t\}$ which is in \mathcal{F}_t since T_n is a s.t. It will be shown now that

$$(3.5a) \quad \{T_\alpha \leq t\} \cap \{T_n > t\} = \{S_\alpha \leq t\} \cap \{T_n > t\}.$$

Since $S_\alpha \geq T_\alpha$, the set on the left is at least as large as the one on the right. Suppose $\omega \in \{T_\alpha \leq t\} \cap \{T_n > t\}$. Then $\mu'_s(\omega) < 0$ for $s \in [t, \alpha]$ and $t < T_n(\omega)$, so that $S_\alpha(\omega) \leq t$, which proves (3.5a).

Thus $\{T_\alpha \mid \alpha \in R_+\}$ is a family of s.t.s and furthermore the sample paths $T_\alpha(\omega)$ are nondecreasing functions of α . By the optional sampling theorem [17, Thm. 11.8, p. 376] applied to the martingale μ_t , the process $\eta_\alpha(\omega) = \mu_{T_\alpha(\omega)}(\omega)$, $\alpha \in R_+$, is a martingale. Also, since $T_\alpha < T_n$, therefore $\mu_{T_\alpha} = \mu'_{T_\alpha}$. Hence $\eta_\alpha \geq 0$. But $\eta_0 = 0$, so that one must have $\eta_\alpha \equiv 0$. In turn this can happen only if $\mu_t \leq 0$ which together with $\mu_0 = 0$ implies $\mu_t \equiv 0$. The lemma is proved. \square

THEOREM 3.2. Let $m_t \in \mathcal{M}_{\text{loc}}^1$ have continuous sample paths. Then $m_t \equiv m_0$.

Proof. The s.t.s $S_k(\omega) = \inf \{t \mid |m_t(\omega)| > k\}$ converge to ∞ and

$$m_{t \wedge S_k} I_{\{S_k > 0\}} \in \mathcal{M}_2,$$

so that by Lemma 3.4, $m_{t \wedge S_k} \equiv m_0$. \square

Thus there are no nontrivial continuous martingales. On the other hand if m_t is a martingale, then its discontinuities occur at the jump times T_n of the process x_t as shown below.

LEMMA 3.5. Let S be a predictable s.t. and let $m_t \in \mathcal{M}^2$. Then $\Delta m_S = m_S - m_{S-} = 0$ a.s.

Proof. By [24, § VIII, Thm. 29] the process $\Delta m_S I_{t \geq S}$ is a martingale. By [25, Prop. 7, p. 159], $E\{\Delta m_S \mid \mathcal{F}_{S-}\} = 0$ a.s. But by Proposition 3.1 and [14, § III, Thm. 51], $\mathcal{F}_{S-} = \mathcal{F}_S$ so that $\Delta m_S = 0$ a.s. \square

The next result gives the first martingale representation theorem. It should be compared with [22, Thm. 4.2 and Prop. 5.2].

THEOREM 3.3. Let $m_t \in \mathcal{M}^2$. Then $m_t - m_0 \in \{f \circ Q \mid f \in L^2(\tilde{\mathcal{P}})\}$.

Proof. It can be assumed without losing generality that $m_0 = 0$. The space $\mathcal{M}_0^2 = \{m_t \in \mathcal{M}^2 \mid m_0 = 0\}$ is a Hilbert space under the norm $\|m\|^2 = E m_\infty^2$ by [16, Thm. 1], and by Lemma 3.3 the set $\mathcal{N} = \{f \circ Q \mid f \in L^2(\tilde{\mathcal{P}})\}$ is a closed linear subspace of \mathcal{M}_0^2 . Furthermore \mathcal{N} is closed under stopping, i.e., if $(f \circ Q)_t \in \mathcal{N}$ and T is a s.t., then $(f \circ Q)_{t \wedge T} \in \mathcal{N}$. This is clear because $(f \circ Q)_{t \wedge T} = (f_T \circ Q)_t$,

where $f_T(t) = f_t I_{\{t < T\}}$. Thus by [27, Thm. 2 and the remark following Def. 4] the theorem is proved if it can be shown that $m_t \equiv 0$ when it is orthogonal to $f \circ Q$ for every $f \in L^2(\tilde{P})$. Let m_t be such a martingale. By [16, Thm. 4], m_t can be decomposed uniquely as

$$m_t = m_t^c + m_t^d,$$

where $m_t^c \in \mathcal{M}_0^2$ is continuous and $m_t^d \in \mathcal{M}_0^2$ is orthogonal to every continuous martingale. By Theorem 3.2, $m_t^c \equiv 0$. By Lemmas 2.2 and 3.5, the discontinuities of m_t^d occur during the stopping times T_n , $n \geq 1$. Therefore, by [16, Thm. 4] again, $m_t = m_t^d$ can be further decomposed as

$$m_t = \sum_{n=1}^{\infty} (M_n I_{t \geq T_n} - a_n(t)) = \sum_{n=1}^{\infty} \mu_{nt} \quad \text{say,}$$

where $M_n = \Delta m_{T_n} = m_{T_n} - m_{T_n-}$, $a_n(t) \in \mathcal{A}$ has continuous sample paths, and $\mu_{nt} \in \mathcal{M}_0^2$. Furthermore the martingale μ_{nt} is orthogonal to every martingale which has no discontinuities at T_n .

To prove that $m_t \equiv 0$ it suffices to show that $M_n = 0$ for each n . Fix n and suppose that $P\{M_n \neq 0\} > 0$. Since M_n is measurable with respect to \mathcal{F}_{T_n} , therefore by Corollary 2.3 there must exist sets $A \in \mathcal{F}_{T_{n-1}}$, $B \in \mathcal{L}$, and $C \in \mathcal{B}[0, \infty)$ such that

$$(3.6) \quad E\{M_n(\omega) I_A(\omega) I_{\{x_{T_n} \in B\}} I_{\{T_n \in C\}}\} \neq 0.$$

Consider the function $f(z, \omega, t)$ defined by

$$f(z, \omega, t) = I_B(z) I_A(\omega) I_C(t) I_{\{T_{n-1} < t \leq T_n\}}.$$

The function $g(z, \omega, t) = I_B(z) I_A(\omega) I_{\{T_{n-1} < t \leq T_n\}}$ has left-continuous paths for fixed (z, ω) and for each fixed z, t the set

$$\{I_A(\omega) I_{\{T_{n-1} < t \leq T_n\}} = 1\} = A \cap \{T_{n-1} < t\} \cap \{t \leq T_n\} \in \mathcal{F}_t$$

since $A \in \mathcal{F}_{T_{n-1}}$. Therefore $g(z, t)$ is adapted, so that $g \in \mathcal{P}$ and hence $f = g I_C(t)$ is also predictable. Also $|f| \leq 1$ and $f(z, t) = 0$ for $t > T_n$ so that $f \in L^2(\tilde{P}) \cap L^1(\tilde{P}) \cap L(P)$. Therefore by Lemma 3.6 below it follows that

$$\begin{aligned} \eta_t &= (f \circ Q)_t = \int_Z \int_{R^+} f(z, s) I_{(0, t]}(s) P(dz, ds) - \int_Z \int_{R^+} f(z, s) I_{(0, t]}(s) \tilde{P}(dz, ds) \\ &= I_A(\omega) I_{\{x_{T_n} \in B\}} I_{\{T_n \in C\}} I_{\{t \geq T_n\}} - a(t), \end{aligned}$$

where $a(t)$ is a continuous process. Thus the discontinuities of η_t occur at T_n . Since m_t is orthogonal to η_t by hypothesis, therefore

$$0 \equiv \langle m, \eta \rangle_t = \sum_{k \neq n} \langle \mu_k, \eta \rangle_t + \langle \mu_n, \eta \rangle_t.$$

Also $\langle \mu_k, \eta \rangle_t \equiv 0$ for $k \neq n$, hence $\langle \mu_n, \eta \rangle_t \equiv 0$ so that $\mu_n \cdot \eta \in \mathcal{M}^1$. By the Corollary in [16, p. 106] and the Definition in [16, p. 87] it follows that $\Delta \mu_{nT_n} \cdot \Delta \eta_{T_n} \cdot I_{t \geq T_n}$ is a martingale so that

$$E\{M_n(\omega) I_A(\omega) I_{\{x_{T_n} \in B\}} I_{\{T_n \in C\}}\} = 0,$$

which contradicts (3.6). The theorem has been proved. \square

Lemma 3.3 provides an obvious extension of the definition of the stochastic integral $(f \circ Q)_t$ to $f \in L^2_{\text{loc}}(\tilde{P})$ and so Theorem 3.3 extends in the following manner.

COROLLARY 3.1. $\{m_t - m_0 | m_t \in \mathcal{M}^2_{\text{loc}}\} = \{(f \circ Q)_t | f \in L^2_{\text{loc}}(\tilde{P})\}$.

To obtain the representation for martingales in $\mathcal{M}^1_{\text{loc}}$, two preliminary results are needed.

LEMMA 3.6. (i) Let $f \in \mathcal{P}$. Then $f \in L^1(P)$ if and only if $f \in L^1(\tilde{P})$. In fact, $\|f\|_1 = \|f\|_1^\sim$. In particular, $L^1(P) = L^1(\tilde{P}) = L^1(Q)$.

(ii) Let $f \in L^2(\tilde{P})$. Then $f \in L^1(\tilde{P})$ and

$$(3.7)^6 \quad (f \circ Q)_t = \int_Z \int_{R^+} f(z, s) I_{(0,t]}(s) Q(dz, ds).$$

(iii) If $f \in L^1(\tilde{P})$, then

$$m_t = \int_Z \int_{R^+} f(z, s) I_{(0,t]}(s) Q(dz, ds) \in \mathcal{M}^1 \cap \mathcal{A}.$$

Proof. By an argument which is almost identical to the proof of [16, Prop. 3], it can be shown that (3.7) holds for $f \in L^2(\tilde{P}) \cap L^1(\tilde{P}) \cap L^1(P)$.

Since $L^2(\tilde{P}) \subset L^1(\tilde{P})$ the second assertion will then follow from the first one. Now let Φ consist of all bounded functions $f(z, t) \in \mathcal{P}$ such that $f(z, t) \equiv 0$ for $t \geq T_n$ for some $n < \infty$. Then certainly $\Phi \subset L^2(\tilde{P}) \cap L^1(\tilde{P}) \cap L^1(P)$. So $(f \circ Q)_t \in \mathcal{M}_2$ for $f \in \Phi$, and in particular, by (3.7),

$$0 = E(|f| \circ Q)_\infty = \|f\|_1 - \|f\|_1^\sim.$$

Thus the identity map, restricted to Φ , from $L^1(P)$ to $L^1(\tilde{P})$ preserves norms. Since Φ is dense in $L^1(P)$ and $L^1(\tilde{P})$, the first assertion follows. To prove the last assertion, let f_k , $k \geq 1$, be a sequence in $L^2(\tilde{P})$ such that $\|f - f_k\|_1$ converges to zero. Then $m_{kt} = (f_k \circ Q)_t \in \mathcal{M}^2$ and by (3.7), $E|m_{kt} - m_t| \leq 2\|f - f_k\|_1$ converges to zero uniformly in t so that $m_t \in \mathcal{M}^1$. \square

PROPOSITION 3.3. Let M be a \mathcal{F}_{T_n} -measurable r.v. for some $n \geq 1$. Suppose $E|M| < \infty$. Then there is a unique $f(z, t) \in L^1(\tilde{P})$ such that

$$(3.8) \quad MI_{t \geq T_n} = \int_Z \int_{R^+} f(z, s) I_{(0,t]}(s) P(dz, ds).$$

Furthermore $f(z, s) = 0$ for $s \leq T_{n-1}$ and $s > T_n$, and

$$(3.9) \quad E|MI_{\{T_n < \infty\}}| = \|f\|_1.$$

Proof. Since $MI_{t \geq T_n} = MI_{\{T_n < \infty\}} I_{\{t \geq T_n\}}$, it can be assumed that $M = MI_{\{T_n < \infty\}}$. By Corollary 2.3 there exist r.v.s M^k of the form

$$M^k(\omega) = \sum_i \alpha_i I_{\{x_{T_n} \in B_i\}} I_{A_i}(\omega) I_{\{T_n \in C_i\}},$$

where $\alpha_i \in R$, $B_i \in \mathcal{L}$, $A_i \in \mathcal{F}_{T_{n-1}}$ and $C_i \in \mathcal{B}[0, \infty)$, such that $E|M - M^k| \rightarrow 0$. If

⁶ It may be worth repeating, to clarify the content of (3.7), that the integral on the right in (3.7) is a Lebesgue-Stieltjes integral whereas that on the left is the stochastic integral as defined in Lemma 3.3.

f^k is defined by

$$f^k(z, \omega, t) = \sum_i \alpha_i I_{B_i}(z) I_{A_i}(\omega) I_{C_i}(t) I_{\{T_{n-1} < t \leq T_n\}},$$

then it is clear that (3.8) and (3.9) hold for M^k and f^k . The assertion now follows by taking limits. \square

LEMMA 3.7. *Let $m_t \in \mathcal{M}^1 \cap \mathcal{A}$. Then there exists $f \in L^1(\tilde{P})$ such that*

$$(3.10) \quad m_t - m_0 = \int_Z \int_{R^+} f(z, s) I_{(0,t]}(s) Q(dz, ds)$$

and

$$(3.11) \quad E \int_0^\infty |dm_t| = 2 \|f\|_1.$$

Proof. m_t has the representation

$$m_t - m_0 = \sum_{n=1}^\infty (M_n I_{t \geq T_n} - a_n(t)) = \sum_{n=1}^\infty \mu_n,$$

where $M_n = \Delta m_{T_n}$, $a_n(t) \in \mathcal{A}$ is continuous, and $\mu_n \in \mathcal{M}$. Since $m_t \in \mathcal{A}$,

$$\infty > E \int_0^\infty |dm_t| > \sum_{n=1}^\infty E |M_n|,$$

so that by Proposition 3.3, there exist functions $f_n(z, t) \in L^1(\tilde{P})$ which vanish outside of $\{T_{n-1} \leq t \leq T_n\}$ such that $E |M_n| = \|f_n\|_1$ and

$$M_n I_{t \geq T_n} = \int_Z \int_{R^+} f_n(z, s) I_{(0,t]}(s) P(dz, ds).$$

By Lemma 3.6,

$$\eta_n(t) = a_n(t) - \int_Z \int_{R^+} f_n(z, s) I_{(0,t]}(s) \tilde{P}(dz, ds) \in \mathcal{M}^1.$$

But $\eta_n(t)$ is continuous so that $\eta_n(t) \equiv 0$ by Theorem 3.2. Therefore (3.10) holds for $f(z, t) = \sum_{n=1}^\infty f_n(z, t)$ and (3.11) follows from Lemma 3.6 and the fact that $f_k(z, t) f_n(z, t) \equiv 0$ for $k \neq n$. \square

THEOREM 3.4. $m_t \in \mathcal{M}_{\text{loc}}^1$ if and only if there exists $f \in L_{\text{loc}}^1(\tilde{P})$ such that

$$(3.12) \quad m_t - m_0 \equiv \int_Z \int_{R^+} f(z, s) I_{(0,t]}(s) Q(dz, ds).$$

Proof. The sufficiency follows readily from Lemma 3.6 (iii). To prove the necessity one starts by noting that by [16, Lemma 3 and Prop. 4] there exists an increasing sequence of s.t.s S_k converging to ∞ such that for each k , $m_{t \wedge S_k} - m_0$ has a decomposition

$$m_{t \wedge S_k} - m_0 = \mu_t^k + \eta_t^k,$$

where $\mu_t^k \in \mathcal{M}_0^2$ and $\eta_t^k \in \mathcal{M}_0^1 \cap \mathcal{A}$. By Lemmas 3.6 (ii) and 3.7, there exists

$f^k \in L^1(\tilde{P})$ such that

$$m_{t \wedge S_k} - m_0 = \int_Z \int_{R^+} f^k(z, s) I_{(0, \infty]}(s) Q(dz, ds).$$

It is clear that $f^k(z, t) = f^{k+1}(z, t)$ for $t \leq S_k$. Thus (3.12) holds for $f \in L^1_{\text{loc}}(\tilde{P})$ defined by $f(z, t) = f^k(z, t)$ for $t \leq S_k$. \square

The results above give a characterization of the classes \mathcal{M}^2 , $\mathcal{M}^2_{\text{loc}}$, $\mathcal{M}^1 \cap \mathcal{A}$ and $\mathcal{M}^1_{\text{loc}}$. It seems much more difficult to obtain a useful characterization of the class \mathcal{M}^1 .

The (local) martingales with respect to $(\Omega, \mathcal{F}_t, P)$ have been represented as sums or integrals of the “basic” martingales $Q(B, t)$. The latter are associated in a one-to-one manner with the counting processes $P(B, t)$ which count those jumps of the underlying process x_t which end in the set B . Thus jumps are distinguished by their final values. Now it is also possible to distinguish jumps by their values. The corresponding counting processes will be of the form $p(A, t)$ which counts those jumps of the x_t process which have values in the set A . The martingales $q(A, t)$ associated with the $p(A, t)$ also form a “basis” for the set of all martingales on $(\Omega, \mathcal{F}_t, P)$ as will be shown below. The alternative representation obtained with this basis can sometimes be more useful since the description of the x_t process is, in practice, often given in terms of a statistical characterization of the jumps of x_t .

For simplicity of notation it will be assumed in the remainder of this section that the x_t process starts at time 0 in a fixed state, i.e., $x_0(\omega) = x_0(\omega')$ for all $\omega, \omega' \in \Omega$.⁷ Next it is assumed that there is given a set Σ of transformations $\sigma: Z \rightarrow Z$ with the following properties:

(i) Σ contains the jumps of the x_t process, i.e., if $x_{s-}(\omega) \neq x_s(\omega)$ for some $s \in R_+$, $\omega \in \Omega$, then there is a unique $\sigma \in \Sigma$ such that $\sigma(x_{s-}(\omega)) = x_s(\omega)$;

(ii) Σ contains a distinguished element σ_0 corresponding to the identity transformation, i.e., $\sigma_0(z) = z$ for all $z \in Z$.

To each sample function $\omega \in \Omega$ of the x_t process is associated a function $\gamma(\omega): R_+ \rightarrow \Sigma$ defined as follows:

$$\gamma_t(\omega) = \begin{cases} \sigma_0 & \text{if } t = 0 \text{ or if } x_t(\omega) = x_{t-}(\omega), \\ \sigma & \text{if } x_t(\omega) \neq x_{t-}(\omega), \end{cases}$$

where $\sigma \in \Sigma$ is the unique element for which $\sigma(x_{t-}(\omega)) = x_t(\omega)$.

Remark. (i) Given a sample path $x_s(\omega)$, $0 \leq s \leq t$, there corresponds in a one-to-one manner a sample path $\gamma_s(\omega)$, $0 \leq s \leq t$.

(ii) The functions $\gamma(\omega)$ are not right continuous. However if $\gamma_t(\omega) = \sigma_0$, then $\gamma_{t-}(\omega) = \sigma_0$. This observation will be used later in an example.

The following “regularity” assumption appears to be necessary. In practice it is readily verifiable.

Assumption. There is a σ -field Ξ on Σ such that \mathcal{F}_t coincides with the σ -field generated by subsets of the form $\{\omega | \gamma_s(\omega) \in A\}$, where $s \leq t$ and $A \in \Xi$.

⁷ It should be noted however that the results below continue to hold in the absence of this simplification.

With the assumptions above it is clear that the processes x_t and γ_t are equivalent alternative descriptions of the same process. In particular they generate the same σ -fields, so that the two processes have the same martingales. The representation theorems derived earlier for the x_t process can be applied to the γ_t process but there is a minor point to be cleared up. Recall that it was assumed that the x_t process was right-continuous whereas γ_t is not. However the assumption of right-continuity was used only to establish the right-continuity of the family \mathcal{F}_t . This continues to hold of course since γ_t and x_t generate the same σ -fields \mathcal{F}_t . Hence one can apply the representation theorems.

DEFINITION 3.4. Let $A \in \Xi$. Let

$$p(A, t) = \sum_{s \leq t} I_{\{\gamma_{s-} \neq \gamma_s\}} I_{\{\gamma_s \in A\}} = \sum_{s \leq t} I_{\{x_{s-} \neq x_s\}} I_{\{\gamma_s \in A\}}$$

be the number of jumps of the x_t process with “values” in A and which occur prior to t .

By Proposition 3.2 there is a unique continuous process $\tilde{p}(A, t) \in \mathcal{A}_{\text{loc}}^+$ such that the process $q(A, t) = p(A, t) - \tilde{p}(A, t)$ is in $\mathcal{M}_{\text{loc}}^2$. In analogy with Definitions 3.2 and 3.3 one can define the subsets of $\mathcal{P}_\Sigma : L^2(\tilde{p}), L_{\text{loc}}^2(\tilde{p}), L^1(\tilde{p}), L^1(p)$ etc.⁸ Lemma 3.3 describes the stochastic integrals $(f \circ q)$ for $f \in L^2(\tilde{p})$. An application of Theorem 3.3, Corollary 3.1, Lemma 3.7 and Theorem 3.4 yields the following representation theorem.

THEOREM 3.5. (i) $m_t \in \mathcal{M}^2(\mathcal{M}_{\text{loc}}^2)$ if and only if $m_t - m_0 = (f \circ q)_t$ for some $f \in L^2(\tilde{p})(L_{\text{loc}}^2(\tilde{p}))$.

(ii) $m_t \in \mathcal{M}^1 \cap \mathcal{A}(\mathcal{M}_{\text{loc}}^1)$ if and only if $m_t - m_0 = \int_\Sigma \int_{R^+} f(\sigma, s) I_{(0,1]}(s) q(d\sigma, ds)$ for some $f \in L^1(\tilde{p})(L_{\text{loc}}^1(\tilde{p}))$.

4. An example. This section consists of a simple example showing how Theorem 3.5 can be applied. The example will be further elaborated in [3]. Let Z be countable and let \mathcal{Z} consist of all subsets of Z . Let x_t be a process with values in Z and satisfying the assumptions listed at the beginning of § 3. Suppose that from each state z the process x_t can jump to one of n states. In terms of a state-transition diagram (see Fig. 1) there are n transitions or links emanating from each state or node. Label these transitions by the symbols $\sigma_1, \dots, \sigma_n$. Let $\Sigma = \{\sigma_0, \dots, \sigma_n\}$. Thus each $\sigma \in \Sigma$ corresponds to a transformation in Z , σ_0 is the identity transformation. Let Ξ be the set of all subsets of Σ . The x_t process defines the process of transitions γ_t . Evidently Σ, Ξ satisfy the assumptions made above.

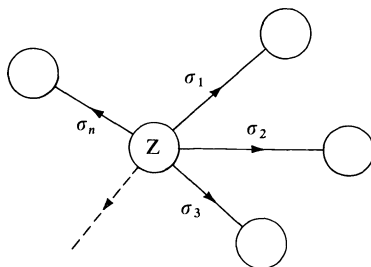


FIG. 1. State-transition diagram for example

⁸ \mathcal{P}_Σ is the set of predictable functions of $(\sigma, \omega, t) \in \Sigma \times \Omega \times R_+$ defined in analogy with Definition 3.2.

Let $p_i(t) = p(\{\sigma_i\}, t)$, $\tilde{p}_i(t) = \tilde{p}_i(\{\sigma_i\}, t)$ and $q_i(t) = q_i(\{\sigma_i\}, t)$, $0 \leq i \leq n$. From a remark made in the last section, $I_{\{x_s \neq x_s\}} \cdot I_{\{\gamma_s = \sigma\}} \equiv 0$. Hence $p_0(t) \equiv 0$ and so $q_0(t) \equiv 0$. Theorem 3.5 simplifies to the following. Here the predictable integrands are functions of (ω, t) only.

THEOREM 4.1. (i) $m_t \in \mathcal{M}^2(\mathcal{M}_{\text{loc}}^2)$ if and only if $m_t - m_0 \equiv \sum_{i=1}^n (f_i \circ q_i)_t$ for some $f_i \in L^2(\tilde{p}_i)(L_{\text{loc}}^2(\tilde{p}_i))$, $1 \leq i \leq n$.

(ii) $m_t \in \mathcal{M}^1 \cap \mathcal{A}(\mathcal{M}_{\text{loc}}^1)$ if and only if $m_t - m_0 = \sum_{i=1}^n \int_{(0,t]} f_i(s) q_i(ds)$ for some $f_i \in L^1(\tilde{p}_i)(L_{\text{loc}}^1(\tilde{p}_i))$, $1 \leq i \leq n$.

Example. Let x_t be a process taking values in a countable state space and of the type described immediately above. From each state the process can make n transitions $\sigma_1, \dots, \sigma_n$ as sketched in Fig. 1. Let $p_i(t)$, $\tilde{p}_i(t)$, $q_i(t)$ be as in Theorem 3.6.

Let $\lambda(t)$, $\rho_1(t)$, \dots , $\rho_n(t)$ be nonnegative predictable processes such that

$$(4.1) \quad \sum_{i=1}^n \rho_i(t) \equiv 1,$$

$$(4.2) \quad p_i(t \wedge T_k) - \int_0^{t \wedge T_k} \rho_i(s) \lambda(s) ds \in \mathcal{M}^1, \quad k = 1, 2, \dots, n.$$

Then the processes $\lambda(t)$, $\rho_i(t)$ have the following interpretation: since from (4.1) and (4.2),

$$(4.3) \quad \left(\sum_{i=1}^n p_i(t \wedge T_k) - \int_0^{t \wedge T_k} \lambda(s) ds \right) \in \mathcal{M}^1$$

and since $\sum_{i=1}^n p_i(t)$ is just the total number of jumps of the process occurring prior to t , therefore the probability that the process x_t makes a transition in the time interval $[t, t+h]$, conditioned on the past \mathcal{F}_t of the process, is equal to $\lambda(t)h + o(h)$. Similarly, $\rho_i(t)$ is the probability that the process makes a transition represented by σ_i , conditioned on \mathcal{F}_t and conditioned on the fact that a transition does occur at t .

Now since the process represented by the indefinite integral in (4.2) has continuous sample paths, it follows quite readily (see, e.g., [25, p. 153]) that the jump times of the process are totally inaccessible. Hence from Theorem 4.1 it can be concluded that every $m_t \in \mathcal{M}_{\text{loc}}^1$ has a representation

$$(4.4) \quad m_t - m_0 = \sum_{i=1}^n \left[\int_0^t f_i(s) dp_i(s) - \int_0^t f_i(s) \rho_i(s) \lambda(s) ds \right]$$

for some predictable processes $f_i \in L_{\text{loc}}^1(\rho_i \lambda)$, i.e., for which

$$\int_0^t f_i(s) \rho_i(s) \lambda(s) ds < \infty \quad \text{a.s. for all } t \in R_+.$$

This result indicates how one can immediately write down the representation results if the process x_t is described in terms of the “rate” processes λ and the “transition” probabilities ρ_i . It should be kept in mind, however, that it has *not* been proved that given processes $\lambda(t)$ and $\rho_i(t)$ there exists a process x_t for which (4.2) holds. This question of existence will be pursued in [3]. The next remark

relates to the representation (4.4), which asserts that the n local martingales in (4.2) indeed form a “basis” for the space of all local martingales $\mathcal{M}_{\text{loc}}^1$. The question is whether n is the minimum number of martingales in every basis of $\mathcal{M}_{\text{loc}}^1$. For the case where x_t is a Gaussian process the minimum number of martingales has been called the “multiplicity” of the process by Cramer [8], [9]. It turns out that this notion of multiplicity extends in a very natural way to arbitrary processes [13]. From the results of [13] the following sufficient condition can be obtained: Suppose that the processes $\rho_i(s)\lambda(s)$ satisfy

$$\rho_i(s)\lambda(s) > 0 \Leftrightarrow \rho_j(s)\lambda(s) > 0 \quad \text{all } i, j.$$

Then n is the minimum number of martingales in a representation of $\mathcal{M}_{\text{loc}}^1$.

Finally, specialize the example still further and assume that x_t is a counting process, i.e., $x_0 = 0$, x_t takes integer values and has unit positive jumps. Then x_t is a direct extension of a Poisson process. The state-transition diagram then simplifies to that of Fig. 2 and since $n = 1$ in (4.1), (4.2) and (4.4), therefore $\rho_1(t) \equiv 1$ and can be omitted. Also $p_1(t) \equiv x_1(t)$ and so the representation (4.4) simplifies to (4.5). Every $m_t \in \mathcal{M}_{\text{loc}}^1$ can be written as

$$(4.5) \quad m_t - m_0 = \int_0^t f(s) dx_s - \int_0^t f(s)\lambda(s) ds,$$

where f is a predictable function such that

$$\int_0^t f(s)\lambda(s) ds < \infty \quad \text{a.s. for all } t \in R_+.$$

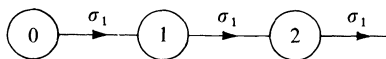


FIG. 2. Transition diagram for counting process

This representation result has been obtained by very different techniques by several authors [4], [5], [11], [12]. However even here the cited references prove (4.5) for the special case where the probability law of the x_t process is mutually absolutely continuous with respect to the probability law of a standard Poisson process. Hence even for this special case, (4.5) is a strict generalization of the available results.

Appendix. The increasing processes $\tilde{P}(A, t)$ and the Lévy system. This section attempts to give an intuitive interpretation of the increasing processes $\tilde{P}(B, t)$ and shows the connection with the Lévy system for Hunt processes.

Begin with the observation that for all $B \in \mathcal{Z}$ the measure $P(B, t)$ is absolutely continuous with respect to the measure $\tilde{P}(Z, t)$, i.e., there exists a predictable function $(\omega, t) \rightarrow n(B, \omega, t)$ such that

$$(A.1) \quad \tilde{P}(B, t) = \int_0^t n(B, \omega, s) P(Z, ds).$$

To see this it is enough to demonstrate that for all predictable functions

$\phi(\omega, s) = \phi^2(\omega, s)$ (i.e., all indicator functions),

$$(A.2) \quad E \int_0^\infty \phi(\omega, s) \tilde{P}(Z, ds) = 0$$

implies

$$(A.3) \quad E \int_0^\infty \phi(\omega, s) \tilde{P}(B, ds) = 0.$$

Suppose (A.2) holds. Then

$$\left\langle \int_0^t \phi(s) dQ(Z, s), \int_0^t \phi(s) dQ(Z, s) \right\rangle \equiv \int_0^t \phi^2(s) \tilde{P}(Z, ds) \equiv 0,$$

and so

$$\begin{aligned} 0 &\equiv \left\langle \int_0^t \phi(s) dQ(B, s), \int_0^t \phi(s) dQ(Z, s) \right\rangle \\ &= \int_0^t \phi^2(s) \tilde{P}(B \cap Z, ds) \quad (\text{by Lemma 3.1}) \\ &= \int_0^t \phi^2(s) \tilde{P}(B, ds), \end{aligned}$$

which proves (A.3).

In exactly the same way as Lemma 3.2 was proved it can be shown that the $n(B, \omega, s)$ considered as a set function in \mathcal{Z} is countably additive in the sense that if B_1, B_2, \dots is a disjoint sequence of sets in \mathcal{Z} , then

$$\tilde{P} \left(\bigcup_i B_i, t \right) \equiv \sum_i \int_0^t n(B_i, s) \tilde{P}(Z, ds).$$

Hence if one sets $\tilde{P}(Z, t) \equiv \Lambda(t) \in \mathcal{A}_{\text{loc}}^+$, then the system $\{n(B, t, \omega), \Lambda(t)\}$ is analogous to a Lévy system for Hunt processes (see [22]), and has a similar interpretation: the probability of x_t having a jump in $[t, t + dt)$ is $d\Lambda(t) + o(dt)$, while $n(A, t, \omega)$ is the chance that $x_t \in A$ given \mathcal{F}_t and given that a jump occurs at t .

Acknowledgment. The authors are very grateful to J. M. C. Clark, M. H. A. Davis and J. H. Van Schuppen for many helpful suggestions and discussions. The authors wish to acknowledge that their efforts were motivated and directed by the pioneering work of P. Bremaud.

REFERENCES

- [1] D. BLACKWELL, *On a class of probability spaces*, Third Symp. on Math. Stat. and Prob., vol. II, Univ. of Calif. Press, Calif., 1956, pp. 1–6.
- [2] B. BENEŠ, *Existence of optimal stochastic control laws*, this Journal, 9 (1971), pp. 446–475.
- [3] R. BOEL, P. VARAIYA AND E. WONG, *Martingales on jump processes. II: Applications*, this Journal, 13 (1975), pp. 1022–1061.

- [4] P. BRÉMAUD, *A martingale approach to point processes*, Electronics Res. Lab., Memo #M-345, Univ. of Calif., Berkeley, 1972.
- [5] ———, *Filtering for point processes*, preprint, Information and Control, submitted.
- [6] J. M. C. CLARK, *The representation of functionals of Brownian motion by stochastic integrals*, Ann. Math. Statist., 41 (1970), pp. 1285–1295.
- [7] P. COURRÈGE AND P. PRIOURET, *Temps d'arrêt d'une fonction aléatoire, théorèmes de décomposition*, Publ. Inst. Statist. Univ. Paris, 14 (1965), pp. 242–274.
- [8] H. CRAMER, *Stochastic processes as curves in Hilbert space*, Theor. Probability Appl., 9 (1964), pp. 195–204.
- [9] ———, *A contribution to the multiplicity theory of stochastic processes*, Fifth Symp. on Math. Stat. and Prob., vol. II, Univ. of Calif. Press, Calif., 1967, pp. 215–221.
- [10] M. H. A. DAVIS, *Detection of signals with point process observations*, Dept. of Computing and Control, Imperial College, London, 1973.
- [11] ———, *Nonlinear filtering with point process observations*, preprint, 1972.
- [12] M. H. A. DAVIS AND P. VARAIYA, *Dynamic programming conditions for partially observable stochastic systems*, this Journal, 11 (1973), pp. 226–261.
- [13] ———, *The multiplicity of an increasing family of σ -fields*, Ann. of Prob., 2 (1974), pp. 958–963.
- [14] C. DELLACHERIE, *Capacités et Processus Stochastiques*, Springer-Verlag, Berlin, 1972.
- [15] C. DOLÉANS-DADE, *Quelques applications de la formule de changement de variables pour les semi-martingales*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 16 (1970), pp. 181–194.
- [16] C. DOLÉANS-DADE AND P. A. MEYER, *Intégrales stochastiques par rapport aux martingales locales*, Séminaire de Probabilités IV, Lecture notes in Mathematics, Springer-Verlag, Berlin, (1970), pp. 77–107.
- [17] J. L. DOOB, *Stochastic Processes*, Wiley, New York, 1953.
- [18] T. DUNCAN, *Evaluation of likelihood functions*, Information and Control, 13 (1968), pp. 62–74.
- [19] T. E. DUNCAN AND P. VARAIYA, *On the solutions of a stochastic control system*, this Journal, 9 (1971), pp. 354–371.
- [20] GIRSANOV, *On transforming a certain class of stochastic processes by absolutely continuous substitution of measures*, Theor. Probability Appl., 5 (1960), pp. 285–301.
- [21] T. KAILATH AND M. ZAKAI, *Absolute continuity and Radon-Nikodym derivatives for certain measures relative to Wiener measure*, Ann. Math. Statist., 42 (1971), pp. 130–140.
- [22] H. KUNITA AND S. WATANABE, *On square integrable martingales*, Nagoya Math. J., 30 (1967), pp. 209–245.
- [23] M. LOÈVE, *Probability Theory*, 2nd ed., Van Nostrand, New York, 1960.
- [24] P. A. MEYER, *Probabilités et potentiel*, Hermann, Paris, 1966, English translation, *Probability and Potentials*, Blaisdell, Waltham, Mass., 1966.
- [25] ———, *Guide détaillé de la théorie "générale" des processus*, Séminaire de Probabilité. II, Lecture Notes in Mathematics, vol. 51, Springer-Verlag, Berlin, 1968, pp. 140–165.
- [26] ———, *Non-square integrable martingales etc.*, Lecture Notes in Mathematics, No. 190, Springer-Verlag, Berlin, 1971, pp. 38–43.
- [27] ———, *Square integrable martingales, a survey*, Lecture Notes in Mathematics, No. 190, Springer-Verlag, Berlin, 1971, pp. 32–37.
- [28] ———, *Temps d'arrêt aléablement prévisibles*, Séminaire de Probabilités VI, Lecture Notes in Mathematics, No. 258, Springer-Verlag, Berlin, 1972, pp. 158–163.
- [29] D. L. SNYDER, *A representation theorem for observed jump processes*, Proc. IEEE Conf. on Decision and Control, New Orleans, La., 1972, p. 218.
- [30] ———, *Information processing for observed jump processes*, Information and Control, 22 (1973), pp. 69–78.
- [31] ———, *Statistical analysis of dynamic tracer data*, IEEE Trans. Biomedical Engr., BME-20 (1973), pp. 11–20.
- [32] J. H. VAN SCHUPPEN AND E. WONG, *Transformations of local martingales under a change of law*, Memo #M-385, Electronics Res. Lab., Univ. of Calif., Berkeley, 1973.
- [33] E. WONG, *Martingale theory and applications to stochastic problems in dynamical systems*, Publication 72/19, Imperial College, Dept. of Computing and Control, 1972.
- [34] ———, *Recent progress in stochastic processes—A survey*, IEEE Trans. Information Theory, IT-19, 1973, pp. 263–275.

- [35] C. S. CHOU AND P. A. MEYER, *Sur la représentation des martingales comme intégrales stochastiques dans les processus ponctuels*, Séminaire de Probabilités VIII, Lecture Notes in Mathematics, Springer-Verlag, Berlin, 1974, to appear.
- [36] J. JACOD, *On the Stochastic Intensity of a Random Point Process over the Half-Line*, Department of Statistics, Princeton University, 1974.

MARTINGALES ON JUMP PROCESSES. II: APPLICATIONS*

R. BOEL, P. VARAIYA AND E. WONG†

1. Introduction and summary. This paper is concerned with applying the theory of martingales of jump processes to various problems arising in communication and control. It parallels the approaches which have been recently discovered in dealing with similar problems where the underlying stochastic process is Brownian motion. Indeed these approaches have recently been extended, starting with the work of Snyder [14], [16], [30] and Brémaud [6], [28], to the case of the Poisson process and its transformations. The paper can then be regarded as a sweeping generalization to this recent work.

The paper can also be considered as an illustration of an abstract view and a set of instructions which must be followed to obtain certain concrete results in the areas of communication and control. It is hoped that this tutorial function will also be served.

Two results from the abstract theory of martingales form the basis of this abstract view. The first consists of the differentiation rule and the associated stochastic calculus for martingales and semi-martingales [1], and its application to the so-called “exponentiation” formula [2]. The second result consists of the earlier Doob–Meyer decomposition theorem for supermartingales [3]. In order to follow the abstract view, one also needs a third set of results, the so-called “martingale representation” theorems for specific processes. These results form a bridge between the abstract theory and the concrete applications. The representation results used here have been obtained in [4], hence the paper can also be viewed as a continuation of that work.

The paper is organized in the following manner. In the next section are presented many definitions, notations and results from [1], [2], [3], [4] which will be used in the succeeding development. These preliminaries are certainly longer than can be considered proper, and are justified partly to serve the tutorial function, partly because there is no consensus of usage in the literature, and lastly because some of the published literature contains errors and inaccurate or misleading statements which can be exposed only within a carefully and completely developed context.

Section 3 is concerned with showing the “global” existence of jump processes over a finite or infinite interval which satisfy certain local descriptions. Existence of such processes is obtained by transforming the laws of “known” processes by an absolutely continuous transformation. We also present a wide class of point processes which can be so transformed to yield solutions to prespecified local

* Received by the editors December 21, 1973, and in revised form August 13, 1974.

† Department of Electrical Engineering and Computer Sciences and the Electronics Research Laboratory, University of California, Berkeley, California 94720. This research was supported by the National Science Foundation under Grant GK-10656X3 and the Army Research Office—Durham under Contract DAHC04-67-C-0046. The work of the first author was also supported by an ESRO-NASA International Fellowship.

descriptions. Sufficient conditions are derived which guarantee when this technique is applicable. The question of uniqueness of the solutions is settled for a wide class of local descriptions.

Section 4 deals with a specific problem in communication theory, namely the calculation of the likelihood ratio of a process which may be governed by one or two absolutely continuous probability laws. The techniques for §§ 3 and 4 are the same. Section 5 is concerned with estimating certain random variables or processes which are statistically related to an observed process. The emphasis here is on obtaining "recursive" filters. As special cases one obtains a "closed form" solution for some of the situations where the estimated process is Markovian. Applications to optimal control will be made in a future paper.

Throughout, there has been an attempt to link up the results with those which have already appeared in the literature in as precise a manner as limitations of space permit. Any omissions are due to oversight of the authors.

2. Preliminaries and formulations. This section describes most of the results from the literature which are necessary to the sequel. Section 2.1 is definitional in nature. Sections 2.2–2.7 are taken mainly from [1], § 2.8 is taken from [2], the remainder is from [4].

2.1. Processes. Throughout Ω is a fixed space, the sample space. The time interval of interest is $R_+ = [0, \infty)$ unless specified otherwise. For each t let \mathcal{F}_t be a σ -field of subsets of Ω . It will always be assumed that the family \mathcal{F}_t , $t \in R_+$, is increasing, i.e., $\mathcal{F}_s \subset \mathcal{F}_t$ for $s \leq t$ and right-continuous, i.e., $\mathcal{F}_t = \bigcap_{s>t} \mathcal{F}_s$. Let $\mathcal{F} = \bigvee_t \mathcal{F}_t$ be the smallest σ -field containing all the \mathcal{F}_t . Let P be a probability measure on (Ω, \mathcal{F}) . Thus one has a family of probability spaces $(\Omega, \mathcal{F}_t, P)$. It will always be assumed that probability spaces are complete.

Let (Z, \mathcal{Z}) be a measurable space. Let $x: \Omega \times R_+ \rightarrow Z$ be a function such that $\{\omega | x_t(\omega) \in B\} \in \mathcal{F}_t$ for all $B \in \mathcal{Z}$, $t \in R_+$. Then (x_t, \mathcal{F}_t, P) is a (stochastic) process. Thus every process has attached to it a family $(\Omega, \mathcal{F}_t, P)$, $t \in R_+$, of probability spaces. The same function x defines a different process if either the family \mathcal{F}_t or the measure P is changed. When the context makes it clear we write (x_t, \mathcal{F}_t) or (x_t, P) or x_t instead of (x_t, \mathcal{F}_t, P) . If (x_t, \mathcal{F}_t, P) is a process, then so is $(x_t, \mathcal{F}_t^x, P)$ where \mathcal{F}_t^x is the sub- σ -field of \mathcal{F}_t generated by x_s , $s \leq t$, and P is the restriction to $\mathcal{F}^x = \bigvee_t \mathcal{F}_t^x$. Two processes (x_t, \mathcal{F}_t, P) and (y_t, \mathcal{F}_t, P) are said to be *modifications* or *versions* of one another if $x_t = y_t$ a.s. P for each t , the set $\{x_t \neq y_t\}$ may vary with t . They are said to be *indistinguishable* if there is a set N with $P(N) = 0$ such that for $\omega \notin N$, $x_t(\omega) = y_t(\omega)$ for all t . Given (Ω, \mathcal{F}, P) , a *random variable*, or r.v., with values in (Z, \mathcal{Z}) is a \mathcal{F} -measurable map from Ω into Z . Unless explicitly stated otherwise all r.v.s and processes take values in $(R \cup \{\infty\}, \mathcal{B})$, where \mathcal{B} is the Borel field.

2.2. Stopping times. Consider a family $(\Omega, \mathcal{F}_t, P)$. A nonnegative r.v. T is a *stopping time*, s.t., of the family if

$$\{T \leq t\} \in \mathcal{F}_t \quad \text{for all } t.$$

The s.t. T is said to be *predictable* if there exists an increasing sequence of s.t.s,

$S_1 \leq S_2 \leq \dots$, such that

$$P\left\{T = 0 \text{ or } S_k < T \text{ for all } k \text{ and } \lim_{k \rightarrow \infty} S_k = T\right\} = 1.$$

The s.t. T is said to be *totally inaccessible* if $T > 0$ a.s. and if for every increasing sequence of s.t.s $S_1 \leq S_2 \leq \dots$,

$$P\left\{S_k < T \text{ for all } k \text{ and } \lim_{k \rightarrow \infty} S_k = T < \infty\right\} = 0.$$

2.3. Martingales and increasing processes. A process (m_t, \mathcal{F}_t, P) is said to be a (uniformly integrable) *martingale* if the collection $\{m_t | t \in R^+\}$ of r.v.s is uniformly integrable, and if $E(m_t | \mathcal{F}_s) = m_s$ a.s. for $s \leq t$. The collection of all such martingales, for which $m_0 = 0$, is denoted $\mathcal{M}^1 = \mathcal{M}^1(\mathcal{F}_t, P)$. (m_t, \mathcal{F}_t, P) is said to be a *local martingale* if there is an increasing sequence of s.t.s S_k , with $S_k \rightarrow \infty$ a.s. such that

$$(m_{t \wedge S_k} I_{\{S_k > 0\}}, \mathcal{F}_t, P) \in \mathcal{M}^1 \text{ for each } k.$$

The collection is denoted $\mathcal{M}_{\text{loc}}^1(\mathcal{F}_t, P)$. (m_t, \mathcal{F}_t, P) is a *square integrable martingale* if $m_t \in \mathcal{M}^1$ and if $\sup_t E m_t^2 < \infty$. The collection is denoted $\mathcal{M}^2(\mathcal{F}_t, P)$ and the class of locally square integrable martingales $\mathcal{M}_{\text{loc}}^2(\mathcal{F}_t, P)$ is defined analogously. It is obvious that $\mathcal{M}_{\text{loc}}^2 \subset \mathcal{M}_{\text{loc}}^1$.

Each $m_t \in \mathcal{M}_{\text{loc}}^1$ has a version whose sample paths are right-continuous and have left-hand limits. Clearly such a version is unique, i.e., unique modulo indistinguishability. It will always be assumed that local martingales have sample paths with this continuity property.

A process (a_t, \mathcal{F}_t, P) is said to be *increasing* if $a_0 = 0$ a.s. and if its sample paths are right-continuous and nondecreasing. The collection is denoted

$$\mathcal{A}_0^+(\mathcal{F}_t, P).$$

$$\mathcal{A}_0 = \mathcal{A}_0^+ - \mathcal{A}_0^+ = \{a_t - a'_t | a_t \in \mathcal{A}_0^+, a'_t \in \mathcal{A}_0^+\}.$$

$$\mathcal{A}^+ = \left\{a_t \in \mathcal{A}_0^+ \mid \sup_t E a_t < \infty\right\}, \quad \mathcal{A} = \mathcal{A}^+ - \mathcal{A}^+.$$

Members of $\mathcal{A}^+(\mathcal{A})$ are said to be *integrable* (or have *integrable variation*). $a_t \in \mathcal{A}_0^+$ is said to be *locally integrable* $a_t \in \mathcal{A}_{\text{loc}}^+$ if there is an increasing sequence of s.t.s $S_k \rightarrow \infty$ a.s. such that $a_{t \wedge S_k} \in \mathcal{A}^+$ for all k . $\mathcal{A}_{\text{loc}} = \mathcal{A}_{\text{loc}}^+ - \mathcal{A}_{\text{loc}}^+$.

Semimartingales. A process (s_t, \mathcal{F}_t, P) is a *semimartingale*, respectively *local semimartingale*, if it can be expressed as $s_t = s_0 + m_t + a_t$, where $m_t \in \mathcal{M}^1(\mathcal{F}_t, P)$ and $a_t \in \mathcal{A}(\mathcal{F}_t, P)$, respectively, $m_t \in \mathcal{M}_{\text{loc}}^1(\mathcal{F}_t, P)$ and $a_t \in \mathcal{A}_0(\mathcal{F}_t, P)$. The families are respectively denoted $\mathcal{S}(\mathcal{F}_t, P)$ and $\mathcal{S}_{\text{loc}}(\mathcal{F}_t, P)$.

2.4. Predictable processes. The family of all processes (y_t, \mathcal{F}_t, P) which have left-continuous sample paths generates a σ -field $\mathcal{P} = \mathcal{P}(\mathcal{F}_t) \subset \mathcal{F} \otimes \mathcal{B}$, where \mathcal{B} is the Borel field of R_+ , with respect to which the functions $(\omega, t) \mapsto y_t(\omega)$ are measurable. \mathcal{P} is called the *predictable* σ -field, and every process (y_t, \mathcal{F}_t, P)

which is \mathcal{P} -measurable is called a *predictable process*. Note that if $\mathcal{F}_t \subset \mathcal{G}_t$, then $\mathcal{P}(\mathcal{F}_t) \subset \mathcal{P}(\mathcal{G}_t)$.

For $(a_t, \mathcal{F}_y, P) \in \mathcal{A}_0$,

$$L^p(a_t) = \left\{ y_t | (y_t, \mathcal{F}_t, P) \text{ is predictable and } E \int_0^\infty |y_t|^p |da_t| < \infty \right\}.$$

$$L_{\text{loc}}^p(a_t) = \{ y_t | \text{there is a sequence of s.t.s } S_k \rightarrow \infty \text{ such that}$$

$$y_t I_{\{t \leq S_k\}} \in L^p(a_t) \text{ for each } k \}.$$

The integrals above are Stieltjes integrals.

2.5. Quadratic variation. Two martingales m_t, n_t in $\mathcal{M}_{\text{loc}}^1$ are *orthogonal* if their product, $m_t n_t \in \mathcal{M}_{\text{loc}}^1$. $m_t \in \mathcal{M}_{\text{loc}}^1$ is *continuous* if its sample paths are continuous; it is said to be *discontinuous* if it is orthogonal to every continuous martingale. Every $m_t \in \mathcal{M}_{\text{loc}}^1$ has a unique *decomposition*,

$$m_t = m_t^c + m_t^d$$

such that m_t^c is continuous and m_t^d is discontinuous. Clearly if $m_t \in \mathcal{M}_{\text{loc}}^1$ is continuous, then it is in $\mathcal{M}_{\text{loc}}^2$. To every path m_t, n_t in $\mathcal{M}_{\text{loc}}^2$ is associated a unique predictable process, denoted $\langle m, n \rangle_t$ or $\langle m_t, n_t \rangle, \mathcal{F}_t, P$ such that $\langle m, n \rangle_t \in \mathcal{A}_{\text{loc}}^+$, and

$$(m_t n_t - \langle m, n \rangle_t) \in \mathcal{M}_{\text{loc}}^1(\mathcal{F}_t, P).$$

$\langle m, n \rangle_t$ is called the *predictable quadratic covariation* of m_t, n_t . For $m_t \in \mathcal{M}_{\text{loc}}^2$, $\langle m \rangle_t = \langle m, m \rangle_t$ is the *predictable quadratic variation* of m_t . Note that generally $\langle m, n \rangle$ depends crucially upon the family (\mathcal{F}_t, P) .

If m_t, n_t in $\mathcal{M}_{\text{loc}}^1$ have the decompositions $m_t = m_t^c + m_t^d, n_t = n_t^c + n_t^d$, then the process

$$[m, n]_t = [m_t, n_t] = \langle m^c, n^c \rangle_t + \sum_{s \leq t} \Delta m_s' \Delta n_s,$$

where $\Delta m_s = m_s - m_{s-}, \Delta n_s = n_s - n_{s-}$, is called the *quadratic covariation* of m_t, n_t and $[m]_t = [m_t, m_t]$ is the *quadratic variation* of m_t . It turns out that

$$m_t n_t - [m, n]_t \in \mathcal{M}_{\text{loc}}^1$$

so that if, furthermore, m_t, n_t are in $\mathcal{M}_{\text{loc}}^2$, then

$$[m, n]_t - \langle m, n \rangle_t \in \mathcal{M}_{\text{loc}}^1.$$

2.6. Stochastic integration. If $m_t \in \mathcal{M}_{\text{loc}}^2(\mathcal{F}_t, P)$ and $\phi_t \in L_{\text{loc}}^2(\langle m \rangle_t)$, then $\phi_t \in L_{\text{loc}}^1(\langle m, n \rangle_t)$ for all $n_t \in \mathcal{M}_{\text{loc}}^2(\mathcal{F}_t, P)$ and there is a unique process, denoted $(\phi \circ m)_t \in \mathcal{M}_{\text{loc}}^2(\mathcal{F}_t, P)$, which satisfies

$$(2.1) \quad \langle \phi \circ m, n \rangle_t = \int_0^t \phi_s d\langle m, n \rangle_s \quad \text{for all } n_t \in \mathcal{M}_{\text{loc}}^2.$$

The integral on the right is a Stieltjes integral. If $m_t \notin \mathcal{M}_{\text{loc}}^2$ then one *cannot* define a stochastic integral in this way. Two other possibilities are open.

If $m_t = m_t^c + m_t^d \in \mathcal{M}_{\text{loc}}^1(\mathcal{F}_t, P)$, if $m_t^d \in \mathcal{A}_{\text{loc}}(\mathcal{F}_t, P)^1$ and if $\phi_t \in L_{\text{loc}}^2(\langle m^c \rangle_t) \cap L_{\text{loc}}^1(m_t^d)$, then the process

$$(2.2) \quad (\phi \circ m)_t = (\phi \circ m^c)_t + \int_0^t \phi_s dm_s^d \in \mathcal{M}_{\text{loc}}^1(\mathcal{F}_t, P),$$

where $(\phi \circ m^c)_t$ is defined as in (2.1) whereas the second integral is a Stieltjes integral.

Finally if $m_t \in \mathcal{M}_{\text{loc}}^1$ and if $(\phi_t, \mathcal{F}_t, P)$ is a *locally bounded*² predictable process, then there exists a unique process $(\phi \circ m)_t \in \mathcal{M}_{\text{loc}}^1$ which satisfies

$$(2.3) \quad [\phi \circ m, n]_t = \int_0^t \phi_s d[m, n]_s \quad \text{for all } n \in \mathcal{M}_{\text{loc}}^1.$$

The integral on the right is not in general a Stieltjes integral unless $[m, n]_t \in \mathcal{A}_{\text{loc}}$. The precise interpretation of this integral is not given here since it is seldom used below. For details see [1].

The process $(\phi \circ m)_t$ is called the *stochastic integral of ϕ with respect to m* . Note that if $(\phi \circ m)$ makes sense according to more than one of the three possibilities (2.1), (2.2) or (2.3), then the resulting stochastic integrals coincide.

2.7. Differentiation formula. Let $s_t = s_0 + m_t + a_t \in \mathcal{S}_{\text{loc}}(\mathcal{F}_t, P)$. The decomposition is *not* unique. If $s_t = s_0 + m'_t + a'_t$ is another decomposition, then the continuous parts m_t^c, m'_t^c of the local martingale are indistinguishable. This unique continuous local martingale is denoted s_t^c .

Let $s_t = (s_t^1, \dots, s_t^n)$ be a process with values in R^n such that $s_t^i \in \mathcal{S}_{\text{loc}}(\mathcal{F}_t, P)$, $i = 1, \dots, n$. Let $F: R^n \rightarrow R$ be a twice continuously differentiable function. Then the following differentiation formula holds:

$$\begin{aligned} F(s_t) = F(s_0) &+ \int_0^t \sum_{i=1}^n \frac{\partial F}{\partial x_i}(s_{\tau-}) ds_{\tau}^i + \frac{1}{2} \int_0^t \sum_{i,j=1}^n \frac{\partial^2 F}{\partial x_i \partial x_j}(s_{\tau-}) d\langle s^{ic}, s^{jc} \rangle_{\tau} \\ &+ \sum_{\tau \leq t} \left[F(s_{\tau}) - F(s_{\tau-}) - \sum_{i=1}^n \frac{\partial F}{\partial x_i}(s_{\tau-})(s_{\tau}^i - s_{\tau-}^i) \right]. \end{aligned}$$

As a special case one obtains the very useful “product” rule. Suppose m_t and n_t are in $\mathcal{M}_{\text{loc}}^1$. Then (since $m_0 = n_0 = 0$), and recalling the definition of $[m, n]_t$,

$$m_t n_t = \int_0^t m_{s-} dn_s + \int_0^t n_{s-} dm_s + [m, n]_t.$$

2.8. The exponentiation formula. Let $s \in \mathcal{S}_{\text{loc}}(\mathcal{F}_t, P)$ with $s_0 = 0$. Then there is a unique process $y_t \in \mathcal{S}_{\text{loc}}(\mathcal{F}_t, P)$ which satisfies the equation

$$y_t = y_0 + \int_0^t y_{\tau-} ds_{\tau}, \quad t \geq 0,$$

¹ This is a nontrivial restriction on m_t^d . It holds for the discontinuous martingales to be introduced in § 2.9 below.

² ϕ_t is locally bounded if there is an increasing sequence of s.t.s $S_k \rightarrow \infty$ such that the process $\phi_t \wedge S_k I_{\{S_k > 0\}}$ is bounded for all k . Note that if ϕ_t is a right-continuous process, having left-hand limits, then the process $\psi_t = \phi_{t-}$ is locally bounded.

for a prespecified \mathcal{F}_0 -measurable y_0 , and y_t is given explicitly by

$$y_t = y_0 \exp(s_t - \frac{1}{2} \langle s^c, s^c \rangle_t) \cdot \prod_{\tau \leq t} (1 + \Delta s_\tau) e^{-\Delta s_\tau},$$

where the second term converges a.s. y_t is called the *exponential of s_t* and is sometimes denoted $y_t = \mathcal{E}(s_t)$. Evidently $\mathcal{E}(s_t) \geq 0$ a.s. if $y_0 \geq 0$ a.s., and if $1 + \Delta s \geq 0$ a.s. If, in addition, $m_0 \geq 0$ and $m_t - m_0 \in \mathcal{M}_{\text{loc}}^1(\mathcal{F}_t, P)$, then $(\mathcal{E}(m_t), \mathcal{F}_t, P)$ is a supermartingale, i.e.,

$$E(\mathcal{E}(m_t) | \mathcal{F}_s) \leq \mathcal{E}(m_s), \quad s \leq t,$$

and so in particular,

$$E(\mathcal{E}(m_t)) \leq E(m_0), \quad t \geq 0.$$

Finally if $m_t \in \mathcal{M}^1$ is bounded, then $\mathcal{E}(m_t)$ is a martingale.

2.9. The fundamental jump process. Let $(\Omega, \mathcal{F}_t, P)$ be a family of spaces and let (x_t, \mathcal{F}_t, P) be a process with values in (Z, \mathcal{Z}) such that all the sample paths of x are piecewise constant and have only a finite number of discontinuities in every finite interval, and such that the sample paths are right-continuous, i.e., for all ω , t there is $\varepsilon_0 > 0$ such that $x_t(\omega) = x_{t+\varepsilon}(\omega)$ for $0 \leq \varepsilon \leq \varepsilon_0$. Let T_n , $n = 0, 1, \dots$, denote the *jump times* of the process, defined inductively by $T_0 \equiv 0$ and

$$T_{n+1}(\omega) = \begin{cases} \inf \{t | t > T_n(\omega), x_t(\omega) \neq x_{T_n}(\omega)\}, & n \geq 0, \\ \infty & \text{if the set above is empty.} \end{cases}$$

(x_t, \mathcal{F}_t, P) is a *fundamental jump process*, or a *fundamental process*, f.p., with values in (Z, \mathcal{Z}) , if in addition,

(i) (Z, \mathcal{Z}) is a Blackwell space, and then it turns out that the jump times are s.t.s, and

(ii) The s.t.s T_n are totally inaccessible.

Evidently if (x_t, \mathcal{F}_t, P) is a f.p., so is $(x_t, \mathcal{F}_t^x, P)$, where \mathcal{F}_t^x is the sub- σ -field of \mathcal{F}_t generated by x_s , $s \leq t$. For each $B \in \mathcal{Z}$, let

$$P(B, t) = \sum_{s \leq t} I_{\{x_s^- \neq x_s\}} I_{\{x_s \in B\}}$$

be the number of jumps of x which occur prior to t and which end in the set B .

Associated with $P(B, t)$ are two unique increasing continuous processes $\tilde{P}(B, t) \in \mathcal{A}_{\text{loc}}^+(\mathcal{F}_t, P)$ and $\tilde{P}^x(B, t) \in \mathcal{A}_{\text{loc}}^+(\mathcal{F}_t^x, P)$ such that

$$Q(B, t) = P(B, t) - \tilde{P}(B, t) \in \mathcal{M}_{\text{loc}}^2(\mathcal{F}_t, P)$$

and $Q^x(B, t) = P(B, t) - \tilde{P}^x(B, t) \in \mathcal{M}_{\text{loc}}^2(\mathcal{F}_t^x, P)$. Furthermore,

$$\langle Q(B_1, t), Q(B_2, t) \rangle = \tilde{P}(B_1 \cap B_2, t),$$

and

$$\langle Q^x(B_1, t), Q^x(B_2, t) \rangle = \tilde{P}^x(B_1 \cap B_2, t).$$

Finally, the functions P , \tilde{P} , \tilde{P}^x , Q , Q^x considered as random set functions on \mathcal{Z} are countably additive.

Note. The condition that the T_n are totally inaccessible is equivalent to the assertion that the $\tilde{P}(B, t)$ are continuous. See [4] for alternative conditions.

A real-valued function $f(z, t) = f(z, \omega, t)$ is said to be *predictable*, and one writes $f \in \mathcal{P}(\mathcal{F}_t)$, if it is measurable with respect to $\mathcal{Z} \otimes \mathcal{F} \otimes \mathcal{B}$ and if for each fixed z , $f(z, \cdot, \cdot)$ is predictable in the sense of § 2.4 above. The family $\mathcal{P}(\mathcal{F}_t^x)$ is defined similarly. If $f \in \mathcal{P}(\mathcal{F}_t)$, respectively $\mathcal{P}(\mathcal{F}_t^x)$, we call f a \mathcal{F}_t -predictable, respectively \mathcal{F}_t^x -predictable, process. The following classes of predictable functions are used in the martingale representation results:

$$\begin{aligned} L^2(\tilde{P}^x) &= \left\{ f \in \mathcal{P}(\mathcal{F}_t^x) \mid \|f\|_2^2 = E \int_Z \int_{R_+} f^2(z, t) \tilde{P}^x(dz, dt) < \infty \right\}, \\ L^1(\tilde{P}^x) &= \left\{ f \in \mathcal{P}(\mathcal{F}_t^x) \mid \|f\|_1 = E \int_Z \int_{R_+} |f(z, t)| \tilde{P}^x(dz, dt) < \infty \right\}, \\ L^1(P) &= \left\{ f \in \mathcal{P}(\mathcal{F}_t^x) \mid \|f\|_1 = E \int_Z \int_{R_+} |f(z, t)| P(dz, dt) < \infty \right\}, \\ L^1(Q^x) &= L^1(P) \cap L^1(\tilde{P}^x). \end{aligned}$$

It turns out that $\|f\|_1 = \|f\|_1^\sim$, hence $L^1(\tilde{P}^x) = L^1(P) = L^1(Q^x)$.

$L_{\text{loc}}^2(\tilde{P}^x) = \{ f \in \mathcal{P}(\mathcal{F}_t^x) \mid \text{there exists a sequence of s.t.s } S_k \rightarrow \infty \text{ such that}$

$$f(z, t)I_{\{t \leq S_k\}} \in L^2(\tilde{P}^x) \text{ for each } k \}.$$

The classes $L_{\text{loc}}^1(P)$ etc. are defined in a similar manner. Evidently,

$$L_{\text{loc}}^1(Q^x) = L_{\text{loc}}^1(P) = L_{\text{loc}}^1(\tilde{P}^x).$$

Let $f(z, \omega, t)$ be a function which is measurable with respect to $\mathcal{Z} \otimes \mathcal{F}^x \otimes \mathcal{B}$ such that $f(z, \cdot, t)$ is \mathcal{F}_t^x -measurable for fixed z and such that

$$E \int_Z \int_{R_+} |f(z, t)| \tilde{P}^x(dz, dt) < \infty.$$

Then there exists a \mathcal{F}_t^x -predictable function \hat{f} such that

$$E \int_Z \int_{R_+} |f - \hat{f}| \tilde{P}^x(dz, dt) = 0.$$

This result follows easily from [22, § V, Thm. 23]. The result will be used in § 4, § 5 in the following context: Let $f \in \mathcal{P}(\mathcal{F}_t)$ and let $\hat{f}(z, t) = E(f(z, t) | \mathcal{F}_t^x)$; it can then be assumed without loss of generality that \hat{f} is \mathcal{F}_t^x -predictable.

2.10. Representation of $\mathcal{M}^2(\mathcal{F}_t^x)$. For each $f \in L^2(Q^x)$ there exists a unique process $(f \circ Q^x)_t \in \mathcal{M}^2(\mathcal{F}_t^x)$ such that for all $g \in L^2(Q^x)$, α and β in R ,

$$(\alpha f + \beta g) \circ Q^x \equiv \alpha(f \circ Q^x) + \beta(g \circ Q^x),$$

$$\langle f \circ Q^x, g \circ Q^x \rangle_t = \int_Z \int_0^t f(z, s) g(z, s) \tilde{P}^x(dz, ds).$$

Conversely if $m_t \in \mathcal{M}^2(\mathcal{F}_t^x)$, then there exists $f \in L^2(Q^x)$ such that

$$m_t = (f \circ Q^x)_t.$$

Similarly, $m_t \in \mathcal{M}_{\text{loc}}^2(\mathcal{F}_t^x)$ if and only if there exists $f \in L_{\text{loc}}^2(Q^x)$ such that

$$m_t = (f \circ Q^x)_t.$$

2.11. Representation of $\mathcal{M}_{\text{loc}}^1(\mathcal{F}_t^x)$. If $f \in L^1(P^x)$, then $(f \circ Q^x)_t \in \mathcal{M}^1(\mathcal{F}_t^x) \cap \mathcal{A}$, where

$$\begin{aligned} (f \circ Q^x)_t &= \int_Z \int_0^t f(z, s) Q^x(dz, ds) = \int_Z \int_0^t f(z, s) P(dz, ds) \\ &\quad - \int_Z \int_0^t f(z, s) \tilde{P}^x(dz, ds), \end{aligned}$$

the integrals on the right being Stieltjes integrals. Conversely if $m_t \in \mathcal{M}^1(\mathcal{F}_t^x) \cap \mathcal{A}$, then there is $f \in L^1(\tilde{P}^x)$ such that

$$m_t = (f \circ Q^x)_t.$$

Finally, $m_t \in \mathcal{M}_{\text{loc}}^1(\mathcal{F}_t^x)$ if and only if there is $f \in L_{\text{loc}}^1(\tilde{P}^x)$ such that

$$m_t = (f \circ Q^x)_t = \int_Z \int_0^t f(z, s) [P(dz, ds) - \tilde{P}^x(dz, ds)].$$

Remark 2.1.1. If $m_t \in \mathcal{M}_{\text{loc}}^1(\mathcal{F}_t^x)$ has continuous sample paths, then $m_t \equiv 0$.

2. If more than one representation above applies then the representations coincide.

2.12. Local description of a fundamental process. Let (x_t, \mathcal{F}_t, P) be a fundamental process with values in (Z, \mathcal{Z}) , and consider the increasing processes $\tilde{P}(B, t)$ and $\tilde{P}^x(B, t)$. Let $\Lambda(t) = \tilde{P}(Z, t)$, $\Lambda^x(t) = \tilde{P}^x(Z, t)$. The countable additivity of these functions with respect to $B \in \mathcal{Z}$ implies that there exist predictable processes $n(B, t)$ and $n^x(B, t)$ such that for all $B \in \mathcal{Z}$,

$$\tilde{P}(B, t) = \int_0^t n(B, s) \Lambda(ds),$$

$$\tilde{P}^x(B, t) = \int_0^t n^x(B, s) \Lambda^x(ds).$$

Evidently it can be assumed that $n(Z, s) = n^x(Z, s) \equiv 1$. The system $\{n(B, t), \Lambda(t)\}$ or $\{n(dz, t), \Lambda(dt)\}$ is analogous to a Lévy system for a Hunt process [5]. The system $\{n(dz, t), \Lambda(dt)\}$ will be called an *extrinsic local description* of x , whereas $\{n^x(dz, t), \Lambda^x(dt)\}$ is called the *intrinsic local description* of x , because of the following interpretation: the probability that x has a jump in $[t, t + dt]$ given \mathcal{F}_t , (respectively \mathcal{F}_t^x) is $\Lambda(dt) + o(dt)$ (respectively $\Lambda^x(dt) + o(dt)$), while $n(B, t)$, (respectively $n^x(B, t)$) is the probability that $x_t \in B$ given $\mathcal{F}_t(\mathcal{F}_t^x)$ and given that a jump occurs at t . For future reference we note the following trivial but important fact.

Fact. Let $\{n(B, t), \Lambda(t)\}$ and $\{n^x(B, t), \Lambda^x(t)\}$ be extrinsic and intrinsic local descriptions. Then for all $B \in \mathcal{Z}$, and $t \in R_+$,

$$(2.4) \quad E \left\{ \int_0^t n(B, s) \Lambda(ds) | \mathcal{F}_t^x \right\} = \int_0^t n^x(B, s) \Lambda^x(ds) \quad \text{a.s.}$$

2.13. Fundamental example. The results in the succeeding sections will be specialized to the following example which covers many practical cases such as Poisson, counting, birth and death, and queueing processes.

Let (x_t, \mathcal{F}_t, P) be a fundamental process with values in (Z, \mathcal{Z}) . Suppose that from each $z \in Z$ the process can make at most n transitions, where n is a fixed finite number. Thus the transitions can be represented by a “state-transition” diagram of Fig. 1, where the transitions are labeled $\sigma_1, \dots, \sigma_n$. Define the counting processes $p_i(t)$, $1 \leq i \leq n$,

$p_i(t)$ = number of transitions of type i made by the process x_t prior to t .

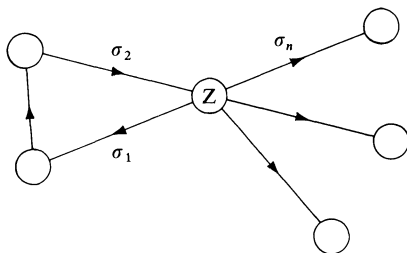


FIG. 1. State-transition diagram for fundamental example

Then there exist increasing processes $\tilde{p}_i(t)$ and $\tilde{p}_i^x(t)$ such that

$$q_i(t) = p_i(t) - \tilde{p}_i(t) \in \mathcal{M}_{\text{loc}}^2(\mathcal{F}_t, P),$$

$$q_i^x(t) = p_i(t) - \tilde{p}_i^x(t) \in \mathcal{M}_{\text{loc}}^2(\mathcal{F}_t^x, P).$$

Furthermore, $m_t \in \mathcal{M}^2(\mathcal{F}_t^x, P)$, respectively $\mathcal{M}_{\text{loc}}^2(\mathcal{F}_t^x, P)$, if and only if there exist $f_i \in L^2(\tilde{p}_i^x)$, respectively $L_{\text{loc}}^2(\tilde{p}_i^x)$, such that

$$(2.5) \quad m_t = \sum_{i=1}^n (f_i \circ q_i^x)_t;$$

and $m_t \in \mathcal{M}^1(\mathcal{F}_t^x, P) \cap \mathcal{A}$, respectively $\mathcal{M}_{\text{loc}}^1(\mathcal{F}_t^x, P)$, if and only if there exist $f_i \in L^1(\tilde{p}_i^x)$, respectively $L_{\text{loc}}^1(\tilde{p}_i^x)$, such that

$$(2.6) \quad m_t = \sum_{i=1}^n (f_i \circ q_i^x)_t,$$

where the integral is a Stieltjes integral.

We call $(\tilde{p}_1, \dots, \tilde{p}_n)$, respectively $(\tilde{p}_1^x, \dots, \tilde{p}_n^x)$, the extrinsic, respectively intrinsic, local descriptions.

Remark 2.2. If (x_t, \mathcal{F}_t, P) is a counting process,³ then Fig. 1 simplifies to Fig. 2 and there is only one transition. Hence in this case $n = 1$ in (2.5) and (2.6).

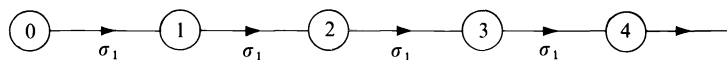


FIG. 2. State-transition diagram for counting processes

For this special case Brémaud [6] has obtained the representation for $\mathcal{M}_{\text{loc}}^2(\mathcal{F}_t^x)$, whereas Davis [7] has extended it to the class $\mathcal{M}_{\text{loc}}^1(\mathcal{F}_t^x)$. However both these results were obtained only for the case where the law of $(x_t, \mathcal{F}_t^x, P)$ is mutually absolutely continuous with respect to the law for a standard Poisson process (see § 3).

3. Solutions to specified local descriptions by change of law. In § 3.1 we present a very useful technique for transforming one fundamental process (x_t, \mathcal{F}_t, P) with a l.d. (local description) (n, Λ) to another process with a different prespecified l.d. The questions of uniqueness of the solution is discussed in § 3.2. Section 3.3 consists of some sufficient conditions which guarantee that the technique is applicable. Finally § 3.4 presents a class of processes which can be transformed into other processes with this technique.

Let $(x_t, \mathcal{F}_t^x, P)$ be a fundamental process with values in (Z, \mathcal{Z}) and with intrinsic local description (l.d.) $(n^x(dz, t), \Lambda^x(ds))$ so that

$$\tilde{P}^x(B, t) = \int_B \int_0^t n^x(dz, s) \Lambda^x(ds), \quad t \in \mathbb{R}_+.$$

Since we will be only dealing with the “intrinsic” σ -field \mathcal{F}_t^x in this section, the superscript x will be omitted here. Hence $\mathcal{F}_t = \mathcal{F}_t^x$, $\tilde{P} = \tilde{P}^x$ etc.

3.1. The transformation technique. Let P_1 be another probability measure on (Ω, \mathcal{F}) and suppose that

$$P_1 \ll P,$$

i.e., P_1 is absolutely continuous with respect to P . It is evident that the same function $x_t(\omega)$ defines another fundamental process $(x_t, \mathcal{F}_t, P_1)$ with a possibly different l.d. $(n_1(B, t), \Lambda_1(t))$ say. We are going to determine the relationship between the two descriptions.

Let $L = dP_1/dP$ be the Radon–Nikodym derivative. The r.v. $L \geq 0$ and $E(L)^4 = 1$. Let $L_t = E(L | \mathcal{F}_t)$. Then (L_t, \mathcal{F}_t, P) is a uniformly integrable martingale, $\lim_{t \rightarrow \infty} L_t = L$ a.s. and in L^1 by [3, remark after § VI, Thm. 6].

PROPOSITION 3.1. (i) If $L > 0$ a.s. P , then for almost all ω , $L_{t-}(\omega) > 0$ and $L_t(\omega) > 0$ for all t .

(ii) Let

$$(3.1) \quad T(\omega) = \inf \{t | L_{t-}(\omega) = 0 \text{ or } L_t(\omega) = 0\}.$$

³ A counting process is an integer-valued process which starts at 0 and has unit jumps.

⁴ E, E_1 , denotes expectation with respect to P, P_1 .

Then for almost all ω , $L_t(\omega) = 0$ for $t \geq T(\omega)$.

Proof. (i) Clearly $L > 0$ a.s. implies $L_t > 0$ a.s. and then the second part of the assertion follows from (ii), and the latter follows from [3, § VI, Thm. 15]. \square

Remarks 3.1. (i) If $L > 0$ a.s. P , then in fact $P \ll P_1$, i.e., the two measures are mutually absolutely continuous.

(ii) It is easy to give examples such that $L_t > 0$ for all t but $P(L = 0) > 0$. For $\varepsilon > 0$ let

$$(3.2) \quad T_\varepsilon(\omega) = \inf \{t \mid L_{t-}(\omega) \leq \varepsilon\}.$$

PROPOSITION 3.2. T_ε is a s.t. for all ε and

$$\lim_{\varepsilon \rightarrow 0} T_\varepsilon(\omega) = T(\omega) \quad \text{a.s. } P$$

Proof. The fact that T_ε is a s.t. follows from the fact that the process L_{t-} is left-continuous and from [3, § IV, Thm. 52]. Now T_ε is clearly nondecreasing with ε . Let

$$T_0(\omega) = \lim_{\varepsilon \rightarrow 0} T_\varepsilon(\omega).$$

Suppose $T(\omega) = \infty$ and per contra $T_0(\omega) < \infty$. Then there exists a sequence t_i increasing to $t_0 < \infty$ such that $L_{t_i-}(\omega) \rightarrow 0$. By left-continuity $L_{t_0-}(\omega) = 0$ and so $T(\omega) \leq t_0$. Next suppose $T(\omega) < \infty$. By Proposition 3.1, for almost all such ω , $T_0(\omega) \leq T(\omega)$. If $T_0(\omega) < T(\omega)$, then a repetition of the previous argument will end in a contradiction. Once again $T_0(\omega) = T(\omega)$. \square

For $\varepsilon > 0$ let

$$L_t^\varepsilon(\omega) = L_{t \wedge T_\varepsilon}(\omega), \quad t \in R_+.$$

Then $L_t^\varepsilon - L_0 \in \mathcal{M}^1(P)$ and $L_{t-}^\varepsilon \geq \varepsilon$ for all t . By § 2.11 there is a predictable function $f^\varepsilon(z, t) \in L_{\text{loc}}^1(P)$ such that

$$(3.3)^5 \quad L_t^\varepsilon = 1 + \int_Z \int_0^t f^\varepsilon(z, s) Q(dz, ds) = \int_Z \int_0^t f^\varepsilon(z, s) [P(dz, ds) - \tilde{P}(dz, ds)].$$

Since $1/L_{t-}^\varepsilon \leq 1/\varepsilon$, therefore the process

$$\phi^\varepsilon(z, s) = [f^\varepsilon(z, s)]/L_{s-}^\varepsilon \in L_{\text{loc}}^1(\tilde{P}),$$

and hence

$$(3.4) \quad m^\varepsilon(t) = \int_Z \int_0^t \phi^\varepsilon(z, s) Q(dz, ds) \in \mathcal{M}_{\text{loc}}^1(\mathcal{F}_t, P)$$

which upon substitution into (3.3) gives

$$L_t^\varepsilon = 1 + \int_0^t L_{s-}^\varepsilon dm_s^\varepsilon.$$

⁵ Here it is being assumed that $L_0 \equiv 1$ which is indeed the case if \mathcal{F}_0 is trivial. Otherwise, in the sequel, replace the martingale L_t by L_t/L_0 .

By the exponentiation formula of § 2.8,

$$(3.5) \quad L_t^\varepsilon = \mathcal{E}(m_t^\varepsilon) = \exp(m_t^\varepsilon - \tfrac{1}{2}\langle m^{\varepsilon,c}, m^{\varepsilon,c} \rangle_t) \prod_{s \leq t} (1 + \Delta m_s^\varepsilon) e^{-\Delta m_s^\varepsilon}.$$

By the Remark in § 2.11, $m^{\varepsilon,c} \equiv 0$, hence (3.5) simplifies to

$$(3.6) \quad L_t^\varepsilon = \exp(m_t^\varepsilon) \prod_{s \leq t} (1 + \Delta m_s^\varepsilon) e^{-\Delta m_s^\varepsilon}.$$

Rewriting (3.4) as

$$(3.7) \quad m^\varepsilon(t) = \int_Z \int_0^t \phi^\varepsilon(z, s) P(dz, ds) - \int_Z \int_0^t \phi^\varepsilon(z, s) \tilde{P}(dz, ds)$$

and acknowledging that the second integral has continuous sample paths (since \tilde{P} is continuous) it follows that for almost all ω ,

$$(3.8) \quad \Delta m_s^\varepsilon(\omega) = m_s^\varepsilon(\omega) - m_{s-}^\varepsilon(\omega) = \int_Z \phi^\varepsilon(z, s)(\omega) [P(dz, s)(\omega) - P(dz, s^-)(\omega)].$$

Also since $P(B, s)(\omega) - P(B, s^-)(\omega)$ equals 1 or 0 depending upon whether or not $x_{s-}(\omega) \neq x_s(\omega)$ and $x_s(\omega) \in B$, therefore the term $(1 + \Delta m_s^\varepsilon)$ in (3.6) can be written as

$$(3.9) \quad (1 + \Delta m_s^\varepsilon)(\omega) = \int_Z (1 + \phi^\varepsilon(z, s)(\omega)) [P(dz, s)(\omega) - P(dz, s^-)(\omega)].$$

From (3.8), (3.9) it follows respectively that

$$\begin{aligned} \sum_{s \leq t} \Delta m_s^\varepsilon(\omega) &= \sum_{\substack{s \leq t \\ x_{s-} \neq x_s}} \phi^\varepsilon(x_s(\omega), s), \\ \prod_{s \leq t} (1 + \Delta m_s^\varepsilon(\omega)) &= \prod_{\substack{s \leq t \\ x_{s-} \neq x_s}} (1 + \phi^\varepsilon(x_s(\omega), s)), \end{aligned}$$

which upon substitution, together with (3.7), into (3.6), yields after some cancellation the first interesting result:

$$(3.10) \quad L_t^\varepsilon = \prod_{\substack{s \leq t \\ x_{s-} \neq x_s}} [1 + \phi^\varepsilon(x_s, s)] \exp \left(- \int_Z \int_0^t \phi^\varepsilon(z, s) \tilde{P}(dz, ds) \right).$$

Finally let $\varepsilon_k > 0$, $k = 1, 2, \dots$, be a sequence decreasing to 0, let $S_0 = 0$, $S_k = T_{\varepsilon_k}$, $k = 1, 2, \dots$, and let

$$\phi(z, s) = \sum_{k=1}^{\infty} \phi^{\varepsilon_k}(z, s) I_{\{S_{k-1} < s \leq S_k\}}.$$

ϕ is predictable since ϕ^{ε_k} is predictable and $I_{\{S_{k-1} < s \leq S_k\}}$ is left-continuous. Since by definition, $L_t^\varepsilon = L_{t \wedge T_\varepsilon}^{\varepsilon'}$ for $\varepsilon' < \varepsilon$, we have proved the following result.

THEOREM 3.1. *Let $P_1 \ll P$, and let $L_t = E(dP_1/dP | \mathcal{F}_t^x)$. Let*

$$T = \inf \{t | L_{t-} = 0 \text{ or } L_t = 0\}.$$

Then there exists a predictable function $\phi(z, s)$ and an increasing sequence S_k of s.t.s converging to T such that

$$\phi_k(z, s) = \phi(z, s)I_{\{s \leq S_k\}} \in L^1_{\text{loc}}(\tilde{P})$$

and

$$(3.11) \quad L_{t \wedge S_k} = \prod_{\substack{s \leq t \\ x_{s-} \neq x_s}} [1 + \phi_k(x_s, s)] \exp \left[- \int_Z \int_0^t \phi_k(z, s) \tilde{P}(dz, ds) \right].$$

The product on the right converges a.s. whereas the integral is a Stieltjes integral.

Remarks 3.2. (i) If $L = dP_1/dP > 0$ a.s., then $T = \infty$ a.s. so that the result above implies that $\phi \in L^1_{\text{loc}}(\tilde{P})$. However if this is not the case then it is *not* true that in general $\phi \in L^1_{\text{loc}}$. Some additional properties of ϕ are given in Theorem 3.2 below. Nevertheless, very loosely speaking, one can interpret (3.11) as

$$(3.12) \quad L_t = \prod_{\substack{s \leq t \\ x_{s-} \neq x_s}} [1 + \phi(x_s, s)] \exp \left[- \int_Z \int_0^t \phi(z, s) \tilde{P}(dz, ds) \right] \quad \text{for } t < T.$$

Indeed some such loose interpretation has to be used in understanding the corresponding formulas of [6], [23].

(ii) The characterization (3.11) has been derived earlier [23], [24] for the case where (x_t, \mathcal{F}_t, P) is a Brownian motion. The techniques for the proof are identical except that in deriving (3.4) one observes that every martingale on a Brownian motion sample space is a stochastic integral of the Brownian motion (see [5]), and that all martingales are continuous so that (3.5) simplifies to

$$L_t^e = \exp(m_t^e - \frac{1}{2} \langle m^{e,c}, m^{e,c} \rangle_t).$$

(iii) For the fundamental example the representation (3.11) becomes, using § 2.12,

$$L_{t \wedge S_k} = \prod_{i=1}^n \left\{ \prod_{\substack{s \leq t \\ (x_{s-}, x_s) \in \sigma_i}} [1 + \phi_k^i(s)] \exp \left[- \int_0^t \phi_k^i(s) \tilde{p}_i^x(ds) \right] \right\}$$

for some predictable $\phi^i(s)$, $1 \leq i \leq n$, such that $\phi_k^i \in L^1_{\text{loc}}(\tilde{p}_i^x)$. Here the notation $(x_{s-}, x_s) \in \sigma_i$ means that x makes a transition of type i at time s .

If $(x_t, \mathcal{F}_t^x, P)$ is a Poisson process then in the above $n = 1$ and, as is well known, $\tilde{p}_1^x(ds) \equiv ds$. For this case the result was first obtained by Brémaud [6] without the integrability condition on ϕ , and for the case $L > 0$ a.s., by Van Schuppen [24], and by Davis [7] who proves in addition that then $\phi \in L^1_{\text{loc}}$. Brémaud [6] also obtains this representation for the case where the example is a Markov chain.

We proceed to obtain the relations between the local descriptions. The next result is well known.

LEMMA 3.1. $m_t \in \mathcal{M}^1_{\text{loc}}(\mathcal{F}_t, P_1)$ if and only if $m_t L_t \in \mathcal{M}^1_{\text{loc}}(\mathcal{F}_t, P)$.

Proof. Let $S_k \rightarrow \infty$ be a sequence of s.t.s such that for each k ,

$$(3.13) \quad m_{t \wedge S_k} L_{t \wedge S_k} I_{\{S_k > 0\}} \in \mathcal{M}^1(P).$$

First of all,

$$\begin{aligned} E_1 |m_t \wedge S_k I_{\{S_k > 0\}}| &= EL |m_t \wedge S_k I_{\{S_k > 0\}}| \\ &= EL_{t \wedge S_k} |m_t \wedge S_k I_{\{S_k > 0\}}| \quad (\text{by (3.13)}) \\ &< \infty. \end{aligned}$$

Next for $s \leq t$,

$$E_1(m_t \wedge S_k I_{\{S_k > 0\}} | \mathcal{F}_s) = \frac{E(m_t \wedge S_k L_{t \wedge S_k} I_{\{S_k > 0\}} | \mathcal{F}_s)}{E(L_{t \wedge S_k} | \mathcal{F}_s)}.$$

From (3.13) and the fact that $L_t \in \mathcal{M}^1(P)$ the right-hand side simplifies to

$$\frac{m_s \wedge S_k L_{s \wedge S_k} I_{\{S_k > 0\}}}{L_{s \wedge S_k}} = m_s \wedge S_k I_{\{S_k > 0\}},$$

which proves the “if” part of the assertion.

Conversely suppose that

$$(3.14) \quad m_t \wedge S_k I_{\{S_k > 0\}} \in \mathcal{M}^1(P_1).$$

It will be shown that for $s \leq t$,

$$(3.15) \quad E(L_{t \wedge S_k} m_t \wedge S_k I_{\{S_k > 0\}} | \mathcal{F}_s) = L_{s \wedge S_k} I_{\{S_k > 0\}} \quad \text{a.s. } P.$$

So let $A \in \mathcal{F}_s$. Then

$$\begin{aligned} E(I_A L_{t \wedge S_k} m_t \wedge S_k I_{\{S_k > 0\}}) &= E_1(I_A m_t \wedge S_k I_{\{S_k > 0\}}) \\ &= E_1(I_A m_{s \wedge S_k} I_{\{S_k > 0\}}) \quad (\text{by (3.14)}) \\ &= E(L_{s \wedge S_k} I_A m_{s \wedge S_k} I_{\{S_k > 0\}}), \end{aligned}$$

which proves (3.15). \square

THEOREM 3.2. Let (x_t, \mathcal{F}_t, P) be a fundamental process with values in (Z, \mathcal{Z}) and with (intrinsic) l.d. $(n(dz, t), \Lambda(dt))$. Let $P_1 \ll P$ and let $L_t = E(dP_1/dP | \mathcal{F}_t)$ have the representation (3.11). Then $(x_t, \mathcal{F}_t, P_1)$ has l.d. $(n_1(dz, t), \Lambda_1(dt))$, where

$$(3.16) \quad \Lambda_1(t) = \Lambda(dt) \quad \text{and} \quad n_1(dz, t) = (1 + \phi(z, t))n(dz, t).$$

Furthermore, it can be assumed that

$$(3.17) \quad (1 + \phi) \geq 0 \quad \text{and} \quad (1 + \phi) \in L_{\text{loc}}^1(\tilde{P})$$

with respect to probability measure P_1 .

Proof. By § 2.9 there exist continuous increasing processes $\tilde{P}_1(B, t) \in \mathcal{A}_{\text{loc}}^+(P_1)$ such that

$$(3.18) \quad Q_1(B, t) = P(B, t) - \tilde{P}_1(B, t) \in \mathcal{M}_{\text{loc}}^2(P_1).$$

Hence to show (3.16) it is equivalent to prove that

$$(3.19) \quad \tilde{P}_1(B, t) = \int_B \int_0^t (1 + \phi(z, s)) \tilde{P}(dz, ds).$$

Let S_i , ϕ_i be as in Theorem 3.1, and let

$$(3.20) \quad Q_1^i(B, t) = Q_1(B, t \wedge S_i) = P(B, t \wedge S_i) - \tilde{P}_1(B, t \wedge S_i),$$

$$(3.21) \quad m_t = P(B, t \wedge S_i) - \int_B \int_0^t (1 + \phi_i(z, s)) \tilde{P}(dz, ds).$$

It will be shown first that $m_t \in \mathcal{M}_{\text{loc}}^1(P_1)$. By Lemma 3.1 it is enough to show that

$$L_t m_t \in \mathcal{M}_{\text{loc}}^1(P).$$

Since $\phi_i \in L_{\text{loc}}^1(P)$, therefore m_t is in $\mathcal{S}_{\text{loc}}(P)$, also $L_t \in \mathcal{M}^1(P) \subset \mathcal{S}_{\text{loc}}(P)$. Hence one can apply the differential formula of § 2.7 to obtain

$$(3.22) \quad L_t m_t = \int_0^t m_{s-} dL_s + \int_0^t L_{s-} dm_s + \sum_{s \leq t} [\Delta(m_s L_s) - m_{s-} \Delta L_s - L_{s-} \Delta m_s].$$

From (3.21),

$$\int_0^t L_{s-} dm_s = \int_0^{t \wedge S_i} L_{s-} P(B, ds) - \int_B \int_0^{t \wedge S_i} L_{s-} (1 + \phi) \tilde{P}(dz, ds),$$

and since $\Delta(m_s L_s) = (m_{s-} + \Delta m_s)(L_{s-} + \Delta L_s) - m_{s-} L_{s-} = m_{s-} \Delta L_s + L_{s-} \Delta m_s + \Delta L_s \Delta m_s$, therefore the last term in (3.22) equals

$$\sum_{s \leq t} \Delta L_s \Delta m_s = \int_B \int_0^{t \wedge S_i} L_{s-} \phi(z, s) P(dz, ds)$$

from (3.11) and (3.21). Substituting these relations back into (3.22) gives

$$\begin{aligned} L_t m_t &= \int_0^t m_{s-} dL_s + \int_B \int_0^{t \wedge S_i} L_{s-} (1 + \phi) P(dz, ds) - \int_B \int_0^{t \wedge S_i} L_{s-} (1 + \phi) \tilde{P}(dz, ds) \\ &= \int_0^t m_{s-} dL_s + \int_B \int_0^{t \wedge S_i} L_{s-} (1 + \phi) Q(dz, ds), \end{aligned}$$

which is clearly in $\mathcal{M}_{\text{loc}}^1(P)$. Hence $m_t \in \mathcal{M}_{\text{loc}}^1(P_1)$. Since $Q_1^i(B, t) \in \mathcal{M}_{\text{loc}}^1(P)$, subtracting (3.21) from (3.20) implies that

$$\tilde{P}_1(B, t \wedge S_i) - \int_B \int_0^{t \wedge S_i} (1 + \phi(z, s)) \tilde{P}(dz, ds) \in \mathcal{M}_{\text{loc}}^1(P_1).$$

But this process has continuous sample paths, hence it must vanish, i.e., for almost all ω (P_1 measure)

$$\tilde{P}_1(B, t \wedge S_i) = \int_B \int_0^{t \wedge S_i} (1 + \phi(z, s)) \tilde{P}(dz, ds) \quad \text{for all } t,$$

which proves (3.19) and thereby (3.16). The assertion contained in (3.17) follows from the fact that \tilde{P}_1 has increasing sample paths and is in $\mathcal{A}_{\text{loc}}^+(P_1)$. \square

Remark 3.3. (i) It has been shown that $\phi \in L_{\text{loc}}^1(\tilde{P})$ in the probability space $(\Omega, \mathcal{F}, P_1)$ and not in (Ω, \mathcal{F}, P) .

(ii) The transformation of l.d. for the case where (x_t, P) and (x_t, P_1) are both Hunt processes has been obtained in [5]. For this case the local description is called a Lévy system.

(iii) For the case of the fundamental example with l.d. $(\tilde{p}_1, \dots, \tilde{p}_n)$ under P , the l.d. under P_1 is $((1 + \phi^1)\tilde{p}_1, \dots, (1 + \phi^n)\tilde{p}_n)$, where the ϕ^i are as in Remark 3.2 (iii).

Theorems 3.1, 3.2 allow us to obtain in certain cases processes which have certain specified l.d. from known processes with other descriptions. Put differently, we have a “synthesis” procedure for obtaining “global” solutions for a class of l.d.s. This is summarized in the following theorem, whose proof is now immediate.

THEOREM 3.3. (Existence of solutions to local descriptions). *Let (x_t, \mathcal{F}_t, P) be a fundamental process with values in (Z, \mathcal{Z}) and with intrinsic l.d. $(n(dz, t), \Lambda(dt))$. Let $\phi(z, s)$ be a predictable function such that*

$$(3.23) \quad \phi(z, s) \in L_{\text{loc}}^1(\tilde{P})$$

and

$$(3.24) \quad \int_{\Omega} L_{\infty} dP = 1,$$

where

$$(3.25) \quad L_t = \prod_{\substack{s \leq t \\ x_s \neq x_s}} [1 + \phi(x_s, s)] \exp \left[- \int_Z \int_0^t \phi(z, s) \tilde{P}(dz, ds) \right].$$

Then $(x_t, \mathcal{F}_t, P_1)$ is a fundamental process with l.d. $(n_1(dz, t), \Lambda(dt))$, where

$$n_1(dz, t) = (1 + \phi(z, t))n(dz, t)$$

and where the probability measure P_1 is given by

$$dP_1 = L_{\infty} dP.$$

Remark 3.4. (i) This result is extremely useful in practice since given an arbitrary l.d. there is no way to determine whether or not there exists a process with such a description. On the other hand, from the viewpoint of dynamical processes, a l.d. is much more natural and useful.

(ii) For the case of Brownian motion the result corresponding to the above was first obtained by Girsanov [9], and the technique was soon adopted in stochastic control problems [10], [11], [12], [13].

(iii) Brémaud [6] was the first to use this result, for the special case where (x_t, \mathcal{F}_t, P) is a Poisson process, to obtain existence of several “self-exciting” counting processes $(x_t, \mathcal{F}_t, P_1)$. Snyder [14] and Rubin [15] introduce several jump processes through their l.d. However they do not discuss whether or not there indeed exist processes with these descriptions. The result above can be used to solve this problem.

(iv) The condition (3.24) is a nontrivial restriction. For the Brownian motion case some sufficient conditions on the local description have been derived which guarantee (3.24). See [10], [11]. For our case similar conditions are given below in § 3.3.

(v) Theorem 3.3 does not address itself to the question of uniqueness of the solution. This question is discussed next.

3.2. Uniqueness of solutions with specified l.d. To discuss uniqueness of laws of solutions it is convenient to assume that Ω is the space of sample functions and that the process x_t on Ω is merely the “evaluation” process, i.e., $x_t(\omega) = \omega_t$. The probability on Ω is then the law of the process. We will be dealing with two such processes, x_t and y_t , with the same set of sample functions but with different laws. Hence we must have two different probability spaces $(\Omega^x, \mathcal{F}_t^x, P^x)$ and $(\Omega^y, \mathcal{F}_t^y, P^y)$, where $(\Omega^x, \mathcal{F}_t^x)$ and $(\Omega^y, \mathcal{F}_t^y)$ are copies of the same family (Ω, \mathcal{F}_t) . In particular, then, x and y are identical functions on $\Omega \times [0, 1]$.

Since we are unable to obtain any interesting results for the infinite time interval, therefore in Theorem 3.4 and Corollary 3.1, $t \in [0, 1]$.

DEFINITION 3.1. An (intrinsic) l.d. (n, Λ) is said to have *unique solutions* if all fundamental processes (x_t, \mathcal{F}_t, P) with l.d. (n, Λ) have the same law.

THEOREM 3.4. Let $x_t, y_t, 0 \leq t \leq 1$, be fundamental processes with values in (Z, \mathcal{Z}) , and on the (sample function) spaces $(\Omega^x, \mathcal{F}_t^x, P^x)$, $(\Omega^y, \mathcal{F}_t^y, P^y)$ respectively. Let (n, Λ) be the l.d. of x and $((1 + \phi)n, \Lambda)$ the l.d. of y for some predictable function ϕ .

Suppose that (n, Λ) has unique solutions, and suppose that for each $\varepsilon > 0$ there exist $Z_\varepsilon \in \mathcal{Z}$ and $k_\varepsilon < \infty$ such that

$$(i) \quad P^x(B_\varepsilon) \geq 1 - \varepsilon,$$

where

$$B_\varepsilon = \{\omega | x_t(\omega) \in Z_\varepsilon \text{ for } 0 \leq t \leq 1\},$$

$$(ii) \quad \int_Z \int_0^1 \left| \frac{\phi(z, \omega, s)}{1 + \phi(z, \omega, s)} \right| (P^y(dz, ds) + \tilde{P}^y(dz, ds)) \leq k_\varepsilon \quad \text{for } \omega \in B_\varepsilon,$$

where these are Stieltjes integrals.

Then

$$\int_\Omega l_1(\omega) P^y(d\omega) = 1 \quad \text{and} \quad dP^x = l_1 dP^y,$$

where

$$l_t = \mathcal{E}((\psi \circ Q^y)_t) \quad \text{and} \quad \psi = -\phi/(1 + \phi).$$

Proof. The process $(m_t, \mathcal{F}_t^y, P^y)$,

$$m_t = \int_Z \int_0^t \psi(z, s) Q^y(dz, ds) = \int_Z \int_0^t \psi(z, s) [P^y(dz, ds) - \tilde{P}^y(dz, ds)]$$

is, by (i) and (ii), well-defined as a Stieltjes integral. Hence $m_t \in \mathcal{A}_0(\mathcal{F}_t^y, P^y)$ so that it is in $\mathcal{S}_{loc}(P^y)$. Therefore by § 2.9 there is a unique process $(l_t, \mathcal{F}_t^y, P^y)$ where

$$l_t = \mathcal{E}(m_t).$$

Let $\varepsilon_n > 0$ be a decreasing sequence converging to 0. Define the predictable functions

$$\psi^n(z, \omega, s) = \begin{cases} \psi(z, \omega, s) & \text{if } z \in Z_{\varepsilon_n}, \\ 0 & \text{otherwise.} \end{cases}$$

Because of (ii), the process $(m_t^n, \mathcal{F}_t^y, P^y)$, where

$$m_t^n = \int_Z \int_0^t \psi^n(z, s) Q^y(dz, ds),$$

is a bounded martingale. Hence by § 2.9 the process $(l_t^n, \mathcal{F}_t^y, P^y)$ is a martingale, where

$$l_t^n = \mathcal{E}(m_t^n).$$

Furthermore from the definition of ψ^n ,

$$(3.26) \quad l_t^n(\omega) = l_t(\omega) \quad \text{for all } t \text{ and } \omega \in B_{\varepsilon_n}.$$

By Theorem 3.3 the fundamental process $(y_t, \mathcal{F}_t^y, P^n)$, where

$$(3.27) \quad \frac{dP^n}{dP^y} = l_1^n,$$

has a l.d. $((1 + \psi^n)(1 + \phi)n, \Lambda)$, and by the definition of ψ^n and ψ ,

$$(1 + \psi^n)(1 + \phi)(z, \omega, t) = 1 \quad \text{for all } t, \omega \in B_{\varepsilon_n} \text{ and } z \in Z_{\varepsilon_n}.$$

Since (n, Λ) has unique solutions, it follows that

$$\begin{aligned} \int_{B_{\varepsilon_n}} P^n(d\omega) &= \int_{B_{\varepsilon_n}} P^x(d\omega) \\ &\geq 1 - \varepsilon \quad (\text{by (i)}). \end{aligned}$$

From (3.26), (3.27) this implies

$$\int_{B_{\varepsilon_n}} l_1(\omega) P^y(d\omega) \geq 1 - \varepsilon,$$

and since $\varepsilon > 0$ is arbitrary, the assertion follows. \square

Note. We must have $P^x(B, t)(\omega) = P^y(B, t)(\omega)$ and $\tilde{P}^y(dz, dt)(\omega) = [1 + \phi(z, t)] \times n(dz, t)\Lambda(dt)$.

COROLLARY 3.1 (uniqueness). *Let (x_t, \mathcal{F}_t, P) be a fundamental process with values in (Z, \mathcal{Z}) and with l.d. (n, Λ) which has unique solutions. Let ϕ be a predictable function such that*

$$\phi(z, s) \in L_{\text{loc}}^1(\tilde{P}^x), \quad E[\mathcal{E}((\phi \circ Q^x)_1)] = 1.$$

Suppose that ϕ satisfies (i) and (ii) of Theorem 3.4. Then the l.d. $((1 + \phi)n, \Lambda)$ has unique solutions.

Proof. By Theorem 3.3 and the hypothesis there is a solution $(y_t, \mathcal{F}_t^y, P_1)$ with l.d. $((1 + \phi)n, \Lambda)$, where $dP_1 = \mathcal{E}((\phi \circ Q^x)_1) dP$.

Suppose $(y_t, \mathcal{F}_t^y, P_2)$ is another solution with l.d. $((1 + \phi)n, \Lambda)$. By Theorem 3.4,

$$dP = l_1 dP_1 = l_2 dP_2,$$

and since $dP_i = L_i dP$ it follows that $l_i > 0$ a.s. P_i , $i = 1, 2$. Evidently then $P_2 = P_1$. \square

Remark 3.5. (i) Theorem 3.4 is inspired by [9, Lemma 7] and the development there suggests how the result can be generalized.

(ii) Conditions (i) and (ii) of Theorem 3.4 are usually easy to verify in practice. Consider a special case of the fundamental example where $(x_t, \mathcal{F}_t^x, P^x)$ is a Poisson process with rate 1. Then Z is the space of integers and y_t is then a counting process with local “intensity” rate $1 + \phi(\omega, t)$. Suppose $\phi(\omega, t)$ is expressed explicitly as a function of the past of x , i.e., $\phi(\omega, t) = f(x_{[0,t]}(\omega), t)$. Then the conditions (i) and (ii) are satisfied if, for instance, there is an increasing function f_0 such that

$$|f(x_{[0,t]}, t)| + \frac{1}{|1 + f(x_{[0,t]}, t)|} \leq f_0(N) \quad \text{when } |x_t| \leq N.$$

For a similar condition in the Brownian motion case see [11].

(iii) Corollary 3.1 extends in an obvious way to the time interval R_+ . However Theorem 3.4 does not.

3.3. Sufficient conditions. Let (x_t, \mathcal{F}_t, P) be a fundamental process with values in (Z, \mathcal{Z}) and with intrinsic description $(n(dz, t), \Lambda(dt))$. Let $\phi(z, t) \in L_{\text{loc}}^1(\tilde{P})$ and define the process L_t , $t \in [0, 1]$, by

$$(3.28) \quad L_t = \prod_{\substack{s \leq t \\ x_s \neq x_s}} [1 + \phi(x_s, s)] \exp \left[- \int_Z \int_0^t \phi(z, s) \tilde{P}(dz, ds) \right].$$

Then L_t also satisfies

$$(3.29) \quad dL_t = L_{t-} dm_t,$$

where

$$(3.30) \quad m_t = (\phi \circ Q)_t.$$

We assume that $1 + \phi(z, t) \geq 0$. Then $L_t \geq 0$, L_t is a supermartingale and

$$E(L_1) \leq 1.$$

The three results below state conditions on ϕ which guarantee

$$E(L_1) = 1.$$

The following assumption is made throughout this subsection.

Assumption 3.1. There exists an increasing function $\mu: R_+ \rightarrow R_+$ such that

$$(3.31) \quad \tilde{P}(Z, t) \leq \mu(t) \quad \text{a.s.}$$

(Note that this implies $\tilde{P}(B, t) \leq \mu(t)$ for all $B \in \mathcal{Z}$).

PROPOSITION 3.3. Suppose that for some $K < \infty$,

$$(3.32) \quad |\phi| \leq K.$$

Then $E(L_1) = 1$.

Proof. From (3.28), (3.31) and (3.32),

$$L_t \leq (K + 1) \exp K\mu(t).$$

Hence

$$|L_t - \phi(z, t)|^2 \leq K^2(K + 1)^2 \exp 2K\mu(t),$$

so that

$$E \int_Z \int_0^1 |L_t - \phi(z, t)|^2 \tilde{P}(dz, dt) \leq E \int_0^1 K^2(K + 1)^2 \exp 2K\mu(t) \mu(dt) < \infty,$$

which implies that

$$L_t - \phi(z, t) \in L^2(\tilde{P}).$$

By § 2.10, L_t is a square-integrable martingale, and in particular $E(L_1) = E(L_0) = 1$. \square

PROPOSITION 3.4. *Suppose that for some $K < \infty$,*

$$(3.33) \quad \int_Z \int_0^1 (1 + \phi(z, t)) [\ln(1 + \phi(z, t))]^2 \tilde{P}(dz, dt) \leq K \quad \text{a.s.}$$

Then $E(L_1) = 1$.

Proof. Define the function ϕ^n so that

$$\phi^n(z, \omega, t) = \begin{cases} \phi(z, \omega, t) & \text{if } 1/n < 1 + \phi(z, \omega, t) < n, \\ 0 & \text{otherwise,} \end{cases}$$

and let L_t^n be obtained from (3.28) by replacing ϕ with ϕ^n . By Proposition 3.3, $E(L_1^n) = 1$ and it is clear that L_1^n converges to L_1 in probability. Hence by [3, § II, Thm. 21], $E(L_1) = 1$ if and only if the set of r.v.s $\{L_1^n | n = 1, 2, \dots\}$ is uniformly integrable. Define the probability measures P_n by

$$\frac{dP_n}{dP}(\omega) = L_1^n(\omega).$$

By Theorem 3.3, $(x_t, \mathcal{F}_t, P_n)$, $t \in [0, 1]$, is a fundamental process with l.d. $((1 + \phi^n)n, \Lambda)$ and so the corresponding margingales are given by

$$Q_n(B, t) = P(B, t) - \tilde{P}_n(B, t) = P(B, t) - \int_B \int_0^t (1 + \phi^n(z, s)) \tilde{P}(dz, ds).$$

Because ϕ^n is bounded and because of (3.31), $Q_n \in \mathcal{M}^2(P_n)$. For later reference define $\xi_t^n \in \mathcal{M}^2(P_n)$ by

$$\xi_t^n = (\ln(1 + \phi^n) \circ Q_n)_t,$$

and note that

$$(3.34) \quad \langle \xi^n, \xi^n \rangle_t = \int_Z \int_0^t [\ln(1 + \phi^n)]^2 (1 + \phi^n) \tilde{P}(dz, ds).$$

We are ready to show that $\{L_1^n\}$ is a uniformly integrable family. Fix $M < \infty$. First,

$$\int_{\{L_1^n > M\}} L_1^n(\omega) P(d\omega) = P_n\{L_1^n > M\}.$$

Next,

$$\begin{aligned} & \{L_1^n > M\} \\ &= \left\{ \exp \left[\int_Z \int_0^1 \ln(1 + \phi^n) P(dz, ds) \right] \cdot \exp \left[- \int_Z \int_0^1 \phi^n \tilde{P}(dz, ds) \right] > M \right\} \\ & \hspace{25em} \text{(from (3.28))} \\ &= \left\{ \exp \left[\int_Z \int_0^1 \ln(1 + \phi^n) [P(dz, ds) - (1 + \phi^n) \tilde{P}(dz, ds)] \right] \right. \\ & \quad \cdot \exp \left[\int_Z \int_0^1 [(1 + \phi^n) \cdot \ln(1 + \phi^n) - \phi^n] \tilde{P}(dz, ds) \right] > M \left. \right\} \\ &\subset \left\{ \int_Z \int_0^1 \ln(1 + \phi^n) [P(dz, ds) - (1 + \phi^n) \tilde{P}(dz, ds)] > \frac{1}{2} \ln M \right\} \\ &\cup \left\{ \int_Z \int_0^1 [(1 + \phi^n) \ln(1 + \phi^n) - \phi^n] \tilde{P}(dz, ds) > \frac{1}{2} \ln M \right\} \\ &= F_1 \cup F_2 \quad \text{say.} \end{aligned}$$

So

$$P_n\{L_1^n > M\} \leq P_n(F_1) + P_n(F_2).$$

From (3.33) it is immediate that $P_n(F_2) = 0$ for all sufficiently large M . On the other hand, $F_2 = \{\xi^n > \frac{1}{2} \ln M\}$, so by the Chebychev inequality,

$$\begin{aligned} P_n(F_2) &\leq \frac{4}{(\ln M)^2} \int_{\Omega} \langle \xi^n, \xi^n \rangle_1 dP_n \\ &\leq \frac{4}{(\ln M)^2} K \quad \text{(by (3.34), (3.33)).} \end{aligned}$$

It follows that for all n , $P_n\{L_1^n > M\} \rightarrow 0$ as $M \rightarrow \infty$, i.e., $\{L^n\}$ is uniformly integrable. \square

For the next proposition express $\tilde{P}(dz, ds) = n(dz, s) \tilde{P}(Z, ds)$ (see § 2.11).

PROPOSITION 3.5. *Suppose that there exist $\alpha > 1$ and K, K' finite such that*

$$(3.35) \quad \int_Z (1 + \phi(z, t))^\alpha n(dz, t) \leq K + K'[P(Z, t) + \tilde{P}(Z, t)] \quad \text{a.s.}$$

and suppose that for all $0 < M < \infty$,

$$(3.36) \quad E \exp [MP(Z, 1)] < \infty.$$

Then for $1 < \gamma \leq \alpha^{1/2}$,

$$(3.37) \quad \sup_{t \in [0,1]} EL_t^\gamma < \infty,$$

in particular $EL_1 = 1$.

Proof. If (3.37) is satisfied then by [8, § II, Thm. 22] the family in $\{L_t; 0 \leq t \leq 1\}$ is uniformly integrable and so by [8, § VI, Thm. 6], L_t is a uniformly integrable martingale, hence $E(L_1) = 1$.

For $\alpha^{1/2} \geq \gamma > 1$, define

$$\begin{aligned} f_t(\gamma) &= \exp \left[\gamma \int_Z \int_0^t \ln(1 + \phi) Q(dz, ds) \right. \\ &\quad \left. + \int_Z \int_0^t \left[\gamma \ln(1 + \phi) + \frac{1}{\gamma} - \frac{(1 + \phi)^{\gamma^2}}{\gamma} \right] \tilde{P}(dz, ds) \right], \\ g_t(\gamma) &= \exp \left[\int_Z \int_0^t \left[-\gamma \phi - \frac{1}{\gamma} + \frac{(1 + \phi)^{\gamma^2}}{\gamma} \right] \tilde{P}(dz, ds) \right], \end{aligned}$$

First of all,

$$\begin{aligned} f_t(\gamma)g_t(\gamma) &= \exp \left[\gamma \int_Z \int_0^t \ln(1 + \phi) Q(dz, ds) - \int_Z \int_0^t \gamma(\phi - \ln(1 + \phi)) \tilde{P}(dz, ds) \right] \\ &= L_t^\gamma \quad (\text{by (3.28)}). \end{aligned}$$

Next it can be checked by substitution in (3.28) that $[f_t(\gamma)]^\gamma$ is obtained from (3.28) by replacing $(1 + \phi)$ with $(1 + \phi)^{\gamma^2}$. Hence if $\gamma^2 < 2$ so that $(1 + \phi)^{\gamma^2} \in L_{\text{loc}}^1(\tilde{P})$, then we must have

$$E[f_t(\gamma)]^\gamma \leq 1 \quad \text{for all } t.$$

Now by Hölder's inequality,

$$EL_t^\gamma \leq (E[f_t(\gamma)]^\gamma)^{1/\gamma} (E[g_t(\gamma)]^{\gamma/(\gamma-1)})^{(\gamma-1)/\gamma},$$

so that

$$EL_t^\gamma \leq (E[g_t(\gamma)]^{\gamma/(\gamma-1)})^{(\gamma-1)/\gamma}.$$

Next,

$$\begin{aligned} [g_t(\gamma)]^{\gamma/(\gamma-1)} &\leq \exp \int_Z \int_0^t \left[\gamma - \frac{1}{\gamma} + \frac{(1 + \phi)^{\gamma^2}}{\gamma} \right] \tilde{P}(dz, ds) \\ &\quad (\text{since } 1 + \phi \geq 0 \text{ implies } -\gamma\phi \leq \gamma) \\ &\leq \exp \left[\frac{\gamma^2 - 1}{\gamma} \mu(t) + \int_0^t \{K + K'(P(Z, s) + \tilde{P}(Z, s))\} \tilde{P}(Z, ds) \right] \\ &\quad (\text{from (3.31), (3.35)}) \\ &\leq \exp \left[\left(\frac{\gamma^2 - 1}{\gamma} + K + \frac{K'}{2} \mu(t) + K'P(Z, t) \right) \mu(t) \right] \quad (\text{from (3.31)}) \\ &\leq \exp \beta \exp K' \mu(1) P(Z, 1) \quad (\text{for some constant } \beta). \end{aligned}$$

Hence,

$$EL_t' \leq (\exp \beta) E[\exp K' \mu(1) P(Z, 1)]$$

and the result follows from (3.36). \square

Remark 3.6. (i) Suppose (x_t, \mathcal{F}_t, P) is as in the fundamental example with corresponding increasing processes $\tilde{p}_1(t), \dots, \tilde{p}_n(t)$. Then Assumption 3.1 translates into the following: there exists an increasing function $\Lambda: R_+ \rightarrow R_+$ such that

$$\sum_{i=1}^n \tilde{P}_i(t) \leq \Lambda(t) \text{ a.s.}$$

Similarly (3.35), (3.36) become: there exist $\alpha > 1$ and K, K' such that

$$(3.38) \quad \sum_{i=1}^n (1 + \phi_i(t))^\alpha \leq K + K' \sum_{i=1}^n (p_i(t) + \tilde{p}_i(t)).$$

(ii) Now suppose that (x_t, \mathcal{F}_t, P) is a standard Poisson process. Then (3.38) becomes

$$(1 + \phi(t))^\alpha \leq K + K'(x(t) + t).$$

Suppose that $\phi(t) = c(x(t_-))^\alpha$ for some $\alpha < 1$. According to Feller [27, p. 452] a counting process x_t with rate $[1 + \phi(t)]$ has infinitely many jumps in a finite interval, so that it *cannot* be a fundamental process. Thus Proposition 3.5 is false if $\alpha < 1$. We have been unable to resolve the case of “linear” growth, i.e., $\alpha = 1$.

Remark 3.7. Propositions 3.3, 3.4, 3.5 are inspired by corresponding results in [6], [24], [28] respectively.

3.4. A class of Poisson-measure processes. In order to apply the transformation technique presented earlier one must begin with a fundamental process (with a known l.d.) whose existence is guaranteed. In this section we present a large class of such processes for which the increasing processes $\tilde{P}(B, t)$ are deterministic.

