SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 1, January 1981

0363-0129/81/1901-0001 $01.00/0

$(A, \mathscr{B})$-INVARIANT DISTRIBUTIONS AND DISTURBANCE DECOUPLING OF NONLINEAR SYSTEMS*

R. M. HIRSCHORN†

Abstract. The concept of (A, B)-invariant subspaces has resulted in a unified approach to many of the basic structural properties of time-invariant linear systems (W. M. Wonham, Lecture Notes in Economics and Mathematical Systems, vol. 101, Springer-Verlag, New York, 1974). The purpose of this paper is to introduce the more general notion of $(A, \mathscr{B})$-invariant distributions on differentiable manifolds and to use this idea to study the disturbance decoupling problem for a class of nonlinear systems which evolve on real analytic manifolds.

1. Introduction.

We consider systems of the form

$$\begin{aligned} \dot{x}(t) &= A(x(t)) + \sum_{i=1}^{m} u_i(t) B_i(x(t)), \qquad x \in M, \\ y(t) &= C(x(t)), \end{aligned} \tag{1}$$

with state space M a connected real analytic manifold and $A, B_1, \cdots, B_m \in V(M)$, the real vector space of real analytic vector fields on M. The input functions $u_1, \cdots, u_m$ belong to U, the class of admissible controls. The elements of U are piecewise real analytic functions on $[0, \infty)$ with values in a path-connected subset Ω of R^m which contains a neighborhood of the origin. We assume that U contains all piecewise constant functions with values in Ω (this assumption is needed to take advantage of the standard accessibility results; cf. [3]). The output map C is a real analytic mapping of M into R^l.

If $u = (u_1, \cdots, u_m)$ is an admissible control vector, then we denote by $x(t, x_0, u)$ the trajectory corresponding to u and the initial condition $x(0) = x_0$. Similarly $y(t, x_0, u)$ denotes the output $C(x(t, x_0, u))$.

The disturbance decoupling problem for (1) arises when a disturbance represented by $v_1, \cdots, v_r \in U$ affects the evolution of the state:

$$\begin{aligned} \dot{x} &= A(x) + \sum_{i=1}^{m} u_i B_i(x) + \sum_{i=1}^{r} v_i D_i(x), \\ y &= C(x). \end{aligned}$$

This system is said to be disturbance decoupled with respect to v and y if $y(\cdot, x_0, u, v) = y(\cdot, x_0, u, \tilde{v})$ for all admissible $u, v, \tilde{v}$, and for all $x_0 \in M$. That is, the output is unaffected by the disturbance. The disturbance decoupling problem is to find a feedback law $u_i \to u_i + k_i(x)$ such that the resulting system is disturbance decoupled with respect to v and y. For time-invariant linear systems with $D_i(x)$ constant vector fields and $k_i(x)$ linear, Wonham [5] defines a unique subspace $\mathscr{V}_C^*$ of $\ker C$ called the supremal (A, B)-invariant subspace of $\ker C$, and shows that the disturbance decoupling problem is solvable if and only if $D_1, \cdots, D_r \in \mathscr{V}_C^*$. The purpose of this paper is to generalize these ideas to study disturbance decoupling in the nonlinear case.

In § 2 the notion of $(A, \mathscr{B})$-invariant distributions on manifolds is introduced and some of the basic properties of these distributions are examined. In § 3 the relationship between $(A, \mathscr{B})$-invariant distributions and disturbance decoupling in nonlinear systems is studied. Our main result is Theorem 3.1, which shows that the local

* Received by the editors July 18, 1979, and in revised form April 14, 1980.
† Department of Mathematics and Statistics, Queen's University, Kingston, Ontario, Canada K7L 3N6.

disturbance decoupling problem is solvable with a nonlinear feedback law if and only if $D_1, \cdots, D_r \in \mathscr{D}_C^*$, the maximal $(A, \mathscr{B})$-invariant distribution with $XC = 0$ for all X belonging to the distribution. Section 4 contains some examples.

We conclude this section with some definitions. Let $X \in V(M)$ be a real analytic vector field. If $(\mathscr{V}, x_1, \cdots, x_n)$ is a coordinate system on M, then locally

$$X|_{\mathscr{V}} = \sum_{i=1}^{n} a_i(x) \frac{\partial}{\partial x_i},$$

where $\{a_i\}$ are real analytic functions on M. We let $t \to X_t \cdot x_0$ denote the integral curve for X passing through x_0 when $t = 0$. Thus $(d/dt)X_t \cdot x_0 = X(X_t \cdot x_0)$ and $X_0 \cdot x_0 = x_0$.