Let (Z, \mathcal{Z}) be any Blackwell space and let μ be any positive measure on the space $(Z \times R_+, \mathcal{Z} \otimes \mathcal{B})$, where \mathcal{B} is the Borel field on R_+ . Suppose that for all $t < \infty$, $\mu(Z \times [0, t]) < \infty$.

Let Ω' be the space of all (nonnegative) integer-valued measures N on $(Z \times R_+, \mathcal{Z} \otimes \mathcal{B})$. For each $T \in R_+$, let \mathcal{F}'_T be the family of all subsets of Ω' of the form

$$\{N \in \Omega' | N(C) \in K\},$$

where $C \in \mathcal{Z} \otimes \mathcal{B}[0, T]$ and $K \subset I_+$, the set of nonnegative integers. Evidently \mathcal{F}'_T is a σ -algebra on Ω' . Let

$$\mathcal{F}' = \bigvee_T \mathcal{F}'_T.$$

Now, for each T define the set function P'_T on $(\Omega', \mathcal{F}'_T)$ by

$$P'_T(N(C) \in K) = \sum_{k \in K} \frac{\mu(C)^k}{k!} e^{-\mu(C)}.$$

Note that $\mu(C) < \infty$ since $C \subset Z \times [0, T]$. By [31], P'_T defines a probability measure on $(\Omega', \mathcal{F}'_T)$. Furthermore if C_1, C_2 are in $Z \times [0, T]$ and $C_1 \cap C_2 = \emptyset$, then the two random variables defined by

$$N \mapsto N(C_1), \quad N \mapsto N(C_2), \quad N \in \Omega',$$

are independent. Finally the random variable $N \mapsto N(C)$ has a Poisson distribution. For $A \in \mathcal{Z}$, consider the counting process $P'(A, t)$, $t \in R_+$, defined on the family $(\Omega', \mathcal{F}'_{t \wedge T}, P'_T)$, by

$$P'(A, t)(N) = N(A \times [0, t \wedge T]).$$

Evidently $E(P'(A, t)) = \mu(A \times [0, t \wedge T])$, and if $A_1 \cap A_2 = \emptyset$, then $P'(A_1, t)$ and $P'(A_2, t)$ are independent processes.

Next by Moyal [32], there exists a jump process x_t , $t \in R_+$, with values in (Z, \mathcal{Z}) , defined on a family $(\Omega, \mathcal{F}^x_t, P_T)$ such that (i) $(\Omega, \mathcal{F}^x_t, P_T)$ is isomorphic to $(\Omega', \mathcal{F}'_{t \wedge T}, P'_T)$ and (ii) the counting processes $P^x(A, t)$ corresponding to x_t are "isomorphic" to the processes $P'(A, t)$ constructed above. Furthermore,

$$\tilde{P}^x(A, t) = \mu(A \times [0, t \wedge T]).$$

To finish the construction we merely note that if $S < T$, then the probability measure P_S on $(\Omega, \mathcal{F}^x_S)$ coincides with the restriction of P_T (defined on \mathcal{F}^x_T) to \mathcal{F}^x_S . By the Kolmogorov consistency theorem, there therefore exists a probability measure P on $(\Omega, \mathcal{F}^x_\infty)$ such that

$$(3.39) \quad \tilde{P}^x(A, t) = \mu(A \times [0, t]), \quad A \in \mathcal{Z}, \quad t \in R_+.$$

However the process x_t may not be a fundamental process. To guarantee this we must be sure that the jump times are totally inaccessible. As mentioned in § 2.9, this is equivalent to the requirement that $\tilde{P}^x(A, t)$ have continuous sample paths, and hence, from (3.39), to the requirement that $\mu(A \times [0, t])$ be continuous in t for each fixed A . We summarize the main conclusions as follows.

THEOREM 3.5. *Let (Z, \mathcal{Z}) be a Blackwell space and let μ be any nonnegative measure on $(Z \times R_+, \mathcal{Z} \otimes \mathcal{B})$ such that*

- (i) $\mu(Z \times [0, t]) < \infty$ for all $t \in R_+$,
- (ii) $\mu(A \times [0, t])$ is continuous in t for all $A \in \mathcal{Z}$.

Then there exists a fundamental process x_t on a family $(\Omega, \mathcal{F}^x_t, P)$ with values in (Z, \mathcal{Z}) such that

$$\tilde{P}^x(A, t) = \mu(A \times [0, t]), \quad A \in \mathcal{Z}, \quad t \in R_+.$$

Remark 3.8. (i) The x_t process has *independent increments* in the sense that the $P(A, t)$ have independent increments. If x_t were vector-valued this would indeed imply that x_t has independent increments in the usual sense.

(ii) The most useful version of this result would be when μ is a product measure $\mu(dz, ds) = n(dz)\Lambda(ds)$, where n is a finite measure on (Z, \mathcal{Z}) and $\Lambda(t)$ is a continuous increasing function on R_+ , in which case (n, Λ) would be a Lévy system.

4. Detection. The prototypical detection problem in communication theory is the following. We observe a sample $x_t(\omega)$, $0 \leq t < \infty$, of a stochastic process.

The process is known to be governed by one of two laws, P or P_1 . Based upon the observed sample one has to decide which of the two hypotheses, P or P_1 , is true. The term “detection” arises from a particular instance of this hypothesis testing model, namely, when the process x has the representation

$$(4.1) \quad \begin{aligned} dx_t &= \text{white noise}, & \text{under } P, \\ dx_t &= \text{white noise} + s_t, & \text{under } P_1, \end{aligned}$$

where s_t is called the “signal”. Thus deciding which hypothesis is true is, for the example, equivalent to “detecting” whether the signal is present (hypothesis P_1) or absent (hypothesis P).

Very recently this problem has been considered for the case where x_t is a counting process under P_1 and a Poisson process under P [6], [7], [15], [16], [17]. The case where x_t is a Markov chain under P has also been discussed [6]. We generalize these results by considering problems where x_t is a fundamental process.

A well-established procedure for judging which hypothesis is true consists in first calculating the “likelihood” ratio $(dP_1/dP)(x(\omega))$ and then in accepting P_1 if $dP_1/dP > \alpha$ and rejecting P_1 otherwise. The selection of the “threshold” α is discussed in [18]. The procedure is often called the “threshold detector”.

Evidently for this procedure to be meaningful one must assume $P_1 \ll P$. Also to obtain results of practical value one must specify precisely how the “signal” affects the observation, as for instance in (4.1), where it is assumed to be additive. We proceed to the mathematical model.

Let (Ω, \mathcal{F}_t) , $t \in R_+$, be a family of spaces and P, P_1 two probabilities on (Ω, \mathcal{F}) . The observed process is a family of measurable functions $x_t : (\Omega, \mathcal{F}_t) \rightarrow (Z, \mathcal{Z})$ such that (x_t, \mathcal{F}_t, P) and $(x_t, \mathcal{F}_t, P_1)$ are both fundamental processes. The processes P, \tilde{P}, Q and \tilde{P}^x, Q^x are the extrinsic and intrinsic (i.e., relative to \mathcal{F}_t^x) processes corresponding to (x_t, P) . Similarly $P_1, \tilde{P}_1, \tilde{P}_1^x$ etc. correspond to (x_t, P_1) . The extrinsic and intrinsic l.d.’s are $(n, \Lambda), (n^x, \Lambda^x)$ for (x_t, P) and $(n_1, \Lambda_1), (n_1^x, \Lambda_1^x)$ for (x_t, P_1) .

We now give the model corresponding to the “signal plus noise” model of (4.1).

Assumption 4.1. There exist \mathcal{F}_t^x -predictable processes $\mu(B, \omega, t)$, $B \in \mathcal{Z}$, and \mathcal{F}_t -predictable processes $g(z, \omega, s)$ and $g_1(z, \omega, s)$ such that $E|g(z, s)| < \infty$ and $E_1|g_1(z, s)| < \infty$ for all z, s , and

$$\tilde{P}(B, t) = n(B, t)\Lambda(t) = \int_B \int_0^t g(z, \omega, s)\mu(dz, \omega, ds),$$

$$\tilde{P}_1(B, t) = n_1(B, t)\Lambda_1(t) = \int_B \int_0^t g_1(z, \omega, s)\mu(dz, \omega, ds),$$

where the integrals are Stieltjes integrals.

Interpretation. In communication theory terms we can say that the “jump rates” $P(B, t)$ are “modulated” by the signal through the functions g, g_1 .

DEFINITION 4.1. Let $E(g(z, t)|\mathcal{F}_t^x) = \hat{g}(z, t)$ and $E_1(g_1(z, t)|\mathcal{F}_t^x) = \hat{g}_1(z, t)$.

PROPOSITION 4.1.

$$\tilde{P}^x(B, t) = \int_B \int_0^t \hat{g}(z, s) \mu(dz, ds) \quad a.s.$$

$$\tilde{P}_1^x(B, t) = \int_B \int_0^t \hat{g}_1(z, s) \mu(dz, ds) \quad a.s.$$

Proof. It is enough to prove the first assertion since the proof for the second is identical. Fix $B \in \mathcal{Z}$. We know that

$$(4.2) \quad Q(B, t) = P(B, t) - \int_B \int_0^t g(z, s) \mu(dz, ds) \in \mathcal{M}_{loc}^2(\mathcal{F}_t, P)$$

$$(4.3) \quad Q^x(B, t) = P(B, t) - \tilde{P}^x(B, t) \in \mathcal{M}_{loc}^2(\mathcal{F}_t^x, P).$$

Let T_n , $n = 0, 1, \dots$, be the jump times of x_t . The T_n are stopping times for the family (\mathcal{F}_t) as well as for (\mathcal{F}_t^x) . Furthermore $E|P(B, t \wedge T_n)| \leq n$. Hence

$$E|Q(B, t \wedge T_n)| < \infty,$$

and we can define a process $(\hat{Q}(B, t), \mathcal{F}_t^x, P)$ such that

$$\hat{Q}(B, t \wedge T_n) = E(Q(B, t \wedge T_n) | \mathcal{F}_t^x)$$

and it is trivial that $\hat{Q}(B, t \wedge T_n) \in \mathcal{M}^1(\mathcal{F}_t^x, P)$. Now $P(B, t)$ and $\mu(z, t)$ are \mathcal{F}_t^x -measurable, hence

$$\hat{Q}(B, t \wedge T_n) = P(B, t \wedge T_n) - \int_B \int_0^{t \wedge T_n} E(g(z, s) | \mathcal{F}_t^x) \mu(dz, ds).$$

Subtracting this from (4.3) implies that

$$\tilde{P}^x(B, t \wedge T_n) - \int_B \int_0^{t \wedge T_n} E(g(z, s) | \mathcal{F}_t^x) \mu(dz, ds) \in \mathcal{M}^1(\mathcal{F}_t^x, P).$$

On the other hand it can be directly verified that

$$\int_B \int_0^{t \wedge T_n} [E(g(z, s) | \mathcal{F}_t^x) - \hat{g}(z, s)] \mu(dz, ds) \in \mathcal{M}^1(\mathcal{F}_t^x, P).$$

Therefore

$$\tilde{P}^x(B, t \wedge T_n) - \int_B \int_0^{t \wedge T_n} \hat{g}(z, s) \mu(dz, ds) \in \mathcal{M}^1(\mathcal{F}_t^x, P).$$

But this is a continuous process. Hence it must vanish, i.e.,

$$\tilde{P}^x(B,) = \int_B \int_0^t \hat{g}(z, s) \mu(dz, ds). \quad \square$$

Remark 4.1. The processes $(\hat{Q}(B, t), \mathcal{F}_t^x, P)$ are called the *innovations* processes of the process (x_t, \mathcal{F}_t, P) , in analogy with the Brownian motion case [21]. These processes will be used in the next section.

THEOREM 4.1. Suppose that $P_1 \ll P$. Let $L_t = E(dP_1/dP|\mathcal{F}_t^x)$ be the likelihood ratio and let

$$T = \inf \{t | L_t = 0 \text{ or } L_{t-} = 0\}.$$

Then there is a sequence of \mathcal{F}_t^x s.t.'s $S_k \uparrow T$ a.s. P such that

$$\frac{\hat{g}_1(z, s)}{\hat{g}(z, s)} I_{\{s \leq S_k\}} \in L_{\text{loc}}^1(\tilde{P}^x)$$

and

$$(4.4) \quad L_{t \wedge S_k} = \prod_{\substack{s \leq t \wedge S_k \\ x_{s-} \neq x_s}} \left[\frac{\hat{g}_1(x_s, s)}{\hat{g}(x_s, s)} \right] \exp \left[- \int_Z \int_0^{t \wedge S_k} \left(\frac{\hat{g}_1(z, s)}{\hat{g}(z, s)} - 1 \right) \hat{g}(z, s) \mu(dz, ds) \right].$$

Proof. By Theorem 3.1 there exists s.t.'s $S_k \uparrow T$ and an \mathcal{F}_t^x -predictable function ϕ such that $L_{t \wedge S_k}$ is given by (3.11), and by Theorem 3.2 the intrinsic l.d. of (x_t, P_1) is $((1 + \phi)n^x, \Lambda^x)$, where (n^x, Λ^x) is the intrinsic l.d. of (x_t, P) ; so from Proposition 4.1 we can conclude that

$$\begin{aligned} (1 + \phi(z, s))n^x(dz, s)\Lambda^x(ds) &= (1 + \phi(z, s))\hat{g}(z, s)\mu(dz, ds) \\ &= \hat{g}_1(z, s)\mu(dz, ds) = n_1^x(dz, s)\Lambda_1^x(ds). \end{aligned}$$

Therefore

$$1 + \phi(z, s) = \frac{\hat{g}_1(z, s)}{\hat{g}(z, s)}$$

which upon substitution into (3.11) yields (4.4). \square

COROLLARY 4.1. Suppose in the above that x_t is as in the fundamental example of § 2.12. Suppose there exists a \mathcal{F}_t^x -predictable process $\mu(t)$, and \mathcal{F}_t -predictable processes $\hat{\lambda}^i(t)$, $\lambda_1^i(t)$, $1 \leq i \leq n$, such that

$$\tilde{p}_i(t) = \int_0^t \lambda^i(s) \mu(ds), \quad \tilde{p}_{i,1}(t) = \int_0^t \lambda_1^i(s) \mu(ds), \quad 1 \leq i \leq n.$$

Then the formula (4.4) changes to

$$(4.5) \quad L_{t \wedge S_k} = \prod_{i=1}^n \left\{ \prod_{\substack{s \leq t \wedge S_k \\ (x_{s-}, x_s) \in \sigma_i}} \left[\frac{\hat{\lambda}_1^i(s)}{\hat{\lambda}^i(s)} \right] \cdot \exp \left[- \int_0^{t \wedge S_k} \left(\frac{\hat{\lambda}_1^i(s)}{\hat{\lambda}^i(s)} - 1 \right) \hat{\lambda}^i(s) \mu(ds) \right] \right\}.$$

Proof. The proof follows from Theorem 4.1 and Remark 3.2 (iii). \square

Remark 4.2. (i) Very special cases of (4.5) have appeared in the recent literature. Suppose in Corollary 4.1 that $(x_t, \mathcal{F}_t^x, P)$ is a Poisson process with rate λ_0 . Then in (4.5), $n = 1$, $\hat{\lambda}(s) \equiv \lambda_0$, $\mu(ds) \equiv ds$ and (4.5) becomes

$$(4.6) \quad L_{t \wedge S_k} = \left[\prod_{\substack{s \leq t \wedge S_k \\ x_{s-} \neq x_s}} \frac{\hat{\lambda}_0(s)}{\lambda_0} \right] \exp \left[- \int_0^{t \wedge S_k} (\hat{\lambda}_1(s) - \lambda_0) ds \right].$$

This version together with the comment in footnote 5 yields the result in [16, p. 95]. Actually in [16] some strong unnecessary assumptions are also imposed.

Formula (4.6) has also been derived in [6] and [7]. Formula (4.5) for the case $n = 1$ and $\mu(ds) \equiv ds$ appears in [15], although the derivation is not satisfactory, and various additional assumptions, some of which are not easily unverifiable, were made there.

(ii) In [6] we can also find (4.5) for the special case where $(x_t, \mathcal{F}_t^x, P)$ is a Markov chain, in which case the $\hat{\lambda}^i$ can be interpreted in terms of various transition probabilities as suggested in § 2.11, § 2.12.

We apply formulas (4.5) and (4.6) to calculate the mutual information between two fundamental processes. Let x_t and x'_t be two such processes on $(\Omega, \mathcal{F}_t, P)$ with values in (Z, \mathcal{Z}) and (Z', \mathcal{Z}') respectively. Let $\mu(dz, ds)$ and $\mu'(dz', ds)$ be \mathcal{F}_t^x - and $\mathcal{F}_t^{x'}$ -predictable processes and $g(z, s)$, $g'(z', s)$ be two \mathcal{F}_t -predictable processes with finite expectation such that

$$\begin{aligned} n(dz, s)\Lambda(ds) &= g(z, s)\mu(dz, ds), \\ n'(dz', s)\Lambda'(ds) &= g'(z', s)\mu'(dz', ds). \end{aligned}$$

Let $P_x, P_{x'}$ denote the restrictions of P to \mathcal{F}^x and $\mathcal{F}^{x'}$ respectively. Assume that $\mathcal{F}_t = \mathcal{F}_t^x \otimes \mathcal{F}_t^{x'}$, the product σ -algebra and let $P_{xx'} = P_x \otimes P_{x'}$ denote the product measure on $\mathcal{F} = \mathcal{F}^x \otimes \mathcal{F}^{x'}$. It is trivial that $P \ll P_{xx'}$. Assume further that $P_{xx'} \ll P$. The mutual information between x, x' is the quantity

$$I(x, x') = E \left(\ln \frac{dP}{dP_{xx'}} \right).$$

Let

$$\begin{aligned} \hat{g}(z, t) &= E(g(z, t) | \mathcal{F}_t^x), \\ \hat{g}'(z', t) &= E(g'(z', t) | \mathcal{F}_t^{x'}). \end{aligned}$$

By Remark 3.2 (i),

$$g/\hat{g} \in L^1_{\text{loc}}(\tilde{P}), \quad g'/\hat{g}' \in L^1_{\text{loc}}(\tilde{P}').$$

Assume further that

$$\ln(g/\hat{g}) \in L^1(\tilde{P}), \quad \ln(g'/\hat{g}') \in L^1(\tilde{P}').$$

Then by Theorem 4.1,

$$\begin{aligned} \frac{dP}{dP_{xx'}} &= \left\{ \prod_{x_s \neq x'_s} \left[\frac{g(x_s, s)}{\hat{g}(x_s, s)} \right] \exp \left[- \int_Z \int_0^\infty \left(\frac{g(z, s)}{\hat{g}(z, s)} - 1 \right) \hat{g}(z, s) \mu(dz, ds) \right] \right\} \\ &\quad \cdot \left\{ \prod_{x'_s \neq x_s} \left[\frac{g'(x'_s, s)}{\hat{g}'(x'_s, s)} \right] \exp \left[- \int_{Z'} \int_0^\infty \left(\frac{g'(z', s)}{\hat{g}'(z', s)} - 1 \right) \hat{g}'(z', s) \mu'(dz', ds) \right] \right\} \end{aligned}$$

so that

$$\begin{aligned} \ln \frac{dP}{dP_{xx'}} &= \sum_{x_s \neq x'_s} \ln \left(\frac{g(x_s, s)}{\hat{g}(x_s, s)} \right) - \int_Z \int_0^\infty \left(\frac{g(z, s)}{\hat{g}(z, s)} - 1 \right) \hat{g}(z, s) \mu(dz, ds) \\ (4.7) \quad &+ \sum_{x'_s \neq x_s} \ln \left(\frac{g'(x'_s, s)}{\hat{g}'(x'_s, s)} \right) - \int_{Z'} \int_0^\infty \left(\frac{g'(z', s)}{\hat{g}'(z', s)} - 1 \right) \hat{g}'(z', s) \mu'(dz', ds). \end{aligned}$$

Since $\ln(g/\hat{g}) \in L^1(\tilde{P})$, therefore

$$\begin{aligned} \sum_{x_s \neq x_s} \ln \left(\frac{g(x_s, s)}{\hat{g}(x_s, s)} \right) - \int_Z \int_0^\infty \ln \left(\frac{g(z, s)}{\hat{g}(z, s)} \right) g(z, s) \mu(dz, ds) \\ = \int_Z \int_0^\infty \ln \left(\frac{g(z, s)}{\hat{g}(z, s)} \right) [P(dz, ds) - \tilde{P}(dz, ds)] \in \mathcal{M}^1(\mathcal{F}_t, P) \end{aligned}$$

so that

$$E \left[\sum_{x_s \neq x_s} \ln \left(\frac{g(x_s, s)}{\hat{g}(x_s, s)} \right) \right] = E \int_Z \int_0^\infty \ln \left(\frac{g(z, s)}{\hat{g}(z, s)} \right) g(z, s) \mu(dz, ds).$$

Similarly,

$$E \left[\sum_{x'_s \neq x'_s} \ln \left(\frac{g'(x'_s, s)}{\hat{g}'(x'_s, s)} \right) \right] = E \int_{Z'} \int_0^\infty \ln \left(\frac{g'(z', s)}{\hat{g}'(z', s)} \right) g'(z', s) \mu'(dz', ds).$$

Taking expectations in (4.7) and substituting these relations gives the following result.

THEOREM 4.2. Suppose $P_{xx'} \ll P$ and $\ln(g/\hat{g}) \in L^1(\tilde{P})$, $\ln(g'/\hat{g}') \in L^1(\tilde{P})$. Then

$$\begin{aligned} I(x, x') = E \left[\int_Z \int_0^\infty \left(\ln \frac{g(z, s)}{\hat{g}(z, s)} + \frac{\hat{g}(z, s)}{g(z, s)} - 1 \right) g(z, s) \mu(dz, ds) \right. \\ \left. + \int_{Z'} \int_0^\infty \left(\ln \frac{g'(z', s)}{\hat{g}'(z', s)} + \frac{\hat{g}'(z', s)}{g'(z', s)} - 1 \right) g'(z', s) \mu'(dz', ds) \right]. \end{aligned} \quad (4.8)$$

Remark 4.3. This result for the case where x, x' are both counting processes has appeared in [6], and our proof is adapted from the one given there.

5. Filtering. A popular model for estimation and filtering problems in communication and control is where the observed process, x_t , depends upon the “signal” or “state” process, y_t , according to

$$\begin{aligned} dy_t &= g(y_t) dt + dB_1(t), \\ dx_t &= f(x_t, y_t) dt + dB_2(t), \end{aligned}$$

where B_1, B_2 are Brownian motions. The problem is to determine $E(y_t | \mathcal{F}_t^x)$. Note that in the above y_t is a semi-martingale.

We begin this section by examining this situation when (x_t, \mathcal{F}_t, P) is a fundamental process with values in (Z, \mathcal{Z}) . We need a preliminary fact.

LEMMA 5.1. Let $(m_t, \mathcal{F}_t, P) \in \mathcal{M}^2(\mathcal{F}_t, P)$. Then there exists an \mathcal{F}_t -predictable process $h(z, t)$ such that

$$(5.1) \quad E \int_Z \int_0^\infty |h(z, t)|^2 \tilde{P}(dz, dt) < \infty$$

and

$$(5.2) \quad \langle m_t, Q(B, t) \rangle = \int_B \int_0^t h(z, s) \tilde{P}(dz, ds) \quad \text{for all } B \in \mathcal{Z}.$$

Proof. The set, say \mathcal{L} , of all processes $(h \circ Q)_t$, where h is any predictable process satisfying (5.1), is easily shown to be a stable subspace of $\mathcal{M}^2(\mathcal{F}_t, P)$ (see [19] for a definition of a *stable* subspace). Therefore by [19], there exists a unique decomposition of m_t , $m_t = n_t + l_t$, with $l_t \in \mathcal{L}$ and $\langle n_t, l'_t \rangle \equiv 0$ for all $l'_t \in \mathcal{L}$. Let $l_t = (h \circ Q)_t$ and the assertion follows. \square

Assumption 5.1. There exist \mathcal{F}_t^x -predictable processes $\mu(B, t)$, $B \in \mathcal{Z}$, and an \mathcal{F}_t -predictable process $g(z, t)$ such that

$$(5.3) \quad \tilde{P}(B, t) = \int_B \int_0^t g(z, s) \mu(dz, ds).$$

Notation. In the following for any process (f_t, \mathcal{F}_t, P) , $\hat{f}_t = E(f_t | \mathcal{F}_t^x)$.

THEOREM 5.1. *Let (x_t, \mathcal{F}_t, P) be a fundamental process satisfying Assumption 5.1. Let $(y_t, \mathcal{F}_t, P) \in \mathcal{S}(\mathcal{F}_t)$ have the representation*

$$(5.4) \quad y_t = y_0 + a_t + m_t$$

with $a_t \in \mathcal{A}(\mathcal{F}_t)$, $m_t \in \mathcal{M}^2(\mathcal{F}_t)$. Then \hat{y}_t satisfies the filtering equation

$$\hat{y}_t = \hat{y}_0 + \eta_t + \int_Z \int_0^t k(z, s) Q^x(dz, ds),$$

where $\eta_t \in \mathcal{A}(\mathcal{F}_t^x)$, $Q^x(B, t) = P(B, t) - \int_B \int_0^t \hat{g}(z, s) \mu(dz, ds)$, and where the \mathcal{F}_t^x -predictable process k satisfies

$$k(z, s) = \frac{\overline{[(y_{s-} - \hat{y}_{s-} + h(z, s))g(z, s)]}}{\hat{g}(z, s)}$$

and h, g are as in (5.2), (5.3) respectively.

Proof. Let $\mu_t = E(m_t | \mathcal{F}_t^x)$. Clearly $\mu_t \in \mathcal{M}^2(\mathcal{F}_t^x)$. Now write $a_t = a_t^+ - a_t^-$ where $a_t^+, a_t^- \in \mathcal{A}_+(\mathcal{F}_t, P)$. It is easy to verify that the \mathcal{F}_t^x -measurable processes $\alpha_t^+ = E(a_t^+ | \mathcal{F}_t^x)$, $\alpha_t^- = E(a_t^- | \mathcal{F}_t^x)$ are submartingales. By the Doob–Meyer decomposition theorem [3], there exist martingales ξ_t^+, ξ_t^- in $\mathcal{M}^1(\mathcal{F}_t^x)$ and \mathcal{F}_t^x -predictable increasing processes η_t^+, η_t^- in $\mathcal{A}_+(\mathcal{F}_t^x)$ such that

$$\alpha_t^+ = \xi_t^+ + \eta_t^+, \quad \alpha_t^- = \xi_t^- + \eta_t^-.$$

Hence

$$(5.5) \quad \begin{aligned} \hat{y}_t &= \hat{y}_0 + \alpha_t^+ - \alpha_t^- + \hat{m}_t \\ &= \hat{y}_0 + (\eta_t^+ - \eta_t^-) + (\xi_t^+ - \xi_t^- + \mu_t) \\ &= \hat{y}_0 + \eta_t + \xi_t, \quad \text{say,} \end{aligned}$$

where $\eta_t \in \mathcal{A}(\mathcal{F}_t^x)$, $\xi_t \in \mathcal{M}^1(\mathcal{F}_t^x)$. By § 2.11 there exists a \mathcal{F}_t^x -predictable process $k(z, s) \in L_{\text{loc}}^1(\tilde{P}^x)$ such that

$$(5.6) \quad \xi_t = \int_Z \int_0^t k(z, s) Q^x(dz, ds).$$

It remains to evaluate k . By the differentiation formula of § 2.7,

$$y_t P(B, t) = \int_0^t y_{s-} P(B, ds) + \int_0^t P(B, s-) dy_s + [m_t, Q(B, t)].$$

Since $P(B, t) - \tilde{P}(B, t)$ and $[m_t, Q(B, t)] - \langle m_t, Q(B, t) \rangle$ are in $\mathcal{M}_{\text{loc}}^1(\mathcal{F}_t)$, therefore, from the above, for some $\gamma_t, \gamma'_t \in \mathcal{M}_{\text{loc}}^1(\mathcal{F}_t)$,

$$\begin{aligned} y_t P(B, t) &= \int_0^t y_{s-} \tilde{P}(B, ds) + \int_0^t P(B, s^-) dy_s + \langle m_t, Q(B, t) \rangle + \gamma_t \\ (5.7) \quad &= \int_B \int_0^t (y_{s-} + h(z, s)) g(z, s) \mu(dz, ds) + \int_0^t P(B, s^-) da_s + \gamma'_t, \end{aligned}$$

using (5.2), (5.3) and (5.4).

Now apply the differential rule to $\hat{y}_t P(B, t)$ to obtain

$$\hat{y}_t P(B, t) = \int_0^t \hat{y}_{s-} P(B, ds) + \int_0^t P(B, s^-) d\hat{y}_s + [\xi_t, Q^x(B, t)].$$

Recalling that $P(B, t) - \tilde{P}^x(B, t)$ and $[\xi_t, Q^x(B, t)] - \langle \xi_t, Q^x(B, t) \rangle$ are in $\mathcal{M}_{\text{loc}}^1(\mathcal{F}_t^x)$, the relation above implies that for some $\delta_t, \delta'_t \in \mathcal{M}_{\text{loc}}^1(\mathcal{F}_t^x)$,

$$\begin{aligned} \hat{y}_t P(B, t) &= \int_0^t \hat{y}_{s-} \tilde{P}^x(B, ds) + \int_0^t P(B, s^-) d\hat{y}_s + \langle \xi_t, Q^x(B, t) \rangle + \delta_t \\ (5.8) \quad &= \int_B \int_0^t (\hat{y}_{s-} + k(z, s)) \hat{g}(z, s) \mu(dz, ds) + \int_0^t P(B, s^-) d\eta_s + \delta'_t, \end{aligned}$$

using Proposition 4.1, (5.5), (5.6).

Next we make the following observations, which can be verified directly from the martingale definition:

$$\begin{aligned} &\left(\int_B \int_0^t (y_{s-} + h(z, s)) g(z, s) \mu(dz, ds) \right) \\ &\quad - \int_B \int_0^t [(\widehat{y_{s-} + h(z, s)) g(z, s)}] \mu(dz, ds) \in \mathcal{M}_{\text{loc}}^1(\mathcal{F}_t^x), \\ &\left(\int_0^t P(B, s^-) da_s \right) - \int_0^t P(B, s^-) d\eta_s \in \mathcal{M}_{\text{loc}}^1(\mathcal{F}_t^x). \end{aligned}$$

Using these facts and the fact that $(\widehat{y_t P(B, t)}) = \hat{y}_t P(B, t)$, we conclude from (5.7), (5.8) that

$$\int_B \int_0^t \{(\hat{y}_{s-} + k(z, s)) \hat{g}(z, s) - [\widehat{(y_{s-} + h(z, s)) g(z, s)}]\} \mu(dz, ds) \in \mathcal{M}_{\text{loc}}^1(\mathcal{F}_t^x),$$

and since this process is continuous, it must vanish identically, so that we may assume

$$\begin{aligned} k(z, s) &= \frac{[(\widehat{y_{s-} + h(z, s)) g(z, s)}]}{\hat{g}(z, s)} - \hat{y}_{s-} \\ &= \frac{[(\widehat{y_{s-} - \hat{y}_{s-} + h(z, s)) g(z, s)}]}{\hat{g}(z, s)}. \end{aligned}$$

□

COROLLARY 5.1. *Suppose in the above that x_t is as in the fundamental example of § 2.12 and that there exists an \mathcal{F}_t^x -predictable process $\mu(t)$ and \mathcal{F}_t -predictable processes λ_t^i such that*

$$\tilde{p}_i(t) = \int_0^t \lambda^i(s) \mu(ds)$$

and let $\langle m_t, q_i(t) \rangle = \int_0^t h_i(s) \tilde{p}_i(ds)$ for some \mathcal{F}_t -predictable processes h_i . Then

$$\hat{y}_t = \hat{y}_0 + \eta_t + \sum_{i=1}^n \int_0^t k_i(s) q_i^x(ds)$$

with

$$k_i(t) = \frac{[(\hat{y}_{t-} - \hat{y}_t + h_i(t))\lambda^i(t)]}{\hat{\lambda}^i(t)}, \quad 1 \leq i \leq n,$$

and

$$q_i^x(t) = p_i(t) - \int_0^t \hat{\lambda}^i(s) ds.$$

Remark 5.1. (i) Suppose in (5.4) that a_t is given as

$$a_t = \int_0^t \beta_s ds$$

for some predictable process β_t in $L^1(\mathcal{F}_t)$. Since $\hat{a}_t - \int_0^t \hat{\beta}_s ds$ is in $\mathcal{M}^1(\mathcal{F}_t^x)$ it follows that in the representation for \hat{y}_t we have the further specification

$$\eta_t = \int_0^t \hat{\beta}_s ds.$$

(ii) Corollary 5.1 has appeared in the literature for the case where (x_t, \mathcal{F}_t, P) is a counting process, i.e., $n = 1$. Even here some additional conditions have been imposed on the y_t process (such as, e.g., y_t is Markov [6], [16]) or on the x_t process (such as, e.g., $(x_t, \mathcal{F}_t^x, P)$ is obtained from a Poisson process by an absolutely continuous change of measure [6], [20]).

(iii) Theorem 5.1 has been inspired largely by the procedures of [21], where the underlying process is Brownian motion. See also [24] for the Brownian motion case.

(iv) While Theorem 5.1 has some value in terms of clarifying the issues involved in obtaining the filtering equations it is of little practical importance since these equations do not lead to a realization by a dynamical system. This is so because the filtering equations contain the terms η_t , k_t and \hat{g}_t which are *not* computable in terms of \hat{y}_t and x_t . In other words, the filtering equation is not recursive. This difficulty persists even when one imposes additional conditions such as y_t is Markov. In the remainder of this section we seek to determine conditions under which the filter is recursive.

We impose conditions on the dependence between the “signal” or “state” process y_t and the “observation” process x_t which are considerably stronger than

those of Assumption 5.1. For the remainder of this section the following assumption holds.

Assumption 5.2. (Ω, \mathcal{F}_t) , $t \in R_+$, is a family of spaces and P, P_1 are two probability measures on (Ω, \mathcal{F}) . x_t and y_t are measurable functions on (Ω, \mathcal{F}_t) with values in (Z, \mathcal{Z}) and (Y, \mathcal{Y}) respectively. The following properties are satisfied.

(i) Z is a Borel subset of R^n , \mathcal{Z} is the Borel field. (The most important practical cases are $Z = R^n$ or Z is the space of all $z \in R^n$ with integer components.) Y is a locally compact Hausdorff space, \mathcal{Y} is the Borel field. $\mathcal{F}_t = \mathcal{F}_t^x \vee \mathcal{F}_t^y$.

(ii) Under the measure P ,

(a) (x_t, \mathcal{F}_t, P) is a fundamental process with *independent increments*, i.e., $x_t - x_s$ is independent of \mathcal{F}_s (under P), for $s \leq t$,

(b) (y_t, \mathcal{F}_t, P) is a *Markov* process whose sample paths are right-continuous and have left-limits, and the jump times of y are totally inaccessible,

(c) the processes x_t and y_t are *independent*, i.e., \mathcal{F}^x and \mathcal{F}^y are independent.

(iii) $P_1 \ll P$, there exists an \mathcal{F}_t -predictable process $f \in L_{\text{loc}}^1(\tilde{P})$ with a representation

$$f(z, \omega, t) = \phi(z, y_{t-}(\omega), \omega, t),$$

where $\phi(\cdot, y, \cdot, \cdot)$ is \mathcal{F}_t^x -predictable for fixed $y \in Y$, and there also exist \mathcal{F}_t^x -predictable processes $\mu(B, t)$ for $B \in \mathcal{Z}$ such that $E(|f(z, t)|) + E_1(|f(z, t)|) < \infty$ for all z, t and

$$L_t = E\left(\frac{dP_1}{dP}\bigg|\mathcal{F}_t\right) = \prod_{\substack{s \leq t \\ x_{s-} \neq x_s}} [1 + \phi(x_s, y_{s-}, s)] \exp\left[-\int_Z \int_0^t \phi(z, y_{s-}, s) \mu(dz, ds)\right].$$

Note that we must have $1 + \phi \geq 0$.

Let Q, \tilde{P} and Q^x, \tilde{P}^x be the processes associated with (x_t, \mathcal{F}_t, P) and $(x_t, \mathcal{F}_t^x, P)$. Similarly let $Q_1, \tilde{P}_1, Q_1^x, \tilde{P}_1^x$ be the processes corresponding with $(x_t, \mathcal{F}_t, P_1)$ and $(x_t, \mathcal{F}_t^x, P_1)$ respectively. From Assumption 5.2 and Proposition 4.1 it is immediate that

$$\tilde{P}(B, t) = \tilde{P}^x(B, t) = \mu(B, t),$$

$$\tilde{P}_1(B, t) = \int_B \int_0^t (1 + f(z, s)) \mu(dz, ds),$$

$$\tilde{P}_1^x(B, t) = \int_B \int_0^t (1 + \hat{f}(z, s)) \mu(dz, ds),$$

where $\hat{f}(z, t) = E_1(f(z, t) | \mathcal{F}_t^x)$.

For any t let $\mathcal{F}_{t-} = \bigvee_{s < t} \mathcal{F}_s$, $\mathcal{F}_{t-}^x = \bigvee_{s < t} \mathcal{F}_s^x$, $\mathcal{F}_{t-}^y = \bigvee_{s < t} \mathcal{F}_s^y$.

PROPOSITION 5.1. For $t \in R_+$, $\mathcal{F}_{t-}^x = \mathcal{F}_t^x$, $\mathcal{F}_{t-}^y = \mathcal{F}_t^y$, $\mathcal{F}_{t-} = \mathcal{F}_t$.

Proof. The jump times of x and y are totally inaccessible, hence by [4, Prop. 3.1] and [22, § III, D38], $\mathcal{F}_{t-}^x = \mathcal{F}_t^x$ and $\mathcal{F}_{t-}^y = \mathcal{F}_t^y$. The last assertion follows because $\mathcal{F}_{t-} = \mathcal{F}_{t-}^x \vee \mathcal{F}_{t-}^y$ and $\mathcal{F}_t = \mathcal{F}_t^x \vee \mathcal{F}_t^y$. \square

PROPOSITION 5.2. $L_{t-} = L_t$ a.s. P .

Proof. The proof follows from [22, § V, Thm. 10] using a stopping time argument. \square

PROPOSITION 5.3. $y_t = y_{t-}$ a.s. P .

Proof. $\text{Prob}\{y_t \neq y_{t-}\} = \text{Prob}\{t \text{ is a jump time}\}$. However, since the jump times are totally inaccessible, this probability must be zero. \square

For a real-valued function g on Y we are interested in determining a (recursive) expression for the process $E_1(g(y_t)|\mathcal{F}_t^x)$. Now

$$(5.9) \quad E_1(g(y_t)|\mathcal{F}_t^x) = \frac{E(g(y_t)L_t|\mathcal{F}_t^x)}{E(L_t|\mathcal{F}_t^x)}.$$

It turns out that the numerator of the expression in the right is much better behaved than the ratio, and, furthermore, the denominator does not depend on g . Hence we will seek to determine instead an expression for $E(g(y_t)L_t|\mathcal{F}_t^x)$.

DEFINITION 5.1. Let \mathcal{G} be the family of all bounded, measurable, real-valued functions g on Y . For $g \in \mathcal{G}$ and $t \in R_+$, let

$$(5.10) \quad \pi_t(g) = E(g(y_t)L_t|\mathcal{F}_t^x).$$

PROPOSITION 5.4. $E(L_t|\mathcal{F}_t^y) = 1$ a.s.

Proof. The proof is immediate from the assumptions that $\mathcal{F}^x, \mathcal{F}^y$ are independent under P and $\mu(B, t)$ is \mathcal{F}_t^x -measurable. \square

Now fix $g \in \mathcal{G}$. Since L_t satisfies

$$L_t = 1 + \int_0^t L_{s-} d(\phi \circ Q)_s,$$

substitution into (5.10) gives

$$(5.11) \quad \begin{aligned} \pi_t(g) &= E(g(y_t)|\mathcal{F}_t^x) + E\left[\int_0^t g(y_t)L_{s-} d(\phi \circ Q)_s|\mathcal{F}_t^x\right] \\ &= E(g(y_t)|\mathcal{F}_t^x) + \int_Z \int_0^t E[g(y_t)L_{s-}\phi(z, y_{s-}, s)|\mathcal{F}_t^x]Q(dz, ds). \end{aligned}$$

Since \mathcal{F}^x and \mathcal{F}^y are independent under P ,

$$(5.12) \quad E(g(y_t)|\mathcal{F}_t^x) = Eg(y_t).$$

Also,

$$(5.13) \quad \begin{aligned} &E[g(y_t)L_{s-}\phi(z, y_{s-}, s)|\mathcal{F}_t^x] \\ &= E[g(y_t)L_s\phi(z, y_s, s)|\mathcal{F}_t^x] && \text{(by Propositions 5.2, 5.3)} \\ &= E[E\{g(y_t)L_s\phi(z, y_s, s)|\mathcal{F}_t^x \vee \mathcal{F}_s^y\}|\mathcal{F}_t^x] \\ &= E[L_s\phi(z, y_s, s)E(g(y_t)|\mathcal{F}_t^x \vee \mathcal{F}_s^y)|\mathcal{F}_t^x] \\ &= E[L_s\phi(z, y_s, s)E(g(y_t)|\mathcal{F}_s^y)|\mathcal{F}_t^x] && \text{(by independence of } \mathcal{F}^x, \mathcal{F}^y) \\ &= E[L_s\phi(z, y_s, s)E(g(y_t)|y_s)|\mathcal{F}_t^x] && \text{(since } y_t \text{ is Markov)} \\ &= E[L_s\phi(z, y_s, s)H_{t,s}(g)|\mathcal{F}_s^x] \end{aligned}$$

since x has independent increments under P and where

$$H_{t,s}(g) = E(g(y_t)|y_s).$$

Substitution of (5.12) and (5.13) into (5.11) gives

$$\pi_t(g) = Eg(y_t) + \int_Z \int_0^t \pi_s(\phi(z, \cdot, s)H_{t,s}(g))Q(dz, ds).$$

Note that the integrand in the above expression is a predictable process for each fixed t , as explained at the end of § 2.9.

We summarize the above.

THEOREM 5.2. *Under Assumption (5.2) the process $\pi_t(g)$ satisfies*

$$(5.14) \quad \pi_t(g) = Eg(y_t) + \int_Z \int_0^t \pi_s(\phi(z, \cdot, s)H_{t,s}(g))Q(dz, ds),$$

where

$$(5.15) \quad H_{t,s}(g) = E(g(y_t)|y_s),$$

and

$$(5.16) \quad Q(B, t) = P(B, t) - \tilde{P}(B, t).$$

Remark 5.2. (i) Because of Proposition 5.4,

$$Eg(y_t) = E_1g(y_t) \quad \text{and} \quad H_{t,s}(g) = E_1(g(y_t)|y_s).$$

(ii) From (5.10), $\pi_t(1) = E(L_t|\mathcal{F}_t^x)$, where 1 denotes the function on Y which is identically equal to unity. Hence from (5.9),

$$\begin{aligned} E_1(g(y_t)|\mathcal{F}_t^x) &= \frac{\pi_t(g)}{\pi_t(1)} \\ &= \frac{Eg(y_t) + \int_Z \int_0^t \pi_s(\phi(z, \cdot, s)H_{t,s}(g))Q(dz, ds)}{1 + \int_Z \int_0^t \pi_s(\phi(z, \cdot, s))Q(dz, ds)} \end{aligned}$$

from (5.14).

(iii) Suppose (x_t, \mathcal{F}_t, P) is as in the fundamental example. Then (5.14) simplifies to

$$(5.17) \quad \pi_t(g) = Eg(y_t) + \sum_{i=1}^n \int_0^t \pi_s(\phi_i(\cdot, s)H_{t,s}(g))[p_i(ds) - \tilde{p}_i(ds)].$$

(iv) We now derive a more familiar-looking version of (5.14). For any set $A \in \mathcal{Y}$,

$$\pi_t(I_A) = E(I_A L_t | \mathcal{F}_t^x).$$

If $P(y_t \in A) = P_1(y_t \in A) = 0$, then $\pi_t(I_A) = 0$ a.s. Hence there exists a measurable function $U_t: Y \rightarrow R$ such that

$$(5.18) \quad \pi_t(A) = \int_A U_t(y)P_t(dy),$$

where P_t is the marginal distribution of y_t under P and P_1 . Evidently if $h \in \mathcal{G}$, then

$$\pi_t(h) = \int_Y h(y)U_t(y)P_t(dy).$$

Next let $P(A, t|y, s)$, $A \in \mathcal{Y}$, $s \leq t$, be the *transition kernel* of the Markov process y so that

$$(H_{t,s}(g))(y) = \int_Y g(y')P(dy', t|y, s)$$

and let $P(A, s|y, t)$, $A \in \mathcal{Y}$, $t \geq s$, be the *backward kernel* so that for $h \in \mathcal{G}$,

$$E(h(y_s)|y_t) = \int_Y h(y')P(dy', s|y_t, t).$$

Substituting these relations into (5.14) leads to

$$\begin{aligned} \int_Y g(y')U_t(y')P_t(dy') &= \int_Y g(y')P_t(dy') + \int_Z \int_0^t \left[\int_Y \left\{ \phi(z, y, s) \right. \right. \\ &\quad \cdot \left. \int_Y g(y')P(dy', t|y, s) \right\} U_s(y)P_s(dy) \Big] \cdot Q(dz, ds) \\ &= \int_Y g(y')P_t(dy') + \int_Y g(y') \\ &\quad \cdot \left[\int_Z \int_0^t \left\{ \int_Y \phi(z, y, s)U_s(y)P(dy, s|y', t) \right\} Q(dz, ds) \right] \cdot P_t(dy'). \end{aligned}$$

Since $g \in \mathcal{G}$ is arbitrary, the process $U_t(y)$ evolves according to

$$(5.19) \quad U_t(y) = 1 + \int_Z \int_0^t \left[\int_Y \phi(z, y', s)U_s(y')P(dy', s|y, t) \right] Q(dz, ds).$$

Remark 5.3. (i) For the case of the fundamental example (see (5.17)) the equation above simplifies to

$$(5.20) \quad U_t(y) = 1 + \sum_{i=1}^n \int_0^t \left\{ \int_Y \phi_i(y', s)U_s(y')P(dy', s|y, t) \right\} [p_i(ds) - \tilde{p}_i(ds)].$$

This equation has been derived in [28] for the special case where (x_t, \mathcal{F}_t, P) is a counting process, so that $n = 1$, and with the additional condition that $\tilde{p}(ds) = ds$.

(ii) Equations (5.14) and (5.15) are *not* yet recursive since the functions $\phi(z, y, t)$, $\phi_i(y, t)$ are allowed to depend on the entire past x_s , $1 \leq s \leq t$. We will see later how under additional conditions these equations become truly recursive.

(iii) Notice that unlike the representation for \hat{y}_t obtained in Theorem 5.1, those for π_t in (5.14) and U_t in (5.19) are not semimartingales because the integrands depend upon t . This dependency can be eliminated by some additional assumptions as follows.

For the remainder of this section the following holds in addition to Assumption 5.2.

Assumption 5.3. The operators $H_{t,s}$ of (5.15) have the following properties:

$$(5.21)(i) \quad \lim_{s \uparrow t} H_{t,s} = I, \text{ the identity operator on } \mathcal{G}.$$

(ii) there exist operators A_t , $t \geq 0$ on \mathcal{G} such that

$$(5.22) \quad \lim_{\varepsilon \downarrow 0} (1/\varepsilon)(H_{t+\varepsilon, s} - H_{t, s})(g) = H_{t, s}A_t(g).$$

We do not elaborate on the precise theoretical status of the operators A_t (i.e., the precise definitions of their domain, range, etc.), since it would take us too far afield and since this topic is well-covered in the semigroup theory of Markov processes (see, e.g., [29]). We merely note that (i) is a continuity assumption, (ii) is a differentiability assumption. The operators A_t are often referred to as the infinitesimal generator, especially when y is a Hunt process. If y is a k -dimensional diffusion, for example, then A_t is just a (partial) differential operator of the form

$$\frac{1}{2} \sum_{i,j=1}^k \sigma_{ij}(y, t) \frac{\partial^2}{\partial y_i \partial y_j} + \sum_{i=1}^k m_i(y, t) \frac{\partial}{\partial y_i}.$$

We now develop the simplifications induced by (5.21), (5.22) in (5.14). First of all, recalling that $P_0(dy)$ is the probability distribution of y_0 and that y is Markov, we get

$$Eg(y_t) = \int_Y (H_{t,0}(g))(y)P_0(dy) = E(H_{t,0}(g)(y_0)).$$

This, together with (5.22), implies that

$$\begin{aligned} E(g(y_{t+\varepsilon}) - g(y_t)) &= \int_Y (H_{t+\varepsilon,0} - H_{t,0})(g)(y)P_0(dy) \\ &\cong \varepsilon \int_Y (H_{t,0}A_t(g))(y)P_0(dy) = \varepsilon E[(A_t(g))(y_t)]. \end{aligned}$$

Substituting this into (5.14), and using (5.21) and (5.22), leads us to

$$\begin{aligned} (\pi_{t+\varepsilon} - \pi_t)(g) &\cong \varepsilon E(A_t(g)) + \varepsilon \int_Z \pi_t(\phi g)Q(dz, dt) + \varepsilon \int_Z \int_0^t \pi_s(\phi H_{t,s}A_t(g))Q(dz, ds) \\ &= \varepsilon \int_Z \pi_t(\phi g)Q(dz, dt) + \varepsilon \pi_t(A_t(g)). \end{aligned}$$

Hence

$$(5.23) \quad \pi_t(g) = \pi_0(g) + \int_0^t \pi_s(A_s g) ds + \int_0^t \int_Z \pi_s(\phi(z, \cdot, s)g)Q(dz, ds).$$

THEOREM 5.3. *Under the additional conditions of Assumption 5.3, the representations (5.14) and (5.17) simplify to (5.23), (5.24) respectively.*

$$(5.24) \quad \pi_t(g) = \pi_0(g) + \int_0^t \pi_s(A_s g) ds + \sum_{i=1}^n \int_0^t \pi_s(\phi_i(\cdot, s)g)[p_i(ds) - \tilde{p}_i(ds)].$$

As an example illustrating (5.24) suppose that under P x_t and y_t are independent standard Poisson processes. Then $Z = Y = I_+$, the set of nonnegative

integers. Also $n = 1$ in (5.24), $p(t) = x_t$ and $\tilde{p}(t) = t$. For $g: I_+ \rightarrow R$,

$$\begin{aligned} H_{t,s}(g)(y) &= E(g(y_t) | y_s = y) \\ &= \sum_{k=0}^{\infty} g(y+k) \frac{(t-s)^k}{k!} e^{-(t-s)}, \end{aligned}$$

so that,

$$\frac{\partial}{\partial t}(H_{t,s}(g))(y) = \sum_{k=0}^{\infty} \frac{(t-s)^k}{k!} e^{-(t-s)} [g(y+k+1) - g(y+k)],$$

and hence

$$(5.25) \quad (A_t g)(y) = (Ag)(y) = g(y+1) - g(y).$$

Consider the “indicator” functions $\delta_k: I_+ \rightarrow R$, where

$$\delta_k(y) = \begin{cases} 1 & \text{if } y = k, \\ 0 & \text{otherwise.} \end{cases}$$

By the linearity of π_t ,

$$\pi_t(g) = \sum_{k=0}^{\infty} g(k) \pi_t(\delta_k),$$

so that it is enough to determine the processes $\pi_t(\delta_k)$, $k = 0, 1, 2, \dots$. Substitution of δ_k for g into (5.24) gives, using (5.25),

$$\begin{aligned} \pi_t(\delta_k) &= \pi_0(\delta_k) + \int_0^t [\pi_s(\delta_{k-1}) - \pi_s(\delta_k)] ds + \int_0^t \pi_s(\phi(\cdot, s) \delta_k)(dx_s - ds) \\ &= \pi_0(\delta_k) + \int_0^t [\pi_s(\delta_{k-1}) - \pi_s(\delta_k)] ds + \int_0^t \phi(k, s) \pi_s(\delta_k)(dx_s - ds), \end{aligned}$$

since $\phi(y, s) \delta_k(y) = \phi(k, s) \delta_k(y)$. Now

$$\pi_0(\delta_k) = E\delta_k(y_0) = \begin{cases} 1 & \text{if } k = 0, \\ 0 & \text{if } k > 0, \end{cases}$$

and $\delta_{-1} \equiv 0$, so that the expression above simplifies to

$$\begin{aligned} \pi_t(\delta_0) &= 1 + \int_0^t \pi_s(\delta_0) ds + \int_0^t \phi(0, s) \pi_s(\delta_0)(dx_s - ds), \\ \pi_t(\delta_k) &= \int_0^t \pi_s(\delta_{k-1}) ds - \int_0^t \pi_s(\delta_k) ds + \int_0^t \phi(k, s) \pi_s(\delta_k)(dx_s - ds), \quad k \geq 1, \end{aligned}$$

and these can be rewritten respectively as

$$\begin{aligned} (5.26) \quad e^t \pi_t(\delta_0) &= 1 + \int_0^t \phi(0, s) e^s \pi_s(\delta_0)(dx_s - ds), \\ e^t \pi_t(\delta_k) &= \int_0^t e^s \pi_s(\delta_{k-1}) ds + \int_0^t \phi(k, s) e^s \pi_s(\delta_k)(dx_s - ds), \quad k \geq 1. \end{aligned}$$

These linear integral equations can now be solved inductively to yield the explicit formulas

$$(5.27) \quad \pi_t(\delta_0) = e^{-t} \prod_{\substack{s \leq t \\ x_s \neq x_s}} [1 + \phi(0, s)] \exp \left[- \int_0^t \phi(0, s) ds \right],$$

$$(5.28) \quad \pi_t(\delta_k) = \int_0^t e^{-(t-s)} \pi_s(\delta_{k-1}) \left\{ \prod_{\substack{s < \tau \leq t \\ x_\tau \neq x_\tau}} [1 + \phi(k, \tau)] \exp \left[- \int_s^t \phi(k, \tau) d\tau \right] \right\} ds,$$

$k \geq 1.$

Remark 5.4. The result just obtained illustrates the power of the formulation of Theorem 5.3 over the more usual formulations which involve obtaining a relation for the conditional density (e.g., [16]). We believe that equations (5.14), (5.20), and (5.23), (5.24) are much more useful since they are *linear* in the “unknown” linear operators π_t , whereas the evolution equations for the conditional density are *nonlinear*. Of course the latter can be easily derived from the former.

REFERENCES

- [1] C. DOLÉANS-DADE AND P. A. MEYER, *Intégrales stochastiques par rapport aux martingales locales*, Séminaire de Probabilités: IV, Lecture Notes in Mathematics, Springer-Verlag, Berlin and New York, 1970, pp. 77–107.
- [2] C. DOLÉANS-DADE, *Quelques applications de la formule de changement de variables pour les semi-martingales*, Z. Wahrscheinlichkeitstheorie Verw. Gebiete, 16 (1970), pp. 181–194.
- [3] P. A. MEYER, *Probabilités et Potentiel*, Hermann, Paris, 1966-English translation: *Probability and Potential*, Blaisdell, Waltham, Mass., 1966.
- [4] R. BOEL, P. VARAIYA AND E. WONG, *Martingales on point processes I: Representation results*, Memo #M-407, Electronics Research Lab., University of California, Berkeley, Calif., 1973.
- [5] H. KUNITA AND S. WATANABE, *On square integrable martingales*, Nagoya Math. J., 30 (1967), pp. 209–245.
- [6] P. M. BRÉMAUD, *A martingale approach to point processes*, Electronics Research Lab., Memo #M-345, University of California, Berkeley, Calif., 1972.
- [7] M. H. A. DAVIS, *Detection of signals with point process observation*, Publication 73/8, Dept. of Computing and Control, Imperial College, London, 1973.
- [8] T. T. KADOTA AND L. A. SHEPP, *Conditions for absolute continuity between a certain pair of probability measures*, Z. Wahrscheinlichkeitstheorie Verw. Gebiete (3), 16 (1960), pp. 13–30.
- [9] I. V. GIRSANOV, *On transforming a certain class of stochastic processes by absolutely continuous substitution of measures*, Theory Probability Appl., 5 (1960), pp. 285–301.
- [10] V. BENEŠ, *Existence of optimal stochastic control laws*, this Journal, 9 (1971), pp. 446–475.
- [11] T. E. DUNCAN AND P. VARAIYA, *On the solutions of a stochastic control system*, this Journal, 9 (1971), pp. 354–371.
- [12] M. H. A. DAVIS AND P. VARAIYA, *Dynamic programming conditions for partially observable stochastic systems*, this Journal, 11 (1973), pp. 226–261.
- [13] R. RISHEL, *Weak solutions of a partial differential equation of dynamic programming*, this Journal, 9 (1971), pp. 519–528.
- [14] D. L. SNYDER, *Information processing for observed jump processes*, Information and Control, 22 (1973), pp. 69–78.
- [15] I. RUBIN, *Regular point processes and their detection*, IEEE Trans. Information Theory IT-18 (1972), pp. 547–557.
- [16] D. L. SNYDER, *Filtering and detection for doubly stochastic Poisson processes*, Ibid., IT-18 (1972), pp. 91–102.
- [17] O. MACCHI AND B. C. PICINBONO, *Estimation and detection of weak optical signals*, Ibid., IT-18 (1972), pp. 562–573.

- [18] E. LEHMANN, *Testing Statistical Hypotheses*, Wiley, New York, 1959.
- [19] P. A. MEYER, *Square integrable martingales, a survey*, Martingales: A report on a Meeting at Oberwolfach, Lecture Notes in Mathematics, No. 190, Springer-Verlag, Berlin, 1970.
- [20] M. H. A. DAVIS, *Nonlinear filtering with point process observations*, preprint.
- [21] M. FUJISAKI, G. KALLIANPUR AND H. KUNITA, *Stochastic differential equations for the nonlinear filtering problem*, Osaka J. Math. (1), 9 (1972), pp. 19–40.
- [22] C. DELLACHERIE, *Capacités et processus stochastiques*, Springer-Verlag, Berlin, 1972.
- [23] T. E. DUNCAN, *On the absolute continuity of measures*, Annals Math. Statist., 41 (1970), pp. 30–38.
- [24] J. H. VAN SCHUPPEN, *Estimation theory for continuous time processes, a martingale approach*, Memo # M-405, Electronics Research Lab., University of California, Berkeley, Calif., 1973.
- [25] J. L. DOOB, *Stochastic Processes*, Wiley, New York, 1953.
- [26] W. FELLER, *An Introduction to Probability Theory and Its Applications*, v.I, Wiley, New York, 1968.
- [27] T. E. DUNCAN AND P. VARAIYA, *On the solutions of a stochastic control system II*, Memo # M406, Electronics Research Lab., Univ. of California, Berkeley, Calif., 1973.
- [28] P. BRÉMAUD, *Filtering for point processes*, preprint, 1973.
- [29] E. B. DYNKIN, *Markov Processes I*, Academic Press, New York, 1965.
- [30] D. L. SNYDER, *Smoothing for doubly stochastic Poisson processes*, IEEE Trans. Information Theory, IT-18 (1972), pp. 558–562.
- [31] A. PRÉKOPA, *On stochastic set functions*, Acta Math. Acad. Sci. Hungar. (2), 7 (1956), pp. 215–263.
- [32] J. E. MOYAL, *The general theory of stochastic population processes*, Acta Math. (Uppsala), 108 (1962), pp. 1–31.
- [33] J. H. VAN SCHUPPEN, *Filtering for counting processes, a martingale approach*, Proc. 4th Symp. Nonlinear Estimation and Appl., San Diego, Calif., 1973.

ON THE CONTROL PROBLEM OF BOLZA IN HILBERT SPACES†

VIOREL BARBU‡

Abstract. The convex control problem of Bolza with state constraints is considered. For this problem the dual problem is studied in relation to certain optimality conditions proved earlier by the author. A subdifferential form of the Hamilton–Jacobi equation is treated in a special case.

1. Introduction. Let V and H be a pair of real Hilbert spaces such that

$$V \subset H \subset V',$$

where each inclusion mapping is continuous and densely defined. Here V' denotes the dual of V and H is identified with its own dual. Let (\cdot, \cdot) be the pairing between v' in V' and v in V ; if $v, v' \in H$ this is the ordinary inner product in H . The norms in V and H are denoted by $\|\cdot\|$ and $|\cdot|$ respectively. We denote by $\|\cdot\|_*$ the norm in V' .

Let $[0, T]$ be a fixed real interval, and let $W(0, T)$ denote the space

$$W(0, T) = \{x \in L^2(0, T; V); x' \in L^2(0, T; V')\},$$

where x' denotes the derivative of x in the sense of vectorial distributions on $]0, T[$.

We denote by $C(0, T; H)$ the usual Banach space of all continuous H -valued functions on $[0, T]$ and by $\mathcal{M}(0, T; H)$ the dual space of $C(0, T; H)$, i.e., the space of all H -valued, regular and bounded measures on $]0, T[$ (see, e.g., [3]). Let K be a closed convex subset of H , and let $\mathcal{K}_0 = \{x \in C(0, T; H); x(t) \in K \text{ for all } t \in [0, T]\}$. By $\mathcal{N}(x, K)$ we shall denote the cone of normals to K at x , i.e.,

$$\mathcal{N}(x, K) = \{\mu \in \mathcal{M}(0, T; H); \mu(x - y) \geq 0 \text{ for all } y \in \mathcal{K}_0\}.$$

Consider the following problem:

Minimize

$$(1.1) \quad \int_0^T L(x(t), u(t)) dt + l(x(0), x(T))$$

subject to

$$(1.2) \quad x'(t) + A(t)x(t) = u(t), \quad 0 < t < T,$$

$$(1.3) \quad x(t) \in K \quad \text{for all } t \in [0, T],$$

where $A(t)$, K , L and l are assumed to satisfy the following conditions:

(A) $\{A(t); 0 \leq t \leq T\}$ is a family of continuous linear operators from V into V' such that

(a) for all u, v in V the function $t \rightarrow (A(t)u, v)$ is measurable on $]0, T[$ and

$$(1.4) \quad \|A(t)u\|_* \leq C\|u\| \quad \text{for } u \in V;$$

† Received by the editors March 19, 1974, and in revised form July 15, 1974.

‡ Department of Mathematics, University of Iasi, Iasi, Romania.

(b) there are α real and $\omega > 0$ such that

$$(1.5) \quad (A(t)u, u) + \alpha|u|^2 \geq \omega\|u\|^2 \quad \text{for all } u \in V.$$

(B) L and l are lower semicontinuous convex functions from $H \times H$ into $]-\infty, +\infty]$ and nonidentically $+\infty$. For every $(x, p) \in H \times H$ the Hamiltonian function

$$H(x, p) = \sup \{(p, v) - L(x, v); v \in H\}$$

is finite.

Let $W_A(0, T)$ denote the Hilbert space of all $x \in W(0, T)$ such that $x' + A(t)x \in L^2(0, T; H)$. We denote by $D(l)$ the effective domain of l , i.e.,

$$D(l) = \{(x_1, x_2) \in H \times H; l(x_1, x_2) < +\infty\}$$

and denote by C_L the set of all pairs $(x(0), x(T)) \in H \times H$, arising from arcs $x \in W_A(0, T)$ which satisfy

$$L(x, x' + A(t)x) \in L^1(0, T), \quad x(t) \in K \quad \text{for } t \in [0, T].$$

The last two assumptions can be stated as follows.

(C) K is a closed convex subset of H . There exists at least one arc $x \in W_A(0, T)$ such that $L(x, x' + A(t)x) \in L^1(0, T)$ and

$$x(t) \in \text{Int } K \quad \text{for every } t \in [0, T], \quad (x(0), x(T)) \in D(l).$$

(D) There exists a pair $(x_1, x_2) \in D(l) \cap C_L$ such that one of the following two conditions holds:

$$(1.6) \quad x_2 \in \text{Int } \{x \in H; (x_1, x) \in D(l)\},$$

$$(1.7) \quad x_2 \in \text{Int } \{x \in H; (x_1, x) \in C_L\}.$$

Let $F: W_A(0, T) \rightarrow]-\infty, +\infty]$ be the lower semicontinuous convex function defined by

$$(1.8) \quad F(x) = \int_0^T L(x, x' + A(t)x) dt + l(x(0), x(T))$$

and let

$$\mathcal{K} = \{x \in W_A(0, T); x(t) \in K \text{ for every } t \in [0, T]\}.$$

Then our control problem can be represented as

$$(1.9) \quad \text{Minimize } F(x) \quad \text{subject to } x \in \mathcal{K}.$$

This problem includes as special cases many types of control problems for systems governed by parabolic differential equations (we refer to [4], [6] for further discussion and examples).

The following characterization of the optimal arcs in the above problem was obtained by the author in [1].

THEOREM 1. *Suppose that assumptions (A), (B), (C) and (D) hold. Then $x \in W_A(0, T)$ is an optimal arc for the problem (1.9) if and only if there exist a function $p \in L^\infty(0, T; H) \cap L^2(0, T; V) \cap BV(0, T; V')$ and a bounded regular*

measure $\mu \in \mathcal{N}(x, K)$ such that

$$(1.10) \quad p' - A^*(t)p - \mu \in L^2(0, T; H)$$

and

$$(1.11) \quad x' + A(t)x - \partial_p H(x, p) \ni 0,$$

$$p' - A^*(t)p + \partial_x H(x, p) \ni \mu,$$

$$(1.12) \quad \{p(0), -p(T)\} \in \partial l(x(0), x(T)).$$

Here $BV(0, T; V')$ denotes the space of all functions $p: [0, T] \rightarrow V'$ of bounded variations on $[0, T]$ and A^* is the adjoint of A . We have denoted by $\partial l \subseteq H \times H$ the subdifferential of l (see, e.g., [2]) and by $\partial H = \{-\partial_x H, \partial_p H\}$ the subdifferential of the concave-convex function H .

It will be proved in § 3 that the functions p which appear in optimality conditions (1.11), (1.12) are minimizing arcs of a certain control problem which can be interpreted as the dual of the problem (1.9). The basic ideas we use to obtain this result are due to R. T. Rockafellar [6], [7]. To this purpose, in § 2 will be proved a variant of Theorem 1, furnishing necessary conditions for optimality. Section 5 is concerned with Hamilton–Jacobi equations for a special case of the problem (1.1).

2. Optimality conditions. Let X be a real reflexive Banach space and let $p: [0, T] \rightarrow X$ be a function of bounded variation. Then $p(t)$ is weakly differentiable a.e. on $[0, T]$, i.e.,

$$\frac{d}{dt}(p(t), x') = (\dot{p}(t), x') \quad \text{for all } x' \in X',$$

and its weak derivative \dot{p} belongs to $L^1(0, T; X)$ (see, e.g., [2, Prop. A.1]).

Let p' denote the derivative of p in the sense of X -valued distributions on $]0, T[$. Then p' is an X -valued bounded regular measure on $]0, T[$ and it can be written as

$$(2.1) \quad p' = p + dp_0,$$

where $p_0(t) = p(t) - \int_0^t \dot{p}(s) ds$ and $dp_0 = p'_0$. The measure dp_0 is just the singular part of p' with respect to Lebesgue measure.

THEOREM 2. *Let the assumptions of Theorem 1 be satisfied. Then, in order that $x \in W_A(0, T)$ be optimal in the problem (1.9), it is necessary and sufficient that there exists a function $p \in L^\infty(0, T; H) \cap L^2(0, T; V) \cap BV(0, T; V')$ such that $\dot{p} - A^*(t)p \in L^1(0, T; H)$ and*

$$(2.2) \quad x'(t) + A(t)x(t) - \partial_p H(x(t), p(t)) \ni 0 \quad \text{a.e. } t \in]0, T[,$$

$$(2.3) \quad \dot{p}(t) - A^*(t)p(t) + \partial_x H(x(t), p(t)) - \partial I_K(x(t)) \ni 0 \quad \text{a.e. } t \in]0, T[,$$

$$(2.4) \quad \{p(0), -p(T)\} \in \partial l(x(0), x(T)),$$

$$(2.5) \quad \text{the singular part of } p' \text{ belongs to } \mathcal{N}(x, K).$$

Here $\partial I_K(x)$ denotes the cone of normals to K at x , i.e.,

$$(2.6) \quad \partial I_K(x) = \{y \in H; (y, x - u) \geq 0 \text{ for all } u \in K\}.$$

Proof of Theorem 2. Let $x \in W_A(0, T)$ be an optimal arc of the problem (1.9). According to Theorem 1 there exist $p \in L^\infty(0, T; H) \cap L^2(0, T; V) \cap BV(0, T; V')$ and a measure $\mu \in \mathcal{N}(x, K)$ satisfying the conditions (1.10), (1.11) and (1.12). In particular (1.10) implies that there exists $q \in BV(0, T; V')$ such that $q' = \mu$. As $\mu \in \mathcal{M}(0, T; H)$ we have

$$|q'(\varphi)| \leq C \|\varphi\|_{C(0, T; H)} \quad \text{for all } \varphi \in C^1(0, T; H),$$

or, by the Hahn–Banach theorem this implies that $q \in L^\infty(0, T; H)$ and

$$\left| \int_0^T (q(t), \varphi'(t)) dt \right| \leq C \|\varphi\|_{C(0, T; H)} \quad \text{for all } \varphi \in C^1(0, T; H).$$

We may conclude therefore (see [2, Prop. A.5]), as an intermediate step, that $q \in BV(0, T; H)$. Then we can write μ as

$$(2.7) \quad \mu = \dot{q} dt + dq_0,$$

where $q, q_0 \in BV(0, T; H)$ and $\dot{q}_0(t) = 0$ a.e. $t \in]0, T[$. Similarly,

$$(2.8) \quad p' = \dot{p} dt + dp_0,$$

where $p_0 \in BV(0, T; V')$ and $\dot{p}_0(t) = 0$ a.e. $t \in]0, T[$. Since $dp_0 - dq_0 \in L^1(0, T; V')$ in virtue of (1.10), we conclude, therefore, that $dp_0 = dq_0$. In particular, it follows that

$$\dot{p} - A^*(t)p \in L^1(0, T; H).$$

Next, we shall prove that

$$(2.9) \quad \dot{q}(t) \in \partial I_K(x(t)) \quad \text{a.e. } t \in]0, T[.$$

Let $t_0 \in]0, T[$ be an arbitrary Lebesgue point of $\dot{q}(t)$ such that $\dot{q}_0(t_0) = 0$. For arbitrary $\varepsilon > 0$ and $u \in K$, define

$$u_\varepsilon(t) = \begin{cases} x(t) & \text{if } |t - t_0| \geq \varepsilon, \\ \left(1 - \frac{t_0 - t}{\varepsilon}\right)u + \frac{t_0 - t}{\varepsilon}x(t_0 - \varepsilon) & \text{if } t \in [t_0 - \varepsilon, t_0], \\ \left(1 - \frac{t - t_0}{\varepsilon}\right)u + \frac{t - t_0}{\varepsilon}x(t_0 + \varepsilon) & \text{if } t \in [t_0, t_0 + \varepsilon]. \end{cases}$$

Clearly $u_\varepsilon \in C(0, T; H)$ and $u_\varepsilon(t) \in K$ for every $t \in [0, T]$. Since $\mu \in \mathcal{N}(x, K)$, by (2.7) it follows that

$$(2.10) \quad \int_0^T (\dot{q}(t), x(t) - u_\varepsilon(t)) dt + dq_0(x - u_\varepsilon) \geq 0.$$

We set $\varphi_\varepsilon(t) = \varepsilon^{-1}(x(t) - u_\varepsilon(t))$ and note that

$$(2.11) \quad \lim_{\varepsilon \rightarrow 0} \int_0^T (\dot{q}(t), \varphi_\varepsilon(t)) dt = (q(t_0), x(t_0) - u).$$

On the other hand, let ψ_ε be the scalar continuous function defined by

$$\psi_\varepsilon(t) = \begin{cases} 0 & \text{if } |t - t_0| \geq \varepsilon, \\ 1 - \frac{t_0 - t}{\varepsilon} & \text{if } t \in [t_0 - \varepsilon, t_0], \\ 1 - \frac{t - t_0}{\varepsilon} & \text{if } t \in [t_0, t_0 + \varepsilon]. \end{cases}$$

Denote by $q_w(t)$ the function $(q_0(t), x(t_0) - u)$. Then one easily deduces that

$$\lim_{\varepsilon \rightarrow 0} dq_0(\varphi_\varepsilon) = \lim_{\varepsilon \rightarrow 0} \varepsilon^{-1} \int_{t_0 - \varepsilon}^{t_0 + \varepsilon} \psi_\varepsilon(t) dq_w(t).$$

Consequently,

$$(2.12) \quad \lim_{\varepsilon \rightarrow 0} dq_0(\varphi_\varepsilon) = 0,$$

because $\dot{q}_w(t_0) = 0$. By (2.10), (2.11) and (2.12) one obtains

$$(\dot{q}(t_0), x(t_0) - u) \geq 0,$$

which implies the desired relation (2.9).

It remains to show that $dq_0 \in \mathcal{N}(x, K)$, i.e.,

$$dq_0(x - y) \geq 0 \quad \text{for all } y \in \mathcal{K}.$$

To this purpose we consider an arbitrary positive measure ν on $[0, T]$ with respect to which dq_0 is absolutely continuous. According to the Lebesgue–Nikodym theorem for vectorial-valued measures (see [3, Chap. IV, Thm. 3]) there exists a ν -integrable function $f: [0, T] \rightarrow H$ such that $dq_0 = f \cdot \nu$.

We have

$$(2.13) \quad \int_0^T (\dot{q}(t), x(t) - y(t)) dt + \int_0^T (f(t), x(t) - y(t)) d\nu(t) \geq 0 \quad \text{for all } y \in K,$$

and this inequality clearly extends to all bounded measurable functions $y: [0, T] \rightarrow H$ such that $y(t) \in K$ for all $t \in [0, T]$.

Since the measure dq_0 is singular with respect to Lebesgue measure on $]0, T[$, there exists $E \subseteq [0, T]$ such that $dq_0(F) = 0$ for any $F \subseteq [0, T] \setminus E$ and $m(E) = 0$ (we have denoted by m Lebesgue measure). Let $y \in C(0, T; H)$ be such that $y(t) \in K$ for any $t \in [0, T]$ and let $\tilde{y}(t) = y(t)$ for $t \in E$ and $y(t) = x(t)$ for $t \in [0, T] \setminus E$.

By (2.13) it follows that

$$dq_0(x - y) = \int_E (f(t), x(t) - y(t)) d\nu(t) \geq 0,$$

as claimed. Moreover, it follows by a standard device that

$$f(t) \in \partial I_K(x(t)), \quad \nu - \text{a.e.} \quad \text{on } [0, T].$$

Conversely, suppose that x and p satisfy the optimality conditions (2.2), (2.3), (2.4) and (2.5). Since H is the Hamiltonian corresponding to L , the conditions (2.2) and (2.3) can be expressed equivalently in the following form (see [1], [6]):

$$\{\dot{p} - A^*(t)p - \partial I_K(x), p\} \in \partial L(x, x' + A(t)x),$$

where ∂L is the subdifferential of the convex function $L: H \times H \rightarrow]-\infty, +\infty]$. Consequently,

$$(2.14) \quad \begin{aligned} &(\dot{p} - A^*(t)p, x - y) + (p, x' + A(t)x - y' - A(t)y) \\ &\geq L(x, x' + A(t)x) - L(y, y' + A(t)y) \end{aligned}$$

for all $y \in \mathcal{K}$. Let $\mu_p = dp_0$ denote the singular part of p' . By (2.5) we have

$$(2.15) \quad \mu_p(x - y) \geq 0 \quad \text{for all } y \in \mathcal{K}.$$

Integrating both sides of (2.13) and adding to (2.14) we get

$$F(x) \leq F(y) \quad \text{for all } y \in \mathcal{K}$$

because

$$\begin{aligned} p'(x - y) + p(x' - y') &= (p(T), x(T) - y(T)) - (p(0), x(0) - y(0)) \\ &\leq -l(x(0), x(T)) + l(y(0), y(T)) \end{aligned}$$

in virtue of the transversality conditions (2.5). This completes the proof of Theorem 2.

Remarks. In particular, condition (2.5) shows that the singular part of p' is concentrated in the set of all $t \in [0, T]$ for which $x(t)$ lies on the boundary of K . Moreover, if t_0 is arbitrary in $[0, T]$, the preceding proof shows that

$$p(t_0^+) - p(t_0^-) \in \partial I_K(x(t_0)).$$

In fact the inequality (2.10) implies

$$\lim_{\varepsilon \rightarrow 0} dq_0(x - u_\varepsilon) = \lim_{\varepsilon \rightarrow 0} \int_{t_0 - \varepsilon}^{t_0 + \varepsilon} \psi_\varepsilon(t) dq_w(t) \geq 0$$

because $\lim_{\varepsilon \rightarrow 0} \int_0^T (\dot{q}(t), x(t) - u_\varepsilon(t)) dt = 0$. Thus by a simple calculation one obtains $q_w(t_0^+) - q_w(t_0^-) \geq 0$. Hence

$$(q_0(t_0^+) - q_0(t_0^-), x(t_0) - u) \geq 0 \quad \text{for all } u \in K.$$

Since $dp_0 = dq_0$, by (2.1) one obtains the desired relation.

3. Dual control problem. Let M and m be the convex functions from $H \times H$ into $]-\infty, +\infty]$ defined by

$$(3.1) \quad M(p, q) = \sup \{(p, v) + (q, x) - L(x, v); x \in K, v \in H\},$$

respectively,

$$(3.2) \quad m(p_1, p_2) = \sup \{(p_1, x_1) - (p_2, x_2) - l(x_1, x_2); x_1, x_2 \in H\}.$$

Let $B_{A^*}(0, T)$ denote the space of all functions $p \in L^2(0, T; V) \cap BV(0, T; V') \cap L^\infty(0, T; H)$ such that

$$(3.3) \quad \dot{p} - A^*(t)p \in L^1(0, T; H), \quad \mu_p \in \mathcal{M}(0, T; H),$$

where \dot{p} is the weak derivative of p and μ_p is the *singular part* of the measure $p' \in \mathcal{M}(0, T; V')$. It is easy to see that if $p \in B_{A^*}(0, T)$, then $p(t_0^+)$ and $p(t_0^-)$ exist (in H) at all points $t_0 \in [0, T[$ (respectively $]0, T]$). In this context we regard $p(0)$ as $p(0^-)$ and $p(T)$ as $p(T^+)$.

We consider the function $\bar{G}: B_{A^*}(0, T) \rightarrow]-\infty, +\infty]$ defined by

$$(3.4) \quad \bar{G}(p) = G(p) + H_K(\mu_p),$$

where

$$(3.5) \quad G(p) = \int_0^T M(p, \dot{p} - A^*(t)p) dt + m(p(0), p(T)),$$

while

$$(3.6) \quad H_K(\mu_p) = \sup \{ \mu_p(y); y \in \mathcal{K}_0 \}.$$

Here \mathcal{K}_0 is the closed convex subset of $C(0, T; H)$ defined by

$$\mathcal{K}_0 = \{ y \in C(0, T; H); y(t) \in K \text{ for all } t \in [0, T] \}.$$

It may be ascertained that \bar{G} is convex and nonidentically $+\infty$.

We call the *dual problem* associated with (1.9) to be that of minimizing over the space $B_{A^*}(0, T)$ the functional \bar{G} defined by (3.4).

The functions p which appear in Theorem 1 are just the solutions of the *dual problem*. More precisely, we have the following.

THEOREM 3. *Let the assumptions of Theorem 1 hold. Then the optimality conditions (1.11) and (1.12) are satisfied by $x \in W_A(0, T)$ and $p \in B_{A^*}(0, T)$ if and only if*

$$(3.7) \quad \bar{G}(p) = \min \{ \bar{G}(q); q \in B_{A^*}(0, T) \} = -F(x) = -\min \{ F(y); y \in \mathcal{K} \}.$$

Proof. Suppose that x and p satisfy the optimality conditions (1.11) and (1.12). By Theorem 2 they satisfy the conditions (2.2) and (2.3) which can be expressed equivalently in the Lagrangian form

$$\{ \dot{p} - A^*(t)p, p \} \in \partial \{ L(x, x' + A(t)x) + I_K(x) \} \quad \text{a.e. } t \in]0, T[.$$

Here I_K denotes the indicator function of the convex K . Since M is the conjugate function of $L(x, v) + I_K(x)$, one gets

$$\{ x' + A(t)x, x \} \in \partial M(p, \dot{p} - A^*(t)p) \quad \text{a.e. } t \in]0, T[.$$

Consequently,

$$(3.8) \quad \begin{aligned} M(p, \dot{p} - A^*(t)p) &\leq M(q, \dot{q} - A^*(t)q) + (x' + A(t)x, p - q) \\ &\quad + (x, \dot{p} - \dot{q} - A^*(t)p + A^*(t)q) \quad \text{a.e. } t \in]0, T[. \end{aligned}$$

Integrating both sides of (3.8), we get

$$\begin{aligned} \int_0^T M(p, \dot{p} - A^*(t)p) dt &\leq \int_0^T M(q, \dot{q} - A^*(t)q) dt + (x(T), p(T) - q(T)) \\ &\quad - (x(0), p(0) - q(0)) - (\mu_p - \mu_q, x) \\ &\quad \text{for all } q \in B_{A^*}(0, T). \end{aligned}$$

By using the transversality conditions (2.4), one finally finds that

$$G(p) + \mu_p(x) \leq G(q) + \mu_q(x) \quad \text{for all } q \in B_{A^*}(0, T).$$

Here, we have used the conjugacy relation (3.2). As $\mu_p \in \mathcal{N}(x, K)$, it follows that $\mu_p(x) = H_K(\mu_p)$ which clearly implies that p is a minimizing arc for \bar{G} on $B_{A^*}(0, T)$.

On the other hand, by (2.2), (2.3) and (2.4) it follows that

$$(3.9) \quad M(p, \dot{p} - A^*(t)p) + L(x, x' + A(t)x) = (p, x' + A(t)x) + (\dot{p} - A^*(t)p, x),$$

respectively,

$$m(p(0), p(T)) + l(x(0), x(T)) = (p(0), x(0)) - (p(T), x(T)).$$

Integrating both sides of (3.9) and adding $\mu_p(x)$, one obtains the desired equality

$$\bar{G}(p) + F(x) = 0.$$

Conversely, suppose that x and p satisfy the conditions (3.7) in Theorem 3. From the definitions it is immediate that

$$(3.10) \quad M(q, \dot{q} - A^*(t)q) + L(y, y' + A(t)y) \geq (y' + A(t)y, q) + (y, \dot{q} - A^*(t)q)$$

and

$$(3.11) \quad l(y(0), y(T)) + m(q(0), q(T)) \geq (y(0), q(0)) - (y(T), q(T))$$

for all $y \in \mathcal{K}_0$ and $q \in B_{A^*}(0, T)$. Integrating (3.10) and adding (3.11) and $H_K(\mu_q)$, one gets

$$(3.12) \quad \bar{G}(q) + F(y) \geq 0$$

for all $y \in \mathcal{K}_0$ and $q \in B_{A^*}(0, T)$. Since $\bar{G}(p) + F(x) = 0$, by (3.10) and (3.11) one finds that

$$(3.13) \quad M(p, \dot{p} - A^*(t)p) + L(x, x' + A(t)x) = (x' + A(t)x, p) + (x, \dot{p} - A^*(t)p),$$

respectively,

$$(3.14) \quad l(x(0), x(T)) + m(p(0), p(T)) = (x(0), p(0)) - (x(T), p(T)).$$

In view of the conjugacy relations (3.1) and (3.2), condition (3.13) is equivalent with (2.2), (2.3) and condition (3.14) is equivalent with (2.4). It remains only to prove the condition (2.5).

By (3.13) and (3.14) it follows that

$$\mu_p(x) = -F(x) - G(p).$$

Together, the conditions (3.7) and this equality imply that $\mu_p(x) = H_K(\mu_p)$. Hence $\mu_p \in \mathcal{N}(x, K)$, which concludes the proof of Theorem 3.

4. An example. Let Ω be a bounded open subset of R^n with sufficiently smooth boundary Γ . Consider the following optimal problem:

Minimize

$$(4.1) \quad F(y, u) = \int_0^T L(y(t), u(t)) dt$$

in $y \in L^2(0, T; H_0^1(\Omega))$ and $u \in L^2(Q)$, $Q =]0, T[\times \Omega$ subject to the constraints

$$(4.2) \quad \partial y / \partial t - \Delta y = u \quad \text{in } Q, \quad y = 0 \text{ in } \Sigma =]0, T[\times \Gamma,$$

$$(4.3) \quad y(0, x) \in Y_0, \quad y(T, x) \in Y_T \quad \text{in } \Omega,$$

and

$$(4.4) \quad \int_{\Omega} |y(t, x)|^2 dx \leq r^2 \quad \text{for } t \in [0, T],$$

where L is a lower semicontinuous convex function from $L^2(\Omega) \times L^2(\Omega)$ to $] -\infty, +\infty]$ and Y_0, Y_T are nonempty closed convex subsets of $L^2(\Omega)$. In addition, the following conditions will be assumed:

(a) $L(y(t), v(t)) \in L^1(0, T)$ for every pair $(y, v) \in L^2(Q) \times L^2(Q)$. Moreover,

$$(4.5) \quad \lim_{\|v\|_{L^2(\Omega)} \rightarrow +\infty} L(y, v) / \|v\|_{L^2(\Omega)} = +\infty, \quad y \in L^2(\Omega);$$

(b) $Y_0 \cap \text{Int } S(0, r) \neq \emptyset$ and $Y_T \cap \text{Int } S(0, r) \neq \emptyset$.

Here $S(0, r)$ denotes the ball with center in 0 and of radius r in $L^2(\Omega)$.

As we observed in [1] the assumptions (A), (B), (C) and (D) are then satisfied with $V = H_0^1(\Omega)$, $H = L^2(\Omega)$ and $K = S(0, r)$. Thus by Theorem 2,

$$y \in L^2(0, T; H_0^1(\Omega))$$

and $u \in L^2(Q)$ is an extremal pair for the problem (4.1)–(4.4) if and only if there exist a function $p \in L^2(0, T; H_0^1(\Omega)) \cap BV(0, T; H^{-1}(\Omega))$ and $\lambda \in L^1(0, T)$ such that

$$(4.6) \quad y_t - \Delta y - \partial_p H(y, p) \ni 0 \quad \text{a.e. in } Q,$$

$$(4.7) \quad \dot{p}_t + \Delta p + \partial_y H(y, p) \ni \lambda y \quad \text{a.e. in } Q,$$

$$(4.8) \quad p(0, x) \in \mathcal{N}_1(y(0, x)), \quad p(T, x) \in \mathcal{N}_2(y(T, x)) \quad \text{a.e. in } \Omega,$$

$$(4.9) \quad \lambda(t)(r - \|y(t)\|_{L^2(\Omega)}) = 0 \quad \text{a.e. } t \in]0, T[,$$

$$(4.10) \quad \mu_p(y) = r \|\mu_p\|_{\mathcal{M}(0, T; L^2(\Omega))}.$$

Here μ_p is the singular part of the measure $p_t \in \mathcal{M}(0, T; L^2(\Omega))$ and \dot{p}_t denotes the weak derivative of the function $t \rightarrow p(t, \cdot)$ defined from $[0, T]$ into $L^2(\Omega)$. $\mathcal{N}_1(y(0, x))$ and $\mathcal{N}_2(y(T, x))$ are the cones of normals to Y_0 (respectively Y_T) at the point $y(0, x)$ (respectively $y(T, x)$).

The functional \bar{G} associated with our problem can be written as

$$\bar{G}(p) = r \|\mu_p\| + \int_0^T M(p, \dot{p}_t + \Delta p) dt + H_1(p(0)) - H_2(p(T)),$$

where $\|\cdot\|$ is the norm in $\mathcal{M}(0, T; H)$, M is the conjugate function of L and H_1 (respectively H_2) is the support function of Y_0 (respectively Y_T).

5. Hamilton–Jacobi equation. This section deals with the Hamilton–Jacobi equation associated with the following control problem:

Minimize

$$(5.1) \quad \int_0^T L(x(t), u(t)) dt$$

subject to

$$(5.2) \quad \begin{aligned} x'(t) + Ax(t) &= u(t) \quad \text{a.e. } t \in]0, T[, \\ x(0) &= x_0, \end{aligned}$$

where $u \in L^2(0, T; H)$ and $x \in W(0, T)$.

We shall make use of the following condition which is stronger than (B):

(i) $H(x, p)$ is finite for all $(x, p) \in H \times H$ and for every $r > 0$ there exists $M_r > 0$ such that

$$(5.3) \quad H(x, p) \leq M_r(1 + |p|^2)$$

and

$$(5.4) \quad |\partial_x H(x, p)| \leq M_r(1 + |p|)$$

for all $p \in H$ and $|x| \leq r$.

We have used the notation

$$|\partial_x H(x, p)| = \sup \{|y|; y \in \partial_x H(x, p)\}.$$

In place of (A) we shall consider the following assumption:

(ii) A is a linear continuous (independent of t) operator from V to V' satisfying

$$(5.5) \quad (Au, u) + \alpha|u|^2 \geq \omega\|u\|^2 \quad \text{for all } u \in V$$

for some real α and some positive ω .

For every $s \in [0, T]$ let us denote by D_s the set of all $h \in H$ such that $[h, h_0] \in C_L^s$ for some $h_0 \in H$. Here C_L^s is defined as in § 1 but with $[s, T]$ instead of $[0, T]$. We note, incidentally, that $D_{s_1} \subset D_{s_2}$ if $s_1 \leq s_2$.

Let J_s be the extended-real-valued function on H defined by

$$J_s(h) = \inf \left\{ \int_s^T L(x, x' + Ax) dt, x \in W_A(s, T); x(s) = h \right\}.$$

LEMMA 1. Suppose that (i) and (ii) hold. Then the function J_s is convex, lower semicontinuous and nowhere $-\infty$. Moreover, for every choice of h in D_s the infimum defining $J_s(h)$ is attained and

$$(5.6) \quad D(\partial J_s) = D(J_s) = D_s, \quad 0 \leq s \leq T.$$

Proof. Let F_s be the lower semicontinuous convex function on $W_A(s, T)$ defined by

$$F_s(x) = \int_s^T L(x, x' + Ax) dt,$$

where

$$D(F_s) = \{x \in W_A(s, T); L(x, x' + Ax) \in L^1(0, T); x(s) = h\}$$

and the norm in $W_A(s, T)$ is

$$(5.7) \quad \|x\|_1^2 = |x(s)|^2 + \int_s^T |x' + Ax|^2 dt.$$

If $h \in D_s$, then $F_s \neq +\infty$ and there exists $\{x_n\} \subset D(F_s)$ such that $F_s(x_n) \rightarrow J_s(h)$ as $n \rightarrow +\infty$. In order to prove that the infimum $J_s(h)$ of F_s is attained, it suffices to show that $\{x_n\}$ is a bounded sequence of $W_A(s, T)$.

From the definition of $H(x, p)$, we have

$$L(x_n, x'_n + Ax_n) \geq (p, x'_n + Ax_n) - H(x_n, p) \quad \text{for all } p \in H.$$

We then have

$$(5.8) \quad L(x_n, x'_n + Ax_n) \geq (p, x'_n + Ax_n) - H(0, p) - (\partial_x H(0, p), x_n).$$

We recall that

$$\partial_x H(x, p) = \{y \in H; H(y, p) \leq H(x, p) + (v, y - x) \text{ for all } y \in H\}.$$

In (5.8) we take $p = \varepsilon(x'_n + Ax_n)$, where ε is positive and sufficiently small. By

(i) above one has

$$(5.9) \quad \int_s^t |x'_n + Ax_n|^2 dt \leq C \left(1 + J_s(h) + \int_s^t |x_n|^2 d\tau - \int_t^T L(x_n, x'_n + Ax_n) d\tau \right)$$

for all $s \leq t \leq T$. (We shall use C to denote several constants all independent of t and n .)

Again using (5.8) one obtains

$$(5.10) \quad \int_s^t |x'_n + Ax_n|^2 dt \leq C \left(1 + J_s(h) + \int_s^t |x_n|^2 d\tau \right) + \rho \int_s^T |x_n|^2 d\tau + C\rho,$$

where ρ is positive and arbitrary small. Denote $x'_n + Ax_n = f_n$. Condition (ii) above then implies that

$$|x_n(t)|^2 + 2\omega \int_s^t \|x_n(\tau)\|^2 d\tau \leq (2\alpha + 1) \int_s^t |x_n(\tau)|^2 d\tau + \int_s^t |f_n(\tau)|^2 d\tau + |h|^2.$$

Making use of (5.10), one finally obtains that

$$|x_n(t)|^2 \leq C \left(1 + C_\rho + \rho \int_s^T |x_n(\tau)|^2 d\tau \right), \quad s \leq t \leq T,$$

which implies that $\{x_n(t)\}$ is uniformly bounded on $[s, T]$. Again using the estimate (5.10), it follows that

$$\int_s^T |x'_n + Ax_n|^2 dt \leq C,$$

as claimed.

In particular, we have proved that $J_s(h) \neq -\infty$ for every $h \in H$. It is also obvious that $D(J_s) = D_s$. We show now that J_s is lower semicontinuous on H .

Let $\{h_n\} \subset D_s$ be such that $h_n \rightarrow h$ and

$$(5.11) \quad J_s(h_n) \leq M.$$

According to the first part of the proof, there exists a sequence $\{x_n\} \subset W_A(s, T)$ such that $x_n(s) = h_n$ and $J_s(h_n) = F_s(x_n)$. Since $\{x_n\}$ is clearly bounded in $W_A(s, T)$, without any loss of generality we may assume that

$$x_n \rightarrow x \quad \text{weakly in } W_A(s, T).$$

Obviously $x(s) = h$ and from (5.11) it follows that

$$F_s(x) \leq M$$

because F_s is lower semicontinuous on $W_A(s, T)$. Thus $J_s(h) \leq M$, which is the desired result.

Let $h \in D_s$ and $x_s \in W_A(s, T)$ be such that $F_s(x_s) = J_s(h)$. The conditions of Theorem 1 being satisfied, there exists at least one function $p_s \in W_{A^*}(s, T)$ such that

$$(5.12) \quad \begin{aligned} x'_s + Ax_s - \partial_p H(x_s, p_s) &\ni 0, \\ p'_s - A^*p_s + \partial_x H(x_s, p_s) &\ni 0, \end{aligned} \quad \text{a.e. } t \in]s, T[,$$

and

$$(5.13) \quad x_s(s) = h, \quad p_s(T) = 0.$$

Thus by a simple calculation involving (5.12) and (5.13) one deduces that

$$(5.14) \quad -p_s(s) \in \partial J_s(h).$$

In particular, this shows that $D(\partial J_s) = D_s$, which completes the proof.

Remark. For every $h \in D_s$, $-\partial J_s(h)$ coincides with the set of all $p(s)$ which appear in (5.12) and (5.13). To prove this, it suffices to show that the mapping $h \rightarrow -p(s)$ is maximal monotone in $H \times H$. But in view of Theorem 1 this is equivalent to the following assertion: For every $a \in H$ the function

$$\int_s^T L(x, x' + Ax) dt + \frac{1}{2}|x(s) - a|^2$$

attains its infimum on $W_A(s, T)$. We observe that the latter follows by using the same argument as that used in the proof of Lemma 1.

We now turn to the problem (5.1), (5.2). If $x_0 \in D_0$, then this problem has at least one optimal arc $x \in W_A(0, T)$ satisfying the equations

$$(5.15) \quad \begin{aligned} x' + Ax - \partial_p H(x, p) &\ni 0, \\ p' - A^*p + \partial_x H(x, p) &\ni 0, \end{aligned} \quad \text{a.e. } t \in]0, T[,$$

while

$$(5.16) \quad x(0) = x_0, \quad p(T) = 0.$$

As we have observed before,

$$p(t) \in -\partial J_t(x(t)) \quad \text{for } t \in [0, T],$$

so that the optimal control $u(t)$ can be expressed as

$$(5.17) \quad u(t) \in \partial_p H(x(t), -\partial J_t(x(t))) \quad \text{a.e. } t \in]0, T[.$$

Thus the mapping ∂J_t can be regarded as the *synthesizing function* for the above control problem. Like in the classical theory of the calculus of variations, it turns out that the function J_t is a solution of a certain functional equation (*Hamilton–Jacobi equation*) associated with the problem (5.1), (5.2).

THEOREM 4. *Assume that conditions (i), (ii) hold. Then, for every $h \in D_0$ such that $Ah \in H$, the function $s \rightarrow J_s(h)$ is Lipschitzian on $[0, T]$ and satisfies*

$$(5.18) \quad \frac{d}{ds} J_s(h) - H(h, -\partial J_s(h)) + (Ah, \partial J_s(h)) \ni 0 \quad \text{a.e. } s \in]0, T[,$$

$$(5.19) \quad J_T(h) = 0.$$

Proof. Let $h \in D_0$. Then by Lemma 1 there exists at least one arc $x_s \in W_A(s, T)$ such that

$$(5.20) \quad J_s(h) = \int_s^T L(x_s, x'_s + Ax_s) dt.$$

In addition, there exists $p_s \in W_{A^*}(s, T)$ satisfying together x_s and the equations (5.15). Equivalently,

$$(5.21) \quad \{p'_s - A^*p_s, p_s\} \in \partial L(x_s, x'_s + Ax_s) \quad \text{a.e. } t \in]s, T[.$$

By assumption, there exists $x \in W_A(0, T)$ such that $x(0) = h$ and $L(x, x' + Ax) \in L^1(0, T)$. Since $J_s(h) \leq \int_0^{T-s} L(x, x' + Ax) dt$, by using the same argument as that used in the proof of Lemma 1, one deduces that

$$(5.22) \quad \int_s^T |x'_s + Ax_s|^2 dt \leq C \quad \text{for all } 0 \leq s \leq T.$$

Since $Ah \in H$, the latter implies that $x'_s \in L^2(s, T; H)$ (see Lions–Magenes [5]) and

$$(5.23) \quad \int_s^T |x'_s|^2 dt \leq C \quad \text{for all } 0 \leq s \leq T,$$

where C is a positive constant independent of s . Let us denote $g_s = p'_s - A^*p_s$. In view of (5.23) the sequence $x_s(t)$ is uniformly bounded on $[s, T]$. Thus by (5.4) and (5.12) one finds that

$$(5.24) \quad |g_s(t)| \leq M(1 + |p_s(t)|) \quad \text{a.e. } t \in]s, T[.$$

On the other hand, from (ii) we deduce that

$$\frac{1}{2} \frac{d}{dt} |p_s(t)|^2 + \alpha |p_s(t)|^2 \geq \omega \|p_s(t)\|^2 - |g_s(t)| |p_s(t)| \quad \text{a.e. } t \in]s, T[.$$

Integrating from t to T and using (5.24) yields

$$\frac{1}{2} |p_s(t)|^2 + \omega \int_t^T \|p_s(\tau)\|^2 d\tau \leq M \int_t^T |p_s(\tau)| (1 + |p_s(\tau)|) d\tau,$$

which implies that

$$(5.25) \quad |p_s(t)| \leq C \quad \text{for all } 0 \leq s \leq t \leq T.$$

Note that (5.24) and (5.25) also imply that

$$(5.26) \quad \int_s^T |p'_s(t)|^2 dt \leq C \quad \text{for all } s \in [0, T].$$

Let s be arbitrary in $[0, T]$ and let $\varepsilon > 0$ be such that $0 < s + \varepsilon < T$.

We have

$$J_{s+\varepsilon}(h) \leq \int_{s+\varepsilon}^T L(x_s(t-\varepsilon), x'_s(t-\varepsilon) + Ax_s(t-\varepsilon)) dt.$$

Consequently,

$$(5.27) \quad J_{s+\varepsilon}(h) - J_s(h) \leq - \int_{T-\varepsilon}^T L(x_s, x'_s + Ax_s) dt.$$

On the other hand, by (5.21) we have

$$(5.28) \quad J_s(h) - J_{s+\varepsilon}(h) \leq \int_s^{s+\varepsilon} L(x_s, x'_s + Ax_s) dt + (p_s(s+\varepsilon), x_s(s) - x_s(s+\varepsilon)).$$

From (5.12) it follows by a standard argument that

$$\frac{d}{dt}((Ax_s(t), p_s(t)) - H(x_s(t), p_s(t))) = 0 \quad \text{a.e. } t \in]s, T[.$$

Therefore $(Ax_s(t), p_s(t)) - H(x_s(t), p_s(t)) = \gamma(s)$ is constant over $[s, T]$. Moreover, from the definition of $H(x, p)$, we have

$$L(x_s, x'_s + Ax_s) = (x'_s + Ax_s, p_s) - H(x_s, p_s) \quad \text{a.e. } t \in]s, T[,$$

because $x'_s + Ax_s \in \partial_p H(x_s, p_s)$.

Thus we conclude that

$$(5.29) \quad L(x_s, x'_s + Ax_s) = \gamma(s) + (x'_s, p_s) \quad \text{a.e. } t \in]s, T[.$$

From (5.27), (5.28) and (5.29) we see that

$$(5.30) \quad |J_{s+\varepsilon}(h) - J_s(h) + \varepsilon\gamma(s)| \leq \sup \left\{ \int_{T-\varepsilon}^T |x'_s| |p_s| dt, \int_s^{s+\varepsilon} |x'_s(t)| |p_s(t) - p_s(s+\varepsilon)| dt \right\}.$$

Noting that

$$|p_s(t)|^2 \leq \varepsilon \int_{T-\varepsilon}^T |p'_s|^2 dt, \quad T - \varepsilon < t < T,$$

and

$$|p_s(t) - p_s(s+\varepsilon)|^2 \leq \varepsilon \int_s^{s+\varepsilon} |p'_s|^2 dt, \quad s < t < s + \varepsilon,$$

the estimate (5.30) together with (5.23) and (5.26) imply that

$$(5.31) \quad |J_{s+\varepsilon}(h) - J_s(h)| \leq C\varepsilon,$$

where C is independent of s and ε .

The inequality (5.30) implies at the same time that

$$\frac{d}{ds} J_s(h) = -\gamma(s) \quad \text{a.e. } s \in]0, T[,$$

as claimed. Theorem 4 is now established.

REFERENCES

- [1] V. BARBU, *Convex control problem of Bolza in Hilbert spaces*, this Journal, 13 (1975), pp. 754–771.
- [2] H. BRÉZIS, *Opérateurs maximaux monotones et semigroupes de contractions dans les espaces de Hilbert*, Math. Studies 5, North-Holland, Amsterdam, 1973.
- [3] N. DINCULEANU, *Vector measures*, Int. Series Monograph. Pure Appl. Math. 95, Pergamon Press, New York, 1967.
- [4] J. L. LIONS, *Contrôle Optimal des Systèmes Gouvernés par des Équations aux Dérivées Partielles*, Dunod, Paris, 1967.
- [5] J. L. LIONS AND E. MAGENES, *Problèmes aux Limites Nonhomogènes et Applications*, Dunod, Paris, 1968.
- [6] R. T. ROCKAFELLAR, *Existence and duality theorems for convex problems of Bolza*, Trans. Amer. Math. Soc., 159 (1971), pp. 1–40.
- [7] ———, *State constraints in convex control problem of Bolza*, this Journal, 10 (1972), pp. 691–715.

ON THE SOLUTIONS OF A STOCHASTIC CONTROL SYSTEM. II*

TYRONE DUNCAN† AND PRAVIN VARAIYA‡

Abstract. This paper presents generalizations of the work in [1], [2] to include controlled stochastic processes which take values in a certain class of Fréchet spaces. The crucial result is an extension of Girsanov's technique for defining solutions of stochastic differential equations by an absolutely continuous transformation of measures. The result is used to prove existence results for stochastic control problems and for a class of two-person zero sum games.

1. Introduction and summary. Consider a controlled stochastic process represented by the stochastic differential equation

$$dX(t) = f(t, X, u(t, X)) dt + dB(t), \quad t \in [0, 1],$$

where $B(t)$ is a Fréchet-valued Brownian motion, $X(t)$ is the state process. The drift f , and the control u , depend at any time t on the past of the state process, $\{X(s), s \leq t\}$. For the case where the Fréchet space is R^n , a satisfactory theory dealing with the problem of existence of solutions of the differential equation and existence of optimal control laws is now available [1], [2]. A crucial building block in this theory consists in defining a solution of the differential equation via an absolutely continuous transformation of measures. Each control thereby defines a solution characterized by its (unique) probability law which is absolutely continuous with respect to Wiener measure. Thus the influence of a control law upon the system is captured in the Radon–Nikodym derivative of the resulting probability law with respect to Wiener measure. Questions dealing with the existence of an optimal control can then be converted into questions about the compactness (in an appropriate sense) of the set of Radon–Nikodym derivatives. The measure transformation technique mentioned above is due originally to Girsanov [16].

This paper deals with these same questions for the case where the state space is infinite-dimensional. The problem of characterizing Brownian motion in infinite-dimensional spaces is a difficult one and has been resolved for certain Fréchet spaces only. This is described in the next section, where some additional properties of such Brownian motion as sample continuity and stochastic integration are established also. In § 3 the result of Girsanov is extended to cover the differential equation under consideration. Once this has been achieved the techniques of [1], [2] apply without change and the existence of optimal controls and saddle points follows easily. This is sketched in § 4.

2. Preliminaries for defining Brownian motion. To define probability measures on Fréchet spaces the results of Dudley–Feldman–LeCam [3] are used. The

* Received by the editors September 25, 1973, and in revised form August 29, 1974.

† Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, Stony Brook, New York 11790. This research was supported in part by the National Science Foundation under Grants GK-24151 and GK-32136.

‡ Department of Electrical Engineering and Computer Science and Electronics Research Laboratory, University of California, Berkeley, California 94720. This research was supported in part by the National Science Foundation under Grant GK-10656X3.

latter are generalizations of the work of Gross [4]. While the generality of the immediately following discussion is not used subsequently, it does serve to indicate some directions along which the results reported here can be further pursued. Some notation and definitions are introduced first.

For a locally convex Hausdorff topological (real) vector space X let X^* denote its topological dual. A *dual system* or *duality* over the reals consists of two vector spaces X, Y and a bilinear form $\langle \cdot, \cdot \rangle: X \times Y \rightarrow R$ that separates points for both X and Y . For a pair X, Y in duality, let $FD(X)$ denote the collection of all finite-dimensional subspaces of X . For $G \subset X$, let $\mathcal{M}(Y, G)$ be the σ -algebra of Y generated by the sets $\{y: \langle x, y \rangle \in B\}$, where $x \in G, B$ is a Borel set in R . Thus $\mathcal{M}(Y, G)$ is the smallest σ -algebra on Y for which each $x \in G$, regarded as a function on Y , is measurable. Finally let $\mathcal{C}(Y, X) = \bigcup \{\mathcal{M}(Y, G) | G \in FD(X)\}$. A *cylinder set measure* on Y is any nonnegative, finitely additive set function m on $\mathcal{C}(Y, X)$ with $m(Y) = 1$ such that m is countably additive on $\mathcal{M}(Y, G)$ for each $G \in FD(X)$.

The following notion of a measurable seminorm given in [3] is crucial to the analysis.

DEFINITION 1. Given a duality $X, Y, \langle \cdot, \cdot \rangle$, and a cylinder set measure m on Y , a seminorm $|\cdot|$ on Y is said to be *m -measurable* if for each $\varepsilon > 0$ there is a $G \in FD(Y)$ such that if $F \in FD(X)$ and $F \perp G$ (i.e., $\langle x, y \rangle = 0$ for $x \in F, y \in G$), then

$$m\{y: |y - F^\perp| \leq \varepsilon\} \geq 1 - \varepsilon$$

or equivalently

$$m^*\{y: |y - G| \leq \varepsilon\} \geq 1 - \varepsilon,$$

where m^* is the outer measure on $\mathcal{C}(Y, X)$ induced by m .

The next result follows from [3, Thm. 2].

FUNDAMENTAL THEOREM. Let $|\cdot|$ be a Mackey-continuous seminorm on Y and suppose that it is m -measurable. Then the cylinder set measure induced by m on the Banach space $Y/|\cdot|$ obtained from Y via the seminorm $|\cdot|$ extends to a regular Borel measure.

From here on the discussion is specialized to a fixed, separable Hilbert space H . It is assumed that there is given for each $t \in R_+$ a cylinder set measure p_t on H such that p_t is a canonical normal distribution on H with variance parameter t (see [4] or [5]).¹ It is also assumed that there is given an increasing family of Mackey-continuous seminorms $|\cdot|_j, j = 1, 2, \dots$, on H . Let F be the Fréchet space obtained from H with respect to the topology defined by the seminorms $|\cdot|_j$ by completion modulo the intersection of their null spaces (which without loss of generality is assumed to equal $\{0\}$). As a corollary to the Fundamental Theorem it is proved in [3, Cor. 2.1] that each $p_t, t \in R_+$, extends to a regular Borel measure, denoted μ_t , on F . For future reference note that $H \subset F, F^* \subset H^*$

¹ This means that if $C \in \mathcal{C}(H, H)$ is of the form $C = P^{-1}(E)$, where P is the orthogonal projection of H onto a subspace $L \in FD(H)$ and E is a Borel subset of L , then

$$p_t(C) = (2\pi t)^{-n/2} \int_E \exp\left(-\frac{1}{2t}|x|_H^2\right) dx,$$

where n is the dimension of L , and $|\cdot|_H$ denotes the norm on H induced by its inner product.

$= H$, and define the maps $i: H \rightarrow F, j: F^* \rightarrow H^*$, as the canonical injections. Finally let $\mathcal{B}(F)$ denote the Borel sets of F . Throughout, H and F denote the spaces introduced here.

The collection $\mu_t, t \in R_+$, will be used now to define a Brownian motion with values in F . For each $t \in R_+$ let F^t be a copy of F . For each finite collection $t_1 < \dots < t_n$, let μ_{t_1, \dots, t_n} be the measure on $(\prod_{i=1}^n F^{t_i}, \prod_{i=1}^n \mathcal{B}(F^{t_i}))$ defined by

$$\int_{\Lambda_1} \dots \int_{\Lambda_n} d\mu_{t_1, \dots, t_n} = \int_{\Lambda_1} \int_{\Lambda_2 - x_1} \dots \int_{\Lambda_n - \sum_{i=1}^{n-1} x_i} d\mu_{t_n - t_{n-1}}(x_n) \dots d\mu_{t_2 - t_1}(x_2) dY_{t_1}(x_1)$$

for $\Lambda_i \in \mathcal{B}(F^{t_i}), i = 1, \dots, n$. Since sets of the form $\prod_{i=1}^n \Lambda_i$ generate the product σ -algebra $\prod_{i=1}^n \mathcal{B}(F^{t_i})$ the measure μ_{t_1, \dots, t_n} is defined. To show that the family of measures μ_{t_1, \dots, t_n} is a projective system, it is necessary to verify consistency. Since F is separable in the topology determined by the countable seminorms $|\cdot|_j$, the measure μ_{t_1, \dots, t_n} is determined by sets of the form $\Lambda_i = (I - P)F \oplus \Gamma_i$, where P is an orthogonal projection with finite dimensional range and Γ_i is a Borel subset of PF . Let $\Lambda_1, \dots, \Lambda_n$ be such sets, and suppose $\Lambda_j = F$. Then

$$\begin{aligned} \int_{\Lambda_1} \dots \int_{\Lambda_n} d\mu_{t_1, \dots, t_n} &= \int_{\Lambda_1} \int_{\Lambda_2 - x_1} \dots \int_{\Lambda_n - \sum_{i=1}^{n-1} x_i} d\mu_{t_n - t_{n-1}}(x_n) \dots d\mu_{t_1}(x_1) \\ &= \int_{\Gamma_1} \int_{\Gamma_2 - x_1} \dots \int_{\Gamma_n - \sum_{i=1}^{n-1} x_i} dp_{t_n - t_{n-1}}(x_n) \dots dp_{t_1}(x_1) \\ &= \int_{\Gamma_1} \dots \int_{\Gamma_{j-1} - \sum_{i=1}^{j-2} x_i} \int_{\Gamma_{j+1} - \sum_{i=1}^j x_i} \dots \int_{\Gamma_n - \sum_{i=1}^{n-1} x_i} dp_{t_n - t_{n-1}}(x_n) \\ &\quad \dots dp_{t_{j+2} - t_{j+1}}(x_{j+1}) dp_{t_{j+1} - t_{j-1}}(x_j) \dots dp_{t_1}(x_1) \\ &= \int_{\Lambda_1} \dots \int_{\Lambda_{j-1}} \int_{\Lambda_{j+1}} \dots \int_{\Lambda_n} d\mu_{t_1, \dots, t_{j-1}, t_{j+1}, \dots, t_n}. \end{aligned}$$

In the above p is the measure of a Brownian motion with values in the finite-dimensional space PF and so the third equality above follows from the Markov property of such a process. Since each of the measures $\mu_t, t \in R_+$, is a regular Borel measure, the projective system of measures admits a projective limit [6, p. 49]. The projective limit thus obtained is denoted (Ω, \mathcal{F}, P) . Evidently, $(\Omega, \mathcal{F}) = \prod_{t \in R_+} (F^t, \mathcal{B}(F^t))$. It will be assumed that (Ω, \mathcal{F}, P) is complete. For each t let X_t denote the F -valued evaluation map on Ω at t . Let \mathcal{F}_t be the smallest completed σ -algebra with respect to which X_s is measurable for $s \leq t$. Then $(X_t, \mathcal{F}_t, P)_{t \in R_+}$ is a Brownian motion as defined below.

DEFINITION 2. A stochastic process $(X_t, \mathcal{F}_t, P)_{t \in R_+}$ (or simply $(X_t, P)_{t \in R_+}$ or (X_t) if there is no ambiguity) is said to be a *Brownian motion with values in the Fréchet space F* induced from a family of canonical normal distributions on a Hilbert space H , if for each $l \in F^*$ (the topological dual of F) the real-valued process $(\langle l, X_t \rangle, \mathcal{F}_t, P)_{t \in R_+}$ is a real-valued Brownian motion with $E\langle l, X_t \rangle^2 = t \|j\|_H^2$. (Here and throughout $|\cdot|_H$ denotes the norm on H and $j: F^* \rightarrow H^*$ is the canonical injection mentioned before.)

The process $(Y_t, P)_{t \in R_+}$ is said to be a *modification* of the process $(X_t, P)_{t \in R_+}$ if for each $t \in R_+$ $X_t(\omega) = Y_t(\omega)$ a.s. (the null set $\{X_t \neq Y_t\}$ may depend on t).

For Fréchet-valued Brownian motion there is a modification which has continuous sample paths as shown by the following lemma.

LEMMA 1. Let $(X_t, \mathcal{F}_t, P)_{t \in \mathbb{R}_+}$ be a Fréchet-valued Brownian motion. There is a modification of X_t with continuous sample paths.

Proof. Since a countable number of seminorms determine the topology of a Fréchet space it suffices to verify the continuity of the sample function of Brownian motion with respect to each of these countable number of seminorms and therefore it is enough to prove the lemma for a Banach-valued Brownian motion.

Fernique [7] has shown that for a Gaussian random variable X on a topological vector space with a measurable seminorm $|\cdot|$ there is an $\alpha > 0$ such that

$$(1) \quad E \exp \alpha |X|^2 < \infty.$$

Combining (1) with a result of Nelson [8, Thm. 2] as used by Gross [9, p. 134] it follows that the Banach valued Brownian motion has a modification with continuous sample paths. \square

Since stochastic integrals will be used subsequently, a family of processes has to be described that can serve as integrands. The following definition gives such a family.

DEFINITION 3. Let \tilde{H} be a separable Hilbert space. An \tilde{H} -valued stochastic process $(\psi_t)_{t \in \mathbb{R}_+}$ on (Ω, \mathcal{F}, P) that is adapted to $(\mathcal{F}_t)_{t \in \mathbb{R}_+}$ is said to be *predictable* if the map $\psi_t(\omega): \mathbb{R}_+ \times \Omega \rightarrow \tilde{H}$ is measurable with respect to the σ -algebra on $\mathbb{R}_+ \times \Omega$ generated by the left-continuous \tilde{H} -valued processes adapted to $(\mathcal{F}_t)_{t \in \mathbb{R}_+}$.

Real-valued stochastic integrals will now be defined from F -valued Brownian motion and predictable H -valued processes.

LEMMA 2. Let $(\psi_t)_{t \in \mathbb{R}_+}$ be a predictable H -valued process with

$$(2) \quad E \int_0^\infty |\psi_t|_H^2 dt < \infty,$$

and let $(B_t, \mathcal{F}_t, P)_{t \in \mathbb{R}_+}$ be an F -valued Brownian motion. Then the real-valued process $(Y_t, \mathcal{F}_t, P)_{t \in \mathbb{R}_+}$,

$$(3) \quad Y_t = \int_0^t \langle \psi_s, dB_s \rangle,$$

is a square integrable martingale which has a modification with continuous sample paths. (The integral (3) is defined in the course of the proof.)

Proof. Let $\varepsilon_n > 0$ be a sequence decreasing to zero and for each n let P_{ε_n} be a projection on H with finite-dimensional range $P_{\varepsilon_n}H \subset jF^*$ such that

$$E \int_0^\infty |(I - P_{\varepsilon_n})\psi_t|_H^2 dt < \varepsilon_n.$$

Let l_1, \dots, l_k be an orthonormal basis for $P_{\varepsilon_n}H$. The process $Y_t(P_{\varepsilon_n})$ defined by

$$(4) \quad \begin{aligned} Y_t(P_{\varepsilon_n}) &= \int_0^t \langle P_{\varepsilon_n}\psi_s, dB_s \rangle \\ &= \int_0^t \sum_{i=1}^k \langle l_i, \psi_s \rangle d\langle l_i, B_s \rangle \end{aligned}$$

is a (continuous) martingale since by the definition of the process (B_t) , $(\langle l_i, B_t \rangle)$ is a real-valued Brownian motion. By Doob's inequality [10, p. 353] the sequence of continuous martingales $(Y_t(P_{\varepsilon_n}))$ converges uniformly on compact subsets of R_+ because

$$E|Y_t(P_{\varepsilon_n}) - Y_t(P_{\varepsilon_m})|^2 = E \int_0^t |(P_{\varepsilon_n} - P_{\varepsilon_m})\psi_s|^2 ds$$

which vanishes as m, n increase to infinity. The integral (3) is defined as the limit of the martingales $Y_t(P_{\varepsilon_n})$. Clearly the limit does not depend on the particular choice of the projections P_{ε_n} and Y_t is evidently square integrable. \square

COROLLARY 1. Let $(\psi_t)_{t \in R_+}$ be a predictable H -valued process with

$$(5) \quad \int_0^t |\psi_s|_H^2 ds < \infty \quad \text{a.s. for all } t.$$

Then the real-valued process Z_t defined by

$$(6) \quad Z_t = \int_0^t \langle \psi_s, dB_s \rangle$$

is a locally square integrable martingale which has a modification with continuous sample paths.²

The following representation of square integrable functionals on the F -valued Brownian motion probability space will be useful subsequently. For R^n -valued Brownian motion, K. Itô [11] has obtained this representation by describing results of Wiener [12] and Cameron–Martin [13] in terms of stochastic integrals.

PROPOSITION 1. Let (Ω, \mathcal{F}, P) be the probability space for an F -valued Brownian motion $(B_t, \mathcal{F}_t)_{t \in R_+}$. Let f be a real-valued square integrable functional on (Ω, \mathcal{F}, P) . Then f can be represented as

$$(7) \quad f = c + \int_0^\infty \langle \psi_s, dB_s \rangle,$$

where $c = Ef$ and $(\psi_t)_{t \in R_+}$ is a predictable H -valued process with

$$E \int_0^\infty |\psi_t|_H^2 dt < \infty.$$

Proof. Since H is separable so is F , hence by the Hahn–Banach theorem, there is a countable family $\Gamma \subset F^*$ that separates points of F . Consider the random variables $\langle \gamma, \int_0^t dB_s \rangle$, where $\gamma \in \Gamma$, $t \in R_+$ and let \mathcal{A} be the algebra of real-valued random variables formed from these and the constant random variables. Since the random variables $\langle \gamma, \int_0^t dB_s \rangle$ are jointly Gaussian, it follows that $\mathcal{A} \subset L^2(P)$ and furthermore

$$\exp \left(\sum_{i=1}^n a_i |f_i| \right) \in L^2(P)$$

² m_t is said to be a locally square integrable martingale if there exists a sequence of stopping times $T_n \uparrow \infty$ a.s. such that $m_{t \wedge T_n}$ is a square integrable martingale for each n .

for $a_i \in R$ and $f_i \in \mathcal{A}$. Since F is separable, the family of random variables $\langle \gamma, \int_0^t dB_s \rangle$, $\gamma \in \Gamma$, $t \in R_+$, generates the σ -algebra \mathcal{F} . By a result of Segal [14, Lemma 2.1] it follows that \mathcal{A} is dense in $L^2(P)$.

Let $f \in L^2(P)$. There exists then a sequence g_n in \mathcal{A} such that $E|f - g_n|^2 \rightarrow 0$. By properties of finite-dimensional Brownian motion, g_n has a representation of the form (7) (see [15]), i.e.,

$$(8) \quad g_n = c_n + \int_0^\infty \langle \psi_s^n, dB_s \rangle,$$

where $c_n = Eg_n$ and (ψ_t^n) is a predictable process with values in some finite-dimensional subspace L of H and with $L \subset jF^*$. Since $L^2(P)$ -convergence implies $L^1(P)$ -convergence it follows that c_n converges to c . Furthermore the stochastic integrals in (8) must be Cauchy; hence

$$\int_0^\infty E|\psi_t^n - \psi_t^m|_H^2 dt \rightarrow 0$$

as $m, n \rightarrow \infty$. Since the sequence of processes (ψ_t^m) are predictable, there must exist a predictable process (ψ_t) such that

$$\int_0^\infty |\psi_t^n - \psi_t|_H^2 dt \rightarrow 0,$$

and evidently (7) is satisfied. \square

3. Transformation of measures. Theorem 1 below describes how Fréchet-valued Brownian motion is transformed by changing the probability measure by an absolutely continuous substitution of the measure. This result was first established by Girsanov [16] for the case of R^n -valued Brownian motion. The result has been presented in [17] and a related result is given in [18].

THEOREM 1. *Let $(B_t, \mathcal{F}_t, P)_{t \in [0,1]}$ be an F -valued Brownian motion and let $(\psi_t)_{t \in [0,1]}$ be a predictable H -valued process such that*

$$(9) \quad \int_0^1 |\psi_t|_H^2 dt < \infty \quad \text{a.s. } P.$$

Define the nonnegative process $(M_t, \mathcal{F}_t, P)_{t \in [0,1]}$ by

$$(10) \quad M_t = \exp \left[\int_0^t \langle \psi_s, dB_s \rangle - \frac{1}{2} \int_0^t |\psi_s|_H^2 ds \right].$$

Then

$$(11) \quad E(M_1) \leq 1.$$

Suppose that

$$(12) \quad E(M_1) = 1;$$

then the process $(\tilde{B}_t, \mathcal{F}_t, \tilde{P})_{t \in [0,1]}$ is an F -valued Brownian motion where

$$\tilde{B}_t = B_t - \int_0^t i\psi_s ds,$$

and the probability measure \tilde{P} is given by

$$\frac{d\tilde{P}}{dP} = M_1.$$

Proof. Define the increasing sequence of stopping times T_n by

$$T_n = \begin{cases} \inf \{t | M_t > n\}, \\ 1 & \text{if the set above is empty.} \end{cases}$$

Because of (9) M_t has (a modification with) continuous sample paths, so $T_n \uparrow 1$ a.s. P , and in particular,

$$(13) \quad \lim_{n \rightarrow \infty} M_{t \wedge T_n} = M_t \quad \text{a.s. } P.$$

By the main result in [19], the processes (M_t) and $(M_{t \wedge T_n})$ satisfy

$$(14) \quad M_t = 1 + \int_0^t M_s \langle \psi_s, dB_s \rangle$$

$$(15) \quad M_{t \wedge T_n} = 1 + \int_0^{t \wedge T_n} M_s \langle \psi_s, dB_s \rangle.$$

Since $(M_{t \wedge T_n})$ is bounded, (15) implies that it is a martingale and hence

$$E(M_{t \wedge T_n}) = E(M_{T_n}) = 1.$$

An application of Fatou's lemma to this result and (13) yields (11).

From now on suppose that (12) holds, so that from (14) it follows that $(M_t, \mathcal{F}_t, P)_{t \in [0, 1]}$ is in fact a martingale. Let $l \in H$ be fixed and consider the process $(N_t, \mathcal{F}_t, \tilde{P})_{t \in [0, 1]}$, where

$$N_t = \langle l, \tilde{B}_t \rangle = \langle l, B_t \rangle - \int_0^t \langle l, \psi_s \rangle ds.$$

The theorem will be proved once it is shown that

$$(16) \quad (N_t, \mathcal{F}_t, \tilde{P}) \quad \text{is a martingale,}$$

$$(17) \quad \tilde{E}(N_t^2 - N_s^2 | \mathcal{F}_s) = \|l\|_H^2(t - s),$$

where \tilde{E} denotes expectation with respect to the measure \tilde{P} . Now to prove (16) it is sufficient to show instead that

$$(18) \quad (M_t N_t, \mathcal{F}_t, P) \quad \text{is a martingale.}$$

Because suppose that (18) is true. Then using the fact that (M_t, P) is a martingale,

$$\begin{aligned} \tilde{E}(N_t | \mathcal{F}_s) &= \frac{E(M_t N_t | \mathcal{F}_s)}{E(M_t | \mathcal{F}_s)} \quad (\text{by [20, p. 345]}) \\ &= \frac{M_s N_s}{M_s} \quad (\text{by (18)}) \\ &= N_s, \end{aligned}$$

which is equivalent to (16). Now

$$M_t N_t = M_t \langle l, B_t \rangle - M_t \int_0^t \langle l, \psi_s \rangle ds.$$

Applying the differentiation formula for continuous martingales [15] gives

$$\begin{aligned} M_t N_t - M_0 N_0 &= \int_0^t M_s \langle l, dB_s \rangle - \int_0^t M_s \langle l, \psi_s \rangle ds \\ &\quad + \int_0^t N_s dM_s + \int_0^t M_s \langle l, \psi_s \rangle ds \\ &= \int_0^t M_s \langle l, dB_s \rangle + \int_0^t N_s dM_s, \end{aligned}$$

which clearly implies (18).

It remains to prove (17). To this end note that

$$(19) \quad \tilde{E}(N_t^2 - N_s^2 | \mathcal{F}_s) = \frac{E(M_t(N_t^2 - N_s^2) | \mathcal{F}_s)}{M_s},$$

and apply the differentiation formula to $M_t N_t^2$ to obtain

$$\begin{aligned} M_t N_t^2 - M_0 N_0^2 &= \int_0^t N_s^2 dM_s + 2 \int_0^t M_s N_s \langle l, dB_s \rangle - 2 \int_0^t M_s N_s \langle l, \psi_s \rangle ds \\ &\quad + 2 \int_0^t N_s M_s \langle l, \psi_s \rangle ds + \int_0^t M_s \|l\|_H^2 ds. \end{aligned}$$

Substitution of this into (19) gives

$$\tilde{E}(N_t^2 - N_s^2 | \mathcal{F}_s) = \frac{E(\int_s^t M_\tau \|l\|_H^2 d\tau | \mathcal{F}_s)}{M_s} = \|l\|_H^2 (t - s),$$

and so (17) is proved. \square

In many applications of the transformation of measures technique of Theorem 1 the crucial difficulty is to verify that $\tilde{P}(\Omega) = 1$. The next result gives a sufficient condition for $\tilde{P}(\Omega) = 1$. It is due to V. Beneš.

LEMMA 3 (Beneš). *Let $(B_t, \mathcal{F}_t, P)_{t \in [0,1]}$ be an F -valued Brownian motion. Let $(\psi_t)_{t \in [0,1]}$ be a predictable H -valued process such that*

$$(20) \quad |\psi(t, B)|_H \leq K + K' \sup_{s \in [0,t]} |B_s|,$$

where $|\cdot|$ is a seminorm on F and K, K' are constants. Then there is $\alpha > 1$ and $M < \infty$, depending only on K and K' , such that

$$(21) \quad E \exp \alpha \left[\int_0^1 \langle \psi_s, dB_s \rangle - \frac{1}{2} \int_0^1 |\psi_s|_H^2 ds \right] < M.$$

In particular,

$$E \exp \left[\int_0^1 \langle \psi_s, dB_s \rangle - \frac{1}{2} \int_0^1 |\psi_s|_H^2 ds \right] = 1.$$

Proof. For any (ψ_t) denote

$$\zeta(\psi) = \int_0^1 \langle \psi_s, dB_s \rangle - \frac{1}{2} \int_0^1 |\psi_s|_H^2 ds.$$

Fix (ψ_t) satisfying (20) and let

$$T_n = \inf \left\{ t \left| \int_0^t |\psi_s|_H^2 ds > n \right. \right\}.$$

Evidently $T_n \uparrow 1$ a.s. and let $\psi_t^n = \psi_{t \wedge T_n}$. Let $\alpha > 1$. Let $\tilde{B}_t^n = B_t - \alpha \int_0^t i\psi_s^n ds$. Since $\int_0^1 |\psi_t^n|_H^2 dt \leq n$, it is easy to use the proof of [16, Lemma 1] to show that $E \exp \zeta(\alpha\psi^n) = 1$. Hence by Theorem 1, $(\tilde{B}_t^n, \mathcal{F}_t, \tilde{P}^n)_{t \in [0,1]}$ is a Brownian motion, where $d\tilde{P}^n = [\exp \zeta(\alpha\psi^n)] dP$. Now

$$|\psi_t^n|_H \leq K + K' \sup_{s \in [0,t]} |B_s|$$

so that

$$\begin{aligned} |B_t| &\leq |\tilde{B}_t^n| + \alpha \int_0^t \left[K + K' \sup_{u \in [0,s]} |B_u| \right] ds, \\ \sup_{s \in [0,t]} |B_s| &\leq \sup_{s \in [0,t]} |\tilde{B}_s^n| + \alpha Kt + \alpha K' \int_0^t \sup_{u \in [0,s]} |B_u| ds, \end{aligned}$$

and hence by Gronwall's inequality,

$$\|B\| \leq (\alpha K + \|\tilde{B}^n\|) \exp \alpha K',$$

where $\|X\| = \sup_{s \in [0,1]} |X_s|$. From (20), $|\psi_s^n|^2 \leq 2K^2 + 2K'\|B\|^2$, so that

$$\int_0^1 |\psi_s^n|_H^2 ds \leq 2K^2 + 4K'(\alpha^2 K^2 + \|\tilde{B}^n\|^2) \exp 2\alpha K'.$$

By the differentiation formula,

$$\begin{aligned} E \exp \alpha \zeta(\psi^n) &= E \exp \left[\zeta(\alpha\psi^n) + \frac{\alpha^2 - \alpha}{2} \int_0^1 |\psi_s|_H^2 ds \right] \\ &\leq E \exp \zeta(\alpha\psi^n) \exp \left(\frac{\alpha^2 - \alpha}{2} \right) [2K^2 + 4K'(\alpha^2 K^2 + \|\tilde{B}^n\|^2) \exp 2\alpha K'] \\ &= NE^n \exp \left[\left(\frac{\alpha^2 - \alpha}{2} \right) (4K' \exp 2\alpha K') \|\tilde{B}^n\|^2 \right], \end{aligned}$$

where N is a constant and E^n denotes expectation with respect to \tilde{P}^n . Since $(\tilde{B}^n, \tilde{P}^n)$ is a Brownian motion by a result of Fernique [7] there exists $\gamma > 0$ such that

$$E^n \exp \gamma \|\tilde{B}^n\|^2 = A < \infty$$

independent of n . Let $\alpha > 1$ be so small that $((\alpha^2 - \alpha)/2)(4K' \exp 2\alpha K') < \gamma$. Then

$$E \exp \alpha \zeta(\psi^n) < NA,$$

and so (21) follows by Fatou's lemma. The final assertion is then immediate because (21) implies that $\exp [\int_0^t \langle \psi_s, dB_s \rangle - \frac{1}{2} \int_0^t |\psi_s|_H^2 ds]$ is a martingale. \square

4. Preliminaries for optimization. The system to be controlled is represented by the stochastic differential equation

$$(22) \quad dX(t) = if(t, X, u(t, X)) dt + dB(t), \quad t \in [0, 1],$$

where B_t is an F -valued Brownian motion, $X_t \in F$ is the state with $X_0 = 0$ a.s., and u is the control law taking values in a prespecified compact subset $U \subset F$ called the *control set*. The function f takes values in H and $i: H \rightarrow F$ is the canonical injection. The first difficulty to be resolved is to define the solution of the differential equation (22) for a large class of control laws. This is achieved in the following manner. One starts with a process $(X_t, \mathcal{F}_t, P_0)_{t \in [0, 1]}$ which is an F -valued Brownian motion. For a given control law u an F -valued process B_t^u is defined by

$$B_t^u = X_t - \int_0^t if(s, X, u(s, X)) ds.$$

Next the probability measure P_0 is replaced by another probability measure P^u such that the process $(B_t^u, \mathcal{F}_t, P^u)_{t \in [0, 1]}$ is an F -valued Brownian motion. The process $(X_t, \mathcal{F}_t, P^u)_{t \in [0, 1]}$ is then regarded as the solution of (22) corresponding to the control law u . To make this procedure precise the following notations and definitions are useful.

DEFINITION 4. (a) \mathcal{C} is the linear space of all F -valued continuous functions, denoted by z , on $[0, 1]$.

(b) For $t \in [0, 1]$, \mathcal{S}_t is the smallest σ -algebra of subsets of \mathcal{C} which contain all sets of the form $\{z \in \mathcal{C} | z(s) \in A\}$, where $s \in [0, t]$ and A is a (topological) Borel subset of F . $\mathcal{S} = \mathcal{S}_1$.

Throughout the remainder of this paper Ω is a fixed probability space with an increasing family of σ -algebras \mathcal{F}_t , $t \in [0, 1]$. $\mathcal{F} = \mathcal{F}_1$. It will be necessary to consider different probability measures on the space (Ω, \mathcal{F}) . If Y_t is a family of measurable functions on (Ω, \mathcal{F}_t) and if P is a probability measure on (Ω, \mathcal{F}) , the stochastic process corresponding to P and the family Y_t will be denoted by $(Y_t, \mathcal{F}_t, P)_{t \in [0, 1]}$. Then the same family Y_t generates different stochastic processes corresponding to different probability measures. Finally let P_0 be a distinguished probability measure, and let X_t be a distinguished family of F -valued measurable functions on (Ω, \mathcal{F}_t) , $t \in [0, 1]$, such that the process $(X_t, \mathcal{F}_t, P_0)_{t \in [0, 1]}$ is an F -valued Brownian motion with continuous sample paths. Unless mentioned otherwise the process X_t refers to this process. Also E_0 will denote expectation with respect to P_0 . The measure induced on $(\mathcal{C}, \mathcal{S})$ by the process X_t is denoted by μ and will be called the *Wiener measure* on $(\mathcal{C}, \mathcal{S})$.

The following conditions are imposed on the function f in (22).

f1. f is a map from $[0, 1] \times \mathcal{C} \times U$ into H and f is measurable with respect to the product σ -algebra $\mathcal{B} \otimes \mathcal{S} \otimes \mathcal{B}_u$, where $\mathcal{B}(\mathcal{B}_u)$ is the family of Borel subsets of $[0, 1](U)$.

f2. For $t \in [0, 1]$, $f(t, \cdot, \cdot)$ is $\mathcal{S}_t \otimes \mathcal{B}_u$ -measurable.

f3. For $(t, z) \in [0, 1] \times \mathcal{C}$, $f(t, z, \cdot): U \rightarrow H$ is continuous.

f4. There is an increasing function $f_0: R_+ \rightarrow R_+$, and a seminorm $|\cdot|$ on F

such that for $(t, z, u) \in [0, 1] \times \mathcal{C} \times U$,

$$|f(t, z, u)|_H \leq f_0(\|z\|),$$

where $\|z\| = \max_{t \in [0, 1]} |z(t)|$. Throughout the remainder the symbols $|\cdot|$ and $\|\cdot\|$ will denote the seminorms assumed here.

f5. For $(t, z) \in [0, 1] \times \mathcal{C}$, $f(t, z, U) = \{f(t, z, u) | u \in U\}$ is a closed and convex subset of H .

DEFINITION 5. (a) An *admissible control* (law) is a map $u: [0, 1] \times \mathcal{C} \rightarrow U$ which is $\mathcal{B} \otimes \mathcal{S}$ -measurable and, further, is such that for each $t \in [0, 1]$, $u(t, \cdot)$ is \mathcal{S}_t -measurable. \mathcal{U} denotes the set of all admissible controls.

(b) The *drift* corresponding to $u \in \mathcal{U}$ is the function g_u given by g_u given by $g_u(t, z) = f(t, z, u(t, z))$. $\mathcal{Y} = \{g_u | u \in U\}$.

(c) For $g \in \mathcal{Y}$ and positive integer n , g^n is the function given by

$$g^n(t, z) = \begin{cases} g(t, z) & \text{if } |z(s)| \leq n \text{ for } s \in [0, t], \\ 0 & \text{otherwise.} \end{cases}$$

DEFINITION 6. A function $\phi: [0, 1] \times \mathcal{C} \rightarrow H$ is *causal* if it is $\mathcal{B} \otimes \mathcal{S}$ -measurable and if $\phi(t, \cdot)$ is \mathcal{S}_t -measurable for $t \in [0, 1]$.

DEFINITION 7. Φ is the collection of all causal functions such that $|\phi(t, z)|_H \leq f_0(\|z\|)$ for $(t, z) \in [0, 1] \times \mathcal{C}$. $\Phi^n = \{\phi \in \Phi | |\phi(t, z)|_H \leq n \text{ for all } (t, z)\}$.

The next result which follows immediately from [21, Lemma 1] gives a very useful characterization of \mathcal{Y} .

LEMMA 4. A causal function g is in \mathcal{Y} if and only if $g(t, z) \in f(t, z, U)$ for all t, z .

DEFINITION 8. Let ϕ be a causal function such that

$$(23) \quad \int_0^1 |\phi(t, z)|_H^2 dt < \infty \quad \text{for all } z \in \mathcal{C}.$$

Then $(\zeta_t(\phi), \mathcal{F}_t, P_0)_{t \in [0, 1]}$ denotes the continuous real-valued process defined by

$$\zeta_t(\phi) = \int_0^t \langle \phi(s, X), dX_s \rangle - \frac{1}{2} \int_0^t |\phi(s, X)|_H^2 ds.$$

Let $\zeta(\phi) = \zeta_1(\phi)$. Note that (23) is always satisfied for $\phi \in \Phi$.

The problem of the existence of solutions of (22) is resolved in the following result which follows immediately from Theorem 1.

THEOREM 2. Let $u \in \mathcal{U}$ be such that

$$(24) \quad E_0 \exp \zeta(g_u) = 1,$$

where E_0 denotes expectation with respect to P_0 . Define the probability measure P_u by

$$dP_u = \exp \zeta(g_u) dP_0.$$

Then the process $(B_t, \mathcal{F}_t, P_u)_{t \in [0, 1]}$ defined by

$$B_t = X_t - \int_0^t if(s, X, u(s, X)) ds$$

is an F -valued Brownian motion.

The next result shows that (24) must be satisfied by every solution of (22).

LEMMA 5. Let $\phi \in \Phi$. Let $(Y_t, \mathcal{F}_t, P)_{t \in [0,1]}$ be any process with continuous sample paths such that the stochastic process $(B_t, \mathcal{F}_t, P)_{t \in [0,1]}$ defined by

$$dB_t = dY_t - i\phi(t, Y) dt$$

is an F -valued Brownian motion. Then

$$\int_{\Omega} \exp \left[- \int_0^1 \langle \phi(s, Y), dB_s \rangle - \frac{1}{2} \int_0^1 |\phi(s, Y)|_H^2 ds \right] dP = 1.$$

Proof. Define the function $\phi_n : [0, 1] \times \mathcal{C} \rightarrow H$ by

$$\phi_n(t, z) = \begin{cases} \phi(t, z) & \text{if } |z(s)| \leq n \text{ for } s \in [0, t], \\ 0 & \text{otherwise.} \end{cases}$$

From the definition of Φ it follows that

$$|\phi_n(t, Y(\omega))|_H \leq f_0(n) \quad \text{for } t \in [0, 1], \quad \omega \in \Omega.$$

By Lemma 3,

$$\int_{\Omega} \exp \zeta(-\phi_n) dP = 1,$$

where

$$\zeta(-\phi_n) = - \int_0^1 \langle \phi_n(s, Y), dB_s \rangle - \frac{1}{2} \int_0^1 |\phi_n(s, Y)|_H^2 ds.$$

By Theorem 1 the process $(Y_n(t), \mathcal{F}_t, \tilde{P}_n)_{t \in [0,1]}$, where

$$Y_n(t) = \int_0^t i\phi_n(s, Y) ds + B(t),$$

$$d\tilde{P}_n = \exp \zeta(-\phi_n) dP,$$

is a Brownian motion with continuous sample paths and hence induces Wiener measure μ on $(\mathcal{C}, \mathcal{S})$. Let $\varepsilon > 0$, and n be so large that

$$\tilde{P}_n(\Omega_n) = \tilde{P}_n(\|Y_n\| < n) = \mu(z \in \mathcal{C} \mid \|z\| < n) \geq 1 - \varepsilon.$$

Now it is clear that $\|Y_n(\omega)\| < n$ only if $\|Y(\omega)\| < n$ and hence

$$Y_n(\omega, t) = Y(\omega, t), \quad \phi_n(\omega, t) = \phi(\omega, t) \quad \text{for } \omega \in \Omega_n, \quad t \in [0, 1],$$

so that

$$\tilde{P}_n(\Omega_n) = \int_{\Omega_n} \exp \zeta(-\phi_n) dP = \int_{\Omega_n} \exp \zeta(-\phi) dP \geq 1 - \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, the result follows. \square

COROLLARY 2. Let $\phi \in \Phi$ and $(Y_t, \mathcal{F}_t, P)_{t \in [0,1]}$ satisfy the hypothesis of Lemma 5. Let ν be the measure induced by the process (Y_t, \mathcal{F}_t, P) on the measurable space

$(\mathcal{C}, \mathcal{S})$. Then ν is mutually absolutely continuous with respect to μ and

$$\frac{d\mu}{d\nu}(Y(\omega)) = \exp \zeta(-\phi)(\omega).$$

Remark. This corollary implies that the solutions of (22) are unique in a weak sense, i.e., all solutions of (22) which have continuous sample paths must induce the same measure on $(\mathcal{C}, \mathcal{S})$. The qualification “weak” is inserted because only the uniqueness of the probability law has been proved.

Recall that $(X_t, \mathcal{F}_t, P_0)_{t \in [0,1]}$ is an F -valued Brownian motion with continuous sample paths. Also recall Definition 8.

DEFINITION 9. For any subset $\Lambda \subset \Phi$ define $\mathcal{D}(\Lambda) \subset L^1(\Omega, \mathcal{F}, P_0)$ by

$$\mathcal{D}(\Lambda) = \{\exp \zeta(\phi) | \phi \in \Lambda\}.$$

The corollary to Lemma 3 implies the next assertion.

PROPOSITION 2. $\mathcal{D}(\Phi^n)$ is a bounded subset of $L^2(\Omega, \mathcal{F}, P_0)$.

LEMMA 6. $\mathcal{D}(\Phi^n)$ is a closed subset of $L^2(\Omega, \mathcal{F}, P_0)$.

Proof. Let ϕ_1, ϕ_2, \dots be a sequence from Φ^n and let ρ be such that

$$(25) \quad \lim_{n \rightarrow \infty} E_0 |\rho - \exp \zeta(\phi_n)|^2 = 0,$$

and

$$(26) \quad \lim_{n \rightarrow \infty} \exp \zeta(\phi_n) = \rho \quad \text{a.s. } P_0.$$

Since $E_0 \exp \zeta(\phi_n) = 1$ for all n , $E_0 \rho = 1$, and so by Proposition 1 there is a functional ψ such that $E_0 \int_0^1 |\psi(t, x)|_H^2 dt < \infty$ and

$$(27) \quad \rho = 1 + \int_0^1 \langle \psi(t, X), dX_s \rangle \quad \text{a.s. } P_0.$$

Define the martingale ρ_t by

$$\rho_t = E_0\{\rho | \mathcal{F}_t\}, \quad t \in [0, 1].$$

By taking modifications if necessary it can be assumed that the martingales ρ_t and $\zeta^t(\phi_n)$ have continuous sample paths so that, by Doob's inequality [10, p. 353], it follows from (25) that

$$(28) \quad \rho_t = \lim_{n \rightarrow \infty} \exp \zeta^t(\phi_n) \quad \text{uniformly on } [0, 1] \text{ a.s. } P_0.$$

Next

$$\exp \zeta(\phi_n) = 1 + \int_0^1 \exp \zeta^t(\phi_n) \langle \phi_n(t), dX(t) \rangle$$

so that from (25), (26),

$$\lim_{n \rightarrow \infty} E_0 \int_0^1 |\exp \zeta^t(\phi_n) \phi_n(t) - \psi(t)|_H^2 dt = 0,$$

and hence by taking subsequences if necessary it can be assumed that

$$(29) \quad \psi(t) = \lim_{n \rightarrow \infty} \exp \zeta^t(\phi_n) \phi_n(t) \quad \text{a.s. } l \otimes P_0,$$

where l denotes Lebesgue measure on $[0, 1]$. Now $\rho > 0$ a.s. P_0 , because if $P_0(A) = P_0\{\rho = 0\} > 0$, then (26), together with the fact that $|\phi_n|_H \leq n$, implies that

$$\lim_{n \rightarrow \infty} \int_0^1 \langle \phi_n(t), dX(t) \rangle = -\infty \quad \text{on } A.$$

But then

$$\infty = E_0 \int_0^1 \left| \langle \phi_n(t), dX(t) \rangle \right|^2 \leq E_0 \int_0^1 |\phi_n(t)|_H^2 dt \leq N^2,$$

so that to avoid the contradiction one must have $\rho > 0$ a.s. P_0 . It follows that $\rho_t > 0$ a.s., and hence combining (28) and (29) gives

$$\frac{\psi(t)}{\rho(t)} = \lim_{n \rightarrow \infty} \phi_n(t) \quad \text{a.s. } l \otimes P_0.$$

Thus there is a causal map $\phi \in \Phi^n$ such that

$$\phi(t, X) = \lim_{n \rightarrow \infty} \phi_n(t, X) \quad \text{a.s. } l \otimes P_0$$

and evidently $\rho = \exp \zeta(\phi)$. \square

The proofs of the next two results are identical respectively with the proofs of [2, Lemma 4] and [2, Thm. 2] with some obvious notational changes. Hence the proofs are omitted.

LEMMA 7. $\mathcal{D}(\Phi^n)$ is a convex subset of $L^2(\Omega, \mathcal{F}, P_0)$.

THEOREM 3. Let

$$\mathcal{Y}^0 = \{g \in \mathcal{Y} | E_0 \exp \zeta(g) = 1\}.$$

Then $\mathcal{D}(\mathcal{Y}^0)$ is a closed, convex subset of $L^1(\Omega, \mathcal{F}, P_0)$.

5. Applications. The results developed above immediately imply the existence of optimal control laws for a broad class of problems. Consider the control system

$$dX(t) = if(t, X, u(t, X)) dt + dB(t), \quad t \in [0, 1],$$

with $X(0) = 0$ a.s. Suppose that f satisfies the assumptions f1 to f5 and in addition the function f_0 in f4 satisfies assumption f6.

f6. There exists K, K' such that $f_0(n) \leq K + K'n$ for all $n \in R_+$.

Let $L: \mathcal{C} \rightarrow R$ be a fixed bounded \mathcal{S} -measurable function. For each $u \in \mathcal{U}$ the cost incurred by u is defined to be

$$(30) \quad J(u) = E_0[(\exp \zeta(g_u))L(X)] = \int_{\Omega} L(X(\omega)) \exp(g_u(\omega)) dP_0.$$

THEOREM 4. Under assumptions f1 to f6, there exists an optimal control $u^* \in \mathcal{U}$, i.e.,

$$J(u^*) \leq J(u), \quad u \in \mathcal{U}.$$

Proof. From Lemma 3 it follows that $E_0 \exp \zeta(g_u) = 1$ for all $u \in \mathcal{U}$ and furthermore there exists $\alpha > 1$ such that

$$\sup_{u \in \mathcal{U}} E_0 \exp \alpha \zeta(g_u) < \infty.$$

Hence by [6, Chap. 2], $\mathcal{D}(\mathcal{Y})$ is a uniformly integrable subset of $L^1(\Omega, \mathcal{F}, P_0)$. By Theorem 3, $\mathcal{D}(\mathcal{Y})$ is convex and strongly closed in $L^1(\Omega, \mathcal{F}, P_0)$. So that by [6, Chap. 2] $\mathcal{D}(\mathcal{Y})$ is weakly compact in $L^1(\Omega, \mathcal{F}, P_0)$. Hence the linear functional $\exp \zeta(g_u) \mapsto E_0 L(X) \exp \zeta(g_u)$ attains a minimum.

In a manner corresponding exactly to the argument developed in [2], the results presented above can be used to obtain the existence of a saddle point for a class of two-person zero-sum games. Since there is nothing new here the details are omitted.

As an example of the class of the Brownian motions described here let $(B_{(t,\tau)}(t,\tau) \in [0,1]^2)$ be a biadditive Gaussian process, i.e., a zero mean Gaussian process with independent increments in each coordinate of the index set such that

$$E[B(t_1, \tau_1)B(t_2, \tau_2)] = (t_1 \wedge t_2)(\tau_1 \wedge \tau_2).$$

This stochastic process has continuous sample paths and when it is considered as indexed by one coordinate of the index set, it is a Brownian motion with values in the Banach space of continuous functions, $C[0,1]$. Optimal control results can be obtained by the previous results for stochastic differential equations with this Brownian motion.

Remark on [2]. It is necessary to make the assumption of uniform integrability of the family of densities to obtain the results in Theorems 4 and 5 of [2]. One sufficient condition for this uniform integrability is that the growth of the drift term is at most linear. This fact is shown by Beneš [1] by verifying that the family of densities have a bounded α th moment for some $\alpha > 1$.

REFERENCES

- [1] V. E. BENEŠ, *Existence of optimal stochastic control laws*, this Journal, 9 (1971), pp. 446–472.
- [2] T. DUNCAN AND P. VARAIYA, *On the solutions of a stochastic control system*, this Journal, 9 (1971), pp. 354–371.
- [3] R. M. DUDLEY, J. FELDMAN AND L. LECAM, *On seminorms and probabilities, and abstract Wiener spaces*, Ann. of Math., 93 (1971), pp. 390–408.
- [4] L. GROSS, *Abstract Wiener spaces*, Proc. Fifth Berkeley Symposium on Math. Stat. and Prob., University of California Press, Berkeley, Calif., 1965.
- [5] I. E. SEGAL, *Distributions in Hilbert space and canonical systems of operators*, Trans. Amer. Math. Soc., 88 (1958), pp. 12–41.
- [6] P. A. MEYER, *Probability and Potentials*, Blaisdell, Waltham, Mass., 1966.
- [7] X. FERNIQUE, *Intégrabilité des vecteurs gaussiens*, C.R. Acad. Sci. Paris Sér. A-B, 270 (1970), pp. 1698–1699.
- [8] E. NELSON, *An existence theorem for second order parabolic equations*, Trans. Amer. Math. Soc., 88 (1958), pp. 414–429.
- [9] L. GROSS, *Potential theory on Hilbert space*, J. Functional Analysis, 1 (1967), pp. 123–181.
- [10] J. L. DOOB, *Stochastic Processes*, John Wiley, New York, 1953.
- [11] K. ITÔ, *Multiple Wiener integral*, J. Math. Soc. Japan, 3 (1951), pp. 157–169.
- [12] N. WIENER, *The homogeneous chaos*, Amer. J. Math., 60 (1938), pp. 897–936.
- [13] R. H. CAMERON AND W. T. MARTIN, *Transformation of Wiener integrals under a general class of linear transformations*, Trans. Amer. Math. Soc., 58 (1945), pp. 184–219.
- [14] I. E. SEGAL, *Tensor algebras over Hilbert space I*, Ibid., 81 (1956), pp. 106–134.

- [15] H. KUNITA AND S. WATANABE, *On square integrable martingales*, Nagoya Math. J., 30 (1967), pp. 209–245.
- [16] I. V. GIRSANOV, *On transforming a certain class of stochastic processes by absolutely continuous substitution of measures*, Theor. Probability Appl., 5 (1960), pp. 285–301.
- [17] T. E. DUNCAN, *Transforming Fréchet-valued Brownian motion by absolute continuity of measures*, unpublished.
- [18] A. BENSOUSSAN, *Généralization du théorème de Girsanov*, to appear.
- [19] C. DOLÉANS-DADE, *Quelques applications de la formule de changement de variables pour les semimartingales*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 16 (1970), pp. 181–194.
- [20] M. LOÉVE, *Probability Theory*, 2nd ed., Van Nostrand, Princeton, N.J., 1960.
- [21] V. E. BENE, *Existence of optimal strategies based on specified information for a class of stochastic decision problems*, this Journal, 8 (1970), pp. 179–188.

ERROR EXPRESSIONS FOR OPTIMAL STATIONARY STATE ESTIMATION*

JAKOV SNYDERS†

Abstract. Explicit expressions are derived for the residual error covariance matrix in optimal causal prediction, filtering and interpolation of a stationary state vector satisfying a linear Ito differential equation with constant coefficients. The estimation is based on the output vector perturbed by noise that is mostly assumed to be white, but extensions to the colored noise case are also indicated. In particular, the error expressions for filtering in white noise provide an explicit solution to a matrix quadratic equation, also known as algebraic Riccati equation. Earlier results obtained by a similar technique are more general (but usually not explicit) expressions for the residual error covariance matrix in linear estimation problems.

1. Introduction. Let an n -dimensional signal x and the m -dimensional observed process y satisfy on $(-\infty, \infty)$ the Ito equations

$$(1) \quad dx(t) = Ax(t)dt + B d\eta(t), \quad x(0) = x_0,$$

$$(2) \quad dy(t) = Cx(t)dt + Q dv(t), \quad y(0) = 0,$$

where A, B, C and Q are constant real matrices, η and v are independent standard Brownian motions and x_0 is a Gaussian vector independent of v and of $\{\eta(t); t \geq 0\}$. Consider the optimal mean-square estimation of $x(t + \tau)$ based on the σ -algebra generated by $\{y(s); s \leq t\}$. It is well known that if x is stationary, then the residual error covariance matrix is constant and, furthermore, the residual filtering (i.e., $\tau = 0$) error covariance matrix E_0 satisfies

$$(3) \quad AE_0 + E_0A' - E_0C'(QQ')^{-1}CE_0 + BB' = 0$$

provided QQ' is nonsingular. This quadratic equation, also known as algebraic Riccati equation, is frequently encountered in control and estimation problems and has been investigated thoroughly [5], [6]. Following a frequency domain approach essentially outlined in [3], we shall derive expressions for E_0 and for the residual error covariance matrix in other estimation problems. This approach is possible because the estimation is actually linear. Unlike most of the results in [3], the expressions obtained below are *explicit* in the sense that they involve only straightforward algebraic manipulations and single integrals. These expressions are apparently not advantageous in comparison with the well-developed time-domain computational techniques, but they are beneficial for investigating the relationship between the residual error matrix and the matrices appearing in the state representation and, furthermore, between the residual error matrices corresponding to various cases. For example, a relatively simple formula connecting the causal and noncausal filtering errors follows from (21) (as explained before Theorem 3), an interesting relationship between filtering in white and colored noises results by comparing (21) with (30), and dependence of the error on time-lag is demonstrated by (24) and (25).

* Received by the editors January 29, 1974, and in revised form July 19, 1974.

† System Science Department, University of California, Los Angeles, California. Now at School of Engineering, Tel-Aviv University, Ramat-Aviv, Tel-Aviv, Israel. This research was supported by the United States Army under Grant KA-75.

The next section contains preliminary considerations. Error expressions for prediction, filtering and interpolation in white noise are derived in § 3. In § 4 prediction and filtering in colored noise are considered. Some related problems are briefly discussed in the last section; in particular a solution to (3) is presented under conditions which do not allow stationarity of x .

2. Preliminaries. Let X stand for the real n -space. The controllable subspace of X associated with (A, B) is denoted $\langle A|B \rangle$ and $\ker C = \{x: Cx = 0\}$. Let $\phi^+(\cdot)$, $\phi^0(\cdot)$ and $\phi^-(\cdot)$ be factors of the minimal polynomial of A having roots exclusively with positive, zero and negative real part, respectively. Then X is decomposable into the direct sum $X = X^+(A) \oplus X^0(A) \oplus X^-(A)$, where $X^x(A) = \ker \phi^x(A)$ with x standing for any superscript. Matrix transposition is indicated by a prime and F^* is the adjoint of F . If F is a matrix-valued function defined on the imaginary axis of a complex plane, then $F \in \mathcal{H}_+^2$ means that *each entry* of F belongs to the Hardy subspace [1] of the Lebesgue space \mathcal{L}^2 , the dimensions being clear from the context. In particular, a matrix-valued function having strictly proper rational entries with poles confined to the open left half-plane belongs to \mathcal{H}_+^2 . We shall frequently consider \mathcal{H}_+^2 -functions extended from their original domain to the right half-plane in the Hardy space theory sense.

Let $K(t, t + \tau)$ be the mean value of $x(t)x'(t + \tau)$. Then for $\tau \geq 0$,

$$(4) \quad \begin{aligned} K(t, t + \tau) &= \left[e^{At} K(0, 0) e^{A'\tau} + \int_0^t e^{As} B B' e^{A's} ds \right] e^{A'\tau}, \\ K(t, t - \tau) &= e^{A\tau} \left[e^{A(t-\tau)} K(0, 0) e^{A'(t-\tau)} + \int_0^{t-\tau} e^{As} B B' e^{A's} ds \right]. \end{aligned}$$

Therefore (1) has a stationary solution if and only if there exists a nonnegative definite K satisfying

$$K = e^{At} K e^{A't} + \int_0^t e^{As} B B' e^{A's} ds$$

for all $t \geq 0$ or, equivalently [4],

$$(5) \quad \langle A|B \rangle \subset X^-(A).$$

If x is stationary, then its spectral density matrix S is given by

$$(6) \quad S(i\omega) = (-A + i\omega I)^{-1} B B' (-A' - i\omega I)^{-1},$$

and the linear estimation error $E(\cdot)$ is expressible as follows:

$$(7) \quad \begin{aligned} E(Y) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} [Y(i\omega)C - e^{i\omega\tau}I] S(i\omega) [C'Y^*(i\omega) - e^{-i\omega\tau}I] d\omega \\ &\quad + \frac{1}{2\pi} \int_{-\infty}^{\infty} Y(i\omega) N(i\omega) Y^*(i\omega) d\omega, \end{aligned}$$

where $N(i\omega) = QQ'$ (in § 4 we shall treat a problem involving nonconstant N). Y is a measurable $n \times m$ "transfer" function defined on the imaginary axis of a complex plane such that $Y(CSC' + N)Y^*$ is integrable. The space formed by

these functions is denoted $\mathcal{L}^2[(CSC' + N) d\omega]$, and its closed subspace of causal transfer functions $\mathcal{H}_+^2[(CSC' + N) d\omega]$ may be defined with the aid of Fourier transforms [3]. We omit the details and mention only that if $N(i\omega) = QQ'$ is nonsingular and CSC' has no poles on the imaginary axis, then

$$Y \in \mathcal{H}_+^2[(CSC' + N) d\omega]$$

if and only if $Y \in \mathcal{H}_+^2$ and is $n \times m$.

By standard techniques it may be shown that the infimum of the trace of $E(\cdot)$ over a closed subspace \mathcal{F} of $\mathcal{L}^2[(CSC' + N) d\omega]$ is attained by a unique function. Denoting this function by Y_0 for the case of $\mathcal{F} = \mathcal{H}_+^2[(CSC' + N) d\omega]$ and writing E for $E(Y_0)$, we have

$$(8) \quad \begin{aligned} E = & \frac{1}{2\pi} \int_{-\infty}^{\infty} [Y_0(i\omega)C - e^{i\omega\tau}I]S(i\omega)[C'Y_0^*(i\omega) - e^{-i\omega\tau}I] d\omega \\ & + \frac{1}{2\pi} \int_{-\infty}^{\infty} Y_0(i\omega)N(i\omega)Y_0^*(i\omega) d\omega. \end{aligned}$$

Further standard steps [3] yield

$$(9) \quad \int_{-\infty}^{\infty} \{[Y_0(i\omega)C - e^{i\omega\tau}I]S(i\omega)C' + Y_0(i\omega)N(i\omega)\} \delta Y^*(i\omega) d\omega = 0$$

for every $\delta Y \in \mathcal{H}_+^2[(CSC' + N) d\omega]$. Setting $\delta Y = Y_0$, we find it follows that

$$(10) \quad E = \frac{1}{2\pi} \int_{-\infty}^{\infty} [I - e^{-i\omega\tau}Y_0(i\omega)C]S(i\omega) d\omega$$

and also

$$E = \frac{1}{2\pi} \int_{-\infty}^{\infty} S(i\omega) d\omega - \frac{1}{2\pi} \int_{-\infty}^{\infty} Y_0(i\omega)[CS(i\omega)C' + N(i\omega)]Y_0^*(i\omega) d\omega.$$

The above (nonexplicit) error expressions are very general; they hold for any nonnegative definite S and N subjected to mild integrability restrictions. In the sequel we shall exploit the particular forms of S and N , and rely heavily on condition (5). Note that S given by (6) is a spectral density whenever $\langle A|B \rangle \cap X^0(A) = 0$, and $E(\cdot)$ in (7) may be formally regarded as an estimation error covariance matrix even without this (considerably milder) restriction. The crucial role that (5) nevertheless plays emphasizes the inherent connection between the problem represented by (7) and the model given by (1) and (2). We shall return to this point in § 5. Another condition imposed for technical reasons, such as convergence of integrals, is nonsingularity of N , although the less restrictive condition $CSC' + N > 0$ could suffice for the major part of the manipulations (see § 4).

The following result is stated under the condition

$$(11) \quad \langle A|B \rangle \cap [X^0(A) \oplus X^+(A)] \subset \ker C.$$

Obviously (5) implies (11), and using (4) it may be checked that (11) is necessary and sufficient for the existence of x_0 such that Cx is stationary. For interpretation

of (11) it is helpful to note that $\ker C$ is replaceable by the unobservable subspace of X associated with (C, A) . If (5) actually holds, then C' may be dropped in (12).

LEMMA 1. Assume (11) and let $F \in \mathcal{H}_+^2$. Then for any number z with positive real part,

$$(12) \quad \frac{1}{2\pi} \int_{-\infty}^{\infty} F(i\omega) B' (-A' - i\omega I)^{-1} C' \frac{d\omega}{z - i\omega} \\ = F(z) B' (-A' - zI)^{-1} C' - \frac{1}{2\pi} \int_{-\infty}^{\infty} F(i\omega) B' (-A' - i\omega I)^{-1} (-A' - zI)^{-1} C' d\omega.$$

Proof. Suppose that $C = I$ and A is in a Jordan canonical form with stable diagonal entries. Then the equality can be proved by applying the scalar version of Lemma 4 in [3] to each entry separately. This implies the general result since, according to (11), a proper choice of basis for X allows the representations

$$C = (C_1 \quad 0 \quad C_2), \quad A = \begin{pmatrix} A_{11} & 0 & A_{31} \\ 0 & A_{22} & A_{32} \\ 0 & 0 & A_{33} \end{pmatrix}, \quad B = \begin{pmatrix} B_1 \\ B_2 \\ 0 \end{pmatrix},$$

where A_{11} is stable and is in Jordan canonical form.

Similar reasoning yields the next result. The technical condition (13) assures convergence of the integrals.

LEMMA 2. If $\mathcal{F} \in \mathcal{H}_+^2$ and

$$(13) \quad \langle A|B \rangle \subset X^-(A) \oplus X^+(A)$$

holds, then for $\tau \geq 0$,

$$\int_{-\infty}^{\infty} e^{i\omega\tau} (-A + i\omega I)^{-1} B F^*(i\omega) d\omega = e^{A\tau} \int_{-\infty}^{\infty} (-A + i\omega I)^{-1} B F^*(i\omega) d\omega.$$

3. Estimation in white noise. Throughout this section it is assumed that E is given by (8) where $N(i\omega) = QQ' > 0$. Introducing the notation $T_A(z) = (-A + zI)^{-1}$, we have $S = T_A B B' T_A^*$ over the imaginary axis, and the extended "spectral density function" is given by $S(z) = T_A(z) B B' T_A'(-z)$. We shall write functions with omitted argument only where the argument is assumed to be imaginary. In the next two theorems, prediction and filtering are considered; Theorems 3 and 4 deal with interpolation.

THEOREM 1. If A is stable and (C, A) is observable, then for $\tau \geq 0$,

$$(14) \quad E = \frac{1}{2\pi} \int_{-\infty}^{\infty} S d\omega - \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{A\tau} S e^{A'\tau} d\omega + e^{A\tau} E_0 e^{A'\tau},$$

where E_0 is given by

$$(15) \quad E_0 = \int_{-\infty}^{\infty} S C' (C S C' + N)^{-1} C T_A d\omega \cdot \left\{ \int_{-\infty}^{\infty} T_A^* C' (C S C' + N)^{-1} C T_A d\omega \right\}^{-1}.$$

Proof. According to (10) and the adjoint version of Lemma 2,

$$(16) \quad E = \frac{1}{2\pi} \int_{-\infty}^{\infty} S(i\omega) d\omega - \frac{1}{2\pi} \int_{-\infty}^{\infty} Y_0(i\omega) C S(i\omega) d\omega \cdot e^{A^* \tau},$$

and (14) follows if we define

$$(17) \quad E_0 = \frac{1}{2\pi} \int_{-\infty}^{\infty} [I - e^{-A^* \tau} Y_0(i\omega) C] S(i\omega) d\omega.$$

Insertion of $\delta Y^*(i\omega) = (z - i\omega)^{-1}$, where $\operatorname{Re}(z) > 0$ (more precisely, $\delta Y^*(i\omega) = ((z - i\omega)^{-1} I - 0)$ if $n > m$ and likewise in other cases), into (9) yields

$$(18) \quad \int_{-\infty}^{\infty} Y_0(i\omega) [C S(i\omega) C' + N] \frac{d\omega}{z - i\omega} = \int_{-\infty}^{\infty} e^{i\omega \tau} S(i\omega) C' \frac{d\omega}{z - i\omega},$$

and by Lemma 2,

$$(19) \quad \int_{-\infty}^{\infty} Y_0(i\omega) [C S(i\omega) C' + N] \frac{d\omega}{z - i\omega} = e^{A^* \tau} \int_{-\infty}^{\infty} S(i\omega) C' \frac{d\omega}{z - i\omega}.$$

Applying Lemma 1 to both sides of (19) and in view of (17),

$$Y_0(z) [C S(z) C' + N] = e^{A^* \tau} S(z) C' - e^{A^* \tau} E_0 (-A' - zI)^{-1} C'.$$

This equality must obviously hold also over the imaginary axis, therefore

$$(20) \quad e^{-A^* \tau} Y_0 = S C' (C S C' + N)^{-1} - E_0 T_A^* C' (C S C' + N)^{-1}.$$

Postmultiplication by $C T_A$ and integration lead to (15). The integrals exist due to stability of A and $(C S C' + N)^{-1} \leq N^{-1}$, and the matrix inversion is possible because (C, A) is observable.

Evidently E_0 is the residual error covariance matrix corresponding to $\tau = 0$. Under milder assumptions, the following expression is available.

THEOREM 2. *If $\langle A|B \rangle \subset X^-(A)$ and $\tau \geq 0$, then E is given by (14) where*

$$(21) \quad E_0 = \frac{1}{2\pi} \int_{-\infty}^{\infty} [S - S C' (C S C' + N)^{-1} C S] d\omega \\ \cdot \left\{ I - \frac{1}{2\pi} \int_{-\infty}^{\infty} T_A^* C' (C S C' + N)^{-1} C S d\omega \right\}^{-1}.$$

Proof. (14) is obtainable as before, and the former proof of (20) is valid also now. Postmultiplying (20) by $C S$ and using (17), we get

$$(22) \quad \frac{1}{2\pi} \int_{-\infty}^{\infty} [S - S C' (C S C' + N)^{-1} C S] d\omega \\ = E_0 \left\{ I - \frac{1}{2\pi} \int_{-\infty}^{\infty} T_A^* C' (C S C' + N)^{-1} C S d\omega \right\},$$

and the result follows. For establishing the required nonsingularity, assume first that (A, B) is controllable and consider $E(\cdot)$ defined by (7). Obviously $E(Y) > 0$ for arbitrary $Y \in \mathcal{L}^2[(C S C' + N) d\omega]$, and setting

$$(23) \quad Y(i\omega) = e^{i\omega\tau} S(i\omega) C' [CS(i\omega) C' + N]^{-1},$$

we get

$$E(Y) = \frac{1}{2\pi} \int_{-\infty}^{\infty} [S - SC'(CSC' + N)^{-1}CS] d\omega > 0.$$

Therefore the right-hand side of (22) is nonsingular (and also $E_0 > 0$). The general case follows by applying this conclusion, using controllability canonical form.

Note that (14) is well known in the following form [2, Thm. 7.2]:

$$E = \int_0^{\tau} e^{As} BB' e^{A's} ds + e^{A\tau} E_0 e^{A'\tau}.$$

Indeed,

$$\begin{aligned} & \frac{1}{2\pi} \int_{-\infty}^{\infty} S(i\omega) d\omega - \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{A\tau} S(i\omega) e^{A'\tau} d\omega \\ &= \int_0^{\infty} e^{As} BB' e^{A's} ds - \int_0^{\infty} e^{A(s+\tau)} BB' e^{A'(s+\tau)} ds \\ &= \int_0^{\tau} e^{As} BB' e^{A's} ds. \end{aligned}$$

It is also interesting that the first factor in (21) is the linear optimal (noncausal) error covariance matrix. This follows by observing that (9) is satisfied for every $\delta Y \in \mathcal{L}^2[(CSC' + N) d\omega]$ if Y given by (23) stands for Y_0 .

THEOREM 3. Assume that A is stable and (C, A) is observable. Then for $\tau \leq 0$,

$$(24) \quad \begin{aligned} E &= \frac{1}{2\pi} \int_{-\infty}^{\infty} [S - SC'(CSC' + N)^{-1}CS] d\omega + P \\ &\quad + \frac{1}{2\pi} \int_{-\infty}^{\infty} D^* T_A^* C' (CSC' + N)^{-1} CS d\omega, \end{aligned}$$

where

$$(25) \quad P = \int_{-\infty}^{\infty} DSC'(CSC' + N)^{-1}CT_A d\omega \left\{ \int_{-\infty}^{\infty} T_A^* C' (CSC' + N)^{-1}CT_A d\omega \right\}^{-1}$$

and $D(i\omega) = e^{i\omega\tau}I$.

Proof. Applying Lemma 1 to both sides of (18) and denoting

$$P = \frac{1}{2\pi} \int_{-\infty}^{\infty} [D(i\omega) - Y_0(i\omega)C]S(i\omega) d\omega,$$

we see that it follows that

$$Y_0(z)[CS(z)C' + N] = D(z)S(z)C' - PT_A^*(-z)C'$$

for $\operatorname{Re}(z) > 0$ and, consequently, also for $\operatorname{Re}(z) = 0$. Thus

$$(26) \quad Y_0 = DSC'(CSC' + N)^{-1} - PT_A^*C'(CSC' + N)^{-1},$$

and substitution into (10) yields (24). Postmultiplication of (26) by CT_A and subsequent integration lead to (25).

The next expression is also derivable by the above applied techniques.

THEOREM 4. *If $\langle A|B \rangle \subset X^-(A)$ and $\tau \leq 0$, then E is given by (24) where*

$$P = \frac{1}{2\pi} \int_{-\infty}^{\infty} D[S - SC'(CSC' + N)^{-1}CS] d\omega \\ \cdot \left\{ I - \frac{1}{2\pi} \int_{-\infty}^{\infty} T_A^* C'(CSC' + N)^{-1}CS d\omega \right\}^{-1}$$

and $D(i\omega) = e^{i\omega\tau}I$.

Obviously, if $\tau = 0$, then $P = E_0$, and consequently $E = E_0$, the filtering error covariance matrix.

4. Estimation in colored noise. The approach followed above is also suitable for problems where N represents colored noise generated by processing white noise through a dynamic system, or a mixture of colored and white noises. However, the resulting error expressions will be, in general, more complicated. We shall demonstrate the procedure by a relatively simple case.

Consider the prediction and filtering problem ($\tau \geq 0$) that involves the noise spectral density:

$$N(i\omega) = (-F + i\omega I)^{-1}GG'(-F' - i\omega I)^{-1},$$

where F and G are constant matrices with proper dimensions. Then $n \times m$ functions with proper (even though not strictly) rational entries and with poles confined to the open left half-plane belong to $\mathcal{H}_+^2[(CSC' + N)d\omega]$. Assume that $\langle A|B \rangle \subset X^-(A)$, $\langle F|G \rangle \subset X^-(F)$ (with $X^-(F)$ standing for the stable subspace of an m -space associated with F , etc.) and $CS(i\omega)C' + N(i\omega) > 0$ except, possibly, at a finite number of points. The former steps leading to (16) are valid also now, and defining

$$(27) \quad E = \frac{1}{2\pi} \int_{-\infty}^{\infty} [I - e^{-A\tau}Y_0(i\omega)C]S(i\omega)d\omega,$$

we again have

$$E = \frac{1}{2\pi} \int_{-\infty}^{\infty} S d\omega - \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{A\tau}S e^{A'\tau} d\omega + e^{A\tau}E_0 e^{A'\tau}.$$

Below we shall obtain an explicit expression for E_0 which will not contain τ ; thus it will follow that E_0 is the filtering error covariance matrix.

Setting $\delta Y = 1$ into (9) and applying Lemma 2,

$$\int_{-\infty}^{\infty} Y_0(i\omega)N(i\omega)d\omega = \int_{-\infty}^{\infty} e^{i\omega\tau}S(i\omega)C' d\omega - \int_{-\infty}^{\infty} Y_0(i\omega)CS(i\omega)C' d\omega \\ = e^{A\tau} \int_{-\infty}^{\infty} S(i\omega)C' d\omega - \int_{-\infty}^{\infty} Y_0(i\omega)CS(i\omega)C' d\omega,$$

and by (27),

$$(28) \quad \frac{1}{2\pi} \int_{-\infty}^{\infty} Y_0(i\omega) N(i\omega) d\omega = e^{A\tau} E_0 C'.$$

The corresponding version of (19) reads

$$\int_{-\infty}^{\infty} Y_0(i\omega) [CS(i\omega)C' + N(i\omega)] \frac{d\omega}{z - i\omega} = e^{A\tau} \int_{-\infty}^{\infty} S(i\omega) C' \frac{d\omega}{z - i\omega},$$

and by Lemma 1, applying (27) and (28),

$$Y_0(z) [CS(z)C' + N] = e^{A\tau} S(z)C' - e^{A\tau} E_0 (-A' - zI)^{-1} C' + e^{A\tau} E_0 C' (-F - zI)^{-1}$$

for $\operatorname{Re}(z) > 0$. With the notations

$$T_A(i\omega) = (-A + i\omega I)^{-1}, \quad T_F(i\omega) = (-F + i\omega I)^{-1},$$

it thus follows

$$e^{-A\tau} Y_0 = SC'(CSC' + N)^{-1} - E_0(T_A^* C' - C' T_F^*)(CSC' + N)^{-1}$$

over the imaginary axis except, possibly, at a finite number of points. Combination of this result with (27) yields

$$(29) \quad \begin{aligned} & \frac{1}{2\pi} \int_{-\infty}^{\infty} [S - SC'(CSC' + N)^{-1} CS] d\omega \\ & = E_0 - \frac{1}{2\pi} \int_{-\infty}^{\infty} E_0(T_A^* C' - C' T_F^*)(CSC' + N)^{-1} CS d\omega. \end{aligned}$$

The integral on the left-hand side of (29) indeed converges since

$$B' T_A^* C' (CSC' + N)^{-1} C T_A B \leq I;$$

therefore the other integral is also bounded. If the condition $CS(i\omega)C' + N(i\omega) > 0$ a.e. is replaced by the stronger requirement $CBB'C' + GG' > 0$, then the entries of $\omega^{-2}[CS(i\omega)C' + N(i\omega)]^{-1}$ are proper rational, and since the entries of $\omega^2[T_A^*(i\omega)C' - C'T_F^*(i\omega)]$ are proper rational, it thus follows that E_0 may be separated from the integrand in the right-hand side of (29). This may still not be sufficient for writing

$$(30) \quad \begin{aligned} E_0 &= \frac{1}{2\pi} \int_{-\infty}^{\infty} [S - SC'(CSC' + N)^{-1} CS] d\omega \\ &\cdot \left\{ I - \frac{1}{2\pi} \int_{-\infty}^{\infty} (T_A^* C' - C' T_F^*)(CSC' + N)^{-1} CS d\omega \right\}^{-1}. \end{aligned}$$

However, (30) does hold if $GG' > 0$; this is provable by taking into account that the left-hand side of (29) is the optimal linear (noncausal) filtering error matrix, as in the proof of Theorem 2.

5. Related problems. The expressions (15) and (21) provide an explicit solution to (3) under the stated assumptions. Since these assumptions are not essential for the existence of a nonnegative definite solution to (3), it is tempting

to work for more general explicit solutions. It would not be proper to consider the causal filtering problem represented by (7) under milder assumptions than $\langle A|B \rangle \subset X^-(A)$, because in case of nonstationary x , (7) cannot stand for the model (1), (2). Our approach is nevertheless applicable for deriving an expression, although not very elegant, for the case where $C = I$, (A, B) is stabilizable (i.e., $\langle A|B \rangle \supset X^0(A) \oplus X^+(A)$) and $X^0(A) = 0$. Indeed, assume that these conditions hold and consider the minimization of the trace of $E(\cdot)$ given by (7), where $\tau = 0$ and $N(i\omega) = QQ' > 0$ over the set

$$\{Y: Y \in \mathcal{H}_+^2 \text{ and } (Y - I)T_A B \in \mathcal{H}_+^2\}.$$

The existence of a unique function Y_0 attaining the infimum can be demonstrated by standard techniques, and it may be assumed that Y_0 is rational. Steps similar to those in § 3 yield that

$$(31) \quad I - Y_0(i\omega) = [N + E_0 T_A^*(i\omega)][S(i\omega) + N]^{-1}$$

and $E_0 = E(Y_0)$ is given by

$$(32) \quad E_0 = \int_{-\infty}^{\infty} [S(i\omega) + N]^{-1} S(i\omega) d\omega \\ \cdot \left\{ \lim_{v \rightarrow \infty} \int_{-\infty}^{\infty} T_A^*(i\omega) [S(i\omega) + N]^{-1} \frac{v d\omega}{v + i\omega} \right\}^{-1}.$$

Assume now, without loss of generality, that

$$A = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}, \quad B = \begin{pmatrix} B_1 \\ 0 \end{pmatrix},$$

where (A_{11}, B_1) is controllable, and let

$$P(i\omega) = \text{diag}((1 + i\omega)(-A_{11} + i\omega I)^{-1}, I).$$

It is possible to factorize

$$S(i\omega) + N = P(i\omega)G^+(i\omega)G^-(i\omega)P^*(i\omega),$$

where the poles of both $G^+(z)$ and $[G^+(z)]^{-1}$ are confined to the open left half-plane and $G^-(i\omega) = [G^+(i\omega)]^*$. Then

$$\delta Y^*(i\omega) = [G^-(i\omega)P^*(i\omega)(z - i\omega)^2]^{-1},$$

where $\text{Re}(z) > 0$, is eligible for (9) because $\delta Y \in \mathcal{H}_+^2$ and $\delta Y T_A B \in \mathcal{H}_+^2$, and we get

$$\int_{-\infty}^{\infty} [I - Y_0(i\omega)]P(i\omega)G^+(i\omega) \frac{d\omega}{(z - i\omega)^2} = \int_{-\infty}^{\infty} N[G^-(i\omega)P^*(i\omega)]^{-1} \frac{d\omega}{(z - i\omega)^2} = 0.$$

Since $[I - Y_0(i\omega)]P(i\omega)G^+(i\omega)(1 + i\omega)^{-1} \in \mathcal{H}_+^2$ by the assumptions on Y_0 , it thus follows that $(I - Y_0)PG^+$ is constant. Obviously this constant is a square root of N , i.e.,

$$(33) \quad (I - Y_0)(S + N)(I - Y_0)^* = N.$$

Substitution of (31) into (33) yields

$$(QQ' + E_0 T_A^*)(S + QQ')^{-1}(QQ' + T_A E_0) = QQ',$$

and (3) readily follows (with $C = I$).

It is straightforward to check that if, in addition to the above assumptions, $\langle A|B \rangle \subset X^-(A)$ (i.e., A is stable), then (21) and (32) coincide.

Let us now examine the problem of minimizing the trace of $E(\cdot)$, where $N(i\omega) = QQ' > 0$ and $\tau \geq 0$, over \mathcal{H}_+^2 under assumption (11). For assuring convergence of integrals, assume also (13) and $X^-(A) \subset \ker C$. Then all the steps in the proof of Theorem 1 up to (19) are valid also now. However, Lemma 1 is not applicable, in general, to the right-hand side of (19), and this renders a complicated form that involves a double integral to the resulting expression for E . On the other hand, for CEC' we have

$$CEC' = \frac{1}{2\pi} \int_{-\infty}^{\infty} C S C' d\omega - \frac{1}{2\pi} \int_{-\infty}^{\infty} C e^{A\tau} S e^{A'\tau} C' d\omega + C e^{A\tau} E_0 e^{A'\tau} C',$$

where E_0 is given by (21). CEC' is, of course, the residual error covariance matrix in prediction of $Cx(\cdot)$ under the assumption that this process is stationary.

REFERENCES

- [1] K. HOFFMAN, *Banach Spaces of Analytic Functions*, Prentice-Hall, Englewood Cliffs, N.J., 1962.
- [2] J. S. MEDITCH, *Stochastic Optimal Linear Estimation and Control*, McGraw-Hill, N.Y., 1969.
- [3] J. SNYDERS, *On the error matrix in optimal linear filtering of stationary processes*, IEEE Trans. Information Theory, IT-19 (1973), pp. 593–599.
- [4] M. ZAKAI AND J. SNYDERS, *Stationary probability measures for linear differential equations driven by white noise*, J. Differential Equations, 8 (1970), pp. 27–33.
- [5] W. M. WONHAM, *On a matrix Riccati equation of stochastic control*, this Journal, 6 (1968), pp. 681–697.
- [6] J. C. WILLEMS, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 621–634.

EXISTENCE AND CONTROL OF MARKOV CHAINS IN SYSTEMS OF DETERMINISTIC MOTION*

JOHN B. MOORE AND S. SANKAR SENGUPTA†

Abstract. If the phase space X of a motion $x_{n+1} = f(x_n)$ is discretized into a space of states X_1, \dots, X_N , then probabilities can be assigned to sample paths in the state space so as to coincide with the ones assigned by a finite Markov chain. Theorems 1 and 2 show how the assignment of such probabilities rests on the properties of $f(\cdot)$ and on the construction of the states. Theorems 3 and 4 extend these results to the case in which $x_{n+1} = f(x_n, \omega)$, $\omega \in \Omega$ being a random event. Theorems 5 and 6 indicate certain applications relating to stochastic systems in which a decision-maker applies some control action which is fully or partially determined by the observed state of the system.

The usual description of a moving system, observed at instants separated by preassigned intervals of time, is in terms of a first order difference equation

$$(1) \quad x_{n+1} = f(x_n), \quad \{x_n\}_{n=0,1,2,\dots} \in X,$$

in the coordinate of the moving object. This is a pointwise description, i.e., a description of motion in the *phase space* X . If one covers the phase space with a finite set of disjoint subsets $\{X_i\}$, $i = 1, 2, \dots, N$, hereafter called *states*, the pointwise motion in phase space induces a state-to-state motion in the state space $\{X_i\}$. If the location of the initial phase point x_0 is described by a probability function μ defined on the sigma field $\sigma(X)$, the motion in state space is a well-defined random process. The first question discussed in this communication is simply, "What properties must be possessed by the probability space $\langle X, \sigma(X), \mu \rangle$, such that the probabilities assigned to sample paths in the state space coincide with the ones assigned by a finite Markov chain?". Theorems 1 and 2 are concerned with this question.

In addition to randomness induced by discretization, there may be randomness in the process itself, i.e., the realization of the motion of the phase point may be contingent upon the occurrence of a random event $\omega \in \Omega$:

$$x_{n+1} = f(x_n, \omega).$$

It is conceivable that the probabilities of ω may—in the sense of conditional probability—depend on x_n . Theorems 3 and 4 extend the ideas of Theorems 1 and 2 to these cases.

1.

1.1. Let $\langle X, \sigma(X), \mu \rangle$ be a probability space. Let $\{X_n\}$, $n = 1, 2, \dots, N$, be a measurable cover of X , i.e.,

$$\bigcup_{i=1}^N X_i = X \quad \text{and} \quad X_i \cap X_j = \emptyset, \quad i \neq j.$$

* Received by the editors March 1, 1973, and in revised form September 3, 1974.

† Department of Management Sciences, Faculty of Engineering, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1. This work was supported in part by the National Research Council under Grants A8236 and A7416.

Let $f: X \rightarrow X$ be $\sigma(X)$ -measurable and let

$$(2) \quad f^{n+1}(x) \equiv f(f^n(x)), \quad n = 1, 2, \dots,$$

where $f^1(x) \equiv f(x)$. Let

$$(3) \quad X_{i_0 i_1 i_2 \dots} \equiv \{x: x \in X_{i_0}, f(x) \in X_{i_1}, f^2(x) \in X_{i_2}, \dots\}.$$

The motion in state space will be described by a family of functions

$$Z_n: X \rightarrow \{1, 2, 3, \dots, N\}, \quad n = 0, 1, 2, \dots,$$

defined as follows:

$$(4) \quad (Z_n(x) = i) \Leftrightarrow (f^n(x) \in X_i), \quad n = 0, 1, 2, \dots.$$

For any x , the sequence $Z_n(x)$ is well-defined. An immediate consequence of the definitions (3) and (4) is that

$$(5) \quad (x \in X_{i_0 i_1 i_2 \dots}) \Leftrightarrow (Z_0(x) = i_0, Z_1(x) = i_1, Z_2(x) = i_2, \dots).$$

Because sets of the form $X_{i_0 i_1 i_2 \dots}$ are μ -measurable,

$$(6) \quad \Pr [Z_0(x) = i_0, Z_1(x) = i_1, \dots] = \mu\{X_{i_0 i_1 \dots}\}$$

is well-defined for all sequences (i_0, i_1, i_2, \dots) . By defining probabilities in this way, it is easy to prove the motion in state space, i.e., $\{Z_n\}$ is a well-defined random process.

The following theorem deals with conditions which are necessary if $\{Z_n\}$ is a Markov chain. A sufficient condition is stated as a corollary.

THEOREM 1. *Let $x_{n+1} = f(x_n)$ define the motion of a point in phase space. Let $\langle X, \sigma(X), \mu \rangle$ be a probability space. Let $\{X_i\}$, $i = 1, 2, \dots, N$, define the state space and let $\{Z_n\}$, $n = 0, 1, \dots$, be the random process in state space (as defined in (4) and (6)) induced by the motion in phase space. If $\{Z_n\}$ is a Markov chain having an initial probability distribution given by¹*

$$(7) \quad \Pr [Z_0 = i] = p_i, \quad i = 1, 2, \dots, N,$$

and stationary transition probabilities given by²

$$(8) \quad \Pr [Z_{n+1} = j | Z_n = i] = p_{ij}, \quad i, j = 1, 2, \dots, N,$$

then

$$(9) \quad (i) \quad p_i = \mu\{X_i\}, \quad i = 1, 2, \dots, N,$$

$$(10) \quad (ii) \quad p_{ij} = \mu\{X_{ij}\} / \mu\{X_i\},$$

$$(11) \quad (iii) \quad \mu\{X_{i_0, i_1, i_2, \dots, i_n}\} = \frac{\mu\{X_{i_0 i_1}\} \cdot \mu\{X_{i_1 i_2 \dots i_n}\}}{\mu\{X_{i_1}\}}.$$

Proof. (i) and (ii) are immediate consequences of the definitions and are included only for the sake of completeness. The third result can be proved by straightforward induction on n . Only the first step in the induction will be presented here. For any X_{ijk} such that $\mu\{X_{ijk}\} > 0$, it follows from (9) and (10)

¹ As is customary, the notation $\Pr [Z_n(x) = i]$ will be shortened to $\Pr [Z_n = i]$.

² See footnote 1.

that

$$\begin{aligned}
 \mu\{X_{ijk}\} &= \Pr[Z_0 = i, Z_1 = j, Z_2 = k] \\
 &= p_i p_{ij} p_{jk} \\
 &= \mu\{X_{ij}\} \cdot \frac{\mu\{X_{jk}\}}{\mu\{X_j\}} \\
 &= \frac{\mu\{X_{ij}\}}{\mu\{X_j\}} \cdot \mu\{X_{jk}\}.
 \end{aligned}$$

Q.E.D.

COROLLARY 1.1. Condition (iii) is sufficient.

Proof. If $\mu\{X_{i_0 i_1 i_2 \dots i_n}\} > 0$, then from (11),

$$\begin{aligned}
 \mu\{X_{i_0 i_1 i_2 \dots i_n}\} &= \frac{\mu\{X_{i_0 i_1}\}}{\mu\{X_{i_1}\}} \cdot \frac{\mu\{X_{i_1 i_2}\}}{\mu\{X_{i_2}\}} \dots \frac{\mu\{X_{i_{n-1} i_n}\}}{\mu\{X_{i_n}\}} \cdot \mu\{X_{i_n}\} \\
 &= \mu\{X_{i_0}\} \cdot \frac{\mu\{X_{i_0 i_1}\}}{\mu\{X_{i_0}\}} \cdot \frac{\mu\{X_{i_1 i_2}\}}{\mu\{X_{i_1}\}} \dots \frac{\mu\{X_{i_{n-1} i_n}\}}{\mu\{X_{i_{n-1}}\}} \\
 &= p_{i_0} \cdot p_{i_0 i_1} \cdot p_{i_1 i_2} \dots p_{i_{n-1} i_n}.
 \end{aligned}$$

COROLLARY 1.2. If $\langle X, \sigma(X), \mu' \rangle$ is another probability space such that for each $i_0 \in \{1, 2, \dots, N\}$, there exists a constant c_{i_0} such that

$$\mu'\{X_{i_0 i_1 \dots i_n}\} = c_{i_0} \mu\{X_{i_0 i_1 \dots i_n}\},$$

then the random process $\{Z'_n\}$ is again a Markov chain with an initial probability distribution $p_i = \mu'\{X_i\}$ and transition probabilities

$$p_{ij} = \mu\{X_{ij}\} / \mu\{X_i\}.$$

Proof. The result is an immediate consequence of the fact that

$$\frac{\mu'\{X_{i_0 i_1 \dots i_n}\}}{\mu'\{X_{i_0 i_1 \dots i_{n-1}}\}} = \frac{\mu\{X_{i_0 i_1 \dots i_n}\}}{\mu\{X_{i_0 i_1 \dots i_{n-1}}\}}.$$

Q.E.D.

Remarks. The probability measure μ determines both the initial probability distribution $\{p_i\}$ and the transition probabilities $\{p_{ij}\}$. Since the p_{ij} values are assumed to be stationary, i.e., $\Pr[Z_{n+1} = j | Z_n = i] = \Pr[Z_{m+1} = j | Z_m = i]$, $m, n = 1, 2, \dots$, the condition $\mu\{X_i\} > 0$ is unavoidable if $\mu\{X_{ij}\} / \mu\{X_i\}$ is to be well-defined for all i, j values. However, because the initial distribution is fixed by μ , it may well be that for some i , $\Pr[Z_0 = i] = 0$ and yet $\Pr[Z_n = i] \neq 0$ for some $n > 0$. In this case, the transition probabilities can be defined as follows. Suppose $\Pr[Z_n = i] = 0$ for $n = 0, 1, 2, \dots, m-1$ and $\Pr[Z_m = i] \neq 0$. Then define $p_i = 0$ and define

$$(12) \quad p_{ij} = \frac{\mu\{x : f^m(x) \in X_i, f^{m+1}(x) \in X_j\}}{\mu\{x : f^m(x) \in X_i\}}.$$

If the transition probabilities are defined in this way, one may still claim that the Markov chain has stationary transition probabilities since $\Pr[Z_{n+1} = j | Z_n = i]$ is undefined for $n = 0, 1, \dots, m-1$.

The reason for not using this definition of transition probabilities in Theorem 1 was simply to remove any ambiguity associated with the term "stationary transition probabilities". In the sections of this communication which follow, it will be assumed that $\mu\{X_{ij}\}/\mu\{X_i\}$ is well-defined for all i, j . The results obtained can easily be extended to permit the more general definition of p_{ij} values found in (12).

1.2. In this section the probability space $\langle X, \sigma(X), \mu \rangle$ is specialized to the space $\langle [0, 1], B, \mu \rangle$, where B denotes the Borel field of subintervals of $[0, 1]$ and μ is the Lebesgue measure. Sufficient conditions are found which will insure that the motion in state space is a Markov chain for an arbitrary partition of $[0, 1]$. The conditions are, loosely put, that if there are some points in X_i which map into X_j , the collection of all such points must map onto X_j and $f(\cdot)$ defined on this subset X_{ij} must be linear. The reader will no doubt feel, as the authors do, that these are fairly strong conditions. However, the authors' investigations indicate that these conditions are probably necessary (in the mathematical sense) if the Markov chain is to be ergodic (see [2, Chap. 15]). In fact, to avoid chains having transient and periodic states (see [2, Chap. 15]) may require the given conditions to be true. This latter statement remains a conjecture yet to be proved.

THEOREM 2. Suppose $\langle X, B(X), \mu \rangle$ is a probability space where X is a $[0, 1]$ interval of real numbers, $B(X)$ is the Borel field of subsets of X and μ is the Lebesgue measure defined on $B(X)$. Suppose X is the phase space of motion

$$x_{n+1} = f(x_n), \quad n = 0, 1, 2, \dots$$

Let $\{Z_n\}$ be the motion induced in the state space $\{X_i\}$. Then sufficient conditions for $\{Z_n\}$ to be a Markov chain are that if X_{ij} is nonempty, then

$$(13) \quad (a) \quad f(X_{ij}) \stackrel{\text{a.e.}}{=} X_j, \quad 1 \leq i, j \leq N,$$

$$(14) \quad (b) \quad f(x)|_{x \in X_{ij}} = \alpha_{ij}x, \quad \alpha_{ij} \geq 0, \quad 1 \leq i, j \leq N.$$

If (13) and (14) are fulfilled, the probabilities associated with the Markov chain are

$$\Pr[Z_0 = i] = \mu\{X_i\}, \quad \Pr[Z_{n+1} = j | Z_n = i] = \frac{\mu\{X_{ij}\}}{\mu\{X_i\}}.$$

Remarks. In view of Theorem 1, Theorem 2 will be proved if it can be shown that

$$(15) \quad \mu\{X_{i_0 i_1 i_2 \dots i_n}\} = \frac{\mu\{X_{i_0 i_1}\}}{\mu\{X_{i_1}\}} \cdot \mu\{X_{i_1 i_2 i_3 \dots i_n}\}$$

whenever $X_{i_0 i_1 i_2 \dots i_n}$ is nonempty.

Proof. We begin by showing inductively that

$$(16) \quad f(X_{i_0 i_1 \dots i_n}) \stackrel{\text{a.e.}}{=} X_{i_1 i_2 \dots i_n}$$

whenever $X_{i_0 i_1 \dots i_n}$ is nonempty. Since $f(X_{ijk}) \subset X_{jk}$, $\mu\{f(X_{ijk})\} \leq \mu\{X_{jk}\}$. Since $X_{ij} = \bigcup_k X_{ijk}$, it follows that $f(X_{ij}) = \bigcup_k f(X_{ijk})$. Therefore

$$X_j = f(X_{ij}) = \bigcup_k f(X_{ijk}).$$

From the linearity of $f(\cdot)$ on X_{ij} it follows that

$$\begin{aligned}
 \mu\{X_j\} &= \mu \bigcup_k f(X_{ijk}) \\
 (17) \qquad &= \sum_k \mu\{f(X_{ijk})\} \\
 &= \sum_k \alpha_{ij} \mu\{X_{ijk}\}.
 \end{aligned}$$

Now $\mu\{X_j\} = \sum_k \mu\{X_{jk}\}$, and $\alpha_{ij} \mu\{X_{ijk}\} \leq \mu\{X_{jk}\}$. Hence

$$(18) \qquad \sum_k \mu\{X_{jk}\} = \sum_k \alpha_{ij} \mu\{X_{ijk}\}$$

can be true only if $\mu\{X_{jk}\} = \alpha_{ij} \mu\{X_{ijk}\}$. Since $f(\cdot)$ has a constant slope on X_{ij} , this is equivalent to

$$(19) \qquad f(X_{ijk}) \stackrel{\text{a.e.}}{=} X_{jk}.$$

To complete the induction, suppose for all nonempty $X_{i_1 i_2 \dots i_m}$ that

$$f(X_{i_1 i_2 \dots i_m}) \stackrel{\text{a.e.}}{=} X_{i_2 i_3 \dots i_m}$$

for all $m \leq n$. Consider any nonempty set $X_{i_0 i_2 \dots i_{n-1}}$. Then proceeding as above, we have $f(X_{i_0 \dots i_{n-1}}) \subset X_{i_1 i_2 \dots i_{n-1}}$ and therefore $\alpha_{i_0 i_1} \mu\{X_{i_0 i_1 \dots i_{n-1}}\} \leq \mu\{X_{i_1 i_2 \dots i_{n-1}}\}$. On the other hand, $X_{i_0 i_1 \dots i_{n-1}} = \bigcup_{i_n} X_{i_0 i_1 \dots i_n}$ and since the elements of the union are disjoint it follows that $f(X_{i_0 i_1 \dots i_{n-1}}) = \bigcup_{i_n} f(X_{i_0 i_1 \dots i_n})$. According to the induction hypothesis, this implies

$$X_{i_1 i_2 \dots i_{n-1}} = \bigcup_{i_n} f(X_{i_0 i_1 \dots i_n});$$

thus one has

$$(20) \qquad \sum_{i_n} \mu\{X_{i_1 \dots i_n}\} = \sum_{i_n} \alpha_{i_0 i_1} \mu\{X_{i_0 \dots i_n}\}.$$

Since no individual term in the right member of (20) can exceed the corresponding term in the left member of (20), equation (20) can hold only if

$$\mu\{X_{i_1 \dots i_n}\} = \alpha_{i_0 i_1} \mu\{X_{i_0 \dots i_n}\}.$$

Because of the linearity of $f(\cdot)$, one must have

$$f(X_{i_0 i_1 \dots i_n}) \stackrel{\text{a.e.}}{=} X_{i_1 i_2 \dots i_n}.$$

This completes the induction.

Now observe that

$$(21) \qquad \mu\{X_{i_0 i_1 \dots i_n}\} = \frac{\mu\{X_{i_1 \dots i_n}\}}{\alpha_{i_0 i_1}}.$$

But $\alpha_{ij} = \mu\{X_j\}/\mu\{X_{ij}\}$ whenever X_{ij} is nonempty. Therefore (21) may be written as

$$\mu\{X_{i_0 i_1 \dots i_n}\} = \frac{\mu\{X_{i_0 i_1}\}}{\mu\{X_{i_1}\}} \cdot \mu\{X_{i_1 i_2 \dots i_n}\}.$$

This, according to Corollary 1.1, is sufficient to guarantee that $\{Z_n\}$ is a Markov chain having the probabilities stated in the hypothesis. Q.E.D.

COROLLARY 2.1. *Theorem 2 is true when α_{ij} is replaced by $-\alpha_{ij}$.*

Proof. If $\alpha_{ij} < 0$, one need only replace α_{ij} with $\|\alpha_{ij}\|$ (the absolute value of the slope) in the proof of Theorem 2.

COROLLARY 2.2. *Let $\{f_k\}$, $k = 1, 2, \dots, K$, be a set of functions each of which fulfills the conditions Theorem 2. Let α_{ij}^k denote the slope of f_k on X_{ij} . Define*

$$(22) \quad X_{i_0 i_1 \dots i_n}^{k_1 k_2 \dots k_n} = \{x: x \in X_{i_0}, f_{k_1}(x) \in X_{i_1}, f_{k_2}(f_{k_1}(x)) \in X_{i_2}, \dots\}.$$

Then the method of induction used in Theorem 2 can be directly applied to establish

$$(23) \quad f_{k_1}(X_{i_0 i_1 \dots i_n}^{k_1 k_2 \dots k_n}) \stackrel{\text{a.e.}}{=} X_{i_1 i_2 \dots i_n}^{k_2 k_3 \dots k_n}$$

whenever $X_{i_0 i_1 \dots i_n}^{k_1 k_2 \dots k_n}$ is nonempty.

1.3. The following numerical example illustrates the ideas of Theorem 2. Suppose $X = [0, 1]$, $X_1 = \{x: 0 \leq x \leq \frac{1}{3}\}$ and $X_2 = \{x: \frac{1}{3} < x \leq 1\}$, and consider two functions

$$f_1(x) = \begin{cases} 3x, & x \in X_1, \\ 7/3 - 4x, & 1/3 \leq x \leq 1/2, \\ 2/3 - 2x/3, & 1/2 < x \leq 1, \end{cases} \quad f_2(x) = \begin{cases} 1 - 2x, & x \in X_1, \\ x/2 - 1/6, & x \in X_2. \end{cases}$$

Both of these functions satisfy the requirements of Theorem 2, and the motion in state space is a Markov chain in both cases. In each case, the values of $p_{ij}^{(k)}$, $k = 1, 2$, are

$$\begin{array}{cc} (p_{ij}^{(1)}) & (p_{ij}^{(2)}) \\ j = 1, \quad j = 2 & j = 1 \quad j = 2 \\ i = 1 \left(\begin{array}{cc} 1/3 & 2/3 \end{array} \right), & i = 1 \left(\begin{array}{cc} 0 & 1 \end{array} \right) \\ i = 2 \left(\begin{array}{cc} 3/4 & 1/4 \end{array} \right), & i = 2 \left(\begin{array}{cc} 1 & 0 \end{array} \right). \end{array}$$

Under f_1 , for instance, $\Pr[x_0 \in X_2, x_1 \in X_1, x_2 \in X_2]$ can be calculated indifferently as $\mu(X_{212}) = \mu[x: \frac{5}{6} \leq x \leq 1] = \frac{1}{6}$ or as $p_2 \cdot p_{21} \cdot p_{12} = \frac{2}{3} \cdot \frac{3}{4} \cdot \frac{1}{3} = \frac{1}{6}$.

2. The scheme considered in § 1 can be extended to the case in which the motion of the phase point is governed by

$$(24) \quad x_{n+1} = f(x_n, \omega), \quad n = 0, 1, 2, \dots$$

In view of the intended applications (to be indicated in § 3) it will be supposed that $\omega \in \Omega \equiv \{\omega_1, \omega_2, \dots, \omega_K\}$, $K < \infty$. The situation to be discussed is of the following nature: an "event" $\omega_k \in \Omega$ occurs (with known probability) and, in consequence, a mapping $f_k: X \rightarrow X$ is determined, i.e., a rule is selected for describing the manner in which the phase of the system is to change. Clearly, one has to consider infinite sequences of "events" belonging to Ω and, hence, infinite sequences of mappings belonging to $\{f_k\}_{k=1,2,\dots,K}$. It is equally clear that the f_k are to be well-behaved (as in § 1) in the sense that if X_{ij}^k is nonempty, then

$$(25) \quad f_k(X_{ij}^k) \stackrel{\text{a.e.}}{=} X_j, \quad f_k(x) = \alpha_{ij}^k x \quad \text{for } x \in X_{ij}^k.$$

The conditional probability,

$$(26) \quad q(x, \omega_k) \equiv \Pr [\omega = \omega_k | x \in X],$$

will be the subject of the discussions to follow.

2.1. It will be shown, first, that if each $f_k(\cdot)$ defined by (25) fulfills the conditions of Theorem 2, and if the value of $q(x, \omega_k)$ depends only on the index k , then the motion in the state space has, again, the Markov property.

THEOREM 3. *Let the motion of a phase point be governed by equation (24), in which*

$$\Pr [f(\cdot) = f_k(\cdot)] = q(\cdot, \omega_k) \equiv q_k.$$

Suppose (employing the preceding notation) that for each k , $k = 1, 2, \dots, K$, the mapping $f_k: X \rightarrow X$ has the property that if X_{ij}^k is nonempty then

$$(a) \quad f_k(X_{ij}^k) = X_j, \text{ a.e., and}$$

$$(b) \quad f_k(x)|_{x \in X_{ij}^k} = \alpha_{ij}^k x.$$

Then $\{Z_n\}$ (as defined in (4)) is a Markov chain with an initial probability distribution given by $\Pr [Z_0 = i] = \mu\{X_i\}$ and with transition probabilities given by

$$\Pr [Z_{n+1} = j | Z_n = i] \equiv p_{ij} = \sum_k q_k p_{ij}^{(k)}.$$

Remarks. The proof will employ the following device. A probability measure, $\hat{\mu}$, will be defined on the field $B(Y)$ generated by the set, Y , of all sequences representing the motion of phase points. Then a mapping $g: Y \rightarrow Y$ will be considered such that, for $y \in Y$, the values of $g(y)$, $g(g(y))$, \dots generate the sequence of phase points which begin their motion with y . Finally, an appropriate cover $\{Y_i\}$ is constructed such that the motion on the $\{Y_i\}$ has the Markov property; the desired result is immediate upon interpreting this chain as a motion in the state space $\{X_i\}$.

Proof. Since the motion of a phase point is completely determined by the initial value and by the sequence ω', ω'', \dots of events occurring in time, a representative sequence can be denoted by a point $(x, \omega', \omega'', \dots)$, and the collection of possible sequences can be denoted by the product $Y = X \times \Omega^\infty$. Observe that according to the hypothesis, the ω 's are independent of the phase points; thus the probability of a (sample) sequence y can be defined as

$$\hat{\mu}\{dy\} = \mu\{dx\} \prod_{k=1}^{\infty} q(\omega^k).$$

Now consider the product field $B(Y) = B(X) \times B(\Omega^\infty)$ and consider, in particular, the projection A_x for fixed x and the projection A_w for fixed sequence, w , in Ω^∞ . For arbitrary $A \in B(Y)$, define

$$(27) \quad \hat{\mu}\{A\} = \sum_w \int_X X_{A_w}(x) X_{A_x}(w) \prod_{K=1}^{\infty} q(\omega^K) \mu\{dx\};$$

it is quite clear that $\hat{\mu}\{\cdot\}$ is a measure on $B(Y)$.

Next, consider a system of covers $\{Y_i\}$ on Y such that $Y_i \equiv X_i \times \Omega^\infty$, $i = 1, 2, \dots, N$, and let $g(\cdot)$ be a mapping from Y to Y such that

$$(28) \quad g(x, \omega^{k_1}, \omega^{k_2}, \dots) = (f_{k_1}(x), \omega^{k_2}, \omega^{k_3}, \dots).$$

Define the following symbols, namely,

$$Y_{i_0 i_1 \dots i_n} = \{y : y \in Y_{i_0}, g(y) \in Y_{i_1}, \dots, g^n(y) \in Y_{i_n}\},$$

$$Y_{i_0 i_1 \dots i_n}^{k_1 k_2 \dots k_n} = \{y : y \in Y_{i_0 i_1 \dots i_n}, \omega^{(1)} = \omega_{k_1}, \dots, \omega^{(n)} = \omega_{k_n}\}$$

$$V_n : \text{the set of ordered } n\text{-tuples } \mathbf{k}_n \text{ chosen from } (1, 2, \dots, K).$$

From Theorem 2 (Corollary), it is seen that

$$f_{k_1}(X_{i_0 i_1 \dots i_n}^{k_1 k_2 \dots k_n}) = X_{i_1 i_2 \dots i_n}^{k_2 k_3 \dots k_n}$$

almost everywhere; therefore, for any n, \mathbf{k}_n and \mathbf{i}_n it follows that

$$g(Y_{i_0 i_1 \dots i_n}^{k_1 k_2 \dots k_n}) \stackrel{\text{a.e.}}{=} Y_{i_1 i_2 \dots i_n}^{k_2 k_3 \dots k_n}.$$

The definition (27) gives

$$(29) \quad \hat{\mu}\{Y_{i_0 \dots i_n}^{k_1 \dots k_n}\} = \frac{q(\omega_{k_1})}{\alpha_{i_0 i_1}^{k_1}} \hat{\mu}\{Y_{i_1 i_2 \dots i_n}^{k_2 k_3 \dots k_n}\}.$$

Making use of the fact that $\hat{\mu}\{Y_{i_0 \dots i_n}\} = \sum_{\mathbf{k}_n} \hat{\mu}\{Y_{i_n}^{\mathbf{k}_n}\}$, it is seen that

$$(30) \quad \hat{\mu}\{Y_{i_0 \dots i_n}\} = \sum_{k=1}^K \frac{q(\omega_k)}{\alpha_{i_0 i_1}^{k_1}} \hat{\mu}\{Y_{i_1 \dots i_n}\}.$$

But note that

$$1/\alpha_{i_0 i_1}^{k_1} = \frac{\mu\{X_{i_0 i_1}^k\}}{\mu\{X_{i_0}\}}, \quad \sum_{k=1}^K q(\omega_k) \frac{\mu\{X_{i_0 i_0}^k\}}{\mu\{X_{i_0}\}} = \frac{\hat{\mu}\{Y_{i_0 i_1}\}}{\hat{\mu}\{Y_{i_0}\}}.$$

Hence an application of Theorem 1 (Corollary) shows that the function $g(\cdot)$ induces a Markov chain on the cover $\{Y_i\}_{i=1,2,\dots,N}$, and that the chain has an initial probability distribution

$$p_i = \hat{\mu}\{Y_i\}$$

and transition probabilities

$$p_{ij} = \hat{\mu}\{Y_{ij}\}/\hat{\mu}\{Y_i\}.$$

From the definition of $Y_i (\equiv X_i \times \Omega^\infty)$ it follows that

$$\Pr[Z_0 = i] = \hat{\mu}\{Y_i\} = p_i$$

and that

$$\begin{aligned} \Pr[Z_{n+1} = j | Z_n = i] &= \Pr[y_{n+1} \in X_j | y_n \in X_i] \\ &= \sum_{k=1}^K \frac{q_k \mu\{X_{ij}^k\}}{\mu\{X_i\}} = \sum_{k=1}^K q_k p_{ij}^{(k)}. \end{aligned}$$

This shows that $\{Z_n\}_{n=0,1,2,\dots}$ is a Markov chain with the required transition probabilities. Q.E.D.

2.2. It is now an easy step to consider the case in which the occurrence of an "event" ω_k may depend on the state of the system. This is the case when, for instance, following each observation of the system, the observer takes some action which affects the probabilities with which events of certain types can occur.³ Here, again, the motion in state space has the Markov property; this is the contention of the following theorem which can be proved routinely along the steps employed in proving Theorem 3.

THEOREM 4. *Let $q: X \times S \rightarrow [0, 1]$ be a transition probability such that $q(x, \omega_k) \equiv q_{ik}$ for all $x \in X_i$. Let the motion of phase points be given by*

$$x_{n+1} = f(x_n), \quad n = 0, 1, 2, \dots,$$

in which the function $f: X \rightarrow X$ coincides, with probability q_{ik} , with the function $f_k: X \rightarrow X$ for all $x \in X_i$. Suppose each $f_k(\cdot)$ satisfies the conditions of Theorem 2. Then $\{Z_n\}$ is a Markov chain having an initial probability distribution given by $\Pr[X_0 = i] = \mu\{X_{i0}\}$ and transition probabilities given by

$$(31) \quad \Pr[Z_{n+1} = i | Z_n = j] = \sum_{k=1}^K q_{ik} p_{ij}^{(k)}.$$

Proof. The proof is omitted.

3. The aim of this concluding section is to indicate a class of applications of which the notions have been presented in §§ 1–2. These applications relate to stochastic systems in which a decision-maker applies some (control) action contingent upon the observed state of the system.⁴ Here, again, one should distinguish between two classes of situations: one in which the choice of action is completely determined by the observed state; and the other in which this is not so. In the former case, then, one can speak of a *pure policy*, i.e., a mapping $h: X \rightarrow \Psi$, Ψ being the set of all possible (available) acts. For the sake of generality and practical considerations one may suppose that

$$(32) \quad \Psi = \bigcup_{i=1}^M \{\Psi_i\},$$

each subset $\{\Psi_i\}$ consisting of all acts which produce an identical effect (in terms of transition probabilities) on the motion of the system. Thus it will be legitimate to speak of an element of $\{\Psi_i\}$ as an *act* and, therefore, to visualize the function h as taking on the same value on all the elements of a given $\{\Psi_i\}$.

³ For the sake of concreteness, one may visualize a queuing system and suppose that when the queue is short, the observer (who is also the decision-maker) announces to all arriving customers that the expected waiting time would be very short or, perhaps, that the price of service would be reduced and, analogously, announce long expected waiting time when the queue is seen to be long. In all likelihood, such action will affect the probabilities of transitions between states.

⁴ Feldbaum [1, pp. 202–216, 217–237, 382–397] and Howard [4], among others, have considered the control of systems whose motion is assumed to have the Markov property. The characterization of the class of cases in which this assumption fits reality is new.

3.1. Observe, first, that an act causes *some* event ω to happen in the immediately following interval. Suppose, now, that the conditional probabilities

$$q_{mk} \equiv \Pr[\omega = \omega_k | \Psi_m], \quad m = 1, 2, \dots, M, \quad k = 1, 2, \dots, K,$$

are given. It can be shown that due to a pure policy, the motion of the system in the state space $\{X_i\}$ and the (induced) motion in the action space $\{\Psi_i\}$ both have the Markov property; this is the contention of the following theorem.

THEOREM 5. Let $\Psi = \{\psi : 0 \leq \psi \leq 1\} = \bigcup_{m=1}^M \Psi_m$, and let $\sigma(\psi)$ be the Borel field of $\{\psi_m\}$. Suppose $q: X \times \sigma(\psi) \rightarrow [0, 1]$ is a transition probability such that for each k , $q(\psi, \omega_k) = q_{mk}$ for all $\psi \in \Psi_m$, and suppose that in the equation

$$x_{n+1} = f(x_n), \quad n = 0, 1, \dots,$$

of motion of phase points, $f(\cdot)$ takes on the value $f_k(\cdot)$ with a probability $q(\psi, \omega_k)$. Assume each $f_k(\cdot)$ satisfies the conditions of Theorem 2. Then $\{Z_n\}_{n=0,1,\dots}$ is a Markov chain with initial probabilities $\Pr[Z_0 = i] = \mu\{X_i\}$ and transition probabilities

$$(33) \quad \Pr[Z_{n+1} = j | Z_n = i] = \sum_{k=1}^K \sum_{m=1}^M v_{im}^h q_{mk} p_{ij}^{(k)}$$

in which

$$(34) \quad v_{im}^h \equiv \begin{cases} 0 & \text{if } x \in X_i \text{ \& } h(x) \notin \Psi_m, \\ 1 & \text{if } x \in X_i \text{ \& } h(x) \in \Psi_m. \end{cases}$$

Proof. Being piecewise constant, the function $h: X \rightarrow \Psi$ is $\sigma(X)$ -measurable and, hence, $\mu\{x: x \in X_i, h(x) \in \Psi_m\}$ is well-defined for all $i, 1 \leq i \leq N$, and $m, 1 \leq m \leq M$. Also,

$$\Pr[h(x) \in \Psi | x \in X_i] = \frac{\mu\{x: x \in X_i, h(x) \in \Psi_m\}}{\mu\{x: x \in X_i\}} \equiv v_{im}^h.$$

According to the law of total probability, $\sum_m v_{im}^h q_{mk}$ must represent the probability $\Pr[\omega = \omega_k | x \in X_i]$. Therefore, the motion of the phase point is isomorphic to the one appearing in the hypothesis of Theorem 4; hence $\{Z_n\}_{n=0,1,2,\dots}$ is a Markov chain with the stated properties. Q.E.D.

3.2.1. Next, consider the second of the two classes of situations mentioned in the introduction to this section. Here the action taken is contingent on the observed state of the system *and* on the output of some arbitrary random device. Interest in this class of situations may be attributed to the following practical consideration. If the choice of policy is to be restricted to one of the $(M)^N$ pure ones, then at most $(M)^N$ levels of system performance can be attained, although the “desired” level of performance may not coincide with any of these attainable ones; hence, it is meaningful to inquire if a larger number of attainable performance levels could be secured by some procedure analogous to that of “randomization” or “mixing”.

3.2.2. In order to lend some measure of concreteness to this inquiry, it will be agreed that a policy is to be viewed, henceforth, as any element of the *space*

H of the functions $h(\cdot)$ representing the pure policies, together with all possible convex linear combinations thereof. The *pure* policies have already been seen (Theorem 5) to generate sequences of states which have the Markovian property. The aim of the following theorem is to examine if this is true for *all* elements of H .

THEOREM 6. Let $\{h_t: X \rightarrow \Psi\}$, $t = 1, 2, \dots, T$, be a set of $(M)^N$ simple functions, each taking on a constant value on each X_i . For any t , define

$$(35) \quad v_{im}^t = \begin{cases} 0 & \text{if } x \in X_i \text{ \& } h_t(x) \notin \Psi_m, \\ 1 & \text{if } x \in X_i \text{ \& } h_t(x) \in \Psi_m, \end{cases}$$

and assume that for any two distinct indices t and t' there exist at least one pair of indices (i, m) such that $v_{im}^t \neq v_{im}^{t'}$. Let $\boldsymbol{\eta} = (\eta_1, \eta_2, \dots, \eta_T)$ be a T -component probability vector, and let

$$h: X \rightarrow \Psi$$

be a function whose values coincide with those of h_t with a probability of η_t . If the motion of a phase point is governed by the equation

$$x_{n+1} = f(x_n), \quad n = 0, 1, 2, \dots,$$

in which $f(\cdot)$ is $f_k(\cdot)$ with a probability $q(h(x), \omega_k)$, and each $f_k(\cdot)$ satisfies the conditions of Theorem 2, then the sequence $\{Z_n\}$ is a Markov chain for which the initial probabilities are $\Pr[Z_0 = i] = \mu\{X_i\}$ and the transition probabilities are

$$\Pr[Z_{n+1} = j | Z_n = i] = \sum_{t=1}^T \eta_t \sum_{k=1}^K \sum_{m=1}^M v_{im}^t q_{mk} p_{ij}^{(k)}.$$

Proof. Applying the law of total probability twice in succession we see that

$$(36) \quad \Pr[\omega = \omega_k | x_n \in X_i] \equiv q_{ik}^* \equiv \left(\sum_{t=1}^T \eta_t v_{im}^t \right) q_{mk}.$$

Thus the motion of phase points is governed by the equation

$$x_{n+1} = f(x_n), \quad n = 0, 1, 2, \dots,$$

in which $f(\cdot)$ coincides with $f_k(\cdot)$ with a probability q_{ik}^* whenever $x_n \in X_i$. Consequently, the motion of phase points is isomorphic to the one in the hypothesis of Theorem 3 and, therefore, $\{Z_n\}$ is a Markov chain with the stated transition probabilities. Q.E.D.

COROLLARY 6.1. Let $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots, \gamma_T)$ be a probability vector, and let $P(\cdot, \boldsymbol{\gamma})$ denote the $N \times N$ matrix of transition probabilities obtained when $h: X \rightarrow \Psi$ coincides with h_t with a probability of γ_t . Let $\{\boldsymbol{\eta}^{(i)}\}$, $i = 1, 2, \dots$, be a set of T -component probability vectors and let

$$\boldsymbol{\eta} = \sum_{i=1}^{\infty} \lambda_i \boldsymbol{\eta}^{(i)}$$

be an arbitrary convex combination of the elements of $\{\boldsymbol{\eta}^{(i)}\}$. Then

$$P(\cdot, \boldsymbol{\eta}) = \sum_{i=1}^{\infty} \lambda_i P(\cdot, \boldsymbol{\eta}^{(i)}).$$

To sum up: if a policy is constructed by taking a convex combination of the elements of any set of policies, then the transition probabilities due to the convex combination are the convex combinations of the transition probabilities associated with the individual component policies.

REFERENCES

- [1] A. A. FELDBAUM, *Optimal Control Systems*, Academic Press, New York, 1965.
- [2] W. FELLER, *An Introduction to Probability Theory and its Applications*, vol. 1, John Wiley, New York, 1966.
- [3] I. I. GIKHMAN AND A. V. SKOROKHOD, *Introduction To The Theory of Random Processes*, W. B. Saunders, Philadelphia, 1965.
- [4] R. A. HOWARD, *Dynamic Programming and Markov Processes*, Technology Press and Wiley Press, New York, 1960.
- [5] R. E. KALMAN, *Nonlinear aspects of sampled data control systems*, Proc. Symposium on Non-Linear Circuit Analysis, Polytechnic Institute of Brooklyn, 1956, pp. 273–315.
- [6] YU. V. PROHOROV AND YU. A. ROXANOV, *Probability Theory*, Springer-Verlag, New York, 1969, pp. 293–301.

A SEMIGROUP REPRESENTATION OF THE MAXIMUM EXPECTED REWARD VECTOR IN CONTINUOUS PARAMETER MARKOV DECISION THEORY*

STANLEY R. PLISKA†

Abstract. The maximum expected reward vector that arises in continuous parameter Markov decision problems is frequently characterized as the unique solution of a certain Cauchy problem. This paper generalizes this characterization by viewing the maximum expected reward vector as a nonlinear semigroup in an appropriate Banach space. This perspective has several advantages. First, the semigroup may exist even though the corresponding Cauchy problem does not have a solution. Second, this approach is often useful in showing when the Cauchy problem does have a solution. Third, these methods are useful in the study of the method of successive approximations. Finally, these methods appear likely to unify some diverse results in Markov decision theory.

The results in this paper are very general. First, sufficient conditions are given for the semigroup to exist. The discounted reward case is studied next; a certain operator is shown to have a unique singular point that is the strong limit of the semigroup as the parameter $t \rightarrow \infty$. The asymptotic properties of the semigroup in the case of undiscounted rewards are studied with the aid of some fixed point theorems for monotone and nonexpansive operators; the transient, positive, negative and optimal stopping cases are studied in this context. The paper concludes with two examples. The first is a controlled diffusion process on a compact interval of the real line. The second is a controlled jump process with general state and action spaces.

1. Introduction. Continuous parameter Markov decision theory has, in general, proceeded along two lines. On the one hand, Miller [16], [17], Kakumanu [12], Stone [22], Pliska [20] and others have considered controlled jump process (including, specifically, controlled Markov chains). On the other hand, Mandl [15], Pliska [19], Puterman [21], Fleming [10] and numerous others have considered controlled diffusion processes. Both finite and infinite planning horizon problems were considered in both cases. The purpose of this paper is to make a preliminary attempt at a general theory of continuous parameter Markov decision theory by studying the maximum expected reward vector.

Let S be the state space and let A be the set of admissible actions. Let X be a Banach space of real-valued functions on S . Let \mathcal{T} be a set of linear transformations from a subset \mathcal{D} of X to X . Throughout this paper, it is assumed that each $\mathcal{A} \in \mathcal{T}$ is the infinitesimal generator of a Markov process on S . By the Hille–Yosida–Phillips theorem, this will be the case with $\mathcal{A} \in \mathcal{T}$, for example, if \mathcal{D} is dense in X and the equation $\lambda v - \mathcal{A}v = g$ has a unique solution $v \in \mathcal{D}$ with $\|\lambda v - \mathcal{A}v\| \geq \|\lambda v\|$ for every $\lambda > 0$ and every $g \in X$.

Let $\mathcal{A}(\cdot)$ be a mapping from A into \mathcal{T} . Let r be a real-valued function on $S \times A$. The value $r(s, a)$ can be interpreted as the reward rate for being in state $s \in S$ while choosing action $a \in A$. Consider the mapping $\phi: D(\phi) \rightarrow R(\phi)$ defined for each $s \in S$ and $v \in D(\phi)$ by

$$(1) \quad \phi v(s) = \sup_{a \in A} \{r(s, a) + \mathcal{A}(a)v(s)\},$$

* Received by the editors May 14, 1974, and in revised form September 29, 1974.

† Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois 60201.

where $D(\phi) \subset \mathcal{D}$ and $R(\phi)$ is some space of real-valued functions on S . The study of the operator ϕ and its properties is one of the central topics of this paper. As will be seen later, ϕ need not be linear or even continuous.

The operator ϕ allows one to define the Cauchy problem

$$(2) \quad \frac{du}{dt} = \phi u, \quad u(0) = x \in D(\phi).$$

A function $u: [0, \tau] \rightarrow X$ is said to be a (strong) solution of (2) on the interval $[0, \tau]$ if u is Lipschitz continuous on compact subsets of $[0, \tau]$, u is strongly differentiable almost everywhere on $[0, \tau]$, and (2) is satisfied almost everywhere.

In several of the papers cited at the beginning of this paper, a Cauchy problem that is a special case of (2) is shown to have a unique solution that equals the maximum expected reward vector. In other words, a typical result is that if u is the solution of an equation like (2), then $u(s, t)$ is the supremum of the expected future rewards over a suitable class of deterministic, memoryless, nonstationary policies given the process is presently in state s with time t remaining until the planning horizon when the terminal reward x is received. Thus the study of the general Cauchy problem (2) would seem germane to a general theory of continuous parameter Markov decision theory.

According to Crandall and Liggett [6], if ϕ satisfies certain assumptions and if u is a solution of (2), then by setting $T(t)x = u(t)$, one obtains a semigroup T whose infinitesimal generator is ϕ . This semigroup was mentioned by Veinott [23, § 5] in connection with finite state continuous parameter Markov decision chains. The subject of this paper is the study of nonlinear semigroups whose infinitesimal generators are of the same form as ϕ in (1). The key result in this regard is Theorem 1 in § 2. An adaptation of Crandall and Liggett's [6] results, this theorem gives sufficient conditions for the existence of—and characterizes—the semigroup T . Theorem 1 is followed by a discussion of these conditions, and it is shown that, with one additional assumption, $T(t)x$ is greater than or equal to the expected reward corresponding to any piecewise constant policy.

There are at least three reasons for studying the semigroup T rather than the solution of the Cauchy problem (2). First, although the semigroup T of Theorem 1 might exist, the function $u(t) = T(t)x$ need not be differentiable anywhere, and (2) need not have any solution. Section 2 concludes with an example of such a situation where the maximum expected reward is indeed given by T .

The second reason for this study is that semigroup methods are frequently useful to show that there does exist a solution to the Cauchy problem (2). The most common differential equation existence methods are along the lines of showing ϕ is a Lipschitz continuous map on some Banach space into itself. For some Markov decision problems, these conditions may fail to hold even though semigroup methods can be used to demonstrate the maximum expected reward vector does satisfy (2). This will be the case if the semigroup T exists and certain extra conditions are satisfied (e.g., ϕ has a closed graph and Lipschitz continuous X -valued functions are always differentiable almost everywhere, according to Crandall and Liggett [6]). Section 6 provides a Markov decision problem example of this point.

The third reason for the study of the semigroup T is to investigate the asymptotic properties of the maximum expected reward vector as the time horizon

increases to infinity. This investigation is carried out in §§ 3 and 4, where the discounted and undiscounted cases are examined, respectively. The emphasis is on the study of when $T(t)x$ converges to a singular point of the operator ϕ as $t \rightarrow \infty$. This behavior is analogous to the method of successive approximations in discrete time Markov decision theory.

In the case of discounted rewards, let ϕ be the infinitesimal generator in the case of undiscounted rewards, and let $\alpha > 0$ be the discount factor. Then the semigroup \tilde{T} corresponding to the infinitesimal generator $\tilde{\phi} = \phi - \alpha I$ will be the maximum expected discounted reward with $\tilde{T}(t) = e^{-\alpha t} T(t)$. Moreover, there exists a unique singular point, say z of $\tilde{\phi}$ such that $\tilde{T}(t)x \rightarrow z$ strongly as $t \rightarrow \infty$ for all $x \in \overline{D(\phi)}$, and z equals the maximum expected discounted infinite horizon reward.

In the case of undiscounted rewards, we first provide a monotonicity condition which ensures that $\lim_{t \rightarrow \infty} T(t)x$ exists. Recognizing that $T(t)$ is typically a non-expansive operator, we study the fixed point properties of $T(t)$ in uniformly convex spaces and obtain a condition that ensures there exists a singular point of ϕ . With additional assumptions we show $T(t)x$ converges weakly to a singular point of ϕ . The application of these results to transient, positive, negative and optimal stopping decision problems is explained.

The final two sections provide two examples; one is a controlled jump process and the other is a controlled diffusion process. The main emphasis is on showing that the operator ϕ satisfies the various requirements developed in this paper. For the most part, these examples are generalizations of the problems considered in the literature that is cited at the beginning of this paper.

2. The semigroup representation. It is appropriate to precisely define some notions pertaining to semigroups that will be used in this paper. Following Crandall and Liggett [6], if $C \subset X$, a semigroup on C is a function T on $[0, \infty)$ such that $T(t)$ maps C into C for each $t \geq 0$ and satisfies

$$\begin{aligned} T(t + \tau) &= T(t)T(\tau) & \text{for } t, \tau \geq 0, \\ \lim_{t \downarrow 0} T(t)x &= T(0)x = x & \text{for } x \in C. \end{aligned}$$

In addition, the notation $T \in Q_w(C)$ will mean that T is a semigroup on C and w is a real number such that

$$(3) \quad \|T(t)x - T(t)y\| \leq e^{wt} \|x - y\|$$

for all $t \geq 0$ and $x, y \in C$.

If Ψ is a function on a subset of X into X , let $D(\Psi)$ and $R(\Psi)$ denote the domain and range of Ψ , respectively. Let I be the identity map on X . The function Ψ is called *dissipative* if, for $u, v \in D(\Psi)$ and all $\varepsilon \geq 0$, then $\|(I - \varepsilon\Psi)u - (I - \varepsilon\Psi)v\| \geq \|u - v\|$. Notice that if Ψ is dissipative, then $(I - \varepsilon\Psi)$ is 1-1 on $D(\Psi)$ and $(I - \varepsilon\Psi)^{-1}$ is nonexpansive on $R(I - \varepsilon\Psi)$, that is, if $u, v \in R(I - \varepsilon\Psi)$, then $\|(I - \varepsilon\Psi)^{-1}u - (I - \varepsilon\Psi)^{-1}v\| \leq \|u - v\|$.

According to Crandall and Liggett [6, Thm. 1], it is necessary to make some assumptions on ϕ in order to guarantee the existence of a corresponding semigroup. The appropriate results are summarized in the following.

THEOREM 1. Suppose ϕ satisfies the following three assumptions:

- (I) $R(\phi) \subset X$.
- (II) $\phi - wI$ is dissipative for some real number w .
- (III) $R(I - \varepsilon\phi) \supset \overline{D(\phi)}$ for all sufficiently small positive ε .

Then

$$(4) \quad \lim_{n \rightarrow \infty} \left(I - \frac{t}{n} \phi \right)^{-n} x$$

exists for each $x \in \overline{D(\phi)}$ and $t > 0$. Moreover, if $T(t)x$ is defined as the limit in (4), then $T \in Q_w(\overline{D(\phi)})$.