If $X, Y \in V(M)$ then the vector field $\mathrm{ad}_X Y$ or $[X, Y] = XY - YX$ is called the *Lie Bracket* of X and Y. If on $\mathscr{V}$ $X = \sum_{i=1}^{n} a_i \, \partial/\partial x_i$ and $Y = \sum_{i=1}^{n} b_i \, \partial/\partial x_i$, then

$$[X, Y] = \sum_{i,j=1}^{n} a_i \frac{\partial b_j}{\partial x_i} \frac{\partial}{\partial x_j} - \sum_{i,j=1}^{n} b_i \frac{\partial a_j}{\partial x_i} \frac{\partial}{\partial x_j}.$$

It follows that for $f, g \in C^w(M)$, the ring of real analytic functions on M, $[fX, gY] = fg[X, Y] - g(Yf)X + f(Xg)Y$, where locally $Xg = \sum a_i \, \partial g/\partial x_i$ (cf. [4]). In particular X defines a linear mapping from $C^w(M)$ into $C^w(M)$, and it is easy to see that for any $f \in C^w(M)$, $Xf(x) = df_x X(x)$ for all $x \in M$, where $df_x : T_x(M) \to T_{f(x)}(R)$ is the differential of f, a linear mapping of tangent spaces. If $f = (f_1, \cdots, f_p)$ is a vector valued real analytic function on M we let $Xf = (Xf_1, \cdots, Xf_p)$.

A distribution $\mathscr{D}$ on M is a choice of a subspace of the tangent space $T_x(M)$ for each $x \in M$. $\mathscr{D}$ is *k-dimensional* if $\dim \mathscr{D}(x) = k$ for all $x \in M$. $\mathscr{D}$ is real *analytic* if for each $x_0 \in M$ there exist a neighborhood $\mathscr{U}_0$ of x_0 and vector fields $X_1, \cdots, X_q \in V(M)$ such that $\mathscr{D}(x) = \mathrm{span}\{X_1(x), \cdots, X_q(x)\}$ for all $x \in \mathscr{U}_0$. We say that a vector field X *belongs to* $\mathscr{D}$, $X \in D$, if $X(x) \in D(x)$ for all $x \in M$. A real analytic distribution $\mathscr{D}$ is *involutive* if for all $X, Y \in \mathscr{D}$, $[X, Y] \in \mathscr{D}$. If $X \in V(M)$ then we set $\mathrm{ad}_X \mathscr{D}(x) = \{\mathrm{ad}_X Y(x) | Y \in \mathscr{D}\}$, a real analytic distribution on M. If $\mathscr{D}_1, \mathscr{D}_2$ are real analytic distributions then $\mathscr{D}_1 \subseteq \mathscr{D}_2$ if for all vector fields $X \in \mathscr{D}_1$, $X \in \mathscr{D}_2$; and $(\mathscr{D}_1 + \mathscr{D}_2)(x) = \mathrm{span}\{\mathscr{D}_1(x) \cup \mathscr{D}_2(x)\}$ is the *distribution generated by* $\mathscr{D}_1$ *and* $\mathscr{D}_2$. The *involutive distribution generated by* $\mathscr{D}_1$ *and* $\mathscr{D}_2$ is the smallest involutive distribution containing $(\mathscr{D}_1 + \mathscr{D}_2)$. Finally, if $\mathscr{D}$ is a real analytic distribution on M and $x \in M$, $I(\mathscr{D}, x)$ will denote the *maximal integral manifold of* $\mathscr{D}$ *through* x—the largest connected submanifold N of M with $x \in N$ and $T_y(N) = \mathscr{D}(y)$ for all $y \in N$.

2. Invariant distributions. For time-invariant linear systems the notion of (A, B)-invariant subspaces due to Wonham and others (cf., [5], [6]) is a central concept in the geometric approach to disturbance isolation, output decoupling, etc. In studying similar structural properties for nonlinear systems of the form (1) it is natural to look for some invariant which generalizes the notion of (A, B)-invariant subspaces and exposes some of the internal structure of nonlinear systems.