Assumptions (I)–(III) are satisfied by many Markov decision problems. Moreover, they are usually easy to check. See the last two sections of this paper for two examples. For most Markov decision problems with undiscounted rewards, assumption (II) will be true with $w = 0$.

Assumption (I) is technical in nature. Throughout much of the literature cited at the beginning of this paper, various restrictions are placed on S , A , r , etc., in order to guarantee that (I) is satisfied.

To verify (III), it suffices to show $(I - \varepsilon\phi)^{-1}x \in D(\phi)$ for each $x \in \overline{D(\phi)}$. An important interpretation of $(I - \varepsilon\phi)^{-1}x$ is frequently useful in the verification of this condition as well as assumption (II) with $w = 0$. Consider the equation $(I - \varepsilon\phi)v = x$. Dividing through by ε and rearranging terms, one obtains

$$(5) \quad \varepsilon^{-1}v(s) - \sup_{a \in A} \{r(s, a) + \varepsilon^{-1}x(s) + \mathcal{A}(a)v(s)\} = 0.$$

Upon comparing (5) with various results in the Markov decision theory literature, it becomes apparent that in a variety of circumstances one may interpret $(I - \varepsilon\phi)^{-1}x$ as the maximum (over some class of admissible stationary policies) expected infinite horizon discounted reward corresponding to reward rate $r + \varepsilon^{-1}x$ and discount factor ε^{-1} .

To show the connection between the assumptions and this interpretation, it is necessary to be more specific. Let M denote a class of stationary policies, that is, each $f \in M$ is a function on S into A such that $\mathcal{A}(f)$ is the infinitesimal generator, with domain $D(\phi)$, of a stationary Markov process. By the theory of resolvent operators, $(I - \varepsilon\mathcal{A}(f))^{-1}x \in D(\phi)$ for each $f \in M$, $x \in \overline{D(\phi)}$, and $\varepsilon > 0$.

From now until Proposition 2, let $x \in \overline{D(\phi)}$ and $\varepsilon > 0$ be fixed and arbitrary. If $r(f) \equiv r(\cdot, f(\cdot)) \in \overline{D(\phi)}$, then there exists a unique $v_f \in D(\phi)$ such that

$$v_f - \varepsilon(r(f) + \mathcal{A}(f)v_f) = x,$$

and v_f may be interpreted as the expected discounted reward under f corresponding to reward rate $r(f) + \varepsilon^{-1}x$ and discount factor ε^{-1} . In many situations, a solution exists with $v_f \in D(\phi)$ for all $f \in M$, and also one has $\sup_{f \in M} v_f = (I - \varepsilon\phi)^{-1}x$. Thus assumption (III) is true whenever $\sup_{f \in M} v_f \in D(\phi)$; in particular, it is true if there exists an optimal policy, that is, some $f \in M$ such that $v_f = \sup_{f \in M} v_f$.

To show the connection between this interpretation and assumption (II) with $w = 0$, recall that for $(I - \varepsilon\phi)^{-1}x$ to be the maximum expected discounted reward means (i) $(I - \varepsilon\phi)^{-1}x$ is greater than or equal to the expected discounted reward corresponding to any $f \in M$ and (ii) for each $s \in S$ there exists some sequence $\{f_n\}$ of policies in M such that the corresponding sequence of expected

discounted rewards evaluated at s converges to $(I - \varepsilon\phi)^{-1}x(s)$ as $n \rightarrow \infty$. These ideas lead to the following.

PROPOSITION 2. *Suppose the supremum norm is used, and for each $x \in R(I - \varepsilon\phi)$, $(I - \varepsilon\phi)^{-1}x$ can be interpreted as the maximum expected discounted reward over some class of stationary policies and corresponding to reward rate $r + \varepsilon^{-1}x$ and discount factor ε^{-1} . Then assumption (II) is satisfied with $w = 0$.*

Proof. Let $\varepsilon > 0$, $\delta > 0$, $u, v \in R(I - \varepsilon\phi)$ and $s \in S$ all be arbitrary. By hypothesis, the inverse exists, so without loss of generality assume $(I - \varepsilon\phi)^{-1}u(s) - (I - \varepsilon\phi)^{-1}v(s) \geq 0$. Then it suffices to show

$$(6) \quad (I - \varepsilon\phi)^{-1}u(s) - (I - \varepsilon\phi)^{-1}v(s) \leq \|u - v\| + \delta.$$

Let $W_x(s, f)$ denote the expected discounted reward corresponding to some policy f and reward rate $r(f) + \varepsilon^{-1}x$. By hypothesis, there exists some policy f such that $(I - \varepsilon\phi)^{-1}u(s) - W_u(s, f) < \delta$. If s_t denotes the sample path of the process, then by definition,

$$W_u(s, f) = E_s \int_0^\infty e^{-t/\varepsilon} r(s_t, f(s_t)) dt + E_s \int_0^\infty e^{-t/\varepsilon} \varepsilon^{-1} u(s_t) dt.$$

Since $W_v(s, f)$ is defined similarly, it follows that

$$\begin{aligned} W_u(s, f) - W_v(s, f) &\leq E_s \int_0^\infty e^{-t/\varepsilon} \varepsilon^{-1} (u(s_t) - v(s_t)) dt \\ &\leq E_s \int_0^\infty e^{-t/\varepsilon} \varepsilon^{-1} \|u - v\| dt \leq \|u - v\|. \end{aligned}$$

Thus (6) follows by the choice of f and because $W_v(s, f) \leq (I - \varepsilon\phi)^{-1}v(s)$.

It would be nice to show that, in a very general way, $T(t)x$ is indeed the maximum expected reward with time t remaining until the planning horizon. In other words, one would like to specify some general class of policies such that each policy corresponds to a Markov process, the expected reward corresponding to each policy is less than or equal to $T(t)x$, and there exists a sequence of policies whose corresponding expected rewards converge in an appropriate sense to $T(t)x$. If $T(t)x$ does satisfy the Cauchy problem, then it is usually possible to show that $T(t)x$ is indeed the maximum expected reward; for example, this was done in Pliska [20] for a problem with a controlled jump process (see also the last section of this paper).

In the more general case, where $T(t)x$ does not necessarily satisfy the Cauchy problem, only partial results have been obtained, namely, if the hypotheses of Proposition 2 hold, then $T(t)x$ is greater than or equal to the expected reward corresponding to any admissible piecewise constant policy. To see this, let f denote a policy that is admissible as far as the interpretation of $(I - \varepsilon\phi)^{-1}x$ is concerned. If the problem has been formulated appropriately, then one can define the function $\Psi: D(\phi) \rightarrow X$ by $\Psi v = r(f) + \mathcal{A}(f)v$ and Ψ will satisfy assumptions (I)–(III). The corresponding semigroup, say U , exists by Theorem 1, and $U(t)x$ can be interpreted as the expected reward under f with time t remaining until the terminal reward x is received. Moreover, $(I - \varepsilon\Psi)^{-1}x$ can be interpreted as the expected

infinite horizon discounted reward under f with reward rate $r(f) + \varepsilon^{-1}x$ and discount factor ε^{-1} . By hypothesis, $(I - \varepsilon\phi)^{-1}x \geq (I - \varepsilon\Psi)^{-1}x$ for each $x \in \overline{D(\phi)}$. Moreover, if $x \geq y$, then $(I - \varepsilon\Psi)^{-1}x \geq (I - \varepsilon\Psi)^{-1}y$. Consequently, for each fixed t , $(I - (t/n)\phi)^{-n}x \geq (I - (t/n)\Psi)^{-n}x$ and $T(t)x \geq U(t)x$. A simple argument suffices to extend this result to the case of piecewise constant policies.

We conclude this section with an example of a Markov decision process whose maximum expected reward vector is specified by the semigroup of Theorem 1 in spite of the fact that the maximum expected reward vector does not satisfy the Cauchy problem (2). One might suppose an example of this kind might need to be rather complicated, but, quite the contrary, we shall consider a "deterministic" Markov process for which there is no choice of controls.

Let $S = [0, 1]$, d be a positive number and s_t denote the sample path of the process. For any initial state s_0 we have $s_t = s_0 - td$ for $t \leq s_0/d$. The boundary $s = 0$ is absorbing, i.e., the process terminates when it arrives there. The infinitesimal generator of this process is specified by $\mathcal{A}v = -dv'$ and has domain $D(\phi)$ equal to all bounded continuous functions on $[0, 1]$ which equal zero at $s = 0$ and which have piecewise continuous derivatives. Finally, we take $X = L_\infty([0, 1])$ and r constant.

Clearly assumption (I) is true. Now $\overline{D(\phi)}$ consists of all bounded continuous functions on $[0, 1]$ which equal zero at $s = 0$, so by the theory of ordinary differential equations, $(I - \varepsilon\phi)^{-1}x \in D(\phi)$ for each $x \in \overline{D(\phi)}$ and assumption (III) is true. Assumption (II) is true for $w = 0$ (as the reader can verify), so the semigroup T of Theorem 1 exists.

For the terminal reward we shall take $x = 0$. If $v(t, s)$ denotes the expected reward when starting in state s with time horizon t , then we clearly have

$$v(t, s) = \begin{cases} tr, & s \geq td, \\ rs/d, & s \leq td. \end{cases}$$

Note that $v(t, \cdot)$ is not strongly differentiable at any $t \leq d^{-1}$. In fact, neither the right- nor left-hand strong derivatives exist there, either. Hence $v(t, \cdot)$ cannot satisfy the Cauchy problem (2).

To complete this example, it remains to show that, indeed, $T(t)x = v(t, \cdot)$. We do this by computing (4). The reader can verify that

$$\left(I - \frac{t}{n}\right)^{-n} x = tr - tr e^{-nk} \left[\sum_{i=0}^{n-1} \frac{(nk)^i}{i!} - k \sum_{i=0}^{n-2} \frac{(nk)^i}{i!} \right],$$

where $k = s/td$. Letting $n \rightarrow \infty$, and using the identity

$$\lim_{n \rightarrow \infty} e^{-nk} \sum_{i=0}^n \frac{(kn)^i}{i!} = \begin{cases} 0, & k > 1, \\ \frac{1}{2}, & k = 1, \\ 1, & 0 \leq k < 1, \end{cases}$$

(see Abramowitz and Stegun [1, p. 263]), we obtain the desired result.

3. The discounted case. The discounted case is easy to dispose of. Our objective is to show there exists a unique singular point, say z , of a certain operator $\tilde{\phi}$ such that the maximum expected discounted reward $\tilde{T}(t)x \rightarrow z$ for all $x \in \overline{D(\tilde{\phi})}$ as

$t \rightarrow \infty$. This result is analogous to the method of successive approximations as used in discrete time dynamic programming.

Throughout this section, assume ϕ satisfies (I)–(III) with $w = 0$. Suppose the rewards are discounted at the constant rate $\alpha > 0$. If $T(t)x$ is the maximum expected undiscounted reward for some problem, then $\tilde{\phi} = \phi - \alpha I$ will be the infinitesimal generator of a semigroup \tilde{T} such that $\tilde{T}(t)x$ will equal the maximum expected discounted reward. In particular, $\tilde{\phi}$ will satisfy (I)–(III) with $\tilde{w} = -\alpha$, $D(\phi) = D(\tilde{\phi})$, and $\tilde{T}(t) = e^{-\alpha t}T(t)$.

In view of (3), for any fixed t , $\tilde{T}(t)$ will be a strict contraction. Thus, for some $z \in \overline{D(\phi)}$, $[\tilde{T}(t)]^n x$ will converge strongly to z independent of x . By the semigroup property and continuity of $\tilde{T}(t)x$ in t , z is also independent of t , so $\tilde{T}(t)x$ will converge strongly to some $z \in \overline{D(\phi)}$ as $t \rightarrow \infty$, independent of x .

In view of Crandall and Liggett [6], the function $J_\varepsilon = (I - \varepsilon\tilde{\phi})^{-1}$ is a strict contraction on $R(I - \varepsilon\tilde{\phi})$ into $\overline{D(\phi)}$ for all small enough $\varepsilon > 0$, so $J_\varepsilon^n y$ converges strongly to some $\bar{z} \in \overline{D(\phi)}$. Moreover, \bar{z} is the unique fixed point of J_ε , that is, $(I - \varepsilon\tilde{\phi})^{-1}\bar{z} = \bar{z}$. But this implies $\bar{z} \in D(\phi)$ and $\tilde{\phi}\bar{z} = 0$. In this case, $\tilde{T}(t)\bar{z} = \bar{z}$ for all $t \geq 0$ and $\bar{z} = z$. Since \bar{z} is the unique fixed point of J_ε , it must be the only point such that $\tilde{\phi}\bar{z} = 0$. Hence $\tilde{T}(t)x$ converges strongly as $t \rightarrow \infty$ to some $z \in D(\phi)$, independent of x , where z is the unique point such that $\phi z = \alpha z$.

One value of this result is in knowing that if there exists some stationary policy, say f , whose expected infinite horizon discounted reward equals z , and if $\tilde{T}(t)x$ is the maximum expected finite horizon discounted reward with respect to some class, say \tilde{M} , of possibly nonstationary policies, then f will be optimal in the infinite horizon case with respect to \tilde{M} . This is simply because the expected discounted reward corresponding to any policy in \tilde{M} will be less than or equal to $\lim_{t \rightarrow \infty} \tilde{T}(t)0$.

4. The undiscounted case. As in the case of discounted rewards, we want to know the asymptotic properties of the semigroup T . In the case of undiscounted rewards, however, it is apparently necessary to make additional assumptions about the problem in order to show $T(t)x$ converges to some \bar{x} with $\bar{x} \in D(\phi)$ and $\phi\bar{x} = 0$. One possibility is to make some kind of continuity assumption about ϕ . This usually leads to consideration of the Cauchy problem. For results along these lines, see the last section on controlled jump processes as well as Pliska [20].

A different approach will be taken here. If assumption (II) is satisfied with $w = 0$, then by (3), for each $t \geq 0$, $T(t)$ will be, by definition, a nonexpansive operator. The fixed point properties of nonexpansive operators in a Banach space have been studied extensively in the literature. In particular, Browder [3] as well as Goebel and Kirk [11] provide sufficient conditions for the existence of a fixed point, and Browder and Petryshyn [4], Petryshyn and Williamson [18], and Kirk [13] discuss when the Picard iterates of the method of successive approximations converges to a fixed point. A large portion of this literature on nonexpansive operators deals with closed, convex and bounded subsets of a uniformly convex Banach space, and this is the approach that will be taken here.

Let \leq be the natural partial order on X . An operator K on a subset D of X into itself is said to be *monotone increasing* on D if $Kx \leq Ky$ for all $x \leq y$ in D . Crucial to the results of this section is the following assumption:

(IV) $(I - \varepsilon\phi)^{-1}$ is monotone increasing on $\overline{D(\phi)}$ for all small enough $\varepsilon > 0$.

Assumption (IV) is satisfied in most Markov decision problems; indeed, it can be verified (with the same line of argument as outlined in § 2 to verify assumption (II)) immediately if $(I - \varepsilon\phi)^{-1}x$ can be interpreted as the maximum expected discounted reward over some class of stationary policies with discount factor ε^{-1} and reward rate $r + \varepsilon^{-1}x$ (one need not be restricted to the supremum norm here).

The results in this section are organized in four stages. First, if assumption (IV) holds, then $T(t)$ is monotone increasing on $\overline{D(\phi)}$ for all $t \geq 0$. Second, if in addition there exist two points $\tilde{x}, \tilde{y} \in D(\phi)$ satisfying $\tilde{x} \leq \tilde{y}$ and $\phi\tilde{x} \geq 0 \geq \phi\tilde{y}$, then $T(t)\tilde{x}$ is nondecreasing in t , $T(t)\tilde{y}$ is nonincreasing in t , and each function converges pointwise to a function on S . Third, if in addition assumption (II) is satisfied with $w = 0$ and $\overline{D(\phi)}$ is a convex subset of a uniformly convex Banach space, then there exists some $z \in [\tilde{x}, \tilde{y}] = \{x \in \overline{D(\phi)} : \tilde{x} \leq x \leq \tilde{y}\}$ satisfying $\phi z = 0$, that is, $T(t)z = z$ for all $t \geq 0$. Finally, if in addition ϕ has a closed graph and there exists a unique $z \in [\tilde{x}, \tilde{y}]$ satisfying $\phi z = 0$, then $T(t)x$ converges weakly to z as $t \rightarrow \infty$ for all $x \in [\tilde{x}, \tilde{y}]$. The application of these results to a wide variety of Markov decision problems will be explained. The following result is true for any w .

LEMMA 3. Let $T \in Q_w(\overline{D(\phi)})$ be the semigroup of Theorem 1, and suppose assumption (IV) holds. Then $T(t)$ is monotone increasing on $\overline{D(\phi)}$ for all $t \geq 0$.

Proof. Let $x \leq y$ and t be arbitrary. If $\varepsilon > 0$ is small enough, then one can show by induction that $(I - \varepsilon\phi)^{-n}x \leq (I - \varepsilon\phi)^{-n}y$ for all n . Substituting $\varepsilon = t/n$ and letting $n \rightarrow \infty$, one obtains the desired result from (4).

THEOREM 4. Let $T \in Q_w(\overline{D(\phi)})$ be the semigroup of Theorem 1, and suppose assumption (IV) holds. If $\tilde{x}, \tilde{y} \in D(\phi)$ are such that $(I - \varepsilon\phi)^{-1}\tilde{x} \geq \tilde{x}$ and $(I - \varepsilon\phi)^{-1}\tilde{y} \leq \tilde{y}$ for all small enough $\varepsilon > 0$, then $T(t)\tilde{x} : [0, \infty) \rightarrow X$ is a nondecreasing function and $T(t)\tilde{y} : [0, \infty) \rightarrow X$ is a nonincreasing function. If, in addition, $\tilde{x} \leq \tilde{y}$, then $T(t)\tilde{x}$ and $T(t)\tilde{y}$ each converge pointwise to some \bar{x} and \bar{y} , respectively, where $\tilde{x} \leq \bar{x} \leq \bar{y} \leq \tilde{y}$.

Proof. Of the two assertions of the theorem, the second is an immediate consequence of the first and Lemma 3. To prove the first assertion, only the case $(I - \varepsilon\phi)^{-1}\tilde{x} \geq \tilde{x}$ will be considered, leaving the other to the reader.

Using assumption (IV) inductively, one concludes $\tilde{x} \leq (I - \varepsilon\phi)^{-n}\tilde{x}$ for all integers n , so for any $t > 0$, $T(t)\tilde{x} = \lim_{n \rightarrow \infty} (I - (t/n)\phi)^{-n}\tilde{x} \geq \tilde{x}$. For arbitrary t and τ , $T(t + \tau)\tilde{x} = T(t)T(\tau)\tilde{x} \geq T(t)\tilde{x}$ by Lemma 3 and the semigroup property, completing the proof.

It should be remarked that if $\tilde{x} \in D(\phi)$ and $\phi\tilde{x} \geq 0$, then $\tilde{x} - \varepsilon\phi\tilde{x} \leq \tilde{x}$ and by (IV) $\tilde{x} \leq (I - \varepsilon\phi)^{-1}\tilde{x}$ for all $\varepsilon > 0$. Hence if $\tilde{x}, \tilde{y} \in D(\phi)$, then the hypotheses in Theorem 4 may be changed in an obvious manner.

It should also be remarked that there is no guarantee the limits \bar{x} and \bar{y} in Theorem 4 are elements of $D(\phi)$, $\overline{D(\phi)}$, or even X . It appears that some additional assumptions are required, for example, that X is uniformly convex.

Following Clarkson [5], a Banach space X is said to be *uniformly convex* if to each ε , $0 < \varepsilon \leq 2$, there corresponds a $\delta(\varepsilon) > 0$ such that $\|x\| = \|y\| = 1$, $\|x - y\| \geq \varepsilon$, and $x, y \in X$ imply $(1/2)\|x - y\| \leq 1 - \delta(\varepsilon)$. Euclidean spaces of all dimensions, Hilbert spaces, l_p -spaces ($1 < p < \infty$), and L_p -spaces ($1 < p < \infty$) are all uniformly convex. On the other hand, l_∞ , L_∞ , the space of continuous functions,

and the space of convergent sequences are not uniformly convex, so the implications of a uniformly convex space, unfortunately, do not apply to the important cases of controlled diffusion and jump processes (see the next two sections). Nevertheless, the notion of uniform convexity yields new results for finite state as well as other Markov decision problems.

A uniformly convex Banach space X enjoys the property that a function $u: [0, \infty) \rightarrow X$ which is Lipschitz continuous on compact subsets of $[0, \infty)$ is differentiable almost everywhere. Crandall and Liggett [6] show that with the semigroup T of Theorem 1 and with $x \in D(\phi)$, the function $T(t)x$ is Lipschitz continuous on compact subsets of $[0, \infty)$. According to another result of theirs, if $T(t)x$ is differentiable almost everywhere and ϕ has a closed graph, then $T(t)x$ satisfies the corresponding Cauchy problem (2). Putting these facts together, if X is uniformly convex, ϕ has a closed graph, T is the semigroup of Theorem 1, and $x \in D(\phi)$, then $T(t)x$ satisfies (2).

Another reason for bringing in the concept of a uniformly convex space is illustrated by the following. Note that now $w = 0$.

THEOREM 5. *Let $T \in Q_0(\overline{D(\phi)})$ be the semigroup of Theorem 1, let $\tilde{x} \leq \tilde{y}$ be as in Theorem 4, and suppose assumption (IV) holds. If X is uniformly convex and $\overline{D(\phi)}$ is convex, then there exists some $z \in [\tilde{x}, \tilde{y}] \equiv \{x \in \overline{D(\phi)}: \tilde{x} \leq x \leq \tilde{y}\}$ satisfying $z \in D(\phi)$, $\phi z = 0$, that is, for any $t > 0$, z is a fixed point of $T(t)$.*

Proof. This result is a consequence of the fixed point theorem in Browder [3]. Let $\varepsilon > 0$ be fixed. The operator $(I - \varepsilon\phi)^{-1}$ maps $[\tilde{x}, \tilde{y}]$ into itself by hypothesis. Moreover, $(I - \varepsilon\phi)^{-1}$ is nonexpansive according to Crandall and Liggett [6]. The subset $[\tilde{x}, \tilde{y}]$ is closed, bounded and convex, so by Browder's fixed point theorem, $(I - \varepsilon\phi)^{-1}z = z$ for some $z \in [\tilde{x}, \tilde{y}]$. But this implies $z = (I - \varepsilon\phi)z$, that is, $z \in D(\phi)$ and $\phi z = 0$. Hence $T(t)z = z$ for all $t \geq 0$ by Theorem 4.

Theorem 4 indicates when $T(t)x$ converges (pointwise) to some function on S , and Theorem 5 indicates when there exists a singular point of ϕ . The following theorem will combine these two ideas and provide a condition that ensures $T(t)x$ converges (weakly) to a singular point of ϕ . This result is analogous to the method of successive approximations as used in discrete time dynamic programming which states that the finite horizon maximum expected reward converges to the fixed point of the optimal return operator as the time horizon diverges to infinity.

THEOREM 6. *If, in addition to the hypotheses of Theorem 5, ϕ has a closed graph and there exists at most one singular point, say, z , of ϕ in $[\tilde{x}, \tilde{y}]$, then for each $x \in [\tilde{x}, \tilde{y}]$ the function $T(t)x$ converges weakly to z as $t \rightarrow \infty$.*

Proof. This result is an immediate consequence of Kirk [13, Thm. 3] provided $T(t)$ has a unique fixed point in $[\tilde{x}, \tilde{y}]$. We know z is a fixed point of $T(t)$, so it suffices to show there does not exist another.

Suppose $w \in [\tilde{x}, \tilde{y}]$ is a fixed point of $T(t)$. Clearly, $T(t)w$ is differentiable; indeed, $T'(t)w = 0$. Consequently, it suffices to show $T(t)w$ satisfies the Cauchy problem

$$(7) \quad u'(t) = \phi u(t), \quad u(0) = w,$$

for then $\phi w = 0$ and either $z = w$ or one has a contradiction.

According to Crandall and Liggett [6, Thm. 2], if T is as in Theorem 1, ϕ has a closed graph, $T(t)w$ is differentiable and $w \in D(\phi)$, then $T(t)w$ satisfies (7). Thus this

proof would be complete except for the fact we do not know $w \in D(\phi)$, only that $w \in \overline{D(\phi)}$. But the proof of Theorem 2 in Crandall and Liggett [6] utilizes the hypothesis $w \in D(\phi)$ only to show $T(t)w$ is Lipschitz continuous in t on bounded t -sets, and this we know, so we are, in fact, done.

The theorems in this section can be applied to a variety of Markov decision problems. The ideas here will only be sketched, since to be specific necessitates giving details about the actual problems.

For an application of Theorem 4, suppose there exists some $v \in D(\phi)$ satisfying $\phi v = 0$ which can be interpreted as the maximum expected infinite horizon reward over some class of admissible stationary policies. Suppose there exists one such policy f such that $r(f) \geq 0$ (using the same notation and formulation as in § 2). If $0 \in D(\phi)$, this implies $\phi 0 \geq 0$. Moreover, $v \geq 0$. By Theorem 4, $T(t)0$ converges upward to \bar{x} , say, with $\bar{x} \leq v$. By the interpretation of v , however, there exists some stationary policy whose expected infinite horizon reward, say w , is arbitrarily close to v . Thus $v = \bar{x}$, because $T(t)0$ is greater than or arbitrarily close to w for all large enough t .

Theorems 5 and 6 can aid in the analysis of three cases that arise frequently in the Markov decision theory literature, namely, the transient, positive and negative cases. Let M be a suitable set of stationary policies f . A problem is said to be the transient case if there exists some number N such that the expected termination time of the process given initial state s and stationary policy f is less than N for all $s \in S$ and $f \in M$. After defining $D(\phi)$, one can usually show there exists a unique $z \in D(\phi)$ satisfying $\phi z = 0$. Moreover, for any $x \in \overline{D(\phi)}$ one can usually find $\tilde{x}, \tilde{y} \in D(\phi)$ as in Theorem 4 satisfying $\tilde{x} \leq x \leq \tilde{y}$ (e.g., see Pliska [20]). In this case, $T(t)x \rightarrow z$ for all $x \in \overline{D(\phi)}$.

A problem is said to be the positive case if there exists some $f \in M$ such that $r(f) \geq 0$ (this was the case mentioned just above). If $0 \in D(\phi)$, then $\phi 0 \geq 0$, so one can take $\tilde{x} = 0$. Suppose there exists a smallest singular point, say \tilde{y} , greater than or equal to 0 (i.e., if $z \geq 0$ is a singular point then $z \geq \tilde{y}$). Then $[0, \tilde{y}]$ possesses a unique singular point and $T(t)x \rightarrow \tilde{y}$ for all $x \in [0, \tilde{y}]$.

A problem is said to be the negative case if $r(f) \leq 0$ for all $f \in M$. This problem is symmetrical with the positive case. In particular, if $0 \in D(\phi)$, then $\phi 0 \leq 0$ and one can set $\tilde{y} = 0$. For \tilde{x} one takes the greatest singular point that is less than or equal to 0.

For a final example of the results of this section, we consider an optimal stopping problem. The optimal stopping of a continuous parameter Markov process has been considered by Fakeev [8], [9] and others. It is possible to formulate an important class of optimal stopping problems in the terms of this paper.

Consider a Markov process on S with infinitesimal generator \mathcal{A} which has domain $D(\mathcal{A})$ and range X . As usual, $D(\mathcal{A})$ is dense in the Banach space X . If the process is intentionally stopped in state s at any particular time, then the reward $v(s)$ is received, where $v \geq 0$ and $v \in D(\mathcal{A})$. The problem is to maximize the expected terminal reward over the class of all Markov stopping rules.

To formulate this as a Markov decision problem, suppose there are two actions for each state, namely "stop" and "continue". Thus $\phi x(s) = \max \{0, \mathcal{A}x(s)\} = 0 \vee \mathcal{A}x(s)$, $D(\phi) = D(\mathcal{A})$, the reward rate is zero, and the terminal reward is v .

Suppose assumptions (II)–(IV) hold; by definition, $\mathcal{A}x \in X$ for $x \in D(\mathcal{A})$, so

$\phi x \in X$ and ϕ satisfies (I). According to Fakeev [9], the maximum expected terminal reward in the infinite horizon case is the smallest excessive majorant of v . Hence, one would expect that $\lim_{t \rightarrow \infty} T(t)v$ satisfies this property, and this, indeed, is the case.

To see this, suppose the excessive function $z \in D(\phi)$ majorizes v . According to Dynkin [7, p. 44], a function $z \in D(\phi)$ is excessive if and only if $\mathcal{A}z \leq 0$ and $z \geq 0$. Then $\phi z = 0$, and since $\phi v \geq 0$, one concludes by Theorem 4 that $\bar{v} = \lim_{t \rightarrow \infty} T(t)v$ exists. In particular, suppose z is the smallest excessive majorant of v . If the hypotheses of Theorem 6 hold, then one must have $z = \bar{v}$.

5. Controlled diffusion processes. This section examines the controlled diffusion processes that were considered by Mandl [15] and Pliska [19]. The state space $S = [s_0, s_1]$ is a compact interval of the real line, and the set A of admissible actions is a compact subset of the real line. The diffusion coefficient $d(s, a) > 0$, the drift coefficient $b(s, a)$, and the reward $r(s, a)$ are continuous real-valued functions. The set M of admissible stationary policies consists of all piecewise continuous functions on S into A .

Each $f \in M$ defines a Markov process according to the differential operator

$$\mathcal{A}(f) = d(s, f(s)) \frac{d^2}{ds^2} + b(s, f(s)) \frac{d}{ds}$$

together with the Fellerian boundary condition

$$\kappa_j v(s_j) + \theta_j \left(v(s_j) - \int_S v(s) d\mu_j(s) \right) - (-1)^j \pi_j v'(s_j) + \sigma_j \mathcal{A}(f)v(s_j) = 0,$$

$j = 0, 1$, where $v(s)$ is a real-valued function whose second derivative is piecewise continuous on S . The four parameters $\kappa_j, \sigma_j, \pi_j$ and θ_j correspond, respectively, to the boundary phenomena of absorption, adhesion, reflection and instantaneous return. To complete the specification of the reward structure, suppose there arises the reward λ_j whenever the process is absorbed at boundary $s_j, j < 0, 1$. Finally, corresponding to θ_j , the process “jumps” from s_j into (s_0, s_1) according to the probability distribution $\mu_j, j = 0, 1$. Let v_j be a real-valued function on S which is integrable with respect to μ_j . If the process jumps from boundary s_j to $s \in (r_0, r_1)$, then there arises the reward $v_j(s)$.

It is now appropriate to formulate this controlled diffusion process problem in the context of this paper. Let X be the space of all bounded continuous real-valued functions on S under the supremum norm. For the purpose of this section, assume the process is nonconservative and neither boundary is purely adhesive, that is, $\kappa_0 + \kappa_1 > 0, \kappa_j + \pi_j + \theta_j > 0, j = 0, 1$. Then the expected undiscounted infinite-horizon reward corresponding to each $f \in M$ will be finite, and for $D(\phi)$ it is appropriate to take all C^2 -functions in X satisfying the boundary conditions

$$(\theta_j + \kappa_j)v(s_j) - \theta_j \int_S (v(s) + v_j(s)) d\mu_j(s) - (-1)^j \pi_j v'(s_j) - \sigma_j \gamma_j - \kappa_j \lambda_j = 0,$$

$j = 0, 1$, where

$$\gamma_j = \max_{a \in A} r(s_j, a), \quad j = 0, 1.$$

For technical reasons, if $\sigma_j > 0$, then without loss of generality, assume M is such that $r(s_j, f(s_j)) = \gamma_j$ for all $f \in M$. The operator ϕ is defined by (1). Note that $D(\phi)$ is convex.

The next step is to verify the various assumptions of this paper. Note that $D(\phi)$ does not contain any open subsets and that ϕ is not continuous on X . Assumption (I) holds by a standard result; see, for example, Berge [2, p. 115].

Before proceeding to the other assumptions, it is necessary to verify some preliminary results. First, with the scalar $\lambda \geq 0$ and $x \in X$, $v \in D(\phi)$ satisfies

$$(8) \quad v''(s) + \max_{a \in A} \{d(s, a)^{-1}[b(s, a)v'(s) - \lambda v(s) + x(s) + r(s, a)]\} = 0$$

if and only if, for each $s \in S$, there exists some $\bar{a} \in A$ such that

$$d(s, \bar{a})v''(s) + b(s, \bar{a})v'(s) - \lambda v(s) + r(s, \bar{a}) = -x(s)$$

and

$$d(s, a)v''(s) + b(s, a)v'(s) - \lambda v(s) + r(s, a) \leq -x(s)$$

for all $a \in A$. And this, in turn, is true if and only if v satisfies $(\lambda I - \phi)v = x$. Moreover, a straightforward generalization of the methods in Pliska [19] yields the fact that there exists a unique $v \in D(\phi)$ satisfying (8). Hence, for each $x \in X$ and any $\varepsilon > 0$, there exists a unique $v \in D(\phi)$ such that

$$(9) \quad (I - \varepsilon\phi)v = x.$$

In particular, $R(I - \varepsilon\phi) = X$ and assumption (III) holds. Also, for each $x \in X$, there exists a unique $v \in D(\phi)$ satisfying $\phi v = x$. Finally, comparing (9) with the results in Pliska [19], it is apparent that v can be interpreted as the maximum expected discounted reward (with respect to stationary policies) with discount factor $\lambda = \varepsilon^{-1}$ and reward rate $r + \varepsilon^{-1}x$. This interpretation immediately implies assumption (IV) is true.

In order to show assumption (II) is true for any $w \geq 0$, two cases will be examined, leaving the others to the reader. First, suppose $\|u - v\| = u(s) - v(s)$ for some $s \in (s_0, s_1)$, in which case $u''(s) - v''(s) \leq 0$ and $u'(s) - v'(s) = 0$. Let $a \in A$ be such that $\phi u(s) = r(s, a) + \mathcal{A}(a)u(s)$. Then $\mathcal{A}(a)u(s) \leq \mathcal{A}(a)v(s)$, so $\phi u(s) \leq \mathcal{A}(a)v(s) + r(s, a) \leq \phi v(s)$. Hence, for any $\varepsilon \geq 0$ and $w \geq 0$,

$$\begin{aligned} \|u - v\| &\leq (1 + \varepsilon w)(u(s) - v(s)) - \varepsilon(\phi u(s) - \phi v(s)) \\ &\leq \|(I - \varepsilon(\phi - wI)) - (I - \varepsilon(\phi - wI))v\|. \end{aligned}$$

For the second case, suppose $\|u - v\| = u(s_0) - v(s_0)$, in which case $u'(s_0) - v'(s_0) \leq 0$ and $u(s_0) - v(s_0) - \int (u(s) - v(s)) d\mu_0(s) \geq 0$. Subtracting boundary conditions yields

$$\begin{aligned} 0 = \kappa_0(u(s_0) - v(s_0)) + \theta_0 \left[u(s_0) - v(s_0) - \int (u(s) - v(s)) d\mu_0(s) \right] \\ - \pi_0[u'(s_0) - v'(s_0)]. \end{aligned}$$

Each term is nonnegative, so all are zero. In particular, $\kappa_0 > 0$ implies $u(s_0) = v(s_0)$ and $\|u - v\| = 0$, $\theta_0 > 0$ implies $\|u - v\| = u(s) - v(s)$ for some $s \in (s_0, s_1)$,

and $\pi_0 > 0$ implies $u'(s_0) = v'(s_0)$ and $u''(s_0) - v'(s_0) \leq 0$. In the latter two situations, one can proceed as with the first case; consequently, assumption (II) is true.

It should be remarked that it is not known whether $T(t)x$ satisfies the corresponding Cauchy problem. The Cauchy problem would have a solution if either ϕ is continuous or X is reflexive, but neither condition holds in this case.

6. Controlled jump processes. This section examines a controlled jump process that is considered in more detail in Pliska [20]. This process is a generalization of ones considered by Miller [16], [17], Kakumanu [12] and others. The state space S , with Borel σ -algebra \mathcal{S} , is a Borel subset of a Polish space. The action space A is a compact Borel subset of a Polish space. The process is specified by a real-valued function λ and a sub-Markov kernel Q on $S \times A$ satisfying the following conditions:

- (i) λ is continuous on $S \times A$.
- (ii) $0 < \lambda(s, a) < N$ for some number $N < \infty$ and all $(s, a) \in S \times A$.
- (iii) λ is uniformly continuous in a , i.e., given $\varepsilon > 0$, there exists some $\delta > 0$ such that $|a_1 - a_2| < \delta$ implies $\sup_{s \in S} |\lambda(s, a_1) - \lambda(s, a_2)| < \varepsilon$.
- (iv) For each $(s, a) \in S \times A$, $S' \rightarrow Q(S'|s, a)$ is a subprobability measure on \mathcal{S} with $Q(S|s, a) \leq 1$.
- (v) For each $S' \in \mathcal{S}$, $(s_n, a_n) \rightarrow (s, a)$ implies $Q(S'|s_n, a_n) \rightarrow Q(S'|s, a)$.
- (vi) For each $(s, a) \in S \times A$, $Q(\{s\}|s, a) = 0$.
- (vii) $Q(S|s, a)$ is uniformly continuous in a , i.e., given $\varepsilon > 0$, there exists some $\delta > 0$ such that $|a_1 - a_2| < \delta$ implies $\sup_{s \in S} |Q(S|s, a_1) - Q(S|s, a_2)| < \varepsilon$.

The set of admissible stationary policies consists of all Borel measurable functions on S into A . The reward rate r is a bounded upper semicontinuous real-valued function on $S \times A$ which is uniformly continuous in a in the sense that, given $\varepsilon > 0$, there exists some $\delta > 0$ such that $|a_1 - a_2| < \delta$ implies $\sup_{s \in S} |r(s, a_1) - r(s, a_2)| < \varepsilon$.

Let X be the Banach space of bounded Borel measurable functions on S under the supremum norm. The process has the infinitesimal generator

$$\mathcal{A}(a)v(s) = -\lambda(s, a) \left[v(s) - \int v(z)Q(dz|s, a) \right],$$

where $a \in A$ and $v \in X$. Let $D(\phi)$ denote the set of all upper semicontinuous functions in X . In Pliska [20], it is shown that ϕ maps $D(\phi)$ into X (i.e., assumption (I) holds) and ϕ is Lipschitz continuous. In view of Maitra [14, Lemma 4.2], $D(\phi) = \overline{D(\phi)}$. Note that ϕ neither maps $D(\phi)$ into itself nor X into itself.

According to Pliska [20], for $x \in D(\phi)$, $(\lambda I - \phi)^{-1}x \in D(\phi)$ is the maximum expected discounted reward with respect to stationary policies, with discount factor λ and with reward rate $r + x$. Hence for $x \in D(\phi)$, $(I - \varepsilon\phi)^{-1}x \in D(\phi)$ is the maximum expected discounted reward with respect to stationary policies with discount factor ε^{-1} and with reward rate $r + \varepsilon^{-1}x$. In particular, assumptions (III) and (IV) hold. Assumption (II) is true for all $w \geq 0$, because if $s \in S$ is such that $\|u - v\| = u(s) - v(s)$ for arbitrary $u, v \in D(\phi)$, then $u(s) - v(s) \geq \int [u(z) - v(z)] \cdot Q(dz|s, a)$, $\phi u(s) - \phi v(s) \leq 0$, and one can proceed as in the case of controlled diffusions to obtain the desired result. On the other hand, if there does not exist

any $s \in S$ such that $\|u - v\| = u(s) - v(s)$, then a suitable limiting argument yields the appropriate result.

To briefly conclude, by Theorem 1, ϕ is the infinitesimal generator of a semigroup $T \in Q_w(D(\phi))$. The operator ϕ is single-valued and continuous from the strong to the weak topology and $D(\phi) = \overline{D(\phi)}$, so by a result in Crandall and Liggett [6, § 2], we also know that $T(t)x$ satisfies the corresponding Cauchy problem. Finally, Pliska [20] verifies that $T(t)x$ is indeed the maximum expected reward vector.

Acknowledgment. The author is grateful to the referee for a number of valuable comments and suggestions.

REFERENCES

- [1] M. ABRAMOWITZ AND I. A. STEGUN, eds., *Handbook of Mathematical Functions*, Applied Mathematics Series 55, National Bureau of Standards, Washington, D.C., 1964.
- [2] C. BERGE, *Topological Spaces*, Oliver and Boyd, London, 1963.
- [3] F. E. BROWDER, *Nonexpansive nonlinear operators in a Banach space*, Proc. Nat. Acad. Sci. USA, 54 (1965), pp. 1041–1044.
- [4] F. E. BROWDER AND W. PETRYSHYN, *The solution by iteration of non-linear functional equations in Banach spaces*, Bull. Amer. Math. Soc., 72 (1966), pp. 571–575.
- [5] J. A. CLARKSON, *Uniformly convex spaces*, Trans. Amer. Math. Soc., 40 (1936), pp. 396–414.
- [6] M. G. CRANDALL AND T. M. LIGGETT, *Generation of semigroups of non-linear transformations on general Banach spaces*, Amer. J. Math., 43 (1971), pp. 265–298.
- [7] E. B. DYNKIN, *Markov Processes*, vol. 1, Springer-Verlag, Berlin, 1965.
- [8] A. G. FAKEEV, *Optimal stopping rules for stochastic processes with continuous parameter*, Theor. Probability Appl., 15 (1970), pp. 324–331.
- [9] ———, *Optimal stopping of a Markov process*, Ibid., 16 (1971), pp. 694–696.
- [10] W. H. FLEMING, *Optimal continuous-parameter stochastic control*, SIAM Rev., 11 (1969), pp. 470–509.
- [11] K. GOEBEL AND W. KIRK, *A fixed point theorem for asymptotically non-expansive mappings*, Proc. Amer. Math. Soc., 35 (1972), pp. 171–174.
- [12] P. KAKUMANU, *Continuously discounted Markov decision models with countable state and action space*, Ann. Math. Statist., 42 (1971), pp. 919–926.
- [13] W. KIRK, *Successive approximations for nonexpansive mappings in Banach spaces*, Glasgow Math. J., 12 (1971), pp. 6–9.
- [14] A. MAITRA, *Discounted dynamic programming on compact metric spaces*, Sankhyā Ser. A, 30 (1968), pp. 211–216.
- [15] P. MANDL, *Analytical Treatment of One-dimensional Markov Processes*, Springer-Verlag, New York, 1968.
- [16] B. L. MILLER, *Finite state continuous time Markov decision processes with a finite planning horizon*, this Journal, 6 (1968), pp. 266–280.
- [17] ———, *Finite state continuous time Markov decision processes with an infinite planning horizon*, J. Math. Anal. Appl., 22 (1968), pp. 552–569.
- [18] W. PETRYSHYN AND T. WILLIAMSON, *A necessary and sufficient condition for convergence of a sequence of iterates for quasi-nonexpansive mappings*, Bull. Amer. Math. Soc., 78 (1972), pp. 1027–1031.
- [19] S. R. PLISKA, *Single-person controlled diffusions with discounted costs*, J. Optimization Theory Appl., 12 (1973), pp. 248–255.
- [20] ———, *Controlled jump processes*, Stochastic Processes Appl., to appear.
- [21] M. L. PUTERMAN, *On the optimal control of diffusion processes*, Tech. Rep. 14, Department of Operations Research, Stanford Univ., Stanford, Calif., 1972.
- [22] L. D. STONE, *Necessary and sufficient conditions for optimal control of semiMarkov jump processes*, this Journal, 11 (1973), pp. 187–201.

- [23] A. F. VEINOTT, JR., *Discrete dynamic programming with sensitive discount optimality criteria*, Ann. Math. Statist., 40 (1969), pp. 1635–1660.

THE INFINITE-DIMENSIONAL RICCATI EQUATION WITH APPLICATIONS TO AFFINE HEREDITARY DIFFERENTIAL SYSTEMS*

RUTH F. CURTAIN†

Abstract. The infinite-dimensional versions of the linear quadratic cost control problem and of the linear filtering problem lead to an infinite-dimensional Riccati equation with unbounded operators. Existence and uniqueness theorems for mild solutions of these were established in *The infinite dimensional Riccati equations*, Ruth F. Curtain and A. J. Pritchard (to appear in J. Math. Anal. Appl.) using a semigroup and evolution operator approach. Although this formulation was very general, covering a large class of parabolic partial differential control systems, it does not cover the semigroup formulation of linear hereditary differential equations introduced by Delfour and Mitter. This paper remedies this and applies the theory to the linear quadratic cost control problem for the affine linear hereditary differential case.

Introduction. The study of the infinite-dimensional linear quadratic cost control problem and the infinite-dimensional filtering problem both lead to an infinite-dimensional Riccati equation with unbounded operators. The original motivation for considering infinite-dimensional systems came from distributed parameter systems, and the best known solution to the linear quadratic cost control problem for linear parabolic partial differential equations is by Lions [14]. Balakrishnan [1] was one of the first to use a semigroup approach to solve distributed parameter control problems, which transforms the problem into an abstract one in a Hilbert space. Lukes and Russell in [15] and Datko in [6] both use this semigroup approach to study the time-invariant linear quadratic cost control problem in a Hilbert space, though by very different methods. Finally, in [4], Curtain and Pritchard gave very general conditions for existence and uniqueness of solutions of the Riccati equation for the time-dependent case.

Although these theorems were formulated for an abstract Hilbert space and were used to solve an infinite-dimensional filtering problem in [2], the conditions were formulated with the parabolic partial differential equations in mind. More recently, Delfour and Mitter in [8] and [9] have studied a class of affine hereditary differential equations, by converting the problem to an abstract one in a suitable Hilbert space, again using a "semigroup" approach.

In [9] they gave a complete theory of the linear quadratic cost control problem for this class of systems obtaining the expected Riccati equation. Although they proved the existence and uniqueness of an integrated version and of a time-invariant version using a result of da Prato, the proof of the differentiated version is incomplete (see Lemma 8 below). This paper fills this gap and also provides an interesting alternative treatment of the linear quadratic cost control problem for affine hereditary differential systems along the lines of Pritchard [16] and Curtain and Pritchard [4].¹

* Received by the editors January 22, 1974, and in revised form June 17, 1974.

† Control Theory Centre, University of Warwick, Coventry CV4 7AL, England. The Control Centre is supported by the Leverhulme Trust and by the Science Research Council under Grant B/SR/9186.

¹ See [3] for applications to the filtering problem for affine hereditary differential systems.

At the same time it is shown that the linear quadratic cost control problem in infinite dimensions has a solution under minimal conditions on the system operator $\mathcal{A}(t)$, the restrictions only being required when one wants a differentiated version of the Riccati equation.

1. Preliminaries on evolution operators. Let X be a real Banach space and consider the following homogeneous abstract evolution equation :

$$(1.1) \quad \begin{aligned} \frac{dz(t)}{dt} &= \mathcal{A}(t)z(t), \\ z(s) &= z_0, \quad 0 \leq s \leq T, \end{aligned}$$

where $t \in [0, T] = T$, a real finite-time interval, $\mathcal{A}(t)$ a closed operator on X for each t and $z_0 \in X$.

Then there are conditions on $\mathcal{A}(t)$ which guarantee the existence of a strong evolution operator $\mathcal{U}(t, s) \in \mathcal{L}(X)$ for $0 \leq s \leq t \leq T$ with the properties :

$$(1.2) \quad \begin{aligned} &\text{(i) } \mathcal{U}(t, t) = \mathcal{I}, \text{ the identity operator,} \\ &\text{(ii) } \mathcal{U}(t, s)\mathcal{U}(s, \sigma) = \mathcal{U}(t, \sigma), \quad \sigma \leq s \leq t. \\ &\text{(iii) } \mathcal{U}(t, s): \mathcal{D}(\mathcal{A}(s)) \rightarrow \mathcal{D}(\mathcal{A}(t)), \\ &\text{(iv) } \frac{\partial}{\partial t}(\mathcal{U}(t, s)z_0) = \mathcal{A}(t)\mathcal{U}(t, s)z_0 \quad \text{for } z_0 \in \mathcal{D}(\mathcal{A}(s)), \\ &\text{(v) } \mathcal{U}(t, s) \text{ is strongly continuous in } s \text{ and } t \text{ for } 0 \leq s < t \leq T, \\ &\text{(vi) } \frac{\partial}{\partial s}(\mathcal{U}(t, s)z_0) = -\mathcal{U}(t, s)\mathcal{A}(s)z_0 \quad \text{for } z_0 \in \mathcal{D}(\mathcal{A}(s)) \\ &\quad \text{(follows from (i), (ii), (iv) and (v)).} \end{aligned}$$

When (1.1) represents a parabolic partial differential equation, the standard conditions are that \mathcal{A} be the infinitesimal generator of a strongly continuous semigroup in the time-independent case, and for the time-dependent case, see Kato [12] or [13] or the end of this section. When (1.1) represents a linear hereditary system, sufficient conditions for the existence of such a $\mathcal{U}(t, s)$ are given in [8]. In general, sufficient conditions for the existence of such a $\mathcal{U}(t, s)$ are that (1.1) should have a unique continuous solution $z(t)$ for $t \in [s, T]$ and $z_0 \in \mathcal{D}(\mathcal{A}(s))$ and $z(t)$ should depend continuously on the initial conditions. Then if we write $z(t) = \mathcal{U}(t, s)z_0$, it is easily deduced that $\mathcal{U}(t, s) \in \mathcal{L}(X)$ and has the properties (i)–(v). Unfortunately, in some applications, we often need $\mathcal{U}(t, s)$ to have additional properties. We now summarize the main results on abstract evolution equations which we need in the sequel. Consider now the inhomogeneous equation

$$(1.3) \quad \begin{aligned} \frac{dz(t)}{dt} &= \mathcal{A}(t)z(t) + f(t), \\ z(0) &= z_0. \end{aligned}$$

Then (1.3) has the unique solution $z(t) = \mathcal{U}(t, 0)z_0 + \int_0^t \mathcal{U}(t, s)f(s) ds$ provided

(a) $z_0 \in \mathcal{D}(\mathcal{A}(t))$, $f(s) \in \mathcal{D}(\mathcal{A}(t)) \forall t \in [0, T]$ and (b) $f(\cdot)$, $\mathcal{A}(t)f(\cdot)$ are Bochner integrable on $[0, T]$ (see [1]).

Remark. If $f(\cdot)$ is merely Bochner integrable, then $z(t) = \mathcal{U}(t, 0)z_0 + \int_0^t \mathcal{U}(t, s)f(s)ds$ is still a well-defined element of X , $\forall t \in [0, T]$ but it may not satisfy (1.3).

DEFINITION 1.1. We then call $z(t)$ a mild solution of (1.3).

LEMMA 1. Consider the perturbed equation

$$(1.4) \quad \begin{aligned} \frac{dz(t)}{dt} &= \mathcal{A}(t)z(t) + \mathcal{B}(t)z(t), & 0 \leq s < t \leq T, \\ z(s) &= z_0 & \text{for } z_0 \in \mathcal{D}(\mathcal{A}(s)), \end{aligned}$$

where $\mathcal{A}(t)$ is a closed operator on X generating the strong evolution operator $\mathcal{U}(t, s)$ satisfying (1.2) (i)–(v). Then if $\mathcal{B}(t) \in L_\infty(T; \mathcal{L}(X))$,

$$(1.5) \quad \mathcal{Y}(t, s) = \mathcal{U}(t, s) + \int_0^t \mathcal{U}(t, \alpha)\mathcal{B}(\alpha)\mathcal{Y}(\alpha, s) d\alpha$$

has the unique solution $\mathcal{Y}(t, s) \in \mathcal{L}(X)$ which is strongly continuous in s and t , with $\sup_{s \leq t \in T} \|\mathcal{Y}(t, s)\| \leq M e^{M\alpha T}$, where M, α are constants such that $\sup_{s < t \in T} \|\mathcal{U}(t, s)\| \leq M$ and $\sup_{t \in T} \|\mathcal{B}(t)\| \leq \alpha$.

Proof. (1.5) is a Volterra equation on $\mathcal{L}(X)$ and, for fixed s , has the form $\mathcal{Y}(\cdot, s) = \mathcal{U}(\cdot, s) + K\mathcal{Y}(\cdot, s)$, where K is a compact operator of norm < 1 on the space of strongly continuous operators on $\mathcal{L}(X)$. So it has the unique strongly continuous solution $\mathcal{Y}(\cdot, s)$. The inequalities follow from the standard iterative estimates on K :

$$\sup_{t \in T} \|K^n \mathcal{Y}(t, s)\|_{\mathcal{L}(X)} \leq \frac{(M\alpha(T-s))^n}{n!},$$

where

$$\sup_{0 \leq s < t \in T} \|\mathcal{U}(t, s)\| \leq M \quad \text{and} \quad \sup_{t \in T} \|\mathcal{B}(t)\| \leq \alpha$$

and

$$\mathcal{Y}(t, s) = (I - K)^{-1} \mathcal{U}(t, s) = \sum_{n=0}^{\infty} K^n \mathcal{U}(t, s)$$

so

$$\sup_{0 \leq s < t \in T} \|\mathcal{Y}(t, s)\| \leq M e^{M\alpha T}.$$

DEFINITION 1.2. Consider (1.5) where $\mathcal{B}(\cdot) \in L_\infty(T; \mathcal{L}(X))$ and $\mathcal{A}(t)$ is the generator of the strong evolution operator $\mathcal{U}(t, s)$. Then we call the unique solution $\mathcal{Y}(t, s)$ of (1.7) the mild evolution operator generated by $\mathcal{A}(t) + \mathcal{B}(t)$.

We note that if (1.4) has a unique strong solution, then it is necessarily of the form $z(t) = \mathcal{Y}(t, s)z_0$, i.e., $\mathcal{Y}(t, s)$ is then a strong evolution operator. We give two sufficient conditions for $\mathcal{Y}(t, s)$ to be a strong evolution operator.

LEMMA 2. Under the alternative extra assumptions (1.6)(a) or (1.6)(b), the mild evolution operator $\mathcal{Y}(t, s)$ for (1.5) is actually a strong evolution operator.

(1.6) (a) *The following inhomogeneous equation has a unique solution for any measurable X -valued function $f \in L_2(X)$*

$$\dot{z}(t) = \mathcal{A}(t)z(t) + \mathcal{B}(t)f(t),$$

$$z(s) = z_0;$$

$$(1.6) \text{ (b)} \quad \text{(i) } \mathcal{B}(s): X \rightarrow \mathcal{D}(\mathcal{A}(t)) \quad \forall s < t \in T,$$

$$\text{(ii) } \sup_{s, t \in T} \|\mathcal{A}(t)\mathcal{B}(s)x\| \text{ exists} \quad \forall x \in X.$$

Proof.

(a) (1.6)(a) implies that

$$\frac{\partial}{\partial t} \left[\int_s^t \mathcal{U}(t, \alpha) \mathcal{B}(\alpha) f(\alpha) d\alpha \right] = \mathcal{A}(t) \int_s^t \mathcal{U}(t, \alpha) \mathcal{B}(\alpha) f(\alpha) d\alpha + \mathcal{B}(t)f(t)$$

for any continuous function $f(\cdot)$.

$$(1.7) \quad \therefore \frac{\partial}{\partial t} \left[\int_0^t \mathcal{U}(t, \alpha) \mathcal{B}(\alpha) \mathcal{Y}(\alpha, s) z_0 d\alpha \right] = \mathcal{A}(t) \int_0^t \mathcal{U}(t, \alpha) \mathcal{B}(\alpha) \mathcal{Y}(\alpha, s) z_0 d\alpha + \mathcal{B}(t) \mathcal{Y}(t, s) z_0.$$

By Lemma 1, $z(t) = \mathcal{Y}(t, s)z_0$ is the solution to the integral equation

$$(1.8) \quad z(t) = \mathcal{U}(t, s)z_0 + \int_s^t \mathcal{U}(t, \alpha) \mathcal{B}(\alpha) z(\alpha) d\alpha,$$

and by (1.7)

$$\mathcal{Y}(t, s)z_0 \text{ satisfies } \dot{z}(t) = \mathcal{A}(t)z(t) + \mathcal{B}(t)z(t),$$

i.e., $\mathcal{Y}(t, s)$ is a strong evolution operator

(b) Under assumptions (1.6)(b), (1.4) is equivalent to the integral equation (1.8) (cf. [1]) and (1.8) has the unique solution $z(t) = \mathcal{Y}(t, s)z_0$ by Lemma 1. So $\mathcal{Y}(t, s)$ is again a strong evolution operator.

LEMMA 3. *Consider the sequence of abstract evolution equations*

$$(1.9) \quad \frac{dz_k(t)}{dt} = (\mathcal{A}(t) + \mathcal{B}_k(t))z(t),$$

$$z(0) = z_0, \quad z_0 \in \mathcal{D}(\mathcal{A}(0)),$$

where $\mathcal{A}(t)$ is as in Lemma 2 and $\mathcal{B}_k(\cdot) \in L_\infty(T; \mathcal{L}(X)) \forall k, \mathcal{B}_k(t) \rightarrow \mathcal{B}_\infty(t)$ strongly as $k \rightarrow \infty$ and $\|\mathcal{B}_k(t)\| \leq C$ uniformly in k and in t on T . Then if $\mathcal{A}(t) + \mathcal{B}_k(t)$ generates the mild evolution operator $\mathcal{U}_k(t, s)$ for each k , $\mathcal{U}_k(t, s) \rightarrow \mathcal{U}_\infty(t, s)$ strongly as $k \rightarrow \infty$, where $\mathcal{U}_\infty(t, s)$ is the mild evolution operator generated by $\mathcal{A}(t) + \mathcal{B}_\infty(t)$.

Proof. From Lemma 1, $\mathcal{U}_k(t, s)$ is the unique solution of the integral equation

$$\mathcal{U}_k(t, s) = \mathcal{U}(t, s) + \int_s^t \mathcal{U}(t, \alpha) \mathcal{B}_k(\alpha) \mathcal{U}_k(\alpha, s) d\alpha \quad \forall k < \infty.$$

Since all terms are uniformly bounded in norm in k and t , we can take limits as $k \rightarrow \infty$, and so $\mathcal{U}_k(t, s)$ converges strongly to the solution of

$$\mathcal{Y}(t, s) = \mathcal{U}(t, s) + \int_s^t \mathcal{U}(t, \alpha) \mathcal{B}_\infty(\alpha) \mathcal{Y}(\alpha, s) ds,$$

which by uniqueness is just $\mathcal{U}_\infty(t, s)$.

Remarks. In [4] similar lemmas were proved but with stronger conditions on $\mathcal{A}(t)$, which we state here for comparison.

Conditions on $\mathcal{A}(t)$. (i) $\mathcal{A}(t)$ is a densely-defined, closed, linear operator on X , whose spectrum is contained in the fixed sector: $\Sigma: |\arg \gamma| < \theta < \pi/2$ and $\|(\gamma \mathcal{I} - \mathcal{A}(t))^{-1}\| \leq \varepsilon/|\gamma|$ for $\gamma \notin \Sigma$;

(ii) $\mathcal{A}(t)^{-1} \in \mathcal{L}(X)$ and is Hölder continuously differentiable in t in the uniform operator topology;

$$(iii) \left\| \frac{d}{dt}(\gamma \mathcal{I} - \mathcal{A}(t))^{-1} \right\| \leq \frac{n}{|\gamma|^{1-\rho}} \text{ for } \gamma \notin \Sigma, 0 \leq \rho \leq 1.$$

(This gives much stronger conditions on $\mathcal{U}(t, s)$, namely,

$$\mathcal{U}(t, s): X \rightarrow \mathcal{D}(\mathcal{A}(t)) \quad \text{and} \quad \|\mathcal{A}(t)\mathcal{U}(t, s)\| \leq c/|t - s|.)$$

These conditions imply that $\mathcal{A}(t)$ generates an analytic semigroup for each t and impose smoothness conditions with respect to t . There are alternative conditions on $\mathcal{A}(t)$ one could use (see, for example, [12]), but these again are stronger and apply to parabolic partial differential equations. In these cases the perturbation result of Lemma 2 holds without requiring conditions (ii) and (iii) on $\mathcal{B}(\cdot)$ and $\mathcal{A}(t) + \mathcal{B}(t)$ always generates an evolution operator with similar conditions to $\mathcal{U}(t, s)$ provided $\mathcal{B}(\cdot) \in L_\infty(T; \mathcal{L}(X))$.

In [4], the time-independent case was treated separately and assumed only that \mathcal{A} was the generator of a strongly continuous semigroup, but $\mathcal{B}(t)$ was required to be strongly continuously differentiable in t to ensure Lemma 2.

2. The linear quadratic cost control problem and the integrated version of the infinite-dimensional Riccati equation. Consider the infinite-dimensional system

$$(2.1) \quad \begin{aligned} \frac{dz}{dt} &= \mathcal{A}(t)z(t) + \mathcal{B}(t)u(t), \\ z(0) &= z_0 \end{aligned} \quad t \in T,$$

with the cost functional

$$(2.2) \quad \mathcal{C}(u) = \langle z(T), \mathcal{G}z(T) \rangle + \int_0^T [\langle z(s), \mathcal{W}(s)z(s) \rangle + \langle u(s), \mathcal{R}(s)u(s) \rangle] ds,$$

where $\mathcal{A}(t)$ is a closed operator on a Hilbert space \mathcal{H} , $z_0 \in \mathcal{H}$, and $\mathcal{A}(t)$ generates the evolution operator $\mathcal{U}(t, s)$ with properties (1.2) (i)–(v). The control $u \in L_2(T; \mathcal{V})$, where \mathcal{V} is another Hilbert space and $\mathcal{B}(t) \in \mathcal{L}(\mathcal{V}, \mathcal{H})$, $\mathcal{G}, \mathcal{W}(t) \in \mathcal{L}(\mathcal{H})$ are self-adjoint and positive semi-definite operators; $\mathcal{R}(t), \mathcal{R}(t)^{-1} \in \mathcal{L}(\mathcal{V})$ are self-adjoint and positive definite. $\mathcal{B}(t)$, $\mathcal{W}(t)$, $\mathcal{R}(t)$ and $\mathcal{R}(t)^{-1}$ are all assumed uniformly bounded in norm on T .

(2.1) and (2.2) define the infinite-dimensional version of the linear quadratic cost control problem and lead to the following integrated Riccati equation:

$$(2.3) \quad \mathcal{Q}_\infty(t) = \mathcal{U}_\infty^*(T, t) \mathcal{G} \mathcal{U}_\infty(T, t) + \int_t^T \mathcal{U}_\infty^*(s, t) \mathcal{W}_\infty(s) \mathcal{U}_\infty(s, t) ds,$$

whose existence and uniqueness we now prove.

We follow the procedure used by Curtain and Pritchard in [4]. Consider the sequence of control problems generated by a sequence $\{u_k(t)\}$ of admissible controls of the form $u_k(t) = -\mathcal{F}_k(t)z(t)$:

$$(2.4) \quad \begin{aligned} \frac{dz(t)}{dt} &= \mathcal{A}_k(t)z(t) + \mathcal{B}(t)\bar{u}(t), \\ z(0) &= z_0, \end{aligned}$$

where

$$(2.5) \quad \begin{aligned} \mathcal{A}_k(t) &= \mathcal{A}(t) - \mathcal{B}(t)\mathcal{F}_k(t) \quad \text{and} \quad \mathcal{F}_0(t) = 0, \\ \mathcal{F}_k(t) &= \mathcal{R}^{-1}(t)\mathcal{B}^*(t)\mathcal{Q}_{k-1}(t), \\ \mathcal{W}_k(t) &= \mathcal{W}(t) + \mathcal{F}_k^*(t)\mathcal{R}(t)\mathcal{F}_k(t), \\ \mathcal{Q}_k(t) &= \mathcal{U}_k^*(T, t) \mathcal{G} \mathcal{U}_k(T, t) + \int_t^T \mathcal{U}_k^*(s, t) \mathcal{W}_k(s) \mathcal{U}_k(s, t) ds, \end{aligned}$$

where $\mathcal{U}_k(t, s)$ is the mild evolution operator generated by $\mathcal{A}_k(t)$. Then (2.4) has the mild solution

$$(2.6) \quad z(t) = \mathcal{U}_k(t, 0)z_0 + \int_0^t \mathcal{U}_k(t, s)\mathcal{B}(s)\bar{u}(s) ds.$$

LEMMA 4.

$$\begin{aligned} \langle z(t), \mathcal{Q}_k(t)z(t) \rangle &= \langle z(T), \mathcal{G}z(T) \rangle \\ &+ \int_t^T [\langle z(s), \mathcal{W}_k(s)z(s) \rangle - \langle z(s), \mathcal{Q}_k(s)\mathcal{B}(s)\bar{u}(s) \rangle - \langle \mathcal{Q}_k(s)\mathcal{B}(s)\bar{u}(s), z(s) \rangle] ds. \end{aligned}$$

Proof. The proof is by substitution from (2.4) and (2.6).

LEMMA 5. $\mathcal{Q}_k(t)$ converges strongly to the self-adjoint operator $\mathcal{Q}_\infty(t) \in \mathcal{L}(\mathcal{H})$ which is uniformly bounded in norm.

Proof. The cost for the system (2.4) with $\bar{u} \equiv 0$ is given by $\mathcal{C}(u_k) = \langle z_0, \mathcal{Q}_k(0)z_0 \rangle$, using Lemma 4. Letting $\bar{u} = u_{k+1} + \mathcal{F}_k z_k$ in (2.4) gives

$$\begin{aligned} \mathcal{C}(u_{k+1}) - \mathcal{C}(u_k) &= \langle z(T), \mathcal{G}z(T) \rangle - \langle z_0, \mathcal{Q}_k(0)z_0 \rangle \\ &+ \int_0^T [\langle z(s), \mathcal{W}(s)z(s) \rangle + \langle (\bar{u} - \mathcal{F}_k z), \mathcal{R}(s)(\bar{u} - \mathcal{F}_k(s)z(s)) \rangle] ds, \end{aligned}$$

where $z(t)$ is the solution of (2.4). Applying Lemma 4 with $t = 0$, and $\bar{u} = -\mathcal{R}^{-1}(s) \cdot [\mathcal{B}^*(s)\mathcal{Q}_k(s) - \mathcal{R}(s)\mathcal{F}_k(s)]z(s)$ yields $\mathcal{C}(u_{k+1}) - \mathcal{C}(u_k) \leq 0$, i.e., $\langle z_0, \mathcal{Q}_k(0)z_0 \rangle$ is decreasing in $k \forall z_0 \in \mathcal{H}$. By a similar argument, the cost of controlling (2.4) with $\bar{u} \equiv 0$ and initial condition z_0 at time t_0 is just $\langle z_0, \mathcal{Q}_k(t_0)z_0 \rangle$, and this is also decreasing in k for each fixed t .

By definition of \mathcal{C} ,

$$\begin{aligned}\langle z_0, \mathcal{Q}_k(t)z_0 \rangle &\leq \langle z_0, \mathcal{Q}_k(0)z_0 \rangle \\ &\leq \langle z_0, \mathcal{Q}_0(0)z_0 \rangle \quad \forall z_0 \in \mathcal{H},\end{aligned}$$

so $\{\mathcal{Q}_k(t)\}$ is a sequence of positive semidefinite, self-adjoint operators nonincreasing in k , weakly continuous in t and uniformly bounded in norm in k and on $[0, T]$. Therefore $\mathcal{Q}_k(t)$ converges strongly to a self-adjoint operator which is uniformly bounded in norm on T .

LEMMA 6. *The unique optimal control for the linear quadratic cost control problem (2.1), (2.2) is the feedback control*

$$u(t) = -\mathcal{R}^{-1}(t)\mathcal{B}^*(t)\mathcal{Q}_\infty(t)z(t)$$

with the minimum cost $\langle z_0, \mathcal{Q}_\infty(0)z_0 \rangle$.

Proof. See [16].

$\mathcal{F}_k(t)$, $\mathcal{W}_k(t)$ are sequences of weakly continuous operators bounded in norm uniformly in k and t and so converge strongly to

$$\mathcal{F}_\infty(t) = \mathcal{R}^{-1}\mathcal{B}^*(t)\mathcal{Q}_\infty(t) \quad \text{and} \quad \mathcal{W}_\infty(t) = \mathcal{W}(t) + \mathcal{F}_\infty^*(t)\mathcal{R}(t)\mathcal{F}_\infty(t)$$

respectively. Considering Lemma 4, we see that all operators are bounded in norm uniformly in k and t , and so we can let $k \rightarrow \infty$, by the Lebesgue dominated convergence theorem. Letting $t = 0$ and $u(t) = \bar{u}(t) - \mathcal{R}^{-1}(t)\mathcal{B}^*(t)\mathcal{Q}_\infty(t)z(t)$ in the resulting expression where u is any admissible control, we obtain

$$\mathcal{C}(u) = \langle z_0, \mathcal{Q}_\infty(0)z_0 \rangle + \int_0^T \langle \bar{u}(s), \mathcal{R}(s)\bar{u}(s) \rangle ds > \langle z_0, \mathcal{Q}_\infty(0)z_0 \rangle$$

since \mathcal{R} is positive definite.

LEMMA 7. $\mathcal{Q}_\infty(t)$ is the unique weakly continuous solution of the integral equation (2.7)

$$\begin{aligned}\mathcal{Q}_\infty(t) &= \mathcal{U}_\infty^*(T, t)\mathcal{G}\mathcal{U}_\infty(T, t) \\ &\quad + \int_0^T \mathcal{U}_\infty^*(s, t)[\mathcal{W}(s) + \mathcal{Q}_\infty(s)\mathcal{B}(s)\mathcal{R}^{-1}(s)\mathcal{B}^*(s)\mathcal{Q}_\infty(s)]\mathcal{U}_\infty(s, t) ds,\end{aligned}$$

where $\mathcal{U}_\infty(t, s)$ is the mild evolution operator generated by

$$\mathcal{A}(t) - \mathcal{B}(t)\mathcal{R}^{-1}(t)\mathcal{B}^*(t)\mathcal{Q}_\infty(t).$$

Proof. In (2.5) all operators are at least integrable in t and are uniformly bounded in k on $[0, T]$, since

$$\sup_{0 \leq s \leq t \leq T} \|\mathcal{U}_k(t, s)\| \leq M e^{M\alpha FT}$$

by Corollary 1, where $\|\mathcal{U}(t, s)\| \leq M$, $\|\mathcal{B}(t)\| \leq \alpha$, $\|\mathcal{F}_k(t)\| \leq F$, all uniformly in t, s and k on $[0, T]$. So by the Lebesgue dominated convergence theorem, we can take limits as $k \rightarrow \infty$ of

$$\langle \mathcal{Q}_k(t)x, y \rangle = \langle \mathcal{U}_k(T, t)x, \mathcal{G}\mathcal{U}_k(T, t)y \rangle + \int_t^T \langle \mathcal{U}_k(s, t)x, \mathcal{W}_k(s)\mathcal{U}_k(s, t)y \rangle ds.$$

Since it holds $\forall x, y \in \mathcal{H}$, we can dispense with the inner product. Lemma 3 ensures us that $\mathcal{U}_k(t, s)$ converges strongly to $\mathcal{U}_\infty(t, s)$, the mild evolution operator generated by $\mathcal{A}(t) - \mathcal{B}(t)\mathcal{R}^{-1}(t)\mathcal{B}^*(t)\mathcal{Q}_\infty(t)$. Uniqueness follows as in [4] using similar arguments as in Lemma 6.

We summarize these results thus generalizing those which were proved in [4] and [16], under stronger conditions on $\mathcal{A}(t)$, with parabolic partial differential equations in mind.

THEOREM 1. *Let $\mathcal{A}(t)$ be a closed operator on a Hilbert space \mathcal{H} which generates a strong evolution operator $\mathcal{U}(t, s)$ with properties (1.2) (i)–(v). Let \mathcal{V} be another Hilbert space and suppose $\mathcal{B}(\cdot) \in L_\infty(T, \mathcal{L}(\mathcal{V}, \mathcal{H}))$, $\mathcal{G} \in \mathcal{L}(\mathcal{H})$, $\mathcal{W}(\cdot) \in L_\infty(T; \mathcal{L}(\mathcal{H}))$, $\mathcal{R}(\cdot)$ and $\mathcal{R}(\cdot)^{-1} \in L_\infty(T; \mathcal{L}(\mathcal{V}))$, and $\mathcal{B}(t)$, \mathcal{G} , $\mathcal{W}(t)$ are self-adjoint, positive semidefinite operators, $\mathcal{R}(t)$ and $\mathcal{R}^{-1}(t)$ are self-adjoint and positive definite. Under the above conditions, the linear quadratic cost control problem (2.1) and (2.2) has a unique minimizing feedback control $u(t) = -\mathcal{R}^{-1}(t)\mathcal{B}^*(t)\mathcal{Q}_\infty(t)z(t)$ with minimum cost $\langle z_0, \mathcal{Q}_\infty(0)z_0 \rangle$, where $\mathcal{Q}_\infty(t)$ is the unique weakly continuous solution of the following integral version of the Riccati equation:*

$$\begin{aligned} \mathcal{Q}_\infty(t) = & \mathcal{U}_\infty^*(T, t)\mathcal{G}\mathcal{U}_\infty(T, t) + \int_t^T \mathcal{U}_\infty^*(s, t)[\mathcal{W}(s)]\mathcal{U}_\infty(s, t) ds \\ (2.3) \quad & + \int_t^T \mathcal{U}_\infty^*(s, t)\mathcal{Q}_\infty(s)\mathcal{B}(s)\mathcal{R}^{-1}(s)\mathcal{B}^*(s)\mathcal{Q}_\infty(s)\mathcal{U}_\infty(s, t) ds, \end{aligned}$$

where $\mathcal{U}_\infty(t, s)$ is the mild evolution operator generated by

$$\mathcal{A}(t) - \mathcal{B}(t)\mathcal{R}^{-1}(t)\mathcal{B}^*(t)\mathcal{Q}_\infty(t).$$

We remark that these are extremely general conditions on $\mathcal{A}(t)$, essentially requiring that the abstract evolution equation

$$\begin{aligned} (1.1) \quad & \frac{dz(t)}{dt} = \mathcal{A}(t)z(t), \\ & z(0) = z_0, \quad z_0 \in \mathcal{D}(\mathcal{A}(0)), \end{aligned}$$

has a unique continuous solution with continuous dependence on initial conditions. In order to obtain a differentiated version of the Riccati equation, however, we do need stronger conditions.

3. The inner product Riccati equation. What we would like to be able to do is to differentiate (2.3) to obtain a differentiated version of the Riccati equation, but in general $\mathcal{U}_\infty(t, s)$ is only a mild evolution operator and there are problems involved with the domain of $\mathcal{A}(t)$. The crux of this problem is clarified in the following simple lemma.

LEMMA 8. *Consider $f(t) = \int_t^T \mathcal{C}(s)\mathcal{U}(s, t)y ds$, where $\mathcal{U}(t, s)$ is the evolution operator with properties (1.2) (i)–(v) generated by a closed operator $\mathcal{A}(t)$ on a Banach space X and $\mathcal{C}(\cdot) \in L_\infty(T; \mathcal{L}(X))$. If $y \in \bigcap_{t \in T} \mathcal{D}(\mathcal{A}(t))$ and $\sup_{t \in T} \|\mathcal{A}(t)y\|$ exists for each $y \in \bigcap_{t \in T} \mathcal{D}(\mathcal{A}(t))$, then f is differentiable and*

$$-\frac{df}{dt} = \mathcal{C}(t)y + \int_t^T \mathcal{C}(s)\mathcal{U}(s, t)\mathcal{A}(t)y ds.$$

Proof. $f(t)$ is well-defined by (1.2) (v) and assumption on \mathcal{C} . Now

$$\frac{df}{dt} = -\mathcal{C}(t)\mathcal{U}(t, t)y + \int_t^T \frac{\partial}{\partial t} [\mathcal{C}(s)\mathcal{U}(s, t)y] ds,$$

provided $(\partial/\partial t)(\mathcal{C}(s)\mathcal{U}(s, t)y)$ exists and is Bochner integrable and $\|(\partial/\partial t)(\mathcal{C}(s)\mathcal{U}(s, t)y)\| \leq \alpha(s)$, an integrable function independent of t . But $(\partial/\partial t)(\mathcal{C}(s)\mathcal{U}(s, t)y) = -\mathcal{C}(s)\mathcal{U}(s, t)\mathcal{A}(t)y$ by (1.2) (v) since $y \in \bigcap_{t \in T} \mathcal{D}(\mathcal{A}(t))$ and

$$\|\mathcal{C}(s)\mathcal{U}(s, t)\mathcal{A}(t)y\| \leq \text{const.} \|\mathcal{A}(t)y\|$$

(by (1.2)(v) and assumption on \mathcal{C})

$\leq \text{const.}$ by assumption.

$$\therefore -\frac{df}{dt} = \mathcal{C}(t)y + \int_t^T \mathcal{C}(s)\mathcal{U}(s, t)\mathcal{A}(t)y ds.$$

We now show that under fairly mild assumptions on (t) , the integral Riccati equation may be differentiated to obtain a differential inner product Riccati equation.

THEOREM 2. Assume the conditions of Theorem 1 and in addition

- (3.1) (a) $\sup_{t \in T} \|\mathcal{A}(t)y\|$ exists for each $y \in \bigcap_{t \in T} \mathcal{D}(\mathcal{A}(t))$,
 (b) $\mathcal{A}(t) - \mathcal{B}(t)\mathcal{R}^{-1}\mathcal{B}^*(t)\mathcal{P}(t)$ generates a strong evolution operator for any weakly continuous $\mathcal{P}(t) \in \mathcal{L}(\mathcal{H})$.

Then the following inner product Riccati equation has a weakly continuous solution $\mathcal{Q}_\infty(t) \in \mathcal{L}(\mathcal{H})$.

$$(3.2) \quad \left\langle \left[\frac{d\mathcal{Q}_\infty(t)}{dt} + \mathcal{A}^*(t)\mathcal{Q}_\infty(t) + \mathcal{Q}_\infty(t)\mathcal{A}(t) + \mathcal{W}(t) - \mathcal{Q}_\infty(t)\mathcal{B}(t)\mathcal{R}^{-1}(t)\mathcal{B}^*(t)\mathcal{Q}_\infty(t) \right] y, x \right\rangle = 0,$$

where $x, y \in \bigcap_{t \in T} \mathcal{D}(\mathcal{A}(t))$.

Proof. Assumption (3.1) (b) ensures that $\mathcal{U}_\infty(t, s)$ is a strong evolution operator with properties (1.2) (i)–(v). Let

$$\begin{aligned} Y_1(t) &= \langle \mathcal{U}_\infty^*(T, t)\mathcal{G}\mathcal{U}_\infty(T, t)x, y \rangle \quad \text{for } x, y \in \bigcap_{t \in T} \mathcal{D}(\mathcal{A}(t)) \\ &= \langle \mathcal{G}\mathcal{U}_\infty(T, t)x, \mathcal{U}_\infty(T, t)y \rangle. \end{aligned}$$

$$(3.3) \quad \therefore \frac{dY_1(t)}{dt} = -\langle \mathcal{G}\mathcal{U}_\infty(T, t)\mathcal{A}(t)x, \mathcal{U}_\infty(T, t)y \rangle - \langle \mathcal{G}\mathcal{U}_\infty(T, t)x, \mathcal{U}_\infty(T, t)\mathcal{A}(t)y \rangle$$

by (1.2) (v).

Let

$$\begin{aligned}
 Y_2(t) &= \left\langle \int_t^T \mathcal{U}_\infty^*(s, t) \mathcal{W}(s) \mathcal{U}_\infty(s, t) ds x, y \right\rangle \\
 &= \int_t^T \langle \mathcal{W}(s) \mathcal{U}_\infty(s, t)x, \mathcal{U}_\infty(s, t)y \rangle ds. \\
 \therefore \frac{dY_2(t)}{dt} &= -\langle \mathcal{W}(t)x, y \rangle - \int_t^T \langle \mathcal{W}(s) \mathcal{U}_\infty(s, t)\mathcal{A}(t)x, \mathcal{U}_\infty(s, t)y \rangle ds \\
 &\quad - \int_t^T \langle \mathcal{W}(s) \mathcal{U}_\infty(s, t)x, \mathcal{U}_\infty(s, t)\mathcal{A}(t)y \rangle ds
 \end{aligned}$$

by Lemma 7 since

$$\begin{aligned}
 &|\langle \mathcal{W}(s) \mathcal{U}_\infty(s, t)\mathcal{A}(t)x, \mathcal{U}_\infty(s, t)y \rangle| \\
 &\leq \|\mathcal{W}(s)\| \|\mathcal{U}_\infty(s, t)\| \|\mathcal{U}_\infty(s, t)\| \|y\| \|\mathcal{A}(t)x\| \\
 &< \text{const. independent of } t \text{ by (1.2)(v) and 3.1(a),}
 \end{aligned}$$

and similarly for the other term. Let

$$\begin{aligned}
 Y_3(t) &= \left\langle \int_t^T \mathcal{U}_\infty^*(s, t) \mathcal{Q}_\infty(s) \mathcal{B}(s) \mathcal{R}^{-1}(s) \mathcal{B}^*(s) \mathcal{Q}_\infty(s) \mathcal{U}_\infty(s, t) ds x, y \right\rangle \\
 &= \int_t^T \langle \mathcal{Q}_\infty(s) \mathcal{R}^{+1}(s) \mathcal{R}^{-1}(s) \mathcal{B}^*(s) \mathcal{Q}_\infty(s) \mathcal{U}_\infty(s, t)x, \mathcal{U}_\infty(s, t)y \rangle ds. \\
 \therefore \frac{dY_3(t)}{dt} &= -\langle \mathcal{Q}_\infty(t) \mathcal{B}(t) \mathcal{R}^{-1}(t) \mathcal{B}^*(t) \mathcal{Q}_\infty(t)x, y \rangle \\
 &\quad - \int_t^T \langle \mathcal{Q}_\infty(s) \mathcal{B}(s) \mathcal{R}^{-1}(s) \mathcal{B}^*(s) \mathcal{Q}_\infty(s) \mathcal{U}_\infty(s, t)\mathcal{A}(t)x, \mathcal{U}_\infty(s, t)y \rangle ds \\
 &\quad - \int_t^T \langle \mathcal{Q}_\infty(s) \mathcal{B}(s) \mathcal{R}^{-1}(s) \mathcal{B}^*(s) \mathcal{Q}_\infty(s) \mathcal{U}_\infty(s, t)x, \mathcal{U}_\infty(s, t)\mathcal{A}(t)y \rangle ds
 \end{aligned}$$

by Lemma 8 since

$$\begin{aligned}
 &|\langle \mathcal{Q}_\infty(s) \mathcal{B}(s) \mathcal{R}^{-1}(s) \mathcal{B}^*(s) \mathcal{Q}_\infty(s) \mathcal{U}_\infty(s, t)\mathcal{A}(t)x, \mathcal{U}_\infty(s, t)y \rangle| \\
 &\leq \text{const.} \|\mathcal{A}(t)x\| \|y\|, \quad \text{since all operators are uniformly bounded} \\
 &< \text{const. by (3.1)(a),} \quad \text{in norm on } T
 \end{aligned}$$

and similarly for the other term.

$$\begin{aligned}
 \therefore \left\langle \frac{d\mathcal{Q}_\infty(t)}{dt} x, y \right\rangle &= \frac{dY_1}{dt} + \frac{dY_2}{dt} + \frac{dY_3}{dt} \\
 &= -\langle \mathcal{Q}_\infty(t)\mathcal{A}(t)x, y \rangle - \langle \mathcal{Q}_\infty(t)x, \mathcal{A}(t)y \rangle \\
 &\quad - \langle [\mathcal{W}(t) + \mathcal{Q}_\infty(t) \mathcal{B}(t) \mathcal{R}^{-1}(t) \mathcal{B}^*(t) \mathcal{Q}_\infty(t)]x, y \rangle,
 \end{aligned}$$

where we have taken $\mathcal{A}(t)$ outside the integral, which is allowed since both versions exist and $\mathcal{A}(t)$ is closed.

This theorem is now applied to the Riccati equation arising in the linear quadratic cost problem for affine hereditary differential systems. For application to the dual Riccati equation arising in the filtering problem, see [3].

4. Application to the linear quadratic cost control problem for affine hereditary differential systems. We consider the linear quadratic cost control problem considered by Delfour and Mitter in [9]. The affine hereditary differential system is

$$(4.1) \quad \begin{aligned} \frac{dx(t)}{dt} &= A_{00}(t)x(t) + \sum_{i=1}^N A_i(t) \begin{cases} x(t + \theta_i), & t + \theta_i \geq 0 \\ h(t + \theta_i), & t + \theta_i < 0 \end{cases} \\ &+ \mathcal{B}(t)u(t) + \int_{-b}^0 A_{01}(t, \theta) \begin{cases} x(t + \theta), & t + \theta \geq 0 \\ h(t + \theta), & t + \theta < 0 \end{cases} d\theta, \\ x(0) &= h(0), \end{aligned}$$

where $t \in T = [0, T]$, $A_{00}, A_i \in L_\infty(T; \mathcal{L}(\mathbb{R}^n))$, $A_{01} \in L_\infty(T; [-b, 0]; \mathcal{L}(\mathbb{R}^n))$, $\mathcal{B}(\cdot) \in L_\infty(T; \mathcal{L}(\mathbb{R}^m, \mathbb{R}^n))$, $u(\cdot) \in L_\infty(T; \mathbb{R}^m)$ and

$$-b < -\theta_N < -\theta_{N-1} < \cdots < -\theta_1 < -\theta_0 = 0.$$

The spaces \mathcal{M}^2 and \mathcal{AC}^2 are defined as follows. Consider $\mathcal{L}^2(-b, 0; \mathbb{R}^n)$, the space of maps: $[-b, 0] \rightarrow \mathbb{R}^n$ under the seminorm $\|y\|_{\mathcal{M}^2} = [|y(0)|^2 + \int_{-b}^0 |y(\theta)|^2 d\theta]^{1/2}$, where $|\cdot|$ is the Euclidean norm on \mathbb{R}^n .

Then $\mathcal{M}^2(-b, 0; \mathbb{R}^n)$ is the quotient space of $\mathcal{L}^2(-b, 0; \mathbb{R}^n)$ generated by the equivalence classes under $\|\cdot\|_{\mathcal{M}^2}$, \mathcal{M}^2 is a Hilbert space and is isometrically isomorphic to $\mathbb{R}^n \times L_2(-b, 0; \mathbb{R}^n)$, $\mathcal{AC}^2(t_1, t; \mathbb{R}^n)$ is the space of absolutely continuous maps: $[t_0, t] \rightarrow \mathbb{R}^n$ with derivative in $L_2(t_0, t; \mathbb{R}^n)$ under the norm

$$\|x\|_{\mathcal{AC}^2} = \left[|x(t_0)|^2 + \int_{t_0}^t \left| \frac{dx(s)}{ds} \right|^2 ds \right]^{1/2},$$

and \mathcal{AC}^2 is a Hilbert space and is isometrically isomorphic to $\mathbb{R}^n \times L_2[t_0, t_1; \mathbb{R}^n]$.

In [8], Delfour and Mitter prove the following fundamental result for (4.1).

For the homogeneous case where $\mathcal{B} \equiv 0$ and the initial datum is $x(s) = h(0)$ at time s and $h \in \mathcal{D} = \mathcal{M}^2(-b, 0; \mathbb{R}^n) \cap \mathcal{AC}^2(-b, 0; \mathbb{R}^n)$, then (4.1) has a unique solution $\tilde{\phi}_s(\cdot; h) \in \mathcal{AC}^2(s, T; \mathbb{R}^n)$. The map: $(t, s) \rightarrow \tilde{\phi}_s(t; h)$ generates the 2-parameter semigroup $\tilde{\Phi}(t, s)$ satisfying:

$$(4.2) \quad \begin{aligned} &\text{(i) } \tilde{\Phi}(t, s) \in \mathcal{L}(\mathcal{M}^2) \quad \text{for } t \geq s \geq 0, \\ &\text{(ii) } \tilde{\Phi}(t, r)\tilde{\Phi}(r, s) = \tilde{\Phi}(t, s); \quad t \geq r \geq s \geq 0, \\ &\text{(iii) } \tilde{\Phi}(t, t) = \mathcal{I}, \text{ the identity operator in } \mathcal{L}(\mathcal{M}^2), \\ &\text{(iv) } \tilde{\Phi}(t, s) \text{ is strongly continuous in } s \text{ and } t \text{ for } s \leq t, \\ &\text{(v) } \tilde{\Phi}(t, s): \mathcal{D} \rightarrow \mathcal{D} \text{ and } \frac{\partial}{\partial t} \tilde{\Phi}(t, s)y = \tilde{\mathcal{A}}(t)\tilde{\Phi}(t, s)y \quad \forall y \in \mathcal{D}, \end{aligned}$$

where $\tilde{\mathcal{A}}(t)$ is a closed operator on \mathcal{M}^2 with domain \mathcal{D} and is defined by

$$[\tilde{\mathcal{A}}(t)h](\theta) = \begin{cases} \tilde{\mathcal{A}}^0(t)h & \text{for } \theta = 0, \\ [\tilde{\mathcal{A}}^1h](\theta) & \text{for } \theta \neq 0, \end{cases}$$

where $\tilde{\mathcal{A}}^0(t): \mathcal{D} \rightarrow \mathbb{R}^n$ and $\tilde{\mathcal{A}}^1: \mathcal{D} \rightarrow L_2[-b, 0; \mathbb{R}^n]$ are given by

$$\tilde{\mathcal{A}}^0(t)h = A_{00}(t)h(0) + \sum_{i=1}^N A_i(t)h(\theta_i) + \int_{-b}^0 A_{01}(t, \theta)h(\theta) d\theta$$

and

$$(\tilde{\mathcal{A}}^1h)(\theta) = \frac{dh(\theta)}{d\theta}.$$

So we see that this means that the homogeneous form of (4.1) may be expressed as an abstract evolution equation on \mathcal{M}^2 , where $\tilde{\mathcal{A}}(t)$ generates the strong evolution operator $\tilde{\Phi}(t, s)$:

$$\begin{aligned} \dot{\tilde{\phi}}(t) &= \tilde{\mathcal{A}}(t)\tilde{\phi}(t), \\ \tilde{\phi}(0) &= h \in \mathcal{D}. \end{aligned} \quad (4.3)$$

The inhomogeneous equation (4.1) may be expressed similarly:

$$\begin{aligned} \dot{\tilde{\phi}}(t) &= \tilde{\mathcal{A}}(t)\tilde{\phi}(t) + \tilde{\mathcal{B}}(t)u(t), \\ \tilde{\phi}(0) &= h \in \mathcal{D}, \end{aligned} \quad (4.4)$$

where $\tilde{\mathcal{B}}(\cdot) \in L_\infty(T; \mathcal{L}(\mathbb{R}^m, \mathcal{M}^2))$ is defined by

$$[\tilde{\mathcal{B}}(t)v](\theta) = \begin{cases} \mathcal{B}(t)v, & \theta = 0 \quad \text{for } v \in \mathbb{R}^m, \\ 0, & \theta \neq 0. \end{cases} \quad (4.5)$$

The quadratic cost for system (4.1) is

$$\begin{aligned} J(u) &= \langle x(T), \mathcal{G}x(T) \rangle + \int_0^T \langle x(s), \mathcal{W}(s)x(s) \rangle ds \\ &\quad + \int_0^T \langle u(s), \mathcal{R}(s)u(s) \rangle ds, \end{aligned} \quad (4.6)$$

where the inner product here is in \mathbb{R}^n or \mathbb{R}^m and the usual assumptions are that \mathcal{G} , $\mathcal{W}(\cdot)$ are symmetric and positive matrices, \mathcal{R} and $\mathcal{R}^{-1}(\cdot)$ are symmetric and strictly positive and \mathcal{W} and \mathcal{R} are L_∞ in s on T .

If we define $\tilde{\mathcal{G}}$, $\tilde{\mathcal{W}}(t)$, as operators in $\mathcal{L}(\mathcal{M}^2)$ corresponding to \mathcal{G} , and $\mathcal{W}(t)$, respectively, as we defined $\tilde{\mathcal{B}}$ from \mathcal{B} in (4.5), then the \mathcal{M}^2 -version of (4.6) is

$$\begin{aligned} J(u) &= \langle \tilde{\phi}(T), \tilde{\mathcal{G}}\tilde{\phi}(T) \rangle + \int_0^T \langle \tilde{\phi}(s), \tilde{\mathcal{W}}(s)\tilde{\phi}(s) \rangle ds \\ &\quad + \int_0^T \langle u(s), \tilde{\mathcal{R}}(s)u(s) \rangle ds, \end{aligned} \quad (4.7)$$

where the first two inner products are now in \mathcal{M}^2 .

We now have a linear quadratic cost control problem in \mathcal{M}^2 , defined by (4.4) and (4.7), which satisfies all the assumptions of Theorem 1 and also of Theorem 2 which follows from the following lemma.

LEMMA 9. (a) $\sup_{t \in T} \|\tilde{\mathcal{A}}(t)h\|_{\mathcal{M}^2}$ exists for each $h \in \mathcal{D}$.

(b) The inhomogeneous equation (4.4) in \mathcal{M}^2 has a unique solution for $u(t) \in L_2(T; \mathbb{R}^n)$.

Proof. (a)

$$\|\tilde{\mathcal{A}}(t)h\|_{\mathcal{M}^2}^2 = \left| A_{00}(t)h(0) + \sum_{i=1}^N A_i(t)h(\theta_i) + \int_{-b}^0 A_{01}(t, \theta)h(\theta) d\theta \right|^2 + \int_{-b}^0 \left| \frac{dh}{d\theta} \right|^2 d\theta$$

$< \text{const. independent of } t$, since A_{00} , A_i and A_{01} are L_∞ in t on T .

(b) See Delfour and Mitter [8].

So the condition (1.6)(a) of Lemma 2 is satisfied and $\mathcal{A}(t) - \tilde{\mathcal{B}}(t)\mathcal{R}^{-1}(t) \cdot \tilde{\mathcal{B}}^*(t)\mathcal{P}(t)$ generates a strong evolution operator for any weakly continuous $\mathcal{P}(t)$ (taking $u(t) = -\mathcal{R}^{-1}(t)\tilde{\mathcal{B}}^*(t)\mathcal{P}(t)\phi(t)$).

So applying Theorems 1 and 2, we have proved that the \mathcal{M}^2 -version of (4.4), (4.7) of the linear quadratic cost control problem for the affine hereditary differential system has a unique minimizing control $u(t) = -\tilde{\mathcal{R}}^{-1}(t)\tilde{\mathcal{B}}^*(t)\mathcal{P}(t)\tilde{\phi}(t)$ and minimum cost $\langle h, \mathcal{P}_\infty(0)h \rangle_{\mathcal{M}^2}$, where $\mathcal{P}(t)$ is a self-adjoint weakly continuous operator in $\mathcal{L}(\mathcal{M}^2)$, which is the unique solution of the integral Riccati equation and satisfies:

$$\left\langle \left[\frac{d\mathcal{P}}{dt} + \tilde{\mathcal{A}}^*(t)\mathcal{P}(t) + \mathcal{P}(t)\tilde{\mathcal{A}}(t) + \tilde{\mathcal{W}}(t) - \mathcal{P}(t)\tilde{\mathcal{B}}(t)\tilde{\mathcal{R}}^{-1}(t)\tilde{\mathcal{B}}^*(t)\mathcal{P}(t) \right] y, x \right\rangle_{\mathcal{M}^2} = 0,$$

where $x, y \in \mathcal{D}$, the domain of $\tilde{\mathcal{A}}(t)$.

This agrees exactly with the results of Delfour and Mitter in [9].

Author's note. In this paper, by $L_\infty(T; \mathcal{L}(X))$ we mean the space of $\mathcal{L}(X)$ -valued functions which are strongly measurable and uniformly bounded in norm on T . Terms like

$$\int_s^t \mathcal{U}(t, \alpha)\mathcal{B}(\alpha)\mathcal{Y}(\alpha, s) d\alpha$$

are shorthand notations for the strong Bochner integral. Finally, $\langle \mathcal{A}^*(t)\mathcal{Q}_\infty(t)x, y \rangle$ in (3.2) of course means $\langle \mathcal{Q}_\infty(t)y, \mathcal{A}(t)x \rangle$.

Acknowledgments. I would like to thank Richard Vinter for his helpful criticisms.

REFERENCES

- [1] A. V. BALAKRISHNAN, *Optimal control problems in Banach spaces*, this Journal, 3 (1965), pp. 152–180.
- [2] RUTH F. CURTAIN, *Infinite-dimensional filtering*, this Journal, 13 (1975), pp. 89–104.
- [3] ———, *A Kalman Bucy filtering theory for affine hereditary differential systems*, Control Theory Centre Rep. 25, University of Warwick, Coventry, England, 1973.
- [4] RUTH F. CURTAIN AND A. J. PRITCHARD, *The infinite-dimensional Riccati equation*, Control Theory Centre Rep. 11, University of Warwick, 1972, J. Math. Anal Appl., to appear.

- [5] G. DA PRATO, *Equations d'évolution dans des algèbres d'opérateurs et application à des équations quasi-linéaires*, J. Math. Pures Appl., 48 (1969), pp. 59–107.
- [6] R. DATKO, *A linear control problem in abstract Hilbert space*, J. Differential Equations, 9 (1971), pp. 346–359.
- [7] ———, *Unconstrained problems with quadratic cost*, this Journal, 11 (1973), pp. 32–52.
- [8] M. C. DELFOUR AND S. K. MITTER, *Hereditary differential systems with constant delays I—General case*, J. Differential Equations, 12 (1972), pp. 213–235; and *II—A class of affine systems and the adjoint problem*, Ibid., to appear.
- [9] ———, *Controllability, observability and optimal feedback control of hereditary differential systems*, this Journal, 10 (1972), pp. 298–328.
- [10] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Part 1: General Theory*, Interscience, New York, 1958.
- [11] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semi-groups*, Amer. Math. Soc. Coll. Publ., 31, revised Ed., 1957.
- [12] T. KATO, *Abstract Evolution Equation of Parabolic type in Banach and Hilbert Space*, Nagoya Math. J., 19 (1961), pp. 93–125.
- [13] T. KATO AND H. TANABE, *On the abstract evolution equation*, Osaka Math. J., 14 (1962), pp. 107–133.
- [14] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.
- [15] D. L. LUKES AND D. L. RUSSELL, *The quadratic criterion for distributed systems*, this Journal, 7 (1969), pp. 101–121.
- [16] A. J. PRITCHARD, *Stability and control of distributed parameter systems governed by wave equations*, IFAC Conference on Distributed Parameter Systems, Banff, Canada, 1971.

COPRIME FACTORIZATIONS AND STABILITY OF MULTIVARIABLE DISTRIBUTED FEEDBACK SYSTEMS*

MATHUKUMALLI VIDYASAGAR†

Abstract. The stability of multivariable feedback systems presents different problems from the stability of single loop feedback systems, owing mainly to the complexities of “pole”-“zero” cancellation in the multivariate case. In this paper, the “coprime factorization” of a nonrational transfer function matrix is defined and is used in studying the stability of multivariable distributed feedback systems. However, the stability results based on coprime factorizations, though they are quite elegant, do not lead to readily applicable testing procedures. For this reason, we introduce the notion of “pseudo-coprime” factorizations. These also lead to many stability theorems. As a special case of these stability results, we obtain explicit necessary and sufficient conditions for the stability of a multivariable feedback system whose open loop transfer function contains a finite number of poles in the closed right half-plane, but is otherwise stable. These results significantly generalize those of Callier and Desoer [8].

1. Introduction. The stability of multivariable feedback systems presents different problems from the stability of single loop feedback systems, owing mainly to the difficulties of “pole”-“zero” cancellations in the multivariable case. For systems with rational transfer functions, the stability question is more or less completely resolved in some recent work of Desoer and Schulman [9], [10], who use the concept of a coprime factorization of a rational transfer function matrix. It is possible to define a coprime factorization of nonrational transfer matrices also and to study the stability of distributed multivariable feedback systems using such factorizations. This is done in the present paper. However, the stability results based on coprime factorizations of nonrational transfer matrices, though they are quite elegant, do not lead to readily applicable testing procedures. For this reason, we introduce the notion of “pseudo-coprime” factorizations. These also lead to many stability theorems, which are quite analogous to results obtained using coprime factorizations, except that the proofs are somewhat messier. As a special case of the stability theorems developed using pseudo-coprime factorizations, we obtain explicit necessary and sufficient conditions for the stability of a multivariable feedback system whose open loop transfer function contains a finite number of poles in the closed right half-plane but is otherwise stable. These conditions significantly generalize those of Callier and Desoer [8].

2. Summary of known results. We consider first the rational case. Let $\hat{G}(\cdot)$ and $\hat{F}(\cdot)$ be $n \times n$ matrices whose elements are proper rational functions of the complex variable s , and let $\hat{H}(\cdot)$ be defined by $\hat{H}(s) = \hat{G}(s)(I + \hat{F}(s) \cdot \hat{G}(s))^{-1}$. Then the elements of $\hat{H}(\cdot)$ are also proper rational functions, and furthermore, one can prove the following identity [2], [11]:

$$(1) \quad \det(I + \hat{F}(s)\hat{G}(s)) = \det(I + \hat{F}(\infty)\hat{G}(\infty)) \prod_{i=1}^n \frac{(s - p_h^{(i)})}{(s - p_f^{(i)})(s - p_g^{(i)})},$$

where $p_f^{(i)}$, $p_g^{(i)}$ and $p_h^{(i)}$ are the poles of $\hat{F}(\cdot)$, $\hat{G}(\cdot)$, and $\hat{H}(\cdot)$, respectively. It is well known that under the present assumptions, a system with the transfer function

* Received by the editors March 29, 1974.

† Department of Electrical Engineering, Concordia University, Montreal, Quebec, Canada.

$\hat{H}(\cdot)$ is BIBO stable if and only if $\operatorname{Re} p_h^{(i)} < 0$ for all i , i.e., if $\hat{H}(\cdot)$ has no poles in the closed right half-plane $C_+ = \{s: \operatorname{Re} s \geq 0\}$. In view of (1), it is clear that a necessary condition for $\hat{H}(\cdot)$ to have no poles in C_+ is that $\det(I + \hat{F}(\cdot)\hat{G}(\cdot))$ has no zeros in C_+ . However, the condition that $\det(I + \hat{F}(\cdot)\hat{G}(\cdot))$ has no zeros in C_+ is not sufficient to insure that $\hat{H}(\cdot)$ has no zeros in C_+ , because it is possible for some $p_h^{(i)}$ to coincide with some $p_g^{(i)}$ or $p_f^{(i)}$. In other words, both the open loop transfer function and the closed loop transfer function of a multivariable system can have poles at the same point, and such a pole cannot be detected by studying $\det(I + \hat{F}(\cdot)\hat{G}(\cdot))$.

As an example, let

$$\hat{G}(s) = \begin{bmatrix} \frac{1}{s-1} & 0 \\ 0 & \frac{1}{s+1} \end{bmatrix}, \quad \hat{F}(s) = \begin{bmatrix} 2 & 2 \\ 2 & 0 \end{bmatrix}.$$

Then

$$I + \hat{F}(s)\hat{G}(s) = \begin{bmatrix} \frac{s+1}{s-1} & \frac{2}{s+1} \\ \frac{2}{s-1} & 1 \end{bmatrix}, \quad \det[I + \hat{F}(s)\hat{G}(s)] = \frac{s+3}{s+1},$$

so $\det[I + \hat{F}(s)\hat{G}(s)]$ has no zeros in C_+ ; yet we have

$$\hat{H}(s) = \hat{G}(s)[I + \hat{F}(s)\hat{G}(s)]^{-1} = \begin{bmatrix} \frac{s+1}{(s-1)(s+3)} & \frac{-2}{(s-1)(s+3)} \\ \frac{-2}{(s-1)(s+3)} & \frac{s+1}{(s-1)(s+3)} \end{bmatrix},$$

which clearly represents an unstable system.

The solution proposed by Desoer and Schulman [9], [10] is to express both $\hat{G}(\cdot)$ and $\hat{F}(\cdot)$ in the form

$$(2) \quad \hat{F}(s) = \hat{N}_F(s) \cdot [\hat{D}_F(s)]^{-1},$$

$$(3) \quad \hat{G}(s) = \hat{N}_G(s) \cdot [\hat{D}_G(s)]^{-1},$$

where \hat{N}_F , \hat{N}_G , \hat{D}_F , \hat{D}_G are polynomial matrices, and further the pairs (\hat{N}_F, \hat{D}_F) , (\hat{N}_G, \hat{D}_G) are right-coprime, i.e., there exist polynomial matrices \hat{P}_F , \hat{Q}_F , \hat{P}_G , \hat{Q}_G such that

$$(4) \quad \hat{P}_F(s) \cdot \hat{N}_F(s) + \hat{Q}_F(s) \cdot \hat{D}_F(s) = I \quad \text{for all } s,$$

$$(5) \quad \hat{P}_G(s) \cdot \hat{N}_G(s) + \hat{Q}_G(s) \cdot \hat{D}_G(s) = I \quad \text{for all } s,$$

where I is the identity matrix. The conditions (4) and (5) can be interpreted very simply: The set of all rational $n \times n$ matrices constitutes a noncommutative ring, and (4) implies that \hat{N}_F and \hat{D}_F have no common right divisor other than one whose inverse is also an element of this ring. Equation (5) can be interpreted similarly. The ordered pair (\hat{N}_F, \hat{D}_F) is called a right-coprime factorization of \hat{F} ; similarly,

the ordered pair (\hat{N}_G, \hat{D}_G) is called a right-coprime factorization of \hat{G} . Once \hat{F} and \hat{G} are suitably factorized, (an algorithm for finding a suitable factorization is given by Wang [5]), Desoer and Schulman proceed to give necessary and sufficient conditions for $\hat{G}(I + \hat{F}\hat{G})^{-1}$ to correspond to a stable system.

We turn now to distributed systems. The following notation is used to facilitate the presentation:

C = set of complex numbers;

$C_+ = \{s: \operatorname{Re} s \geq 0\}$;

$\hat{\cdot}$: indicates Laplace transform;

$\mathcal{A}^{n \times n}, \hat{\mathcal{A}}^{n \times n}$: Banach algebras defined in [4];

$*$: convolution;

L.t.d. = Laplace transformable distribution;

condition (N): an ordered pair of functions (\hat{a}, \hat{b}) , where $\hat{a}, \hat{b}: C \rightarrow C$ is said to satisfy condition (N) if, whenever $(s_i)_{i=1}^\infty$ is a sequence in C_+ with $\lim_{i \rightarrow \infty} \hat{a}(s_i) = 0$, we have $\liminf_{i \rightarrow \infty} |\hat{b}(s_i)| > 0$.

The following theorems are slight generalizations of [6, Thm. 1] and [7, Thms. 1 and 2].

THEOREM A. Let $G(\cdot)$ be a matrix of distributions with support in $[0, \infty)$, and let $F(\cdot) \in \mathcal{A}^{n \times n}$. Suppose that in some neighborhood of the origin, G contains at most impulse functions, and suppose further that the equation

$$(6) \quad H + H * F * G = G$$

has a unique solution for $H(\cdot)$. Under these conditions, if $H(\cdot) \in \mathcal{A}^{n \times n}$, then

(i) $G(\cdot)$ is Laplace transformable; for some $\alpha > 0$, $G(\cdot) \in \mathcal{A}^{n \times n}(\alpha)$, $G(\cdot)$ is analytic in $\operatorname{Re} s > \alpha$ and can be continued to a meromorphic function in $\operatorname{Re} s > 0$;

(ii) $\hat{G}(s)$ is of the form

$$(7) \quad \hat{G}(s) = \hat{N}(s)[\hat{D}(s)]^{-1} \quad \text{for all } s \in C_+,$$

where $\hat{N}(\cdot), \hat{D}(\cdot) \in \hat{\mathcal{A}}^{n \times n}$;

(iii)

$$(8) \quad \inf_{s \in C_+} |\det(I + F(s)G(s))| > 0.$$

THEOREM B. Suppose $G(\cdot)$ is an L.t.d., $F(\cdot) \in \mathcal{A}^{n \times n}$, and suppose (6) determines H uniquely. (Hence $\hat{H} = \hat{G}(I + \hat{F}\hat{G})^{-1}$ over the common half-plane of convergence). Under these conditions, $\hat{H}(\cdot) \in \mathcal{A}^{n \times n}$ if and only if

(i) there exist \hat{N}, \hat{D} in $\hat{\mathcal{A}}^{n \times n}$ such that

$$(9) \quad \hat{G}(s) = \hat{N}(s)[\hat{D}(s)]^{-1} \quad \forall s \in C_+$$

and such that the ordered pair $(\det \hat{D}, \det(\hat{D} + \hat{F}\hat{N}))$ satisfies condition (N), and

(ii)

$$(10) \quad \inf_{s \in C_+} |\det(I + \hat{F}(s)\hat{G}(s))| > 0.$$

Since the proofs of these theorems are only minor modifications of those of corresponding theorems in [6], [7], we omit them here in the interests of brevity.

The situation as it exists for distributed multivariable systems vis-a-vis that for lumped multivariable systems is as follows: If $\hat{G}(s)$ and $\hat{F}(\cdot)$ are proper rational matrices, then one can apply Wang's algorithm [5] to obtain polynomial right-coprime factorizations for both $\hat{G}(\cdot)$ and $\hat{F}(\cdot)$ and apply the definitive tests of Desoer and Schulman [9], [10]. In the distributed case, the situation is more complicated. Suppose we are given an L.t.d. $G(\cdot)$ and $F(\cdot) \in \mathcal{A}^{n \times n}$. In order to determine whether or not $\hat{H} = \hat{G}(I + \hat{F}\hat{G})^{-1}$ belongs to $\mathcal{A}^{n \times n}$, we would first check the condition (8). If (8) fails, then definitely $H \notin \mathcal{A}^{n \times n}$, in view of Theorem A. On the other hand, suppose (8) holds. This alone does not imply that $\hat{H} \in \mathcal{A}^{n \times n}$. (Indeed, the example given earlier serves to demonstrate this fact.) Thus in order to conclude that $H \in \mathcal{A}^{n \times n}$, in addition to (2), we use the added assumption that \hat{G} is the form (9), where the appropriate condition (N) is satisfied. This brings us to the following question: Suppose we express a given \hat{G} in the form (9), but the corresponding condition (N) fails to hold; what, if anything, can we conclude? The answer is, in general, nothing. However, if the ordered pair (\hat{N}, \hat{D}) constitutes a so-called right-coprime factorization of \hat{G} , or a pseudo-right-coprime factorization of \hat{G} , then definitive conclusions can be drawn. The purpose of this paper is to introduce these two concepts and to prove several stability results based on such factorizations.

3. Right-coprime factorizations. We first introduce the notion of right-coprime factorizations.

DEFINITION 1. A pair of elements (N, D) in $\mathcal{A}^{n \times n}$ is said to be *right-coprime* in $\mathcal{A}^{n \times n}$ if there exist elements \hat{P}, \hat{Q} in $\mathcal{A}^{n \times n}$ such that

$$(11) \quad \hat{P}(s)\hat{N}(s) + \hat{Q}(s)\hat{D}(s) = I \quad \text{for all } s \in C_+.$$

DEFINITION 2. Given an L.t.d. $G(\cdot)$ with support in $[0, \infty)$, the ordered pair (\hat{N}, \hat{D}) is said to be a *right-coprime factorization* (r.c.f.) of \hat{G} in $\mathcal{A}^{n \times n}$ if

(i)

$$(12) \quad \hat{G}(s) = \hat{N}(s)[\hat{D}(s)]^{-1} \quad \text{for all } s \in C_+$$

and

(ii) the pair (\hat{N}, \hat{D}) is right-coprime in $\mathcal{A}^{n \times n}$.

Remarks. The concept of an r.c.f. in $\mathcal{A}^{n \times n}$ of a transfer-function matrix should not be confused with the similar yet distinct concept of *polynomial* coprime factorizations as detailed in § 2. The concepts are similar in that $\mathcal{A}^{n \times n}$ is a non-commutative ring, just as the set of $n \times n$ polynomial matrices also is a non-commutative ring. But an important fact to note is the following: Suppose $\hat{G}(\cdot)$ is a matrix whose elements are proper rational functions. Then it is known [2] that $\hat{G}(\cdot)$ has a polynomial right-coprime factorization, and in fact, an algorithm exists for finding such a factorization. In contrast, it is not even known whether or not $\hat{G}(\cdot)$ has, in general, an r.c.f. in $\mathcal{A}^{n \times n}$.

We now state a few results that lead to stability conditions.

PROPOSITION 1. Let $G(\cdot)$ be an L.t.d. with support in $[0, \infty)$, and let (\hat{N}, \hat{D}) be an r.c.f. in $\mathcal{A}^{n \times n}$ of G . Then $\hat{G} \in \mathcal{A}^{n \times n}$ if and only if

$$(13) \quad \inf_{s \in C_+} |\det \hat{D}(s)| > 0.$$

Proof. If. Suppose (13) holds. Then by [1, Thm. 4.18.6, p. 150], it follows that $[1/\det \hat{D}(\cdot)] \in \mathcal{A}$, whence $[\hat{D}(\cdot)]^{-1} \in \mathcal{A}^{n \times n}$. Therefore \hat{G} is a product of two elements of $\mathcal{A}^{n \times n}$, and as such itself belongs to $\mathcal{A}^{n \times n}$.

Only if. Suppose (13) is violated, and let $(s_i)_{i=1}^\infty$ be a sequence in C_+ such that $\det \hat{D}(s_i) \rightarrow 0$. Since \hat{N} and \hat{D} are right-coprime, there exist \hat{P} and \hat{Q} in $\mathcal{A}^{n \times n}$ such that

$$\hat{P}(s)\hat{N}(s) + \hat{Q}(s)\hat{D}(s) = I \quad \text{for all } s \in C_+.$$

Postmultiplying both sides by $[\hat{D}(s)]^{-1}$, we get

$$(14) \quad \hat{P}(s)\hat{G}(s) + \hat{Q}(s) = [\hat{D}(s)]^{-1}.$$

If we replace s by s_i in (14) and let $i \rightarrow \infty$, the right side becomes unbounded, whence so must the left side. However, $\hat{P}(\cdot)$ and $\hat{Q}(\cdot)$ are both bounded over C_+ since they belong to $\mathcal{A}^{n \times n}$. This shows that $\hat{G}(s_i)$ must become unbounded as $i \rightarrow \infty$, whence $\hat{G} \notin \mathcal{A}^{n \times n}$. \square

Remarks. The point of Proposition 1 can be explained as follows: Suppose $\hat{G}(s) = [\hat{D}(s)]^{-1}$ where $\hat{N}, \hat{D} \in \mathcal{A}^{n \times n}$, but we assume nothing more. Then $\hat{G} \in \mathcal{A}^{n \times n}$ if (13) holds; but failure of (13) does not in general imply that $\hat{G} \notin \mathcal{A}^{n \times n}$ because, in general, $\hat{G}(s)$ need not become unbounded as $\det \hat{D}(s) \rightarrow 0$ with s in C_+ . However, if we add the assumption that \hat{N}, \hat{D} are right-coprime, then (13) becomes a necessary and sufficient condition for \hat{G} to belong to $\mathcal{A}^{n \times n}$. In other words, once we find an r.c.f. for \hat{G} , we can state that $\hat{G} \in \mathcal{A}^{n \times n}$ if and only if the determinant of its “denominator” is bounded away from zero over C_+ .

LEMMA 1. Let $G(\cdot)$ be an L.t.d. with support in $[0, \infty)$; let $\hat{F} \in \mathcal{A}^{n \times n}$. Let (\hat{N}, \hat{D}) be an r.c.f. in $\mathcal{A}^{n \times n}$ of \hat{G} , and let $\hat{H} = \hat{G}(I + \hat{F}\hat{G})^{-1}$. Then the ordered pair $(\hat{N}, \hat{D} + \hat{F}\hat{N})$ is an r.c.f. in $\mathcal{A}^{n \times n}$ of \hat{H} .

Proof. Clearly $\hat{H} = \hat{N}(\hat{D} + \hat{F}\hat{N})^{-1}$, and it only remains to show that the pair $(\hat{N}, \hat{D} + \hat{F}\hat{N})$ is right-coprime. Since \hat{N} and \hat{D} are right-coprime by assumption, there exist \hat{P}, \hat{Q} in $\mathcal{A}^{n \times n}$ such that

$$\hat{P}(s)\hat{N}(s) + \hat{Q}(s)\hat{D}(s) = I \quad \text{for all } s \in C_+.$$

A little manipulation yields that

$$[\hat{P}(s) - \hat{Q}(s)\hat{F}(s)]\hat{N}(s) + \hat{Q}(s)[\hat{D}(s) + \hat{F}(s)\hat{N}(s)] = I \quad \text{for all } s \in C_+.$$

Since $\hat{P} - \hat{Q}\hat{F}$ belongs to $\mathcal{A}^{n \times n}$, this shows that the pair $(\hat{N}, \hat{D} + \hat{F}\hat{N})$ is right-coprime. \square

Remarks. Consider a system with \hat{G} in the forward path and \hat{F} in the feedback path. Then $\hat{H} = \hat{G}(I + \hat{F}\hat{G})$ is clearly the gain of the closed-loop system. Lemma 1 shows that if we can find an r.c.f. in $\mathcal{A}^{n \times n}$ of \hat{G} and if $\hat{F} \in \hat{\mathcal{A}}^{n \times n}$, then we can readily find an r.c.f. of \hat{H} . In view of Proposition 1, once we have an r.c.f. in $\hat{\mathcal{A}}^{n \times n}$ of \hat{H} , we can readily ascertain whether or not $H \in \hat{\mathcal{A}}^{n \times n}$.

The main stability result based on r.c.f.'s is given next.

THEOREM 1. *Let $G(\cdot)$ be an L.t.d. with support in $[0, \infty)$; let $F \in \hat{\mathcal{A}}^{n \times n}$, and let $\hat{H} = \hat{G}(I + \hat{F}\hat{G})^{-1}$. The $\hat{H} \in \hat{\mathcal{A}}^{n \times n}$ if and only if \hat{G} has an r.c.f. (\hat{N}, \hat{D}) in $\hat{\mathcal{A}}^{n \times n}$ such that*

$$(15) \quad \inf_{s \in C_+} |\det(\hat{D}(s) + \hat{F}(s)\hat{N}(s))| > 0.$$

Proof. If. Suppose that (\hat{N}, \hat{D}) is an r.c.f. in $\hat{\mathcal{A}}^{n \times n}$ of \hat{G} and that (15) holds. By Lemma 1, we have that $(\hat{N}, \hat{D} + \hat{F}\hat{N})$ is an r.c.f. in $\hat{\mathcal{A}}^{n \times n}$ of \hat{H} . Thus by (15) and Proposition 1, it follows that $\hat{H} \in \hat{\mathcal{A}}^{n \times n}$.

Only if. Suppose $\hat{H} \in \hat{\mathcal{A}}^{n \times n}$. Since $\hat{H} + \hat{G}(I + \hat{F}\hat{G})^{-1}$, we also have $\hat{G} = \hat{H}(I - \hat{F}\hat{H})^{-1}$. Accordingly, define $\hat{N} = \hat{H}$; $\hat{D} = I - \hat{F}\hat{H}$. Then $\hat{N}, \hat{D} \in \hat{\mathcal{A}}^{n \times n}$, $\hat{G} = \hat{N}\hat{D}^{-1}$, and moreover,

$$\hat{F}(s)\hat{N}(s) + \hat{D}(s) = I \quad \text{for all } s.$$

This shows that (i) the pair (\hat{N}, \hat{D}) is right-coprime whence the ordered pair (\hat{N}, \hat{D}) is an r.c.f. in $\hat{\mathcal{A}}^{n \times n}$ of \hat{G} , and (ii) (15) is satisfied. \square

COROLLARY 1. *Let (\hat{N}, \hat{D}) be an r.c.f. in $\hat{\mathcal{A}}^{n \times n}$ of \hat{G} ; let $\hat{F} \in \hat{\mathcal{A}}^{n \times n}$, and let $\hat{H} = \hat{G}(I + \hat{F}\hat{G})^{-1}$. Then $\hat{H} \in \hat{\mathcal{A}}^{n \times n}$ if and only if (15) holds.*

It is desirable to restate Theorem 1 in such a way that in (15) the dependence on G is brought out more explicitly. The following lemma is useful for this purpose.

LEMMA 2. *Let $\hat{D}, \hat{F}, \hat{N}$ be arbitrary elements of $\hat{\mathcal{A}}^{n \times n}$, and let $\hat{G} = \hat{N}\hat{D}^{-1}$. Then*

$$(16) \quad \inf_{s \in C_+} |\det(\hat{D}(s) + \hat{F}(s)\hat{N}(s))| > 0$$

if and only if

$$(17) \quad \inf_{s \in C_+} |\det(I + \hat{F}(s)\hat{G}(s))| > 0$$

and the ordered pair

$$(18) \quad \det \hat{D}, \det(\hat{D} + \hat{F}\hat{N})$$

satisfies the condition (N).

Proof. If. This involves only routine arguments, and a proof can be found following [6, Thm. 1].

Only if. First, suppose (17) fails. Since $\det(\hat{D} + \hat{F}\hat{N}) = \det \hat{D} \cdot \det(I + \hat{F}\hat{G})$, and since $\det \hat{D}(\cdot)$ is bounded over C_+ , we see that (16) also fails. Next, suppose (18) fails. Then there exists a sequence $(s_i)_{i=1}^\infty$ in C_+ such that $\det \hat{D}(s_i) \rightarrow 0$ and $\det(\hat{D}(s_i) + \hat{F}(s_i)\hat{N}(s_i)) \rightarrow 0$ as $i \rightarrow \infty$. But this means that (16) fails. \square

Combining Lemma 2 and Theorem 1, we get what is essentially an alternate form of Theorem 1.

THEOREM 2. Let G, F, \hat{H} be as in Theorem 1. Then $\hat{H} \in \mathcal{A}^{n \times n}$ if and only if there exists an r.c.f. (\hat{N}, \hat{D}) in $\mathcal{A}^{n \times n}$ of \hat{G} such that (17) and (18) hold.

Proof. The proof is immediate from Lemma 2 and Theorem 1.

COROLLARY 2. Let $G, F, \hat{N}, \hat{D}, \hat{H}$ be as in Corollary 1. Then $\hat{H} \in \mathcal{A}^{n \times n}$ if and only if (17) and (18) hold.

The stability theorems above demonstrate the importance of being able to determine an r.c.f. in $\mathcal{A}^{n \times n}$ for a given transfer matrix \hat{G} . This task is rather difficult in general, but is made a little simpler by the following result, which shows that it is only necessary to find an r.c.f. in $\mathcal{A}^{n \times n}$ for the “unstable part” of \hat{G} .

PROPOSITION 2. Let $\hat{G} = \hat{G}_b + \hat{G}_u$, where $\hat{G}_b \in \mathcal{A}^{n \times n}$, and suppose the ordered pair (\hat{N}, \hat{D}) is an r.c.f. in $\mathcal{A}^{n \times n}$ of \hat{G}_u . Then $(\hat{N} + \hat{G}_b \hat{D}, \hat{D})$ is an r.c.f. in $\mathcal{A}^{n \times n}$ of \hat{G} .

Proof. Clearly, $\hat{N} + \hat{G}_b \hat{D}$ and \hat{D} both belong to $\mathcal{A}^{n \times n}$, and $\hat{G} = (\hat{N} + \hat{G}_b \hat{D}) \cdot \hat{D}^{-1}$. Thus it only remains to show that $\hat{N} + \hat{G}_b \hat{D}$ and \hat{D} are right-coprime. By assumption, there exist \hat{P}, \hat{Q} in $\mathcal{A}^{n \times n}$ such that

$$\hat{P}(s)\hat{N}(s) + \hat{Q}(s)\hat{D}(s) = I \quad \text{for all } s \in C_+,$$

so we have

$$\hat{P}(s)[\hat{N}(s) + \hat{G}_b(s)\hat{D}(s)] + [\hat{Q}(s) - \hat{P}(s)\hat{G}_b(s)]\hat{D}(s) = I \quad \text{for all } s \in C_+,$$

whence $\hat{N} + \hat{G}_b \hat{D}$ and \hat{D} are right-coprime. \square

4. Pseudo-right-coprime factorizations. While the stability theorems based on r.c.f.’s are quite elegant, they do not seem to be easy to use for testing stability in specific situations, owing mainly to the difficulties of finding an r.c.f. in $\mathcal{A}^{n \times n}$ for a given transfer function. As mentioned before, it is not clear how to find an r.c.f. in $\mathcal{A}^{n \times n}$ for a given \hat{G} even in the simple (and practically significant) case where all elements of \hat{G} are proper rational matrices. To overcome these difficulties, we introduce pseudo-right-coprime factorizations. These lead to a theory that is more clumsy than that involving r.c.f.’s, but is at the same time more readily applicable.

DEFINITION 3. A pair of elements (\hat{N}, \hat{D}) , where $\hat{N}, \hat{D} \in \mathcal{A}^{n \times n}$, is said to be *pseudo-right-coprime* (p.r.c.) in $\mathcal{A}^{n \times n}$ if there exist elements $\hat{U}, \hat{V}, \hat{W}$ in $\mathcal{A}^{n \times n}$ such that (i) $\det \hat{W}(s) \neq 0$ whenever $s \in C_+$ and (ii)

$$(19) \quad \hat{U}(s)\hat{N}(s) + \hat{V}(s)\hat{D}(s) = \hat{W}(s) \quad \text{for all } s \in C_+.$$

DEFINITION 4. Given an L.t.d. G with support in $[0, \infty)$, the ordered pair (N, D) is said to be a *pseudo-right-coprime factorization* (p.r.c.f.) of \hat{G} in $\mathcal{A}^{n \times n}$ if

- (i) $\hat{G}(s) = \hat{N}(s)[\hat{D}(s)]^{-1}$ for all $s \in C_+$,
- (ii) the pair (\hat{N}, \hat{D}) is p.r.c.,
- (iii) whenever $(s_i)_{i=1}^\infty$ is a sequence in C_+ with $|s_i| \rightarrow \infty$, we have $\lim_{i \rightarrow \infty} \inf |\det \hat{D}(s_i)| > 0$.

Remarks. Comparing Definition 1 with Definition 3, we see that the main difference between an r.c. pair and a p.r.c. pair is that the identity matrix I in the right-hand side of (11) is replaced by $\hat{W}(s)$ in (19), where $\det \hat{W}(s) \neq 0$ whenever

$s \in C_+$. So, loosely speaking, $\hat{W}(s)$ can be inverted at all $s \in C_+$ and therefore behaves like the identity matrix; but it is possible that $\det \hat{W}(s) \rightarrow 0$ as $|s| \rightarrow \infty$ with $s \in C_+$. Similarly, comparing Definitions 2 and 4, we see that a *necessary* condition for a given Laplace transform $\hat{G}(\cdot)$ to have a p.r.c.f. is that all singularities of $\hat{G}(\cdot)$ in C_+ are contained in a bounded subset of C_+ .

The practical significance p.r.c.f.'s is illustrated by the following result.

PROPOSITION 3. *Let $\hat{G}(s)$ be an $n \times n$ matrix whose elements are proper rational functions of s . Then \hat{G} has a p.r.c.f.*

Proof. Under the given hypothesis on \hat{G} , there exists a polynomial r.c.f. of \hat{G} , say (Λ, Γ) . In other words, Λ and Γ are polynomial matrices such that (i) $\hat{G}(s) = \Lambda(s)[\Gamma(s)]^{-1}$ and (ii) there exist polynomial matrices $\Phi(s)$ and $\Omega(s)$ such that

$$(20) \quad \Phi(s)\Lambda(s) + \Omega(s)\Gamma(s) = I.$$

Moreover, we can assume without loss of generality that $\Gamma(s)$ is column proper. Let $\det \Gamma(s) = \prod_{i=1}^k (s - p_i)^{m_i}$, and let δ_i denote the highest power of s appearing in the i th column of $\Gamma(s)$. Since $\Gamma(s)$ is column proper, we have $\sum_{i=1}^k m_i = \sum_{j=1}^n \delta_j$. Let $M(s) = \text{diag} \{ (s+1)^{-\delta_1}, \dots, (s+1)^{-\delta_n} \}$, and define

$$\hat{N}(s) = \Lambda(s)[M(s)]^{-1}, \quad \hat{D}(s) = \Gamma(s)[M(s)]^{-1}.$$

Then \hat{N} and \hat{D} are proper rational matrices with poles only at $s = -1$, and hence $\hat{N}, \hat{D} \in \mathcal{A}^{n \times n}$. Moreover, $\hat{G}(s) = \hat{N}(s)[\hat{D}(s)]^{-1}$. Next, define

$$\hat{U}(s) = \Phi(s) \cdot (s+1)^{-\alpha}, \quad \hat{V}(s) = \Omega(s) \cdot (s+1)^{-\alpha},$$

where α is greater than or equal to the degree of any element of Φ and Ω . Then $\hat{U}, \hat{V} \in \mathcal{A}^{n \times n}$. Also, from (20), we have

$$\hat{U}(s)\hat{N}(s) + \hat{V}(s)\hat{D}(s) = [M(s)]^{-1} \cdot (s+1)^{-\alpha} \triangleq W(s).$$

Clearly, $\hat{W}(s) \in \mathcal{A}^{n \times n}$. Moreover, $\det \hat{W}(s) = (s+1)^{-l}$, where l is an integer (in fact, $l = n\alpha + \sum_{i=1}^n \delta_i$), so $\det \hat{W}(s) \neq 0$ whenever $s \in C_+$. Hence the pair (\hat{N}, \hat{D}) is p.r.c. Finally, we have

$$\det \hat{D}(s) = \det \Gamma(s) \cdot [\det M(s)]^{-1} = \prod_{i=1}^k (s - p_i)^{m_i} \cdot \prod_{i=1}^n (s+1)^{-\delta_i},$$

so $\det \hat{D}(s) \rightarrow 1$ whenever $|s| \rightarrow \infty$. Thus the ordered pair (\hat{N}, \hat{D}) constitutes a p.r.c.f. of \hat{G} according to Definition 4. \square

We now proceed to derive some stability results based on p.r.c.f.'s. These theorems are quite similar to those based on r.c.f.'s, but the details are slightly different. To bring out the similarities more clearly, we use the additional symbol "p". For instance, "Theorem 1p" is the p.r.c.f. analogue of "Theorem 1".

PROPOSITION 1p. *Let $G(\cdot)$ be an L.t.d. with support in $[0, \infty)$, and let (\hat{N}, \hat{D}) be a p.r.c.f. of \hat{G} . Then $\hat{G} \in \mathcal{A}^{n \times n}$ if and only if*

$$(21) \quad \det \hat{D}(s) \neq 0 \quad \text{whenever } s \in C_+.$$

Proof. If. Suppose (31) holds. Since (\hat{N}, \hat{D}) is a p.r.c.f. of \hat{G} , we have that $\liminf |\det \hat{D}(s_i)| > 0$ whenever $(s_i)_{i=1}^\infty$ is a sequence in C_+ such that $|s_i| \rightarrow \infty$;

so this fact, together with (21), implies that

$$\inf_{s \in C_+} |\det \hat{D}(s)| > 0.$$

Hence, by [1, Thm. 4.18.6, p. 150], it follows that $1/\det \hat{D} \in \mathcal{A}^{n \times n}$, whence $\hat{D}^{-1} \in \mathcal{A}^{n \times n}$. Hence $\hat{G} = \hat{N}\hat{D}^{-1}$ also belongs to $\mathcal{A}^{n \times n}$.

Only if. Suppose (21) is violated, and accordingly, suppose $\det \hat{D}(s_0) = 0$ for some $s_0 \in C_+$. Since \hat{N} and \hat{D} are p.r.c., there exist \hat{U} , \hat{V} and \hat{W} in $\mathcal{A}^{n \times n}$ such that $\det \hat{W}(s) \neq 0$ whenever $s \in C_+$ and such that

$$\hat{U}(s)\hat{N}(s) + \hat{V}(s)\hat{D}(s) = \hat{W}(s) \quad \text{for all } s \in C_+.$$

Hence

$$(22) \quad \hat{U}(s)\hat{G}(s) + \hat{V}(s) = \hat{W}(s)[\hat{D}(s)]^{-1}.$$

As $s \rightarrow s_0$, the right-hand side of (22) becomes unbounded, since $\det \hat{W}(s_0) \neq 0$. This implies (as in the proof of Proposition 1) that $\hat{G}(s)$ becomes unbounded as $s \rightarrow s_0$, whence $\hat{G} \notin \mathcal{A}^{n \times n}$. \square

LEMMA 1p. *Let $\hat{G}(\cdot)$ be an L.t.d. with support in $[0, \infty)$; let $F(\cdot) \in \mathcal{A}^{n \times n}$; let (\hat{N}, \hat{D}) be a p.r.c.f. of \hat{G} , and let $\hat{H} = \hat{G}(I + \hat{F}\hat{G})^{-1}$. Suppose we have that $\liminf |\det(I + \hat{F}(s_i)\hat{G}(s_i))| > 0$ whenever $(s_i)_{i=1}^\infty$ is a sequence in C_+ such that $|s_i| \rightarrow \infty$. Then the ordered pair $(\hat{N}, \hat{D} + \hat{F}\hat{N})$ is a p.r.c.f. of \hat{H} .*

Proof. Clearly, $\hat{H} = \hat{N}(\hat{D} + \hat{F}\hat{N})^{-1}$. Also, it can be shown as in the proof of Lemma 1 that \hat{N} and $\hat{D} + \hat{F}\hat{N}$ are p.r.c.; so in order to show that $(\hat{N}, \hat{D} + \hat{F}\hat{N})$ is a p.r.c.f. of \hat{H} , it only remains to show that $\liminf |\det(\hat{D}(s_i) + \hat{F}(s_i)\hat{N}(s_i))| > 0$ whenever $(s_i)_{i=1}^\infty$ is a sequence in C_+ such that $|s_i| \rightarrow \infty$. But this is immediate since $\det(\hat{D} + \hat{F}\hat{N}) = \det \hat{D} \cdot \det(I + \hat{F}\hat{G})$ and since $\liminf |\det \hat{D}(s_i)| > 0$ whenever $(s_i)_{i=1}^\infty$ is such a sequence (by virtue of the fact that (\hat{N}, \hat{D}) is a p.r.c.f. of \hat{G}). \square

Remarks. Comparing Lemma 1 with Lemma 1p, we notice a significant difference. If (\hat{N}, \hat{D}) is an r.c.f. of \hat{G} , then $(\hat{N}, \hat{D} + \hat{F}\hat{N})$ is automatically an r.c.f. of \hat{H} . However, if (\hat{N}, \hat{D}) is only a p.r.c.f. of \hat{G} , then $(\hat{N}, \hat{D} + \hat{F}\hat{N})$ is a p.r.c.f. of \hat{H} only under some additional hypotheses, which essentially insure that all singularities of \hat{H} in C_+ are contained in some bounded subset thereof.

We now present the main stability theorems based on p.r.c.f.'s.

THEOREM 1p. *Let $G(\cdot)$ be an L.t.d. with support in $[0, \infty)$, and let $F(\cdot) \in \mathcal{A}^{n \times n}$. Suppose all singularities of \hat{G} in C_+ are contained within a bounded subset of C_+ . Then $\hat{H} = \hat{G}(I + \hat{F}\hat{G})^{-1} \in \mathcal{A}^{n \times n}$ if and only if there exists a p.r.c.f. (\hat{N}, \hat{D}) of \hat{G} such that*

$$\inf_{s \in C_+} |\det(\hat{D}(s) + \hat{F}(s)\hat{N}(s))| > 0.$$

Proof. The “if” part is obvious. To prove the “only if” part, suppose $\hat{H} \in \mathcal{A}^{n \times n}$ and define

$$\hat{N} = \hat{H}, \quad \hat{D} = I - \hat{F}\hat{H}.$$

Then (\hat{N}, \hat{D}) is a right-coprime pair and is therefore also p.r.c. It only remains to show that \hat{D} satisfies condition (iii) of Definition 4. Towards this end, suppose by way of contradiction that $(s_i)_{i=1}^\infty$ is a sequence in C_+ such that $|s_i| \rightarrow \infty$ and

$\det \hat{D}(s_i) \rightarrow 0$. Now we have

$$\hat{F}(s)\hat{N}(s) + \hat{D}(s) = I,$$

so that

$$(23) \quad \hat{F}(s) \cdot \hat{G}(s) + I = [\hat{D}(s)]^{-1}.$$

So, if we replace s by s_i in (23) and let $i \rightarrow \infty$, the right side becomes unbounded, whence so must the left side. This in turn implies that $\hat{G}(s_i)$ becomes unbounded, which contradicts the hypothesis that all singularities of \hat{G} in C_+ are contained in a bounded subset of C_+ . Hence no such sequence (s_i) can exist, and as a result, the ordered pair (\hat{N}, \hat{D}) constitutes a p.r.c.f. of \hat{G} . \square

COROLLARY 1p. *Let $G(\cdot)$ be an L.t.d. with support in $[0, \infty)$; let $F(\cdot) \in \mathcal{A}^{n \times n}$; let (\hat{N}, \hat{D}) be a p.r.c.f. of \hat{G} , and let $\hat{H} = \hat{G}(I + \hat{F}\hat{G})^{-1}$. Then $\hat{H} \in \mathcal{A}^{n \times n}$ if and only if (23) holds.*

From an applications point of view, Theorem 1p is not particularly illuminating. Theorem 2p below is much better in this respect. The proofs are omitted in the interests of brevity.

THEOREM 2p. *Let G, F and \hat{H} be as in Theorem 1p. Then $\hat{H} \in \mathcal{A}^{n \times n}$ if and only if*

$$(24) \quad \inf_{s \in C_+} |\det(I + F(s)G(s))| > 0,$$

(ii) *there exists a p.r.c.f. (\hat{N}, \hat{D}) of \hat{G} such that*

$$(25) \quad \det(\hat{D}(s) + \hat{F}(s)\hat{N}(s)) \neq 0 \quad \text{whenever } \det \hat{D}(s) = 0 \text{ and } s \in C_+.$$

COROLLARY 2p. *Let $G, F, \hat{H}, \hat{N}, \hat{D}$ be as in Corollary 1p. Then $H \in \mathcal{A}^{n \times n}$ if and only if (24) and (25) hold.*

Finally, as is the case with r.c.f.'s, it is only necessary to find a p.r.c.f. of the "unstable part" of a given transfer function matrix.

PROPOSITION 2p. *Let $\hat{G} = \hat{G}_b + \hat{G}_u$, where $\hat{G}_b \in \mathcal{A}^{n \times n}$ and the ordered pair (\hat{N}, \hat{D}) constitutes a p.r.c.f. of \hat{G}_u . Then $(\hat{N} + \hat{G}_b\hat{D}, \hat{D})$ is a p.r.c.f. of \hat{G} .*

5. Special case. As an application of Theorem 2p, we derive explicit necessary and sufficient conditions for a class of multivariable feedback systems to be stable.

THEOREM 3. *Let $G(\cdot)$ be an L.t.d. with support in $[0, \infty)$, and suppose $\hat{G}(\cdot)$ is of the form*

$$G(s) = \sum_{i=1}^k \sum_{j=1}^{m_i} R_{ij}/(s - p_i)^j + \hat{G}_b(s) \triangleq \hat{G}_u(s) + \hat{G}_b(s),$$

where $\operatorname{Re} p_i \geq 0$ for all i and $G_b(\cdot) \in \mathcal{A}^{n \times n}$. Let

$$\hat{G}_u(s) = \Lambda(s)[\Gamma(s)]^{-1}$$

be a right-coprime polynomial factorization of $\hat{G}_u(\cdot)$, and suppose $\Gamma(\cdot)$ is column

proper. Let $\hat{F} \in \mathcal{A}^{n \times n}$ and let $\hat{H} = \hat{G}(I + \hat{F}\hat{G})^{-1}$. Then $\hat{H}(\cdot) \in \mathcal{A}^{n \times n}$ if and only if

$$(26) \quad \inf_{s \in C_+} |\det(I + \hat{F}(s)\hat{G}(s))| > 0,$$

(ii)

$$(27) \quad \det[\Gamma(p_i) + \hat{F}(p_i)\Lambda(p_i) + \hat{F}(p_i)\hat{G}_b(p_i)\Gamma(p_i)] \neq 0$$

for $i = 1, \dots, k$.

Proof. Given Λ and Γ , define \hat{N} and \hat{D} as in Proposition 3. Then, the ordered pair (\hat{N}, \hat{D}) is a p.r.c.f. of \hat{G}_u , and therefore, by Proposition 2p, $(\hat{N} + \hat{G}_b\hat{D}, \hat{D})$ is a p.r.c.f. of \hat{G} . Hence, by Theorem 2p, $H \in \mathcal{A}^{n \times n}$ if and only if (26) holds, and in addition

$$(28) \quad \det[\hat{D}(s) + \hat{F}(s)\hat{N}(s) + \hat{F}(s)\hat{G}_b(s)\hat{D}(s)] \neq 0$$

whenever $\det \hat{D}(s) = 0, \quad s \in C_+.$

However, it is clear from the way that \hat{N} and \hat{D} are constructed that the only points s in C_+ such that $\det \hat{D}(s) = 0$ are $s = p_i, i = 1, \dots, k$. Furthermore, $\hat{N}(s) = \Lambda(s)M^{-1}(s)$, $\hat{D}(s) = \Gamma(s)M^{-1}(s)$, and $\det M(p_i) \neq 0$ for $i = 1, \dots, k$. Therefore, (28) simplifies to (27).

Remarks. Theorem 3 is a generalization of some results of Callier and Desoer [9], who assume that $\hat{F}(\cdot)$ is a constant nonsingular matrix.

When \hat{G} has several poles in C_+ , finding the polynomial right-coprime factorization of \hat{G}_u may be quite cumbersome. Since the condition (26) is purely local, it is possible to deal with each pole individually. Let

$$\hat{R}_i(s) = \sum_{j=1}^{m_i} R_{ij}/(s - p_i)^j,$$

$$\hat{L}_i(s) = \hat{G}(s) - \hat{R}_i(s).$$

Then \hat{L}_i is analytic in a sufficiently small disk centered at p_i . Also, we have

$$\hat{G}(s) = [\hat{R}_i(s) + \hat{L}_i(p_i)] + [\hat{L}_i(s) - \hat{L}_i(p_i)].$$

The first bracket is a proper rational function of s ; the second bracket is analytic in a sufficiently small disk centered at p_i and vanishes at $s = p_i$. Now, suppose

$$\hat{R}_i(s) = \Lambda_i(s)[\Gamma_i(s)]^{-1}$$

is a polynomial r.c.f. of $\hat{R}_i(s)$. Then, since $\hat{L}_i(p_i)$ is a constant matrix, one can show by methods entirely analogous to those of Propositions 2 and 2p that the ordered pair $(\Lambda_i(\cdot) + \hat{L}_i(p_i)\Gamma_i(\cdot), \Gamma_i(\cdot))$ is a polynomial r.c.f. of $[\hat{R}_i(s) + \hat{L}_i(p_i)]$. Using this decomposition, (26) can be simplified to

$$\det(\Gamma(p_i) + \hat{F}(p_i)\Lambda_i(p_i) + \hat{F}(p_i)\hat{L}_i(p_i)\Gamma_i(p_i)) \neq 0, \quad i = 1, \dots, k.$$

6. Conclusions. In the interests of brevity, we do not state the stability theorems for discrete-time systems. We merely note that for discrete time systems, there is no loss of generality in dealing with p.r.c.f.'s rather than with r.c.f.'s,

owing to the power-series nature of the z -transform. The complete details can be found in [11].

In this paper, we have presented several stability theorems based on right-coprime factorizations and pseudo-right-coprime factorizations that can be used to test the stability of *multivariable* feedback systems. It goes without saying that completely analogous results can be derived using left-coprime factorizations and the formula $\hat{H} = (I + \hat{G}\hat{F})^{-1}\hat{G}$.

Acknowledgment. The author gratefully acknowledges many fruitful discussions with Professor C. A. Desoer, who also made many valuable suggestions regarding the structure of this paper.

REFERENCES

- [1] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semigroups*, American Mathematical Society, Providence, R.I., 1953.
- [2] C. H. HSU AND C. T. CHEN, *A proof of the stability of multivariable feedback systems*, Proceedings IEEE, vol. 56, IEEE, New York, 1968, pp. 2061–2062.
- [3] H. H. ROSENBRCK, *State-space and Multivariable Theory*, Nelson, London, 1970.
- [4] C. A. DESOER AND M. VIDYASAGAR, *General necessary conditions for input-output stability*, Proceedings IEEE, vol. 59, IEEE, New York, 1971, pp. 1255–1256.
- [5] S. H. WANG, *Design of linear multivariable systems*, Memo. ERL-M309, College of Engrg., Univ. of Calif, Berkeley, 1971.
- [6] M. VIDYASAGAR, *Input-output stability of a broad class of linear time-invariant multivariable systems*, this Journal, 10 (1972), pp. 203–209.
- [7] C. A. DESOER AND F. M. CALLIER, *Convolution feedback systems*, this Journal, 10 (1972), pp. 737–746.
- [8] F. M. CALLIER AND C. A. DESOER, *Necessary and sufficient conditions for stability of n -input- n -output convolution feedback systems with a finite number of unstable poles*, IEEE Trans. Automatic Control, AC-18 (1973), pp. 295–298.
- [9] C. A. DESOER AND J. D. SCHULMAN, *Cancellations in multivariable continuous-time and discrete-time feedback systems treated by greatest common divisor extraction*, Ibid., AC-18 (1973), pp. 401–402.
- [10] ———, *Zeros and poles of matrix transfer functions and their dynamical interpretation*, IEEE Trans. Circuits and Systems, CAS-21 (1974), pp. 3–8.
- [11] C. A. DESOER AND M. VIDYASAGAR, *Feedback Systems: Input-output Properties*, Academic Press, New York, 1975.

STABILITY CRITERIA FOR TIME-VARYING SYSTEMS IN HILBERT SPACE*

ANDREW ACKER†

Abstract. The Freedman and Zames logarithmic variation criterion for stability [1] is extended to two types of systems in Hilbert space: one involving a generalized causal convolution operator and a time-varying gain, the other involving a real causal convolution operator and a time-varying unbounded operator. Also, the conditions on the Nyquist diagram are relaxed, and it is shown that a certain bound on either the average logarithmic increase or the average logarithmic decrease of the gain functions insures stability.

1. Introduction. This paper contains a variety of independent improvements of the Freedman and Zames logarithmic variation criterion [1]. On the one hand, two separate generalizations in the Hilbert space context are presented. On the other hand, improved stability criteria are found whose significance is independent of the Hilbert space generalizations.

The main application (within the context of real feedback systems) of these latter stability criteria occurs in the case (not considered in [1]) in which the positive gain function is not uniformly logarithmically bounded over all time, i.e., approaches 0 or ∞ or oscillates with both as limit points as $t \rightarrow \infty$. In this context, we find that if the Nyquist diagram of the time-invariant part \mathbf{G} of the system avoids $(-\infty, 0]$, then an appropriate bound on either the average logarithmic increase or the average logarithmic decrease of the gain suffices to insure stability. Furthermore, if the Nyquist diagram intersects the origin, but avoids an arbitrarily thin angular sector about the negative real axis, then stability is implied by a bound on the average logarithmic decrease of the gain provided that

$$\limsup_{|\omega| \rightarrow \infty} |\arg \hat{\mathbf{G}}(i\omega)| \leq \pi/2,$$

i.e., provided the Nyquist diagram approaches the origin as $|\omega| \rightarrow \infty$ at an angle which in the limit does not exceed 90° . Although, for unity of notation, the following sections have been written mostly in the Hilbert space context, the reader who is primarily interested in these results can understand them quite easily by reading Lemma 4.4 and §§ 5 and 6. In § 8, the assumption that the gain is not uniformly logarithmically bounded yields naturally to the assumption that a time-varying operator, which replaces the gain, is unbounded.

The main lemmas in the new stability theorems are generalizations of two operator factorization lemmas in [1]. Namely, in § 4, the generalized transform theory of Freedman and Falb [2] (which is reviewed in § 3) is utilized in the proof of a generalized multiplier lemma for causal convolution operators in Hilbert space. Of particular interest is the fact that the new proof does not rely on the "operators with prescribed phase characteristics" technique used in the original proof in [1], but rather relies on a lemma relating the Nyquist diagram of an operator to its spectrum in the space of causal, time-invariant operators. In § 5, the

* Received by the editors December 27, 1973, and in revised form July 10, 1974.

† Mathematisches Institut I, Universität Karlsruhe, Karlsruhe, Germany.

logarithmic variation condition for the factorization of the gain function ([1, Lemma 4]) is weakened to necessary and sufficient conditions which involve bounds on the average logarithmic increase or decrease of the gain.

In § 7, a new positivity lemma is proven for integrals involving a time-varying self-adjoint operator and a real causal convolution operator.

The stability results in §§ 6 and 8 are generalizations to the cases where \mathbf{G} becomes a generalized causal convolution operator in H and where the gain is replaced by a time-varying, unbounded self-adjoint operator. The results in the second case are particularly well adapted to the determination of stability of solutions of boundary and initial value problems in partial differential equations. For this application the new conditions concerning behavior of the Nyquist diagram in a neighborhood of the origin are essential.

2. The feedback system. H is a separable Hilbert space which, unless specifically specified, can be either real or complex. If H is real, then $H^c = H \oplus iH$ is the complexification of H . $\mathcal{L}(H, H)$ denotes the Banach algebra of bounded linear operators in H , and $\|\cdot\|$ is the norm on $\mathcal{L}(H, H)$. For any function $f(t)$ defined on $[0, \infty)$ and $T \geq 0$, $f_T(t)$ is the truncation of $f(t)$ to the interval $[0, T]$.

DEFINITION 2.1. $L_{2e}([0, \infty), H)$ is the space of functions f defined on $[0, \infty)$ such that f_T is in $L_2([0, \infty), H)$ for all $T \geq 0$.

DEFINITION 2.2. For any real number σ , $L_{2\sigma}([0, \infty), H)$ is the space of functions $x(t)$ defined on $[0, \infty)$ for which $x(t) \exp(\sigma t)$ is in $L_2([0, \infty), H)$. $L_{2\sigma}([0, \infty), H)$ is a Banach space in the norm

$$\|x\|_{\sigma} = \left| \int_0^{\infty} |x(t) \exp(\sigma t)|^2 dt \right|^{1/2}.$$

DEFINITION 2.3. An operator $\mathbf{G}: L_{2e}([0, \infty), H) \rightarrow L_{2e}([0, \infty), H)$ is *causal* if, for all functions $x(t)$ in $L_{2e}([0, \infty), H)$ and all $T \geq 0$, we have

$$(\mathbf{G}x)_T = (\mathbf{G}x_T)_T.$$

DEFINITION 2.4. $W(H, \sigma)$, for any real number σ , is the space of *causal linear operators*

$$\mathbf{G}: L_{2\sigma}([0, \infty), H) \rightarrow L_{2\sigma}([0, \infty), H).$$

$W(H, \sigma)$ is a Banach algebra in the norm

$$\|\mathbf{G}\|_{W, \sigma} = \sup \{ \|\mathbf{G}x\|_{\sigma} \mid \|x\|_{\sigma} = 1 \}.$$

DEFINITION 2.5. For any real number σ (i.e., σ can be positive or negative), $B(H, \sigma)$ is the space of causal operators $\mathbf{G}: L_{2e}([0, \infty), H) \rightarrow L_{2e}([0, \infty), H)$ which are of the form $\mathbf{G} = g_0 \Delta + \mathbf{g}$ where Δ is the identity in $L_{2e}([0, \infty), H)$, \mathbf{g} is a convolution operator, i.e.,

$$(\mathbf{g}e)(t) = \int_0^t g(t - \tau)e(\tau) d\tau,$$

where $g(t) \exp(\sigma t)$ is in $L_1([0, \infty), \mathcal{L}(H, H))$, and g_0 is a real or complex number depending on whether H is real or complex. $B(H, \sigma)$ is a (noncommutative) Banach algebra under the norm $\|\mathbf{G}\|_{B, \sigma} = |g_0| + \int_0^{\infty} |g(t)| \exp(\sigma t) dt$. Moreover, $B(H, \sigma) \subset W(H, \sigma')$ whenever $\sigma' \leq \sigma$.

DEFINITION 2.6. $\overline{B(H, \sigma)}$ is the *closure* of $B(H, \sigma)$ in $W(H, \sigma)$.

Remark 2.7. The operators in $\overline{B(H, \sigma)}$ are causal. They are also time invariant in the sense that if $f \in L_{2\sigma}([0, \infty), H)$ and if $f(t)$ and its translation $f_a(t) = f(a + t)$ both have support in $[0, \infty)$, then $\mathbf{G}f_a = (\mathbf{G}f)_a$.

The stability results in § 6 will concern the general feedback system represented for $t \geq 0$ by the equations:

$$(1) \quad e(t) = f(t) - \phi(t)y(t),$$

$$(2) \quad y(t) = (\mathbf{G}e)(t),$$

where the following assumptions hold.

ASSUMPTION 1. H is a real, separable Hilbert space and \mathbf{G} is an operator in $\overline{B(H, \sigma)}$ for some positive number σ .

ASSUMPTION 2. $\phi(t)$ is a real function defined on $[0, \infty)$. (Properties such as continuity are explicitly not assumed.)

ASSUMPTION 3. For every input $f(t)$ in $L_{2\alpha}([0, \infty), H)$ (α arbitrary), the solutions $e(t)$ and $y(t)$ exist in $L_{2\alpha}([0, \infty), H)$.

DEFINITION 2.8. For any real α , the feedback system of (1) and (2) is $L_{2\alpha}$ -stable with respect to y if for every input $f(t)$ in $L_{2\alpha}([0, \infty), H)$ the corresponding output $y(t)$ is also in $L_{2\alpha}([0, \infty), H)$, and if there exists a constant A such that $\|y\|_\alpha \leq A\|f\|_\alpha$ uniformly for all input-output pairs $(f(t), y(t))$ in $L_{2\alpha}([0, \infty), H)$. $L_{2\alpha}$ -stability with respect to the output $e(t)$ is defined analogously. If $\alpha = 0$, the system is simply called L_2 -stable with respect to y or e .

3. Generalized transform theory. The transform theory developed by Freedman and Falb in [2] is briefly reviewed and adapted to the present context. C represents the complex numbers. For any real number σ , $C(\sigma) = \{z \in C \mid \operatorname{Re} z > \sigma\}$ and $\overline{C}(\sigma)$ is the closure of $C(\sigma)$.

DEFINITION 3.1. If $\mathbf{G} = g_0\Delta + \mathbf{g} \in B(H, \sigma)$, then the Laplace transform $\hat{\mathbf{G}}(z)$ is defined on $\overline{C(-\sigma)}$ by $\hat{\mathbf{G}}(z) = g_0I + \hat{\mathbf{g}}(z)$ where $\hat{\mathbf{g}}(z) = \int_0^\infty g(t) \exp(-zt) dt$ and the identity I are operators in H^c . If $\mathbf{G} \in \overline{B(H, \sigma)}$, then there is a sequence $\{\mathbf{G}_n\}$ of operators in $B(H, \sigma)$ such that $\|\mathbf{G}_n - \mathbf{G}\|_{W, \sigma} \rightarrow 0$. The transforms $\hat{\mathbf{G}}_n(z)$ are uniformly Cauchy on $\overline{C(-\sigma)}$ and converge uniformly on $\overline{C(-\sigma)}$ to a function $\hat{\mathbf{G}}(z)$ which is then defined as the Laplace transform of \mathbf{G} .

Remark 3.2. For any $\mathbf{G} \in \overline{B(H, \sigma)}$, $\hat{\mathbf{G}}(z)$ is analytic on $C(-\sigma)$ and uniformly continuous on $\overline{C(-\sigma)}$. Also, there is a point $g_0 \in C$ (g_0 real if H is real) such that $\hat{\mathbf{G}}(z) \rightarrow g_0I$ uniformly on $\overline{C(-\sigma)}$ as $|z| \rightarrow \infty$. We can write $\mathbf{G} = g_0\Delta + \mathbf{g}$ where $\mathbf{g} = \mathbf{G} - g_0\Delta$ is the limit of a sequence of convolution operators without adjoined identity.

DEFINITION 3.3. (See [2, Def. 5.5].) Assume $\mathbf{G} = g_0\Delta + \mathbf{g} \in \overline{B(H, \sigma)}$ and let $\{e_1, e_2, \dots\}$ be an orthonormal basis for H . Let H_n and H_n^c be the spans of $\{e_1, \dots, e_n\}$ in H and H^c and let E_n be the projection of H, H^c into H_n, H_n^c . Let $\hat{\mathbf{g}}_n(z) = E_n \hat{\mathbf{g}}(z) E_n$ for each $z \in \overline{C(-\sigma)}$. Then \mathbf{G} is *approximable* in $\overline{B(H, \sigma)}$ if $\hat{\mathbf{g}}_n(z)$ converges uniformly to $\hat{\mathbf{g}}(z)$ on $\overline{C(-\sigma)}$.

Remark 3.4. It is shown in [2, Prop. 5.6] that $\mathbf{G} = g_0\Delta + \mathbf{g}$ is approximable in $\overline{B(H, \sigma)}$ if and only if $\hat{\mathbf{g}}(z)$ is a compact operator in H^c for all $z \in \overline{C(-\sigma)}$.

DEFINITION 3.5. For $\mathbf{G} \in \overline{B(H, \sigma)}$, the σ -shifted Nyquist diagram of \mathbf{G} is $N(\mathbf{G}, \sigma) = \bigcup_{\omega \in [-\infty, \infty]} \operatorname{spec} \hat{\mathbf{G}}(i\omega - \sigma)$. (This is written $N_H(\mathbf{G}, \sigma)$ if H is ambiguous.) $N(\mathbf{G}, \sigma)$ is a compact subset of C which, if H is real, is symmetric to the real axis.

DEFINITION 3.6. For $\mathbf{G} = g_0\mathbf{A} + \mathbf{g} \in \overline{B(H, \sigma)}$, we define $\sum(\mathbf{G}, \sigma) = (\cup_{z \in \overline{C}(-\sigma)} \text{spec } \hat{\mathbf{G}}(z)) \cup \{g_0\}$. (This is written $\sum_H(\mathbf{G}, \sigma)$ when H is ambiguous.) $\sum(\mathbf{G}, \sigma)$ is compact and is also symmetric to the real axis when H is real.

Remark 3.7. Freedman and Falb proved in [2, Cor. 5.8] that

$$\text{spec}_{\overline{B(H, 0)}} \mathbf{G} = \sum_H(\mathbf{G}, 0)$$

when \mathbf{G} is approximable and H is complex. If H is real and \mathbf{G} is approximable in $\overline{B(H, \sigma)}$, then this result becomes

$$\text{spec}_{\overline{B(H^c, \sigma)}} \mathbf{G} = \sum_{H^c}(\mathbf{G}, \sigma) = \sum_H(\mathbf{G}, \sigma).$$

4. Multiplier lemmas for operators in $\overline{B(H, \sigma)}$. Lemma 4.3 extends the multiplier lemma of Freedman and Zames ([1, Lemma 3]) to Hilbert space. Also, the condition on the Nyquist diagram is reduced from $(-\infty, 0] \cap N(\mathbf{G}, \sigma) = \emptyset$ to $R_\varepsilon \cap N(\mathbf{G}, \sigma) = \emptyset$ for ε sufficiently small, where $R_\varepsilon = (-\infty, 0) \cup \{z | |z + \varepsilon| < \varepsilon\}$. Thus, Lemma 4.3 applies to operators whose Nyquist diagram contains the origin, such as convolution operators without an adjoined identity.

Lemma 3 of [1] was proved by application of the method of "Construction of multipliers with prescribed phase characteristics" ([1, Lemma 2]). This method appears not to generalize readily beyond the case where g_0 and $g(t)$ are real-valued. It is shown here that the required multiplier can be obtained by a simpler method which is unaffected by dimension. Namely, if $R_\varepsilon \cap N(\mathbf{G}, \sigma) = \emptyset$, then the multiplier is $(\varepsilon\mathbf{A} + \mathbf{G})^{-1/2}$. Here, the main problem becomes that of showing that the condition on $N(\mathbf{G}, \sigma)$ confines the spectrum $\sum(\mathbf{G}, \sigma)$ of \mathbf{G} in $\overline{B(H, \sigma)}$ sufficiently so that $(\varepsilon\mathbf{A} + \mathbf{G})^{-1/2}$ exists as a causal operator. This problem is solved by Proposition 4.2.

In Lemma 4.4 we apply Lemma 2 of [1] to derive alternative sufficient conditions which restrict $N(\mathbf{G}, \sigma)$ somewhat less in the vicinity of the origin, but which only apply when g_0 and $g(t)$ are real-valued.

PROPOSITION 4.1. Assume $\mathbf{G} \in \overline{B(H, 0)}$, where H is n -dimensional. Then for $z_0 \in C$, $z_0 \notin \sum(\mathbf{G}, 0)$ if and only if (a) $z_0 \notin N(\mathbf{G}, 0)$ and (b) the sum of the winding numbers about z_0 of the n (continuously chosen) eigenvalue functions of $\hat{\mathbf{G}}(i\omega)$ on $[-\infty, \infty]$ is 0.

Proof. If $\{e_1, \dots, e_n\}$ is an orthogonal basis for H , then for $z \in \overline{C}(0)$ the matrix $G_0(z)$ defined by

$$[G_0(z)]_{i,j} = (g_0 - z_0)^{-1} \langle e_i, (\hat{\mathbf{G}}(z) - z_0 I) e_j \rangle, \quad i, j = 1, \dots, n,$$

is invertible if and only if $\hat{\mathbf{G}}(z) - z_0 I$ is invertible. The winding numbers about 0 of $\det G_0(i\omega)$ on $[-\infty, \infty]$ is the sum of the winding numbers about 0 of the n continuously chosen eigenvalue functions $\lambda_k(\omega)$, $k = 1, \dots, n$, of $G_0(i\omega)$ on $[-\infty, \infty]$. Since $\det G_0(z)$ is analytic on $C(0)$ and $\det G_0(z) \rightarrow 1$ uniformly as $|z| \rightarrow \infty$ in $\overline{C}(0)$, the proposition follows from the argument principle.

PROPOSITION 4.2. Let $\mathbf{G} = g_0\mathbf{A} + \mathbf{g}$ be an approximable operator in $\overline{B(H, 0)}$. For points $z_0, z_1 \in C$, assume that $z_0 \notin \sum(\mathbf{G}, 0)$ and that there exists a polygonal arc Γ with z_0 and z_1 as endpoints such that $\Gamma \cap N(\mathbf{G}, 0) = \emptyset$. Then $z_1 \notin \sum(\mathbf{G}, 0)$.

Proof. If $\Gamma' = \{z - g_0 | z \in \Gamma\}$, then $\Gamma' \cap N(\mathbf{g}, 0) = \emptyset$. Thus $(\hat{\mathbf{g}}(i\omega) - \lambda I)^{-1}$ exists for $(\omega, \lambda) \in [-\infty, \infty] \times \Gamma'$, and in fact $|(\hat{\mathbf{g}}(i\omega) - \lambda I)^{-1}|$ is uniformly bounded

over this set. Thus if $\hat{\mathbf{g}}_n(z) = E_n \hat{\mathbf{g}}(z) E_n$, then the equation

$$(\hat{\mathbf{g}}_n(i\omega) - \lambda I) = [I + (\hat{\mathbf{g}}_n(i\omega) - \hat{\mathbf{g}}(i\omega))(\hat{\mathbf{g}}(i\omega) - \lambda I)^{-1}](\hat{\mathbf{g}}(i\omega) - \lambda I)$$

shows that there is an integer $n_0 > 0$ such that $(\hat{\mathbf{g}}_n(i\omega) - \lambda I)^{-1}$ exists and is uniformly bounded over all (ω, λ, n) with $(\omega, \lambda) \in [-\infty, \infty] \times \Gamma'$ and $n \geq n_0$. For $n \geq n_0$ and $H_n = E_n H$, the condition $\Gamma' \cap N_H(\mathbf{g}_n, 0) = \emptyset$ implies $\Gamma' \cap \sum_{H_n}(\mathbf{g}_n, 0) = \emptyset$ by Proposition 4.1. Therefore, $\Gamma' \cap \sum_H(\mathbf{g}_n, 0) = \emptyset$, since $\sum_H(\mathbf{g}_n, 0) = \sum_{H_n}(\mathbf{g}_n, 0)$. Thus, for $n \geq n_0$ and $\lambda \in \Gamma'$, $(\hat{\mathbf{g}}_n(z) - \lambda I)^{-1}$ is analytic on $C(0)$, uniformly continuous on $\bar{C}(0)$, and uniformly convergent to $(-1/\lambda)$ as $|z| \rightarrow \infty$ in $\bar{C}(0)$. The norm $|(\hat{\mathbf{g}}_n(z) - \lambda I)^{-1}|$ must take on its maximum value somewhere along the line $z = i\omega$, $\omega \in [-\infty, \infty]$, because the operator norm function is subharmonic on $C(0)$. Thus $\sup_{z \in C(0)} |(\hat{\mathbf{g}}_n(z) - \lambda I)^{-1}|$ is uniformly bounded over all $n \geq n_0$. From the equation

$$(\hat{\mathbf{g}}(z) - \lambda I) = [I + (\hat{\mathbf{g}}(z) - \hat{\mathbf{g}}_n(z))(\hat{\mathbf{g}}_n(z) - \lambda I)^{-1}](\hat{\mathbf{g}}_n(z) - \lambda I),$$

we conclude that $(\hat{\mathbf{g}}(z) - \lambda I)^{-1}$ exists for all $z \in \bar{C}(0)$. Therefore, $\Gamma \cap \sum_H(\mathbf{G}, 0) = \Gamma' \cap \sum_H(\mathbf{g}, 0) = \emptyset$, which implies the result.

For any $\varepsilon > 0$, we let $R_\varepsilon = (-\infty, 0) \cup \{z | |z + \varepsilon| < \varepsilon\}$.

LEMMA 4.3 (the generalized multiplier lemma). *Let H be real and let \mathbf{G} be an approximable operator in $\overline{B(H, \sigma)}$ such that $\hat{\mathbf{G}}(z)$ is a normal operator in $\mathcal{L}(H^c, H^c)$ for each $z \in \bar{C}(-\sigma)$. Then if $R_\varepsilon \cap N(\mathbf{G}, \sigma) = \emptyset$ for some $\varepsilon > 0$, then the operator $\mathbf{M}_\varepsilon = (\varepsilon \Delta + \mathbf{G})^{-1/2}$ exists in $\overline{B(H, \sigma)}$ and satisfies (for a fixed positive number δ) the inequalities*

$$(a1) \quad \int_0^T \exp(rt) \langle y(t), (\mathbf{M}_\varepsilon y)(t) \rangle dt \geq \delta \int_0^T |y(t)|^2 dt$$

and

$$(a2) \quad \int_0^T \exp(rt) \langle y(t), (\mathbf{M}_\varepsilon \mathbf{G} y)(t) \rangle dt \geq 0,$$

whenever $T \in [0, \infty)$, $r \in [0, 2\sigma]$, and $y \in L_{2e}([0, \infty), H)$. If, furthermore, $(-\infty, 0] \cap N(\mathbf{G}, \sigma) = \emptyset$, then the operator $\mathbf{M}_0 = \mathbf{G}^{-1/2}$ exists in $\overline{B(H, \sigma)}$ and satisfies (for a fixed positive number δ) the inequalities

$$(b1) \quad \int_0^T \exp(rt) \langle y(t), (\mathbf{M}_0 y)(t) \rangle dt \geq \delta \int_0^T |y(t)|^2 dt$$

and

$$(b2) \quad \int_0^T \exp(rt) \langle y(t), (\mathbf{M}_0 \mathbf{G} y)(t) \rangle dt \geq \delta \int_0^T |y(t)|^2 dt,$$

whenever $T \in [0, \infty)$, $r \in [0, 2\sigma]$ and $y \in L_{2e}([0, \infty), H)$.

Proof. For any real α and β , the transformation $T(\alpha): \overline{B(H, \beta)} \rightarrow \overline{B(H, \alpha + \beta)}$ is defined as follows. For $\mathbf{G} = g_0 \Delta + \mathbf{g}$ in $\overline{B(H, \beta)}$, $T(\alpha)(g_0 \Delta + \mathbf{g}) = g_0 \Delta + \mathbf{g}_\alpha$ where $g_\alpha(t) = g(t) \exp(-\alpha t)$. $T(\alpha)$ is extended to $\overline{B(H, \beta)}$ by taking the closure.

Now, given $\mathbf{G} \in \overline{B(H, \sigma)}$, define $\mathbf{G}_0 = T(-\sigma)\mathbf{G}$. Then $\hat{\mathbf{G}}_0(z) = \hat{\mathbf{G}}(z - \sigma)$ in $\bar{C}(0)$, so that $R_\varepsilon \cap N(\mathbf{G}_0, 0) = \emptyset$ or $(-\infty, 0] \cap N(\mathbf{G}_0, 0) = \emptyset$ in the respective cases.

Thus, by Proposition 4.2, either $R_\varepsilon \cap \sum(\mathbf{G}_0, 0) = \emptyset$ or $(-\infty, 0] \cap \sum(\mathbf{G}_0, 0) = \emptyset$. For all $\alpha \geq 0$, let $\mathbf{G}_\alpha = T(\alpha)\mathbf{G}_0$. Then $\mathbf{G}_\alpha \in B(H, 0)$ and $\sum(\mathbf{G}_\alpha, 0) \subset \sum(\mathbf{G}_0, 0)$.

For $\varepsilon' \in [0, \varepsilon]$, define on the complement of $(-\infty, -\varepsilon']$ the analytic functions $f_{\varepsilon'}(z) = (\varepsilon' + z)^{-1/2}$ and $h_{\varepsilon'}(z) = zf_{\varepsilon'}(z)$, where $z^{-1/2}$ is the principal square root of z^{-1} . Then $\operatorname{Re} f_{\varepsilon'}(z) > 0$ and $\operatorname{Re} h_{\varepsilon'}(z) \geq 0$ in the complement of R_ε , and $\operatorname{Re} f_0(z) > 0$ and $\operatorname{Re} h_0(z) > 0$ in the complement of $(-\infty, 0]$. Thus if $R_\varepsilon \cap N(\mathbf{G}, \sigma) = \emptyset$, then there is a $\delta > 0$ such that $\operatorname{Re} f_{\varepsilon'}(z) \geq 2\delta$ and $\operatorname{Re} h_{\varepsilon'}(z) \geq 0$ on $\sum(\mathbf{G}_0, 0)$. Similarly, if $(-\infty, 0] \cap N(\mathbf{G}, \sigma) = \emptyset$, then there is a $\delta > 0$ such that $\operatorname{Re} f_0(z) \geq 2\delta$ and $\operatorname{Re} h_0(z) \geq 2\delta$ on $\sum(\mathbf{G}_0, 0)$.

For $\alpha \geq 0$ and $\varepsilon' \in [0, \varepsilon]$ define the operator $\mathbf{M}_{\varepsilon', \alpha}$ in $\overline{B(H, \alpha)}$ by $\mathbf{M}_{\varepsilon', \alpha} = f_{\varepsilon'}(\mathbf{G}_\alpha)$ (where $f_{\varepsilon'}(\mathbf{G}_\alpha)$ is defined by means of a Cauchy integral as in [3, (22)]). The integral exists as a consequence of Remark 3.7). Then under the respective conditions on $N(\mathbf{G}, \sigma)$ we obtain $\mathbf{M}_{\varepsilon', \alpha} \geq \delta\Delta$ or $\mathbf{M}_{0, \alpha} \geq \delta\Delta$ in $\overline{B(H, 0)}$ by [3, Lemma 4.5]. Let the multipliers \mathbf{M}_ε and \mathbf{M}_0 asserted in the lemma be given by $\mathbf{M}_\varepsilon = \mathbf{M}_{\varepsilon, \sigma}$ and $\mathbf{M}_0 = \mathbf{M}_{0, \sigma}$. Then for $r \in [0, 2\sigma]$ and $y_r(t) = y(t) \exp((rt)/2)$ we have

$$\begin{aligned} \int_0^T \exp(rt) \langle y(t), (\mathbf{M}_\varepsilon y)(t) \rangle dt &= \int_0^T \langle y_r(t), (\mathbf{M}_{\varepsilon, \sigma - (r/2)} y_r)(t) \rangle dt \\ &\geq \delta \int_0^T |y_r(t)|^2 dt \geq \delta \int_0^T |y(t)|^2 dt, \end{aligned}$$

which proves (a1).

The proof of (b1) is similar.

Also, $\mathbf{M}_{\varepsilon, \alpha} \mathbf{G}_\alpha = h_\varepsilon(\mathbf{G}_\alpha)$ and $\mathbf{M}_{0, \alpha} \mathbf{G}_\alpha = h_0(\mathbf{G}_\alpha)$, so that the inequalities (a2) and (b2) follow by the same arguments applied to the functions h_ε and h_0 .

LEMMA 4.4. *Let \mathbf{G} be an operator in $\overline{B(R, \sigma)}$. ($R =$ the reals.) For an arbitrary fixed constant $\sigma' < \sigma$, define $\psi(\omega)$ on $[-\infty, \infty]$ by $\psi(\omega) = \arg \hat{\mathbf{G}}(i\omega - \sigma')$ (and by continuity at the isolated zeros of $\hat{\mathbf{G}}$) and define $\Delta(\omega) = \text{distance } \{\psi(\omega), [-\pi/2, \pi/2]\}$. Assume that $\Delta(\omega)$ and $\Delta'(\omega)$ are in $L_2((-\infty, \infty), R)$ and that there is a constant β such that $\Delta(\omega) \leq \beta < \pi/2$ on $(-\infty, \infty)$. Then there exists an operator $\mathbf{M} \in B(R, \sigma')$ such that the inequalities (a1) and (a2) in Lemma 4.3 hold for $r \in [0, 2\sigma']$.*

Proof. Define the function: $s(\omega) = -\Delta(\omega) \operatorname{sign}(\psi(\omega))$ (where $\operatorname{sign}(0) = 0$). Properties of $\hat{\mathbf{G}}(i\omega - \sigma')$ imply that $\psi(\omega)$ is odd, $\Delta(\omega)$ is even and both are locally absolutely continuous on $(-\infty, \infty)$. Also, $\psi(\omega_0) = 0$ only if $\Delta(\omega) \equiv 0$ in a neighborhood of ω_0 , so that $s(\omega)$ is locally absolutely continuous and $s'(\omega) = -\Delta'(\omega) \cdot \operatorname{sign}(\psi(\omega))$. Thus $s(\omega)$ and $s'(\omega)$ are in $L_2((-\infty, \infty), R)$ and $|s(\omega)| \leq \beta < \pi/2$. By [1, Lemma 2], there is an operator $\mathbf{Q} \in B(R, 0)$ such that $s(\omega) = \arg \hat{\mathbf{Q}}(i\omega)$ and such that $0 \notin N(\mathbf{Q}, 0)$ (in fact $\hat{\mathbf{Q}}(i\omega) \rightarrow 1$ as $|\omega| \rightarrow \infty$). Define $\mathbf{M} \in B(R, \sigma')$ by $\mathbf{M} = T(\sigma')\mathbf{Q}$ (see Lemma 4.3) so that $\hat{\mathbf{M}}(z - \sigma') = \hat{\mathbf{Q}}(z)$ on $C(0)$. Then $|\arg \hat{\mathbf{M}}(i\omega - \sigma') \hat{\mathbf{G}}(i\omega - \sigma')| = |s(\omega) + \psi(\omega)| \leq \pi/2$ implies $\hat{\mathbf{M}}(i\omega - \sigma') \hat{\mathbf{G}}(i\omega - \sigma') \in \bar{C}(0)$ for all $\omega \in [-\infty, \infty]$. Similarly the conditions $|\arg \hat{\mathbf{M}}(i\omega - \sigma')| = |s(\omega)| \leq \beta$ and $0 \notin N(\mathbf{M}, \sigma')$ imply $\hat{\mathbf{M}}(i\omega - \sigma') \in \bar{C}(\delta)$ for all $\omega \in [-\infty, \infty]$, where $\delta = \cos(\beta) \inf \{|z| | z \in N(\mathbf{M}, \sigma')\}$. One concludes that $\hat{\mathbf{M}}(z) \in \bar{C}(\delta)$ and $\hat{\mathbf{M}}(z) \hat{\mathbf{G}}(z) \in \bar{C}(0)$ for all $z \in \bar{C}(-\sigma')$ by the maximum principle. The integral

inequalities (a) and (b) are now proved by using the causality of \mathbf{M} and \mathbf{MG} and the exponentially weighted form of Parseval's theorem (see [1, pp. 505–506]).

Remark 4.5. The condition in Lemma 4.4 that $\Delta(\omega)$ and $\Delta'(\omega)$ are in $L_2((-\infty, \infty), R)$ means essentially that $\Delta(\omega) \rightarrow 0$ as $|\omega| \rightarrow \infty$, i.e., that $\limsup_{|\omega| \rightarrow \infty} |\arg \hat{\mathbf{G}}(i\omega - \sigma)| \leq \pi/2$. The condition in Lemma 4.3 that $R_\varepsilon \cap N(\mathbf{G}, \sigma) = \emptyset$ is stronger in that whenever $\{z_n\}$ is a null sequence of points in $N(\mathbf{G}, \sigma)$ (without regard to the values of ω to which the individual z_n correspond) one must have $\limsup_{n \rightarrow \infty} |\arg z_n| \leq \pi/2$.

It is readily seen in one dimension that the condition $\limsup_{|\omega| \rightarrow \infty} |\arg \hat{\mathbf{G}}(i\omega - \sigma)| \leq \pi/2$ is also necessary. Assume to the contrary for some $\mathbf{G} \in B(R, \sigma)$ that there were an $\varepsilon > 0$ and a sequence $\{\omega_n\}$ with $|\omega_n| \rightarrow \infty$ such that $|\arg \hat{\mathbf{G}}(i\omega_n - \sigma)| \geq (\pi/2) + \varepsilon$ for each n . The multiplier must satisfy $|\arg(\mathbf{MG})^\wedge(i\omega_n - \sigma)| \leq \pi/2$ for all n . But $\arg \hat{\mathbf{M}}(i\omega_n - \sigma) \rightarrow 0$ as $n \rightarrow \infty$ since $\hat{\mathbf{M}}(i\omega_n - \sigma)$ tends to a positive real constant. This gives a contradiction, since

$$|\arg \hat{\mathbf{G}}(i\omega_n - \sigma)| \leq |\arg \hat{\mathbf{M}}(i\omega_n - \sigma)| + |\arg(\mathbf{MG})^\wedge(i\omega_n - \sigma)|.$$

5. Extensions of the gain factorization lemma. In this section, we generalize the gain factorization lemma of Freedman and Zames [1, Lemma 4]. In particular, the average logarithmic variation condition is weakened to necessary and sufficient conditions for factorizations which involve the average logarithmic increase and decrease of the gain.

We let $\text{Var}(l, t_0, t_1)$, $\text{Inc}(l, t_0, t_1)$, and $\text{Dec}(l, t_0, t_1)$ be respectively the total variation, total increase, and total decrease of a real function $l(t)$ on the interval $[t_0, t_1]$.

LEMMA 5.1. *For a fixed positive real function $\phi(t)$ defined on $[0, \infty)$, the following two statements are equivalent.*

- (i) $\text{Inc}(\log \phi, t_0, t_1) \leq r(t_1 - t_0) + b$ whenever $0 \leq t_0 \leq t_1 < \infty$.
- (ii) *There exists a function $\phi_-(t)$ such that (a) $\phi_-(t) \exp(-rt)$ and $\phi_-(t)\phi(t) \exp(-rt)$ are monotone nonincreasing on $[0, \infty)$ and (b) $\exp(-b/2) \leq \phi_-(t) \leq \exp(b/2)$ throughout $[0, \infty)$.*

Proof. Assume (ii) holds, and let $l(t) = \log \phi(t)$ and $l_-(t) = \log \phi_-(t)$. Then (a') $l_-(t) - rt$ and $l_-(t) + l(t) - rt$ are monotone nonincreasing on $[0, \infty)$ and (b') $|l_-(t)| \leq b/2$ for $t \geq 0$. Condition (a') is equivalent to: (a'') $l_-(t_1) - l_-(t_0) \leq r(t_1 - t_0) - \text{Inc}(l, t_0, t_1)$ whenever $0 \leq t_0 \leq t_1 < \infty$. Thus, using (b'), we obtain $\text{Inc}(l, t_0, t_1) \leq r(t_1 - t_0) + l_-(t_0) - l_-(t_1) \leq r(t_1 - t_0) + b$, which is condition (i). Conversely, assume (i) holds. For any $t \geq 0$, define the function $l_p(t_i)$ on any partition $P = \{t_0 = 0, t_1, \dots, t_n = t\}$ of $[0, t]$ inductively as follows:

$$l_p(0) = \frac{b}{2},$$

$$l_p(t_{i+1}) = \min \{b/2, l_p(t_i) + r(t_{i+1} - t_i) - \max \{0, l(t_{i+1}) - l(t_i)\}\},$$

$i = 0, 1, \dots, n-1$. Then define $l_-(t) = \limsup_{|P| \rightarrow 0} l_p(t)$, where $|P|$ is the maximum distance between points of the partition. Then $l_-(t)$ satisfies (a'') and therefore (a'). Also, $l_-(t) \leq b/2$ follows from the definition. We now prove that $l_-(t) \geq -b/2$. Let t be any point in $[0, \infty)$ with $l_-(t) < b/2$, and let $t_0 = \sup \{\tau \in [0, t] | l_-(\tau) = b/2\}$. Assume $t_0 < t$. Then $l_-(\tau) < b/2$ on $(t_0, t]$, and for every $\delta > 0$ there is a

point $t^* \in (t_0 - \delta, t_0]$ at which $l_-(t^*) = b/2$. Then (for $\delta < t - t_0$) $l_-(t_0 + \delta) - l_-(t^*) \geq -\text{Inc}(l, t^*, t_0 + \delta)$ and $l_-(t) - l_-(t_0 + \delta) = r(t - t_0 - \delta) - \text{Inc}(l, t_0 + \delta, t)$. Therefore, $l_-(t) \geq l_-(t^*) + r(t - t^*) - \text{Inc}(l, t^*, t) - 2r\delta \geq -(b/2) - 2r\delta$. Since δ is arbitrarily small, we have $l_-(t) \geq -b/2$. Thus (ii) is satisfied by $\phi_-(t) = \exp(l_-(t))$.

LEMMA 5.2. For a fixed positive real function $\phi(t)$ defined on $[0, \infty)$, the following two statements are equivalent:

- (i) $\text{Dec}(\log \phi, t_0, t_1) \leq r(t_1 - t_0) + b$ whenever $0 \leq t_0 \leq t_1 < \infty$.
- (ii) There exists a function $\phi_-^*(t)$ such that (a) $\phi_-^*(t) \exp(-rt)$ and $(\phi_-^*(t)/\phi(t)) \exp(-rt)$ are monotone nonincreasing on $[0, \infty)$ and (b) $\exp(-b/2) \leq \phi_-^*(t) \leq \exp(b/2)$ on $[0, \infty)$.

Proof. This follows from Lemma 5.1 and the identity $\text{Dec}(\log \phi, t_0, t_1) = \text{Inc}(\log(1/\phi), t_0, t_1)$.

COROLLARY 5.3. Let $\phi(t)$ be a positive function defined on $[0, \infty)$ such that (a) $|\log \phi(t)| \leq \beta$ on $[0, \infty)$ and (b) $\text{Var}(\log \phi, t_0, t_1) \leq 2r(t_1 - t_0) + b_0$ whenever $0 \leq t_0 \leq t_1 < \infty$. Then statements (ii) of Lemmas 5.1 and 5.2 hold, where $b = (1/2) \cdot (b_0 + 2\beta)$.

Proof. Let $l(t) = \log(\phi(t))$. From the equation

$$l(t_1) - l(t_0) = \text{Inc}(l, t_0, t_1) - \text{Dec}(l, t_0, t_1),$$

one obtains

$$\text{Inc}(l, t_0, t_1) = \frac{1}{2}(\text{Var}(l, t_0, t_1) + l(t_1) - l(t_0))$$

and

$$\text{Dec}(l, t_0, t_1) = \frac{1}{2}(\text{Var}(l, t_0, t_1) + l(t_0) - l(t_1)).$$

Thus, the assumptions imply $\text{Inc}(l, t_0, t_1) \leq r(t_1 - t_0) + (1/2)(b_0 + 2\beta)$ and $\text{Dec}(l, t_0, t_1) \leq r(t_1 - t_0) + (1/2)(b_0 + 2\beta)$ when $0 \leq t_0 \leq t_1 < \infty$. Thus, the result follows from Lemmas 5.1 and 5.2.

COROLLARY 5.4. Assume for the positive real function $\phi(t)$ defined on $[0, \infty)$ that there is a positive constant T such that $\text{Inc}(\log \phi, kT, (k+1)T) \leq rT$ for $k = 0, 1, 2, \dots$. Then there exists a multiplier $\phi_-(t)$ with the properties stated in Lemma 5.1, statement (ii) (where $b = 2rT$). If, on the other hand, the aforementioned multiplier $\phi_-(t)$ exists, then for every $\varepsilon > 0$ there is a T_0 so large that $\sup_{t \geq 0} \text{Inc}(\log \phi, t, t+T) < (r + \varepsilon)T$ whenever $T \geq T_0$.

Proof. For the first part, assume $[t_0, t_1]$ intersects exactly the intervals $[kT, (k+1)T]$ for $k_0 \leq k \leq k_1$. Then $\text{Inc}(\log \phi, t_0, t_1) \leq \text{Inc}(\log \phi, k_0T, (k_0+1)T) + \dots + \text{Inc}(\log \phi, k_1T, (k_1+1)T) \leq (k_1 - k_0 + 1)rT \leq (t_1 - t_0)r + 2rT$, since $(k_1 - k_0 - 1)T \leq (t_1 - t_0)$. Thus $\phi_-(t)$ exists by Lemma 5.1. Part 2 follows immediately from statement (i) of Lemma 5.1 when $\phi_-(t)$ exists.

COROLLARY 5.5. Corollary 5.4 holds when Inc is replaced by Dec and $\phi_-(t)$ by $\phi_-^*(t)$.

Remark 5.6. Corollaries 5.4 and 5.5 show that the appropriate multiplier $\phi_-(t)$ or $\phi_-^*(t)$ exists if the average logarithmic increase or decrease of $\phi(t)$ over the intervals $[kT, (k+1)T]$ is bounded by r , and on the other hand, if $\phi_-(t)$ and $\phi_-^*(t)$ exist, then the average logarithmic increase or decrease of $\phi(t)$ exceeds r by at most an arbitrarily small amount on sufficiently long intervals.

6. The stability results. In Theorems 6.1 and 6.2, Theorem 1 of Freedman and Zames [1] is extended to the systems of equations (1) and (2). The theorems naturally include the case where $\phi(t)$ is not uniformly logarithmically bounded on $[0, \infty)$, and more general conditions involving the average logarithmic increase and decrease replace the average logarithmic variation condition on $\phi(t)$. Also, in Theorem 6.2 the assumption that $N(\mathbf{G}, \sigma)$ does not intersect the origin is removed. Theorem 6.4 generalizes Theorem 2 of [1] in the case where $\phi(t)$ lies between two infinite bounds, and shows in particular that when $\phi(t)$ is uniformly bounded on $[0, \infty)$, it is sufficient for $N(\mathbf{G}, \sigma)$ to not intersect a certain proper subinterval $(-\infty, -\varepsilon]$ of $(-\infty, 0]$. Thus the improved condition concerning the origin in Theorem 6.2 is of no significance in this case. However, when $\phi(t)$ is unbounded, the transformation [1, p. 502] by which Theorem 6.4 was obtained from Theorem 6.1 is not valid. Thus Theorem 6.2 is the only result which applies to systems which involve a convolution operator without an adjoined identity and for which the gain has no uniform bound on $[0, \infty)$.

THEOREM 6.1. *Assume in the feedback system of (1) and (2) that Assumptions 1, 2 and 3 hold. Further assume*

(a) \mathbf{G} is approximable in $\overline{B(H, \sigma)}$ (for $\sigma > 0$) and $\hat{\mathbf{G}}(z)$ is a normal operator for each $z \in \bar{C}(-\sigma)$;

(b) $(-\infty, 0] \cap N(\mathbf{G}, \sigma) = \emptyset$, and

(c) $\phi(t) > 0$ and there exist nonnegative constants r and b with $r < 2\sigma$ such that one of the following conditions holds:

(c1) $\text{Inc}(\log \phi, t_0, t_1) \leq r(t_1 - t_0) + b$ whenever $0 \leq t_0 \leq t_1 < \infty$.

(c2) $\text{Dec}(\log \phi, t_0, t_1) \leq r(t_1 - t_0) + b$ whenever $0 \leq t_0 \leq t_1 < \infty$.

Then the system is L_2 -stable with respect to $y(t)$ and $e(t)$.

Proof. Equations (1) and (2) can be combined in the forms

$$(3) \quad f = e + \Phi \mathbf{G} e$$

and

$$(4) \quad Gf = y + \mathbf{G}\phi y,$$

where $(\Phi y)(t) = \phi(t)y(t)$. From (3) and (4), one obtains, respectively, the conservation of energy equations

$$(5) \quad \int_0^T \phi_-(t) \langle f(t), (\mathbf{M}_0 \mathbf{G} e)(t) \rangle dt = p \int_0^T \langle e(t), (\mathbf{M}_0 \mathbf{G} e)(t) \rangle dt \\ + \int_0^T (\phi_-(t) - p) \langle e(t), (\mathbf{M}_0 \mathbf{G} e)(t) \rangle dt + \int_0^T \phi_-(t) \phi(t) \langle y(t), (\mathbf{M}_0 y)(t) \rangle dt,$$

$$(6) \quad \int_0^T \phi_+^*(t) \langle y(t), (\mathbf{M}_0 \mathbf{G} f)(t) \rangle dt = p \int_0^T \langle y(t), (\mathbf{M}_0 y)(t) \rangle dt \\ + \int_0^T (\phi_+^*(t) - p) \langle y(t), (\mathbf{M}_0 y)(t) \rangle dt + \int_0^T (\phi_+^*(t)/\phi(t)) \langle u(t), (\mathbf{M}_0 \mathbf{G} u)(t) \rangle dt,$$

where \mathbf{M}_0 has the properties stated in Lemma 4.3 and $\phi_-(t)$ and $\phi_+^*(t)$ (having the properties stated in parts (ii) of Lemmas 5.1 and 5.2) exist respectively under conditions (c1) and (c2). Also, $u(t) = \phi(t)y(t)$, $p = (1 - r/(2\sigma)) \exp(-b/2)$, and

$T \geq 0$. The functions $(\phi_-(t) - p) \exp(-2\sigma t)$ and $(\phi^*(t) - p) \exp(-2\sigma t)$ are both positive and monotone nonincreasing on $[0, \infty)$. Therefore, under the respective conditions (c1) and (c2), all the integrals on the right-hand sides of (5) and (6) are nonnegative as a consequence of (b1) and (b2) in Lemma 4.3 and the second mean value theorem. By further use of Lemma 4.3, we obtain

$$(7) \quad p\delta \int_0^T |e(t)|^2 dt \leq \int_0^T \phi_-(t) \langle f(t), (\mathbf{M}_0 \mathbf{G}e)(t) \rangle dt$$

under condition (c1), and

$$(8) \quad p\delta \int_0^T |y(t)|^2 dt \leq \int_0^T \phi^*(t) \langle y(t), (\mathbf{M}_0 \mathbf{G}f)(t) \rangle dt$$

under condition (c2). Therefore, if

$$A = \left| \frac{\exp(b) \|\mathbf{M}_0 \mathbf{G}\|_{w,0}}{(1 - r/(2\sigma))} \right|,$$

then $\|e_T\|_0 \leq A \|f\|_0$ for all $T \geq 0$ in the case (c1) or $\|y_T\|_0 \leq A \|f\|_0$ for all $T \geq 0$ in the case (c2). The stability of the system with respect to e in the case (c1) or with respect to y in the case (c2) follows from the fact that A is independent of T in these inequalities. However, stability with respect to e implies stability with respect to y and vice versa in the present context, because $y = \mathbf{G}e$ where \mathbf{G}^{-1} exists in $B(H, \sigma)$.

THEOREM 6.2. *In Theorem 6.1, let assumptions (b) and (c) be replaced by:*

(b') *There is an $\varepsilon > 0$ so small that $R_\varepsilon \cap N(\mathbf{G}, \sigma) = \emptyset$.*

(c') *$\phi(t) > 0$ and there exist nonnegative constants r and b , with $r < 2\sigma$, such that $\text{Dec}(\log \phi, t_0, t_1) \leq r(t_1 - t_0) + b$ whenever $0 \leq t_0 \leq t_1 < \infty$.*

Then under the new assumptions the system is L_2 -stable with respect to y .

Proof. The proof is the same as that of Theorem 6.1 in the case (c2) with the exception that the multiplier \mathbf{M}_ε satisfying the inequalities (a1) and (a2) of Lemma 4.3 is used in the place of \mathbf{M}_0 .

DEFINITION 6.3. The *critical region* $R(x_1, x_2)$ for the interval (x_1, x_2) is the set of real numbers x such that $-1/x \in [x_1, x_2]$.

THEOREM 6.4. *In Theorem 6.1, let conditions (b) and (c) be replaced by:*

(b') *$R(x_1, x_2) \cap N(\mathbf{G}, \sigma) = \emptyset$ for real numbers x_1 and x_2 , and there is at least one point in $R(x_1, x_2)$ which is not in $\sum(\mathbf{G}, \sigma)$.*

(c') *$x_1 < \phi(t) < x_2$ for all $t \geq 0$, and there exist nonnegative constants r and b with $r < 2\sigma$ such that $\phi^*(t) = (\phi(t) - x_1)(x_2 - \phi(t))^{-1}$ satisfies one of the conditions (c1) and (c2) of Theorem 6.1.*

Then the system of equations (1) and (2) is L_2 -stable with respect to both e and y .

Proof. This is obtained from Theorem 6.1 by means of the transformation in [1, p. 502].

COROLLARY 6.5. *In Theorems 6.1, 6.2 and 6.4, let σ , r , and b be arbitrary with the exceptions that $r, b \geq 0$. (The restrictions $\sigma > 0$ and $r < 2\sigma$ are removed.) Then in Theorems 6.1 and 6.4 the system is $L_{2\alpha}$ -stable with respect to e and y whenever $\alpha < \sigma - (r/2)$. Similarly, in Theorem 6.2 the system is $L_{2\alpha}$ -stable with respect to y if $\alpha < \sigma - (r/2)$.*

Proof. The situation is transformed into that of Theorems 6.1, 6.2 and 6.4 as follows. Let f_x, e_x and y_x be defined according to the rule $f_x(t) = f(t) \exp(\alpha t)$. Define $\mathbf{G}_x \in \overline{B(H, \sigma - \alpha)}$ by $\mathbf{G}_x = T(-\alpha)\mathbf{G}$, where $T(\alpha)$ was used in the proof of Lemma 4.3. Then $e_x = f_x - \phi y_x$ and $y_x = \mathbf{G}_x e_x$.

Remarks 6.6.

(a) The conditions involving the average logarithmic increase and decrease in the theorems of this section can be replaced by possibly more convenient conditions given in Corollaries 5.3, 5.4, and 5.5. In particular, in the case where $\phi(t)$ is uniformly logarithmically bounded, the bound of r on the average logarithmic increase or decrease in Theorem 6.1 is seen to be essentially equivalent to the bound of $2r$ on the average logarithmic variation in [1, Thm. 1].

(b) A bound on the average logarithmic increase of $\phi(t)$ places no restriction on the manner in which it decreases, and similarly a bound on the average logarithmic decrease in no way restricts the manner in which it increases. Thus, under the conditions (a) and (b) of Theorem 6.1, the only positive gain function which could cause instability would be one which oscillates so strongly that for every $T > 0$ there exist arbitrarily positive intervals of length T on which the average logarithmic increase is at least 2σ and other arbitrarily positive intervals on which the average logarithmic decrease is at least 2σ .

(c) In Theorem 6.4 the second part of condition (b') is fulfilled automatically unless x_1 and x_2 are either both positive or both negative, since in the other cases $R(x_1, x_2)$ consists of semi-infinite intervals.

(d) In Theorem 6.2, if \mathbf{G} is also in $B(R, \sigma)$, then application of Lemma 4.4 shows that condition (b') can be generalized to the following condition.

(b'') There is a point $\sigma' \in (0, \sigma)$ such that

$$N(\mathbf{G}, \sigma') \cap \{z = x + iy | x < 0 \text{ and } |y| < \alpha|x|\} = \emptyset$$

for a sufficiently small positive number α and such that $\Delta(\omega)$ and $\Delta'(\omega)$ (at σ') are in $L_2((-\infty, \infty), R)$. (It is then required in condition (c') that $r < 2\sigma'$).

Since $N(\mathbf{G}, \sigma')$ is a continuous closed curve when $\mathbf{G} \in \overline{B(R, \sigma)}$, the condition (b'') means essentially that if the curve $\hat{\mathbf{G}}(i\omega - \sigma')$ passes through 0 at a finite ω it must in doing so avoid a thin cone about the negative real axis, whereas if $\hat{\mathbf{G}}(i\omega - \sigma') \rightarrow 0$ as $|\omega| \rightarrow \infty$, then the limiting approach angle with the positive real axis must not exceed 90° .

(e) Theorem 6.2 also applies to the equation $f(t) = y(t) + (\mathbf{G}\phi y)(t)$, where $(\phi y)(t) = \phi(t)y(t)$. This equation is equivalent to (1) and (2) except in the case where \mathbf{G}^{-1} fails to exist, as can occur under the conditions of Theorem 6.2.

7. Positive integrals.

LEMMA 7.1. Assume in a real Hilbert space H that:

(a) $K(t)$ is a self-adjoint function in $L_1([0, T], \mathcal{L}(H, H))$ such that $K(t) \geq 0$ and $K(t) \exp(-2\sigma t)$ is monotone nonincreasing on $[0, T]$.

(b) $\mathbf{G} \in \overline{B(R, \sigma)}$ and $N(\mathbf{G}, \sigma) \subset \overline{C}(0)$.

Then

$$J(y) = \int_0^T \langle (\mathbf{G}y)(t), K(t)y(t) \rangle dt \geq 0$$

for all functions $y(t)$ in $L_2([0, T], H)$.

Proof. One obtains $J(y) = \int_0^T \langle (\mathbf{G}_\sigma y_\sigma)(t), K_\sigma(t) y_\sigma(t) \rangle dt$, where $\mathbf{G}_\sigma = T(-\sigma)\mathbf{G}$, $K_\sigma(t) = \mathbf{K}(t) \exp(-2\sigma t)$ and $y_\sigma(t) = y(t) \exp(\sigma t)$. ($T(-\sigma)$ is defined in the proof of Lemma 4.3.) In the case where $K_\sigma(t)$ is a constant, we obtain $J(y) = \int_0^T \langle (\mathbf{G}_\sigma u_\sigma)(t), u_\sigma(t) \rangle dt$, where $u_\sigma(t) = \sqrt{K_\sigma} y_\sigma(t)$. Therefore, $J(y) \geq 0$ by Lemma 4.4. The general case reduces to this as follows. Under the assumptions in (a), $K_\sigma(t)$ can be arbitrarily closely approximated in the norm on $L_1([0, T], \mathcal{L}(H, H))$ by operator functions which are nonnegative, nonincreasing and piecewise constant on $[0, T]$. If $K^*(t)$ is such a function and $K^*(t) = K_i$ for $t \in [t_i, t_{i+1})$ where $t_0 = 0 < t_1 < t_2 \cdots < t_n = T$, then

$$\int_0^T \langle (\mathbf{G}_\sigma y_\sigma)(t), K^*(t) y_\sigma(t) \rangle dt = \sum_{i=1}^n \int_0^{t_i} \langle (\mathbf{G}_\sigma y_\sigma)(t), (K_{i-1} - K_i) y_\sigma(t) \rangle dt.$$

All the terms in the sum are nonnegative.

Remark 7.2. If $K(t)$ is absolutely continuous in the operator norm, then Lemma 7.1 can be proven by using the equation:

$$\begin{aligned} & \int_0^T \langle (\mathbf{G}y)(t), K(t)y(t) \rangle dt \\ &= \int_0^T \langle (\mathbf{G}y)(t), K(T)y(t) \rangle dt + \int_0^T \int_0^t \langle (\mathbf{G}y)(\tau), (-K'(\tau))y(\tau) \rangle d\tau dt. \end{aligned}$$

LEMMA 7.3. Assume (where H is real) that:

- (a) $K(t)$ is an in general unbounded self-adjoint operator function on $[0, T]$, with $K(t) > 0$ at each t . The bounded operator $K^{-1}(t)$ is in $L_1([0, T], \mathcal{L}(H, H))$ and $K^{-1}(t) \exp(2\sigma t)$ is monotone nondecreasing.
- (b) $\mathbf{G} \in \overline{B(R, \sigma)}$ and $N(\mathbf{G}, \sigma) \subset \overline{C}(0)$.

Then

$$J(y) = \int_0^T \langle (\mathbf{G}y)(t), K(t)y(t) \rangle dt \geq 0$$

whenever $y(t) \in \text{domain } (K(t))$ a.e. and the functions $y(t)$ and $K(t)y(t)$ are in $L_2([0, T], H)$.

Proof. One sees through the substitutions $u(t) = K(t)y(t)$ and $v(t) = (K^{-1}(t) + \delta I)u(t)$ that it is sufficient to prove $J(\delta, v) = \int_0^T \langle (\mathbf{G}v)(t), (K^{-1}(t) + \delta I)^{-1}v(t) \rangle dt \geq 0$ for all $\delta > 0$. But for all $\delta > 0$, $(K^{-1}(t) + \delta I)^{-1}$ is a bounded operator and $(K^{-1}(t) + \delta I)^{-1} \exp(-2\sigma t)$ is monotone nonincreasing.

8. Stability theory for systems involving a time-varying unbounded operator.

The following system is considered:

$$(9) \quad e(t) = f(t) - K(t)y(t),$$

$$(10) \quad y(t) = (\mathbf{G}e)(t),$$

where Assumptions 1–3 are replaced by:

ASSUMPTION 1'. \mathbf{G} is an operator in $\overline{B(R, \sigma)}$ for a positive number σ . (\mathbf{G} is then also an operator in $\overline{B(H, \sigma)}$.)

ASSUMPTION 2'. At each $t \geq 0$, $K(t)$ is a self-adjoint, in general unbounded operator in the real Hilbert space H .

ASSUMPTION 3'. For every input $f(t)$ in $L_{2e}([0, \infty), H)$, the solutions $e(t)$ and $y(t)$ exist in $L_{2e}([0, \infty), H)$. This means in particular that $y(t) \in \text{domain}(K(t))$ a.e. and that $K(t)y(t)$ is a function in $L_{2e}([0, \infty), H)$.

The proofs of Theorem 6.1 and 6.2 are easily adapted through application of Lemmas 4.4 and 7.1 and 7.3 to yield the following results.

THEOREM 8.1. Assume in the feedback system of equations (9) and (10) that Assumptions 1', 2' and 3' hold. Further assume:

(a) $(-\infty, 0] \cap N(\mathbf{G}, \sigma) = \emptyset$;

(b) $K(t)$ is a function in $L_{1e}([0, \infty), \mathcal{L}(H, H))$ such that $K(t) \geq 0$ for all $t \geq 0$. There exists a self-adjoint operator function $K^*(t)$ in $L_{1e}([0, \infty), \mathcal{L}(H, H))$ such that $K(t)K^*(t) = K^*(t)K(t)$ and $0 < \underline{b}I \leq K^*(t) \leq \bar{b}I < \infty$ at each $t \geq 0$ and such that, for some $r < 2\sigma$, one of the following conditions holds:

(c1) $K^*(t)\exp(-rt)$ and $K^*(t)K(t)\exp(-rt)$ are monotone nonincreasing on $[0, \infty)$;

(c2) $K^*(t)\exp(rt)$ and $K^*(t)K(t)\exp(rt)$ are monotone nondecreasing on $[0, \infty)$. Then the system is L_2 -stable with respect to $y(t)$ and $e(t)$.

Remark 8.2. The case where $K(t) \geq 0$ is an unbounded operator can be reduced to the case of Theorem 8.1 as follows. For $\delta > 0$, let $K_\delta(t) = (\delta I + K(t))^{-1}$, $\mathbf{G}_\delta = \mathbf{G}^{-1}(\mathbf{A} - \delta \mathbf{G})$, $e_\delta(t) = e(t) - \delta y(t)$ and $f_\delta(t) = K_\delta(t)f(t)$. Then (9) and (10) are equivalent to $f_\delta(t) = y(t) + K_\delta(t)e_\delta(t)$ and $e_\delta(t) = (\mathbf{G}_\delta y)(t)$. If $(-\infty, 0] \cap N(\mathbf{G}, \sigma) = \emptyset$, then there is a $\delta > 0$ so small that $(-\infty, 0] \cap N(\mathbf{G}_\delta, \sigma) = \emptyset$. Then the system is L_2 -stable with respect to $y(t)$ and $e(t)$ if $K_\delta(t)$ satisfies condition (b) in Theorem 8.1. Of course one can set $\delta = 0$ if $K(t)$ is already uniformly strongly positive.

THEOREM 8.3. Assume that there are real constants x_1 and x_2 such that $x_1 I < K(t) < x_2 I$. Then the conclusion of Theorem 8.1 holds if:

(a) $R(x_1, x_2) \cap N(\mathbf{G}, \sigma) = \emptyset$ and the winding number of $N(\mathbf{G}, \sigma)$ about $R(x_1, x_2)$ is 0;

(b) $L(t) = (K(t) - x_1 I)(x_2 I - K(t))^{-1}$ satisfies the conditions of Theorem 8.1, statement (b).

In the case where the Nyquist Diagram intersects the origin, we have the following.

THEOREM 8.4. Assume in the feedback system of (9) and (10) that Assumptions 1', 2' and 3' hold. Further assume:

(a) There is a $\sigma' \in (0, \sigma)$ such that

$$\{z = x + iy \mid x < 0 \text{ and } |y| < \alpha|x|\} \cap N(\mathbf{G}, \sigma') = \emptyset$$

for α sufficiently small and such that $\Delta(\omega)$ and $\Delta'(\omega)$ are in $L_2((-\infty, \infty), \mathbf{R})$, where $\Delta(\omega) = \text{distance} \{ \arg \hat{\mathbf{G}}(i\omega - \sigma'), [-\pi/2, \pi/2] \}$;

(b) $K(t) > 0$ for each $t \geq 0$, and $K^{-1}(t)$ is a function in $L_{1e}([0, \infty), \mathcal{L}(H, H))$. There exists a self-adjoint operator function $K_-(t)$ in $L_{1e}([0, \infty), \mathcal{L}(H, H))$ such that $K^{-1}(t)K_-(t) = K_-(t)K^{-1}(t)$ and $0 < \underline{b}I \leq K_-(t) \leq \bar{b}I < \infty$ at each $t \geq 0$ and such that $K_-(t)\exp(-rt)$ and $K_-(t)K^{-1}(t)\exp(-rt)$ are both monotone nonincreasing on $[0, \infty)$.

Then the system is L_2 -stable with respect to $y(t)$.

Remarks 8.5.

(a) If $K(t) \geq 0$ in Theorem 8.4, then it suffices to have the operator function $\delta I + K(t)$ satisfying the conditions (b) for all $\delta > 0$.

(b) In the case where $K(t)$ is unbounded, the conditions on the behavior of $N(\mathbf{G}, \sigma')$ near the origin in Theorem 8.4 are critical because there is no extension of Theorem 8.1 (in the unbounded operator case, see Remark 8.2) of the type given in Theorem 8.3.

(c) Corollary 6.5 also applies to Theorems 8.1, 8.3, and 8.4.

(d) The same proof shows that Theorem 8.4 also applies to the equation $f(t) = y(t) + (\mathbf{G}\mathbf{K}y)(t)$ (where $(\mathbf{K}y)(t) = K(t)y(t)$). This equation is equivalent to (9) and (10) in the case where \mathbf{G} is invertible, but not otherwise, as in the case of Theorem 8.4.

It would be ideal, for use in Theorem 8.1, 8.3, and 8.4 to know necessary and sufficient conditions on the operator function $K(t)$ under which there would exist one of the multipliers $K^*(t)$ with the properties of statement (b) of Theorem 8.1. Such a result would be an extension of Lemmas 5.1 and 5.2 to the operator context. These conditions are not known. However, Lemmas 5.1 and 5.2 are easily extended to give the necessary and sufficient conditions under which $K^*(t)$ exists in the form of a scalar multiplier. This result is based on the following definitions.

DEFINITION 8.6. The *logarithmic increase* $\text{LI}(\mathbf{K}, a, b)$ of operator function $K(t)$ on the interval $[a, b]$ is given by

$$\text{LI}(\mathbf{K}, a, b) = \sup \sum_{i=1}^{m-1} \inf \{ \alpha \geq 0 \mid \exp(\alpha) K(t_i) \geq K(t_{i+1}) \},$$

where the sup is taken over all partitions

$$p = \{t_1 = a < t_2 < \cdots < t_n = b\} \text{ of } [a, b].$$

DEFINITION 8.7. The *average logarithmic decrease* $\text{LD}(\mathbf{K}, a, b)$ of an operator function $K(t)$ on the interval $[a, b]$ is the logarithmic increase of the function $K(-t)$ on the interval $[-b, -a]$.

Remark 8.8. If $K(t)$ is locally absolutely continuous in the operator norm and $K'(t)$ is the derivative, then

$$\text{LI}(\mathbf{K}, a, b) = \int_a^b \alpha(t) dt, \quad \text{where } \alpha(t) = \inf \{ \alpha \geq 0 \mid \alpha K(t) \geq K'(t) \}.$$

Similarly,

$$\text{LD}(\mathbf{K}, a, b) = \int_a^b \alpha(t) dt, \quad \text{where } \alpha(t) = \inf \{ \alpha \geq 0 \mid \alpha K(t) \geq -K'(t) \}.$$

LEMMA 8.9. Let $K(t)$ be a nonnegative bounded operator-valued function on $[0, \infty)$. Then the following two statements are equivalent:

- (i) $\text{LI}(\mathbf{K}, t_0, t_1) \leq r(t_1 - t_0) + b$ whenever $0 \leq t_0 \leq t_1 < \infty$;
- (ii) There exists a real function $\phi_-(t)$ such that
 - (a) $\phi_-(t) \exp(-rt)$ and $\phi_-(t)K(t) \exp(-rt)$ are nonnegative and monotone nonincreasing on $[0, \infty)$ and
 - (b) $\exp(-b/2) \leq \exp(b/2)$ on $[0, \infty)$.

Proof. If $l_-(t) = \log \phi_-(t)$, then condition (a) of statement (ii) is equivalent to (a'): $l_-(t_1) - l_-(t_0) \leq r(t_1 - t_0) - \text{LI}(\mathbf{K}, t_0, t_1)$ whenever $0 \leq t_0 \leq t_1 < \infty$. The proof using this fact is essentially the same as the proof of Lemma 5.1.

LEMMA 8.10. *Let $K(t)$ be a nonnegative bounded operator-valued function on $[0, \infty)$. Then the following statements are equivalent:*

(i) $\text{LD}(\mathbf{K}, t_0, t_1) \leq r(t_1 - t_0) + b$ whenever $0 \leq t_0 \leq t_1 < \infty$.

(ii) *There exists a real function $\phi_+(t)$ such that*

(a) $\phi_+(t) \exp(rt)$ and $\phi_+(t)K(t) \exp(rt)$ are nonnegative and monotone non-decreasing on $[0, \infty)$ and

(b) $\exp(-b/2) \leq \phi_+(t) \leq \exp(b/2)$ on $[0, \infty)$.

The stability theorems in this section are directly applicable to a class of boundary and initial value problems in partial differential equations as well as initial value problems for systems of ordinary differential equations, either of which can be expressed in the form:

$$(11) \quad P(D_t)y(t) + Q(D_t)(K(t)y(t)) = f(t),$$

$$(12) \quad D_t^l(0) = y_l, \quad l = 0, 1, \dots, p-1,$$

where $P(D_t)$ and $Q(D_t)$ are real-coefficient polynomials of respective degrees p and q (with $p \geq q$) in the time derivative D_t and $K(t)$ is a self-adjoint operator function. For the case where $K(t)$ is positive and either unbounded or not uniformly bounded, Theorem 8.4 yields the following result.

COROLLARY 8.11. *Assume for an input $f(t)$ in $L_2([0, \infty), H)$ in (11) and (12) that a solution $y(t)$ exists such that $D_t^i y(t)$ and $D_t^j (K(t)y(t))$ are locally absolutely continuous on $[0, \infty)$ for $i \leq p-1$ and $j \leq q-1$ and in $L_{2e}([0, \infty), H)$ for $i \leq p$ and $j \leq q$. (This requires further assumptions concerning $K(t)$ which we don't go into.) Further assume there exists a $\sigma > 0$ such that all the roots of $P(z)$ lie in the region $\text{Re}(z) < -\sigma$ and such that:*

(a) *there exists an $\alpha > 0$ such that*

$$\left\{ \frac{Q(i\omega - \sigma)}{P(i\omega - \sigma)} \middle| \omega \in [-\infty, \infty] \right\} \cap \{z = x + iy | x < 0 \text{ and } |y| < \alpha|x|\} = \emptyset,$$

(b) *either $p - q$ is odd or else $p - q = 2k$ and the sign of the product of the leading coefficients of $P(z)$ and $Q(z)$ is the same as that of $(-1)^k$ and*

(c) $K(t) > 0$ for each $t \geq 0$, $K^{-1}(t)$ is a function in $L_{1e}([0, \infty), \mathcal{L}(H, H))$, and there exist positive constants r and b , with $r < 2\sigma$, such that $\text{LI}(\mathbf{K}^{-1}, t_0, t_1) \leq r(t_1 - t_0) + b$ whenever $0 \leq t_0 \leq t_1 < \infty$. Then $y(t)$ is in $L_2([0, \infty), H)$.

Proof. $y(t)$ is shown to satisfy an equation of the form $y(t) + (\mathbf{GK}y)(t) = f_0(t)$, where $f_0(t)$ is in $L_2([0, \infty), H)$ and the Nyquist diagram of G is the plot of the function $(Q/P)(i\omega - \sigma)$ over all ω .

Remark 8.12. Only Theorem 8.4 can be applied to the above problem in the case where $q < p$, since then the Nyquist diagram always enters the origin as $|\omega| \rightarrow \infty$. In the case where $q = p$, Theorem 8.1 (with Remark 8.2) yields an analogous result.

REFERENCES

- [1] M. FREEDMAN AND G. ZAMES, *Logarithmic variation criteria for the stability of systems with time-varying gains*, this Journal, 6 (1968), pp. 487–507.
- [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 G. ZAMES, *A Hilbert space stability theory over locally compact Abelian groups*, this Journal, 7 (1969), pp. 479–495.
- [4] A. ACKER, *Stability results for linear systems involving a time-varying unbounded operator*, Doctoral thesis, Boston Univ., Boston, 1972.

ON NORMALITY AND CONJUGATE POINT CRITERIA FOR SINGULAR EXTREMALS*

VIOLET B. HAAS†

Abstract. The Moore–Penrose generalized inverse is employed in the control problem of Lagrange to obtain necessary and sufficient conditions for normality, sufficient conditions for the nonexistence of conjugate points and sufficient conditions for the nonnegativity of the second variation. Our results are valid for both regular and singular extremal arcs.

1. Introduction. Let f be a mapping from $\mathcal{S} = E^1 \times E^n \times E^m$ into E^n , let f_0 be a mapping from \mathcal{S} into E^1 , and suppose f and f_0 have continuous second order partial derivatives with respect to all variables in some open region \mathcal{S}_0 contained in \mathcal{S} . Let \mathcal{U} be the class of mappings, $u: E^1 \rightarrow E^m$ which are piecewise continuous and have a piecewise continuous derivative on the interval $T: [t_0, t_f]$. We consider the problem of minimizing the functional

$$J(u) = \int_{t_0}^{t_f} f_0(t, x(t), u(t)) dt$$

on the class \mathcal{U} subject to the differential equation and endpoint constraints,

$$(1.1) \quad \dot{x} = f(t, x, u)$$

and

$$(1.2) \quad x(t_0) = a, \quad x(t_f) = b.$$

It is assumed, henceforth, that all arcs $x = x(t)$, $u = u(t)$, $t \in T$ for $u \in \mathcal{U}$, which satisfy (1.1) and whose properties are discussed here lie in \mathcal{S}_0 . Such an arc shall be said to be admissible.

Let

$$H = H(t, x, u, p_0, p) = p_0 f_0(t, x, u) + p^T f(t, x, u)$$

denote the Hamiltonian for this problem, where p is a vector, and the superscript T denotes transposition. An extremal arc is a curve \mathcal{C} described by $x = x(t)$, which is admissible and along which there exists a set of multipliers $p_0, p(t)$ with p_0 constant such that

$$(1.3) \quad \dot{p}^T = -H_x$$

and

$$(1.4) \quad H_u = 0.$$

Equations (1.1), (1.3) and (1.4) are known as the Euler–Lagrange equations. Variations of the state x , the control u , and the multiplier p shall be denoted by ξ, η, ζ , respectively

* Received by the editors August 5, 1974, and in revised form November 22, 1974.

† School of Electrical Engineering, Purdue University, West Lafayette, Indiana 47907.

The accessory minimum problem is that of minimizing the second variation,

$$J_2(\eta) = \frac{1}{2} \int_{t_0}^{t_f} 2\omega(t, \xi, \eta) dt,$$

where

$$(1.5) \quad 2\omega = \xi^T H_{xx} \xi + 2\eta^T H_{ux} \xi + \eta^T H_{uu} \eta$$

subject to the constraints

$$(1.6) \quad \dot{\xi} = f_x \xi + f_u \eta$$

and

$$(1.7) \quad \xi(t_0) = \xi(t_f) = 0,$$

where f_x , f_u , H_{xx} , H_{ux} and H_{uu} are evaluated along \mathcal{C} .

A control u or a control variation η shall be called “admissible” if it belongs to the class \mathcal{U} . A variation ξ shall be called “admissible” if it corresponds to an admissible control variation via (1.6). If $x = x(t; u)$ describes an admissible trajectory arc on $[t_0, t_f]$, we shall suppose that the arc described by $x = x(t; u + \varepsilon\eta)$ is admissible for all $|\varepsilon|$ sufficiently small whenever $\eta(t)$ is admissible. An accessory extremal is a variation $\xi(t)$ which together with a multiplier variation $\zeta(t)$ satisfies along \mathcal{C} equation (1.6),

$$(1.8) \quad \dot{\zeta} = -H_{xx} \xi - H_{xu} \eta - f_x^T \zeta$$

and

$$(1.9) \quad H_{uu} \eta + H_{ux} \xi + f_u^T \zeta = 0$$

for some admissible control variation η . We shall refer to equations (1.6), (1.8) and (1.9) as the “Jacobi equations”.

We shall allow H_{uu} to be singular along \mathcal{C} . H^+ shall denote the Moore–Penrose generalized inverse of H_{uu} evaluated along an extremal arc \mathcal{C} . The generalized inverse M^+ of any matrix M is defined by

$$M^+ = \lim_{\delta \rightarrow 0} (M^T M + \delta^2 I)^{-1} M^T,$$

where I is the identity matrix of appropriate dimensions. It can be shown (see [1]) that this limit always exists and that

$$M^+ = \lim_{\delta \rightarrow 0} M^T (M M^T + \delta^2 I)^{-1}.$$

Let $\mathcal{R}(M)$ denote the range of the matrix M and let $\mathcal{N}(M)$ denote its nullity. Any matrix M satisfies $\mathcal{R}(M^T) = \mathcal{R}(M^+)$ and also $M^+ M M^+ = M^+$, $M M^+ M = M$. If M is square and nonsingular, then $M^+ = M^{-1}$ and if M is a zero matrix, so is M^+ . If v is any vector in E^m , then $v = \hat{v} + \tilde{v}$, where \hat{v} is the projection of v on $\mathcal{R}(H_{uu})$, \tilde{v} is the projection of v on $\mathcal{N}(H_{uu})$ and $H^+ H_{uu} v = \hat{v}$.

It is well known (see [2]) that if the admissible control u minimizes J and if \mathcal{C} is the corresponding trajectory satisfying (1.1) and (1.2), then \mathcal{C} must be an extremal and along \mathcal{C} ,

$$\eta^T H_{uu} \eta \geq 0,$$

H is a continuous function of t and if \mathcal{C} is normal then $J_2(\eta) \geq 0$ for all admissible variations η .

In the sequel we shall frequently use the following hypothesis which we shall denote by \mathcal{H} .

\mathcal{H} : There exists a continuously differentiable symmetric matrix $P(t)$ such that along \mathcal{C} ,

$$\mathcal{N}(H_{xu} + Pf_u) \supseteq \mathcal{N}(H_{uu}).$$

This hypothesis implies that

$$\mathcal{R}(H_{ux} + f_u^T P) \subseteq \mathcal{R}(H_{uu}).$$

It will become necessary to imbed a particular extremal arc \mathcal{C} in a one-parameter family of such arcs. In order to accomplish this we shall at times need to suppose that the Euler–Lagrange equations together with the equations $H_u = 0$ and $\partial H/\partial t = dH/dt$ have along \mathcal{C} a solution $u = u(t, x, p)$ which has continuous first order partial derivatives in a neighborhood of \mathcal{C} in t, x, p -space. This will be the case if there exists a nonnegative integer q such that the matrix

$$\left(H_{uu}, \left(\frac{\partial}{\partial u} G \right)^T, \left(\frac{\partial}{\partial u} \frac{d}{dt} G \right)^T, \dots, \left(\frac{\partial}{\partial u} \frac{d^q}{dt^q} G \right)^T \right)$$

has maximal rank on \mathcal{C} , where G is the function $H_x \dot{x} + H_p \dot{p}$. In case $q \neq 0$ we shall, of course, require that H have continuous derivatives on \mathcal{C} of order $q + 2$. The hypothesis needed to enable us to solve for $u = u(t, x, p)$ shall be denoted by \mathcal{H}_0 , and this will be assumed throughout.

2. Normality. A subarc of an extremal arc \mathcal{C} is said to be *abnormal* [3] if there exists a nonidentically vanishing vector p which satisfies

$$(2.1) \quad \dot{p} = -f_x^T p$$

and

$$(2.2) \quad f_u^T p = 0$$

along this subarc. If \mathcal{C} has no abnormal subarcs, then \mathcal{C} is said to be *normal*. If \mathcal{C} has an abnormal subarc on the (nondegenerate) subinterval $[t', t'']$ of T , then there exists on this subinterval a one-parameter family of solutions

$$(2.3) \quad p(t) = p^{(0)}(t) + \mu p^{(1)}(t)$$

of (2.1) and (2.2) with $p^{(1)} \neq 0$. We now invoke hypothesis \mathcal{H}_0 so that the functions u and x also become dependent upon the parameter μ . If we set

$$\xi = \frac{\partial x}{\partial \mu}, \quad \zeta = \frac{\partial p}{\partial \mu}, \quad \eta = \frac{\partial u}{\partial \mu},$$

we obtain the Jacobi equations (1.6), (1.8) and (1.9) and ζ satisfies

$$(2.4) \quad \dot{\zeta} = -f_x^T \zeta, \quad f_u^T \zeta = 0.$$

Suppose that $\xi \equiv 0$ on $[t', t'']$. Then from (1.6) and (1.9) it follows that

$$(2.5) \quad f_u \eta = 0$$

and

$$(2.6) \quad H_{uu} \eta = 0.$$

Thus, $\eta \in \mathcal{N}(H_{uu})$ and by hypothesis \mathcal{H} ,

$$(2.7) \quad (H_{xu} + Pf_u) \eta = 0.$$

It follows from (2.5) and (2.7) that the first of (2.4) is equivalent to (1.8). Thus, if \mathcal{C} has an abnormal subarc on $[t', t'']$, then there exists on this interval a solution (ξ, ζ) of (1.6), (1.8), (1.9) for which $\xi \equiv 0$, $\zeta \neq 0$.

Conversely, suppose that \mathcal{C} is normal and that for some admissible control variation η there exists a solution (ξ, ζ) of (1.6), (1.8), (1.9) for which $\xi \equiv 0$, $\zeta \neq 0$ on some nondegenerate subinterval $[t', t'']$ of $[t_0, t_f]$. Then (2.5) holds and

$$(2.8) \quad \dot{\zeta} = -(H_{xu} + Pf_u) \eta - f_x^T \zeta.$$

From (1.9) it follows that

$$(2.9) \quad \hat{\eta} = -H^+ f_u^T \zeta.$$

If hypothesis \mathcal{H} is satisfied, then from (2.8) we obtain

$$(2.10) \quad \dot{\zeta} = -f_x^T \zeta - (H_{xu} + Pf_u) \hat{\eta}$$

and from (1.9) and (2.5) we obtain $\eta^T H_{uu} \eta = \eta^T f_u^T \zeta = 0$ and so $\hat{\eta} = 0$. Hence, from (2.10) it follows that $\dot{\zeta} = -f_x^T \zeta$. From (1.9) and (2.5) it follows that $f_u^T \zeta \in \mathcal{R}(H_{uu})$ and hence $f_u^T \zeta \in \mathcal{R}(H^+)$. Thus,

$$f_u^T \zeta = 0.$$

Since $\zeta \neq 0$ this implies that \mathcal{C} has an abnormal subarc, and this is a contradiction. We have proved the following theorem.

THEOREM 2.1. *If hypothesis \mathcal{H} holds, then \mathcal{C} is normal if and only if there does not exist a solution (ξ, ζ) of the Jacobi equations for some admissible variation η for which $\xi \equiv 0$, $\zeta \neq 0$ on a nondegenerate subinterval of $[t_0, t_f]$.*

The corollary below is a direct consequence of this theorem.

COROLLARY 2.1. *If hypothesis \mathcal{H} holds and if \mathcal{C} is a normal extremal arc, then there cannot be 2 solutions, $(\xi^{(1)}, \zeta^{(1)}, \eta^{(1)})$ and $(\xi^{(2)}, \zeta^{(2)}, \eta^{(2)})$ of the Jacobi equations with $\xi^{(1)} \equiv \xi^{(2)}$ and $\zeta^{(1)} \neq \zeta^{(2)}$ on any nondegenerate subinterval of $[t_0, t_f]$. Furthermore, if $\xi^{(1)} \equiv \xi^{(2)}$, then $\hat{\eta}^{(1)} \equiv \hat{\eta}^{(2)}$.*

3. The second variation: Preliminaries. By rearranging the terms in equation (1.5) and using (1.6), (1.8) and (1.9), we obtain

$$(3.1) \quad 2\omega = -\xi^T(\dot{\zeta} + f_x^T \zeta) - \eta^T f_u^T \zeta.$$

The following result is a direct consequence of equation (3.1).

Scolium. If \mathcal{C} has an abnormal subarc, then there exists a nonvanishing solution ξ, ζ of the Jacobi equations for which $2\omega \equiv 0$ on a subinterval of T .

By adding and subtracting the term $\xi^T \zeta$ and again making use of (1.6), (3.1) becomes

$$(3.2) \quad 2\omega = -\frac{d}{dt}(\xi^T \zeta).$$

Now suppose that ξ is an extremal variation for which $\xi(t_0) = \xi(t') = 0$, $\xi(t) \neq 0$ for $t_0 \leq t \leq t'$, and $\xi \equiv 0$ for $t' \leq t \leq t_f$. From (3.2) it follows that $J_2 = 0$. Thus, if \mathcal{C} is a normal minimizing arc and if there is an extremal variation ξ for which $\xi(t_0) = \xi(t') = 0$, $\xi \neq 0$ for $t_0 \leq t \leq t'$ and $\xi \equiv 0$ for $t' \leq t \leq t_f$, then such a variation minimizes J_2 .

The values t_0, t' are said to be conjugate on an extremal arc \mathcal{C} if there is an accessory extremal for \mathcal{C} with $\xi(t_0) = \xi(t') = 0$, but $\xi(t) \neq 0$ on $[t_0, t']$.

LEMMA 3.1. *If \mathcal{C} is a normal minimizing arc and if hypothesis \mathcal{H} holds on \mathcal{C} , then any extremal variation ξ which satisfies $\xi(t_0) = \xi(t') = 0$, $\xi(t) \neq 0$ for $t_0 \leq t \leq t'$, $\xi(t) \equiv 0$ on $[t', t_f]$ minimizes J_2 .*

4. The transformation of Clebsch. Here we shall obtain an alternate form for the second variation. Our methods are adopted from [5]. By premultiplying equation (1.9) by H^+ , we obtain

$$(4.1) \quad \hat{\eta} = -H^+(H_{ux}\xi + f_u^T \zeta).$$

Substitution of (4.1) into (1.6) and (1.8) yields

$$(4.2) \quad \dot{\xi} = A\xi - B\zeta + f_u \hat{\eta}$$

and

$$(4.3) \quad \dot{\zeta} = -C\xi - A^T \zeta - H_{xu} \hat{\eta},$$

where

$$A = f_x - f_u H^+ H_{ux}, \quad B = f_u H^+ f_u^T, \quad C = H_{xx} - H_{xu} H^+ H_{ux}.$$

If we add to the integrand of J_2 the term $d/dt(\xi^T W \zeta)$, where W is an arbitrary continuously differentiable symmetric matrix function of t , then the value of J_2 is unaffected. By substitution of (4.2) and (4.3) into (1.5) we obtain

$$(4.4) \quad 2\omega + \frac{d}{dt}(\xi^T W \zeta) = \xi^T (\dot{W} + WA + A^T W - WBW + C)\xi \\ + (\zeta - W\xi)^T B(\zeta - W\xi) + 2\xi^T (H_{xu} + Wf_u)\hat{\eta}.$$

Setting

$$(4.5) \quad W^* = \dot{W} + A^T W + WA - WBW + C$$

and $\sigma = \zeta - W\xi$ yields

$$(4.6) \quad J_2 = \frac{1}{2} \int_{t_0}^{t_f} [\xi^T W^* \xi + \sigma^T B \sigma + 2\xi^T (H_{xu} + Wf_u)\hat{\eta}] dt.$$

We remark that if $H_{uu} = H^+ = 0$, then $B = 0$ and $W^* = \dot{W} + f_x^T W + Wf_x + H_{xx}$.

If hypothesis \mathcal{H} holds, let us choose $W = P$. Then (4.6) becomes

$$(4.7) \quad J_2 = \frac{1}{2} \int_{t_0}^{t_f} (\xi^T P^* \xi + \sigma^T B \sigma) dt.$$

If we add $f_u^T P \xi$ to both sides of (1.9), we obtain

$$(4.8) \quad H_{uu} \eta + (H_{ux} + f_u^T P) \xi = -f_u^T \sigma.$$

Then hypothesis \mathcal{H} implies that

$$(4.9) \quad f_u^T \sigma \in \mathcal{R}(H_{uu}).$$

We have proved the following theorem.

THEOREM 4.1. *If \mathcal{C} is an extremal arc along which hypothesis \mathcal{H} holds and if ξ, ζ are extremal variations, then*

$$J_2 = \frac{1}{2} \int_{t_0}^{t_f} [\xi^T P^* \xi + \sigma^T f_u H^+ f_u^T \sigma] dt,$$

where $\sigma = \zeta - P\xi$. Furthermore, $f_u^T \sigma \in \mathcal{R}(H^+)$.

Theorem 4.2 is a direct consequence of Theorem 4.1.

THEOREM 4.2. *If along an extremal arc \mathcal{C} , H_{uu} is nonnegative definite, hypothesis \mathcal{H} holds and $\xi^T P^* \xi \geq 0$ for all extremal variations, then J_2 is nonnegative.*

LEMMA 4.1. *If along a minimizing arc \mathcal{C} hypothesis \mathcal{H} holds, if $\mathcal{N}(H_{uu})$ is not all of E^m , if either $\xi^T P^* \xi > 0$ for all nontrivial admissible variations or if $\xi^T P^* \xi \geq 0$ for all admissible variations, and there is no nontrivial solution ξ, ζ on T of the Jacobi equations for which $f_u^T \xi \equiv f_u^T P \xi$, then J_2 is positive definite.*

A direct consequence of Lemma 3.1, Lemma 4.1 and equation (4.9) is the following.

THEOREM 4.3. *If \mathcal{C} is a normal minimizing arc along which hypothesis \mathcal{H} holds, if $\mathcal{N}(H_{uu})$ is not all of E^m , if $\xi^T P^* \xi \geq 0$ for all admissible variations, and if there is no nontrivial solution ξ, ζ of the Jacobi equations for which $f_u^T \sigma \equiv 0$ on T , then there can be no value t' conjugate to t_0 on $(t_0, t_f]$.*

We shall now prove the following lemma.

LEMMA 4.2. *Let \mathcal{C} be a normal minimizing arc along which hypothesis \mathcal{H} holds. Let (ξ, ζ) be a solution of the Jacobi equations for which $P^* \xi \equiv 0$ on \mathcal{C} . Then $\sigma \equiv 0$.*

Proof. It follows from the Jacobi equations that

$$(4.10) \quad \dot{\xi} = (A - BP)\xi - B\sigma + f_u \tilde{\eta},$$

$$(4.11) \quad \dot{\sigma} = -P^* \xi - (A^T - PB)\sigma.$$

If $\text{rank } H_{uu}$ is zero, it follows from (4.9) that

$$(4.12) \quad f_u^T \sigma = 0.$$

If $P^* \xi = 0$, it follows from (4.11) and (4.12) that

$$\dot{\sigma} = -f_x^T \sigma.$$

Since \mathcal{C} is normal, σ vanishes identically. If $\text{rank } H_{uu}$ is positive and $P^* \xi = 0$, then

$$(4.13) \quad J_2 = \frac{1}{2} \int_{t_0}^{t_f} \sigma^T B \sigma dt,$$

and $J_2 = 0$ if and only if (4.12) holds. The rest of the argument is the same as above, and this proves the lemma.

LEMMA 4.3. Let \mathcal{C} be an extremal arc defined on T along which hypothesis \mathcal{H} holds. If $P^*\xi \equiv 0$ on \mathcal{C} for all extremal variations, then $\mathcal{N}(f_u) \supseteq \mathcal{N}(H_{uu})$. If furthermore, H_{uu} is nonnegative definite and $\xi(t_1) = \xi(t_2) = 0$ for $t_0 \leq t_1 < t_2 \leq t_f$, then $\xi(t) \equiv 0$ on $[t_1, t_2]$.

Proof. If $P^*\xi \equiv 0$ for all extremal variations, then

$$(4.14) \quad \dot{\sigma} = -(A^T - PB)\sigma.$$

Let ξ, ζ, η be a set of extremal variations and let $\Sigma(t, t_1)$ be that fundamental matrix solution of (4.14) which is the identity matrix at $t = t_1$. If $v(t)$ is any vector in $\mathcal{N}(H_{uu})$ on \mathcal{C} , it follows from (4.8) that

$$(4.15) \quad \Sigma^T(t, t_1)f_u v \equiv 0, \quad t \in T.$$

Since $\Sigma^T(t, t_1)$ is nonsingular, then $f_u v \equiv 0$ on \mathcal{C} .

We thus see that ξ satisfies the differential equation

$$(4.16) \quad \dot{\xi} = (A - BP)\xi - B\sigma.$$

From the variation of parameters formula we obtain

$$\Sigma^T(t, t_1)\xi(t) = \xi(t_1) - \int_{t_1}^t \Sigma^T(\tau, t_1)B\sigma \, d\tau.$$

If $\xi(t_1) = \xi(t_2) = 0$, it follows that

$$\int_{t_1}^{t_2} \Sigma^T(\tau, t_1)B\sigma \, d\tau = 0.$$

Since H_{uu} is nonnegative definite, then so is B . Hence

$$\Sigma^T B\sigma \equiv 0 \quad \text{on } [t_1, t_2],$$

and since $\Sigma^T(t, t_1)$ is nonsingular, $\xi(t)$ vanishes identically, and this proves the lemma.

Now if $P^*\xi \equiv 0$ along \mathcal{C} for all extremal variations and if hypothesis \mathcal{H} holds, then for any vector v in $\mathcal{N}(H_{uu})$ we have

$$H_{xu}v = -Pf_u v = 0.$$

Thus $\mathcal{N}(H_{xu}) \supseteq \mathcal{N}(H_{uu})$, and therefore equations (4.2) and (4.3) become

$$(4.17) \quad \dot{\xi} = A\xi - B\zeta,$$

$$(4.18) \quad \dot{\zeta} = -C\xi - A^T\zeta,$$

respectively.

An extremal arc \mathcal{C} is said to be regular if H_{uu} is nonsingular along it.

THEOREM 4.4. If along \mathcal{C} hypothesis \mathcal{H} holds and $P^* \equiv 0$, then every extremal variation ξ, ζ satisfies (4.17) and (4.18). If furthermore, the matrix (f_u^T, H_{ux}) has maximal rank, then \mathcal{C} is a regular extremal arc.

We need only establish the last part of the theorem. We have just seen that if $P^* \equiv 0$, then $f_u v = H_{xu}v = 0$ for all $v \in H_{uu}$. Hence, if (f_u^T, H_{ux}) has rank m , then $v = 0$ and the nullspace of H_{uu} contains only the zero vector.

LEMMA 4.4. Let \mathcal{C} be a minimizing arc along which $\text{rank } H_{uu} = 0$ and hypothesis \mathcal{H} holds. If $P^* \equiv 0$ and \mathcal{C} is normal, then all extremal variations satisfy $\zeta \equiv P\xi$.

Proof. If $\text{rank } H_{uu} = 0$ and hypothesis \mathcal{H} holds, then from equation (4.9) we obtain

$$(4.19) \quad f_u^T \sigma = 0,$$

and if $P^*\xi = 0$ it follows from (4.14) that

$$\dot{\sigma} = -f_x^T \sigma.$$

Thus, if \mathcal{C} is normal, then $\sigma \equiv 0$.

The following theorem is a direct consequence of Lemma 4.3.

THEOREM 4.5. Let \mathcal{C} be a minimizing arc along which hypothesis \mathcal{H} holds. If $P^*\xi \equiv 0$ on \mathcal{C} for all extremal variations there can be no value t' conjugate to t_0 on $(t_0, t_f]$.

In classical variational theory, extremal variations which satisfy the relation $\zeta = P\xi$ play an important role. The following theorem indicates that hypothesis \mathcal{H} may be a necessary condition in the singular theory.

THEOREM 4.6. Let \mathcal{C} be an extremal arc and let ξ, η, ζ be a set of extremal variations on \mathcal{C} . If $\zeta = P\xi$, then $P^*\xi = 0$ and $\tilde{\eta}$ is in the nullspace of $H_{xu} + Pf_u$.

Proof. If $\zeta = P\xi$ it follows from (4.11) that $P^*\xi = 0$. Differentiating the relation $\zeta = P\xi$ with respect to t , using (4.10) and (4.5) with $W = P$ and comparing the result with (4.11), we find that $(H_{xu} + Pf_u)\tilde{\eta} = 0$.

5. Concerning the Kelley condition: A heuristic approach. We shall now abandon hypothesis \mathcal{H} , but we shall require here that f_0 and f be three times continuously differentiable in \mathcal{S}_0 . Let η be the vector $h\phi_0$, where h is a constant vector in $\mathcal{N}(H_{uu})$ and $\phi_0(t, \tau)$ is the special variation introduced by Kelley [4] and illustrated in Fig. 1. The value \bar{t} is any one in the interior of T . We suppose

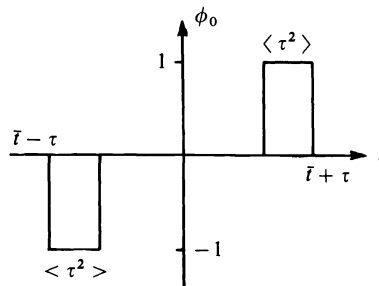


FIG. 1

here that $\mathcal{N}(H_{uu})$ is a fixed subspace of E^m on \mathcal{C} . Let $\phi_k(t, \tau)$ be defined by

$$\frac{d^k \phi_k}{dt^k} = \phi_0, \quad k = 1, 2.$$

Then as in [4] we can show that

$$(5.1) \quad \xi(t) = f_u(t)h\phi_1(t, \tau) + (f_x(t)f_u(t) - \dot{f}_u(t))h\phi_2(t, \tau) + \pi(t),$$

where $\pi(t) = O(\tau^3)$ as $\tau \rightarrow 0$. Substitution of (5.1) into (4.6) yields

$$2J_2 = \int_{t_0}^{t_f} (\xi^T W^* \xi + \sigma^T B \sigma) dt - \tau^5 h^T (Q(\bar{t}, W) + Q^T(\bar{t}, W)) h + o(\tau^5),$$

where

$$Q(t, W) = \frac{d}{dt} [(H_{ux} + f_u^T W) f_u] + 2(H_{ux} + f_u^T W)(f_x f_u - \dot{f}_u).$$

That control variations of the type described here yield admissible trajectories of (1.6) satisfying (1.7) requires an assumption of normality and an argument concerning tangent cones. A proof of this nature is not yet available. The heuristic argument presented here makes the following theorem seem likely.

THEOREM 5.1. *Let \mathcal{C} be a normal minimizing arc on the interval $[t_0, t_f]$. Then $Q(t, W) + Q^T(t, W)$ is nonpositive definite on the nullspace of H_{uu} .*

6. An example. Let $\dot{x}_1 = x_2 + u_1$, $\dot{x}_2 = -u_2$, $x_1(1) = x_2(1) = 0$, and minimize

$$J = \frac{1}{2} \int_0^1 [x_1^2 + (u_1 - u_2)^2] dt.$$

Here

$$H = \frac{1}{2}x_1^2 + \frac{1}{2}(u_1 - u_2)^2 + p_1(x_2 + u_1) - p_2u_2,$$

$$H_u = (u_1 - u_2 + p_1, u_2 - u_1 - p_2),$$

$$H_{uu} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

and

$$\dot{p}_1 = -x_1, \quad \dot{p}_2 = -p_1.$$

Extremals are composed of impulsive arcs along which $(x_1 + x_2)$ is constant and a singular arc which is an arc of an hyperbola, $x_1x_2 = \text{const.}$ as shown in Fig. 2.

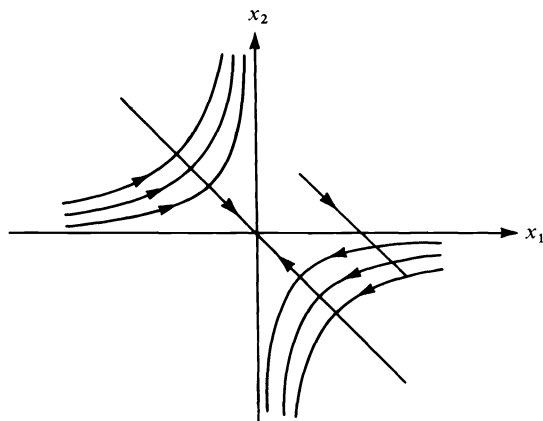


FIG. 2

Singular control is given by

$$u_1 = -x_1 - x_2, \quad u_2 = -x_2.$$

The nullspace of H_{uu} is the set of all 2-vectors v having equal components, and $\mathcal{N}(H_{xu} + Pf_u) \supseteq \mathcal{N}(H_{uu})$ if and only if

$$(6.1) \quad p_{11} = p_{12} = p_{22} = p.$$

For the matrix

$$P = \begin{pmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{pmatrix}$$

whose elements satisfy (6.1) we see that

$$(6.2) \quad P^* = \begin{pmatrix} \dot{p} - p^2 + 1 & \dot{p} - p^2 + p \\ \dot{p} - p^2 + p & \dot{p} - p^2 + 2p \end{pmatrix}.$$

It is not hard to see that there is no solution of $P^* = 0$. If $p = 0$, then

$$P^* = C = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

which is nonnegative definite.

The second variation is given by

$$J_2 = \frac{1}{2} \int_0^1 [\xi_1^2 + (\zeta_1 + \zeta_2)^2] dt.$$

Note that $\sigma = \zeta$ and that $(\zeta_1 + \zeta_2)^2 = 0$ if and only if $\zeta_1 = -\zeta_2$, i.e., if and only if ζ is in the nullspace of B . Furthermore, if $\zeta_1 + \zeta_2 = 0$, then the only solution of the accessory equations for which $\xi(1) = 0$ is the trivial one.

7. Conclusion. In this paper, we have begun the development of a theory of singular extremals by means of hypothesis \mathcal{H} and the generalized inverse of a matrix. It is seen that all our theorems are direct analogs of classical theorems. It is the belief of this author that every classical theorem concerning a Lagrange or Bolza problem has its analog in the singular theory, and in particular that it can be shown by our methods that the nonexistence of a conjugate point on a normal extremal arc \mathcal{C} along which H_{uu} is nonnegative definite implies the positive definiteness of the second variation. In classical regular theory the existence of solutions of the matrix Riccati equation on an interval is equivalent to the nonexistence of conjugate points. The analog of this theorem in the singular theory should be that the nonexistence of conjugate points on an interval is equivalent to the existence on this interval of a matrix P satisfying hypothesis \mathcal{H} and for which the matrix Riccati expression P^* is nonnegative definite.

REFERENCES

- [1] A. ALBERT, *Regression and the Moore-Penrose Inverse*, Academic Press, New York, 1972.
- [2] L. D. BERKOVITZ, *Variational methods in problems of control and programming*, J. Math. Anal. Appl., 3 (1961), pp. 145-169.

- [3] M. R. HESTENES, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.
- [4] H. J. KELLEY, R. E. KOPP AND H. G. MOYER, *Singular extremals*, Topics in Optimization, G. Leitmann, ed., Academic Press, New York, 1967, pp. 63–101.
- [5] J. RADON, *Zum Problem von Lagrange*, Abh. Math. Sem. Univ. Hamburg, 6 (1928), pp. 237–246.

CONSISTENCY OF LEAST-SQUARES ESTIMATES USED IN LINEAR SYSTEMS IDENTIFICATION*

D. O. NORRIS AND L. E. SNYDER†

Abstract. Least-squares estimation of the parameters of a single input-single output linear autonomous system is considered where both plant noise and observation noise are present. It is shown that under fairly general conditions that the estimates converge almost surely to the true system parameters.

1. Introduction. The problem of estimating the parameters of a linear dynamical system has received considerable attention over the past few years. Early work on the problem was done by Mann and Wald [8]. They treated linear stochastic difference equations which are models for autoregressive processes. They assumed the noise for the process consisted of a sequence of independent and identically distributed random variables. They derived an estimator for the parameters by assuming the noise to be Gaussian white noise and showed that this estimator was consistent and asymptotically normal even if the Gaussian assumption is not valid. It is worth noting that the estimator is a linear least-squares estimator. More generally, one is interested in the situation where there are system inputs, observation noise and plant noise. It is common practice to assume that the noise processes are independent Gaussian processes. Under these conditions the method of estimation is usually taken to be maximum likelihood. Problems of this kind have been treated by Aoki and Yue [1], Astrom, Bohlin and Wensmark [2], Levin [3] and Kashyap [4]. None of these estimators is a linear least-squares estimator and therefore determination of the maximum likelihood estimates must be accomplished iteratively. It is the authors' experience that the increase in accuracy achieved by using maximum likelihood estimates rather than linear least-squares estimates (such as appear in [9]) more than compensate for the increase in computational difficulty.

It is well known that the maximum likelihood estimator is a least-squares estimator when the only noise present is observation noise and this noise is an independent, Gaussian process. However, when Gaussian plant noise is also present the maximum likelihood estimator is no longer a least-squares estimator.

Suppose, however, it is not known that the noise processes are independent and Gaussian. In fact, suppose the likelihood function is unknown. Then how should the identification be accomplished? A natural choice, it seems, is to use a least-squares identification procedure which minimizes the modeling error as it will be a maximum likelihood estimator in certain cases. Furthermore, the computational experience of the authors' is that the procedure is superior to linear least-squares estimators. In this paper, the convergence of a least-squares estimator of the system parameters will be considered. It will be shown that under fairly mild conditions the estimator converges almost surely to the true system parameters. The stochastic assumptions made are that the plant noise

* Received by the editors October 23, 1973, and in revised form October 24, 1974.

† Department of Mathematics, Ohio University, Athens, Ohio 45701. This work was sponsored by the United States Air Force under Contract F 33615-73-Q-4009.

and observation noise are independent sequences of independent and identically distributed random variables. If plant noise is present, it is also necessary to assume that the plant noise is almost surely bounded. This is a fairly strong assumption but could not be avoided by the authors.

2. Notation, problem formulation and assumptions. In this paper, systems of the following form

$$(1) \quad \begin{aligned} x_k + a_1^0 x_{k-1} + \cdots + a_n^0 x_{k-n} &= b_1^0 u_{k-1} + \cdots + b_n^0 u_{k-n} \\ &+ d_1^0 \xi_{k-1} + \cdots + d_n^0 \xi_{k-n}, \\ y_k &= x_k + \eta_k, \end{aligned} \quad k = 0, 1, 2, \dots,$$

will be treated, where it is assumed that quantities with negative subscripts are zero. $a_1^0, \dots, a_n^0, b_1^0, \dots, b_n^0$ are the parameters to be identified and for notational convenience will be denoted by θ^0 . $\{\xi_k\}$ and $\{\eta_k\}$ denote random processes. Following the notation of [1], let $\mathbf{x}_N = (x_0, \dots, x_{N-1})^T$ and let

$$S_N = \begin{bmatrix} 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \\ \vdots & & \ddots & \ddots \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix}$$

denote the $N \times N$ right shift matrix with entries $s_{ij} = \delta_{i,j+1}$. Note that S_N^k , $k = 1, 2, \dots, N-1$, is a right shift matrix with entries $s_{ij}^{(k)} = \delta_{i,j+k}$. For convenience let $S_N^0 = I_N$, the $N \times N$ identity matrix. It is a simple matter to verify that

$$(2) \quad \begin{aligned} A_N(\theta^0) \mathbf{x}_N(\omega) &= B_N(\theta^0) \mathbf{u}_N + D_N(\theta^0) \boldsymbol{\xi}_N(\omega), \\ \mathbf{y}_N(\omega) &= \mathbf{x}_N(\omega) + \boldsymbol{\eta}_N(\omega), \end{aligned}$$

where

$$\begin{aligned} A_N(\theta^0) &= \sum_{j=0}^n a_j^0 S_N^j, & a_0^0 &= 1, \\ B_N(\theta^0) &= \sum_{j=1}^n b_j^0 S_N^j, \\ D_N(\theta^0) &= \sum_{j=1}^n d_j^0 S_N^j, \end{aligned}$$

and the variable ω is introduced to emphasize that we are dealing with stochastic processes, i.e., for each $\omega \in \Omega$, Ω a probability space, we have sample functions $\{\xi_k(\omega)\}$ and $\{\eta_k(\omega)\}$ which in turn cause the state variables and output variables to be sample functions of stochastic processes.

In the presentation given in [1], the equivalent dynamical system formulation of (1) is also presented. The authors also treat the problem of estimating the

initial state vector. However, we shall assume that the initial state vector is the zero vector because in order to guarantee convergence of the estimates of the system parameters it is necessary to assume that the system is asymptotically stable, and, consequently, the effect of nonzero initial conditions is negligible, at least for large samples.

Given the input vector \mathbf{u}_N and the output vector $\mathbf{y}_N(\omega)$, we want to choose $\hat{\boldsymbol{\theta}}_N(\omega) = (a_1, \dots, a_n, b_1, \dots, b_n)^T$ so that

$$(3) \quad F(N, \boldsymbol{\theta}, \omega) = \frac{1}{N} \|\mathbf{y}_N(\omega) - A_N(\boldsymbol{\theta})^{-1} B_N(\boldsymbol{\theta}) \mathbf{u}_N\|^2$$

is minimized; i.e., we want to obtain a least-squares estimate for the true system parameters $\boldsymbol{\theta}^0 = (a_n^0, \dots, b_n^0)^T$. The major result of this paper is that under suitable conditions $\lim_{N \rightarrow \infty} \hat{\boldsymbol{\theta}}_N(\cdot) = \boldsymbol{\theta}^0$ almost surely (a.s.).

Note that $\mathbf{x}_N(\omega) = A_N(\boldsymbol{\theta}^0)^{-1} B_N(\boldsymbol{\theta}^0) \mathbf{u}_N + A_N(\boldsymbol{\theta}^0)^{-1} D_N(\boldsymbol{\theta}^0) \boldsymbol{\xi}_N(\omega)$ and therefore,

$$(4) \quad \begin{aligned} F(N, \boldsymbol{\theta}, \omega) = \frac{1}{N} \|(A_N(\boldsymbol{\theta}^0)^{-1} B_N(\boldsymbol{\theta}^0) - A_N(\boldsymbol{\theta})^{-1} B_N(\boldsymbol{\theta})) \mathbf{u}_N \\ + A_N^{-1}(\boldsymbol{\theta}^0) D_N(\boldsymbol{\theta}^0) \boldsymbol{\xi}_N(\omega) + \boldsymbol{\eta}_N(\omega)\|^2. \end{aligned}$$

For notational convenience let $Q_N(\boldsymbol{\theta}) = A_N(\boldsymbol{\theta}^0)^{-1} B_N(\boldsymbol{\theta}^0) - A_N(\boldsymbol{\theta})^{-1} B_N(\boldsymbol{\theta})$ and let $P_N = A_N(\boldsymbol{\theta}^0)^{-1} D_N(\boldsymbol{\theta}^0)$. Then

$$(5) \quad \begin{aligned} F(N, \boldsymbol{\theta}, \omega) &= \frac{1}{N} \|Q_N(\boldsymbol{\theta}) \mathbf{u}_N + P_N \boldsymbol{\xi}_N(\omega) + \boldsymbol{\eta}_N(\omega)\|^2 \\ &= \frac{1}{N} \|Q_N(\boldsymbol{\theta}) \mathbf{u}_N\|^2 + \frac{1}{N} \|P_N \boldsymbol{\xi}_N(\omega)\|^2 + \frac{1}{N} \|\boldsymbol{\eta}_N(\omega)\|^2 \\ &\quad + \frac{2}{N} \langle Q_N(\boldsymbol{\theta}) \mathbf{u}_N, P_N \boldsymbol{\xi}_N(\omega) \rangle + \frac{2}{N} \langle Q_N(\boldsymbol{\theta}) \mathbf{u}_N, \boldsymbol{\eta}_N(\omega) \rangle \\ &\quad + \frac{2}{N} \langle P_N \boldsymbol{\xi}_N(\omega), \boldsymbol{\eta}_N(\omega) \rangle, \end{aligned}$$

where we are using the inner product notation $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}^T \mathbf{b}$.

The specific assumptions used to prove that $\lim_{N \rightarrow \infty} \hat{\boldsymbol{\theta}}_N(\cdot) = \boldsymbol{\theta}^0$, almost surely, will now be discussed. These assumptions will be presented in two groups—the system assumptions and the stochastic assumptions.

System assumptions

Assumption S1. $A(z, \boldsymbol{\theta}) = 1 + \sum_{j=1}^n a_j z^j$ has no zeros in $|z| \leq 1$.

Assumption S2. $B(z, \boldsymbol{\theta}) = \sum_{j=1}^n b_j z^j$ has no zeros in common with $A(z, \boldsymbol{\theta})$ and not all $b_j = 0$.

Assumption S3. The set Θ of admissible parameters $\boldsymbol{\theta}$ is a compact subset of R^{2n} which contains the true system parameter $\boldsymbol{\theta}^0$ in its interior and whose elements satisfy Assumptions S1 and S2.

Assumption S4. The sequence of inputs $\{u_j\}$ is bounded and $\lim_{N \rightarrow \infty} (1/N) \cdot \sum_{n=\min\{i,j\}}^{N-1} u_{n-i} u_{n-j} = \tilde{u}(i, j)$ exists for every integer $i \geq 0$ and $j \geq 0$.

Assumption S5. $\lim_{N \rightarrow \infty} (1/N) U_{N,2n}^T U_{N,2n}$ is positive definite, where $U_{N,2n} = (S_N \mathbf{u}_N, \dots, S_N^{2n} \mathbf{u}_N)$.

We now consider the need for these assumptions and make some specific observations concerning them. Assumption S1 insures that the system represented by (1) is asymptotically stable. Without this assumption, the sequence $\{\hat{\theta}_N(\omega)\}$ of estimates need not converge.

S_N is a linear operator on R^N and its spectrum $\sigma(S_N) = \{0\}$. $A(z, \theta)^{-1}$ is analytic on $|z| \leq 1$ for each fixed θ satisfying Assumption S1 as the zeros of $A(z, \theta)$ are in $|z| > 1$. Therefore, $A(z, \theta)^{-1} = \sum_{j=0}^{\infty} g_j(\theta) z^j$ is valid on $|z| \leq 1$ from which it follows that $\sum_{j=0}^{\infty} |g_j(\theta)| < \infty$. In addition, from spectral theory it can be concluded that $A(S_N, \theta)^{-1} = A_N(\theta)^{-1} = \sum_{j=0}^{\infty} g_j(\theta) S_N^j$. But, S_N is nilpotent of order N , hence $A_N(\theta)^{-1} = \sum_{j=0}^{N-1} g_j(\theta) S_N^j$. From this it easily follows that

$$Q_N(\theta) = \sum_{j=0}^{N-1} q_j(\theta) S_N^j, \quad \text{where } \sum_{j=0}^{\infty} |q_j(\theta)| < \infty$$

and

$$P_N = \sum_{j=0}^{N-1} p_j S_N^j, \quad \text{where } \sum_{j=0}^{\infty} |p_j| < \infty.$$

Assumption S2 is needed for complete controllability of the system represented by (1). If the system to be identified is not completely controllable, then θ^0 is not unique in the sense that $\{\theta: J(\theta) = J(\theta^0)\}$ is not a singleton, where $J(\theta) = \lim_{N \rightarrow \infty} E[F(N, \theta, \cdot)]$. The uniqueness is a vital ingredient of the proof of the almost sure convergence of the least-squares estimates.

Assumption S3 is needed to insure that the sequence $\{\hat{\theta}_N(\omega)\}$ of least-squares estimates has a convergent subsequence.

Assumption S4 is needed to insure the almost sure convergence of $F(N, \theta, \cdot)$. In particular, if the input sequence does not satisfy the Cesaro summability requirement, then $\lim_{N \rightarrow \infty} (1/N) \|Q_N(\theta) \mathbf{u}_N\|^2$ will not exist in all cases. It should also be observed that the Cesaro summability condition is equivalent to the requirement that $\lim_{N \rightarrow \infty} (1/N) \langle S_N^i \mathbf{u}_N, S_N^j \mathbf{u}_N \rangle$ exist for all integers $i \geq 0$ and $j \geq 0$.

Assumption S5 together with Assumption S2 insure that θ^0 is unique in the sense described above (see [1, Thm. 1]). In [2], inputs which satisfy Assumption S4 and S5 are called "persistently exciting."

Stochastic assumptions.

Assumption R1. $\{\xi_j\}$ and $\{\eta_j\}$ are independent random processes of independent and identically distributed random variables such that $E[\xi_n] = E[\eta_n] = 0$ and $E[\xi_n^2] < \infty$ and $E[\eta_n^2] < \infty$ for all integers $n \geq 0$.

Assumption R2. $\{\xi_n\}$ is almost surely bounded.

Assumption R1 is a stronger assumption than is actually needed. However, it is a sufficient condition to guarantee that a strong law of large numbers holds for a variety of sequences which will be encountered in the sequel. Rather than assert, as an assumption, that each of these sequences satisfies a strong law of large numbers it is more economical to use Assumption R1.

For example, we need to know that $\lim_{N \rightarrow \infty} (1/N) \sum_{k=0}^{N-1} \xi_{k+j} \xi_k = \text{var}[\xi_0] \delta_{0j}$ a.s. for every $j = 0, 1, 2, \dots$. The sequence $\{\xi_{k+j} \xi_k\}$ is a sequence of identically

distributed random variables which are j -dependent; i.e., $\xi_{k+j}\xi_k$ and $\xi_{n+j}\xi_n$ are independent if $n > j + k$. Therefore, the sequence $\{\xi_{k+j}\xi_k\}$ is $*$ -mixing and the desired strong law of large numbers holds (see definition of $*$ -mixing and corollary to Theorem 3 in [5]). The same argument applies to the sequence $\{\eta_{k+j}\eta_k\}$.

The sequences $\{u_{n+j}\eta_n\}$, $\{\xi_{n+j}\eta_n\}$ and $\{v_{n+j}(\theta)\xi_n : v_n(\theta) = \sum_{j=0}^n q_j(\theta)u_{n-j}\}$ are all sequences of independent random variables, hence are $*$ -mixing. It is easy to show that the variances of the elements of each sequence are uniformly bounded (for each θ and each $j \geq 0$) and that the random variables in each sequence are uniformly integrable (follows from uniform boundedness of sequences). Therefore, we can conclude (see [5, Thm. 2]) that

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} u_{n+j}\eta_n = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} \xi_{n+j}\eta_n = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} v_{n+j}(\theta)\xi_n = 0.$$

Assumption R2 is a severe restriction to make. For example, Gaussian processes are not a.s. bounded. However, if it is known that only observation noise is present, then, of course, this assumption is unnecessary.

3. Convergence of the estimates. The two main theorems of this section are concerned with (a) the a.s. convergence of the least-squares cost function $F(N, \theta, \cdot)$ and (b) the a.s. convergence of the least-squares estimator $\hat{\theta}_N(\cdot)$ to θ^0 . The proof of these results is quite complex and consequently the two theorems are preceded by three lemmas. The lemmas are used to establish the a.s. convergence of the terms which appear in the expansion of $F(N, \theta, \omega)$ given in (5). For a fixed $\omega \in \Omega$ and fixed $\theta \in \Theta$ the two sequences $\{(1/N)\|Q_N(\theta)u_N\|^2\}$ and $\{(1/N)\|P_N\xi_N(\omega)\|^2\}$ are of the same type. Rather than establish the a.s. convergence of each sequence we shall prove that $\lim_{N \rightarrow \infty} (1/N)\|Q_N(\theta)\xi_N\|^2$ exists almost surely. This result will cover the convergence of both sequences. In addition, the sequences $\{(1/N)\langle Q_N(\theta)u_N, P_N\xi_N(\omega) \rangle\}$, $\{(1/N)\langle Q_N(\theta)u_N, \eta_N(\omega) \rangle\}$ and $\{(1/N)\langle P_N\xi_N(\omega), \eta_N(\omega) \rangle\}$ are all of a similar type for fixed $\omega \in \Omega$ and $\theta \in \Theta$. Each of these sequences has an a.s. limit of zero and the arguments required to show this are essentially the same. Therefore, we will only prove that $\lim_{N \rightarrow \infty} (1/N)\langle Q_N(\theta)u_N, P_N\xi_N \rangle = 0$ almost surely.

LEMMA 1. *For any θ^* satisfying Assumption S1 and $\varepsilon > 0$ there is an integer N_0 and $\delta > 0$ such that*

$$(i) \quad \sum_{j=0}^{\infty} |q_j(\theta)| \leq \sum_{j=0}^{\infty} |q_j(\theta^*)| + 1$$

and

$$(ii) \quad \sum_{j=N_0+1}^{\infty} |q_j(\theta)| \leq \varepsilon$$

for all θ such that $\|\theta - \theta^*\| < \delta$.

Proof. Let $p^* = \min \{|\alpha_1^*|, \dots, |\alpha_n^*|\}$, where $\alpha_1^*, \dots, \alpha_n^*$ are the zeros of $A(\cdot, \theta^*)$. The roots of a polynomial are continuous functions of the coefficients (see [6, Thm. 1-6, p. 3] or [7, Thm. 5.1.4, p. 162]); therefore it is possible to choose δ_1 so that the roots of $A(\cdot, \theta)$ have modulus greater than $(p^* + 1)/2$ for any θ such that $\|\theta - \theta^*\| < \delta_1$.

If a function f is analytic on the disc $\{z: |z| < (p^* + 1)/2\}$, then f has a power series representation on the disc, say, $f(z) = \sum_{j=0}^{\infty} a_j z^j$. $\|f\| = \sum_{j=0}^{\infty} |a_j|$ defines a norm on all such functions which is continuous relative to the topology of uniform convergence on compact sets; i.e., if $\lim_{n \rightarrow \infty} f_n = f$ uniformly on every compact subset of $\{z: |z| < (p^* + 1)/2\}$, then $\lim_{n \rightarrow \infty} \|f_n\| = \|f\|$. Recall that $Q_N(\theta) = \sum_{j=0}^{N-1} q_j(\theta) S_N^j$, where $q_j(\theta)$ are the coefficients in the expansion of $B(z, \theta)/A(z, \theta) - B(z, \theta^0)/A(z, \theta^0)$. Let $Q(z, \theta) = \sum_{j=0}^{\infty} q_j(\theta) z^j$. Then there is a $\delta_2 \leq \delta_1$ such that $\|Q(\cdot, \theta)\| \leq \|Q(\cdot, \theta^*)\| + 1$ for all θ such that $\|\theta - \theta^*\| < \delta_2$, i.e., (i) holds.

Next choose N_0 such that $\sum_{j=N_0+1}^{\infty} |q_j(\theta^*)| < \varepsilon/2$. The functional $\|f\|_{N_0} = \sum_{j=N_0+1}^{\infty} |a_j|$, where $f(z) = \sum_{j=0}^{\infty} a_j z^j$ is a seminorm which is continuous relative to the topology of uniform convergence on compact sets. Therefore, there is a positive number $\delta \leq \delta_2$ such that $\|Q(\cdot, \theta)\|_{N_0} < \varepsilon$ for all θ such that $\|\theta - \theta^*\| < \delta$, i.e., (ii) holds. Q.E.D.

LEMMA 2. Let θ^* be given and suppose Assumptions S1, R1 and R2 hold. There is an almost sure event B such that for each $\omega \in B$ and for each $\varepsilon > 0$ there is a $\delta(\omega, \varepsilon) > 0$ and an integer $N(\omega, \varepsilon)$ such that

$$\left| \frac{1}{N} \|Q_N(\theta) \xi_N(\omega)\|^2 - \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} q_i(\theta) q_j(\theta) \gamma(i, j) \right| < \varepsilon$$

for every θ satisfying $\|\theta - \theta^*\| < \delta(\omega, \varepsilon)$ and $N \geq N(\omega, \varepsilon)$, where $\gamma(i, j) = E[\xi_i \xi_j]$.

(The double sum, although unnecessary, is used here because of analogues of this lemma which will be used later.)

Proof. Let $\varepsilon > 0$ be given such that $0 < \varepsilon < 1$. From Assumption R1 it follows that $\lim_{N \rightarrow \infty} (1/N) \sum_{n=m}^{N-1} \xi_{n-i}(\omega) \xi_{n-j}(\omega) = \gamma(i, j)$ exists almost surely for any fixed $m \geq \min\{i, j\}$. Let $B \subseteq \Omega$ denote the event for which this limit exists and for which $\{\xi_n(\omega)\}$ is bounded. By hypothesis, $P(B) = 1$. Fix $\omega \in B$. Let $T(\omega) = \sup_n \{|\xi_n(\omega)|\}$. By Lemma 1 there is an integer N' and $\delta(\omega, \varepsilon) > 0$ such that $\sum_{j=N'+1}^{\infty} |q_j(\theta)| < \varepsilon/T(\omega)$ for all θ such that $\|\theta - \theta^*\| < \delta(\omega, \varepsilon)$. From this and the fact that $\gamma(i, j)$ is bounded it follows that the double series $\sum_{i=0}^{\infty} \sum_{j=0}^{\infty} q_i(\theta) \cdot q_j(\theta) \gamma(i, j)$ is absolutely convergent. Hence, there is an integer N^* such that

$$\left| \sum_{i=0}^N \sum_{j=0}^N q_i(\theta) q_j(\theta) \gamma(i, j) - \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} q_i(\theta) q_j(\theta) \gamma(i, j) \right| < \varepsilon$$

for $N > N^* + 1$ and all θ such that $\|\theta - \theta^*\| < \delta(\omega, \varepsilon)$. Then, for $N > N^* + 1$,

$$\begin{aligned} \frac{1}{N} \|Q_N(\theta) \xi_N(\omega)\|^2 &= \frac{1}{N} \sum_{n=0}^{N-1} \left(\sum_{i=0}^n q_i(\theta) \xi_{n-i}(\omega) \right)^2 \\ (6) \qquad &= \frac{1}{N} \sum_{n=0}^{N^*} \left(\sum_{i=0}^n q_i(\theta) \xi_{n-i}(\omega) \right)^2 \\ &\quad + \frac{1}{N} \sum_{n=N^*+1}^{N-1} \left(\sum_{i=0}^n q_i(\theta) \xi_{n-i}(\omega) \right)^2. \end{aligned}$$

N^* is fixed, so for N sufficiently large the absolute value of the first term on the right side of (6) can be made less than ε for all θ such that $\|\theta - \theta^*\| < \delta(\omega, \varepsilon)$.

Therefore, we restrict our attention to the second term on the right side of (6).

$$\begin{aligned}
 & \frac{1}{N} \sum_{n=N^*+1}^{N-1} \left(\sum_{i=0}^{N^*} q_i(\boldsymbol{\theta}) \zeta_{n-i}(\omega) + \sum_{i=N^*+1}^n q_i(\boldsymbol{\theta}) \zeta_{n-i}(\omega) \right)^2 \\
 &= \frac{1}{N} \sum_{n=N^*+1}^{N-1} \left(\sum_{i=0}^{N^*} q_i(\boldsymbol{\theta}) \zeta_{n-i}(\omega) \right)^2 \\
 &+ \frac{2}{N} \sum_{n=N^*+1}^{N-1} \left[\left(\sum_{i=0}^{N^*} q_i(\boldsymbol{\theta}) \zeta_{n-i}(\omega) \right) \left(\sum_{j=N^*+1}^n q_j(\boldsymbol{\theta}) \zeta_{n-j}(\omega) \right) \right] \\
 &+ \frac{1}{N} \sum_{n=N^*+1}^{N-1} \left(\sum_{j=N^*+1}^n q_j(\boldsymbol{\theta}) \zeta_{n-j}(\omega) \right)^2.
 \end{aligned}
 \tag{7}$$

From Lemma 1 and the fact that $\omega \in B$ it follows that the absolute value of the third term on the right side of (7) is smaller than

$$\frac{1}{N} \sum_{n=N^*+1}^{N-1} \left(T(\omega) \sum_{j=N^*+1}^n |q_j(\boldsymbol{\theta})| \right)^2 < \frac{1}{N} \sum_{n=N^*+1}^{N-1} \varepsilon^2 < \varepsilon^2 < \varepsilon$$

for all $\boldsymbol{\theta}$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| < \delta(\omega, \varepsilon)$. The absolute value of the second term on the right side of (7) is smaller than

$$\begin{aligned}
 & \frac{2}{N} \sum_{n=N^*+1}^{N-1} \left(\left| \sum_{i=0}^{N^*} q_i(\boldsymbol{\theta}) \zeta_{n-i}(\omega) \right| \cdot T(\omega) \cdot \sum_{j=N^*+1}^n |q_j(\boldsymbol{\theta})| \right) \\
 & \leq \frac{2\varepsilon}{N} \sum_{n=N^*+1}^{N-1} \sum_{i=0}^{N^*} |q_i(\boldsymbol{\theta})| |\zeta_{n-i}(\omega)| \\
 & \leq \frac{2\varepsilon}{N} \cdot T(\omega) \cdot N \cdot \sum_{i=0}^{N^*} |q_i(\boldsymbol{\theta})| \\
 & \leq 2\varepsilon \cdot T(\omega) \cdot \left(\sum_{i=0}^{\infty} |q_i(\boldsymbol{\theta}^*)| + 1 \right) = K\varepsilon
 \end{aligned}$$

for all θ such that $\|\theta - \theta^*\| < \delta(\omega, \varepsilon)$. Finally, from the first term on the right side of (7) we have

$$\begin{aligned}
 & \frac{1}{N} \sum_{n=N^*+1}^{N-1} \left(\sum_{i=0}^{N^*} q_i(\boldsymbol{\theta}) \zeta_{n-i}(\omega) \right)^2 \\
 &= \frac{1}{N} \sum_{n=N^*+1}^{N-1} \sum_{i=0}^{N^*} \sum_{j=0}^{N^*} q_i(\boldsymbol{\theta}) q_j(\boldsymbol{\theta}) \zeta_{n-i}(\omega) \zeta_{n-j}(\omega) \\
 &= \sum_{i=0}^{N^*} \sum_{j=0}^{N^*} q_i(\boldsymbol{\theta}) q_j(\boldsymbol{\theta}) \left[\frac{1}{N} \sum_{n=N^*+1}^{N-1} \zeta_{n-i}(\omega) \zeta_{n-j}(\omega) \right].
 \end{aligned}
 \tag{8}$$

N^* is fixed, so for N sufficiently large,

$$\begin{aligned}
 & \sum_{i=0}^{N^*} \sum_{j=0}^{N^*} \left| q_i(\boldsymbol{\theta}) q_j(\boldsymbol{\theta}) \left[\frac{1}{N} \sum_{n=N^*+1}^{N-1} \zeta_{n-i}(\omega) \zeta_{n-j}(\omega) \right] \right. \\
 & \quad \left. - \sum_{i=0}^{N^*} \sum_{j=0}^{N^*} q_i(\boldsymbol{\theta}) q_j(\boldsymbol{\theta}) \gamma(i, j) \right| < \varepsilon
 \end{aligned}$$

for all θ such that $\|\theta - \theta^*\| < \delta(\omega, \varepsilon)$. Therefore, for N sufficiently large,

$$\left| \frac{1}{N} \|Q_N(\theta)\xi_N(\omega)\|^2 - \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} q_i(\theta)q_j(\theta)\gamma(i, j) \right| < (K + 4)\varepsilon. \quad \text{Q.E.D.}$$

LEMMA 3. Let θ^* be given and suppose Assumptions S1, S4, R1 and R2 hold. Then there is an almost sure event B such that for each $\omega \in B$ and for each $\varepsilon > 0$ there is a $\delta(\omega, \varepsilon) > 0$ and an integer $N(\omega, \varepsilon)$ such that $|\langle (1/N)Q_N(\theta)\mathbf{u}_N(\omega), P_N\xi_N(\omega) \rangle| < \varepsilon$ for every θ satisfying $\|\theta - \theta^*\| < \delta(\omega, \varepsilon)$ and $N \geq N(\omega, \varepsilon)$.

Proof. As a consequence of Assumption R1 there is an event B such that $P(B) = 1$ and $\lim_{N \rightarrow \infty} (1/N) \sum_{n=j}^{N-1} v_n(\theta^*)\xi_{n-j}(\omega) = 0$ for all $\omega \in B$ and all $j = 0, 1, 2, \dots$. Fix $\omega \in B$ such that Assumption R2 holds. Let $\varepsilon > 0$ be given. By Lemma 1 there is an integer N^* and $\delta(\omega, \varepsilon) > 0$ such that $\sum_{j=N^*+1}^{\infty} |q_j(\theta)| < \max(\varepsilon/VT(\omega), \varepsilon/UT(\omega), \varepsilon)$ for all θ such that $\|\theta - \theta^*\| < \delta(\omega, \varepsilon)$ where $V = \sup_n \sup_{\|\theta - \theta^*\| < \delta} \{ |v_n(\theta)| \} = |\sum_{i=j}^n q_i(\theta)u_{n-i}|$, $U = \sup_n \{ |u_n| \}$ and $T(\omega) = \sup_n \{ |\xi_n(\omega)| \}$. Furthermore, N^* can be further restricted so that $\sum_{j=N^*+1}^{\infty} |p_j| < \max(\varepsilon/VT(\omega), \varepsilon/UT(\omega))$. For $N > N^* + 1$ we have

$$\begin{aligned} & \frac{1}{N} \langle Q_N(\theta)\mathbf{u}_N, P_N\xi_N(\omega) \rangle \\ &= \frac{1}{N} \sum_{n=0}^{N-1} v_n(\theta) \sum_{j=0}^n p_j \xi_{n-j}(\omega) \\ (9) \quad &= \frac{1}{N} \sum_{j=0}^{N-1} p_j \sum_{n=j}^{N-1} v_n(\theta) \xi_{n-j}(\omega) \\ &= \sum_{j=0}^{N^*} p_j \left[\frac{1}{N} \sum_{n=j}^{N-1} v_n(\theta) \xi_{n-j}(\omega) \right] + \frac{1}{N} \sum_{j=N^*+1}^{N-1} p_j \sum_{n=j}^{N-1} v_n(\theta) \xi_{n-j}(\omega). \end{aligned}$$

Consider the first term on the right side of (9). We claim $\lim_{N \rightarrow \infty} (1/N) \sum_{n=j}^{N-1} v_n(\theta) \cdot \xi_{n-j}(\omega) = 0$ for $j = 0, 1, \dots, N^*$ and for all θ such that $\|\theta - \theta^*\| < \delta(\omega, \varepsilon)$. To see this, note that

$$\begin{aligned} (10) \quad \left| \frac{1}{N} \sum_{n=j}^{N-1} v_n(\theta) \xi_{n-j}(\omega) \right| &= \left| \frac{1}{N} \sum_{n=j}^{N-1} [v_n(\theta) - v_n(\theta^*)] \xi_{n-j}(\omega) + \frac{1}{N} \sum_{n=j}^{N-1} v_n(\theta^*) \xi_{n-j}(\omega) \right| \\ &\leq \frac{T(\omega)}{N} \sum_{n=j}^{N-1} |v_n(\theta) - v_n(\theta^*)| + \left| \frac{1}{N} \sum_{n=j}^{N-1} v_n(\theta^*) \xi_{n-j}(\omega) \right|. \end{aligned}$$

The second term on the right side of (10) has a limit zero as $N \rightarrow \infty$ because $\omega \in B$. Thus, we need only consider the first term on the right side of (10). In particular,

$$\begin{aligned} \frac{1}{N} \sum_{n=j}^{N-1} |v_n(\theta) - v_n(\theta^*)| &= \sum_{n=j}^{N-1} \frac{1}{N} \left| \sum_{k=0}^n [q_k(\theta) - q_k(\theta^*)] u_{n-k} \right| \\ &\leq U \sum_{n=j}^{N-1} \frac{1}{N} \sum_{k=0}^n |q_k(\theta) - q_k(\theta^*)|. \end{aligned}$$

From the way N^* was chosen we have that

$$\sum_{k=0}^{\infty} |q_k(\theta) - q_k(\theta^*)| \leq \sum_{k=0}^{N^*} |q_k(\theta) - q_k(\theta^*)| + 2\varepsilon \quad \text{for all } \theta$$

satisfying $\|\theta - \theta^*\| < \delta(\omega, \varepsilon)$. By continuity of q_1, \dots, q_{N^*} we can further restrict

$\delta(\omega, \varepsilon)$ so that $|q_k(\boldsymbol{\theta}) - q_k(\boldsymbol{\theta}^*)| < \varepsilon/(N^* + 1)$, $k = 0, 1, \dots, N^*$. Then

$$U \sum_{n=j}^{N-1} \frac{1}{N} \sum_{k=0}^n |q_k(\boldsymbol{\theta}) - q_k(\boldsymbol{\theta}^*)| \leq 3U\varepsilon,$$

and we conclude $\lim_{N \rightarrow \infty} (1/N) \sum_{n=j}^{N-1} v_n(\boldsymbol{\theta}) \xi_{n-j}(\omega) = 0$ for $j = 0, \dots, N^*$ and for all $\boldsymbol{\theta}$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| < \delta(\omega, \varepsilon)$ as was asserted. Clearly, the first term on the right side of (9) can be made arbitrarily small because N^* is fixed. Now consider the second term on the right side of (9):

$$\left| \frac{1}{N} \sum_{j=N^*+1}^{N-1} p_j \sum_{n=j}^{N-1} v_n(\boldsymbol{\theta}) \xi_{n-j}(\omega) \right| \leq \frac{NVT(\omega)}{N} \sum_{j=N^*+1}^{N-1} |p_j| < \varepsilon$$

for all $\boldsymbol{\theta}$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| < \delta(\omega, \varepsilon)$. Q.E.D.

We are now in a position to establish the two main results. First we show that $\lim_{N \rightarrow \infty} F(N, \boldsymbol{\theta}, \cdot) = J(\boldsymbol{\theta}) \triangleq \lim_{N \rightarrow \infty} E[F(N, \boldsymbol{\theta}, \cdot)]$ almost surely and then we will establish that $\lim_{N \rightarrow \infty} \hat{\boldsymbol{\theta}}_N(\cdot) = \boldsymbol{\theta}^0$ almost surely. This will then establish the consistency of the least-squares estimator.

THEOREM 1. *If Assumptions S1, S4, R1 and R2 hold, then there is an event $\Omega^0 \subseteq \Omega$ with $P(\Omega^0) = 1$ such that $\lim_{N \rightarrow \infty} F(N, \boldsymbol{\theta}, \omega) = J(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\omega \in \Omega^0$.*

Proof. First we show that for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\lim_{N \rightarrow \infty} F(N, \boldsymbol{\theta}, \omega) = J(\boldsymbol{\theta})$ for almost all ω . Let $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ be fixed. Then, by taking expectations on both sides of (5) we have

$$\begin{aligned} E[F(N, \boldsymbol{\theta}, \cdot)] &= \frac{1}{N} \|Q_N(\boldsymbol{\theta})\mathbf{u}_N\|^2 + \frac{1}{N} E\|P_N \xi_N\|^2 + \frac{1}{N} E\|\boldsymbol{\eta}_N\|^2 \\ &\quad + \frac{2}{N} E\langle Q_N(\boldsymbol{\theta})\mathbf{u}_N, P_N \xi_N \rangle + \frac{2}{N} E\langle Q_N(\boldsymbol{\theta})\mathbf{u}_N, \boldsymbol{\eta}_N \rangle + \frac{2}{N} E\langle P_N \xi_N, \boldsymbol{\eta}_N \rangle \\ &= \frac{1}{N} \|Q_N(\boldsymbol{\theta})\mathbf{u}_N\|^2 + \frac{1}{N} E\|P_N \xi_N\|^2 + \frac{1}{N} E\|\boldsymbol{\eta}_N\|^2. \end{aligned}$$

From Lemma 2 applied to $\{\|Q_N(\boldsymbol{\theta})\mathbf{u}_N\|^2\}$ and $\{\|P_N \xi_N\|^2\}$ (rather than $\{\|Q_N(\boldsymbol{\theta})\xi_N\|^2\}$) and Assumption R1 it follows that

$$\lim_{N \rightarrow \infty} E[F(N, \boldsymbol{\theta}, \cdot)] = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} q_i(\boldsymbol{\theta}) q_j(\boldsymbol{\theta}) \tilde{u}(i, j) + \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} p_i p_j \gamma(i, j) \text{ var } [\eta_0].$$

Thus, $J(\boldsymbol{\theta})$ exists. Now apply Lemmas 2 and 3 to the terms in the expansion of $F(N, \boldsymbol{\theta}, \cdot)$ in (5) to conclude that $\lim_{N \rightarrow \infty} F(N, \boldsymbol{\theta}, \omega) = J(\boldsymbol{\theta})$ for almost all ω .

For each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, there is an almost sure event for which the convergence takes place, but the events may be different for different $\boldsymbol{\theta}$'s. We show now that there is an almost sure event Ω^0 such that for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\omega \in \Omega^0$, $\lim_{N \rightarrow \infty} F(N, \boldsymbol{\theta}, \omega) = J(\boldsymbol{\theta})$. Let D denote a countable dense subset of the admissible parameter set $\boldsymbol{\Theta}$. Define Ω^0 to be the event $\{\omega \in \Omega: \text{R1 and R2 hold, } \lim_{N \rightarrow \infty} F(N, \boldsymbol{\theta}, \omega) = J(\boldsymbol{\theta}) \text{ for all } \boldsymbol{\theta} \in D\}$. Ω^0 is a countable intersection of almost sure events, hence $P(\Omega^0) = 1$.

Let $\omega \in \Omega^0$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ be fixed. To show that $\lim_{N \rightarrow \infty} F(N, \boldsymbol{\theta}, \omega) = J(\boldsymbol{\theta})$ it is only necessary to establish that

$$\lim_{N \rightarrow \infty} \frac{1}{N} \langle Q_N(\boldsymbol{\theta})\mathbf{u}_N, P_N \xi_N(\omega) \rangle = \lim_{N \rightarrow \infty} \frac{1}{N} \langle Q_N(\boldsymbol{\theta})\mathbf{u}_N, \boldsymbol{\eta}_N(\omega) \rangle = 0$$

as the other terms in the expansion of $F(N, \theta, \omega)$ do not depend on θ or are deterministic.

Let $\varepsilon > 0$ be given. By Lemma 1 there is an integer N^* and a $\delta > 0$ such that (i) $\sum_{j=0}^{\infty} |q_j(\theta^1)| \leq \sum_{j=0}^{\infty} |q_j(\theta)| + 1$ and (ii) $\sum_{j=N^*+1}^{\infty} |q_j(\theta^1)| \leq \varepsilon$ for all θ^1 such that $\|\theta^1 - \theta\| < \delta$. This enables us to choose $\theta^1 \in D$ such that $\|\theta^1 - \theta\| < \delta$ and

$$(11) \quad \left| \sum_{j=0}^n q_j(\theta) u_{n-j} - \sum_{j=0}^n q_j(\theta^1) u_{n-j} \right| < \frac{\varepsilon}{T(\omega) \sum_{j=0}^{\infty} |p_j|}$$

for $n = 0, 1, 2, \dots$, where $T(\omega) = \sup_n \{|\xi_n(\omega)|\}$. Now

$$\begin{aligned} & \frac{1}{N} \langle Q_N(\theta) \mathbf{u}_N, P_N \xi_N(\omega) \rangle \\ &= \frac{1}{N} \sum_{n=0}^{N-1} v_n(\theta) \sum_{j=0}^n p_j \xi_{n-j}(\omega) \\ &= \frac{1}{N} \sum_{j=0}^{N-1} p_j \sum_{n=j}^{N-1} v_n(\theta) \xi_{n-j}(\omega) \\ &= \frac{1}{N} \sum_{j=0}^{N^*} p_j \sum_{n=j}^{N-1} v_n(\theta) \xi_{n-j}(\omega) + \frac{1}{N} \sum_{j=N^*+1}^{N-1} p_j \sum_{n=j}^{N-1} v_n(\theta) \xi_{n-j}(\omega) \\ &= \frac{1}{N} \sum_{j=0}^{N^*} p_j \sum_{n=j}^{N-1} [v_n(\theta) - v_n(\theta^1)] \xi_{n-j}(\omega) + \frac{1}{N} \sum_{j=0}^{N^*} p_j \sum_{n=j}^{N-1} v_n(\theta^1) \xi_{n-j}(\omega) \\ &\quad + \frac{1}{N} \sum_{j=N^*+1}^{N-1} p_j \sum_{n=j}^{N-1} v_n(\theta) \xi_{n-j}(\omega). \end{aligned}$$

The third term on the right side of the last equality can be made small by choosing N^* larger if necessary because $\sum_{j=0}^{\infty} |p_j| < \infty$ and $\{(1/N) \sum_{n=j}^{N-1} v_n(\theta) \xi_{n-j}(\omega)\}$ is bounded. The second term on the right side of the last equality is small for N sufficiently large because $\theta^1 \in D$ and N^* is fixed. Finally, the first term is small for large N because of (11). This establishes that $\lim_{N \rightarrow \infty} (1/N) \langle Q_N(\theta) \mathbf{u}_N, P_N \xi_N(\omega) \rangle = 0$ for $\theta \in \Theta$ and $\omega \in \Omega^0$.

The proof that $\lim_{N \rightarrow \infty} (1/N) \langle Q_N(\theta) \mathbf{u}_N, \eta_N(\omega) \rangle = 0$ for $\theta \in \Theta$ and $\omega \in \Omega^0$ is easier and will be omitted. It suffices to make use of the fact that $\lim_{N \rightarrow \infty} (1/N) \sum_{n=i}^{N-1} u_{n-i} \eta_n(\omega) = 0$ and $\sum_{i=N^*+1}^{\infty} |q_i(\theta)| < \varepsilon$.

THEOREM 2. *If Assumptions S1-S5, R1 and R2 hold, then $\lim_{N \rightarrow \infty} \hat{\theta}_N(\omega) = \theta^0$ for almost all ω .*

Proof. Let $\omega \in \Omega^0$, where Ω^0 is the set of probability one defined in the proof of Theorem 1. The parameter set is compact so $\{\hat{\theta}_N(\omega)\}$ has a convergent subsequence, say $\lim_{j \rightarrow \infty} \hat{\theta}_{N_j}(\omega) = \theta^*$. Suppose $\theta^* \neq \theta^0$. Then $J(\theta^*) > J(\theta^0)$ because J has a unique minimum (see [1; proof of Thm. 1]). Let $\varepsilon = J(\theta^*) - J(\theta^0)$. Now, $F(N_j, \hat{\theta}_{N_j}(\omega), \omega) \leq F(N_j, \theta^0, \omega)$ for each j , hence $\limsup_{j \rightarrow \infty} F(N_j, \hat{\theta}_{N_j}(\omega), \omega) \leq J(\theta^0)$. For j sufficiently large we have $F(N_j, \hat{\theta}_{N_j}(\omega), \omega) \leq J(\theta^0) + \varepsilon/4$. As noted in the proof of Theorem 1,

$$J(\theta) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} q_i(\theta) q_j(\theta) \tilde{u}(i, j) + \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} p_i p_j \gamma(i, j) + \text{var} [\eta_0].$$

With the aid of Lemma 1 it is easy to see that J is continuous in θ , so there is a $\delta_1 > 0$ such that $|J(\theta^*) - J(\theta)| < \varepsilon/8$ for all θ such that $\|\theta - \theta^*\| < \delta_1$. Applying Lemmas 2 and 3 to those terms in the expansion of $F(N, \theta, \omega)$ which depend explicitly on θ , we have that there is an integer $N_0(\omega, \varepsilon)$ and a $\delta_2(\omega, \varepsilon) > 0$ such that $|F(N, \theta, \omega) - J(\theta)| < \varepsilon/8$ for all θ satisfying $\|\theta - \theta^*\| < \delta_2(\omega, \varepsilon)$ and $N \geq N_0(\omega, \varepsilon)$. Let $\delta(\omega, \varepsilon) = \min(\delta_1, \delta_2(\omega, \varepsilon))$. Then for j sufficiently large,

$$\begin{aligned} |J(\theta^*) - F(N_j, \theta_{N_j}(\omega), \omega)| &\leq |J(\theta^*) - J(\hat{\theta}_{N_j}(\omega))| \\ &\quad + |J(\hat{\theta}_{N_j}(\omega) - F(N_j, \hat{\theta}_{N_j}(\omega), \omega)| < \frac{\varepsilon}{8} - \frac{\varepsilon}{8}. \end{aligned}$$

Then, $J(\theta^*) < F(N_j, \hat{\theta}_{N_j}(\omega), \omega) + \varepsilon/4 < J(\theta^0) + \varepsilon/2$ from which it follows that $J(\theta^*) - J(\theta^0) < \varepsilon/2$, contrary to the definition of ε .

REFERENCES

- [1] M. AOKI AND P. C. YUE, *On certain convergence questions in system identification*, this Journal, 8 (1970), pp. 239–256.
- [2] K. T. ASTROM, T. BOHLIN AND S. WENSMARK, *Automatic construction of linear stochastic dynamic models for stationary industrial processes with random disturbances using operating records*, IBM Nordic Laboratory, TP18.150, Lidingö, Sweden, 1965.
- [3] M. LEVIN, *Estimation of system pulse transfer function in the presence of noise*, IEEE Trans. Automatic Control, AC-9 (1964), pp. 229–235.
- [4] R. L. KASHYAP, *Maximum likelihood identification of stochastic linear systems*, Ibid., AC-15 (1970), pp. 25–34.
- [5] J. R. BLUM, D. L. HANSON AND L. H. KOOPMANS, *On the strong law of large numbers for a class of stochastic processes*, Wahrscheinlichkeitstheorie, 2 (1963), pp. 1–11.
- [6] M. MARDEN, *The Geometry of the Zeros of a Polynomial in a Complex Variable*, Mathematical Surveys 3, American Mathematical Society, New York, 1949.
- [7] R. B. ASH, *Complex Variables*, Academic Press, New York, 1971.
- [8] H. B. MANN AND A. WALD, *On the statistical treatment of linear stochastic difference equations*, Econometrica, 11 (1943), pp. 173–220.
- [9] R. C. K. LEE, *Optimal Estimation, Identification, and Control*, MIT Press, Cambridge, Mass., 1964.

AN ARC METHOD FOR NONLINEAR PROGRAMMING*

GARTH P. McCORMICK†

Abstract. An algorithm using second derivatives for solving the optimization problem: minimize $f(x)$ subject to $g_i(x) \geq 0$, $i = 1, \dots, m$, where the g_i are not necessarily linear is presented. The basic idea is to generate a sequence of feasible points with decreasing objective value by movement along piecewise, smooth, quadratic arcs. Cluster points of the sequence generated are shown to be second order Kuhn–Tucker points. If the strict second order sufficiency conditions hold, the rate of convergence is shown to be at least quadratic.

1. Introduction. The mathematical programming problem addressed is

$$(1.1) \quad \text{minimize } f(x)$$

subject to

$$(1.2) \quad g_i(x) \geq 0, \quad i = 1, \dots, m,$$

where the problem functions are at least twice continuously differentiable and $x \in E^n$. The functions $f(x)$, $\{g_i(x)\}$ are allowed to be nonlinear. A partial listing of methods for solving this problem is: a class of algorithms called methods of feasible directions [13], the generalized reduced gradient method [2], sequential unconstrained minimization techniques [4], and the gradient projection method [10]. The algorithm presented here is an extension of the variable reduction method [7], [8], and is a refinement of ideas presented in [9].

In this section notation is presented which will be used throughout the paper. A linear independence assumption in force throughout the paper and a discussion of the first and second order conditions, some necessary, others sufficient, which characterize local minimizers to problem (1.1), (1.2) are also given. Section 2 contains a heuristic and rigorous description of the algorithm. Section 3 contains a proof of the convergence of points generated by the algorithm to points satisfying the first and second order necessary conditions (second order Kuhn–Tucker points). In § 4 it is shown that if a cluster point of the points generated satisfies the strict second order sufficiency conditions, the rate of convergence of the algorithm is at least quadratic. Section 5 contains historical information and a discussion of the theoretical and computational implications of the algorithm.

Notation.

$x^{k,p}$	The value of x at the beginning of the p th arc of the k th iteration.
$r(k, p)$	The number of constraints exactly satisfied at the beginning of the p th arc of the k th iteration.
$R(k, p)$	A permuted ordered set of indices $1, \dots, m$, where the first $r(k, p)$ indices are of the constraints exactly satisfied at the beginning of the p th arc of the k th iteration.

* Received by the editors April 19, 1973, and in revised form November 6, 1974.

† Department of Operations Research, George Washington University, Washington, D.C. 20006. This research was supported by the United States Army Research Office.

$C(k, p)$	A permuted ordered set of the indices $1, \dots, n$, where the variables whose indices are among the first $r(k, p)$ are called "dependent" variables, and the remainder "independent" variables.
$x_D(k, p)$	The $r(k, p) \times 1$ vector of values of the problem variables at the beginning of the p th arc of the k th iteration where the connection with the vector $x^{k,p}$ is given through the first $r(k, p)$ indices in the permutation set $C(k, p)$ —considered the vector of "dependent" variables.
$x_I(k, p)$	The $(n - r(k, p)) \times 1$ vector of "independent" variables.
$g_i(k, p)$	The value of the i th constraint at $x^{k,p}$.
$\nabla g_i(k, p)$	The $n \times 1$ vector of first partial derivatives of the i th constraint evaluated at $x^{k,p}$ and in the order prescribed by the permutation set $C(k, p)$.
$[B(k, p), E(k, p)]$	The $r(k, p) \times n$ matrix of transposed gradients of the constraints, evaluated at $x^{k,p}$, whose indices are the first $r(k, p)$ in the permutation set $R(k, p)$ and whose columns are ordered by the set $C(k, p)$.
$\nabla_D f(k, p)$	The $r(k, p) \times 1$ vector of first partial derivatives of $f(x)$ evaluated at $x^{k,p}$ in the order prescribed by the permutation set $C(k, p)$ —similar notation is used for other quantities which depend on just the "dependent" or "independent" variables.
$a(k, p, t)$	The arc generated at $x^{k,p}$.
$P(k)$	The number of arcs used at iteration k .
$\bar{t}(k, p)$	The step-size scalar.
$\hat{u}(k, 1)$	Estimate of Kuhn–Tucker multipliers.
$s_I(k, p)$	Direction of search of independent variables.
$\alpha(k, p, t)$	Arc generated by Newton's method.

We now formulate a regularity condition which will be in force throughout the paper and which will be used for the proofs of convergence and rate of convergence.

Linear independence assumption A_1 . Let \bar{x} be any point satisfying (1.2). Let $\bar{S} = \{i | g_i(\bar{x}) = 0\}$. Then the set of gradients

$$(1.3) \quad \{\nabla g_i(\bar{x})\}, \quad i \in \bar{S},$$

is *linearly independent*.

We call a point \bar{x} a Kuhn–Tucker point (KTP) if \bar{x} satisfies the following well-known Kuhn–Tucker conditions, (1.4)–(1.7), which are necessary (under the linear independence assumption (1.3)) for \bar{x} to be a local minimizer to the problem (1.1), (1.2): there exist $\bar{u}_i, i = 1, \dots, m$, such that

$$(1.4) \quad \nabla f(\bar{x}) - \sum_{i=1}^m \bar{u}_i \nabla g_i(\bar{x}) = 0,$$

$$(1.5) \quad \bar{u}_i g_i(\bar{x}) = 0, \quad i = 1, \dots, m,$$

$$(1.6) \quad \bar{u}_i \geq 0, \quad i = 1, \dots, m,$$

$$(1.7) \quad g_i(\bar{x}) \geq 0, \quad i = 1, \dots, m.$$

The Kuhn–Tucker conditions, (1.4)–(1.7), are first order conditions in that they involve only the first derivatives of the problem functions. Less well known, but equally important, are second order conditions; conditions involving the second derivatives of the problem functions. We state without proof two lemmas, one stating second order necessary, the other stating second order sufficiency conditions concerning local minimizers of problem (1.1), (1.2). For proofs and a discussion of these lemmas see Chapter 2 of [4].

LEMMA 1. *If the problem functions are twice continuously differentiable, and if the linear independence assumption (1.3) holds, then a necessary condition that a point \bar{x} be a local minimizer to the problem (1.1), (1.2) is that there exists a vector \bar{u} such that (\bar{x}, \bar{u}) satisfies (1.4)–(1.7), and also satisfies*

$$(1.8) \quad z' \nabla^2 L(\bar{x}, \bar{u}) z \geq 0$$

for all z such that

$$z' \nabla g_i(\bar{x}) = 0 \quad \text{for all } i \text{ where } g_i(\bar{x}) = 0,$$

where

$$(1.9) \quad L(x, u) \equiv f(x) - \sum_{i=1}^m u_i g_i(x).$$

A point satisfying (1.4)–(1.8) is called a second order Kuhn–Tucker point (SOKTP).

LEMMA 2. *If the problem functions are twice continuously differentiable, and if associated with some x^* there is a u^* such that (x^*, u^*) satisfies (1.4)–(1.7) and also satisfies*

$$(1.10) \quad z' \nabla^2 L(x^*, u^*) z > 0$$

for all $z \neq 0$ such that

$$z' \nabla g_i(x^*) = 0 \quad \text{for all } i \text{ where } g_i(x^*) = 0,$$

then x^* is an isolated local minimizer of problem (1.1), (1.2).

We note here that these theorems are not the best theorems for characterizing local minimizers, but for the purposes of proving theorems about the convergence and rate of convergence of the algorithm, they are the versions required.

2. The algorithm. To motivate the rigorous statement of the algorithm, an intuitive derivation is presented first. For simplicity the problem to be solved is stated

$$\text{minimize } f(x) \text{ subject to } g_i(x) = 0, \quad i = 1, \dots, r.$$

Suppose a point x^0 satisfying the r equality constraints is given. The following is based on the notion that one should try to “solve for” r of the variables in terms of $n - r$ of them and then find the unconstrained minimizer of f , which can now be thought of as a function only of $n - r$ variables. Another way of looking at this

approach is to generate an arc emanating from x^0 along which the constraints are exactly satisfied and to attempt to minimize the objective function along that arc. To that end we divide the vector x into two parts, x_D , consisting of the first r components of x and considered as the vector of "dependent" variables, those to be "solved for," and x_I , the remaining $n - r$ variables considered as the vector of "independent" variables. We assume the form of the arc as

$$a(t) = \begin{pmatrix} x_D^0 \\ x_I^0 \end{pmatrix} + \begin{pmatrix} s_D \\ s_I \end{pmatrix} t + \begin{pmatrix} v \\ 0 \end{pmatrix} t^2,$$

where s_D is the first order direction of search of the "dependent" variables, and s_I is the first order direction of search of the "independent" variables. The vector v is the second order direction of search of the "dependent" variables. Now an approximation to the value of any constraint along the arc is

$$g_i[a(t)] \doteq g_i(x^0) + \frac{dg_i[a(0)]}{dt}t + \frac{d^2g_i[a(0)]}{dt}t^2/2.$$

Hence, if the quantities s_D and v can be made to depend upon s_I so that $dg_i[a(0)]/dt = 0$ and $d^2g_i[a(0)]/dt^2 = 0$, for $i = 1, \dots, r$, motion close to that required to remain in the space of equality constraints can be maintained. Then s_I can be chosen independently to aid in the "unconstrained" minimization of f .

The necessary algebraic quantities are obtained by a simple application of the chain rule for differentiation. We have, for $i = 1, \dots, r$,

$$(2.1) \quad dg_i[a(0)]/dt = [s'_D, s'_I] \nabla g_i[a(0)],$$

$$(2.2) \quad d^2g_i[a(0)]/dt^2 = [s'_D, s'_I] \nabla^2 g_i[a(0)] [s'_D, s'_I]' + 2[v', 0] \nabla g_i[a(0)].$$

Dividing the matrix of constraint gradients into two parts, B' and E' , where B is an $r \times r$ invertible matrix (we can assume this without loss of generality) and E is an $r \times (n - r)$ matrix, we can solve for s_D in (2.1) as

$$(2.3) \quad s_D = -B^{-1}Es_I.$$

Substituting this in (2.2) and solving for v yields (letting $T' = [-E'(B')^{-1}, I]$)

$$(2.4) \quad v = -B^{-1}(\dots, s'_I T' \nabla^2 g_i[a(0)] T s_I / 2, \dots)'.$$

Thus, both s_D and v are specified for any choice of s_I . Note that if any g_i is a linear function, exact constraint satisfaction results along the arc.

Assuming the same approximation for the value of the objective function along the arc, and substituting the quantities obtained from (2.3) and (2.4), we have

$$(2.5) \quad \begin{aligned} f[a(t)] \doteq & f[a(0)] + s'_I T' \nabla f[a(0)] t + s'_I T' \nabla^2 f[a(0)] T s_I t^2 / 2 \\ & - [\dots, s'_I T' \nabla^2 g_i[a(0)] T s_I, \dots] (B')^{-1} \nabla_D f[a(0)] t^2 / 2. \end{aligned}$$

Defining

$$u = (B')^{-1} \nabla_D f[a(0)], \quad L(x, u) = f(x) - \sum_{i=1}^r u_i g_i(x),$$

$$h' = T' \nabla f[a(0)], \quad H = T' \nabla^2 L(x, u) T,$$

the unconstrained minimization problem becomes

$$(2.6) \quad \underset{(s_I, t)}{\text{minimize}} \quad f[a(0)] + h's_I t + s_I' H s_I t^2 / 2.$$

There are two cases to consider. The first case is that the function in (2.6) has an unconstrained minimizer in s_I, t . A necessary condition for this (except for pathological cases) is that H must be a positive definite matrix. Clearly there are an infinite number of values of s_I, t for which the solution is obtained since the crucial quantity is the product of s_I and t . Arbitrarily it will be assumed that $t^* = 1$, and thus

$$s_I^* = -H^{-1}h$$

is the optimum direction vector for the independent variables.

The second case is when the solution to (2.6) is unbounded, i.e., it is possible to choose s_I^* so that the function in (2.6) goes to $-\infty$ as $t \rightarrow \infty$. This case often arises in the early stages of trying to solve a nonconvex programming problem. A good strategy for this case has been to choose s_I^* to be an eigenvector associated with the minimum eigenvalue of H . In practice, this creates a fast decrease in f and forces the arc into a region where the positive definiteness of H obtains (or, as usually happens when there are inequality constraints, forces the arc into the boundaries of constraints previously well-satisfied).

This is the basic motivation for the algorithm, finding the unconstrained minimizer of f in an equality constrained region. Three matters have not been covered: since the arc does not, in general, maintain exact constraint satisfaction, what procedure is used to do this; for the inequality constrained problem, how should the decision be made on when to move interior to a constraint boundary whose exact satisfaction has been insisted upon; and, when in the course of moving along the arc to find the unconstrained minimizer, what should be done if a constraint, not previously binding, is encountered?

The method used to enforce exact constraint satisfaction is a modification of Newton's method for solving nonlinear equations. The decision to drop a constraint from consideration is based upon periodic estimates of the Kuhn-Tucker multipliers which are required to be nonnegative. If a constraint is encountered before minimization along the arc occurs, the constraint is incorporated into a list of those required to be binding, and unconstrained minimization continues in the subspace of reduced dimension. No constraints are dropped from consideration until an unconstrained minimization has occurred in some subspace.

A rigorous statement of the algorithm for the inequality constrained problem will now be given.

Iteration 0, Step 1. Let x^0 denote the given feasible starting points. Set $r(0, 1)$ equal to the number of constraints which are exactly satisfied at x^0 , i.e., the number of indices for which $g_i(x^0) = 0$, $i = 1, \dots, m$. Set $R(0, 1)$ to be a rearrangement of the constraint indices so that those constraints which are exactly equal to zero correspond to the first $r(0, 1)$ indices. Set $x(0, 1, j) = x_j^0$, for $j = 1, \dots, n$, and $C(0, 1, j) = j$ for $j = 1, \dots, n$. Continue as in the first step of the general iteration k , where $x^{0,1}$ is set equal to x^0 .

Iteration k , Step 1 (for all $k \geq 0$). Let $x^{k,1}$ denote the value of x at the beginning

of the first arc at the k th iteration. At this point the integer $r(k, 1)$ indicates the number of constraints which are exactly satisfied at $x^{k,1}$. The set $R(k, 1)$ is a permutation set of the integers $1, \dots, m$. The constraints whose indices are in the first $r(k, 1)$ numbers in $R(k, 1)$ are those constraints exactly satisfied. The set $C(k, 1)$ is a permutation set of the integers $1, \dots, n$. The variables whose indices are of the first $r(k, 1)$ are called dependent and the variables whose indices are $C(k, 1, r(k, 1) + 1), \dots, C(k, 1, n)$ are called independent variables. The vector $x(k, 1)$ is a permuted $n \times 1$ vector of the values of the vector $x^{k,1}$. They are related through the permutation set $C(k, 1)$ as

$$x(k, 1, j) = x_{C(k, 1, j)}^{k,1}, \quad j = 1, \dots, n.$$

We denote $x_D(k, 1)$ to be the $r(k, 1) \times 1$ vector containing the first $r(k, 1)$ components of $x(k, 1)$, and $x_I(k, 1)$ to denote the $(n - r(k, 1)) \times 1$ vector containing the last components of $x(k, 1)$. Two matrices of constraint derivatives $B(k, 1)$ (which may be thought of as the "basis" matrix) and $E(k, 1)$ are defined as

$$B(k, 1)_{i,j} = \frac{\partial g_{R(k, 1, i)}(x^{k,1})}{\partial x_{C(k, 1, j)}} \\ i = 1, \dots, r(k, 1), \quad j = 1, \dots, r(k, 1), \\ E(k, 1)_{i,j} = \frac{\partial g_{R(k, 1, i)}(x^{k,1})}{\partial x_{C(k, 1, j)}} \\ i = 1, \dots, r(k, 1), \quad j = r(k, 1) + 1, \dots, n.$$

It is important for proving convergence of the algorithm, and more importantly for numerical stability of the process, that the matrix $B(k, 1)$ be well-conditioned. This means that the choice of which variables are dependent must be made with care. For $K \geq 1$, previous selections have been made so that presumably the order of the variable indices in the permutation set $C(k, 1)$ tend to give $B(k, 1)$ numerical stability. Since the constraints are nonlinear, recomputation of the gradient vectors of the binding constraints can result in an ill-conditioned $B(k, 1)$ -matrix. It is necessary, therefore, to test for the stability of $B(k, 1)$ and, if necessary, rearrange the indices in $C(k, 1)$. Other considerations for the choice of which variables are to be dependent are dictated by the form in which the inverse of $B(k, 1)$ is to be represented. Recent work in large scale linear programming [5] has modified many of the traditional views about the use of the product form of the inverse and its many variations. The maintaining of a sparse representation is important. The results of current experimentation in linear programming will have much to do with the selection mechanisms for nonlinearly constrained problems. For this reason it will just be assumed here that an appropriate way of selecting the dependent variables, or changing them if necessary, is done at the first step of each iteration. That a suitable choice exists is ensured by the linear independence assumption (1.3). One of the implications of using a numerically stable method to select the dependent variables is that the columns of $B(k, 1)$ will not be "nearly" linearly dependent. A measure of this is the value of determinant of the matrix $B(k, 1)$. Careful selection coupled with the linear independence assumption and the continuity of the problem function derivatives implies that the absolute value

of the determinant of $B(k, 1)$ is bounded below away from zero. This is important in proving the convergence of the algorithm.

Deletion of constraints from list of those required to remain binding. The list of those constraints equal to zero in value and which are required to remain at zero value is changed by deletion only at the beginning of the first step of each iteration. In this case, one constraint, at most, is removed from the list by the following method. Let $\hat{u}(k, 1) = B'(k, 1)^{-1} \nabla_D f(k, 1)$ (estimates of Kuhn–Tucker multipliers) have components $\hat{u}(k, 1, i)$, $i = 1, \dots, r(k, 1)$. Denote i_1 to be an index such that

$$(2.7) \quad \hat{u}(k, 1, i_1) = \min_i \hat{u}(k, 1, i).$$

If $\hat{u}(k, 1, i_1) < 0$, delete the corresponding constraint from the list of those required to remain binding. When this is done, a dependent variable must be selected to be made into an independent variable. Assume the existence of a numerically stable selection mechanism which decides that the variable with index $C(k, 1, j_1)$ is to become an independent variable. The row and column deletion is accomplished formally by interchanging $C(k, 1, j_1)$ with $C(k, 1, r(k, 1))$, $x(k, 1, j_1)$ with $x(k, 1, r(k, 1))$, $R(k, 1, i_1)$ with $R(k, 1, r(k, 1))$, and setting $r(k, 1) = r(k, 1) - 1$. Now all the quantities in (2.7) must be recomputed.

Iteration k , Step p (for all $k \geq 0$, $p \geq 1$). Let $x^{k,p}$ denote the value of x at the beginning of the p th step (or arc) at the k th iteration. The number of constraints equal to zero and required to remain equal to zero is designated by $r(k, p)$. The set $R(k, p)$ is a permutation set of the integers $1, \dots, m$. The constraints whose indices are in the first $r(k, p)$ numbers in $R(k, p)$ are those constraints exactly satisfied. The set $C(k, p)$ is a permutation set of the integers $1, \dots, n$. The variables whose indices are of the first $r(k, p)$ are called dependent variables, and the variables whose indices are $C(k, p, r(k, p) + 1), \dots, C(k, p, n)$ are called independent variables. The vector $x(k, p)$ is a permuted $n \times 1$ vector of the values of the vector $x^{k,p}$. They are related through the permutation set $C(k, p)$ as

$$x(k, p, j) = x_{C(k,p,j)}^{k,p}, \quad j = 1, \dots, n.$$

We denote $x_D(k, p)$ to be the $r(k, p) \times 1$ vector containing the first $r(k, p)$ components of $x(k, p)$, and $x_I(k, p)$ to denote the $(n - r(k, p)) \times 1$ vector containing the last $(n - r(k, p))$ components of $x(k, p)$. Two matrices of constraint derivatives $B(k, p)$ and $E(k, p)$ are defined as

$$B(k, p)_{i,j} = \frac{\partial g_{R(k,p,i)}(x^{k,p})}{\partial x_{C(k,p,j)}}, \quad i = 1, \dots, r(k, p), \quad j = 1, \dots, r(k, p),$$

$$E(k, p)_{i,j} = \frac{\partial g_{R(k,p,i)}(x^{k,p})}{\partial x_{C(k,p,j)}}, \quad i = 1, \dots, r(k, p), \quad j = r(k, p) + 1, \dots, n.$$

The matrices of second derivatives of the problem functions are defined as $\nabla^2 g_l(k, p)$ with the i, j th component equal to

$$\frac{\partial^2 g_{R(k,p,l)}(x^{k,p})}{\partial x_{C(k,p,i)} \partial x_{C(k,p,j)}}, \quad l = 1, \dots, r(k, p).$$

To prescribe the direction of search at the p th step of the k th iteration, it is necessary to define the following quantities. Let $u(k, p, i)$ be the i th component of

$$(2.8) \quad u(k, p) = B'(k, p)^{-1} \nabla_D f(k, p),$$

$$(2.9) \quad T'(k, p) = [-E'(k, p)B'(k, p)^{-1}, I],$$

$$(2.10) \quad h(k, p) = T'(k, p) \nabla f(k, p),$$

and

$$(2.11) \quad H(k, p) = T'(k, p) \nabla^2 L(k, p) T(k, p),$$

where

$$(2.12) \quad \nabla^2 L(k, p) = \nabla^2 f(k, p) - \sum_{i=1}^{r(k, p)} u(k, p, i) \nabla^2 g_i(k, p).$$

The most important quantity to compute in the arc is the direction of search of the independent variables. The computation takes three forms depending upon the eigenvalues of $H(k, p)$. Let ε_1 be a preselected positive value (presumed "small"). Denote by $\delta(k, p)$ the smallest eigenvalue of the symmetric matrix $H(k, p)$. Let $e(k, p)$ denote an eigenvector associated with $\delta(k, p)$.

Case 1.

$$\delta(k, p) \geq \varepsilon_1 > 0.$$

Then

$$(2.13) \quad s_I(k, p) = -H(k, p)^{-1} h(k, p).$$

Case 2.

$$\varepsilon_1 > \delta(k, p) > 0.$$

Then

$$(2.14) \quad s_I(k, p) = -h(k, p).$$

Case 3.

$$0 \geq \delta(k, p).$$

Then

$$(2.15) \quad s_I(k, p) = -h(k, p) + e(k, p),$$

where, without loss of generality, it can be assumed that

$$(2.16) \quad h'(k, p) e(k, p) \leq 0.$$

We define an $r(k, p) \times 1$ vector $W(k, p)$ whose i th component is

$$(2.17) \quad W(k, p, i) = -s_I'(k, p) T'(k, p) \nabla^2 g_i(k, p) T(k, p) s_I(k, p) / 2,$$

and a vector $v(k, p)$ given as

$$(2.18) \quad v(k, p) = B(k, p)^{-1} W(k, p).$$

The arc $a(k, p, t)$ is now described as

$$(2.19) \quad a(k, p, t) = x(k, p) + T(k, p)s_I(k, p)t + \begin{pmatrix} v(k, p) \\ 0 \end{pmatrix} t^2.$$

For notational convenience this will sometimes be written

$$a(k, p, t) = x(k, p) + b(k, p)t + w(k, p)t^2,$$

where

$$b(k, p) = T(k, p)s_I(k, p)$$

and

$$w(k, p) = \begin{pmatrix} v(k, p) \\ 0 \end{pmatrix}.$$

Maintenance of exact constraint satisfaction by use of a modified form of Newton's method. Movement along the arc given by (2.19) will not, in general, maintain exact constraint satisfaction. To correct for this an arc which is obtained from that given by (2.19) is generated using a form of Newton's method for solving nonlinear equations. We define an arc $\alpha(k, p, t)$ as follows. Set

$$\alpha(k, p, t, 0) = a(k, p, t).$$

Let $\nabla_D g(k, p, t, l)$ denote the submatrix of constraint gradients defined at $\alpha(k, p, t, l)$ by the following equations. The i, j th component of the matrix is

$$\frac{\partial g_{R(k, p, i)}(\alpha(k, p, t, l))}{\partial x_{C(k, p, j)}}, \quad i = 1, \dots, r(k, p), \quad j = 1, \dots, r(k, p).$$

Then $\alpha(k, p, t, l + 1)$ is obtained from $\alpha(k, p, t, l)$ as

$$(2.20) \quad \alpha(k, p, t, l + 1) = \alpha(k, p, t, l) - \begin{bmatrix} \nabla_D' g(k, p, t, l)^{-1} \\ 0 \end{bmatrix} g_{D(k, p, t, l)}.$$

There is a value $t_3(k, p) > 0$ such that for all t where $0 \leq t \leq t_3(k, p)$, the iterations defined by (2.20) converge and define an arc which has all the differentiability properties of $a(k, p, t)$. This arc is denoted by $\alpha(k, p, t)$. In practice, rules are necessary for deciding when to cease iterating on (2.20), either because convergence has been successfully approximated, or because the iterations are diverging. For now, the ability to ascertain $t_3(k, p)$ and $\alpha(k, p, t)$ exactly will be assumed.

Obtaining the next point. Find $t(k, p) = \min [t_1(k, p), t_2(k, p), t_3(k, p)]$, where $t_1(k, p)$ is the smallest local minimizer of

$$(2.21) \quad \begin{aligned} &\text{minimize } f[\alpha(k, p, t)] \\ &\text{subject to } 0 \leq t \leq t_3(k, p), \end{aligned}$$

and $t_2(k, p)$ is the smallest maximizer of

$$\begin{aligned} & \text{maximize } t \\ & \text{subject to } 0 \leq t \leq t_3(k, p), \end{aligned}$$

and

$$g_{R(k, p, i)}[\alpha(k, p, t)] \geq 0, \quad i = r(k, p) + 1, \dots, m.$$

If $t(k, p) = t_2(k, p)$, the number of constraints considered binding must be increased by 1. This is accomplished by first setting $r(k, p) = r(k, p) + 1$, and then interchanging $R(k, p, r(k, p))$ and $R(k, p, i_1)$, where $R(k, p, i_1)$ is the index of the constraint the encountering of which caused cessation of motion along the arc. Also, a variable must be changed from an independent one to a dependent one. Again, it is assumed that an appropriate selection mechanism exists that chooses the variable with index $C(k, p, J_1)$ to become a dependent variable. Then interchange $C(k, p, J_1)$ with $C(k, p, r(k, p))$.

In this case, when $t(k, p) = t_2(k, p)$, set $x(k, p + 1) = \alpha(k, p, t(k, p))$ and begin the $(p + 1)$ st step of the k th iteration. Otherwise, when $t(k, p) < t_2(k, p)$, set $x(k + 1, 1) = \alpha(k, p, t(k, p))$ and begin iteration $k + 1$.

3. Convergence of the algorithm. It is useful to define a continuous vector function which yields the value of x for any step along the arcs during the course of the k th iteration and its successor iteration $k + 1$. Let $P(k)$ denote the number of arcs used at iteration k , and define

$$t^k = \sum_{p=1}^{P(k)} t(k, p).$$

Let $\tau \geq 0$ be given. There are two cases to consider. Suppose $0 \leq \tau \leq t^k$. Let $q(k, \tau)$ be the largest integer from $(1, \dots, P(k))$ for which

$$\sum_{p=1}^{q(k, \tau)-1} t(k, p) \leq \tau.$$

Let

$$\beta(k, \tau) = - \sum_{p=1}^{q(k, \tau)-1} t(k, p) + \tau.$$

Then the continuous vector function $x^k(\tau)$ is defined as

$$\begin{aligned} (3.1) \quad x^k(\tau) = & x(k, 1) + \sum_{p=1}^{q(k, \tau)-1} b(k, p)t(k, p) + w(k, p)t^2(k, p) \\ & + b(k, q(k, \tau))\beta(k, \tau) + w(k, q(k, \tau))\beta^2(k, \tau). \end{aligned}$$

Suppose $\tau > t^k$. Then set $\tau = \tau - t^k$ and follow the above procedure, where k is replaced by $k + 1$ everywhere except for the left-hand side of (3.1).

Our next lemma states (in effect) that if \bar{x} , any cluster point of the sequence of points beginning each iteration is not a SOKTP, then successor cluster points are "not close" to it. In order to simplify the presentation of the proofs we will assume that the quadratic approximation to the boundaries of the constraints is

exact and that no modified Newton iterations are required to maintain exact constraint satisfaction. A series of lemmas will establish the convergence of the algorithm under these assumptions. Later we will show that the algorithm is valid with the modified Newton corrections.

LEMMA 3. *If $f, \{g_i\}$ are three times continuously differentiable, if the linear independence assumption holds, and if for a subsequence of $\{x(k, 1)\}$ converging to some cluster point \bar{x} ,*

$$(3.2) \quad \liminf_{k \rightarrow \infty} \max_{0 \leq t \leq t^k + t^k + 1} \|x^k(t) - x^k(0)\| = 0,$$

then \bar{x} is a SOKTP (where $x^k(0)$ is also used to denote the subsequence converging to \bar{x}).

Proof. The possible termination of an iteration because of the failure of the modified Newton method to converge for $t(k, p)$ has been precluded. Hence, motion ceases infinitely often because the directional derivative at $x(k, P(k))$ is nonnegative, i.e.,

$$(3.3) \quad df[a(k, P(k)), t(k, P(k))]/dt \geq 0.$$

Using the definitions of the computed quantities, and the chain rule for differentiation this is

$$(3.4) \quad \nabla' f[a(k, P(k)), t(k, P(k))] [b(k, P(k)) + w(k, P(k))t(k, P(k))] \geq 0,$$

where b and w are computed from $s_i(k, P(k))$ which is obtained from one of equations (2.13), (2.14) or (2.15). Because there are a finite number of constraints, and because inequality (3.3) is assumed to occur an infinite number of times, the division into dependent and independent variables, the order of the constraint indices in $R(k, P(k))$, and the value of $r(k, P(k))$ can be assumed to be the same for all k . For simplicity, and without loss of generality, we will assume all quantities in the following converge without formally selecting converging subsequences. Where necessary, of course, we will show that the algorithm is constructed in such a way that the quantities are uniformly bounded. This allows for the selection of converging subsequences.

The continuity of the derivatives of f , and (3.2) assures that $\nabla f(a(k, P(k)), t(k, P(k))) \rightarrow \nabla f(\bar{x})$. The careful selection of the dependent variables and the linear independence assumption (1.3) along with the continuity of the derivatives of the problem functions assure that the absolute value of the determinant of $B(k, P(k))$ is uniformly bounded below away from zero. Let r be the assumed constant value or $r(k, P(k))$ for large k . Then

$$B(k, P(k)) \rightarrow \bar{B}, \quad \text{an } r \times r \text{ invertible matrix,}$$

$$E(k, P(k)) \rightarrow \bar{E}, \quad \text{an } r \times (n - r) \text{ matrix,}$$

and all quantities in (2.8)–(2.21) converge to limits denoted, e.g., by \bar{u} , \bar{h} , \bar{H} and \bar{e} .

Because of (3.2), the term involving t in inequality (3.4) vanishes as $k \rightarrow \infty$. Taking Cases 1–3 (equations (2.13)–(2.15)) in order, and using inequality (3.4), we will show that $\nabla L(\bar{x}, \bar{u}) = 0$ by the following arguments.

If Case 1 yields $s_I(k, P(k))$ infinitely often, inequality (3.4) implies that

$$-\bar{h}'\bar{H}^{-1}\bar{h} \geq 0,$$

where the smallest eigenvalue of \bar{H} is greater than or equal to ε_1 which is greater than 0. This implies that $\bar{h} = 0$ which is equivalent to $\nabla L(\bar{x}, \bar{u}) = 0$.

If Case 2 is used to generate $s_I(k, P(k))$ infinitely often, inequality (3.4) yields

$$-\bar{h}'\bar{h} \geq 0,$$

which implies that \bar{h} and hence $\nabla L(\bar{x}, \bar{u})$ both vanish.

If Case 3 is used to generate $s_I(k, P(k))$ infinitely often, (3.4) yields

$$-\bar{h}'\bar{h} + \bar{h}'\bar{e} \geq 0,$$

where, by virtue of (2.16), $\bar{h}'\bar{e} \leq 0$. As above, it is immediate that $\nabla L(\bar{x}, \bar{u}) = 0$.

Thus, for all possible cases, (\bar{x}, \bar{u}) satisfies $\nabla L(\bar{x}, \bar{u}) = 0$. Next, we need to show that $\bar{u} \geq 0$. To do this, both the point $x(k, 1)$ and its successor point must be considered. Because of (3.2), if $x(k_j, 1) \rightarrow \bar{x}$, so does $x(k_j + 1, 1)$ (as $j \rightarrow \infty$). Hence, the previous result holds for $x(k + 1, P(k + 1))$ (assuming for notational convenience that $x(k, 1)$ represents the converging subsequence instead of $x(k_j, 1)$).

Consider at $x(k, P(k))$ and $x(k + 1, P(k + 1))$ the vectors $u(k, P(k))$ and $u(k + 1, P(k + 1))$. From the proof above it follows that as $k \rightarrow \infty$, these quantities tend to the Lagrange multipliers associated with the stationarity of the Lagrangian function. Let the set S^k be defined as the set of the first $r(k, P(k))$ indices in $R(k, P(k))$, and S^{k+1} as the first indices of binding constraints in $R(k + 1, P(k + 1))$. Assume that the indices in these sets are constant for large k . (This can be done since there are only a finite number of constraints.) If the sets S^k and S^{k+1} contain different sets of integers, the differences can only be that one or the other contains indices of constraints binding at \bar{x} whose limiting multiplier values are zero. Otherwise, the linear independence assumption would be violated. Hence, for k large, S^k and S^{k+1} both contain the indices of all constraints binding at \bar{x} with nonzero Lagrange multipliers which constitute the multiplier values for which the stationarity condition holds.

Assume $\bar{u}_i < 0$ for some $i \in S^k$ infinitely often. Eventually, the index corresponding to one of the multipliers with smallest multiplier value would be eliminated from the set of indices of constraints required to remain binding at the start of iteration $k + 1$. By the argument just made, it must be in S^{k+1} and, hence, it must be picked up at some step between 1 and $P(k + 1)$.

Let \mathcal{J} denote one of the integers satisfying

$$\bar{u}_{\mathcal{J}} = \min_i \bar{u}_i$$

which is dropped from the indices of constraints required to remain binding at $x(k + 1, 1)$ infinitely often. Expanding in a Taylor's series,

$$(3.5) \quad g_{\mathcal{J}}(k + 1, l(k + 1)) = g_{\mathcal{J}}(k + 1, 1) + \left\{ \sum_{p=1}^{l(k+1)-1} b(k + 1, p)t(k + 1, p) \cdot w(k + 1, p)t^2(k + 1, p) \right\} \nabla g_{\mathcal{J}}(\eta),$$

where η is a convex combination of $x(k + 1, l(k + 1))$ and $x(k + 1, 1)$ and where

$l(k+1)$ is the step during the $(k+1)$ st iteration when the \mathcal{J} th constraint is picked up again. Now $g_{\mathcal{J}}(k+1, 1) = 0 = g_{\mathcal{J}}(k+1, l(k+1))$. Dividing (3.5) by

$$\sum_{p=1}^{l(k+1)-1} t(k+1, p)$$

and taking the limit as $k \rightarrow \infty$ yields (using (3.2)) a result from which it can be deduced that

$$(3.6) \quad \tilde{s}'_l[-\tilde{E}'(\tilde{B}')^{-1}, I]\nabla g_{\mathcal{J}}(\bar{x}) = 0,$$

where \tilde{B} is the limit of basis matrices which do not involve the \mathcal{J} th constraint. Now, the matrix in the above is orthogonal to the gradients of all the other constraints with indices in S^k , i.e.,

$$(3.7) \quad [-\tilde{E}'(\tilde{B}')^{-1}, I]\nabla g_i(\bar{x}) = 0 \quad \text{for } i \neq \mathcal{J}, \quad i \in S^k.$$

Because of the Lagrangian stationarity condition we have

$$(3.8) \quad \nabla f(\bar{x}) - \sum_{\substack{i \neq \mathcal{J} \\ i \in S^k}} \bar{u}_i \nabla g_i(\bar{x}) = \bar{u}_{\mathcal{J}} \nabla g_{\mathcal{J}}(\bar{x}).$$

Now using (3.6) and (3.7), multiplying (3.8) by $\tilde{s}'[-\tilde{E}'(\tilde{B}')^{-1}, I]$ yields

$$\tilde{s}'_l[-\tilde{E}'(\tilde{B}')^{-1}, I]\nabla f(\bar{x}) = 0,$$

where \tilde{s}_l is of the form obtained from either (2.13), (2.14) or (2.15). Thus, we have either

$$-\tilde{h}'\tilde{H}^{-1}\tilde{h} = 0 \quad \text{or} \quad -\tilde{h}'\tilde{h} = 0 \quad \text{or} \quad -\tilde{h}'\tilde{h} + \tilde{h}'\tilde{e} = 0,$$

where

$$\tilde{h}'\tilde{e} \leq 0.$$

In each of these cases it can be argued, as in the first portion of the proof of this lemma, that (\bar{x}, \bar{u}) is a stationary point of the Lagrangian function. The fact that the \mathcal{J} th constraint has a multiplier value of zero, since it does not enter into the computation of \bar{u} , means that the smallest multiplier associated with the stationarity of the Lagrangian at \bar{x} is zero, otherwise, the linear independence assumption would be violated.

In the final part of the proof we show that the second order necessary conditions are satisfied by (\bar{x}, \bar{u}) .

The rules for the construction of the arc imply that the directional derivative of f , at the beginning of any arc (with respect to t) is nonpositive. Inequality (3.3) states that the directional derivative at the end of the iteration is nonnegative. Since no new constraints are picked up between the start of the last arc and its termination, the arc, as a function of t , is twice differentiable when the problem functions are. Expanding the directional derivative using Taylor's theorem we obtain

$$(3.9) \quad \frac{df}{dt}[a(k, P(k), t(k, P(k)))] = \frac{df}{dt^+}[a(k, P(k), 0)] + \frac{df^2}{dt^2}[a(k, P(k), \xi^k)] \\ \cdot t(k, P(k)),$$

where ξ^k is a convex combination of 0 and $t(k, P(k))$. Using the facts about the directional derivatives stated just before (3.9) and dividing through by $t(k, P(k))$ yields the inequality

$$\frac{d^2 f}{dt^2}(a(k, P(k), \xi^k)) \geq 0.$$

Because of (3.2), taking the limit as $k \rightarrow \infty$ yields

$$(3.10) \quad \bar{s}'_I \bar{H} \bar{s}_I \geq 0,$$

where \bar{s}_I , \bar{H} , and all the similar quantities with bars are limits of the quantities defined in (2.8)–(2.15). That these limits exist, i.e., that there exist converging subsequences of all those quantities, follows from the selection of the independent variables, the linear independence assumption, and the twice continuous differentiability of the problem functions.

There are three cases for the origin of the quantities in inequality (3.10). If (2.13) were used an infinite number of times to generate the vectors whose limit is \bar{s}_I , then \bar{x} is a SOKTP since

$$z' \bar{H} z \geq \varepsilon_1 \|z\|^2 \quad \text{for all } z.$$

If (2.14) were used an infinite number of times, the same conclusion would follow since

$$z' \bar{H} z \geq 0 \quad \text{for all } z.$$

If (2.15) were used an infinite number of times, because of the stationarity of the Lagrangian (recall $\bar{h} = 0$ was proved in the first part of the lemma) inequality (3.10) yields

$$\bar{e}' \bar{H} \bar{e} \geq 0,$$

where \bar{e} is an eigenvector of \bar{H} associated with its minimum eigenvalue. Thus, \bar{H} is a positive semidefinite matrix. This completes the proof of the lemma.

In order to prove the next lemma as a prelude to proving the convergence of the algorithm, two additional assumptions will be made. The first one says, in effect, that any point which is a local minimizer in an equality constrained subspace is unique locally.

Isolated minimizer assumption A₂. Let \bar{y} be any point feasible to the programming problem (1.1), (1.2) which satisfies the second order conditions necessary for the point to be a minimizer in the subspace of constraints equal to zero at \bar{y} . Then \bar{y} also satisfies the second order conditions sufficient for \bar{y} to be an isolated local minimizer in the equality constrained subspace.

The second order conditions, necessary ones, and sufficient ones, for an equality constrained problem are exactly those for the inequality constrained problem (see (1.4)–(1.10)) except that the multiplier values need not be nonnegative.

Without this assumption it is not possible to prove convergence of the two-part algorithm presented in this paper to a SOKTP. A different algorithm, with an autonomous move where all the terms necessary to satisfy the second order necessary conditions for an inequality constrained problem are used in the first

step and movement away from boundaries is allowed, can be shown (see [9]) to converge without the above assumption. That method, however, is not very efficient on the computer in solving problems. For safety's sake, one could include at intervals a move such as that described in [9] to guarantee convergence. However, if one is to assume that the situation where a point satisfying the second order necessary conditions in an equality constrained subspace is not an isolated minimizer is a rare occurrence, then the algorithm described herein is preferable.

Another nondegeneracy assumption is required for the convergence proof. This assumption rules out the possibility of a Lagrangian stationary point having multiplier values associated with binding constraints equal to zero.

Strict complementary slackness assumption A_3 . Let y be any point which satisfies (1.2) and for which there exists values $\{w_i(y)\}$ satisfying

$$\nabla f(y) = \sum_{i \in B(y)} \nabla g_i(y) w_i(y),$$

where

$$B(y) = \{i | g_i(y) = 0, i = 1, \dots, m\}.$$

Then

$$w_i(y) \neq 0 \quad \text{for } i \in B(y).$$

The next lemma shows that the directions chosen by the algorithm are such that in a neighborhood of a point, say \bar{x} which is not a SOKTP, the outcome after two iterations of the algorithm will yield a value of f less than $f(\bar{x})$ if a small distance is moved.

LEMMA 4. *Suppose \bar{x} is a cluster point of $\{x(k, 1)\}$. Let $\{x(k_j, 1)\}$ denote the subsequence converging to \bar{x} . Assume, as in the previous lemma, that the problem functions are three times continuously differentiable, and that the linear independence assumption holds. Assume further that assumptions A_2 and A_3 hold. Now, if for every $\varepsilon > 0$ there is a $\tau, 0 < \tau \leq \varepsilon$, such that for all k_j large,*

$$(3.11) \quad f[x^{k_j}(\tau)] \leq f(\bar{x}),$$

then \bar{x} is a SOKTP.

Proof. We can assume that ε is small enough so that only constraints binding at \bar{x} are considered in the analysis following. We can also assume that $(t^{k_j} + t^{k_j+1})$ is bounded below away from zero, otherwise the conclusion would follow from the previous lemma. Because of this we can assume that $x^{k_j}(\tau)$ defines some point actually traversed during iteration k_j or $k_j + 1$.

Now, if $\liminf_{j \rightarrow \infty} t^{k_j} = 0$, it follows from the arguments in the previous lemma that \bar{x} is a Lagrangian stationary point although the associated Lagrange multipliers need not be nonnegative. Assume therefore that $\liminf_{j \rightarrow \infty} t^{k_j} > 0$. Then ε can be considered small enough that for k_j large, $0 \leq \tau \leq t^{k_j}$ prevails in

the following analysis. Using (3.11), (3.1) and Taylor's theorem we have

$$(3.12) \quad f(\bar{x}) \leq f[x^{k_j}(\tau)] = f[x(k_j, 1)] + \left[\sum_{p=1}^{q(k_j, \tau)-1} \{b(k_j, p)t(k_j, p) + w(k_j, p)t^2(k, p)\} \right. \\ \left. + b(k_j, q(k_j, \tau))\beta(k_j, \tau) + w(k_j, q(k_j, \tau))\beta^2(k_j, \tau) \right] \nabla f(\eta(k_j, \tau)),$$

where $\eta(k_j, \tau)$ is a convex combination of $x^{k_j}(\tau)$ and $x(k_j, 1)$.

Because the number of variables and constraints is finite, and because of the rules for adding and subtracting from $r(k, p)$, and for including new constraints in the set of those considered to be binding and required to remain binding, we can assume, without loss of generality, that for k_j large enough, the number of steps $P(k_j)$ is the same, the order of the indices in the permutation sets $R(k_j, p)$, $C(k_j, p)$ is the same, for $p = 1, \dots, P(k_j)$, (for $k_j + 1$ also), and that the index $q(k_j, \tau)$ is the same. Call this constant value $q(\tau)$. Taking the limit (as $j \rightarrow \infty$),

$$(3.13) \quad f(\bar{x}) \leq f(\bar{x}) + \left[\sum_{p=1}^{q(\tau)-1} \{\tilde{b}(p)\tilde{t}(p) + \tilde{w}(p)\tilde{t}^2(p)\} + \tilde{b}(q(\tau))\tilde{\beta}(\tau) \right. \\ \left. + \tilde{w}(q(\tau))\tilde{\beta}^2(\tau) \right] \nabla f(\tilde{\eta}(\tau)),$$

where $\tilde{\eta}(\tau)$ is a convex combination of $\tilde{x}(\tau)$ and \bar{x} . (The fact that all the quantities in (3.12) are uniformly bounded allows us to extract converging subsequences.) Now,

$$\sum_{p=1}^{q(\tau)-1} \tilde{t}(p) + \tilde{\beta}(\tau) > 0.$$

Canceling $f(\bar{x})$ in (3.13), dividing by τ and taking the limit as $\tau \rightarrow 0$ yields

$$0 \leq \sum_{p=1}^{\bar{q}} \bar{\alpha}_p(\bar{b}(p)\nabla f(\bar{x})),$$

where $\bar{\alpha}_p \geq 0$, for all p , and $\sum_{p=1}^{\bar{q}} \bar{\alpha}_p = 1$, where each $\bar{b}(p)$ is a vector computed from quantities evaluated at \bar{x} using only information from constraints equal to zero at \bar{x} . The analysis used in the proof of the previous lemma can now be used to show that \bar{x} is a Lagrangian stationary point. Thus, since for both cases, when $\liminf_{j \rightarrow \infty} t^{k_j} = 0$ and when $\liminf_{j \rightarrow \infty} t^{k_j} > 0$, \bar{x} is a Lagrangian stationary point that portion of the proof is complete.

Next we will show that \bar{x} is a second order Lagrangian stationary point; i.e., it satisfies the second order necessary conditions for a local minimizer in an equality constrained subspace.

If \bar{b}_p for some p where $\bar{\alpha}_p > 0$ came from the computation of s_I using (2.13) or (2.14) an infinite number of times, it follows that \bar{x} is a second order Lagrangian stationary point. We need consider, then, only the case when s_I was computed from (2.15) an infinite number of times.

Now, inequality (3.12) could have been written

$$f(\bar{x}) \leq f[x^{k_j}(\tau)] = f[x(k_j, 1)] + \sum_{p=1}^{q(k_j, \tau)-1} f[x(k_j, p+1)] - f[x(k_j, p)] \\ + f[x^{k_j}(\tau)] - f[x(k_j, q(k_j, \tau))].$$

Expanding each of the differences in a two-term Taylor's series yields

$$(3.14) \quad f(\bar{x}) \leq f[x^{k_j}(\tau)] = f[x(k_j, 1)] + \sum_{p=1}^{q(k_j, \tau)-1} \{z(k_j, p)' \nabla f(k_j, p)\} \\ + z(k_j, \tau)' \nabla f(k_j, q(k_j, \tau)) \\ + \sum_{p=1}^{q(k_j, \tau)-1} \{z(k_j, p)' \nabla^2 f(\eta(k_j, p)) z(k_j, p)\} / 2 \\ + z(k_j, \tau)' \nabla^2 f(\eta(k_j, q(k_j, \tau))) z(k_j, \tau) / 2,$$

where

$$z(k_j, p) = b(k_j, p) t(k_j, p) + w(k_j, p) t^2(k_j, p), \quad p = 1, \dots, q(k_j, \tau) - 1, \\ z(k_j, \tau) = b(k_j, q(k_j, \tau)) \beta(k_j, \tau) + w(k_j, q(k_j, \tau)) \beta^2(k_j, \tau),$$

where each $\eta(k_j, p)$ is a convex combination of $x(k_j, p+1)$ and $x(k_j, p)$ and where $\eta(k_j, q(k_j, \tau))$ is a convex combination of $x(k_j, q(k_j, \tau))$ and $x^{k_j}(\tau)$.

Taking the limit in inequality (3.14) as $j \rightarrow \infty$ (and extracting appropriate subsequences) we have

$$(3.15) \quad f(\bar{x}) \leq f[\tilde{x}(\tau)] = f(\bar{x}) + \sum_{p=1}^{q(\tau)-1} \{\tilde{z}(p)' \nabla \tilde{f}(p)\} + \tilde{z}(\tau)' \nabla f(q(\tau)) \\ + \sum_{p=1}^{q(\tau)-1} \{\tilde{z}(p)' \nabla^2 f(\eta(p)) \tilde{z}(p) / 2\} + \tilde{z}(\tau)' \nabla^2 f(\eta(q(\tau))) \tilde{z}(\tau) / 2,$$

where

$$\tilde{z}(p) = (\tilde{b}(p) \tilde{t}(p) + \tilde{w}(p) \tilde{t}^2(p)), \quad p = 1, \dots, q(\tau), \\ \tilde{z}(\tau) = (\tilde{b}(q(\tau)) \tilde{\beta}(\tau) + \tilde{w}(q(\tau)) \tilde{\beta}^2(\tau))$$

and where each $\tilde{\eta}(p)$ is a convex combination of $\tilde{x}(p+1)$ and $\tilde{x}(p)$, and $\tilde{\eta}(q(\tau))$ is a convex combination of $\tilde{x}(q(\tau))$ and $\tilde{x}(\tau)$. Because of the rules for the computation of the arcs,

$$(3.16) \quad \tilde{b}(p) \nabla \tilde{f}(p) \leq 0, \quad \tilde{b}(q(\tau)) \nabla \tilde{f}(q(\tau)) \leq 0.$$

Then canceling $f(\bar{x})$ in (3.15), deleting the terms of (3.16) since they tend to reinforce the inequality, dividing by $\tau^2/2$, taking the limit as $\tau \rightarrow 0$ yields

$$0 \leq \sum_{p=1}^q \bar{\alpha}_p \bar{e}'_p \bar{H}(p) \bar{e}_p,$$

where $\bar{\alpha}_p \geq 0$ for all p , and $\sum_{p=1}^q \bar{\alpha}_p = 1$, and \bar{e}_p is an eigenvector of $\bar{H}(p)$ associated

with its minimum eigenvalue. Hence \bar{x} satisfies the second order conditions necessary for \bar{x} to be a local minimizer in the set of constraints equal to zero at \bar{x} .

The preceding arguments showed that any set of constraints corresponding to some $\bar{\alpha}_p > 0$ constituted a set for which the point \bar{x} was a second order Lagrangian stationary point. Thus, the limiting value $\bar{\alpha}_p$ must have been zero for any set of constraints which did not include all those for which the corresponding multiplier value at \bar{x} was unequal to zero. This means that the $t(k_j, p)$ associated with such a set of constraints must have approached zero as $j \rightarrow \infty$. Because of the assumption of strict complementarity slackness (A_3) only one index can be associated with the set of constraints whose limiting α_p value is nonzero, the index $P(k_j)$, the final step of each iteration. Then it follows that if $x(k_j, 1) \rightarrow \bar{x}$, so must $x(k_j, 2), \dots, x(k_j, P(k_j))$. It then follows from the isolated minimizer assumption A_2 that the distance that can be moved away from $x(k_j, P(k_j))$ and decrease the function value tends to zero. This implies that the point $x(k_j + 1, 1)$ tends to \bar{x} and that initially the set S^{k+1} contains the indices of all constraints binding at \bar{x} . We are now in a position to prove that all the multipliers associated with the Lagrangian stationary point are nonnegative.

If all the \bar{u}_i 's are not nonnegative, then, for j large enough, some index is deleted from the set S^{k+1} at the beginning of the $(k_j + 1)$ st iteration. The arguments above implied that $t^{k_j} \rightarrow 0$. Hence, the proof that \bar{x} was a Lagrangian stationary point for the indices in the set of those binding and required to remain binding at $x(k_j, P(k_j))$ apply here. Therefore, for some step between the start and the end of the $(k_j + 1)$ st iteration the constraint whose index was dropped will be re-encountered. All the $\bar{\alpha}_p$'s associated with sets of indices of constraints required to be binding are zero unless all indices of constraints binding at \bar{x} are included. It follows that the constraint index is picked up at points arbitrarily close to \bar{x} as $j \rightarrow \infty$. Thus, the analysis from (3.5) on can be repeated to show that all the multipliers associated with the Lagrangian stationary point must be nonnegative. This completes the proof that \bar{x} is a SOKTP.

THEOREM 1 (Convergence of the algorithm). *If the problem functions are three times continuously differentiable, if the linear independence assumption is satisfied, and if the strict complementary slackness and isolated minimizer assumption hold, then every cluster point of $\{x(k, 1)\}$ is a SOKTP.*

Proof. Suppose \bar{x} is a cluster point of $\{x(k, 1)\}$. Let $\{x(k, 1)\}$ denote also the subsequence converging to \bar{x} . Assume the contrary, i.e., that \bar{x} is not a SOKTP. From Lemma 3 it follows that

$$(3.17) \quad \liminf_{k \rightarrow \infty} \max_{0 \leq t \leq (t^k + t^{k+1})} \|x^k(t) - x^k(0)\| = \delta > 0.$$

Using the denial of Lemma 4, we know that there is an $\varepsilon > 0$ such that for every τ where $0 < \tau \leq \varepsilon$, for k large enough,

$$f[x^k(\tau)] < f(\bar{x}).$$

Let τ_1 be smaller than one-half the value for which δ in (3.17) is attained and small enough so that Lemma 4 applies. Then, eventually,

$$f[x^k(t^k + t^{k+1})] < f[x^k(\tau_1)] < f(\bar{x}).$$

But $f(\bar{x}) < f[x^k(t^k + t^{k+1})]$ by the fact that the $\{f[x(k, 1)]\}$ form a strictly decreasing sequence. This contradiction proves the theorem.

Modifications required to correct for the fact that Newton iterations are required to attain the boundary. In the proof of convergence, for the sake of simplicity, it was assumed that the arc movement exactly followed the boundary of the constraint region. In general, this is not the case. The following important lemma is required in proving that the method converges when the Newton method is used to attain the boundary. The lemma shows how close the Newton generated arc is to the original quadratic arc $a(k, p, t)$.

LEMMA 5. *If the problem functions are three times continuously differentiable, then*

$$(3.18) \quad \|\alpha(k, p, t) - a(k, p, t)\| = O(\|s_I(k, p)\|^3 t^3),$$

$$(3.19) \quad \|d\alpha(k, p, t)/dt - da(k, p, t)/dt\| = O(\|s_I(k, p)\|^2 t^2),$$

and

$$(3.20) \quad \|d^2\alpha(k, p, t)/dt^2 - d^2a(k, p, t)/dt^2\| = O(\|s_I(k, p)\|t).$$

These assumptions will not be proved here as they are a trivial consequence of the properties of Newton's method for solving nonlinear equations.

We are now in a position to show how the proofs of Lemmas 3 and 4, and Theorem 1 can be modified.

It follows from the linear independence assumption and the assumed care with which the dependent variables are chosen that $t_3(k, p)$ (for all k, p) is bounded below away from zero, i.e.,

$$t_3(k, p) \geq \varepsilon_3 > 0 \quad \text{for all } k, p.$$

In Lemma 3, statement (3.2) should be

$$\liminf_{k \rightarrow \infty} \max_{0 \leq t \leq (t^k + t^{k+1})} \|\alpha^k(t) - \alpha^k(0)\| = 0,$$

where $\alpha^k(t)$ is the result of the modified Newton method (2.20) applied to $x^k(t)$ defined in (3.1). Clearly for this subsequence converging to the infimum,

$$t(k, P(k)) < t_3(k, P(k)),$$

since by the above, $t_3(k, P(k)) > \varepsilon_3$, and because of the main hypothesis of Lemma 3 that $t(k, P(k)) \rightarrow 0$. Thus, the directional derivative in (3.3) is applied to

$$df[\alpha(k, P(k), t(k, P(k)))]/dt \geq 0.$$

Because of Lemma 5, equation (3.3) and all following equations must be modified by adding a term $O(\|s_I(k, P(k))\|^2 t(k, P(k))^2)$. Then the analysis goes through since the added term vanishes under the assumptions of the lemma. Similarly, the equation following equation (3.9) is modified by adding (using (3.20) of Lemma 5) a term $O(\|s_I(k, P(k))\| \xi^k)$. Again, the proof continues on since the added term vanishes.

For Lemma 4, inequality (3.11) should read

$$f[\alpha^{k_j}(\tau)] \geq f(\bar{x}).$$

Again, the possibility that $t(k, P(k)) = t_3(k, P(k))$ is precluded by the assumption there that ε is "small" and by the fact that $t_3(k, P(k)) \geq \varepsilon_3$.

Now, because of (3.18) of Lemma 5, (3.12) must be modified by a term $O(\|\tau\|^3)$. Then, finally, after division by τ , and taking the limit as $\tau \rightarrow 0$, the same conclusion holds. Similar analysis holds for (3.14) on; ultimate division by τ^2 , still allows for the error term to vanish as $\tau \rightarrow 0$ since it is of the order of the cube of τ .

In Theorem 1, all statements using $x^k(\tau)$ must be replaced by $\alpha^k(\tau)$, and the same conclusion holds since the two crucial lemmas are still valid.

4. Rate of convergence of the algorithm. If there were no constraints on the problem, the algorithm would reduce to the Newton method (as revised in [4]) for minimizing an unconstrained function. This is very important for the constrained case in accelerating the *rate* at which the points generated by the algorithm converge to an isolated local minimizer.

THEOREM 2 (Quadratic rate of convergence of the algorithm). *If the problem functions are three times differentiable, the linear independence assumption holds, the strict complementary slackness and isolated minimizer assumptions are satisfied, then, if the sequence $\{x(k, 1)\}$ has at least one cluster point, that point is the only cluster point, and it is an isolated local minimizer to problem (1.1), (1.2). Furthermore, if ε_1 used in Case 1 (see (2.13)) is "sufficiently small",¹ the rate of convergence of the sequence to that point is ultimately at least quadratic; i.e., there is an M , independent of k such that*

$$\|x(k+1, 1) - x^*\| \leq \|x(k, 1) - x^*\|^2 M.$$

(Here x^* denotes the limit point of $\{x(k, 1)\}$.)

Proof. From Theorem 1 it follows that every cluster point of the sequence $\{x(k, 1)\}$ satisfies the second order necessary conditions. Because of the isolated minimizer assumption, it follows that that cluster point also satisfies the second order sufficiency conditions and is, therefore, an isolated local minimizer to the problem. Because the algorithm always selects the nearest local minimizer in any search along the arc, it follows that once the sequence of points $\{x(k, 1)\}$ gets close to x^* , it can never generate points far enough away to allow for a second cluster point. It further follows from the arguments used to prove Lemma 4 that all the indices of constraints binding at x^* must be in the set S^k when k is large. The strict complementary slackness condition implies that all multipliers associated with the necessary conditions are strictly positive for binding constraints. Thus, no constraint index will be removed from the list of those binding and required to remain binding when k is large. Motion in the space of the constraints binding at x^* will be the only motion when k is large; hence, for k large, $P(k) = 1$, i.e., only one arc movement is made.

We will now argue that the equation used to generate the direction of the independent variables for large k is always (2.13). This follows from the fact that when the second order sufficiency conditions (1.10) hold along with the other

¹ The parameter ε_1 is used to prevent the possibility of nonconvergence to a constrained stationary point. If it is too large, the rate of convergence properties of the algorithm are destroyed. In practice, no difficulties are encountered in setting ε_1 to a "sufficiently small" value.

assumptions, the matrix

$$[-E^*(B^*)^{-1}, I]\nabla^2 L(x^*, u^*)[-(B^*)^{-1}E^*]$$

is positive definite, and the continuity of the second derivatives which implies that for k large, $H(k, 1)$ is therefore positive definite. For ε_1 "sufficiently small" then, eventually (2.13) will always be used.

The remainder of the proof will not be given in detail here. It is contained in [9]. A summary of the reasoning follows.

Under the assumptions stated above, the algorithm ultimately reduces to minimizing a uniformly strictly convex function in the space of independent variables. See (2.6) for an approximate form of the function to be minimized. It can be easily shown that without the adjustment for feasibility using Newton's modified method for solving nonlinear equations, and without the optimization along the arc, the new points would have an at least quadratic rate of convergence. Next, it must be shown that the feasibility adjustment (with a step size equal to 1), would not change the rate of convergence. Asymptotic arguments using (3.18) can be used to prove this. Finally, the use of the step size algorithm wherein the modified arc is used and the objective function minimizer along the arc found can be shown to generate points asymptotically close to those on the modified arc. Combining these steps shows that the rate of convergence to the isolated minimizer is at least quadratic.

5. Discussion. The ideas for direction generation presented in this paper are easily modified to accommodate different approaches for solving under-determined sets of linear equations. Using Rosen's gradient projection point of view [10] would yield different formulas, but the general approach would be just as valid. Equality constraints can be accommodated by ignoring the sign of the multiplier estimate which, according to theory, can be either plus or minus. Once a linear equality constraint is satisfied, after going through an appropriate feasibility phase, it will remain satisfied. Nonlinear equality constraints will, in general, be satisfied only after the Newton iterations generate the arc of feasible points.

The algorithm presented here has elements in common with many of those suggested in the past. The idea, for linearly constrained problems, of "eliminating" some variables from the problem and attempting to minimize the resulting unconstrained function is contained in Zoutendijk [13], and Wolfe [12]. In both cases the method of unconstrained minimization is a form of the method of steepest descent. In this paper the method reduces to the method of Newton for unconstrained minimization problems. In [1], [2], Abadie, et al., modified the Reduced gradient method to handle nonlinear constrained problems. The method they used to prescribe the direction of search of the independent variables was that of the method of conjugate gradients. To maintain feasibility they, as did Rosen [10], used a modified form of Newton's method for solving nonlinear equations. In the present paper, Newton's method is also used, but iterations are done from a quadratic arc which is closer to the constraint boundary than the straight line generated by the other methods.

Historically, there have been two different philosophies for deciding on how many constraint boundaries (for inequality constrained problems) to attempt to leave at the beginning of an iteration. The philosophy of Wolfe [11] is to leave the boundaries of all currently binding constraints with nonpositive multiplier estimates. That was the approach taken by the present author in [9]. Computer experience showed this to be very inefficient. Usually, most of these were immediately encountered again and the computations to reenter them into the "basis" matrix were time-consuming. The other approach, suggested by Rosen, is to minimize the objective function in a fixed subspace before attempting to leave the boundary of any constraint. This is clearly not the answer since a lot of unnecessary work can be done when one is far from the solution. In this paper, an attempt is made to leave the boundary of the constraint, whose multiplier estimate is most negative, only at the beginning of an iteration. It may be immediately reencountered. After the first step though, no boundaries are relinquished until a minimization of the objective function along some arc has occurred. As new constraint boundaries are encountered they are incorporated into the computations. This process, an anti-zig-zagging device, (see Zoutendijk [13] for the first mention of this problem) was developed by the author in [6]. It turns out to be a computationally efficient method although the original motivation was to avoid the problem of theoretical nonconvergence of the algorithm. (See Wolfe [11] for an example of how lack of anti-zig-zagging devices can result in premature convergence.)

By using arcs, and explicitly introducing the curvature of a nonlinear constraint as measured by its second derivative matrix, this method allows for movement nearer the boundary of the constraint region, making efforts to attain the boundary easier. The choice of the direction of search of the independent variables yields the attainment of ultimately an "at least" every-step quadratic rate of convergence.

Computational experience. An experimental code has been written to implement this algorithm for separable programming problems. A standard test problem to try out algorithms for solving nonlinearly constrained optimization problems is the "Shell dual" problem (number 2 in the study by Colville [3]). It was converted to an equivalent separable problem by the author and in that form is characterized by having 20 variables, five nontrivial nonlinear inequality constraints, five linear equality constraints, a nonlinear objective function, and lower bounds of zero on the first fifteen variables. Using the traditional starting point of $x_j = 0$, $j \neq 7$ and $x_7 = 60$, the time to attain seven place objective function solution accuracy was 4.4 seconds on the CDC 6400.

REFERENCES

- [1] J. ABADIE, J. CARPENTER AND C. HENSGEN, *Generalization of the Wolfe reduced gradient method to the case of nonlinear constraints*, Paper presented at the Joint European Meeting of the Econometric Society, The Institute of Management Science, Warsaw, 1966.
- [2] J. ABADIE AND J. CARPENTIER, *Generalization of the Wolfe reduced gradient method to the case of nonlinear constraints*, R. Fletcher, ed., Academic Press, New York, 1969, pp. 37-47.
- [3] A. R. COLVILLE, *A comparative study on nonlinear programming codes*, Rep. 320-2949, Revised 1970, IBM New York Scientific Center, 1968.
- [4] A. V. FIACCO AND G. P. MCCORMICK, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley, New York, 1968.

- [5] J. J. H. FORREST AND J. A. TOMLIN, *Updated triangular factors of the basis to maintain sparsity in the product form simplex method*, Math. Prog., 2 (1972), pp. 263–278.
- [6] G. P. McCORMICK, *Anti-zig-zagging by bending*, Management Sci., 15 (1969), pp. 315–320.
- [7] ———, *A second-order method for the linearly constrained nonlinear programming problem*, Nonlinear Programming, J. B. Rosen, O. L. Mangasarian and K. Ritter, eds., Academic Press, New York, 1970, pp. 207–243.
- [8] ———, *The variable reduction method for nonlinear programming*, Management Sci., 17 (1970), pp. 146–160.
- [9] ———, *An arc method for nonlinearly constrained programming problems*, MRC Tech. Summary Rep. 1073, Mathematics Research Center, U.S. Army, University of Wisconsin, Madison, Wisconsin, 1970.
- [10] J. B. ROSEN, *The gradient projection method for nonlinear programming. Part II: Nonlinear constraints*, J. Soc. Indust. Appl. Math., 9 (1961), pp. 514–532.
- [11] P. WOLFE, *On the convergence of gradient methods under descent*, IBM J. Res. Develop., 16 (1972), pp. 407–411.
- [12] ———, *Methods of nonlinear programming*, Recent Advances in Mathematical Programming, R. L. Graves and P. Wolfe, eds., McGraw-Hill, New York, pp. 67–86.
- [13] G. ZOUTENDIJK, *Methods of Feasible Directions*, Elsevier, Amsterdam and New York, 1960.

ON OPTIMAL STOCHASTIC CONTROL OF DISCRETE-TIME SYSTEMS IN HILBERT SPACE*

JERZY ZABCZYK†

Abstract. A general system described by a linear difference equation in a Hilbert space is considered. Three types of disturbances, control-dependent noise, state-dependent noise and purely additive noise, are taken into account. The cost function is assumed to be quadratic. The existence of an optimal stationary strategy and the uniqueness of the stationary measure related to this strategy are proved.

Special attention is paid to the related Riccati operator difference equation and the asymptotic behavior of the solution of such an equation is investigated. Under certain assumptions, the existence and uniqueness of the solution of the algebraic Riccati equation are proved, too.

1. Introduction. In this paper the existence of an optimal control for the problem with a given quadratic cost function is investigated. The system under consideration is described by a difference equation in a Hilbert space, and the disturbances present are of three types: control-dependent noise, state-dependent noise and purely additive noise. The problem is considered on the infinite interval $[0, 1, 2, \dots)$. Such models arise in the fields of sampled-data distributed parameter systems (see [14] where deterministic systems were studied). The analogous models in the case of continuous-time and finite-dimensional space were investigated in [11], [19], [7] and [8], whereas the *general* discrete-time systems, as far as we know, have never been examined even in the case of finite-dimensional space (see, for instance, [17], [13]).

Special attention is paid to the stationary optimal control law. The main difficulty is connected with the infinite dimension of state space. Two methods which overcome this difficulty are presented.

The existence of an optimal control law is closely connected with the behavior of the Riccati operator difference equation associated with the problem, and, therefore, the latter equation is studied in detail. In this connection, we introduce a notion of *stochastic observability* which plays a rather important role. This part of the study was inspired by Wonham's paper [18]. Another novelty of the paper is an application to these questions of a method based on Krasnosel'skiĭ's theory of I-concave functions. The same method (but in simpler case) was used in [20].

The discrete-time *matrix* Riccati equation has been extensively studied (the most recent references are [1], [2], [13], [12], [9]). Nevertheless, the author believes that some of the results contained in this paper are new when specialized to the finite-dimensional case.

2. Preliminaries.

2.1. Let $H, U, V, H_\zeta, H_\xi, H_\eta$ be separable Hilbert spaces, and let $\mathcal{E}, \mathcal{E}_U$ be the Banach spaces of all self-adjoint operators belonging to the spaces $L(H, H)$, $L(U, U)$ of all bounded linear operators transforming H into H and U into U . Let $\mathcal{K} \subset \mathcal{E}$, $\mathcal{K}_U \subset \mathcal{E}_U$ be the cones of all positive semidefinite operators. By Φ, D, Q, R, C we shall respectively denote operators belonging to $L(H, H)$,

* Received by the editors January 11, 1974, and in revised form August 18, 1974.

† Institute of Mathematics, Polish Academy of Sciences, Warsaw, Poland.

$L(U, H)$, $L(H, V)$, $L(U, U)$, $L(H_\eta, H)$ and by $A(\cdot, \cdot)$, $B(\cdot, \cdot)$ bilinear, continuous transformations: $A: H \times H_\zeta \rightarrow H$, $B: U \times H_\zeta \rightarrow H$. R is assumed to be an invertible element of \mathcal{K}_U .

The control system is given by

$$(2.1) \quad X_{n+1} = \Phi X_n + Du_n + A(X_n, \zeta_n) + B(u_n, \xi_n) + C\eta_n.$$

Here $X_0, \zeta_0, \zeta_1, \dots, \xi_0, \xi_1, \dots, \eta_0, \eta_1, \dots$ are mutually independent random variables with finite second moments taking their values in the appropriate Hilbert spaces. The random variables $\zeta_0, \zeta_1, \zeta_2, \dots$ (as well as ξ_0, ξ_1, \dots and η_0, η_1, \dots) are assumed to be identically distributed, and R_ζ (respectively, R_ξ and R_η) is their covariance operator. The covariance operator of a random variable ζ taking values in H_ζ is a linear operator $R_\zeta: H_\zeta \rightarrow H_\zeta$ such that $(R_\zeta x, y) = E(\zeta, x)(\zeta, y)$ for all $x, y \in H_\zeta$. Its trace $\text{tr } R_\zeta = \sum_{n=1}^{+\infty} (R_\zeta e_n, e_n)$, where $e_n, n = 1, 2, \dots$, is an orthonormal basis in H_ζ , is equal to $E|\zeta|^2$. Moreover, R_ζ is a nuclear (hence compact), self-adjoint operator. We assume $E(\zeta_n) = 0$, $E(\xi_n) = 0$, $E(\eta_n) = 0$, $n = 0, 1, \dots$.

LEMMA 2.1. Let $\{(\lambda_n, e_n); n = 1, 2, \dots\}$ be the sequence of eigenvalues and eigenvectors of the operator R_ζ , and let a transformation $\pi_1 \in L(\mathcal{E}, \mathcal{E})$ be given by the formula

$$\pi_1(K) = \sum_{n=1}^{+\infty} \lambda_n A_n^* K A_n, \quad K \in \mathcal{E},$$

where $A_n(\cdot) = A(\cdot, e_n)$, $n = 1, 2, \dots$. Then for all $K \in \mathcal{E}$ and $x, y \in H$,

$$(2.2) \quad E\{(KA(x, \zeta), A(y, \zeta))\} = (\pi_1(K)x, y).$$

Analogously, if $\{(\mu_n, f_n), n = 1, 2, \dots\}$ is the sequence of eigenvalues and eigenvectors of the operator R_ξ and a transformation $\pi_2 \in L(\mathcal{E}, \mathcal{E}_U)$ is given by the formula

$$\pi_2(K) = \sum_{n=1}^{+\infty} \mu_n B_n^* K B_n, \quad K \in \mathcal{E},$$

where $B_n(\cdot) = B(\cdot, e_n)$, $n = 1, 2, \dots$, then for all $K \in \mathcal{E}$ and $u, v \in U$,

$$(2.3) \quad E\{(KB(u, \xi), B(v, \xi))\} = (\pi_2(K)u, v).$$

Proof. Let us consider, for example, the formula (2.2). If we fix $x, y \in H$, then the transformations $A_x(\cdot) = A(x, \cdot)$, $A_y(\cdot) = A(y, \cdot)$ are linear and $E\{(KA(x, \zeta), A(y, \zeta))\} = E\{(A_y^* K A_x \zeta, \zeta)\}$. Since $E\{(A_y^* K A_x \zeta, \zeta)\} = \sum_{n=1}^{+\infty} (A_y^* K A_x R_\zeta e_n, e_n)$ and $R_\zeta e_n = \lambda_n e_n$, $n = 1, 2, \dots$, therefore,

$$\begin{aligned} E\{(KA(x, \zeta), A(y, \zeta))\} &= \sum_{n=1}^{+\infty} \lambda_n (K A_n x, A_n y) \\ &= \sum_{n=1}^{+\infty} (\lambda_n A_n^* K A_n x, y) = (\pi_1(K)x, y). \end{aligned}$$

This completes the proof of (2.2). The proof that (2.3) holds is analogous.

Remark 2.1. Let us notice that the operators π_1, π_2 are monotonic (see § 3) and satisfy the following condition.

HYPOTHESIS 1. *If $K_n \in \mathcal{K}$ and $K_n \uparrow K$ strongly (as $n \rightarrow +\infty$), then $\pi_1(K_n) \uparrow \pi_1(K)$ and $\pi_2(K_n) \uparrow \pi_2(K)$ strongly too.*

It is not difficult to construct a transformation $\pi \in L(\mathcal{E}, \mathcal{E})$ such that $\pi(\mathcal{K}) \subset \mathcal{K}$, but for some strongly convergent sequence $K_n \uparrow K$, $K_n \in \mathcal{K}$, the sequence $\{\pi(K_n); n = 1, 2, \dots\}$ does not converge strongly to $\pi(K)$.

For $W \in L(H, U)$, let us define the mappings $T_W: H \rightarrow H$, $G_W: \mathcal{E} \rightarrow \mathcal{E}$, $F_W: \mathcal{E} \rightarrow \mathcal{E}$ and $\mathcal{A}: \mathcal{K} \rightarrow \mathcal{K}$ by means of the formulas:

$$(2.4) \quad T_W = \Phi - DW, \quad G_W(K) = \pi_1(K) + W^* \pi_2(K) W + T_W^* K T_W,$$

$$(2.5) \quad F_W(K) = Q^* Q + W^* R W + G_W(K),$$

$$(2.6) \quad \mathcal{A}(K) = Q^* Q + \pi_1(K) + \Phi^* K (I + D [R + \pi_2(K)]^{-1} D^* K)^{-1}.$$

By virtue of Lemma 3.1 in [20], the transformation \mathcal{A} is well-defined and $\mathcal{A}(\mathcal{K}) \subset \mathcal{K}$.

2.2. Control on a finite interval $[0, 1, \dots, N-1]$. Let us consider any non-anticipating control $\{u_0, u_1, \dots, u_{N-1}\}$. We denote by $\{X_n^x; n = 0, 1, \dots, N-1\}$ the solution of (2.1) with initial distribution $\delta_{\{x\}}$. It is easy to prove the following theorem using the method of dynamic programming (see [13], [20]).

THEOREM 2.1. *For any nonanticipating control $\{u_0, u_1, \dots, u_{N-1}\}$ and $x \in H$, the solution $\{X_n^x, n = 0, \dots, N-1\}$ of (2.1) satisfies*

$$(2.7) \quad \begin{aligned} E \left\{ \sum_{n=0}^{N-1} [(QX_n^x, QX_n^x) + (Ru_n, u_n)] + (KX_N^x, X_N^x) \right\} \\ \geq (K_N x, x) + \sum_{n=0}^{N-1} \text{tr } C^* K_n C R_{\eta}, \end{aligned}$$

where $K_0 = K$ and $K_{n+1} = \mathcal{A}(K_n)$, $n = 0, 1, \dots, N-1$. Equality in (2.7) holds if $u_n = -W_{K_{N-n-1}} X_n^x$, where

$$W_K = [R + \pi_2(K) + D^* K D]^{-1} D^* K \Phi, \quad K \in \mathcal{K}.$$

The lemma below gives a probabilistic interpretation of the transformation F_W .

LEMMA 2.2. *Let $u_n = -W_{N-n-1} X_n^x$, where $W_n \in L(H, U)$, $n = 0, 1, \dots, N-1$. Then for all $x \in H$,*

$$(2.8) \quad \begin{aligned} E \left\{ \sum_{n=0}^{N-1} [(QX_n^x, QX_n^x) + (Ru_n, u_n)] + (KX_N^x, X_N^x) \right\} \\ = (F_{W_{N-1}} \cdots F_{W_1} (F_{W_0}(K)) x, x) + \text{tr } C^* K C R_{\eta} \\ + \sum_{n=0}^{N-2} \text{tr } C^* F_{W_n} \cdots F_{W_0}(K) C R_{\eta}. \end{aligned}$$

If $W_n = W$ for $n = 0, 1, \dots, N-1$ and $F_W(K) = K$, then (in particular)

$$(2.9) \quad \begin{aligned} E \left\{ \sum_{n=0}^{N-1} [(QX_n^x, QX_n^x) + (RWX_n^x, WX_n^x)] + (KX_N^x, X_N^x) \right\} \\ = (Kx, x) + N \text{tr } C^* K C R_{\eta}. \end{aligned}$$

Proof. The proof of Lemma 2.2 follows by induction and straightforward calculations and will be omitted.

3. The asymptotic behavior of the solution of the Riccati equation. The algebraic Riccati equation.

3.1. Let \mathcal{A} be the map given by (2.6). We say that a sequence $\{K_n; n = 0, 1, \dots\}$, $K_n \in \mathcal{K}$, satisfies the *Riccati operator difference equation* if and only if

$$K_{n+1} = \mathcal{A}(K_n), \quad n = 0, 1, \dots$$

Thus any solution of the Riccati equation has the form $K_n = \mathcal{A}^n(K_0)$, $n = 0, 1, \dots$. The equation

$$(3.1) \quad K = \mathcal{A}(K), \quad K \in \mathcal{K},$$

is called the *algebraic Riccati equation*.

In this section, we are interested in the asymptotic behavior of the iterates $\mathcal{A}^n(K_0)$, as well as in the problem of the existence and uniqueness of the solution of the equation (3.1).

3.2. Properties of the mapping \mathcal{A} . We start with the following lemma.

LEMMA 3.1. For all $W \in L(H, U)$ and $K \in \mathcal{K}$, we have

$$(3.2) \quad F_W(K) = F_{W_K}(K) + (W - W_K)^*[R + \pi_2(K) + D^*KD](W - W_K),$$

where

$$W_K = [R + \pi_2(K) + D^*KD]^{-1}D^*K\Phi.$$

In addition,

$$(3.3) \quad \mathcal{A}(K) = F_{W_K}(K).$$

Proof. Let us note that for $\bar{W} \in L(H, U)$,

$$\begin{aligned} F_W(K) &= Q^*Q + \pi_1(K) + W^*[R + \pi_2(K)]W + T_W^*KT_W \\ &= F_{\bar{W}}(K) + (W - \bar{W})^*[R + \pi_2(K) + D^*KD](W - \bar{W}) \\ &\quad + (W - \bar{W})^*[(R + \pi_2(K))\bar{W} - D^*KT_{\bar{W}}] \\ &\quad + [(R + \pi_2(K))\bar{W} - D^*KT_{\bar{W}}]^*(W - \bar{W}). \end{aligned}$$

But

$$(R + \pi_2(K))\bar{W} - D^*K(\Phi - D\bar{W}) = [R + \pi_2(K) + D^*KD]\bar{W} - D^*K\Phi.$$

Therefore, for $\bar{W} = W_K$, we have (3.2). To prove (3.3), let us introduce the notation $\bar{Q} = Q^*Q + \pi_1(K)$, $\bar{R} = R + \pi_2(K)$. Then

$$F_{W_K}(K) = \bar{Q} + W_K^*\bar{R}W_K + (\Phi - DW_K)^*K(\Phi - DW_K),$$

$$W_K = (I + \bar{R}^{-1}D^*KD)^{-1}\bar{R}^{-1}D^*K\Phi = \bar{R}^{-1}D^*K(I + D\bar{R}^{-1}D^*K)^{-1}\Phi,$$

and thus

$$\begin{aligned} \Phi - DW_K &= (I - D\bar{R}^{-1}D^*K(I + D\bar{R}^{-1}D^*K)^{-1})\Phi \\ &= (I + D\bar{R}^{-1}D^*K)^{-1}\Phi. \end{aligned}$$

Consequently,

$$\begin{aligned}
 F_{W_K}(K) &= \bar{Q} + [\bar{R}^{-1}D^*K(I + D\bar{R}^{-1}D^*K)^{-1}\Phi]^* \\
 &\quad \cdot \bar{R}[\bar{R}^{-1}D^*K(I + D\bar{R}^{-1}D^*K)^{-1}\Phi] \\
 &\quad + [(I + D\bar{R}^{-1}D^*K)^{-1}\Phi]^*K[(I + D\bar{R}^{-1}D^*K)^{-1}\Phi] \\
 &= \bar{Q} + \Phi^*(I + KD\bar{R}^{-1}D^*)^{-1}(KD\bar{R}^{-1}D^* + I)K(I + D\bar{R}^{-1}D^*K)^{-1}\Phi \\
 &= \mathcal{A}(K).
 \end{aligned}$$

COROLLARY 3.1. For any sequence $\{W_n; n = 0, 1, \dots\}$, $W_n \in L(H, U)$ and $K \in \mathcal{K}$, the following inequalities are satisfied:

$$(3.4) \quad F_{W_{N-1}} \cdots F_{W_1}(F_{W_0}(K)) \geq \mathcal{A}^N(K), \quad N = 1, 2, \dots$$

Proof. The inequality (3.4) is satisfied for $N = 1$ because of Lemma 3.1. Since $F_{W_N}(K) \geq F_{W_N}(L)$ for $K \geq L$, we obtain, by induction,

$$\begin{aligned}
 F_{W_N}(F_{W_{N-1}} \cdots F_{W_0}(K)) &\geq F_{W_N}(\mathcal{A}^N(K)) \geq F_{W_{\mathcal{A}^N(K)}}(\mathcal{A}^N(K)) \\
 &\geq \mathcal{A}(\mathcal{A}^N(K)) = \mathcal{A}^{N+1}(K).
 \end{aligned}$$

We recall (see [10]) that a transformation $\mathcal{A}: \mathcal{K} \rightarrow \mathcal{K}$ is said to be *monotonic* if and only if $\mathcal{A}(K) \geq \mathcal{A}(L)$ for all $K \geq L \geq 0$, and it is said to be *concave* if and only if $\mathcal{A}(\alpha K + \beta L) \geq \alpha \mathcal{A}(K) + \beta \mathcal{A}(L)$ for $\alpha, \beta \geq 0$, $\alpha + \beta = 1$ and $K, L \in \mathcal{K}$. A monotonic transformation $\mathcal{A}: \mathcal{K} \rightarrow \mathcal{K}$ is said to be *I-concave* if and only if

1° for all $K \in \mathcal{K}$, there exist positive numbers $\alpha = \alpha(K) > 0$, $\beta = \beta(K) > 0$ such that $\alpha I \leq \mathcal{A}(K) \leq \beta I$,

2° for every number $t \in (0, 1)$ and operator $K \in \mathcal{K}$, there exists $\varepsilon_0 > 0$ such that $\mathcal{A}(tK) \geq (t + \varepsilon_0)\mathcal{A}(K)$.

THEOREM 3.1. The transformation \mathcal{A} given by (2.6) is

(a) *monotonic, concave and continuous,*

(b) *if Hypothesis 1 is satisfied, and $K_n \uparrow K$, then also $\mathcal{A}(K_n) \uparrow \mathcal{A}(K)$ as $n \rightarrow +\infty$.*

(c) *if $\mathcal{A}^{n_0}(0)$ is an invertible operator, then the transformation $\mathcal{A}^{n_0}: \mathcal{K} \rightarrow \mathcal{K}$ is I-concave.*

Proof. (a) To prove monotonicity, suppose $K \geq L \geq 0$. Then, for any operator $W \in L(U, H)$,

$$F_W(K) \geq F_W(L).$$

Therefore,

$$\mathcal{A}(K) = F_{W_K}(K) \geq F_{W_K}(L) \geq F_{W_L}(L) = \mathcal{A}(L)$$

because of Lemma 3.1.

Suppose now that $\alpha, \beta \geq 0$, $\alpha + \beta = 1$ and $K, L \in \mathcal{K}$. We obtain from Lemma 3.1 that

$$\begin{aligned}
 \mathcal{A}(\alpha K + \beta L) &= F_{W_{\alpha K + \beta L}}(\alpha K + \beta L) = \alpha F_{W_{\alpha K + \beta L}}(K) + \beta F_{W_{\alpha K + \beta L}}(L) \\
 &\geq \alpha \mathcal{A}(K) + \beta \mathcal{A}(L).
 \end{aligned}$$

The continuity of \mathcal{A} follows from [4, Part I, Lemma 1, p. 584].

(b) Let $0 \leq K_n \uparrow K$ ($n \rightarrow +\infty$). Using [4, Part II, Thm. 2, p. 992], we see that $W_{K_n} \rightarrow W_K$ and $W_{K_n}^* \rightarrow W_K^*$ strongly. Since

$$\mathcal{A}(K_n) = Q^*Q + \pi_1(K_n) + W_{K_n}^*(R + \pi_2(K_n))W_{K_n} + (\Phi - DW_{K_n})^*K_n(\Phi - DW_{K_n}),$$

therefore, $\mathcal{A}(K_n) \rightarrow \mathcal{A}(K)$ strongly too.

(c) Since \mathcal{A} is a monotonic and concave transformation, therefore, the transformation \mathcal{A}^{n_0} is concave and monotonic too. If $0 < t < 1$, then

$$\mathcal{A}^{n_0}(tK + (1-t)0) \geq t\mathcal{A}^{n_0}(K) + (1-t)\mathcal{A}^{n_0}(0) \geq (t+\varepsilon)\mathcal{A}^{n_0}(K)$$

for an $\varepsilon > 0$ such that

$$\mathcal{A}^{n_0}(0) \geq \frac{\varepsilon}{(1-t)}\mathcal{A}^{n_0}(K).$$

3.3. Two stability lemmas. In the next sections, we shall need some facts concerning stability. They are formulated as Lemma 3.2 and Lemma 3.3.

Let $\Psi \in L(H, H)$. Then by $r(\Psi)$ we shall denote the spectral radius of Ψ : $r(\Psi) = \lim_n |\Psi^n|^{1/n}$. The pair (Φ, D) is said to be *stabilizable* if and only if there exists an operator $W \in L(H, U)$ such that $r(\Phi - DW) < 1$. The pair (Φ, Q) is said to be *detectable* if and only if there exists an operator $P \in L(V, H)$ such that $r(\Phi - PQ) < 1$.

LEMMA 3.2. If $\mathcal{B}: \mathcal{E} \rightarrow \mathcal{E}$ is a linear, monotonic operator such that, for some invertible $L \in \mathcal{K}$ the inequality

$$(3.5) \quad K \geq \mathcal{B}(K) + L$$

has a positive solution, then

$$r(\mathcal{B}) < 1.$$

Proof. Since $K = \mathcal{B}(K) + (K - \mathcal{B}(K))$ and $K - \mathcal{B}(K) \geq L$, we can assume that $K = \mathcal{B}(K) + L$. It is evident that the operator $\tilde{\mathcal{B}}: \mathcal{K} \rightarrow \mathcal{K}$, $\tilde{\mathcal{B}}(K) = \mathcal{B}(K) + L$ is monotonic and I-concave. Monotonicity is obvious, and concavity results from the observation that for $\varepsilon > 0$ sufficiently small and $\tilde{K} \in \mathcal{K}$,

$$\tilde{\mathcal{B}}(t\tilde{K}) - (t+\varepsilon)\tilde{\mathcal{B}}(\tilde{K}) = (1-t-\varepsilon)L - \varepsilon\mathcal{B}(\tilde{K}) \geq \frac{1}{2}(1-t-\varepsilon)L.$$

Therefore, because of Krasnosel'skiĭ's Theorem 6.7 of [10], the sequence of the successive approximations $\tilde{\mathcal{B}}^n(0)$; $n = 1, 2, \dots$, tends in the operator norm to the unique, nonnegative solution of (3.5). But $\tilde{\mathcal{B}}^n(0) = L + \mathcal{B}(L) + \dots + \mathcal{B}^{n-1}(L)$; thus, we conclude that $|\mathcal{B}^n(L)| \rightarrow 0$. On the other hand, for some $\delta > 0$, $I < \delta L$, this gives $|\mathcal{B}^n| = |\mathcal{B}^n(I)| \leq \delta |\mathcal{B}^n(L)| \rightarrow 0$. Thus $r(\mathcal{B}) < 1$.

Let us remark that this lemma is a generalization of the well-known Liapunov type theorem (see [3, p. 90] or [20]).

If for some invertible positive operator L , the equation

$$K = T^*KT + L, \quad K \in \mathcal{K},$$

has a nonnegative solution, then $r(T) < 1$.

LEMMA 3.3. Suppose there exist $K \in \mathcal{K}$, $W \in L(H, U)$ such that the following inequality is satisfied:

$$(3.6) \quad K \geq Q^*Q + W^*RW + T_W^*KT_W.$$

If the pair (Φ, Q) is detectable, then $r(T_W) < 1$.

Proof. It is sufficient to show (see [20]) that for an arbitrary $x \in H$,

$$\sum_{k=0}^{+\infty} |x_k|^2 < +\infty,$$

where $x_k = T_W^k x$, $k = 0, 1, \dots$. Because of the inequality (3.6), the sequences

$$\left\{ \sum_{k=0}^n T_W^{*k} Q^* Q T_W^k; n = 0, 1, \dots \right\}, \quad \left\{ \sum_{k=0}^n T_W^{*k} W^* R W T_W^k; n = 0, 1, \dots \right\}$$

are bounded from above by K . Therefore,

$$(3.7) \quad \sum_{k=0}^{+\infty} |Qx_k|^2 < +\infty, \quad \sum_{k=0}^{+\infty} |Wx_k|^2 < +\infty.$$

(We recall the general assumptions: R -invertible, positive operator.)

Let us remark that

$$\begin{aligned} x_{k+1} &= (\Phi - DW)T_W^k x = (\Phi - PQ)T_W^k x + PQT_W^k x - DW T_W^k x \\ &= (\Phi - PQ)x_k + u_k, \end{aligned}$$

where $u_k = PQx_k - DWx_k$. Using (3.7), we see that $\sum_{k=0}^{+\infty} |u_k|^2 < +\infty$. By virtue of Lemma 7.1 in [20], we conclude that $r(\Phi - DW) < 1$.

3.4. Main theorems. In this new section, we prove theorems which answer the questions formulated in 3.1.

LEMMA 3.4. Suppose Hypothesis 1 is satisfied. A nonnegative solution of (3.1) exists if and only if there exist operators $K \in \mathcal{K}$, $W \in L(H, U)$ such that

$$(3.8) \quad F_W(K) = K.$$

Proof. If $K \in \mathcal{K}$ is a solution of (3.1), then $K = \mathcal{A}(K) = F_{W_K}(K)$ (see Lemma 3.1). Let us suppose that $F_W(K) = K$, $K \in \mathcal{K}$. Then $K = F_W^n(K) \geq \mathcal{A}^n(K) \geq \mathcal{A}^n(0)$, $n = 1, 2, \dots$. Therefore, the monotonic sequence $\mathcal{A}^n(0)$, $n = 1, 2, \dots$, tends strongly to an operator $\bar{K} \in \mathcal{K}$. From this it follows (see Theorem 3.1, (2)) that $\bar{K} = \mathcal{A}(\bar{K})$.

THEOREM 3.2. Suppose there exists an operator W such that $r(G_W) < 1$.

Then

- (i) for any $K \in \mathcal{K}$, the sequence $\{\mathcal{A}^n(K); n = 1, \dots\}$ is bounded;
- (ii) if, in addition, Hypothesis 1 is satisfied, then there exists at least one non-negative solution of (3.1).

Proof. Since $r(G_W) < 1$, there exist numbers $M, \alpha > 0$, $0 < \alpha < 1$, such that $|G_W^n| \leq M\alpha^n$, $n = 1, 2, \dots$. But $F_W^n(K) = S + G_W(S) + \dots + G_W^{n-1}(S) + G_W^n(K)$ where $S = Q^*Q + W^*RW$. Thus

$$|F_W^n(K)| \leq M(1 - \alpha)^{-1}|S| + M|K|, \quad n = 1, 2, \dots.$$

From this and Corollary 3.1, we obtain that $|\mathcal{A}^n(K)| \leq |F_W^n(K)| \leq M(1 - \alpha)^{-1}|S|$

+ $M|K|$. This proves (i). The assumption $r(G_W) < 1$ implies that (3.8) has a non-negative solution. The existence of a nonnegative solution of (3.1) follows now from Lemma 3.4.

THEOREM 3.3. *Let us assume that there exists a solution $\bar{K} \geq 0$ of the equation (3.1).*

(i) *If the pair (Φ, Q) is detectable, then $r(T_{W_{\bar{K}}}) < 1$.*

(ii) *If for some n_0 , $\mathcal{A}^{n_0}(0)$ is an invertible operator, then $r(G_{W_{\bar{K}}}) < 1$, \bar{K} is the unique solution of (3.1) and $|\mathcal{A}^n(K) - \bar{K}| \rightarrow 0$ geometrically fast.*

Proof. (i). Since $\bar{K} = \mathcal{A}(\bar{K}) = F_{W_{\bar{K}}}(\bar{K})$, we see that

$$\bar{K} = (Q^*Q + W_{\bar{K}}^*RW_{\bar{K}} + T_{W_{\bar{K}}}^*\bar{K}T_{W_{\bar{K}}}) + (\pi_1(\bar{K}) + W_{\bar{K}}^*\pi_2(\bar{K})W_{\bar{K}}).$$

Using Lemma 3.3, we obtain $r(T_{W_{\bar{K}}}) < 1$.

(ii). Since $\mathcal{A}^{n_0}(0) \leq F_{W_{\bar{K}}}^{n_0}(0)$, then the operator $F_{W_{\bar{K}}}^{n_0}(0)$ is invertible. But

$$\bar{K} = F_{W_{\bar{K}}}(\bar{K}) = \dots = F_{W_{\bar{K}}}^{n_0}(\bar{K})$$

and

$$F_{W_{\bar{K}}}^{n_0}(\bar{K}) = F_{W_{\bar{K}}}^{n_0}(0) + G_{W_{\bar{K}}}^{n_0}(\bar{K}).$$

Thus $\bar{K} = F_{W_{\bar{K}}}^{n_0}(0) + G_{W_{\bar{K}}}^{n_0}(\bar{K})$. By virtue of Lemma 3.2, $r(G_{W_{\bar{K}}}^{n_0}) < 1$, and therefore, $r(G_{W_{\bar{K}}}) < 1$, too.

Let us define $\mathcal{W}(K) = \mathcal{A}^n(K)$, $K \in \mathcal{K}$. Then \mathcal{W} is an I-concave transformation (see Theorem 3.1) and consequently, has at most one solution (see [10, Thm. 6.3]). From this, it follows that (3.1) has in \mathcal{K} the unique solution \bar{K} . Let $0 < t < 1$, and define $K_0 = t\bar{K}$, $K_{n+1} = \mathcal{W}(K_n)$, $\rho_n = \sup \{s: s\bar{K} \leq K_n, n = 0, 1, 2, \dots\}$. Then

$$\begin{aligned} K_{n+1} = \mathcal{W}(K_n) &\geq \mathcal{W}(\rho_n \bar{K}) \geq \rho_n \mathcal{W}(\bar{K}) + (1 - \rho_n) \mathcal{W}(0) \\ &\geq \rho_n \bar{K} + (1 - \rho_n) \delta \bar{K}, \end{aligned}$$

where $\delta > 0$ is a number such that $\mathcal{W}(0) \geq \delta \bar{K}$. It follows from this that

$$\rho_{n+1} \geq \rho_n + (1 - \rho_n)\delta, \quad n = 0, 1, \dots$$

Consequently,

$$1 - \rho_{n+1} \leq (1 - \rho_n)(1 - \delta) \leq (1 - \delta)^{n+1}(1 - \rho_0).$$

The definition of the numbers ρ_0, ρ_1, \dots implies that

$$0 \leq \bar{K} - K_n \leq (1 - \rho_n)\bar{K} \leq (1 - \delta)^n(1 - t)\bar{K}.$$

Thus

$$|\bar{K} - K_n| \leq (1 - \delta)^n(1 - t)|\bar{K}|.$$

Analogously, if $u > 1$, $K^0 = u\bar{K}$, $K^{n+1} = \mathcal{W}(K^n)$, $n = 0, 1, \dots$, then

$$|\bar{K} - K^n| \leq (1 - \delta)^n(u - 1)|\bar{K}|.$$

Let K be any element of \mathcal{K} such that $|K| \leq R$. For some $u > 1$,

$$K_0 = \delta \bar{K} \leq \mathcal{W}(0) \leq \mathcal{W}(K) \leq u\bar{K} = K^0.$$

Therefore, there exists a number $M(R)$ such that

$$|\mathcal{W}^n(K) - \bar{K}| \leq M(R)(1 - \delta)^n, \quad n = 0, 1, \dots$$

If $|\mathcal{A}^i(K)| \leq R$ for $i = 0, 1, \dots, n_0 - 1$, then

$$|\mathcal{A}^n(K) - \bar{K}| \leq M(R)(1 - \delta)^{\lfloor n/n_0 \rfloor}.$$

Thus $|\mathcal{A}^n(K) - \bar{K}| \rightarrow 0$ geometrically fast uniformly on $\{K \in \mathcal{K}; |K| \leq R\}$.

Remark 3.1. The proof of the geometrically fast convergence is a modification of a proof given in [10, Thm. 6.7]. In fact, we proved the following result.

If $\mathcal{W}: \mathcal{K} \rightarrow \mathcal{K}$ is a monotonic and concave transformation such that $\mathcal{W}(0)$ is an invertible element of \mathcal{K} , then the equation

$$(3.9) \quad \mathcal{W}(K) = K, \quad K \in \mathcal{K},$$

has at most one solution. If \bar{K} is the solution of (3.9), then for all $K \in \mathcal{K}$, $|\mathcal{W}^n(K) - \bar{K}| \rightarrow 0$ geometrically fast.

The facts 1° and 2° below are true if $\dim H < +\infty$, but are not valid, in general, if $\dim H = +\infty$ (see Proposition 3.1).

1°. If the equation (3.1) has the unique solution \bar{K} , then the pair (Φ, Q) is detectable [12, Thm. 6].

2°. If the unique solution \bar{K} is an invertible operator, then for some n_0 , the operator $\mathcal{A}^{n_0}(0)$ is invertible, and $|\mathcal{A}^n(K) - \bar{K}| \rightarrow 0$ as $n \rightarrow +\infty$ for all $K \in \mathcal{K}$.

PROPOSITION 3.1. *Let $\dim H = +\infty$, and define*

$$(3.10) \quad \mathcal{A}(K) = Q + K(I + Q(I - Q)^{-1}K)^{-1}, \quad K \in \mathcal{K},$$

where for some α , $0 < \alpha < 1$, and all $x \in H$, $x \neq 0$, $0 < (Qx, x) < \alpha(x, x)$.

(i) *The equation (3.1) has exactly one nonnegative solution \bar{K} and $\bar{K} = I$. For any $K \in \mathcal{K}$, $\mathcal{A}^n(K) \rightarrow \bar{K}$ strongly.*

(ii) *If Q, K are compact operators, then $\mathcal{A}^n(K)$ is also a compact operator and $|\mathcal{A}^n(K) - \bar{K}| \not\rightarrow 0$.*

(iii) *The pairs $(I, Q^{1/2})$, $(I, Q^{1/2}(I - Q)^{1/2})$ are neither stabilizable nor detectable provided Q is compact.*

Proof. (i) Evidently the operator $\bar{K} = I$ satisfies (3.1) with \mathcal{A} defined by (3.10). Let us assume that $K \in \mathcal{K}$ is also the solution to (3.1) and that $KQ = QK$. By straightforward calculations, $(K - Q/2)^2 = (I - Q/2)^2$. Since $K \geq \lim_n \mathcal{A}^n(0) \geq Q$, therefore, $K = I$. From this, we conclude that $\bar{K} = I$ is the least solution of (3.1), and $\mathcal{A}^n(0) \rightarrow K$ strongly. Further, $W_I = Q^{1/2}(I - Q)^{1/2}$, $T_{W_I} = I - Q$. Thus for $K \in \mathcal{K}$,

$$\begin{aligned} F(K) &\triangleq F_{W_I}(K) = (Q + Q(I - Q)) + (I - Q)K(I - Q) \\ &= S + (1 - Q)K(1 - Q), \quad \text{where } S = Q + Q(I - Q). \end{aligned}$$

Therefore,

$$\begin{aligned} F^n(K) &= S + (I - Q)S(I - Q) + \dots + (I - Q)^{n-1}S(I - Q)^{n-1} \\ &\quad + (I - Q)^n K (I - Q)^n = (I - (I - Q)^{2n})(2I - Q)^{-1} \\ &\quad + (I - (I - Q)^{2n})(2I - Q)^{-1}(I - Q) \\ &\quad + (I - Q)^n K (I - Q)^n. \end{aligned}$$

By virtue of our assumptions, $(I - Q)^n \rightarrow 0$ strongly. It follows that $F^n(K) \rightarrow I$ strongly too. For all $K \in \mathcal{K}$ and $n = 1, 2, \dots$,

$$F^n(K) \geq \mathcal{A}^n(K) \geq \mathcal{A}^n(0).$$

Hence, $\mathcal{A}^n(K) \rightarrow I$ strongly, and $\bar{K} = I$ is the unique solution of (3.1) as required.

(ii) If Q, K are compact operators, then $\mathcal{A}^n(K)$ is a compact operator (as a sum of two compact operators). If $\dim H = +\infty$, then $\bar{K} = I$ is not compact and therefore, $\|\mathcal{A}^n(K) - I\| \not\rightarrow 0$.

(iii) Let us consider, for instance, the pair $(I, Q^{1/2})$. If $P \in L(H, H)$, then $PQ^{1/2}$ and $Q^{1/2}P$ are compact operators. Hence, $0 \in \sigma(PQ^{1/2}) \cap \sigma(Q^{1/2}P)$. Consequently, $1 \in \sigma(I - PQ^{1/2}) \cap \sigma(I - Q^{1/2}P)$ and $r(I - PQ^{1/2}) > 1$, $r(I - Q^{1/2}P) > 1$. This is the desired result.

In the context of the linear stochastic regulator problem, the assumption (which will be explained in 3.5) “ $\mathcal{A}^n(0)$ is an invertible operator” is quite natural but, if $\dim H = +\infty$, it is never satisfied in the problem of filtering. In the latter case, Q^*Q has to be a nuclear operator, and consequently, $\mathcal{A}^n(0)$ as a compact operator is not invertible. In the case where $\mathcal{A}^n(0)$ is not an invertible operator, Lemma 3.4 or Theorem 3.4 below may be useful.

THEOREM 3.4. *Let us suppose $\pi_1 \equiv 0$, $\pi_2 \equiv 0$.*

(i) *If the pair (Φ, Q) is detectable, then there exists at most one solution of (3.1).*

(ii) *If, in addition, (Φ, D) is stabilizable, then the sequence $\{\mathcal{A}^n(K); n = 1, 2, \dots\}$ converges strongly to the unique, nonnegative solution of (3.1) for $K \in \mathcal{K}$.*

Proof. (i) Suppose $K, L \in \mathcal{K}$ are two solutions of (3.1). We know by Lemma 3.1 that $K = F_{W_K}(K)$, $L = F_{W_L}(L)$ and $L - K = T_{W_L}^*(L - K)T_{W_L} + S$, where

$$S = (W_L - W_K)^*[R + D^*KD](W_L - W_K).$$

Lemma 3.3 implies $r(T_{W_L}) < 1$; therefore, $L - K = \sum_{n=0}^{+\infty} T_{W_L}^n(S)T_{W_L}^n \in \mathcal{K}$. Thus $L \geq K$. In the same way we obtain $K \geq L$ and finally $L = K$.

(ii) Suppose $\bar{K} \in \mathcal{K}$ is the unique solution of (3.1), and define the mapping $\bar{F}: \mathcal{E} \rightarrow \mathcal{E}$ as follows:

$$\bar{F}(K) = Q^*Q + W_K^*RW_K + T_{W_K}^*KT_{W_K} = F_{W_K}(K).$$

Since $r(T_{W_K}) < 1$, the iterates $\bar{F}^n(K)$ tend to the unique solution of the equation $\bar{F}(K) = K$, $K \in \mathcal{K}$. But $\bar{F}(\bar{K}) = \bar{K}$; therefore, the sequence $\{\bar{F}^n(K); n = 1, \dots\}$ tends strongly to \bar{K} . On the other hand, the bounded monotone sequence $\{\mathcal{A}^n(0); n = 1, \dots\}$ tends also to \bar{K} . Taking into account Corollary 3.1 and the monotonicity of \mathcal{A}^n , we see that $\mathcal{A}^n(0) \leq \mathcal{A}^n(K) \leq \bar{F}^n(K)$; therefore, $\mathcal{A}^n(K) \rightarrow \bar{K}$ strongly.

COROLLARY 3.2. *If $r(\Phi) < 1$, then (3.1) has exactly one solution (compare [20, Sec. 7]).*

3.5. Stochastic observability. Let us consider a stochastic system

$$X_{n+1} = \Phi X_n + A(X_n, \zeta_n),$$

$$Y_n = QX_n, \quad n = 0, 1, 2, \dots$$

We shall denote it as $(\Phi, \{A_n\}, Q)$, where the operators A_n were defined in Lemma 2.1. The system $(\Phi, \{A_n\}, Q)$ is said to be *stochastically observable* if and only if, for every initial state $x \in H$, $x \neq 0$, $\sup_n |QX_n^x| > 0$ with positive probability.

THEOREM 3.5. *The system $(\Phi, \{A_n\}, Q)$ is stochastically observable if and only if, for all $x \in H, x \neq 0$,*

$$\sup \{(F_0^N(0)x, x); N = 0, 1, \dots\} > 0,$$

and if and only if

$$\sup \{(\mathcal{A}^N(0)x, x); N = 0, 1, \dots\} > 0.$$

Proof. The first part of the theorem follows from the identity (see Lemma 2.2):

$$E \left\{ \sum_{n=0}^{N-1} |QX_n^x|^2 \right\} = (F_0^N(0)x, x).$$

To prove the latter part of the theorem, it is sufficient to prove that the equality $(\mathcal{A}^N(0)x, x) = 0$ implies $(F_0^N(0)x, x) = 0$. (The opposite implication is also true because $F_0^N(0) \geq \mathcal{A}^N(0)$). Let $(\mathcal{A}^N(0)x, x) = 0$. By virtue of Theorem 2.1 with $C = 0$,

$$E \left\{ \sum_{n=0}^{N-1} [(QX_n^x, QX_n^x) + (Ru_n, u_n)] \right\} = (\mathcal{A}^N(0)x, x) = 0,$$

where $\{u_n; n = 0, 1, \dots, N-1\}$ is the optimal control. This implies that the process $\{X_n^x; n = 0, 1, \dots\}$ satisfies

$$X_{n+1}^x = \Phi X_n^x - DWX_n^x + A(X_n^x, \zeta_n) - B(WX_n^x, \zeta_n),$$

with $W = 0$, and therefore,

$$(F_0^N(0)x, x) = (\mathcal{A}^N(0)x, x) = 0.$$

COROLLARY 3.3. *Let us assume that $\dim H = \bar{n} < +\infty$. It is not difficult to verify that the system $(\Phi, \{A_n\}, Q)$ is stochastically observable if and only if the matrix $F_0^{\bar{n}}(0)$ is invertible and if and only if the matrix $\mathcal{A}^{\bar{n}}(0)$ is invertible. Therefore, the condition “ $\mathcal{A}^n(0)$ is an invertible operator for some n ” is equivalent to the statement “ $\mathcal{A}^{\bar{n}}(0)$ is invertible.” Evidently if $A = 0$, then the stochastic observability is exactly the usual observability.*

COROLLARY 3.4. *Taking into account Corollary 3.1, Theorem 3.2 and Theorem 3.3, we obtain the following result. Let $\dim H = \bar{n} < +\infty$, and let $F_0^{\bar{n}}(0)$ be an invertible matrix. Then, there exists a solution to (3.1) if and only if there exists an operator $W \in L(U, H)$ such that $r(G_W) < 1$, (stabilizability).*

COROLLARY 3.5. *Let $\dim H \leq +\infty$. If K is a solution to (3.1) and the system is stochastically observable, then $(Kx, x) > 0$ for all $x \in H, x \neq 0$.*

4. Control on the infinite interval $[0, 1, 2, \dots]$.

4.1. Statement of the problem. Let φ be a Borel measurable mapping (see [16]) from H into U . The mapping φ generates the following Markov chain $\{X_n^\mu; n = 0, 1, \dots\}$,

$$(4.1) \quad X_{n+1}^\mu = \Phi X_n^\mu + D\varphi(X_n^\mu) + A(X_n^\mu, \zeta_n) + B(\varphi(X_n^\mu), \zeta_n) + C\eta_n$$

with initial distribution μ . If $\mu = \delta_{\{x\}}$, we write shortly $\{X_n^x; n = 0, 1, \dots\}$. The analogous solutions of (4.1) with $C = 0$ we denote, respectively, $\{Z_n^\mu; n = 0, 1, \dots\}$ and $\{Z_n^x; n = 0, 1, \dots\}$.

We distinguish two cases.

Case 1. $C = 0$. By an *admissible control law (strategy)* we mean the mapping φ such that for all $x \in H$,

$$(4.2) \quad \mathcal{W}^\varphi(x) \stackrel{\text{df}}{=} E \left\{ \sum_{n=0}^{+\infty} [(QZ_n^x, QZ_n^x) + (R\varphi(Z_n^x), \varphi(Z_n^x))] \right\} < +\infty.$$

The control problem is formulated as follows. Find an admissible control law $\bar{\varphi}$ such that

$$\mathcal{W}^{\bar{\varphi}}(x) \leq \mathcal{W}^{\varphi}(x)$$

for the arbitrary control law φ and arbitrary $x \in H$.

Case 2. $C \neq 0$. A Borel measurable mapping $\varphi: H \rightarrow U$ is said to be an *admissible control law* (in Case 2) if and only if there exists exactly one stationary measure for the Markov chain (4.1) and this measure has a finite second moment. Let us define the cost functional \mathcal{V}^{φ} as

$$(4.3) \quad \mathcal{V}^{\varphi} = \int_H [(Qx, Qx) + (R\varphi(x), \varphi(x))] \mu_{\varphi}(dx),$$

where μ_{φ} is the stationary measure corresponding to φ . An admissible control law $\bar{\varphi}$ is said to be optimal if and only if

$$\mathcal{V}^{\bar{\varphi}} \leq \mathcal{V}^{\varphi}.$$

4.2. Control on the infinite interval. Case 1. The theorem below gives a solution to the control problem in Case 1.

THEOREM 4.1. Assume $C = 0$.

(i) An admissible control law exists if and only if there exist operators $K \in \mathcal{K}$, $W \in L(H, U)$ such that (3.8) holds.

(ii) If (3.8) holds, then the function $\varphi = -W$ is an admissible control law and

$$E \left\{ \sum_{n=0}^{+\infty} [(QZ_n^x, QZ_n^x) + (RWZ_n^x, WZ_n^x)] \right\} = (\tilde{K}x, x),$$

where $\tilde{K} = \lim_n F_W^n(0)$ is the least nonnegative solution to (3.8). If, in addition, \tilde{K} is an invertible operator, then for arbitrary $x \in H$, $Z_n^x \rightarrow 0$ with probability one (as $n \rightarrow +\infty$).

(iii) The optimal control law ϕ is given by the formula $\phi = -W_{\bar{K}}$, where $\bar{K} = \lim_n \mathcal{A}^n(0)$ is the least solution of (3.1).

Proof. Let $F_W(K) = K$, $W \in L(H, U)$, $K \in \mathcal{K}$. By virtue of Lemma 2.2,

$$E \left\{ \sum_{n=0}^{N-1} [(QZ_n^x, QZ_n^x) + (RWZ_n^x, WZ_n^x)] \right\} \leq (Kx, x).$$

Thus

$$E \left\{ \sum_{n=0}^{+\infty} [(QZ_n^x, QZ_n^x) + (RWZ_n^x, WZ_n^x)] \right\} < +\infty,$$

and the function $\varphi = -W$ is an admissible control law.

(ii) Let $\tilde{K} = \lim_n F_W^n(0)$. Then $F_W(\tilde{K}) = (\tilde{K})$, and

$$\begin{aligned} E \left\{ \sum_{n=0}^{N-1} [(QZ_n^x, QZ_n^x) + (RWZ_n^x, WZ_n^x)] + (\tilde{K}Z_N^x, Z_N^x) \right\} \\ = (F_W^N(0)x, x) + E\{(\tilde{K}Z_N^x, Z_N^x)\} = (\tilde{K}x, x), \end{aligned}$$

because of Lemma 2.2. From this we conclude that $E\{\tilde{K}Z_N^x, Z_N^x\} \rightarrow 0$. Thus

$$E\left\{\sum_{n=0}^{+\infty} [(QZ_n^x, QZ_n^x) + (RWZ_n^x, WZ_n^x)]\right\} = (\tilde{K}x, x).$$

The inequality

$$E\{(\tilde{K}Z_1^x, Z_1^x)\} \leq (\tilde{K}Z_0^x, Z_0^x) = (\tilde{K}x, x),$$

valid for $x \in H$ implies that the sequence $\{(\tilde{K}Z_n^x, Z_n^x); n = 0, 1, \dots\}$ is a supermartingale. But $E\{(\tilde{K}Z_n^x, Z_n^x)\} \rightarrow 0$; therefore (see [15]), $(\tilde{K}Z_n^x, Z_n^x) \rightarrow 0$ with probability one. If \tilde{K} is an invertible operator, then $Z_n^x \rightarrow 0$ with probability one too, because for a suitable number $\gamma > 0$,

$$0 \leq |Z_n^x| \leq \gamma(\tilde{K}Z_n^x, Z_n^x), \quad n = 0, 1, \dots$$

(iii) By virtue of Theorem 2.1 for any nonanticipating control law and $x \in H$,

$$(4.4) \quad E\left\{\sum_{n=0}^{+\infty} [(QZ_n^x, QZ_n^x) + (Ru_n, u_n)]\right\} \geq \lim_n (\mathcal{A}^n(0)x, x).$$

But \bar{K} is the least solution of the equation $F_{W\bar{K}}(K) = K$, $K \in \mathcal{K}$; therefore,

$$(4.5) \quad E\left\{\sum_{n=0}^{+\infty} [(QZ_n^x, QZ_n^x) + (RW_{\bar{K}}Z_n^x, W_{\bar{K}}Z_n^x)]\right\} = (\bar{K}x, x)$$

because of the point (ii). The inequality (4.4) and the formula (4.5) show that the mapping $\bar{\varphi} = -W_{\bar{K}}$ is the optimal control law.

Remark 4.1. The above theorem shows that the equations (3.1), (3.8) which were carefully investigated in §3 play an essential role in the stochastic control.

4.3. Ergodic properties of a class of Markov chain. Before proceeding to Case 2, it is necessary to investigate properties of a Markov chain represented by the following difference equation:

$$(4.6) \quad X_{n+1} = FX_n + G(X_n, \zeta_n) + \eta_n, \quad n = 0, 1, \dots$$

Here, $F \in L(H, H)$, $G: H \times H_\zeta \rightarrow H$ is a bilinear transformation and the sequences ζ_n ; $n = 0, 1, \dots$, η_n ; $n = 0, \dots$ were described in §2.1. As in §4.1, we denote by $\{X_n^\mu; n = 0, 1, \dots\}$ the solution of (4.6) with initial distribution μ ; if $\mu = \delta_{\{x\}}$, then we write $\{X_n^x; n = 0, 1, \dots\}$. Let us remark that the process $\{Z_n^\mu; n = 0, 1, \dots\}$, $Z_n^\mu = X_n^\mu - X_n^0$, $n = 0, 1, \dots$, satisfies the equation

$$(4.7) \quad \begin{aligned} Z_{n+1} &= FZ_n + G(Z_n, \zeta_n), \\ Z_0 &= X_0. \end{aligned}$$

PROPOSITION 4.1. *If for all $x \in H$, $Z_n^x \rightarrow 0$ with probability one (as $n \rightarrow +\infty$), then the Markov chain (4.6) has at most one stationary distribution.*

Proof. Let f be a bounded, uniformly continuous function on E and let μ be a stationary measure for the Markov chain (4.6). Obviously, $E\{f(X_n^\mu)\} = \int_H f(x) \cdot \mu(dx)$ for $n = 0, 1, \dots$, and

$$\begin{aligned} |E\{f(X_n^\mu) - f(X_n^0)\}| &= |E\{f(X_n^0 + Z_n^\mu) - f(X_n^0)\}| \\ &\leq E\{|f(X_n^0 + Z_n^\mu) - f(X_n^0)|\}. \end{aligned}$$

Since f is uniformly continuous and $Z_n^\mu \rightarrow 0$ with probability one, therefore, $f(X_n^0 + Z_n^\mu) - f(X_n^0) \rightarrow 0$ with probability one, too. Consequently, $E\{f(X_n^0)\} \rightarrow \int_H f(x)\mu(dx)$ and therefore, $P^n(0, \cdot) \rightarrow \mu$ (weakly) (see [16, p. 40]) where $P(x, \cdot)$, $x \in H$, denotes the transition function of the Markov chain (4.6). This completes the proof.

COROLLARY 4.1. *If μ is a stationary measure of the Markov chain (4.6) and the assumption of Proposition 4.1 is satisfied, then for all initial distributions ν ,*

$$P^n(\nu, \cdot) \rightarrow \mu(\cdot)$$

(weakly); here, $P(\nu, \cdot) = \int_H \nu(dx)P(x, \cdot)$.

LEMMA 4.1.¹ *For any bounded and weakly continuous function $f: H \rightarrow R^1$, the function $Pf: H \rightarrow R^1$, $Pf(x) = \int_H f(y)P(x, dy)$ is a bounded and weakly continuous function too.*

Proof. Let us note that

$$Pf(x) = E\{f(Fx + G(x, \zeta_0) + \eta_0)\}.$$

For a fixed sample event ω and a sequence $\{x_n; n = 1, 2, \dots\}$ which converges weakly to x , the sequence $Fx_n + G(x_n, \zeta_0(\omega)) + \eta_0(\omega)$, $n = 1, 2, \dots$, tends to $Fx + G(x, \zeta_0(\omega)) + \eta_0(\omega)$ weakly too. This follows from the fact that linear, continuous operators are at the same time weakly continuous. Thus

$$f(Fx_n + G(x_n, \zeta_0) + \eta_0) \rightarrow f(Fx + G(x, \zeta_0) + \eta_0)$$

almost everywhere, and consequently, (f is a bounded function!)

$$\begin{aligned} Pf(x_n) &= E\{f(Fx_n + G(x_n, \zeta_0) + \eta_0)\} \\ &\rightarrow E\{f(Fx + G(x, \zeta_0) + \eta_0)\} = Pf(x). \end{aligned}$$

THEOREM 4.2. *If for some $x \in H$ the sequence $\{N^{-1}E\{\sum_{n=0}^{N-1} |X_n^x|^2\}; N = 1, 2, \dots\}$ is bounded, then there exists at least one stationary measure μ for the Markov chain (4.6).*

Proof. Let us define the sequence of measures $\{\nu_N; N = 1, 2, \dots\}$ by the formula $\nu_N(\cdot) = N^{-1} \sum_{n=0}^{N-1} P^n(x, \cdot)$. By our assumption, the sequence $\{\int_H |y|^2 \cdot \nu_N(dy); N = 1, 2, \dots\}$ is bounded from above, and therefore, for every $\varepsilon > 0$, there exists $r_\varepsilon > 0$ such that

$$\nu_N\{y: |y| > r_\varepsilon\} \leq \frac{\int_H |y|^2 \nu_N(dy)}{r_\varepsilon^2} \leq \varepsilon, \quad N = 1, 2, \dots$$

Since H is a separable Hilbert space, therefore, closed and weakly closed subsets of H generate the same σ -algebra of Borel measurable subsets of H . Moreover, the weak topology of the balls $B_\varepsilon = \{y: |y| \leq r_\varepsilon\}$ is metrizable. Using these remarks and the fact that the balls B_ε are weakly compact, we can find a subsequence $\{\nu_{N_k}; k = 1, 2, \dots\}$ and a probability measure ν such that for all bounded and weakly continuous real functions f on H ,

$$\int_H f(x) \nu_{N_k}(dx) \rightarrow \int_H f(x) \nu(dx) \quad \text{as } k \rightarrow +\infty.$$

¹ The idea to apply this lemma to the problems considered in the paper was suggested to me by S. Kwapien.

The classical method due to Krylov and Bogolubov (see [6, pp. 100–101]) implies that for all bounded weakly continuous functions f on H , $\int_H f(x) \nu(dx) = \int_H Pf(x) \cdot \nu(dx)$, or equivalently, $\int_H f(x) \nu(dx) = \int_H f(x) \nu P(dx)$. From this we obtain $\nu = \nu P$ (see [5, pp. 382–383]). Thus ν is a stationary measure for the Markov chain (4.6).

Remark 4.2. It is possible to prove Theorem 4.2 by a different method. Let $x = 0$. Then the theorem follows from the following observations:

(i) The sequence $\{R_n; n = 0, 1, 2, \dots\}$ of covariance operators of random variables X_2^0, X_1^0, \dots is monotonically increasing.

(ii) The sequence $\{\text{tr } R_n; n = 0, 1, \dots\}$ is bounded from above.

(iii) Families of measures $\{P^n(0, \cdot); n = 0, 1, \dots\}$, $\{N^{-1} \sum_{n=0}^{N-1} P^n(0, \cdot); N = 1, \dots\}$ are uniformly tight (see [16, p. 154]).

4.4. Control on the infinite interval. Case II.

THEOREM 4.3. Let us assume $C \neq 0$ and that $\mathcal{A}^{n_0}(0)$ is an invertible operator for some n_0 .

(i) If there exist $K \in \mathcal{K}$ and $W \in L(H, U)$ such that (3.8) holds, then the mapping $\varphi = -W$ is an admissible control law. In addition,

$$(4.8) \quad \mathcal{V}^\varphi = \text{tr } C^* K C R_\eta.$$

(ii) If \bar{K} is the solution to (3.1), then the mapping $\bar{\varphi} = -W_{\bar{K}}$ is the optimal control law and the minimal cost

$$\mathcal{V}^{\bar{\varphi}} = \text{tr } C \bar{K} C R.$$

Proof. (i) First, we prove that the sequence $\{N^{-1} E\{\sum_{n=0}^{N-1} |X_n^x|^2\}; N = 1, \dots\}$ is bounded from above. By virtue of Corollary 3.1, $F_W^{n_0}(0) \geq \mathcal{A}^{n_0}(0)$; therefore, $F_W^{n_0}(0)$ is an invertible operator. But

$$K = F_W^{n_0}(K) = F_W^{n_0}(0) + G_W^{n_0}(K),$$

and Lemma 3.2 implies $r(G_W^{n_0}) = (r(G_W^{n_0}))^{n_0} < 1$. Consequently, K is the unique solution of (3.8), and there exist numbers $M > 0$, $0 < \alpha < 1$ such that $|F_W^N(0) - K| \leq M\alpha^N$, $N = 0, 1, \dots$. It follows from Lemma 2.2 that

$$\begin{aligned} & E\left\{\sum_{n=0}^{N-1} [(QX_n^x, QX_n^x) + (RWX_n^x, WX_n^x)]\right\} + E\{(KX_N^x, X_N^x)\} \\ &= (F_W^N(0)x, x) + \sum_{n=0}^{N-1} \text{tr } C^* F_W^n(0) C R_\eta + E\{(KX_N^x, X_N^x)\} \\ &= (Kx, x) + N \text{tr } C^* K C R_\eta. \end{aligned}$$

Thus

$$\begin{aligned} E\{(KX_N^x, X_N^x)\} &= ((K - F_W^N(0))x, x) + \sum_{n=0}^{N-1} \text{tr } C^* (K - F_W^n(0)) C R_\eta \\ &\leq M|x|^2 + \frac{M}{1-\alpha} |C|^2 \text{tr } R_\eta. \end{aligned}$$

This inequality implies boundedness of the sequence $\{N^{-1} E\{\sum_{n=0}^{N-1} |X_n^x|^2\}; N = 1, 2, \dots\}$ because K is an invertible operator.

Let us define $F = \Phi - DW$, $G(x, (\zeta, \xi)) = A(x, \zeta) - B(Wx, \xi)$ for $x \in H$, $(\zeta, \xi) \in H_\zeta \times H_\xi$ and ζ_n as (ζ_n, ξ_n) , η_n as $C\eta_n$. Applying Theorem 4.2, we obtain that there exists a stationary measure μ_φ for the Markov chain (4.1). This stationary measure is unique. Really, K is the unique and invertible solution of (3.8). It follows from Theorem 4.1 (ii) that for all $x \in H$, $Z_n^x \rightarrow 0$ with probability one, and Proposition 4.1 implies that the stationary measure for (4.6) is unique. Since for some \bar{M} and $N = 1, 2, \dots$,

$$\int_H |y|^2 P^N(x, dy) = E\{|X_N^x|^2\} \leq \bar{M}$$

and $P^N(x, \cdot) \rightarrow \mu_\varphi$ weakly (see Corollary 4.1), therefore, $\int_H |y|^2 \mu_\varphi(dy) < +\infty$ and the measure μ_φ has finite second moment. To prove (4.8), we start with the formula

$$\begin{aligned} E\left\{\sum_{n=0}^{N-1} [(QX_n^x, QX_n^x) + (RWX_n^x, WX_n^x)]\right\} \\ = (F_W^N(0)x, x) + \sum_{n=0}^{N-1} \text{tr } C^* F_W^n(0) C R_\eta. \end{aligned}$$

From it we have that

$$\begin{aligned} \int_H [(Qx, Qx) + (RWx, Wx)] \mu_\varphi(dx) \\ = \frac{1}{N} \int_H (F_W^N(0)x, x) \mu_\varphi(dx) + \frac{1}{N} \sum_{n=0}^{N-1} \text{tr } C^* F_W^n(0) C R_\eta. \end{aligned}$$

Since

$$\frac{1}{N} \int_H (F_W^N(0)x, x) \mu_\varphi(dx) \leq \frac{1}{N} \int_H (Kx, x) \mu_\varphi(dx) \rightarrow 0,$$

therefore,

$$\mathcal{V}^\varphi = \lim_N \frac{1}{N} \sum_{n=0}^{N-1} \text{tr } C^* F_W^n(0) C R_\eta = \text{tr } C^* K C R_\eta.$$

(ii) For any admissible control law φ we have that, (see Thm. 2.1),

$$\begin{aligned} \int_H [(Qx, Qx) + (R\varphi(x), \varphi(x))] \mu_\varphi(dx) \\ \geq \frac{1}{N} \int_H (\mathcal{A}^N(0)x, x) \mu_\varphi(dx) + \frac{1}{N} \sum_{n=0}^{N-1} \text{tr } C^* \mathcal{A}^n(0) C R_\eta, \end{aligned}$$

and consequently,

$$\mathcal{V}^\varphi \geq \text{tr } C^* \bar{K} C R_\eta = \mathcal{V}^{\bar{\varphi}}.$$

4.5. Generalization. In many applications, it is necessary to consider more general systems and more general cost functions than those given by (2.1) and

(4.3). Let, for instance, the control system be represented by the equation

$$(4.9) \quad \begin{aligned} X_{n+1} &= \Phi X_n + Du_n + A(X_n, \zeta_n) + B(u_n, \xi_n) + C\eta_n + d \\ u_n &= \varphi(X_n), \end{aligned} \quad n = 0, 1, \dots,$$

and let the cost function have the form

$$\int_H [(Q(x - q), (x - q)) + (R(\varphi(x) - r), \varphi(x) - r)] \mu_\varphi(dx).$$

In this case, the optimal control law is no longer a linear function of the state, but it is of the form $\varphi = -Wx + w$. We shall restrict ourselves to the proposition below, which shows that under the same assumption as in Theorem 4.3, the mapping $\varphi = -Wx + w$ is an admissible control law for any $w \in H$. The proof of this proposition is analogous to that of Theorem 4.3 and will be omitted.

PROPOSITION 4.2. *Let the control system be given by (4.9). Let us assume that $\mathcal{A}^n(0)$ is an invertible operator for some n and let $F_W(K) = K$ for some $K \in \mathcal{K}$, $W \in L(H, U)$. Then the mapping $\varphi = -Wx + w$ is an admissible control law for all $w \in H$.*

4.6. Applications. Model (2.1) can be applied when we want to control the solution of a parabolic equation on the boundary $\partial\Omega$ of a region Ω . To see this, let H, U be Hilbert spaces consisting of functions defined respectively on Ω and $\partial\Omega$. Let $\{T_t; t \geq 0\}$ be a strongly continuous semigroup in H such that, roughly speaking, for $t > 0$, $T_t = 0$ on the boundary $\partial\Omega$, and let A be its infinitesimal operator. If $\tilde{D} \in L(U, H)$ and for all $u \in U$, $\tilde{D}u \in \mathcal{D}(A)$ and satisfies $A\tilde{D}u = 0$ with the boundary condition u , then the function $z_t = T_t(x_0 - \tilde{D}u) + \tilde{D}u$, $t > 0$, is the solution of the equation

$$\frac{dz_t}{dt} = Az_t$$

with boundary conditions: $z_t \rightarrow x_0$, if $t \downarrow 0$ and $z_t = u$ on the boundary $\partial\Omega$ for $t > 0$. Suppose that a controller chooses control functions u_0, u_1, \dots at moments $0, h, 2h, \dots$; then, for $t \in (nh, (n+1)h]$,

$$z_t = T_{t-nh}(z_{nh} - \tilde{D}u_n) + \tilde{D}u_n, \quad n = 0, 1, \dots$$

Let us define $X_n = Z_{nh}$, and suppose that the control functions u_0, u_1, \dots are "linearly" disturbed; then, for example,

$$X_{n+1} = T_h X_n + D(u_n + \zeta_n)$$

or

$$X_{n+1} = T_h X_n + DB(u_n, \xi_n), \quad n = 0, 1, 2, \dots$$

More generally, if a control system is represented by a stochastic equation with partial derivatives, and boundary conditions are described by a stochastic differential equation or by a stochastic equation with partial derivatives, then after introducing sampling we obtain (2.1).

REFERENCES

- [1] R. S. BUCY, *Linear and non-linear filtering*, Proc. IEEE, 58 (1970), pp. 854–864.
- [2] ———, *A priori bounds for the Riccati equation*, Proc. 6th Berkeley Symp., vol. 3, 1971, pp. 645–656.
- [3] I. DALECKIĬ AND M. KREIN, *Stability of Solutions of Differential Equations in Banach Space*, Izdat. Nauka, Moscow, 1970. (In Russian.)
- [4] N. DUNFORD AND T. SCHWARTZ, *Linear Operators*, Interscience, New York; Part I, 1958; Part II, 1963.
- [5] I. I. GICHMAN AND A. W. SKOROCHOD, *Theory of Random Processes*, vol. 1, Izdat. Nauka, Moscow, 1971. (In Russian.)
- [6] R. Z. HASMINSKI, *Stability of Differential Equations Disturbed by Stochastic Processes*, Izdat. Nauka, Moscow, 1969. (In Russian.)
- [7] U. G. HAUSSMAN, *Optimal stationary control with state and control dependent noise*, this Journal, 9 (1971), pp. 184–198.
- [8] ———, *Stability of linear systems with control dependent noise*, this Journal, 11 (1973), pp. 382–394.
- [9] G. A. HEWER, *Analysis of a discrete matrix Riccati equation of linear control and Kalman filtering*, J. Math. Anal. Appl., 42 (1973), pp. 226–236.
- [10] M. A. KRASNOSEĬSKIĬ, *Positive Solutions of Operator Equations*, P. Noordhoff, Groningen, the Netherlands, 1964.
- [11] N. N. KRASOVSKIĬ, *Stabilization of systems in which noise is dependent on the value of control signal*, Engrg. Cybernetics, 2 (1965), pp. 94–102.
- [12] V. KUČERA, *The discrete Riccati equation of optimal control*, Kybernetika, 8 (1972), pp. 430–447.
- [13] H. KUSHNER, *Introduction to Stochastic Control*, Holt, Rinehart and Winston, New York, 1971.
- [14] K. Y. LEE, S. CHOW AND R. BARR, *On the control of discrete-time distributed parameter systems*, this Journal, 10 (1972), pp. 361–376.
- [15] I. NEVEU, *Bases Mathématiques du Calcul des Probabilités*, Masson, Paris, 1964.
- [16] K. R. PARATHASARATHY, *Probability Measures on Metric Spaces*, Academic Press, New York, 1967.
- [17] H. W. SORENSON, *Controllability and observability of linear stochastic, time-discrete control systems*, Advances in Control Systems, vol. 6, Academic Press, New York, 1968, pp. 95–156.
- [18] W. M. WONHAM, *On a matrix Riccati equation of stochastic control*, this Journal, 6 (1968), pp. 681–697.
- [19] ———, *Random differential equations in control theory*, Probabilistic Methods in Applied Mathematics, vol. 2, Academic Press, New York, 1970, pp. 131–220.
- [20] J. ZABCZYK, *Remarks on the control of discrete-time distributed parameter systems*, this Journal, 12 (1974), pp. 721–735.

PERIODICITY, DETECTABILITY AND THE MATRIX RICCATI EQUATION*

G. A. HEWER†

Abstract. This paper discusses the periodic solution of matrix Riccati differential equations with periodic coefficients. Such equations arise in linear filtering and control and in many other applications. The principal result: the existence of a periodic solution is equivalent to detectability and stabilizability of certain coefficient pairs. This result generalizes the Kalman–Wonham–Kucera theorem for algebraic Riccati equations. Among the numerous preliminaries is a discussion, apparently new, of detectability for linear periodic control systems. Another important result, for a linear matrix differential equation, is the equivalence of a bounded solution, an exponentially stable solution and a periodic solution. Finally, the periodic solution is shown to be an equilibrium solution in the sense of Kalman.

1. Introduction. Consider the matrix Riccati differential equation

$$(1.1) \quad \dot{R}(t) + \psi(R(t), A(t), H(t)) - R(t)B(t)B(t)'R(t) = 0,$$

$-\infty < t < \infty$, subject to the terminal condition $R(0) = R \geq 0$, where $\psi(R(t), A(t), H(t)) = A(t)'R(t) + R(t)A(t) + H(t)'H(t)$. The real-valued matrices $A(t)$, $B(t)$, $H(t)$ are defined, periodic with real period ω ($\omega > 0$) and of dimension $n \times n$, $n \times m$, $r \times n$, respectively. Here $A(t)'$ is the transpose of the matrix $A(t)$. For any two square symmetric matrices M , N , the notation $M > N$ (\geq) means that the matrix $M - N$ is positive (semi) definite [3]. While the products $B(t)B(t)'$ and $H(t)'H(t)$ can be positive semidefinite, we assume that they are not identically zero on any interval of positive length.

Since the coefficients in (1.1) are continuous and periodic, the general solution $R(t, R_0)$, which satisfies $R(0, R_0) = R_0 \geq 0$, exists on some nonempty open interval containing the origin [30, p. 10].

DEFINITION 1.2. The solution $R(t, R_0)$ of (1.1) is periodic with real period ω ($\omega > 0$) if $R(t, R_0)$ ($R_0 \neq 0$) is defined and $R(t + \omega, R_0) = R(t, R_0)$ for $-\infty < t < \infty$.

The main purpose of this paper is to prove the following theorem, whose undefined terms are defined subsequently.

THEOREM 1.3. *A necessary and sufficient condition for the existence of one and only one positive semidefinite periodic solution $R(t, R_0)$ of (1.1) such that every solution of the ordinary differential equation*

$$(1.4) \quad \dot{x} = (A(t) - B(t)B(t)'R(t, R_0))x$$

is asymptotically stable is that $(A(t), B(t))$ is stabilizable and $(H(t), A(t))$ is detectable.

Theorem 1.3 is a generalization of the following theorem.

THEOREM 1.5. *Let A , B and H be constant matrices. A necessary and sufficient condition for the existence of one and only one positive semidefinite solution P of the*

* Received by the editors February 22, 1974, and in revised form September 30, 1974.

† Michelson Laboratories, Naval Weapons Center, China Lake, California 93555.

algebraic Riccati equation

$$(1.6) \quad \psi(P, A, H) - PBB'P = 0$$

such that the eigenvalues of the matrix $(A - BB'P)$ have real parts less than zero is that (A, B) is stabilizable and (H, A) is detectable.

The necessity of stabilizability and detectability was first obtained by Kucera [20], while the sufficiency part with stabilizability and detectability replaced by controllability and observability was first established by Kalman [14]. Reid [30] and Coppel [8] discuss the importance of controllability and the classical concept of normality in studying the matrix Riccati equation. The introduction of stabilizability and detectability is due to Wonham [34], who subsequently used these concepts in weakening Kalman's sufficiency part in Theorem 1.5.

As shown in §4, the periodic solution $R(t, R_0)$ of Theorem 1.3 is also an equilibrium solution, in the sense that every general solution of (1.1) converges to $R(t, R_0)$ as $t \rightarrow -\infty$. Moreover, the hypotheses in Theorem 1.3 for the existence of this equilibrium solution are weaker than those given by Kalman [14], who was the first to establish the existence of formal equilibrium solutions of (1.1). When (1.1) is an autonomous system (A, B, H are constant matrices), Kalman [14] proved that the equilibrium solution of (1.6) is a bona fide equilibrium solution. Since constant matrices are always periodic, Theorem 1.3 clearly subsumes Theorem 1.5.

Often the matrix Riccati equation can be associated with the $2n$ -dimensional Hamiltonian system

$$(1.7) \quad \begin{aligned} \dot{V} &= A(t)V - B(t)B(t)'U, \\ \dot{U} &= -H(t)H(t)'V - A(t)'U \end{aligned}$$

in the following manner. If $(V(t), U(t))$ is a solution of (1.7) such that $V^{-1}(t)$ exists ($V^{-1}(t)$ denotes the inverse of the matrix $V(t)$) on some interval, then the matrix $R(t) = U(t)V^{-1}(t)$ is a solution of (1.1) on that interval. Hence, any criterion which insures that the product $U(t)V^{-1}(t)$ is periodic and defined establishes the existence of a periodic solution of (1.1). The converse, however, could be false, for the product $U(t)V^{-1}(t)$ can be periodic with period ω , although the factors do not have period ω . Further discussion of the relation between (1.1) and (1.7) is beyond the scope of this paper, and the interested reader is referred to Reid [29] or Coppel [8].

Besides the implications of Theorem 1.3 for Hamiltonian systems, there are at least two other immediate applications of this theorem. The first application is to Fredholm integral equations. Schumitzky [31] has shown that the solution of every matrix Riccati equation can be generated by the resolvent of a certain Fredholm integral operator, and conversely, this resolvent can be determined from the corresponding Riccati solution. The second application is to linear filtering (Bucy et al. [7]) and linear quadratic control (Kalman et al. [17]). In this context, the existence of a terminal matrix R_0 such that $R(t, R_0)$ is periodic means that the coefficient matrix in the "closed loop" system (1.4) is also periodic. This is not only an appealing result, but is useful in applying stability or perturbation results, because the hypothesis of periodic coefficients often implies stronger

conclusions in Liapunov stability theorems (Yoshizawa [35, p. 45]). Moreover, the stabilizability and detectability hypotheses in Theorem 1.3 are considerably weaker than the usual conditions in Bucy [6] for the closed loop system to be asymptotically stable. Recently, periodic controls have been discussed by L. Markus [24] as a way of controlling the amplitude of limit cycles; the control of limit cycles in chemical systems has been discussed by S. Bittanti [5] and M. Fjeld [9]. It is tempting to infer from these articles that Theorem 1.3, when applied with linear quadratic control theory, could be used for the stability enhancement of limit cycles.

Another consequence of Theorem 1.3 in filtering and control is that the periodic equilibrium solution of (1.1) can play the same role numerically that the solution of the algebraic Riccati equation plays in the autonomous case. In other words, the period solution can be computed over one period "off line" and used as a suboptimal control just as the solution of (1.6) is often used for suboptimal control. As Kleinman et al. [19] demonstrate, the use of an equilibrium solution as a suboptimal control is often justified by comparing the computational advantages with the "extra cost" of using a suboptimal control.

Consider the linear dynamical control system

$$(1.8) \quad \dot{x} = A(t)x + B(t)u,$$

with the observation equation

$$(1.9) \quad y = H(t)x,$$

where x is an n -dimensional state vector, u is an m -dimensional control vector and y is an r -dimensional output vector. In § 2 a new characterization (Theorem 2.13) of controllability and observability for (1.8) and (1.9) is established via Floquet theory. This result is an extension to periodic systems of a theorem for autonomous systems due to Hautus [11]. Also a proof that the intervals of controllability and periodicity always coincide (whenever they are both defined) is given. This extends a theorem of Kalman [17, p. 35], who states (without proof) that the interval of controllability is no greater than the interval of periodicity. The importance of knowing that these two intervals are exactly equal is clearly illustrated in proving Theorem 1.3 in § 4.

In § 3, we consider the following linear matrix differential equation

$$(1.10) \quad \dot{K}(t) + \psi(K(t), A(t), H(t)) = 0, \quad -\infty < t < \infty,$$

with general solution $K(0, K_0) = K_0 \geq 0$, where $\psi(\cdot, \cdot, \cdot)$ is the matrix defined in (1.1). The concept of detectability for (1.8)–(1.9) is introduced in this section and is a crucial hypothesis in Theorem 3.7. This theorem establishes the equivalence of the following conditions:

- (i) the fundamental matrix solution of (1.8) with no control term is exponentially stable;
- (ii) equation (1.10) has a periodic solution;
- (iii) equation (1.10) has at least one bounded solution.

For ordinary vector linear differential equations, the equivalence of (ii) and (iii) is due to Massera [26]. This equivalence for (1.10) can be inferred from his results by the use of Kronecker sums [3] (an artifice that transforms (1.10) into a system

of $2n$ -dimensional linear ordinary differential equations). However, the detectability hypothesis is essential in proving the equivalence of (ii) and (i). This connection differs from the standard result for reasons which are given before the statement of Theorem 3.7. Another key ingredient in proving Theorem 3.7 is the observation—also new—that there exists a periodic solution of (1.10) if and only if an associated algebraic matrix equation has a positive semidefinite solution.

The other results of §3 are natural generalizations to the periodic case of some results of Wonham [34] on detectability.

In §4, the sufficiency part of Theorem 1.3 is proven as well as Theorem 4.11, which proves that the periodic solution of (1.1) is also an equilibrium solution. The existence of a periodic solution of (1.1) is established by the use of quasilinearization, a technique due to Bellman [4]. Quasilinearization as applied here means that every solution of (1.1) can be obtained as a limit of a sequence of solutions of equations like (1.10).

In §5, the necessity of Theorem 1.3 is established by the use of the canonical structure theorems of Kalman [16] and Weiss [33] and the eigenvalue–eigenvector characterization of detectability and stabilizability developed in other sections of this paper. The importance of this characterization in the proof is evident from Theorem 5.5, which shows that detectability and stabilizability satisfy the “duality principle.” The duality principle, as first stated by Kalman [15], is based on the observation that certain control and estimation concepts are in one-to-one correspondence. This correspondence is established by a time reversal, and by either defining or proving that a control (estimation) concept for (1.8)–(1.9) has a parallel estimation (control) interpretation for the system dual (adjoint)

$$(1.11) \quad \dot{z} = -A(t)'z + H(t)'u,$$

$$(1.12) \quad y = B(t)z.$$

The other lemmas are generalizations of corresponding lemmas for autonomous systems found in Payne [27] who outlined a simple proof of the necessity part of Theorem 1.5.

2. Controllability and periodic systems. Recall that Kalman's original definition of controllability [14] for constant matrices A and B is equivalent to verifying that the rank of the $n \times nm$ controllability matrix $[B, AB, \dots, A^{n-1}B]$ is n . The following characterization of controllability for constant matrices of any dimension is due to Hautus [11] and for B of dimension $n \times 1$ is found in Johnson [13].

THEOREM 2.1. *The pair (A, B) is controllable if and only if for each eigenvalue λ of A and each (possibly complex) left eigenvector η ($\eta A = \lambda \eta$), the equation $\eta B = 0$ implies $\eta = 0$.*

The main result of this section is to extend Hautus's theorem to the periodic system (1.8), but first some additional notation and results are needed.

Let $X(t)$ denote the fundamental matrix solution of

$$(2.2) \quad \dot{x} = A(t)x, \quad X(0) = I, \quad I = \text{identity matrix}.$$

The “observability matrix” for system (1.8)–(1.9) is

$$(2.3) \quad \mathcal{O}_\tau^\sigma(X(\cdot), H(\cdot)) = \int_\tau^\sigma (H(s)X(s)X^{-1}(\tau))'(H(s)X(s)X^{-1}(\tau)) ds$$

for $\sigma \geq \tau$. By the duality principle [16], the matrix $\mathcal{O}_\tau^\sigma(X(t), B(t))$ is the controllability matrix for (1.8).

Weiss proved [33] that the linear independence of the rows of the matrix $X^{-1}(\cdot)B(\cdot)$ on some smallest interval is equivalent to the controllability matrix being of rank n on that interval. This rank condition is equivalent to Kalman's original definition of controllability. Subsequently, Brunovsky [2] proved that the following definition is also equivalent to Kalman's.

DEFINITION 2.4. System (1.8) is K-W-B controllable if the rows of the matrix $X^{-1}(t)B(t)$ are linearly independent on the interval $[0, \omega]$ for some positive integer ι .

As shown at the end of this section (in Theorem 2.15) the interval of K-W-B controllability (when it exists) is always $[0, \omega]$. Thus, we shall assume, unless otherwise noted, that the interval of K-W-B controllability is $[0, \omega]$.

A more direct test of controllability was established by Silverman and Meadows [32]. Let $A(t)$ and $B(t)$ have continuous derivatives of order $n-2, n-1$ respectively, and define the $n \times nm$ controllability matrix, which also occurs in [28], $Q_{(1.8)}(t)$ for $(A(t), B(t))$

$$Q_{(1.8)}(t) = [P_0(t), P_1(t), \dots, P_{n-1}(t)],$$

with $P_0(t) = B(t)$, $P_i(t) = -A(t)P_{i-1}(t) + \dot{P}_{i-1}(t)$ for $i = 1, \dots, n-1$.

DEFINITION 2.5. System (1.8) is S-M controllable if the matrix $Q_{(1.8)}(t)$ is defined and continuous for all $t \in [0, \omega]$, and if for some t in this interval the rank of $Q_{(1.8)}(t)$ is n .

When $Q_{(1.8)}(t)$ is defined and continuous it will be periodic; thus its maximal rank will be achieved in one period. Consequently, the choice of interval in this definition is not restrictive.

Silverman and Meadows [32] prove that if a system is S-M controllable, then it is K-W-B controllable, and that the two definitions are equivalent when $A(t)$ and $B(t)$ are analytic matrices. Silverman [32] shows that the two definitions are not generally equivalent. When A and B are constant matrices, then both definitions coincide and $Q_{(1.8)}(t)$ is the well-known controllability matrix

$$[B, AB, \dots, A^{n-1}B].$$

The following two theorems are proved in Reid [29, p. 442] and are due to Floquet.

THEOREM 2.6. *The fundamental matrix solution of (2.2) can be written in the factored form*

$$(2.7) \quad X(t) = F(t) e^{Jt},$$

where $F(t)$ is a nonsingular matrix of period ω and J is a constant (possibly complex) matrix satisfying the equation $e^{J\omega} = X(\omega)$.

Since the coefficient matrix $A(t)$ and the initial matrix I determine $X(t)$ uniquely in (2.2), the matrices $F(t)$ and J which satisfy (2.7) will be called, not inappropriately, the *Floquet factors* of $A(t)$.

THEOREM 2.8. *If the matrices in (2.7) satisfy the matrix differential equation*

$$(2.9) \quad \dot{F}(t) + F(t)J - A(t)F(t) = 0, \quad -\infty < t < \infty, \quad F(0) = I,$$

then under the Floquet transformation $x = F(t)w$, (1.8) and (1.9) become

$$(2.10) \quad \dot{w} = Jw + F^{-1}(t)B(t)u,$$

$$(2.11) \quad y = H(t)F(t)w.$$

THEOREM 2.12. *System (1.8) is K-W-B(S-M) controllable if and only if (2.10) is also.*

Proof. The claim can be easily established by noting that controllability is a rank condition on $X^{-1}(t)B(t)$ or $Q_{(1,8)}(t)$, and the Floquet transformation is rank preserving because $F(t)$ is nonsingular.

Each eigenvalue of the monodromy matrix $X(\omega)$ is called a characteristic multiplier of (2.2). If any other initial matrix is prescribed in (2.2), the characteristic multipliers are related to those of $X(\omega)$ by a similarity transformation. Any κ which satisfies any of the equations $\rho = e^{\kappa\omega}$ is called a characteristic exponent of (2.2), and only the real part of each characteristic exponent is unique.

Part (b) of the following theorem extends Hautus's theorem to periodic systems.

THEOREM 2.13. (a) *If system (1.8) is K-W-B controllable on $[0, \omega]$, then for each characteristic exponent κ of the monodromy matrix $X(t)$ associated with $A(t)$ and associated eigenvector $\eta(\eta J = \kappa\eta)$, the equation*

$$(2.14) \quad \eta F^{-1}(t)B(t) = 0 \quad \text{for a.e. } t \in [0, \omega] \text{ implies } \eta = 0.$$

(b) *System (1.8) is S-M controllable if and only if condition (2.14) is satisfied.*

Proof. (a) If there exists a nonzero eigenvector η which satisfies (2.14), then $\eta(e^{-Jt}F^{-1}(t)B(t)) = e^{-\kappa t}(\eta F^{-1}(t)B(t)) = 0$ for a.e. $t \in [0, \omega]$. Now by equation (2.7), $X^{-1}(t) = e^{-Jt}F^{-1}(t)$, and so the rank of $X^{-1}(t)B(t)$ is less than n on $[0, \omega]$. Therefore system (1.8) is not K-W-B controllable.

(b) The proof in one direction follows from part (a). Suppose that (2.14) is satisfied and (1.8) is not S-M controllable. By Theorem 2.8, the controllability matrix $Q_{(2.10)}(t)$ has rank $< n$ for all $t \in [0, \omega]$. Thus, there exists a nonzero vector ξ such that $\xi Q_{(2.10)}(t) = 0$ or equivalently for $P_0(t) = F^{-1}(t)B(t)$, $\xi P_i(t) = 0$ for all $t \in [0, \omega]$ and $i = 0, 1, \dots, n-1$. Now following Hautus [11] let ψ be a polynomial of minimal degree such that $\xi\psi(J) = 0$. This polynomial clearly exists and has degree d with $1 \leq d \leq n$. For some κ and some polynomial $\phi(z)$ of degree $d-1$, $\psi(z) = \phi(z)(z - \kappa)$. The vector $\eta = \xi\phi(J)$ is a nonzero eigenvector of J and satisfies $\eta e^{J\omega} = \eta e^{\kappa\omega}$. By considering successive terms in the equation $\xi P_i(t) = 0$, we find that $\xi J^i P_0(t) = 0$ ($i = 1, \dots, n-1$); thus, $\eta F^{-1}(t)B(t) = \xi\phi(J)P_0(t) = 0$ for all $t \in [0, \omega]$. This contradicts (2.14) and completes the proof.

Let $|\cdot|$ denote both the Euclidean norm of a vector and the compatible matrix norm, i.e., the spectral norm [22, p. 210].

THEOREM 2.15. *System (1.8) is K-W-B controllable on $[0, \omega]$ if and only if it is K-W-B controllable.*

Proof. Since the implication is immediate in one direction, we suppose that (1.8) is K-W-B controllable on $[0, \iota\omega]$ for some $\iota > 1$, but is not controllable on $[0, \omega]$. This means that there exists a vector x such that

$$X^{-1}(t)B(t)x = 0 \quad \text{for a.e. } t \in [0, \omega].$$

Let t_1 be any point in the interval $[\omega, 2\omega]$, and choose $t \in [0, \omega]$ such that $t + \omega = t_1$. Now

$$e^{-Jt_1}F^{-1}(t_1)B(t_1)x = e^{-J\omega}[e^{-Jt}F^{-1}(t)B(t)x] = 0$$

for a.e. $t_1 \in [\omega, 2\omega]$. Since i is a positive integer, this implies that (1.8) is not K-W-B controllable $[0, \omega]$. This contradiction completes the proof.

In the sequel when convenient, any property that is attributed to the matrix pair $(A(t), B(t))$, for example, controllability, is understood by making the obvious association with (1.8), and vice versa. By this convention and the duality principle, the operational meaning of the statement “ $(H(t), A(t))$ is K-W-B(S-M) controllable” should be clear. Thus the previous theorems and the discussion of this section remain valid when “controllability” is replaced throughout by “observability.”

3. Periodicity and detectability. We introduce the following definition of detectability for periodic systems, which for autonomous systems originated with Wonham [34].

Let $\text{Re } \kappa$ denote the real part of the characteristic exponent κ .

DEFINITION 3.1. Let $F(t)$ and J be the Floquet factors of $A(t)$. $(H(t), A(t))$ is *detectable* if for each characteristic exponent κ with $\text{Re } \kappa \geq 0$ and each associated nonzero eigenvector $\eta (J\eta = \kappa\eta)$,

$$(3.2) \quad H(t)F(t)\eta = 0 \quad \text{for a.e. } t \in [0, \omega] \text{ implies } \eta = 0.$$

Clearly, when $(H(t), A(t))$ is S-M observable, it is also detectable. The converse is false as illustrated by $(0, -I)$. However, by Theorem 2.13(b), the following concepts are equivalent: $(H(t), A(t))$ is detectable if and only if every unstable ($\text{Re } \kappa \geq 0$) characteristic exponent is S-M observable (satisfies condition (2.14)). This latter characterization of detectability motivated Wonham's original definitions [34].

Let $r(M)$ be the spectral radius (the maximum of the moduli of its eigenvalues) of the matrix M and η^* be the conjugate transpose of the eigenvector η .

LEMMA 3.3. *If there exists a $K_0 \geq 0$ ($K_0 \neq 0$), which satisfies the algebraic matrix equation,*

$$(3.4) \quad K_0 - X(\omega)'K_0X(\omega) = \mathcal{O}_0^w(X(\cdot), H(\cdot)),$$

then (1.10) has a periodic solution $K(t, K_0)$.

Proof. Let $K_1(t, \omega, K_0)$ be the general solution of (1.10), where K_0 satisfies (3.4). This unique solution has the integral representation

$$(3.5) \quad K_1(t, \omega, K_0) = X^{-1}(t)'X(\omega)'K_0X(\omega)X^{-1}(t) + \mathcal{O}_t^w(X(\cdot), H(\cdot)).$$

Since K_0 satisfies (3.4),

$$K_1(0, \omega, K_0) = X(\omega)'K_0X(\omega) + \mathcal{O}_0^w(X(\cdot), H(\cdot)) = K_0.$$

The general solution $K(t, K_0)$ defined by the identity

$$K(t, K_0) = K_1(t + W, \omega, K_0), \quad -\infty < t < \infty,$$

is a periodic solution of (1.10).

Lemma 3.3 is the matrix analog of the following well-known algebraic criterion [10, p. 223] for the existence of periodic solutions for ordinary linear differential equations. If there is a nonzero solution X_0 of the linear system of equations $(X(\omega) - I)x = \int_0^\omega X(s)f(s) ds$, then

$$(3.6) \quad \dot{X} = A(t)X + f(t), \quad f(t + \omega) = f(t)$$

has a periodic solution. By the usual determinantal condition, a sufficient (but not necessary) condition for (3.6) to have a nonzero solution is that $r(X(\omega)) \neq 1$. For this reason, the main significance of the following theorem is that the spectral radius condition on the monodromy matrix is equivalent to the existence of a periodic solution of (1.10). Under the conditions of this theorem, this is a new result, even for (3.6).

THEOREM 3.7. *If $(H(t), A(t))$ is detectable, then the following are equivalent:*

- (i) $r(X(\omega)) < 1$;
- (ii) *there exists a periodic solution $K(t, K_0)$ of (1.10);*
- (iii) *equation (1.10) has at least one solution which is defined and bounded on $(-\infty, \omega]$, with $K_0 \geq 0$ ($K_0 \neq 0$)[~].*

Proof. (i) implies (ii). Since $r(X(\omega)) < 1$, the solution of (3.4) is given by the convergent series

$$K_0 = \sum_{j=0}^{\infty} (X(\omega))'^j \mathcal{O}_0^\omega(X(\cdot), H(\cdot))(X(\omega))^j.$$

Each term in this series is positive semidefinite because $X(\omega)$ is nonsingular and $H(t)H(t)'$ is not identically zero on some positive interval, and in particular, $\mathcal{O}_0^\omega(X(\cdot), H(\cdot)) \neq 0$. This implies that the matrix K_0 is nonzero, and by Lemma 3.3, (1.10) has a periodic solution.

(ii) implies (i). Let $K(t, K_0)$ be the periodic solution of (1.10), and suppose that $r(X(\omega)) \geq 1$. The solution $K(t, \omega, K_0)$ defined by (3.5) agrees with $K(t, K_0)$ at $t = \omega$, and both solutions satisfy (1.10) on $[0, \omega]$. Thus by uniqueness, $K_1(t, \omega, K_0)$ satisfies (1.10). There are now two cases: 1. $\text{Re } \kappa = 0$ and 2. $\text{Re } \kappa > 0$.

1. Let η be an associated eigenvector, and consider the equation

$$(3.8) \quad \eta^* K_0 \eta - \eta^* e^{J\omega} K_0 \eta e^{J\omega} = \eta^* \mathcal{O}_0^\omega(X(\cdot), H(\cdot)) \eta.$$

Since $\text{Re } \kappa = 0$, this equation is valid only if $H(t)F(t)\eta = 0$ for a.e. $t \in [0, \omega]$, which contradicts the detectability.

2. Again let η be an eigenvector associated with $\text{Re } \kappa > 0$. Equation (3.8) now yields

$$(1 - \exp((2 \text{Re } \kappa)\omega)) \eta^* K_0 \eta = \eta^* \mathcal{O}_0^\omega(X(\cdot), H(\cdot)) \eta.$$

This equation is valid only if both quadratic forms are zero, which again contradicts the detectability assumption. Thus $r(X(\omega)) < 1$.

Clearly (ii) implies (iii). To obtain the converse, suppose that there is no periodic solution of (1.10). Since (i) and (ii) are already equivalent, this means there exists a characteristic exponent κ with $\text{Re } \kappa \geq 0$. Let η be a corresponding nonzero eigenvector η of J . Define the mapping $T(\cdot)$ from the set \mathcal{M} of $n \times n$ real-valued

positive semidefinite matrices into itself by

$$T(M) = X(\omega)'MX(\omega) + \mathcal{O}_0^{\omega}(X(\cdot), H(\cdot)).$$

By induction on $j, j = 1, 2, 3, \dots$,

$$\begin{aligned} T^j(M) &= X(j\omega)'MX(j\omega) + X((j-1)\omega)'\mathcal{O}_0^{\omega}(X(\cdot), H(\cdot))X((j-1)\omega) \\ &\quad + \dots + \mathcal{O}_0^{\omega}(X(\cdot), H(\cdot)). \end{aligned}$$

We have used the identity $X(j\omega) = X^j(\omega)$ which follows by uniqueness. Now

$$\begin{aligned} \eta^* T^j(M) \eta &= \exp(2(\operatorname{Re} \kappa)j\omega) \eta^* M \eta + (\exp(2(\operatorname{Re} \kappa)(j-1)\omega) + \dots + 1) \\ &\quad \cdot \eta^* \mathcal{O}_0^{\omega}(X(\cdot), H(\cdot)) \eta. \end{aligned}$$

Since $(H(t), A(t))$ is detectable, the observability matrix is positive definite, and so $\lim \eta^* T^j(M) \eta = \infty$ as $j \rightarrow \infty$. Let $K(t, K_0)$ be a bounded solution of (1.10). Again using $X^j(\omega) = X(j\omega)$, it is easy to verify that $T^j(K_0) = K(j\omega, K_0)$. This implies that $K(t, K_0)$ cannot be a bounded solution; so there must exist a periodic solution.

The next lemma is inspired by a similar lemma in Wonham [34] for constant matrices.

LEMMA 3.9. *Let $K(t)$ be a continuous periodic matrix of period ω and let*

$$(3.10) \quad K(t)'K(t) + H(t)'H(t) = L(t)'L(t).$$

If $(H(t), A(t))$ is detectable, then $(L(t), A(t) + G(t)K(t))$ is detectable for any periodic matrix $G(t)$ (with period ω) of suitable dimension.

Proof. By an application of Theorem 2.8, the pair $(H(t)F(t), J)$ is detectable, also. Furthermore, using the same theorem, the detectability of $(L(t)F(t), J + F^{-1}(t)G(t)K(t)F(t))$ implies the same behavior for $(L(t), A(t) + G(t)K(t))$. Thus, we need only prove that the former pair is detectable. If this claim is false, then there exists a characteristic exponent κ_1 of the monodromy matrix for $J + F^{-1}(t) \cdot G(t)K(t)F(t)$ with Floquet factors J_1 and $F_1(t)$ such that for any eigenvector $\eta(J_1\eta = \kappa_1\eta)L(t)F(t)F_1(t)\eta = 0$ a.e. $t \in [0, \omega]$.

As can be inferred from Theorem 2.8, the Floquet factors satisfy the matrix differential equation

$$(3.11) \quad \dot{F}_1(t) + F_1(t)J_1 - (J + F^{-1}(t)G(t)K(t)F(t))F_1(t) = 0$$

with initial value $F_1(0) = I$.

Multiplying (3.10) by the nonsingular matrix $F(t)$ yields

$$(K(t)F(t))'K(t)F(t) + (H(t)F(t))'H(t)F(t) = (L(t)F(t))'L(t)F(t).$$

Using this equation, we have, for all $t \in [0, \omega]$, $K(t)F(t)F_1(t)\eta = 0$ and $H(t)F(t)F_1(t)\eta = 0$. Since (3.11) is true for every t and $F(0) = I$, all these results imply that

$$J\eta = \kappa_1\eta.$$

The vector $F_1(t)\eta$ satisfies the ordinary differential equation

$$r + \kappa_1 r - Jr = 0$$

with initial value at $t = 0$ equal to η . Since the vector $e^{(J-\kappa)t}\eta$ is also a solution of this equation, by uniqueness these two solutions are equal for all t . Combining all these observations, we have

$$H(t)F(t)F(t)\eta = H(t)F(t)(e^{(J-\kappa)t}\eta) = H(t)F(t)\eta = 0$$

for a.e. $t \in [0, \omega]$, which contradicts the detectability of $(H(t)F(t), J)$.

4. Periodic solutions of the Riccati equation. In this section, the sufficiency part of Theorem 1.3 is established, and the equilibrium nature of the periodic solution is described. First some additional notation and concepts are introduced.

DEFINITION 4.1. System (1.8) is stabilizable if there exists a continuous periodic matrix $Q(t)$ (with period ω) of suitable dimension such that the fundamental matrix solution $\Phi(t)$ of the system

$$(4.2) \quad \dot{x} = (A(t) + B(t)Q(t))x$$

satisfies $r(\Phi(\omega)) < 1$.

Informally, this definition means that (1.8) is stabilized by the "feedback control law" $u = Q(t)x$ if the fundamental matrix solution of the "closed loop system" (4.2) satisfies the inequality

$$(4.3) \quad |\Phi(t)| \leq \beta e^{-\kappa t}, \quad t \geq 0,$$

for some positive constants β and κ . By (4.3), the fundamental matrix solution tends to zero as $t \rightarrow \infty$ at an exponential rate, and for that reason, matrix solutions satisfying this inequality are called *exponentially stable*.

The equivalence of exponential stability and $r(\Phi(\omega)) < 1$ is found in Reid [30, p. 445]. As established by Brunovsky [2], a stabilizing matrix $Q(t)$ exists whenever $(A(t), B(t))$ is K-W-B controllable.

For completeness we include the following lemma, which is found more generally in [18].

LEMMA 4.4 (Monotone lemma). Let $\{P_v : v = 1, 2, \dots\}$ be a sequence of $n \times n$ symmetric matrices such that $P_1 \geq P_2 \geq \dots \geq 0$. Then $P_\infty = \lim P_v$ as $v \rightarrow \infty$ exists and $P_\infty \geq 0$.

Consider the quasilinearization identity

$$(4.5) \quad \begin{aligned} \psi(R(t), A(t) - B(t)K_0(t), K_0(t)) &= \psi(R(t), A(t) - B(t)K(t), K(t)) \\ &\quad - (K(t) - K_0(t))'(K(t) - K_0(t)), \end{aligned}$$

where $K_0(t) = B(t)'R(t)$ and $K(t)$ is any well-defined matrix of suitable dimension. This quasilinearization identity, which in scalar form is due to Bellman [4], expresses the fact that the right-hand side of (4.5) as a function of $K(t)$ is minimized by $K_0(t)$.

Consider the differential equations for $v = 1, 2, 3, \dots$,

$$(4.6) \quad \dot{R}_v(t) + \psi(R_v(t), A_v(t), L_v(t)) = 0,$$

where $K_v(t)$ is any well-defined continuous matrix of suitable dimension, the matrix $A_v(t)$ is defined by the equation

$$(4.7) \quad A_v(t) = A(t) - B(t)K_v(t),$$

and $L_v(t)$ is the "square root" of the equation

$$L_v(t)'L_v(t) = K_v(t)'K_v(t) + H(t)'H(t).$$

The fundamental matrix $X_v(t)$ satisfies the differential equation

$$(4.8) \quad \dot{X}_v(t) = A_v(t)X_v(t), \quad X_v(0) = I.$$

The following equation for the difference of any two solutions of (4.6) is easily obtained by an application of the quasilinearization identity:

$$(4.9) \quad \dot{R}_v(t) - \dot{R}_{v+1}(t) + \psi(R_v(t) - R_{v+1}(t), A_v(t), K_v(t) - K_{v+1}(t)) = 0.$$

As the proofs in this section indicate, an equation like (4.9) for the difference of any two solutions of (4.6) is an important and useful consequence of quasilinearization in the study of Riccati equations.

To prove the sufficiency part of Theorem 1.3, we shall first construct a sequence of periodic solutions of (4.6), which pointwise satisfy the conditions of the monotone lemma. Then it will be shown by an application of the quasilinearization identity that the limit of this convergent sequence is the desired periodic solution. For convenience, the key steps in the proof are labeled and described.

Proof. (a) For $v = 1$, there exists a periodic solution of (4.6). To prove this statement, choose a periodic continuous matrix $K_1(t)$ in (4.7) such that the fundamental matrix solution $X_1(t)$ of (4.8) satisfies $r(X_1(\omega)) < 1$. Such a choice is possible, because $(A(t), B(t))$ is stabilizable. Since $(H(t), A(t))$ is detectable, we now apply Theorem 3.7 to conclude that there exists a periodic solution $R_1(t, R_0(1))$ of (4.6) with $R_0(1) \geq 0$.

(b) For $v = 2$ and $K_2(t) \equiv B(t)'R_1(t, R_0(1))$, (4.6) has a periodic solution $R_2(t, R_0(2))$ with $R_0(2) \geq 0$. This claim will be verified by showing that condition (iii) of Theorem 3.7 is satisfied. Since (4.6) is a linear matrix differential equation, there exists a solution $\hat{R}(t, R_0)$ ($R_0 \geq 0$) defined on some interval containing zero. Because the difference $R_1(t, R_0(1)) - \hat{R}(t, R_0)$ satisfies (4.9), it has the equivalent Volterra integral representation for $t \leq \omega$,

$$(4.10) \quad R_1(t, R_0(1)) - \hat{R}(t, R_0) = X_1^{-1}(t)'X_1(\omega)'(R_0(1) - R_0)X_1(\omega)X_1^{-1}(t) \\ + \mathcal{O}_t^\omega(X_1(\cdot), K_1(\cdot) - K_2(\cdot)).$$

Because the matrix $K_1(t) - K_2(t)$ is continuous and periodic, it is also bounded for all t . By part (a), $r(X_1(\omega)) < 1$, which is equivalent to (4.3), and this means that $X_1^{-1}(t)$ is exponentially stable on $(-\infty, \omega]$ [29, p. 445]. Using these two results, it is easy to conclude that the right-hand side of (4.10) is defined and bounded on $(-\infty, \omega]$ and, since $R_1(t, R_0(1))$ is periodic, $\hat{R}(t, R_0)$ is defined and bounded on this same interval.

Recall that $(H(t), A(t))$ is detectable, and so by Lemma 3.9, $(L_2(t), A_2(t))$ is detectable. The claim is now clearly established.

(c) The successive initial value matrices $R_0(1)$ and $R_0(2)$, associated with the periodic solutions $R_1(t, R_0(1))$ and $R_2(t, R_0(2))$ obtained in parts (a) and (b), satisfy the inequality $R_0(1) \geq R_0(2)$.

Clearly, the difference of these two periodic solutions satisfies (4.9) with $v = 1$. Again, use the integral representation (evaluated at $t = 0$) for the solution of

(4.9) to obtain the following algebraic matrix equation:

$$R_0(1) - R_0(2) = X_1(\omega)'(R_0(1) - R_0(2))X_1(\omega) + \mathcal{O}_0^\omega(X_1(\cdot), K_1(\cdot) - K_2(\cdot)).$$

Because $r(X_1(\omega)) < 1$, the following series is convergent:

$$R_0(1) - R_0(2) = \sum_{j=1}^{\infty} (X_1(\omega))' \mathcal{O}_0^\omega(X_1(\cdot), K_1(\cdot) - K_2(\cdot)) (X_1(\omega))^j.$$

By inspecting this series, the validity of $R_0(1) \geq R_0(2)$ is easily established.

(d) The successive periodic solutions obtained in parts (a) and (b) satisfy pointwise the inequality

$$R_1(t, R_0(1)) \geq R_2(t, R_0(2)).$$

Using part (c), $R_0(1) \geq R_0(2)$, and again the Volterra integral representation for the solution of (4.9) with $v = 1$, this result follows easily.

This chain of arguments for $v = 2, 3, 4, \dots$ can be repeated with $K_{v+1}(t) \equiv B(t)'R_v(t, R_0(v))$ to obtain a sequence $R_v(t, R_0(v))$ of periodic solutions of (4.6). The initial matrices $R_0(v)$ by part (c) are monotone nonincreasing and bounded below by the null matrix. By part (d), the periodic solution matrices have this property for each $t \in (-\infty, \omega]$, also. Thus by the monotone lemma, the matrices R_0 , $R(t, R_0)$ and $K(t)$ are well-defined by the following pointwise limits as $v \rightarrow \infty$:

$$\lim R_0(v) = R_0, \quad \lim R_v(t, R_0(v)) = R(t, R_0),$$

and

$$\lim K_v(t).$$

Furthermore, as constructed, $R_0 \geq 0$, and both $K(t)$ and $R(t, R_0)$ are periodic matrices. For completeness, the remainder of the proof which can be found in Wonham [34] is included.

Since $|R_v(t, R_0(v))| \leq \max \{|R_1(t, R_0(1))| : 0 \leq t \leq \omega\}$, $v = 1, 2, 3, \dots$, it follows that the sequences $\{K_v(t)\}$ and $\{L_v(t)\}$ are uniformly bounded for all t . Also, the sequence of solution matrices $\{X_v(t)\}$ of (4.8) is uniformly bounded, because $r(X_v(\omega)) < 1$ for each v . The solutions $R_v(t, R_0(v))$ of (4.6) have the following integral representation for $-\infty < t \leq \omega$:

$$R_v(t, R_0(v)) = X_v^{-1}(t)' X_v(\omega)' R_0(v) X_v(\omega) X_v^{-1}(t) + \mathcal{O}_t^\omega(X_v(\cdot), L_v(\cdot)).$$

Using Dini's theorem [1] and the Lebesgue dominated convergence theorem, the matrices $R_v(t, R_0(v))$, $X_v(t)$ and $L_v(t)$ can be replaced in the equation by $R(t, R_0)$, $X(t)$ and $L(t)$. By the quasilinearization identity, this means that (1.1) has a periodic solution. The uniqueness of this solution is proven in Wonham [34].

The exponential stability of (1.4) is a consequence of Theorem 3.7 and Lemma 3.9.

The next theorem shows that the periodic solution of (1.1) is an equilibrium solution as discussed in the Introduction. Yoshizawa [35] has a discussion (including definitions) of boundedness for ordinary differential equations.

THEOREM 4.11. *If $(A(t), B(t))$ is stabilizable and $(H(t), A(t))$ is detectable, then every solution $R(t, R_1)$ ($R_1 \geq 0$) is bounded on the interval $(-\infty, \omega]$. Moreover, the*

periodic solution $R(t, R_0)$ is an equilibrium solution, i.e.,

$$|R(t, R_1) - R(t, R_0)| \rightarrow 0 \quad \text{as } t \rightarrow -\infty.$$

Proof. Since $(A(t), B(t))$ is stabilizable, there exists a matrix $K_0(t)$ such that the solution $X_0(t)$ of (4.8) is exponentially stable. For any $R_1 \geq 0$, let $\hat{R}(t, R_1)$ satisfy the equation

$$(4.12) \quad \dot{\hat{R}}(t, R_1) + \psi(\hat{R}(t, R_1), A_0(t), L_0(t)) = 0.$$

Since $X_0(t)$ is exponentially stable and the coefficient matrices in (4.12) are bounded, by using the Volterra representation of the solution of (4.12), it is easy to verify that $\hat{R}(t, R_0)$ is bounded on $(-\infty, \omega]$.

By the minimum property (Wonham [34]), the solution $R(t, R_1)$ of (1.1) satisfies the inequality $0 \leq R(t, R_1) \leq \hat{R}(t, R_1)$. Since the spectral norm is a monotone norm [25], this inequality implies that the norm of $R(t, R_1)$ is bounded on $(-\infty, \omega]$, also.

To prove the final statement, let $K_0(t) \equiv B(t)'R(t, R_0)$ and $K_1(t) \equiv B(t)'R(t, R_1)$. These two solutions both satisfy (1.1), and so their difference satisfies the equation

$$\begin{aligned} \dot{R}(t, R_0) - \dot{R}(t, R_1) + A_1(t)'(R(t, R_0) - R(t, R_1)) \\ + (R(t, R_0) - R(t, R_1))A_0(t) = 0. \end{aligned}$$

By uniqueness,

$$(4.13) \quad R(t, R_0) - R(t, R_1) = X_1^{-1}(t)' X_1(\omega)' (R_0 - R_1) X_0(\omega) X_0^{-1}(t)$$

for $t \in (-\infty, \omega]$.

By means of the hypotheses of Theorem 4.11, Lemma 3.9, and Theorem 3.7, both fundamental matrices in (4.13) are seen to be exponentially stable, which completes the proof.

5. Proof of the necessity of Theorem 1.3. The necessity of stabilizability and detectability in Theorem 1.3 can be inferred from the following lemma.

LEMMA 5.1. *If there is one and only one positive semidefinite periodic solution $R(t, R_0)$ of (1.1) such that the solution of (1.4) is exponentially stable, then $(H(t), A(t))$ is detectable.*

An essential component in the proof of this lemma is the canonical structure theorem. This was first stated for the time varying case by Kalman [16] and proved by Weiss [33]. The main features of Weiss's paper are included below.

When $(H(t), A(t))$ is not observable, then there exists a nonsingular matrix $S(t)$ which transforms (independent of time) the triple (A, B, H) into observable canonical form $(\hat{A}, \hat{B}, \hat{H})$,

$$\hat{A} = \begin{bmatrix} \hat{A}_{11} & 0 \\ \hat{A}_{21} & \hat{A}_{22} \end{bmatrix}, \quad \hat{B} = \begin{bmatrix} \hat{B}_1 \\ \hat{B}_2 \end{bmatrix}, \quad \hat{H} = [\hat{H}_1 \quad 0],$$

so that $(\hat{H}_1, \hat{A}_{11})$ is K-W-B controllable.

The Floquet factors J and F associated with A are transformed by the similarity transformation $S(t)$ into the compatibly partitioned matrices \hat{J} and \hat{F} ,

$$\hat{J} = \begin{bmatrix} \hat{J}_{11} & 0 \\ \hat{J}_{21} & \hat{J}_{22} \end{bmatrix}, \quad \hat{F} = \begin{bmatrix} \hat{F}_{11} & 0 \\ \hat{F}_{21} & \hat{F}_{22} \end{bmatrix}.$$

The matrix $S(t)$ satisfies the matrix differential equation

$$(5.2) \quad \dot{S}(t) + A(t)S(t) = S(t)\hat{A}(t).$$

The matrices \hat{B} and \hat{H} satisfy the equations

$$(5.3) \quad S^{-1}B = \hat{B}, \quad \hat{H} = HS.$$

LEMMA 5.4. *If $(\hat{H}_1(t), \hat{A}_{11}(t))$ is detectable, then $(H(t), A(t))$ is detectable if and only if $r(e^{J_{22}\omega}) < 1$.*

Proof. Since $(\hat{H}_1(t), \hat{A}_{11}(t))$ is detectable and \hat{J} is a block triangular matrix, the only undetectable characteristic exponents of \hat{J} that can occur are in the submatrix \hat{J}_{22} . The nonnull eigenvectors for those undetectable characteristic exponents can be expressed as the direct sum of a null vector and the eigenvectors of \hat{J}_{22} . The proof can now be finished by recalling the definition of detectability.

Detectability, as originally defined by Wonham [34], means that the dual system (A', H') is stabilizable. This concept is based on the observation that a minimal control requirement for a system is that (at least) the unstable eigenvalues can be modified by a feedback control. Lemma 5.4 and the following theorem demonstrate that our definition of a detectable pair reflects this philosophy also.

In fact, the next theorem shows that detectability and stabilizability, as defined here, obey the duality principle.

THEOREM 5.5. *$(A(t), B(t))$ is stabilizable if and only if the dual $(B(t)', A(t)')$ is detectable.*

Proof. Suppose that the dual system $(B(t)', A(t)')$ is not detectable. The Floquet factors of $A(t)'$ for the dual system (1.11)–(1.12) are $-J'$ and $F^{-1}(t)'$. Thus there exists a characteristic exponent κ with $\text{Re } \kappa \geq 0$ ($J\eta = \kappa\eta$) such that the equation $B(t)'F^{-1}(t)'\eta = 0$ is satisfied for $\eta \neq 0$.

Let $Q(t)$ be any suitable matrix in (4.2). By using (2.9), we obtain, after a Floquet transformation, the equivalent system

$$\dot{w} = (J + F^{-1}(t)B(t)Q(t))w.$$

The general solution $w(t, t_0, \eta)$ of this equation satisfies the integral equation

$$w(t, t_0, \eta) = e^{J(t-t_0)}\eta + \int_{t_0}^t e^{J(t-s)}F^{-1}(s)B(s)Q(s)ds.$$

Multiply this integral equation on the left by η' and then apply Schwarz's inequality to obtain

$$|\eta|^2 e^{\text{Re } \kappa(t-t_0)} \leq (|\eta|^2 |w(t, t_0, \eta)|^2)^{1/2}.$$

This implies that $w(t, t_0, \eta)$ is unbounded as $t \rightarrow \infty$, and since $Q(t)$ is arbitrary, this contradicts the stabilizability of $(A(t), B(t))$.

Now suppose that $(A(t), B(t))$ is not stabilizable. By the contrapositive of Brunovsky's theorem [2], $(A(t), B(t))$ is not K-W-B controllable. So again by Weiss, (A, B) can be transformed by a nonsingular matrix $S(t)$ into control canonical form (\tilde{A}, \tilde{B}) (independent of time),

$$\tilde{A} = \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ 0 & \tilde{A}_{22} \end{bmatrix}, \quad \tilde{B} = \begin{bmatrix} \tilde{B}_1 \\ 0 \end{bmatrix},$$

so that (\tilde{A}_{11}, B_1) is K-W-B controllable. Also, of course, the Floquet factors J and F become \tilde{J} and \tilde{F} ,

$$\tilde{J} = \begin{bmatrix} \tilde{J}_{11} & \tilde{J}_{12} \\ 0 & \tilde{J}_{22} \end{bmatrix}, \quad \tilde{F} = \begin{bmatrix} \tilde{F}_{11} & \tilde{F}_{12} \\ 0 & \tilde{F}_{22} \end{bmatrix}.$$

If our assumption is valid, the matrix \tilde{J}_{22} must have an unstable characteristic exponent, because $(\tilde{A}_{11}, \tilde{B}_{11})$ is stabilizable and $S(t)$ operates as a similarity transformation on the Floquet factors. By the duality principle and Theorem 2.13 (a), $(\tilde{B}_1(t)', \tilde{A}_{11}(t)')$ is detectable. Since the time direction is reversed in the dual system, by Lemma 5.4, $(B(t)', A(t)')$ is not detectable, which completes the proof.

By applying Theorem 5.5, the following lemma is easily proved by contradiction.

LEMMA 5.6. *If $(A(t), B(t))$ is stabilizable, then $(\hat{A}_{11}(t), \hat{B}_1(t))$ is stabilizable.*

We now present the proof of Lemma 5.1.

Proof. Let $R(t, R_0)$ be the periodic solution of (1.1), and let $\hat{R}(t, \hat{R}_0) = S(t)'R(t, R_0)S(t)$. By using (5.2)–(5.3), it is easy to verify that $\hat{R}(t_0, \hat{R}_0)$ is a solution of the Riccati equation

$$(5.7) \quad \dot{\hat{R}}(t) + \hat{A}(t)' \hat{R}(t) + \hat{R}(t) \hat{A}(t) - \hat{R}(t) \hat{B}(t) \hat{B}(t)' \hat{R}(t) + \hat{H}(t)' \hat{H}(t) = 0.$$

The following equation has a periodic solution, because $(\hat{A}_{11}(t), \hat{B}_1(t))$ is stabilizable (by Lemma 5.6) and $(\hat{H}_{11}(t), \hat{A}_{11}(t))$ is K-W-B observable:

$$(5.8) \quad \begin{aligned} &\dot{\hat{R}}_{11}(t) + \hat{A}_{11}(t)' \hat{R}_{11}(t) + \hat{R}_{11}(t) \hat{A}_{11}(t) \\ &- \hat{R}_{11}(t) \hat{B}_1(t) \hat{B}_1(t)' \hat{R}_{11}(t) + \hat{H}_1(t)' \hat{H}_1(t) = 0. \end{aligned}$$

The periodic solution matrix of (5.8), when bordered by appropriately dimensioned null matrices, is also a periodic solution of (5.7). By Theorem 4.12 these two periodic solutions agree, and so $\hat{R}(t, \hat{R}_0)$ is the conformably partitioned matrix

$$R(\hat{t}, \hat{R}_0) = \begin{bmatrix} \hat{R}_{11}(t, (\hat{R}_{11})_0) & 0 \\ 0 & 0 \end{bmatrix}.$$

The vector $\hat{x} = S(t)x$ satisfies the equation

$$(5.9) \quad \dot{\hat{x}} = (\hat{A}(t) - \hat{B}(t) \hat{B}(t)' \hat{R}(t, \hat{R}_0)) \hat{x}$$

for x satisfying (1.4).

The coefficient matrix in (5.7) is

$$(5.10) \quad \begin{bmatrix} \hat{A}_{11}(t) - \hat{B}_1(t) \hat{B}_1(t)' \hat{R}_{11}(t, (\hat{R}_{11})_0) & 0 \\ \hat{A}_{21}(t) - \hat{B}_2(t) \hat{B}_1(t)' \hat{R}_{11}(t, (\hat{R}_{11})_0) & \hat{A}_{22}(t) \end{bmatrix}.$$

Since $(H_{11}(t), A_{11}(t))$ is K-W-B observable and thus, by Theorem 2.13(a), detectable, the proof will be completed by proving that $r(e_{22}^J) \leq 1$ and applying Lemma 5.4. By assumption, the fundamental matrix solution (1.4) is exponentially stable, and since $S(t)$ is nonsingular for all t , the matrix $e^{J_{22}t}$, which is a solution of (5.9) as (5.10) shows, satisfies $r(e^{J_{22}t}) < 1$.

An obvious sequel to this paper is a similar investigation of matrix Riccati equations with almost periodic coefficients. As is well known, when the coefficients are almost periodic, the Floquet representation is no longer generally valid, and so a complete and parallel theory is impossible. However, with suitable hypotheses, the existence of an almost periodic solution of (1.1) is certainly assured.

In a different direction, preliminary numerical studies in controlling an unstable Mathieu equation indicate that the numerical convergence to the equilibrium periodic solution of the Riccati equation is quite rapid. This numerical behavior is consistent with the observed convergence rates for constant coefficient Riccati equations.

Acknowledgment. The author would like to thank Dr. E. A. Fay, Dr. D. E. Zilmer and Dr. R. B. Leipnik for their advice and encouragement in the preparation of this paper.

Note added in proof. Any of a set of additional conditions appears to be required to establish the existence of a fixed nulling vector for $Q_{(2,10)}(t)$ in the sufficiency portion of Theorem 2.13(b), as a relatively simple example shows. Namely, the equivalent Floquet systems ((2.10), (2.11)) must satisfy either

- (i) F equivalence (in the sense of Brunovsky, *Kybernetika*, 6 (1970)) to a detectable autonomous system (DAT),
- (ii) Lyapunov equivalence (Wolovich, *IEEE Trans., Automatic Control*, AC-13 (1968)) to a DAT, or
- (iii) commutivity of coefficient matrices.

All other relevant results remain valid under this replacement. Alternatively, if the hypotheses of K-W-B controllability replaces detectability and one of the above conditions, then parts (a) and (b) of the proof of Theorem 1.3 hold, by mimicking the constructions in Theorem 3.7 and Lemma 3.9. Note that Theorem 1.5 does not require observability. The last two remarks suggest that a slightly tightened definition of detectability would allow the restrictions (i), (ii) or (iii) to be removed.

REFERENCES

- [1] T. APOSTEL, *Mathematical Analysis*, Addison-Wesley, Reading, Mass., 1957.
- [2] P. BRUNOVSKY, *Controllability and linear closed-loop controls in linear periodic systems*, *J. Differential Equations*, 6 (1969), pp. 296–313.
- [3] R. BELLMAN, *Introduction to Matrix Analysis*, 2nd ed., McGraw-Hill, New York, 1970.
- [4] ———, *Functional Equations in the theory of dynamic programming, positivity and quasilinearization*, *Proc. Nat. Acad. Sci. U.S.A.*, 41 (1955), pp. 743–746.
- [5] S. BITTANTI, G. FRONZA AND G. GUARDABASSI, *Periodic optimization of linear systems under control power constraints*, *Automatica—J. IFAC*, 9 (1973), pp. 269–271.
- [6] R. S. BUCY, *The Riccati equation and its bounds*, *J. Comput. System Sci.*, 6 (1972), pp. 343–353.
- [7] R. S. BUCY AND P. D. JOSEPH, *Filtering for Stochastic Processes With Applications to Guidance*, Interscience, New York, 1968.
- [8] W. A. COPPEL, *Disconjugacy*, *Lecture Notes in Mathematics*, Springer-Verlag, New York, 1971.
- [9] M. FJELD, *Optimal control of multivariable periodic processes*, *Automatica—J. IFAC*, 5 (1969), pp. 497–506.
- [10] A. HALANAY, *Differential Equations: Stability, Oscillations, Time Lags*, Academic Press, New York, 1966.
- [11] M. L. J. HAUTUS, *Controllability and observability conditions of linear autonomous systems*, *Indag. Math.*, 72 (1969), pp. 443–448.

- [12] ———, *Stabilization, controllability, and observability of linear autonomous systems*, Ibid., 32 (1970), pp. 448–455.
- [13] C. D. JOHNSON, *Invariant hyperplanes for linear dynamical systems*, IEEE Trans. Automatic Control, 11 (1966), pp. 113–116.
- [14] R. E. KALMAN, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana, 5 (1960), pp. 102–119.
- [15] ———, *On the general theory of control*, Proc. 1st Internat. Congr. on Automatic Control, vol. 1, Butterworth, London, 1960.
- [16] ———, *Mathematical description of linear dynamical systems*, this Journal, 4 (1968), pp. 152–192.
- [17] R. E. KALMAN, P. L. FALB AND M. A. ARBIB, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1969.
- [18] L. V. KANTOROVICH AND G. P. AKILOV, *Functional Analysis in Normed Spaces*, Macmillan, New York, 1964.
- [19] D. L. KLEINMAN, T. FORTMAN AND M. ATHANS, *On the design of linear systems with piecewise constant feedback gains*, IEEE Trans. Automatic Control, 13 (1968), pp. 354–361.
- [20] V. KUCERA, *A contribution to matrix quadratic equations*, Ibid., 17 (1972), pp. 344–347.
- [21] ———, *On nonnegative definite solutions to matrix quadratic equations*, Automatica—J. IFAC, 8 (1972), pp. 413–423.
- [22] P. LANCASTER, *Theory of Matrices*, Academic Press, New York, 1969.
- [23] ———, *Explicit solutions of linear matrix equation*, SIAM Rev., 12 (1970), pp. 544–566.
- [24] L. MARKUS, *Optimal control of limit cycles or what control theory can do to cure a heart attack or cause one*, Lecture Notes in Mathematics, Springer-Verlag, New York, 1973, pp. 108–134.
- [25] A. W. MARSHALL AND I. OLKIN, *Norms and inequalities for condition numbers, II*, Linear Algebra and Appl., 2 (1969), pp. 167–172.
- [26] JOSE L. MASSERA, *The existence of periodic solutions of systems of differential equations*, Duke Math. J., 17 (1950), pp. 457–475.
- [27] H. J. PAYNE, *An alternative proof related to the algebraic Riccati equation*, IEEE Trans. Automatic Control, 17 (1972), p. 822.
- [28] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.
- [29] W. T. REID, *Ordinary Differential Equations*, John Wiley, New York, 1971.
- [30] ———, *Riccati Differential Equation*, Academic Press, New York, 1972.
- [31] A. SCHUMITZKY, *On the equivalence between matrix Riccati equations and Fredholm resolvents*, J. Comput. System. Sci., 2 (1968), pp. 76–87.
- [32] L. SILVERMAN AND H. E. MEADOWS, *Controllability and observability in time-variable linear systems*, this Journal, 5 (1967), pp. 64–72.
- [33] L. WEISS, *On the structural theory of linear differential systems*, this Journal, 6 (1968), pp. 659–680.
- [34] W. N. WONHAM, *Matrix Riccati equations of stochastic control*, this Journal, 6 (1968), pp. 681–697.
- [35] T. YOSHIZAWA, *Stability Theory by Liapunov's Second Method*, Mathematical Society of Japan, Tokyo, 1966.

TIME-VARYING SYSTEMS*

MICHAEL A. ARBIB† AND ERNEST G. MANES‡

Abstract. Section 1 provides a theory of reachability, observability, minimal realization and duality for time-varying linear systems, using only the basic language of linear algebra. Section 2 uses category theory to show that time-varying dynamics for adjoint processes in a category \mathcal{K} may be defined as adjoint processes in a suitable new category \mathcal{K}^z .

Introduction. This paper on time-varying systems is made up of two sections, which may be read independently of one another.

Those readers, still relatively few, who have studied the application of category theory to control problems may turn directly to § 2 in which we show that any adjoint process admits a theory of time-varying systems, defined by a new process which is itself adjoint.

The first section is written for the reader who finds the preceding paragraph obscure. Basically, it caters to those familiar with time-invariant linear systems, and shows how to generalize such concepts as reachability, observability, minimal realization and duality, to the time-varying case. A number of these results are familiar from the literature. Our main contribution is to give the material an algebraic formulation which is not only natural for linear systems, but which is applicable to all those systems which we have termed “decomposable” in [2], including the group machines of Brockett and Willsky [5]. Thus, while § 1 provides a concise algebraic treatment of basic topics of time-varying linear systems, it has a more general applicability which may be appreciated by the reader familiar with [2].

1. Linear systems. In this first section, we recall the main elements of our formulation of the theory of time-invariant linear systems, and then show how naturally it generalizes to a theory of time-varying linear systems. While the section is written for the reader with a basic knowledge of linear algebra, but no knowledge of category theory, we shall so structure the section that a reader acquainted with our “Foundations of System Theory: Decomposable Systems” [2] may appreciate the more general applicability of our theory, e.g., to linear systems over R -modules, or to group machines. To this end, we start by listing the basic concepts for the general theory, together with their interpretation for linear systems over vector spaces.

1.1. “The category \mathcal{K} ” specializes to: the collection of all vector spaces and linear maps between them. A \mathcal{K} -object is then simply a vector space, while a \mathcal{K} -morphism $f : A \rightarrow B$ is a linear map from vector space A to vector space B .

1.2. “ I has a countable copower $(\text{in}_j : I \rightarrow I^{\mathbb{N}} | j \in \mathbb{N})$ ” specializes to: Given a vector space I we form a new vector space

$$I^{\mathbb{N}} = \{(\cdot \cdot \cdot, i_j, \cdot \cdot \cdot, i_1, i_0) | \text{each } i_j \in I; \text{ only finitely many } i_j \text{ are nonzero}\}$$

* Received by the editors November 19, 1973 and in revised form July 14, 1974. This research was supported in part by the National Science Foundation under Grant GJ35759.

† Department of Computer and Information Science, University of Massachusetts, Amherst, Massachusetts 01002.

‡ Department of Mathematics, University of Massachusetts, Amherst, Massachusetts 01002.

comprising all left-infinite I -sequences of finite support. Addition and multiplication by a scalar are component-wise: $(\lambda i + \lambda' i')_j = \lambda i_j + \lambda' i'_j$. I^\S is equipped with the linear maps

$$\text{in}_j : I \rightarrow I^\S : i_0 \mapsto (\cdots, 0, \cdots, \underset{\substack{\uparrow \\ j \text{ position}}}{i_0}, \cdots, 0)$$

(injection to the j position) as well as a left-shift

$$z : I^\S \rightarrow I^\S : (\cdots, i_j, \cdots, i_1, i_0) \mapsto (\cdots, i_{j-1}, \cdots, i_0, 0).$$

The crucial property is that given any sequence of \mathcal{K} -morphisms (linear maps!) $f_j : I \rightarrow Q$, there exists a unique linear map f such that the diagram

$$(1.1) \quad \begin{array}{ccc} I & \xrightarrow{\text{in}_j} & I^\S \\ & \searrow f_j & \downarrow f \\ & & Q \end{array}$$

commutes for every $j \in \mathbf{N}$ (i.e., $f \cdot \text{in}_j = f_j$ for each j ; we dash the f -arrow to indicate that f is to be constructed so as to make the diagram commute). f is simply defined by the rule $f(\cdots, i_j, \cdots, i_1, i_0) = \sum_j f_j(i_j)$, the sum being well-defined since elements of I^\S have finite support. Note that z is constructed in just this way for the choice $Q = I^\S$, $f_j = \text{in}_{j+1}$, as the reader may easily verify.

1.3. “ Y has a countable power ($\pi_k : Y_\S \rightarrow Y | k \in \mathbf{N}$)” specializes to: Given a vector space Y , we form a new vector space

$$Y_\S = \{(y_0, y_1, \cdots, y_k, \cdots) \mid \text{each } y_k \in Y\}$$

comprising all right-infinite Y -sequences under component-wise addition and multiplication by a scalar. Y_\S is equipped with the linear maps

$$\pi_k : Y_\S \rightarrow Y : (y_0, y_1, \cdots, y_k, \cdots) \mapsto y_k$$

(projection from the k position) as well as a left shift

$$z : Y_\S \rightarrow Y_\S : (y_0, y_1, \cdots, y_k, \cdots) \mapsto (y_1, y_2, \cdots, y_{k+1}, \cdots).$$

The crucial property is that, given any sequence of \mathcal{K} -morphisms $g_k : Q \rightarrow Y$ there exists a unique linear map g such that the diagram

$$(1.2) \quad \begin{array}{ccc} Y & \xleftarrow{\pi_k} & Y_\S \\ & \nwarrow g_k & \uparrow g \\ & & Q \end{array}$$

commutes for every $k \in \mathbf{N}$. (Note that (1.2) looks just like (1.1), save that we have reversed all the arrows. This pairing of concepts under arrow-reversal is what category-theorists call *duality*. The duality of (1.1) and (1.2) will prove, below, to be the heart of the well-known reachability-observability duality of linear system theory.) g is simply defined by the rule $g(q) = (g_0(q), g_1(q), \dots, g_k(q), \dots)$. Note that $z : Y_{\S} \rightarrow Y_{\S}$ is constructed in just this way for the choice $Q = Y_{\S}$, $f_k = \pi_{k+1}$, as the reader may easily verify.

With this terminology we may recall, in conversational style, the basic concepts of [2]. [In the exemplary case of vector spaces and linear maps, this is simply a recasting of the approach of Kalman [8].]

A (time-invariant, decomposable) *system dynamics* is simply a pair (Q, F) , where $F : Q \rightarrow Q$ is a linear map (zero-input state-transition map). Given two dynamics (Q, F) and (Q', F') we say a linear map $g : Q \rightarrow Q'$ is a *dynamorphism from (Q, F) to (Q', F')* if the diagram

$$\begin{array}{ccc} Q & \xrightarrow{F} & Q \\ g \downarrow & & \downarrow g \\ Q' & \xrightarrow{F'} & Q' \end{array}$$

commutes ($F'g = gF$), i.e., g is compatible with the dynamics. To get a *system* we simply add to the dynamics an *input map* $G : I \rightarrow Q$ and an *output map* $H : Q \rightarrow Y$. The system $M = (Q, F, I, G, Y, H)$ then has *reachability map* $r : I^{\S} \rightarrow Q$ defined by the (1.1)-type diagram

$$\begin{array}{ccc} I & \xrightarrow{\text{in}_i} & I^{\S} \\ & \searrow F^j G & \downarrow r \\ & & Q \end{array}$$

(so that the state *reached* from the zero state by applying input sequence $(\dots, i_j, \dots, i_1, i_0)$ is $\sum_i F^j G i_j$, a familiar formula); while M 's *observability map* $\sigma : Q \rightarrow Y_{\S}$ is defined by the (1.2)-type diagram

$$\begin{array}{ccc} Y & \xleftarrow{\pi_k} & Y_{\S} \\ & \nwarrow HF^k & \uparrow \sigma \\ & & Q \end{array}$$

(so that the sequence *observed* if M is started in state q and fed only zeros for input is just $\sigma(q) = (Hq, HFq, \dots, HF^k q, \dots)$, again a familiar formula).

Two specific dynamics are the input dynamics (I^\S, z) and the output dynamics (Y_\S, z) for each of which the zero-input transition is just the left shift. Of particular interest is the fact that *the reachability map is a dynamorphism* $r : (I^\S, z) \rightarrow (Q, F)$

$$\begin{array}{ccc} I^\S & \xrightarrow{z} & I^\S \\ r \downarrow & & \downarrow r \\ Q & \xrightarrow{F} & Q \end{array}$$

and *the observability map is a dynamorphism* $\sigma : (Q, F) \rightarrow (Y_\S, z)$

$$\begin{array}{ccc} Y_\S & \xleftarrow{z} & Y_\S \\ \sigma \uparrow & & \uparrow \sigma \\ Q & \xleftarrow{F} & Q \end{array}$$

as the reader may easily check. (As we spell out in [2], this corresponds to Kalman's observations on $K[z]$ -homomorphisms.)

We are now ready to present our theory of time-varying linear systems, in which the input space I and output space Y are fixed, but the state-space may vary with time, being Q_k at time k , and the behavior of the system is described by the equations

$$\begin{aligned} q_{k+1} &= F_k q_k + G_{k+1} i_k, & q_{k+1} &\in Q_{k+1}, & q_k &\in Q_k, & i_k &\in I, \\ y_k &= H_k q_k, & y_k &\in Y, \end{aligned}$$

for linear maps $F_k : Q_k \rightarrow Q_{k+1}$, $G_{k+1} : I \rightarrow Q_{k+1}$ and $H_k : Q_k \rightarrow Y$.

As before, we shall write for the reader unfamiliar with category theory. A few comments, enclosed in $\langle\langle \cdot \cdot \cdot \rangle\rangle$ will be addressed to the reader familiar with [2], and should be omitted by readers familiar only with linear algebra.

1.4. A (time-varying) *system dynamics* in \mathcal{K} is a sequence $(Q_k, F_k | k \in \mathbf{Z}) = (Q, F)$, where each Q_k is a \mathcal{K} -object, and each F_k is a \mathcal{K} -morphism $Q_k \rightarrow Q_{k+1}$.¹ A dynamorphism $g : (Q, F) \rightarrow (Q', F')$ is a sequence of \mathcal{K} -morphisms $(g_k : Q_k \rightarrow Q'_k | k \in \mathbf{Z})$ such that $F' \cdot g = g \cdot F$ in the sense that

¹ $\langle\langle$ In case $\mathcal{K} = \mathbf{R}\text{-Mod}$, note that a dynamics which satisfies $F_{k+1}F_k = 0$ is what homologists call a *chain complex*. $\rangle\rangle$

$$\begin{array}{ccc}
 Q_k & \xrightarrow{F_k} & Q_{k+1} \\
 g_k \downarrow & & \downarrow g_{k+1} \\
 Q'_k & \xrightarrow{F'_k} & Q'_{k+1}
 \end{array}$$

commutes for all $k \in \mathbf{Z}$. «It is obvious that system dynamics and dynamorphisms form a category $\text{Dyn}_{\mathbf{Z}}(\mathcal{K})$:» Clearly, $(id_{Q_k} | k \in \mathbf{Z})$ is a dynamorphism $(Q, F) \rightarrow (Q, F)$; while if $g : (Q, F) \rightarrow (Q', F')$ and $h : (Q', F') \rightarrow (Q'', F'')$ are dynamorphisms, then their composition $h \cdot g$ defined by $(h \cdot g)_k = h_k \cdot g_k$ is a dynamorphism $(Q, F) \rightarrow (Q'', F'')$.

1.5. A (time-varying) (*decomposable*) system in \mathcal{K} is a 6-tuple $M = (Q, F, I, Y, G, H)$ such that (Q, F) is a time-varying system dynamics in \mathcal{K} , G is a sequence of \mathcal{K} -morphisms of the form $(G_k : I \rightarrow Q_k | k \in \mathbf{Z})$ (the *input map*) and H is a sequence of \mathcal{K} -morphisms of the form $(H_k : Q_k \rightarrow Y | k \in \mathbf{Z})$ (the *output map*).

«As in the theory of time-invariant systems, we shall assume \mathcal{K} such that I has a countable copower $(in_j : I \rightarrow I^{\mathbb{N}} | j \in \mathbf{N})$ while Y has a countable power $(\pi_k : Y_{\mathbb{N}} \rightarrow Y | k \in \mathbf{N})$. However, we shall make crucial use of the full properties of the copower and power, not just the simple recursion and simple corecursion used in the theory of time-invariant systems.»

We shall identify the time-invariant (Q, F) with time-varying (Q, F) for which $Q_k = Q$ and $F_k = F$ for all $k \in \mathbf{Z}$; while a dynamorphism $g : (Q, F) \rightarrow (Q', F')$ of time-invariant systems may be viewed as the time-varying g for which $g_k = g : Q \rightarrow Q'$ for all $k \in \mathbf{Z}$. «Thus, we may regard $\text{Dyn}(\mathcal{K})$ as a subcategory of $\text{Dyn}_{\mathbf{Z}}(\mathcal{K})$.»

Thus we have the *input dynamics* $(I^{\mathbb{N}}, z)$ and the *output dynamics* $(Y_{\mathbb{N}}, z)$ of the time-invariant theory available in our more general setting. With this observation, it is straightforward to define the *reachability* and *observability* maps of any system.

The reader may find the following intuition for the linear case useful.

For each input sequence $i = (\cdots, i_j, \cdots, i_1, i_0)$ in $I^{\mathbb{N}}$, $r_k(i)$ in Q_k is the state reached at time k if the sequence i is applied through time $k-1$, i.e., just in case i_j is applied at time $k-j-1$ for all $j \geq 0$. We may call $r_k : I^{\mathbb{N}} \rightarrow Q_k$ the *reachability map* at time k .

For each state q_k in Q_k , the sequence $\sigma_k(q_k) = (y_0, y_1, y_2, \cdots, y_j, \cdots)$ in $Y_{\mathbb{N}}$ is the sequence of outputs generated by the system if started in state q_k at time k , and fed zero inputs thereafter with y_j being the output emitted at time $k+j$. We may call $\sigma_k : Q_k \rightarrow Y_{\mathbb{N}}$ the *observability map* at time k .

For each input sequence i in $I^{\mathbb{N}}$, $f_k^{\bullet}(i)$ in $Y_{\mathbb{N}}$ is the sequence of outputs emitted from time k on if the system is fed input sequence i through time $k-1$, and 0 inputs from time k on. We may call $f_k^{\bullet} : I^{\mathbb{N}} \rightarrow Y_{\mathbb{N}}$ the *total response* at time k on.

However, useful though these intuitions are, the interesting point is that we define r , σ and f^{\bullet} as dynamorphisms, using the crucial properties (1.1) and (1.2) to ensure that all components may be defined simultaneously.

1.6. THEOREM. Given a decomposable system $M = (Q, F, I, Y, G, H)$, we define its reachability map $r : (I^{\S}, z) \rightarrow (Q, F)$ to be the sequence $r_k : I^{\S} \rightarrow Q_k$ defined by the diagrams

$$(1.3) \quad \begin{array}{ccccc} I & \xrightarrow{\text{in}_0} & I^{\S} & \xrightarrow{z} & I^{\S} \\ & \searrow G_k & \downarrow r_k & & \downarrow r_{k+1} \\ & & Q_k & \xrightarrow{F_k} & Q_{k+1} \end{array}$$

Then r is uniquely defined and a dynamorphism.

Proof. For all k , set $\Phi_{k,k} = \text{id}_{Q_k}$, while for $k > l$ set $\Phi_{k,l} = F_{k-1} \cdots F_{l+1} F_l : Q_l \rightarrow Q_k$. Now, for (1.3) to hold we must have $r_k \cdot \text{in}_j = \Phi_{k,k-j} G_{k-j}$ for each $k \in \mathbb{Z}$ for all $j \geq 0$

$$\begin{array}{ccc} I & \xrightarrow{\text{in}_j} & I^{\S} \\ & \searrow \Phi_{k,k-j} G_{k-j} & \downarrow r_k \\ & & Q_k \end{array}$$

and, by diagram (1.1), this defines each r_k , and thus r , uniquely. But then this r clearly satisfies (1.3), and the square in (1.3) says that $r : (I^{\S}, z) \rightarrow (Q, F)$ is a dynamorphism. \square

1.7. THEOREM. Given a decomposable system $M = (Q, F, I, Y, G, H)$, we define its observability map $\sigma : (Q, F) \rightarrow (Y^{\S}, z)$ to be the sequence $\sigma_k : Q_k \rightarrow Y^{\S}$ defined by the diagrams

$$(1.4) \quad \begin{array}{ccccc} Y & \xleftarrow{\pi_0} & Y^{\S} & \xleftarrow{z} & Y^{\S} \\ & \nwarrow H_k & \uparrow \sigma_k & & \uparrow \sigma_{k-1} \\ & & Q_k & \xleftarrow{F_{k-1}} & Q_{k-1} \end{array}$$

Then σ is uniquely defined and a dynamorphism.

Proof. As above, but noting from (1.4) that $\pi_j \cdot \sigma_k = H_{k+j} \Phi_{k+j,k}$ so that we may use (1.2). \square

1.8. DEFINITION. The total response of M is the composition

$$f^{\Delta} = \sigma \cdot r : (I^{\S}, z) \rightarrow (Y^{\S}, z).$$

Note that even though f^{Δ} is, as the composition of the dynamorphism r and σ , a dynamorphism of time-invariant systems, it is *not*, in general, time-invariant since for each j, k, l ,

$$\begin{aligned}
 \pi_j \cdot f_k^\Delta \cdot \text{in}_l &= (\pi_j \cdot \sigma_k) \cdot (r_k \cdot \text{in}_l) \\
 &= (H_{k+j} \cdot \Phi_{k+j,k}) \cdot (\Phi_{k,k-l} \cdot G_{k-l}) \\
 &= H_{k+j} \cdot \Phi_{k+j,k-l} \cdot G_{k-l},
 \end{aligned}$$

which reduces to the familiar $HF^{j+l}G$ in the time-invariant case. «Thus $\text{Dyn}(\mathcal{H})$ is *not* a full subcategory of $\text{Dyn}_{\mathbf{Z}}(\mathcal{H})$.» The above formula is already known in the linear system case (see, e.g., [10], noting that Weiss denotes our H_{k+j} by H_{k+j+1} and our $\Phi_{k+j,k-l}$ by $\Phi_{k+j,k-l+1}$). Our point, of course, is that this formula was obtained by diagram-chasing in a way which makes it applicable, as the reader of [2] will appreciate, to a number of nonlinear cases, including the group machines of [5].

As in the time-invariant case, we can associate two nondynamorphisms with M , from either of which we can reconstitute f^Δ .

1.9. DEFINITION. The *response map of M at time k* is

$$f_k = \pi_0 \cdot f_k^\Delta : I^\S \rightarrow Y$$

while the *impulse response of M from time k* is

$$\hat{f}_k = f_k^\Delta \cdot \text{in}_0 : I \rightarrow Y_\S.$$

Let \mathcal{E} be the set of *onto* linear maps, and \mathcal{M} be the set of *one-to-one* linear maps. Then *any* linear map $g : A \rightarrow B$ may be factored as

$$A \xrightarrow{g} B = A \xrightarrow{e} g(A) \xrightarrow{m} B,$$

where $e(a) = g(a)$ for each $a \in A$, while $m(b) = b$ for each $b \in g(A) \subset B$; and we note that $e \in \mathcal{E}$ and $m \in \mathcal{M}$. We refer to *any* pair (e', m') such that $m' \in \mathcal{M}$ and $e' \in \mathcal{E}$ with $g = m' \cdot e'$ as an $\mathcal{E} - \mathcal{M}$ factorization of g . «To define *reachability* and *observability* we must fix a choice $(\mathcal{E}, \mathcal{M})$ of an *image factorization system* on \mathcal{H} .» We then have the following.

1.10. DEFINITION. We say M is *reachable at time k* if $r_k \in \mathcal{E}$; *completely reachable* if every $r_k \in \mathcal{E}$; *observable at time k* if $\sigma_k \in \mathcal{M}$; *completely observable* if every $\sigma_k \in \mathcal{M}$.

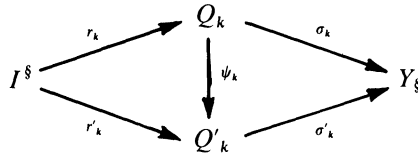
We now approach the realization problem.

1.11. DEFINITION. Fixing I and Y , but letting the state-space $(Q_k | k \in \mathbf{Z})$ vary, given systems $M = (Q, F, G, H)$ and $M' = (Q', F', G', H')$, a *simulation* $\psi : M \rightarrow M'$ (we say M *simulates* M') is a dynamorphism $\psi : (Q, F) \rightarrow (Q', F')$ which commutes with the input and output

$$\begin{array}{ccccc}
 & & Q_k & \xrightarrow{F_k} & Q_{k+1} & & \\
 & \nearrow G_k & \downarrow \psi_k & & \downarrow \psi_{k+1} & \searrow H_{k+1} & \\
 (1.5) \quad I & & & & & & Y \\
 & \searrow G'_k & Q'_k & \xrightarrow{F'_k} & Q'_{k+1} & \nearrow H'_{k+1} &
 \end{array}$$

We say M is *isomorphic* to M' if there exists a simulation ψ such that each $\psi_k : Q_k \rightarrow Q'_k$ is one-to-one and onto.

It is an immediate consequence of the definitions of countable power and copower that a dynamorphism ψ satisfies (1.5) if and only if it commutes with the reachability and observability maps



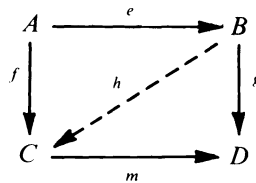
1.12. DEFINITION. We say a system M is a *realization* of a total response f^Δ if f^Δ is the total response of M . We say M is a *minimal realization* of f^Δ if it is a completely reachable realization of f^Δ which can be simulated by every other completely reachable realization of f^Δ .

It is an easy consequence of this definition that if M and M' are both minimal realizations of f^Δ they must be isomorphic.

1.13. DEFINITION. M is a *canonical realization* of the total response f^Δ if it is a realization of f^Δ which is completely reachable and completely observable.

We now recall [8, p. 256].

1.14. FILL-IN LEMMA. *Given the commutative square*

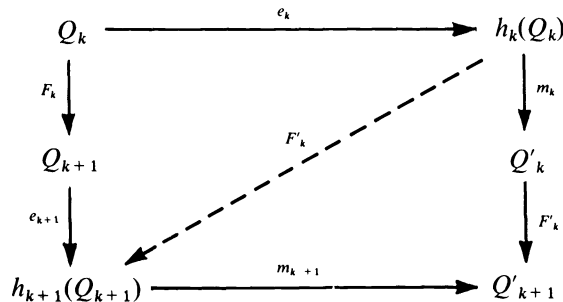


with $e \in \mathcal{E}$ and $m \in \mathcal{M}$, there exists a unique h such that the entire diagram commutes.

With this we can prove the two lemmas crucial to our minimal realization theorem.

1.15. DYNAMORPHIC IMAGE LEMMA. *Let $h : (Q, F) \rightarrow (Q', F')$ be a dynamorphism, and let each (e_k, m_k) be an $\mathcal{E} - \mathcal{M}$ factorization of $h_k : Q_k \rightarrow Q'_k$ for each $k \in \mathbb{Z}$. Then there exists a unique dynamical structure $F'' = (F''_k : h_k(Q_k) \rightarrow h_{k+1}(Q_{k+1}))$ on $h(Q)$ such that $e : (Q, F) \rightarrow (h(Q), F'')$ and $m : (h(Q), F'') \rightarrow (Q', F')$ are dynamorphisms.*

Proof. The proof is immediate from the diagonal fill-in lemma and the diagram below.



□

Again, a straightforward extension of the time-invariant argument yields the following.

1.16. SIMULATION LEMMA. *Let M be a completely reachable realization of $f^\Delta : I^\S \rightarrow Y_\S$ and let M' be a completely observable realization of f^Δ . Then there exists a unique simulation $\psi : M \rightarrow M'$.*

1.17. MINIMAL REALIZATION THEOREM FOR TIME-VARYING DECOMPOSABLE SYSTEMS. *Every dynamorphism $f^\Delta : (I^\S, z) \rightarrow (Y_\S, z)$ has a minimal realization M_f . A system M is a minimal realization of f^Δ if and only if M is a canonical realization of f^Δ .*

Proof. For each k , let (r_k^f, σ_k^f) be an $\mathcal{E}\text{-}\mathcal{M}$ factorization of f_k^Δ , and let $Q_k^f = f_k^\Delta(I^\S)$ be the corresponding image. By the dynamorphic image lemma there exists a unique dynamics F_f on Q^f such that $r^f : (I^\S, z) \rightarrow (Q^f, F_f)$ and $\sigma_f : (Q^f, F_f) \rightarrow (Y_\S, z)$ are dynamorphisms. Define $G^f = (G_k^f I \rightarrow Q_k^f | k \in \mathbb{Z})$ and $H^f = (H_k^f : Q_k \rightarrow Y | k \in \mathbb{Z})$ by

$$\begin{array}{ccc} I & \xrightarrow{\text{in}_\circ} & I_\S \\ & \searrow G_k^f & \downarrow r_k^f \\ & & Q_k^f \end{array} \quad \text{and} \quad \begin{array}{ccc} Y & \xleftarrow{\pi_\circ} & Y_\S \\ & \nwarrow H_k^f & \uparrow \sigma_k^f \\ & & Q_k^f \end{array}$$

Then $M_f = (Q^f, F_f, G^f, H^f)$ is a system whose reachability map is r^f and whose observability map is σ_f , and is thus a canonical realization of f^Δ . Since M_f is completely observable, it follows from the simulation lemma that M_f is minimal—and since all minimal realizations of f^Δ are isomorphic, they are certainly canonical.

Finally, let M be an arbitrary canonical realization of f . By the simulation lemma there exist unique simulations $\psi : M \rightarrow M_f$, $\psi' : M_f \rightarrow M$. That $\psi\psi' = id$ and $\psi'\psi = id$ also follows from the simulation lemma, and so M and M_f are isomorphic. \square

«Turning now to duality, recall that in the time-invariant case, we defined the dual of the system M ,

$$I \xrightarrow{G} Q \xrightarrow{F} Q \xrightarrow{H} Y,$$

in \mathcal{K} to be the system M^{op} ,

$$Y \xleftarrow{H} Q \xleftarrow{F} Q \xleftarrow{G} I,$$

in \mathcal{K}^{op} .»

For finite-dimensional time-invariant linear systems we define the dual of the system M characterized by

$$I \xrightarrow{G} Q \xrightarrow{F} Q \xrightarrow{H} Y$$

to be that M^* characterized by

$$Y \xrightarrow{H^*} Q \xrightarrow{F^*} Q \xrightarrow{G^*} 1.$$

However, in the time-varying case, we note that

$$I \xrightarrow{G_k} Q_k \xrightarrow{F_k} Q_{k+1} \xrightarrow{H_{k+1}} Y$$

is such that “time” increases from k to $k+1$ as we move from left to right. Thus in the duality theory for time-varying systems we must not only reverse the arrows in our diagrams but must *reverse the direction of “time”* as well.

«Given a system $M = (Q, F, I, G, Y, H)$ in \mathcal{K} , its *dual* is the system $M^{\text{op}} = (Q^{\text{op}}, F^{\text{op}}, Y, H^{\text{op}}, I, G^{\text{op}})$ in \mathcal{K}^{op} , where

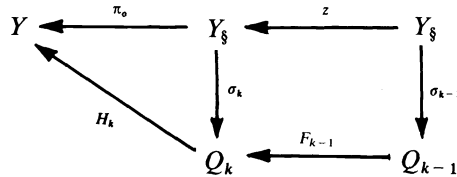
$$Q_k^{\text{op}} = Q_{-k},$$

$$F_k^{\text{op}} : Q_k^{\text{op}} \multimap Q_{k+1}^{\text{op}} = F_{-k-1} : Q_{-k-1} \rightarrow Q_{-k},$$

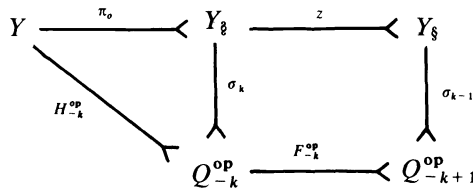
$$H_k^{\text{op}} : Y \multimap Q_k^{\text{op}} = H_{-k} : Q_{-k} \rightarrow Y,$$

$$G_k^{\text{op}} : Q_k^{\text{op}} \multimap I = G_{-k} : I \rightarrow Q_{-k}.$$

Using these substitutions, we see that the observability diagram for M



becomes

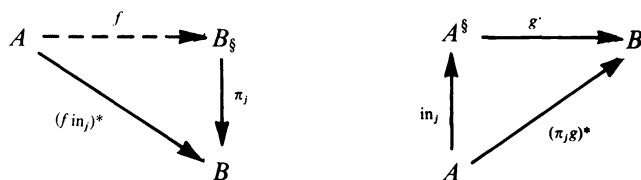


and, recalling that $(\pi_k : Y \multimap Y_§)$ is the *copower* in \mathcal{K}^{op} , we see that—consistent with time reversal—the reachability map of M^{op} at time $-k$ is just $r_{-k}^{\text{op}} = \sigma_k : Y_§ \multimap Q_{-k}^{\text{op}}$. Since $(M^{\text{op}})^{\text{op}} = M$, the observability map of M^{op} at time $-k$ must also be $\sigma_{-k}^{\text{op}} = r_k : Q_{-k}^{\text{op}} \multimap I^{\S}.$ »

Now the operation of taking the transpose only holds for maps $h : A \rightarrow B$ with finite-dimensional vector spaces A and B ; while I^{\S} and $Y_§$ are infinite-dimensional (unless I or Y is $\{0\}$). However, given

$$f : B^{\S} \rightarrow A \quad \text{and} \quad g : B \rightarrow A_§$$

we can define “pseudo-transposes” $f^* : A \rightarrow B_\S$ and $g^* : A^\S \rightarrow B$ by the diagrams



and it can be easily seen [2, Lemma 5.10] that $f \in \mathcal{E}$ if and only if $f^* \in \mathcal{M}$ while (dually) $g \in \mathcal{M}$ if and only if $g^* \in \mathcal{E}$. Note that if r is the reachability map of M , then r^* is the observability map of M^* ; while σ^* is the reachability map of M^* for σ the observability map of M .

«Now recall that in the time-invariant case of [2] we fixed our category \mathcal{K} with countable powers and copowers and an image factorization system $(\mathcal{E}, \mathcal{M})$ to be such that there was a class \mathcal{F} of “finite-dimensional” objects of \mathcal{K} together with maps $*$: $\mathcal{K}(A, B) \rightarrow \mathcal{K}(B, A)$, for all A, B in \mathcal{F} , subject to the axioms:

1. Given A, B, C in \mathcal{F} and $f : A \rightarrow B$, $g : B \rightarrow C$, then $(gf)^* = f^*g^*$, $(\text{id}_A)^* = \text{id}_A$ and $(f^*)^* = f$.
2. If $e : A \rightarrow B$ is in \mathcal{E} with $A \in \mathcal{F}$, then $B \in \mathcal{F}$. If $m : A \rightarrow B \in \mathcal{M}$ with $B \in \mathcal{F}$, then $A \in \mathcal{F}$.
3. Given $f : A \rightarrow B$ with $A, B \in \mathcal{F}$, then $f : A \rightarrow B \in \mathcal{M} \Leftrightarrow f^* : B \rightarrow A \in \mathcal{E}$.

Given A, B in \mathcal{F} , and $f : B^\S \rightarrow A$ and $g : B \rightarrow A_\S$, we extended the $*$ correspondence to them by defining $f^* : A \rightarrow B_\S$ and $g^* : A^\S \rightarrow B$ by the diagrams



and we observed that for A, B, C in \mathcal{F} , for $f : A^\S \rightarrow B$, $g : A \rightarrow B_\S$, $t : B \rightarrow C$ and $u : C \rightarrow B$, we have

1. $f^{**} = f$ and $g^{**} = g$,
2. $(tf)^* = f^*t^*$ and $(gu)^* = u^*g^*$,
3. f is in $\mathcal{E} \Leftrightarrow f^*$ is in \mathcal{M} .

Let us now extend to the time-varying case the duality theory we based on these observations.

1.18. DEFINITION. «Let $(\mathcal{K}, \mathcal{E}, \mathcal{M}, \mathcal{F}, *)$ be as above.» A system $M = (Q, F, I, G, Y, H)$ is *finite-dimensional* if I, Y and each $(Q_k (k \in \mathbb{Z}))$ are finite-dimensional «i.e., in \mathcal{F} ». If M is finite-dimensional, we define its *dual system*

$$M^* = (Q^*, F^*, Y, H^*, I, G^*)$$

by the rules

$$\begin{aligned} Q_k^* &= Q_{-k}, \\ F_k^* &= (F_{-k-1})^* : Q_k^* \rightarrow Q_{k+1}^*, \\ H_k^* &= (H_{-k})^* : Y \rightarrow Q_k^*, \\ G_k^* &= (G_{-k})^* : Q_k^* \rightarrow I. \end{aligned}$$

Then M^* is finite-dimensional, and $(M^*)^* = M$.

1.19. DUALITY THEOREM FOR FINITE-DIMENSIONAL TIME-VARYING SYSTEMS. Let $M = (Q, F, I, G, Y, H)$ be a finite-dimensional system in \mathcal{K} with dual M^* , and with reachability map $r : (I^\S, z) \rightarrow (Q, F)$ and observability map $\sigma : (Q, F) \rightarrow (Y_\S, z)$. Then

- (i) $(r_{-k})^* : Q_{-k} \rightarrow I_\S$ is the observability map of M^* at time k ,
- (ii) $(\sigma_{-k})^* : Y_\S \rightarrow Q_{-k}$ is the reachability map of M^* at time k ,
- (iii) M is reachable (resp. observable) at time k if and only if M^* is observable (resp. reachable) at time $-k$. Thus M is completely reachable (resp. completely observable) if and only if M^* is completely observable (resp. completely reachable).

Proof. Recall the formulas $r_k \cdot in_j = \Phi_{k,k-j} G_{k-j}$ and $\pi_j \cdot \sigma_k = H_{k+j} \Phi_{k+j,k}$; and note that if we define Φ^* by $\Phi_{k,l}^* = F_{k-1}^* \cdots F_{l+1}^* F_l^*$ ($k > l$) while $\Phi_{k,k}^* = \text{id}_{Q_k}$, then $\Phi_{k,l}^* = (\Phi_{-l,-k})^*$. We thus have that

$$\begin{aligned} \pi_j \cdot (r_{-k})^* &= (r_{-k} \cdot in_j)^* \quad (\text{by definition of } ^*) \\ &= (\Phi_{-k,-k-j} G_{-k-j})^* \\ &= (G_{-k-j})^* \cdot (\Phi_{-k,-k-j})^* \\ &= G_{k+j}^* \Phi_{k+j,k}^* \end{aligned}$$

so that $(r_{-k})^*$ is indeed the observability map of M^* at time k . Part (ii) follows by duality. Part (iii) then follows from the observation that an f is in \mathcal{E} if and only if its f^* is in \mathcal{M} , and its dual. \square

2. Adjoint machines. In this section, we abandon our attempt to make the material accessible to readers unacquainted with category theory, and instead demonstrate that the theory of §1 is a special case of the theory of adjoint machines developed in [3]. (As we spelled out in [3], our theory bears interesting relations to the approach of Bainbridge [4], and generalizes the theory of Goguen [7].) In particular, we unify the minimal realization theory for time-varying linear systems with time-varying automata [6]. We show the richness of this concept in [3]—with adjoint machines including sequential machines, nondeterministic machines, Boolean machines, metric machines and topological machines. However, for now we need only assume the reader to be acquainted with our more introductory paper “Machines in a Category” [1]. In that paper, we gave the category theorist’s definition of functors and of left adjoints, and said that the input structure of a machine should not be regarded as a *set* of applicable inputs, but rather as a *process* which transforms the state-space Q into a new object QX on which the dynamics can act.

2.1. DEFINITION. Given a functor $X : \mathcal{K} \rightarrow \mathcal{K}$, $\text{Dyn}(X)$ denotes the category of X -dynamics whose objects are pairs (Q, δ) , where Q is a \mathcal{K} -object and $\delta : QX \rightarrow Q$ is a \mathcal{K} -morphism; while *dynamorphisms* $g : (Q, \delta) \rightarrow (Q', \delta')$ are \mathcal{K} -morphisms $g : Q \rightarrow Q'$ for which the diagram

$$\begin{array}{ccc} QX & \xrightarrow{\delta} & Q \\ gX \downarrow & & \downarrow g \\ Q'X & \xrightarrow{\delta'} & Q' \end{array}$$

commutes.

We then said that for X to be interesting, it must be an input process in the following sense.

2.2. DEFINITION. X is an *input process* if the forgetful functor $\text{Dyn}(X) \rightarrow \mathcal{K} : (Q, \delta) \mapsto Q$ has a left adjoint; i.e., if for each $Q \in \mathcal{K}$ there exists a free dynamics $\mu_0 : (QX^{\otimes})X \rightarrow QX^{\otimes}$ with a \mathcal{K} -morphism $\eta : Q \rightarrow QX^{\otimes}$ such that given any X -dynamics (Q', δ') and any \mathcal{K} -morphism $f : Q \rightarrow Q'$, there exists a unique dynamorphism $\psi : (QX^{\otimes}, \mu_0) \rightarrow (Q', \delta')$ such that $\psi \cdot \eta = f$:

$$\begin{array}{ccc} Q & \xrightarrow{\eta} & QX^{\otimes} \\ & \searrow f & \downarrow \psi \\ & & Q' \end{array} \qquad \begin{array}{ccc} (QX^{\otimes})X & \xrightarrow{\mu_0} & QX^{\otimes} \\ \psi X \downarrow & & \downarrow \psi \\ Q'X & \xrightarrow{\delta'} & Q' \end{array}$$

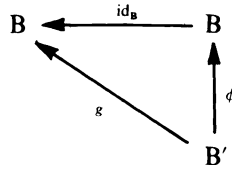
We assume that the reader knows (or can look up in MacLane [9]) the categorical definition of coproducts. We shall next define what it means for a functor to have a *right adjoint*, and establish broad conditions under which an $X : \mathcal{K} \rightarrow \mathcal{K}$ with right adjoint will be an input process.

2.3. DEFINITION. A functor $F : \mathcal{A} \rightarrow \mathcal{B}$ has a *right adjoint* if there exists a functor $F^* : \mathcal{B} \rightarrow \mathcal{A}$ (the right adjoint of F) such that to each B in \mathcal{B} there corresponds a \mathcal{B} -morphism $BF^*F \rightarrow B$ such that to each \mathcal{B} -morphism $g : B'F \rightarrow B$ there corresponds a unique \mathcal{A} -morphism $\phi : B' \rightarrow BF^*$ such that

$$(2.1) \quad \begin{array}{ccc} B & \xleftarrow{\varepsilon} & BF^*F \\ & \nwarrow g & \uparrow \phi F \\ & & B'F \end{array} \qquad \begin{array}{c} BF^* \\ \uparrow \phi \\ B' \end{array}$$

commutes.

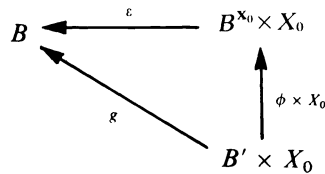
2.4. Example. Let $\text{Vect} = \langle \text{Vector spaces and Linear maps} \rangle$, and let $X : \text{Vect} \rightarrow \text{Vect}$ be the identity functor ($Q \mapsto Q; f \mapsto f$). Then an X -dynamics is just a linear map $F : QX = Q \rightarrow Q$, as in the linear time-invariant systems of § 1. Clearly X is its own right adjoint—setting $F = F^* = X$, $\varepsilon = id_B$, we have that (2.1) is satisfied with $\phi = g$:



2.5. *Example.* Let $\text{Set} = \langle \text{Sets and Maps} \rangle$, and let $X = - \times X_0 : \text{Set} \rightarrow \text{Set}$ be the functor $Q \mapsto Q \times X_0$; $f \mapsto f \times X_0$, where

$$f \times X_0 : Q \times X_0 \rightarrow Q' \times X_0 : (q, x) \mapsto (f(q), x).$$

Then an X -dynamics is just a map $\delta : Q \times X_0 \rightarrow Q$, the next-state function of a sequential machine. $X = - \times X_0$ has right adjoint $(-)^{X_0}$ which sends Q to the set Q^{X_0} of all maps from X_0 to Q . $\varepsilon : B^{X_0} \times X_0 \rightarrow B$ is the *evaluation* $(f, x) \mapsto f(x)$, and we have that (2.1) is satisfied on taking $\phi(b') : X_0 \rightarrow B : x \mapsto g(b', x)$:



We recall the following standard result.

2.6. LEMMA. *If $F : \mathcal{A} \rightarrow \mathcal{B}$ has a right adjoint, then F preserves coproducts.*

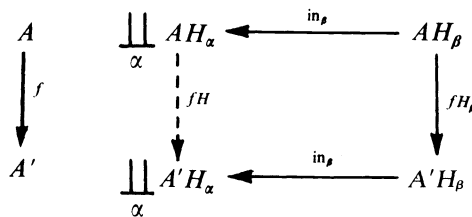
Let $(\mathcal{A} \xrightarrow{H_\alpha} \mathcal{B} : \alpha \in I)$ be a collection of functors; and assume that \mathcal{B} has I -indexed coproducts. Then we can form the functor

$$H = \coprod_{\alpha \in I} H_\alpha$$

(uniquely up to isomorphism) by

$$AH = \coprod_{\alpha} AH_\alpha$$

while



It is then straightforward to check that $\coprod_{\alpha} H_{\alpha}$ is a functor. Now, given any functor $X : \mathcal{K} \rightarrow \mathcal{K}$, define X^n for any $n \geq 0$ by the rules

$$X_0 = id, \quad X^{n+1} = X^n \cdot X \quad \text{for } n \geq 0,$$

and then set

$$X^* = \coprod_{n \geq 0} X^n.$$

2.7. THEOREM. *Let \mathcal{K} have and let $X : \mathcal{K} \rightarrow \mathcal{K}$ preserve countable coproducts. Then X is an input process and $X^{\oplus} = X^*$.*

Proof. If $IX^{\oplus} = IX^*$ we must define $\mu_0 : IX^*X \rightarrow IX^*$ and $\eta : I \rightarrow IX^*$ as in 2.2. Since X preserves the coproduct $IX^n \rightarrow IX^*$, we have that

$$IX^{n+1} = IX^n X \xrightarrow{\text{in}_n X} IX^* X$$

is also a coproduct. Thus we may define $I\mu_0$ by the obvious rule

$$\begin{array}{ccc} IX^* X & \xrightarrow{I\mu_0} & IX^* \\ \uparrow \text{in}_n X & \nearrow \text{in}_{n+1} & \\ IX^{n+1} & & \end{array}$$

which certainly reduces to the familiar story in case $X = - \times X_0 : \text{Set} \rightarrow \text{Set}$. We define $\eta : I \rightarrow IX^*$ to be simply in_0 . Let us check that this works, i.e., that the diagrams

$$\begin{array}{ccc} I & \xrightarrow{\text{in}_0} & IX^* \\ & \searrow f & \downarrow \psi \\ & & Q \end{array} \qquad \begin{array}{ccccc} & & IX^{n+1} & & \\ & \swarrow \text{in}_n X & & \searrow \text{in}_{n+1} & \\ IX^* X & \xrightarrow{\mu_0} & IX^* & & \\ \downarrow \psi X & & \downarrow \psi & & \\ QX & \xrightarrow{\delta} & Q & & \end{array}$$

define a unique $\psi : IX^* \rightarrow Q$. But the left-hand diagram says

$$\psi \cdot \text{in}_0 = f$$

while the right-hand diagram asserts that

$$\psi \cdot \text{in}_{n+1} = \delta \cdot \psi X \cdot \text{in}_n X = \delta \cdot (\psi \text{ in}_n) X, \quad n \geq 0,$$

and these equations define the unique ψ which satisfies the diagrams. \square

2.8. COROLLARY. If \mathcal{K} has countable coproducts and $X : \mathcal{K} \rightarrow \mathcal{K}$ has a right adjoint, then X is an input process, and $X^\circledast = X^*$. We say such an X is an adjoint process.

[Goguen [7] has a result equivalent to this for the special case in which X is of the form $-\otimes X_0$, for X_0 an object of some closed category (K, \otimes, I) .]

We now show that if X is an adjoint process in \mathcal{K} , then X induces an adjoint process \bar{X} which yields “time-varying” X -dynamics. The construction of the category $\mathcal{K}^{\mathbb{Z}}$ in which \bar{X} lives is clearly motivated by the treatment of the identity process X of Example 2.4 as given in Definition 1.4:

$\mathcal{K}^{\mathbb{Z}}$ -Objects: Sequences $Q = (Q_k | k \in \mathbb{Z})$ with each Q_k in \mathcal{K} .

$\mathcal{K}^{\mathbb{Z}}$ -Morphisms: $f : Q \rightarrow Q' = (f_k : Q_k \rightarrow Q'_k | k \in \mathbb{Z})$ with each f_k in \mathcal{K} .

$\mathcal{K}^{\mathbb{Z}}$ is a category with $(f \cdot g)_k = f_k \cdot g_k$; $(id_Q)_k = (id_{Q_k})$.

[Note: The underlying scheme could be far more complex than \mathbb{Z} . The tool for developing the implications of this observation is the notion of a *functor category* (see [9] for a discussion).]

2.9. THEOREM. Let $X : \mathcal{K} \rightarrow \mathcal{K}$ where \mathcal{K} has countable coproducts. Then so too does $\mathcal{K}^{\mathbb{Z}}$, and if X

(i) preserves countable coproducts; or

(ii) has a right adjoint,

then so does the functor $\bar{X} : \mathcal{K}^{\mathbb{Z}} \rightarrow \mathcal{K}^{\mathbb{Z}}$ defined by

$$(\bar{X})_k = Q_{k-1}X$$

and

$$(f\bar{X})_k = f_{k-1}X : Q_{k-1}X \rightarrow Q'_{k-1}X.$$

Proof. Given $\mathcal{K}^{\mathbb{Z}}$ -objects Q^n , one for each $n \in \mathbb{N}$ [with $(Q^n)_k = Q_k^n$], we define

$$\left(\coprod_n Q^n \right)_k = \coprod_n Q_k^n \quad \text{for each } k \in \mathbb{Z}$$

and it is easy to check, with the obvious injections, that this is a coproduct in $\mathcal{K}^{\mathbb{Z}}$.

Now, to say that \bar{X} has a right adjoint means we can solve

$$\frac{A\bar{X} \rightarrow B}{A \rightarrow B\bar{X}^*}$$

for suitable $B\bar{X}^*$. But note that we have the correspondence

$$\frac{A_{k-1}X \rightarrow B_k}{A_{k-1} \rightarrow B_k X^*}$$

so that we define

$$(B\bar{X}^*)_k = B_{k+1}X^*,$$

and a straightforward computation shows that \bar{X}^* is indeed the right adjoint of \bar{X} . Note that taking the adjoint “reverses” the direction of “time”.

We omit the straightforward verification that if \mathcal{K} has and X preserves countable coproducts, then \bar{X} too preserves countable coproducts. \square

Now an element of $\text{Dyn}(\bar{X})$ is just a $\mathcal{H}^{\mathbb{Z}}$ -morphism

$$\delta : Q\bar{X} \rightarrow Q$$

defined by $\delta_k : Q_{k-1}X \rightarrow Q_k$ for each k —i.e., a *time-varying X -dynamics!*

2.10. COROLLARY. *If $X : \mathcal{H} \rightarrow \mathcal{H}$ is an adjoint process, then the time-varying X -process \bar{X} is an adjoint process, with*

$$(\bar{X})^{\circledast} = \bar{X}^*.$$

Thus

$$(2.2) \quad [Q\bar{X}^*]_k = \left[\coprod_{n \geq 0} QX^n \right]_k = \coprod_{n \geq 0} Q_{k-n}X^n.$$

2.11. *Example.* If $X = id_{\text{Vect}}$, then $\bar{X} \neq id_{\text{Vect}^{\mathbb{Z}}}$. In this case (2.2) yields

$$(2.3) \quad [Q\bar{X}^*]_k = \coprod_{n \geq 0} Q_{k-n}.$$

In the case that all Q_k are equal, say to I , (2.3) reduces to the familiar $I^{\mathbb{Z}}$ of § 1.

If $X = - \times X_0 : \text{Set} \rightarrow \text{Set}$, then

$$[Q\bar{X}^*]_k = \coprod_{n \geq 0} Q_{k-n} \times X_0^n,$$

so that a state of the free dynamics at time k records the q in Q_{k-n} in which the machine “started”, and the string w of X_0^n of inputs received since then.

With this, the theory of time-varying X -dynamics for an adjoint process $X : \mathcal{H} \rightarrow \mathcal{H}$ reduces to the theory of the adjoint process $\bar{X} : \mathcal{H}^{\mathbb{Z}} \rightarrow \mathcal{H}^{\mathbb{Z}}$. Thus for the theory of reachability, observability, realization and duality of such systems, the reader may turn to [3]. [Bear in mind that if $(\mathcal{E}, \mathcal{M})$ is an image factorization system for \mathcal{H} , then $(\bar{\mathcal{E}}, \bar{\mathcal{M}})$ is an image factorization system for $\mathcal{H}^{\mathbb{Z}}$, where $\bar{\mathcal{E}} = \{e | \text{each } e_k \text{ is in } \mathcal{E}\}$ and $\bar{\mathcal{M}} = \{m | \text{each } m_k \text{ is in } \mathcal{M}\}$.]

To make this a little more concrete, we give an example of a minimal realization for a time-varying system.

2.12. *Example.* Consider the time-varying linear system of Fig. 1. It may be described by the matrices

$$G_k = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad F_k = \begin{bmatrix} 0 & a \\ \bar{a} & 0 \end{bmatrix} \text{ for all } k,$$

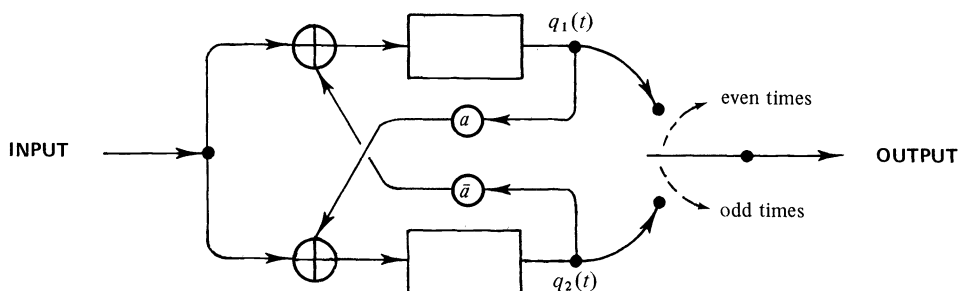


FIG. 1. A time-varying system—the output is $q_1(t)$ at even times and $q_2(t)$ at odd times

$$H_k = \begin{cases} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} & \text{for even } k, \\ \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} & \text{for odd } k. \end{cases}$$

The response at time k to a sequence $\omega = (\cdots, i_j, \cdots, i_1, i_0)$, with input i_j applied at $k-1-j$ is

$$\alpha_k(\omega) = \begin{cases} i_0 + ai_1 + \bar{a}ai_2 + a\bar{a}ai_3 + \cdots & \text{if } k \text{ is even,} \\ i_0 + \bar{a}i_1 + a\bar{a}i_2 + \bar{a}a\bar{a}i_3 + \cdots & \text{if } k \text{ is odd,} \end{cases}$$

and thus the total response at time k is clearly

$$f_k^\Delta : I^\S \rightarrow Y_\S : \omega \mapsto \begin{cases} (\alpha_k(\omega), \bar{a}\alpha_k(\omega), a\bar{a}\alpha_k(\omega), \bar{a}a\bar{a}\alpha_k(\omega), \cdots) & \text{if } k \text{ is even,} \\ (\alpha_k(\omega), a\alpha_k(\omega), \bar{a}a\alpha_k(\omega), a\bar{a}a\alpha_k(\omega), \cdots) & \text{if } k \text{ is odd.} \end{cases}$$

Now we have observed (and the theory of § 1 is a special case) that an $(\mathcal{E}, \mathcal{M})$ -factorization for f^Δ in $\mathcal{H}^\mathbb{Z}$ is simply a collection of $(\mathcal{E}, \mathcal{M})$ -factorizations—i.e., epi-mono factorizations—with one for each f_k^Δ . Now it is clear that for each k the image of f_k^Δ is a one-dimensional subspace of Y_\S , being spanned by $(1, \bar{a}, a\bar{a}, \bar{a}a\bar{a}, \cdots)$ for even k and by $(1, a, \bar{a}a, a\bar{a}a, \cdots)$ for odd k . Thus, we let $Q_k = \mathbf{R}$ for all k , interpreting q in Q_k as $(q, \bar{a}q, a\bar{a}q, \bar{a}a\bar{a}q, \cdots)$ in Y_\S for even k and as $(q, aq, \bar{a}aq, a\bar{a}aq, \cdots)$ in Y_\S for odd k . It is then straightforward to read off the values of \hat{G}_k, \hat{F}_k and \hat{H}_k for the minimal realization

$$\hat{G}_k = [1], \quad \hat{H}_k = [1] \quad \text{for all } k,$$

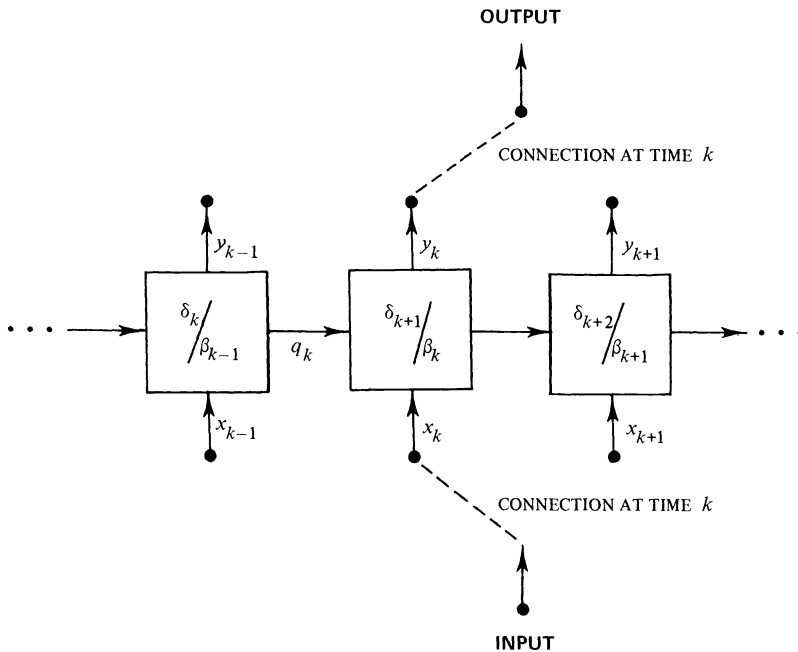


FIG. 2. General representation of a time-varying sequential machine as an infinite chain of time-invariant sequential machines, with time-varying read-in and read-out

while

$$\hat{F}_k = \begin{cases} [a] & \text{for even } k, \\ [\bar{a}] & \text{for odd } k. \end{cases}$$

Thus our general procedure has reduced the 2-dimensional system of Fig. 1 to a 1-dimensional system.

Incidentally, the system of Fig. 1 can be seen as a special case of a more general construction. We see this in Fig. 2 where we have represented a time-varying sequential machine as an infinite sequence of conventional sequential machines, only one of which is "activated" at time t . Note that if the $(\delta_{k+1}|\beta_k)$ dependency is periodic with period p then we can replace the infinite chain by a chain of p time-invariant machines. Figure 1 corresponds, for a linear system, to the case $p = 2$.

REFERENCES

- [1] M. A. ARBIB AND E. G. MANES, *Machines in a category: An expository introduction*, SIAM Rev., 16, (1974), pp. 163–192.
- [2] ———, *Foundations of system theory: Decomposable systems*, Automatica, 10 (1974), pp. 285–302.
- [3] ———, *Adjoint machines, state-behavior machines, and duality*, J. Pure Appl. Algebra, (1975), to appear.
- [4] E. S. BAINBRIDGE, *A unified minimal realization theory, with duality, for machines in a hyperdoctrine*, dissertation, Univ. of Michigan, Ann Arbor, 1972.
- [5] R. W. BROCKETT AND A. S. WILLSKY, *Finite state homomorphic sequential machines*, IEEE Trans. Automatic Control, 17 (1973), pp. 483–490.
- [6] W. DAUSCHA, G. NÜRNBERG, P. H. STARKE AND K. D. WINKLER, *Theorie der determinierten zeitvariablen Automaten*, Elektron. Informationsverarbeitung. Kybernetik, 9 (1973), pp. 455–511.
- [7] J. A. GOGUEN, *Minimal realizations of machines in closed categories*, Bull. Amer. Math. Soc., 78 (1972), pp. 777–783.
- [8] R. E. KALMAN, *Algebraic theory of linear systems*, Topics in Mathematical System Theory, R. E. Kalman, P. L. Falb and M. A. Arbib, McGraw-Hill, New York, 1969.
- [9] S. MAC LANE, *Categories for the Working Mathematician*, Springer-Verlag, Berlin, 1972.
- [10] L. WEISS, *Controllability, realization, and stability of discrete-time systems*, this Journal, 10 (1972), pp. 230–251.

ON CONTROLLABILITY BY MEANS OF TWO VECTOR FIELDS*

NORMAN LEVITT AND HÉCTOR J. SUSSMANN†

Abstract. A set S of vector fields on a differentiable manifold M is said to be *completely controllable* if for every pair (m, m') of points of M there exists a trajectory of S from m to m' . Here a trajectory of S is a curve which is an integral curve of some $X \in S$ or a finite concatenation of such curves so that, in general, a trajectory of S run in reverse is no longer a trajectory. Our main theorem is: on every connected paracompact manifold of class C^k , $2 \leq k \leq \infty$, or $k = \omega$, there exists a completely controllable set S consisting of two vector fields of class C^{k-1} .

1. Introduction. An autonomous control system on a smooth manifold M is, roughly, the same as a set of vector fields on M . By analogy with linear system theory, let us call a set S of vector fields *completely controllable* if, given any two points m_1 and m_2 of M , there exists a trajectory of S that steers m_1 to m_2 (the precise definition is given below in § 2). Since vector fields in general do not commute, it is clear that a finite set S of r vector fields on the n -dimensional manifold M can be completely controllable even if $r < n$. The purpose of this paper is to show that $r = 2$ is always sufficient to achieve controllability, as long as the obvious necessary condition that M be connected is satisfied. We shall prove the following.

THEOREM 1. *On every connected paracompact manifold M of class C^k , $2 \leq k \leq \infty$, or $k = \omega$, there exists a completely controllable pair $\{X, Y\}$ of vector fields of class C^{k-1} (when $k = \infty$ or $k = \omega$ it is understood that $k - 1 = k$).*

Results that are related to but weaker than our Theorem 1 were given by Lobry in [4] and by Sussmann in [10] and [11].

In [4] Lobry proved, using a different definition of controllability, that controllability is a generic property of pairs of C^k vector fields in the C^∞ manifold M , provided k is large enough (and $\leq \infty$). His proof is based on the well-known Lie algebra criteria for controllability (cf., for instance, [1], [2], [3] and [8]), and it requires that k be large because a large number of Lie brackets is involved.

Sussmann gives, in [10], a proof that the set of pairs (X, Y) of C^{k-1} vector fields that have the *accessibility property* is open and dense in the fine C^{k-1} topology of the set of all pairs of C^{k-1} vector fields on the C^k manifold M , if $2 \leq k \leq \infty$.

If M is connected, the accessibility property implies controllability in Lobry's sense, and the latter property implies controllability, in the sense of this paper, of the set consisting of the given vector fields together with their negatives. Hence, the results described above imply that controllability can always be achieved with *four* vector fields. It was shown in [11] how to bring this number down to three.

From now on, we shall only use the word "controllability" in the sense of the definition of § 2 below. The proof of Theorem 1 for $2 \leq k \leq \infty$ will be given in §§ 3 to 7, after the introduction, in § 2, of our basic definitions and notations. The proof for the real analytic case is given in § 8. We now give a brief sketch of the proof for $k \neq \omega$.

* Received by the editors May 17, 1974, and in revised form November 24, 1974.

† Department of Mathematics, Rutgers University, University Heights Campus, New Brunswick, New Jersey 08903. This research was supported in part by National Science Foundation Grants GP-33945 and GP-37488, respectively.

Our construction of a controllable pair $\{X, Y\}$ is based on the following idea: if f is a proper Morse function on M with a unique relative minimum, then the integral curves of the gradient of $-f$ (with respect to a suitable Riemannian metric) can be used to move from almost every point $m \in M$ to points arbitrarily close to the minimum. The set of $m \in M$ for which the integral trajectory of $-\text{grad } f$ does not approach the minimum consists of those points for which the trajectory approaches one of the other critical points. Therefore, the pair $\{\text{grad } f, -\text{grad } f\}$, though not controllable, is in some sense “approximately controllable”. We start with such a pair and modify it in two stages to achieve a controllable one. The first modification is described in § 5, and it enables us to obtain a pair $\{X^0, Y^0\}$ such that, for some sphere S in M , every point of M can be connected to a point of S by a trajectory. In § 6 we prove a lemma on controllability on spheres, which we use in § 7 to modify X^0 and Y^0 in an annulus, and prove Theorem 1.

We also indicate briefly (cf. remarks at end of § 5 and § 7) how to modify our argument so as to obtain the following stronger result.

THEOREM 2. *For every positive integer n there is a positive integer N such that, if $\infty \geq k \geq 2$, then on every connected paracompact n -dimensional C^k manifold M there is a pair $\{X, Y\}$ of C^{k-1} vector fields with the property that, if m_1, m_2 are in M , then there exists a trajectory of $\{X, Y\}$ from m_1 to m_2 which involves at most N switchings.*

2. Notations and basic definitions. Throughout this paper, we use M to denote a connected paracompact differentiable manifold of class C^k (where $2 \leq k \leq \omega$). If X is a C^{k-1} vector field on M , we use $\{X_t\}$ to denote the one-parameter family of local diffeomorphisms generated by X . If S is a set of vector fields on M , a *trajectory* of S is a continuous curve σ on M which is a finite concatenation of curves σ^i that are integral curves of vector fields $X^i \in S$. A point which is of the form $\gamma(t)$ for some $t \geq 0$ and some trajectory γ of S for which $\gamma(0) = m$, is said to be *S-reachable* from m . The set of all points that are *S-reachable* from m is called the *positive S-orbit* of m , and is denoted by $O^+(S, m)$. A system S is said to be (completely) *controllable* if $O^+(S, m) = M$ for every $m \in M$. We observe that the relation of *S-reachability* is reflexive and transitive, but (in general) not symmetric, for a trajectory “run in reverse” is not necessarily a trajectory.

3. Morse functions. We now recall some basic facts concerning Morse functions. If $f: M \rightarrow \mathbb{R}$ is a C^k function, then a point $m \in M$ is called a *critical point* of f if the differential of f vanishes at m . If $m \in M$ and if $\{x_1, \dots, x_n\}$ is a chart about m , we can define the *Hessian matrix* $H = (\partial^2 f / (\partial x_j \partial x_i)(m))_{1 \leq i, j \leq n}$. If m is a critical point of f , then the property that H is nonsingular is independent of the choice of coordinates. When this property holds, m is called a *nondegenerate critical point*.

If m_0 is a nondegenerate critical point of f , then there exists a coordinate chart (x_1, \dots, x_n) defined on a neighborhood U of m and an integer λ ($0 \leq \lambda \leq n$) such that $x_1(m_0) = \dots = x_n(m_0) = 0$ and that

$$f(m) = f(m_0) - x_1^2 - \dots - x_\lambda^2 + x_{\lambda+1}^2 + \dots + x_n^2$$

for $m \in U$ (Morse’s lemma, cf. [6, Lemma 2.2]). This reference considers the case

$k = \infty$, but the proof is actually valid for $k \geq 3$; cf. also the remark in § 4). The integer λ is called the *index* of the critical point m_0 .

A C^k function $f: M \rightarrow \mathbb{R}$ all of whose critical points are nondegenerate is called a *Morse function*. It is well known that on every paracompact manifold there exists a Morse function. We shall use the fact that on every connected paracompact M there is a Morse function f such that

- (i) for every real a , the set $M_a = \{m \in M, f(m) \leq a\}$ is compact;
- (ii) f has only one critical point of index zero, and
- (iii) if c and c' are critical points with $c \neq c'$, then $f(c) \neq f(c')$.

We give a brief sketch of the proof that such an f exists. It is proved in [6, Cor. 6.7] that there is a Morse function g for which (i) holds. From this it follows easily that M is the union of an increasing sequence M_0, M_1, \dots of compact connected submanifolds with smooth boundary, of which M_0 can be taken to be a ball. (Indeed, choose M_0 to be an arbitrary ball in M . For a real a , let $N_a = \{m \in M: g(m) \leq a\}$, and let $\{a_j\}_{j=1,2,\dots}$ be an unbounded increasing sequence of reals that are not critical levels of g and such that $M_0 \subseteq N_{a_1}$. Then choose $M_j =$ connected component of N_{a_j} that contains M_0 , for $j = 1, 2, \dots$) Use V_j to denote the closure of $M_{j+1} - M_j$. Then the triad $(V_j, \partial M_j, \partial M_{j+1})$ satisfies the condition $H_0(V_j, \partial M_j) = 0$. It follows from Theorem 8.1 of [5] that there is a Morse function

$$f_j: (V_j; \partial M_j, \partial M_{j+1}) \rightarrow ([j, j+1]; \{j\}, \{j+1\})$$

with no critical points of index zero. It is easy to see that all these functions, and the obvious Morse function

$$f_0: (M_0, \partial M_0) \rightarrow ([-1, 0], \{0\})$$

can be pieced together (after modifying them, if necessary, on neighborhoods of the ∂M_j). The resulting Morse function f satisfies (i) and (ii), and it is a trivial matter to modify it so that (iii) will also hold.

4. The first step. We let f be a Morse function on M for which conditions (i), (ii) and (iii) of § 3 hold, and we let c_0, c_1, \dots denote the sequence of critical points of f , arranged so that

$$f(c_0) < f(c_1) < f(c_2) < \dots$$

It is clear that we can add an arbitrary constant to f , so we shall assume from now on that

$$f(c_0) = 0.$$

Let λ_j denote the index of c_j . Clearly $\lambda_0 = 0$, $\lambda_j > 0$ for $j > 0$. Choose, for each j , a chart $\{x_1^j, \dots, x_n^j\}$ defined on a neighborhood U_j of c_j , such that $x_1^j(c_j) = \dots = x_n^j(c_j) = 0$ and that

$$f(m) = f(c_j) - (x_1^j)^2 - \dots - (x_{\lambda_j}^j)^2 + (x_{\lambda_j+1}^j)^2 + \dots + (x_n^j)^2$$

for $m \in U_j$.

Remark. The choice of coordinates x_1^j, \dots, x_n^j with the above properties is possible because of Morse's lemma. An anonymous referee has brought to our attention the fact that the proof of this lemma that appears in [6, Lemma 2.2] requires f to be C^3 . On the other hand, here we are only assuming that M is a C^k

manifold with $k \geq 2$. For the benefit of those readers who might be concerned about this difficulty, we observe that any C^k manifold, $k \geq 1$, has a unique C^∞ structure which is compatible with the given C^k structure (cf., for instance, [17]). Therefore, it makes sense to talk of a C^∞ Morse function on M even if M is C^2 . We could have chosen f to be C^∞ , and then Morse's lemma undoubtedly applies.

We assume that the U_j are chosen so that U_j is mapped diffeomorphically by (x_1^j, \dots, x_n^j) onto the ball with radius α_j . Moreover, we assume that

$$\alpha_j + \alpha_{j+1} < f(c_{j+1}) - f(c_j)$$

so that, in particular,

$$f(m) < f(m') \quad \text{for } m \in U_j, \quad m' \in U_{j+1}.$$

Choose, in an arbitrary fashion, numbers β_j such that $0 < \beta_j < \alpha_j$, and let V_j denote the set of points in U_j for which

$$(x_1^j)^2 + \dots + (x_n^j)^2 < \beta_j^2.$$

Now choose a Riemannian metric on M , in such a way that

$$\left\langle \frac{\partial}{\partial x_k^j}, \frac{\partial}{\partial x_l^j} \right\rangle \equiv \delta_{kl} \quad \text{on } V_j.$$

Let

$$X^0 = -\text{grad } f,$$

where the gradient is taken with respect to the chosen metric.

It is clear from condition (i) of § 3 that $X_t^0(m)$ is defined for all $m \in M$ and all times $t \geq 0$. Moreover, the limit

$$L(m) = \lim_{t \rightarrow +\infty} X_t^0(m)$$

exists for every $m \in M$ and is a critical point of f . We let, for $j = 0, 1, \dots$,

$$C_j = \{m \in M, L(m) = c_j\}.$$

Clearly, $C_j \cap V_j$ is the set of all points of V_j for which

$$x_1^j = \dots = x_{\lambda_j}^j = 0.$$

In particular, if $j > 0$, then $C_j \cap V_j$ is a submanifold of M of codimension ≥ 1 . If we let

$$B = \bigcup_{j>0} C_j,$$

then B (the "bad set") has measure zero in M .

5. Definition of Y^0 . For $j = 0, 1, 2, \dots$, let $\Delta_j: V_j \rightarrow V_j$ be a diffeomorphism which coincides with the identity in the complement of a compact subset of V_j . Moreover, if $j > 0$, we choose Δ_j so that the following conditions hold:

- (A) the vector $X^0(\Delta_j^{-1}(c_j))$ is not tangent to the manifold $\Delta_j^{-1}(C_j \cap V_j)$, and
- (B) the vector $X^0(\Delta_j(c_j))$ is not tangent to $\Delta_j(C_j \cap V_j)$.

For $j = 0$, we only require that $\Delta_0(c_0) \neq c_0$.

To prove that such Δ_j exist, let us fix j . Let $\{x_1^j, \dots, x_n^j\}$ be the coordinate chart on V_j that was introduced above, and let a vector field Z be defined on V_j by

$$Z = \phi \hat{Z}, \quad \hat{Z} = \sum_{i=1}^n a_i \frac{\partial}{\partial x_i^j}.$$

Here ϕ is a smooth function in V_j , with a compact support, and equal to one in a neighborhood Ω of c_j . The real constants a_1, \dots, a_n are arbitrary, subject only to the condition $a_1 \neq 0$.

We choose $\varepsilon > 0$ and a neighborhood Ω' of c_j such that $Z_t(m) \in \Omega$ for $m \in \Omega'$, $|t| \leq \varepsilon$. If $m \in \Omega'$, $|t| \leq \varepsilon$, and if the coordinates of m are (x_1^j, \dots, x_n^j) , it follows that the coordinates of $Z_t(m)$ are $(x_1^j + ta_1, \dots, x_n^j + ta_n)$. In particular, the points $Z_{\pm\varepsilon}(c_j)$ have coordinates $(\pm\varepsilon)(a_1, \dots, a_n)$. Moreover, the manifold $Z_{\pm\varepsilon}(C_j \cap V_j)$ is given, in a neighborhood of $Z_{\pm\varepsilon}(c_j)$, by the equations

$$x_1^j = \pm \varepsilon a_1, \dots, x_{\lambda_j}^j = \pm \varepsilon a_{\lambda_j},$$

so that the vector $\sum_{i=1}^n v_i (\partial/\partial x_i)$ at $Z_{\pm\varepsilon}(c_j)$ is tangent to $Z_{\pm\varepsilon}(C_j \cap V_j)$ if and only if $v_1 = \dots = v_{\lambda_j} = 0$. But

$$X^0(Z_{\pm\varepsilon}(c_j)) = (\pm\varepsilon) \sum_{i=1}^n v_i \frac{\partial}{\partial x_i},$$

where $v_i = 2a_i$ for $1 \leq i \leq \lambda_j$, and $v_i = -2a_i$ for $\lambda_j < i \leq n$. Since $a_1 \neq 0$, we conclude that $\Delta_j = Z_\varepsilon$ satisfies the desired conditions.

We let $\Delta: M \rightarrow M$ be the diffeomorphism defined by $\Delta(x) = \Delta_j(x)$ for $x \in V_j$, $\Delta(x) = x$ for $x \notin \bigcup_{j=1}^\infty V_j$. We let $Y^0 = -d\Delta(X^0)$, so that the integral curves of Y^0 are precisely the images under Δ of the integral curves of X^0 , run in reverse.

The pair $\{X^0, Y^0\}$ is not yet controllable, but we show that it is reasonably close to it.

For each real λ such that $0 < \lambda \leq \alpha_0$, let us use $B(\lambda)$ to denote the ball

$$B(\lambda) = \{x: x \in U_0, f(x) \leq \lambda^2\},$$

and use $S(\lambda)$ to denote the boundary of $B(\lambda)$. The complement of $B(\lambda)$ will be denoted by $W(\lambda)$, and a fixed γ_0 such that $0 < \gamma_0 < \beta_0$ is chosen once and for all, in such a way that the diffeomorphism Δ_0 coincides with the identity in the complement of a compact subset of $B(\gamma_0)$.

LEMMA 1. *Let $\gamma_0 < \lambda < \beta_0$. Then for every $m \in M$ there exist $p \in S(\lambda)$, $q \in S(\lambda)$ such that p is reachable from m , and m is reachable from q , by trajectories of $\{X^0, Y^0\}$.*

Proof. We first show that

(a) for every $m \in W(\lambda)$ there is a trajectory of $\{X^0, Y^0\}$ from m to some $p \in S(\lambda)$. This is clear if $m \in C_0$, for then the integral curve of X^0 through m must intersect $S(\lambda)$. We assume that the conclusion is true for $m \in \bigcup_{j < k} C_j$, and prove that it is also true for $m \in C_k$. Since $X^0(\Delta^{-1}(c_k))$ is not tangent to $\Delta^{-1}(C_k \cap V_k)$ at $\Delta^{-1}(c_k)$, it follows that $Y^0(c_k) = -d\Delta(X^0(\Delta^{-1}(c_k)))$ is not tangent to $C_k \cap V_k$ at c_k . Therefore, there is a neighborhood Ω of c_k such that $Y^0(x)$ is not tangent to $C_k \cap V_k$ if $x \in C_k \cap V_k \cap \Omega$. For sufficiently large $t > 0$, the point $X_t^0(m) = x$ is in $C_k \cap V_k \cap \Omega$. Therefore, for some $\tau > 0$, $m' = Y_\tau^0(x)$ belongs to V_k but not to C_k . Since $f(y) < f(c_{k+1})$ for $y \in V_k$, it follows that $m' \in \bigcup_{i < k} C_j$. By the inductive

assumption, there is a trajectory of $\{X^0, Y^0\}$ that takes m' to $S(\lambda)$. Since $m' = Y_t^0 X_t^0(m)$, the conclusion follows, and the proof of (a) is complete.

We now show that

(b) statement (a) remains true if $\{X^0, Y^0\}$ is replaced by $\{-Y^0, -X^0\}$.

To prove this, simply repeat the proof of (a), replacing throughout X^0 by $-Y^0$, Y^0 by $-X^0$, Δ by Δ^{-1} , c_j by $c'_j = \Delta(c_j)$ and C_j by $C'_j = \Delta(C_j)$. Condition (B) of the definition of Y^0 is used in the same way as condition (A) was used in the proof of (a).

Since the trajectories of $\{-Y^0, -X^0\}$ run in reverse are trajectories of $\{X^0, Y^0\}$, the conclusion of our lemma has been proved for $m \in W(\lambda)$. Now consider $m \in B(\lambda)$. If $m \neq \Delta(c_0)$, then the integral curve of Y^0 through m goes through $S(\lambda)$ for some positive time t . If $m = \Delta(c_0)$, then $X_t^0(m) \neq \Delta(c_0)$ for $t > 0$, and therefore there is a trajectory of $\{X^0, Y^0\}$ from m to $S(\lambda)$. A similar argument shows that every $m \in B(\lambda)$ is in a trajectory emanating from some $q \in S(\lambda)$. This completes the proof of Lemma 1.

Remark. A slightly more careful argument would have enabled us to choose Δ in such a way that each c_k has a neighborhood $\Omega_k \subseteq V_k$ such that, for $x \in \Omega_k \cap C_k$, it follows that $Y_t^0(x) \notin B$ for some $t > 0$ (and a similar conclusion with X^0, Y^0, B, c_k, C_k replaced by $-Y^0, -X^0, B', c'_k = \Delta(c_k), C'_k = \Delta(C_k)$). With this choice of Δ , the conclusion of Lemma 1 can be improved by adding the statement that *the trajectories from m to p and from q to m involve at most two switchings.*

6. Controllability on spheres. We shall now prove our theorem for the special case where M is a sphere. We shall need two vector fields that satisfy a condition slightly stronger than controllability. The following lemma states this condition precisely.

LEMMA 2. *Let n be a positive integer and let $T > 0$ be a real number. Then there exist two vector fields X and Y on the n -dimensional sphere S^n , and a positive integer l , with the property that for every pair m, m' of points in S^n there are nonnegative reals $t_1, \dots, t_l, \tau_1, \dots, \tau_l$ such that*

$$m' = X_{t_1} Y_{\tau_1} X_{t_2} Y_{\tau_2} \cdots X_{t_l} Y_{\tau_l}(m)$$

and that

$$(1) \quad t_1 + \cdots + t_l = \tau_1 + \cdots + \tau_l < T.$$

Lemma 2 will follow from Lemma 3.

LEMMA 3. *Let $SO(\mu)$ denote the group of $\mu \times \mu$ orthogonal matrices whose determinant is 1. Then for every $T > 0$ there exist two skew-symmetric matrices A and B such that every $P \in SO(\mu)$ can be expressed as a product*

$$(2) \quad \prod_{i=1}^l e^{t_i A} e^{\tau_i B}$$

with the t_i, τ_i nonnegative and such that (1) holds.

To prove Lemma 3, we use the following.

LEMMA 4. *Let G be a connected Lie group and let \mathcal{G} be its Lie algebra. Let X^1, \dots, X^k be elements of \mathcal{G} which generate \mathcal{G} as a Lie algebra. Then every $g \in G$ is a finite product of exponentials $\exp(X^i t)$, $i = 1, \dots, k$, $-\infty < t < \infty$.*

Proof. Cf. [7, Chap. IV, Thm. 1] or [1, Lemma 6.2].

Proof of Lemma 3. Our conclusion follows rather directly from the results and methods of [1], but we shall give a self-contained proof. We shall use Lemma 4, applied to the Lie group $G = SO(\mu) \times \mathbb{R}$. The Lie algebra of G is the product $so(\mu) \times \mathbb{R}$, where $so(\mu)$ denotes the Lie algebra of all $\mu \times \mu$ skew-symmetric matrices. We let A, B be elements of $so(\mu)$ which generate $so(\mu)$ as a Lie algebra (for instance, let $A = (a_{ij}), B = (b_{ij})$, where $a_{12} = -a_{21} = 1, b_{12} = b_{23} = \cdots = b_{\mu-1, \mu} = -b_{21} = \cdots = -b_{\mu, \mu-1} = 1$ and where all the other a_{ij}, b_{ij} vanish). We let C, D denote the elements of $so(\mu) \times \mathbb{R}$ defined by

$$C = (A, 1) \quad \text{and} \quad D = (B, -1).$$

It is easy to see that C and D generate the Lie algebra $so(\mu) \times \mathbb{R}$ (proof: let Λ denote the Lie subalgebra of $so(\mu) \times \mathbb{R}$ generated by C and D , and let $L = \{E : E \in so(\mu), (E, 0) \in \Lambda\}$. It is clear that $A + B \in L$. Moreover, if $E \in L$, then

$$[C, (E, 0)] = ([A, E], 0)$$

so that $[A, E] \in L$. Similarly, $[B, E] \in L$. Since A and B generate $so(\mu)$, it follows that L is an ideal of $so(\mu)$. Since $so(\mu)$ is simple, we conclude that $L = so(\mu)$. Therefore $so(\mu) \times \{0\} \subseteq \Lambda$. In particular, $(A, 0) \in \Lambda$. Since $(A, 1) \in \Lambda$, it follows that $(0, 1) \in \Lambda$, so that $\{0\} \times \mathbb{R} \subseteq \Lambda$. Therefore $\Lambda = so(\mu) \times \mathbb{R}$, as we wanted to prove).

By Lemma 4, every element of $SO(\mu) \times \mathbb{R}$ can be written as a product

$$\prod_{i=1}^l \exp(t_i C) \exp(\tau_i D),$$

where the t_i, τ_i are real numbers. Clearly, the preceding product is equal to

$$\left(\prod_{i=1}^l e^{t_i A} e^{\tau_i B}, t_1 + \cdots + t_l - \tau_1 - \cdots - \tau_l \right).$$

If we apply this result to the elements of $SO(\mu) \times \mathbb{R}$ that are of the form $(P, 0)$ we conclude that every element of $SO(\mu)$ is a product (2) with $t_1 + \cdots + t_l = \tau_1 + \cdots + \tau_l$.

We must now improve our conclusion to show that the t_i, τ_i can be taken to be nonnegative, and that their sums can be taken to be bounded by a fixed $T > 0$. We let $\Phi_l: \mathbb{R}^{2l} \rightarrow SO(\mu)$ denote the map which to each element $(t_1, \cdots, t_l, \tau_1, \cdots, \tau_l)$ of \mathbb{R}^{2l} assigns the product (2). We let Σ^l denote the subspace of \mathbb{R}^{2l} whose elements satisfy $t_1 + \cdots + t_l = \tau_1 + \cdots + \tau_l$. We have just shown that the union of the $\Phi_l(\Sigma^l)$ (taken over all l) is all of $SO(\mu)$. Since each $\Phi_l(\Sigma^l)$ is a countable union of compact sets, it follows that some $\Phi_l(\Sigma^l)$ has a nonempty interior in $SO(\mu)$.

Choose one value of l for which this holds, so that l is now fixed. By Sard's theorem (cf. [12]) the differential $d\Phi_l(p)$ is surjective at some point p of Σ^l . Since Φ_l is real-analytic, it follows that $d\Phi_l(p)$ is surjective for all points p in an open dense subset of Σ^l . In particular, we can choose $\eta > 0$ and let $\Sigma_+^l(\eta)$ denote the set of all points in Σ^l for which the t_i and the τ_i are positive, and for which $t_1 + \cdots + t_l < \eta$. It follows that $\Phi_l(\Sigma_+^l(\eta))$ contains a nonempty open subset U of $SO(\mu)$. Since $SO(\mu)$

is a compact connected topological group, there is an integer $v > 0$ such that every element of $SO(\mu)$ is a product of v elements of U .

(This can be proved as follows. Let K be a compact connected group and let $U \subseteq K$ be open and nonempty. Let V be the union of the sets U^m , where m ranges over all integers ≥ 1 , and where U^m is the set of all products of m elements of U . Then V is a semigroup, i.e., the product of any two elements of V is again in V . Moreover, V is open. The closure \bar{V} is also a semigroup, and we show that it is in fact a group. Indeed, if $v \in \bar{V}$, then the sequence of powers v^n , $n = 1, 2, \dots$, has a convergent subsequence $\{v^{n(j)}\}$. By taking a subsequence, if necessary, we can assume that $n(j+1) \geq n(j) + 2$. Set $v_j = v^{n(j+1)-n(j)-1}$. Then $v_j \in \bar{V}$. Clearly $\{v^{n(j)}\}$ and $\{v^{n(j+1)}\}$ have the same limit as $j \rightarrow \infty$, so that v_j converges to v^{-1} . Therefore v^{-1} is in \bar{V} . Thus \bar{V} is indeed a subgroup of K . Since \bar{V} has a nonempty interior, the connectedness of K implies that $\bar{V} = K$. This proves that V is dense in K . In particular, V intersects the open set V^{-1} , so that there is a $v \in V$ such that $v^{-1} \in V$. But then the identity, which is equal to $v \cdot v^{-1}$, must belong to V . Therefore, there is a positive integer m such that the open set U^m is a neighborhood of the identity. Since K is compact and connected, $(U^m)^{m'} = K$ for m' sufficiently large. To conclude the proof, take $v = mm'$.)

We have therefore established that every $P \in SO(\mu)$ is a product of the form (2) (with lv instead of l), where the t_i and τ_i are nonnegative, and where condition (1) is satisfied, provided we take $T = v\eta$.

The proof of our lemma is now complete, except for the fact that we have shown that our conclusion holds for *some* $T > 0$, rather than for *every* $T > 0$. But this gap is easily removed. Indeed, if $A, B, T > 0$ are such that the conclusion holds, and if $\lambda > 0$ is arbitrary, it is clear that the conclusion also holds if A, B, T are replaced by $\lambda A, \lambda B, \lambda^{-1}T$, respectively. Q.E.D.

Proof that Lemma 3 implies Lemma 2. The group $SO(n+1)$ acts transitively on the sphere S^n . Regard the elements of S^n as column vectors in \mathbb{R}^{n+1} . The vector fields defined by

$$X(x) = Ax, \quad Y(x) = Bx$$

are tangent to S^n , because A and B are skew-symmetric (so that $\langle Ax, x \rangle = \langle Bx, x \rangle = 0$). Moreover, $X_t(x) = e^{At}x$ and $Y_t(x) = e^{Bt}x$ for $x \in S^n$, $t \in \mathbb{R}$. Therefore, X and Y satisfy the desired conclusion. Q.E.D.

7. Proof of the main theorems for $k \neq \omega$. We are now ready to prove our theorems for an arbitrary manifold M . We let N denote the annulus

$$\{m : m \in U_0, \gamma_0^2 < (x_1^0)^2 + \dots + (x_n^0)^2 < \beta_0^2\}.$$

Regard N as the product of the sphere S^{n-1} with the interval $I = (\frac{1}{2} \log \gamma_0, \frac{1}{2} \log \beta_0)$, by identifying the point $p = (x_1, \dots, x_n; \zeta)$ of $S^{n-1} \times I$ with the point of N whose coordinates are given by $x_i^0 = e^{2\zeta} x_i$. With this identification, the trajectories of X^0, Y^0 are the curves $t \rightarrow (x_1^0, \dots, x_n^0; \zeta^0 \mp t)$ (where the sign is “−” for X^0 , “+” for Y^0).

We let a denote the midpoint of I , and choose $\varepsilon > 0$ so small that $a - 2\varepsilon$ and $a + 2\varepsilon$ belong to I . We let ϕ be a smooth function that vanishes in the complement of $(a - 2\varepsilon, a + 2\varepsilon)$, and which is equal to one on $(a - \varepsilon, a + \varepsilon)$. We let \bar{X}, \bar{Y} denote

vector fields on S^{n-1} for which the conclusion of Lemma 2 holds, with $T = \varepsilon$. We identify \bar{X}, \bar{Y} with vector fields on $S^{n-1} \times I$, and hence on N , in an obvious way. For $p = (x_1, \dots, x_n; \zeta) \in S^{n-1} \times I$, let

$$X(p) = X^0(p) + \phi(\zeta)\bar{X}(p),$$

$$Y(p) = Y^0(p) + \phi(\zeta)\bar{Y}(p).$$

Then X and Y are smooth vector fields on N which coincide with X^0, Y^0 in the complement of $S^{n-1} \times [a - 2\varepsilon, a + 2\varepsilon]$. We can therefore extend them to M by letting

$$X(m) = X^0(m), \quad Y(m) = Y^0(m) \quad \text{for } m \notin N.$$

This completes the construction of X and Y . We now show that they have the desired properties.

LEMMA 5. *Let $S = S^{n-1} \times \{a\}$. Let $m \in M$ be arbitrary. Then there are trajectories Γ_1, Γ_2 of $\{X, Y\}$ such that (i) Γ_1 starts at m and ends at a point of S and (ii) Γ_2 starts at a point of S and ends at m .*

Proof. Along the trajectories of X, Y in N the coordinate ζ satisfies the equations $\dot{\zeta} = -1, \dot{\zeta} = 1$ respectively. The conclusion of our lemma is then trivial if $m \in N$.

If $m \notin N$, Lemma 1 gives us trajectories $\Gamma'_1 = [0, r] \rightarrow M, \Gamma'_2 = [0, s] \rightarrow M$ of $\{X^0, Y^0\}$ such that $\Gamma'_1(0) = \Gamma'_2(s) = \dot{m}$ and that $\Gamma'_1(r)$ and $\Gamma'_2(0)$ belong to S . We can clearly choose r_0, s_0 such that $\Gamma'_1(t)$ ($0 \leq t \leq r_0$), $\Gamma'_2(t)$ ($s_0 \leq t \leq s$) are not in $S^{n-1} \times [a - 2\varepsilon, a + 2\varepsilon]$ and that $m' = \Gamma'_1(r_0), m'' = \Gamma'_2(s_0)$ are in N . Since the restrictions of Γ'_1 to $[0, r_0]$ and of Γ'_2 to $[s_0, s]$ are trajectories of $\{X, Y\}$, the desired conclusion follows from the first part of our proof.

LEMMA 6. *Let S be as in Lemma 5. Let m_1, m_2 be points of S . Then there is a trajectory of $\{X, Y\}$ which starts at m_1 and ends at m_2 .*

Proof. Let $m_i = (\bar{m}_i, a), i = 1, 2, \bar{m}_i \in S^{n-1}$. The vector fields \bar{X}, \bar{Y} were chosen so that there is a trajectory $\bar{\Gamma} = [0, r] \rightarrow S^{n-1}$ of $\{\bar{X}, \bar{Y}\}$ for which $\bar{\Gamma}(0) = \bar{m}_1, \bar{\Gamma}(r) = \bar{m}_2$. Moreover, if we let $\xi(t), \eta(t)$ denote the sum of the lengths of the sub-intervals of $[0, t]$ in which $\bar{\Gamma}$ is an integral curve of \bar{X}, \bar{Y} , respectively, we can assume that $\xi(r) = \eta(r) < \varepsilon$, because of condition (1) of Lemma 2.

Define Γ by

$$\Gamma(t) = (\bar{\Gamma}(t), \zeta(t)), \quad 0 \leq t \leq r, \quad \text{where } \zeta(t) = a + \eta(t) - \xi(t).$$

It is clear that $\Gamma(0) = m_1, \Gamma(r) = m_2$. Moreover, $\dot{\zeta}(t) = 1$ or $\dot{\zeta}(t) = -1$ depending on whether t is in an interval in which $\bar{\Gamma}$ is an integral curve of \bar{Y} or of \bar{X} . Finally, $a - \varepsilon < \zeta(t) < a + \varepsilon$ for all t . Since the function ϕ is identically equal to one on $[a - \varepsilon, a + \varepsilon]$, it follows that Γ is a trajectory of $\{X, Y\}$.

Lemmas 5 and 6 clearly imply that the pair $\{X, Y\}$ is controllable. The proof of Theorem 1 for $k \neq \omega$ is therefore complete.

We now indicate how to prove Theorem 2. This will follow from

(a) the remark at the end of § 5, and

(b) the fact that the integer l of Lemma 2 is independent of m and m' .

If the diffeomorphism Δ is chosen as indicated in the remark of § 5, then arbitrary points m and m' can be joined by a trajectory which

- (i) goes from m to a point p of S . To achieve this, two switchings may be required, to avoid being absorbed by a critical point.
- (ii) goes from p to some other point q in S . This involves at most $2l - 1$ switchings.
- (iii) goes from q to m' , with no more than two switchings.

Thus the trajectory involves at most $2 + (2l - 1) + 2 + 2$ switchings (the last term is due to the two switchings that are needed to change from (i) to (ii) and from (ii) to (iii)). Thus we can take $N = 5 + 2l$, where l is the integer of Lemma 2, which depends only on n . This proves Theorem 2.

8. The real analytic case. We now assume that M is a real analytic manifold. We have already proved that there exists a pair $\{X, Y\}$ of C^∞ vector fields which is completely controllable, and we want to show that X and Y can actually be taken to be real analytic.

We shall use the following fact, due to Morrey [13] for compact M and to Grauert [14] in the general case.

THE MORREY–GRAUERT IMBEDDING THEOREM. *Every separable real analytic manifold M can be imbedded in a Euclidean space \mathbb{R}^l by a regular and proper C^ω mapping.*

Assume that M is so imbedded. Then the tangent bundle $T(M)$ of M is naturally identified with a subbundle of the trivial bundle $M \times \mathbb{R}^l$. This bundle has an obvious Riemannian metric, and there is an associated projection map π from the space $\Gamma(M \times \mathbb{R}^l)$ of C^∞ sections of $M \times \mathbb{R}^l$ into the space $\Gamma(T(M))$ of C^∞ vector fields on M . If both spaces are given the fine C^∞ topology, it is clear that π is continuous. Moreover, if $F \in \Gamma(M \times \mathbb{R}^l)$ is real analytic, so is πF . Now, it follows from the Whitney approximation theorem (cf. [15], [16] or [17]) that every real-valued C^∞ function on M can be approximated arbitrarily close in the fine C^∞ topology by real analytic functions. The same is therefore true for sections of the trivial bundle $M \times \mathbb{R}^l$. If X is a C^∞ vector field on M , then X , considered as a section of $M \times \mathbb{R}^l$, is the limit of a net F_α of real analytic sections. Putting $X_\alpha = \pi F_\alpha$, we get a net of real analytic vector fields that converges to X in the fine C^∞ topology. The validity of Theorem 1 for $k = \omega$ is now an immediate consequence of the following result, proved by Sussmann in [10]: the set of all completely controllable pairs (X, Y) of C^∞ vector fields on a C^∞ manifold M is open in $\Gamma(T(M)) \times \Gamma(T(M))$, if $\Gamma(T(M))$ is given the fine C^∞ topology.

Acknowledgment. An anonymous referee contributed interesting comments, for which the authors are grateful.

REFERENCES

- [1] V. JURDJEVIC AND H. SUSSMANN, *Control systems on Lie groups*, J. Differential Equations, 12 (1972), pp. 313–329.
- [2] A. KRENER, *A generalization of Chow's theorem and the bang-bang theorem to nonlinear control systems*, this Journal, 12 (1974), pp. 43–52.
- [3] C. LOBRY, *Contrôlabilité des systèmes non linéaires*, this Journal, 8 (1970), pp. 573–605.
- [4] ———, *Une propriété générique des couples de champs de vecteurs*, Czechoslovak Math. J., 22 (97) (1972).
- [5] J. MILNOR, *Lectures on the h-Cobordism Theorem*, Princeton University Press, Princeton, N.J., 1965.

- [6] ———, *Morse Theory*, Princeton University Press, Princeton, N.J., 1963.
- [7] R. PALAIS, *A global formulation of the Lie theory of transformation groups*, Amer. Math. Soc. Memoirs No. 22, 1957.
- [8] H. SUSSMANN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95–116.
- [9] H. SUSSMANN, *Orbits of families of vector fields and integrability of distributions*, Trans. Amer. Math. Soc., 180 (1973), pp. 171–183.
- [10] ———, *Some properties of vector fields that are not altered by small perturbations*, J. Differential Equations, to appear.
- [11] ———, *On the number of directions needed to achieve controllability*, this Journal, 13 (1975), pp. 414–419.
- [12] S. STERNBERG, *Lectures on Differential Geometry*, Prentice-Hall, Englewood Cliffs, N.J., 1964.
- [13] C. B. MORREY, *The analytic embedding of abstract real analytic manifolds*, Ann. of Math., 68 (1958), pp. 159–201.
- [14] H. GRAUERT, *On Levi's problem and the imbedding of real analytic manifolds*, Ibid., 68 (1958), pp. 460–472.
- [15] H. WHITNEY, *Differentiable manifolds*, Ibid., 37 (1936), pp. 645–680.
- [16] H. CARTAN, *Variétés analytiques réelles et variétés analytiques complexes*, Bull. Soc. Math. France, 85 (1957), pp. 77–99.
- [17] K. SHIGA, *Some aspects of real analytic manifolds and differentiable manifolds*, J. Math. Soc. Japan, 16 (1964), pp. 128–142.