Consider the time-invariant linear system

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \qquad x \in R^n, \\ y(t) &= Cx(t), \end{aligned} \tag{2}$$

where $y \in R^l$. A subspace $\mathscr{V}$ of the state space R^n is said to be (A, B)-*invariant* if

$$A\mathscr{V} \subseteq \mathscr{V} + \mathscr{B}$$

where $\mathscr{B}$ is the range space of the matrix B (i.e., the linear operator obtained by

restricting A to $\mathcal{V}$ has its range space contained in the subspace $\mathcal{V}+\mathcal{B}$). Wonham has shown that this subspace plays a key role in understanding some of the deeper structural properties of time-invariant linear systems, properties which are not determined by the standard controllability and observability Gramians (cf. [1]). Since differential geometry provides a natural framework for studying nonlinear systems, a restatement of the above definition in the language of differential geometry is the obvious first step in creating an analogous construction for nonlinear systems. The state equation in (2) can be rewritten as

$$\dot{x}(t) = Ax(t) + \sum_{i=1}^{m} u_i(t) b_i \qquad x \in R^n,$$

where $b_1, \cdots, b_m$ are the columns of the matrix B, and we consider $x \mapsto Ax$ and $x \mapsto b_i$ to be real analytic vector fields on the manifold R^n. The subspace $\mathcal{V}$ can be considered as a submanifold of R^n, but then the condition $A\mathcal{V} \subseteq \mathcal{V}+\mathcal{B}$ becomes confusing: $\mathcal{V}$ plays the role of a submanifold so that for each $x \in \mathcal{V}$, $A(x) = Ax \in T_x(R^n)$, but in $\mathcal{V}+\mathcal{B}$ we have $\mathcal{V}$ being considered as a subspace of the tangent space $T_x(R^n)$.

To resolve this difficulty, we let $\mathcal{V}$ define a real analytic distribution on $M = R^n$. In fact $\mathcal{V}$ defines a "flat" real analytic distribution $\mathcal{D}_2$ on $M = R^n$ if we let

$$\mathcal{D}_{\mathcal{V}}(x) = \mathcal{V} \subseteq T_x(R^n),$$

for all $x \in R^n$, where the standard identification of $T_x(R^n)$ with R^n is made. Similarly $\mathcal{B}$ can be interpreted as the ("flat") distribution

$$\mathcal{B}(x) = \mathcal{B} \subseteq T_x(R^n).$$

The (A, B)-invariance of the subspace $\mathcal{V}$ requires that for each $x \in \mathcal{V}$, $Ax \in \mathcal{V}+\mathcal{B}$, or, if $\{v_1, \cdots, v_k\}$ is a basis for $\mathcal{V}$, that

$$Av_i \in \mathcal{V}+\mathcal{B} \quad \text{for } i = 1, \cdots, k.$$

To state this result in the language of distributions, we let $\{X_i(x) = v_i | i = 1, 2, \cdots, k\}$ denote k constant vector fields on R^n with the property that

$$\mathcal{D}_{\mathcal{V}}(x) = \text{span}\,\{X_1(x), \cdots, X_k(x)\}.$$

Thus

$$\text{ad}_A X_i(x) = (dX_i)_x Ax - (dA)_x X_i = 0 - Av_i \in \mathcal{D}_{\mathcal{V}} + \mathcal{B},$$

and if X is any real analytic vector field belonging to $\mathcal{D}_{\mathcal{V}}$ then $X(x) = \sum_{i=1}^{k} a_i(x) X_i(x)$ for some $a_i \in C^w(R^n)$ and

$$\begin{aligned} \text{ad}_A X(x) &= \left[A, \sum_{i=1}^{k} a_i X_i\right](x) \\ &= \sum_{i=1}^{k} (Aa_i)(x) X_i(x) + \sum_{i=1}^{k} a_i(x)(\text{ad}_A X_i(x)). \end{aligned}$$

Since $X_i(x) = v_i \in \mathcal{D}_{\mathcal{V}}$ and $\text{ad}_A X_i \in \mathcal{D}_{\mathcal{V}} + \mathcal{B}$, we have

$$\text{ad}_A X \in \mathcal{D}_{\mathcal{V}} + \mathcal{B} \quad \text{for all } X \in \mathcal{D}_{\mathcal{V}}.$$

This can conveniently be written as $\text{ad}_A \mathcal{D}_{\mathcal{V}} \subseteq \mathcal{D}_{\mathcal{V}} + \mathcal{B}$. Thus saying $\mathcal{V}$ is an (A, B)-invariant subspace is equivalent to saying that the "flat" distribution $\mathcal{D}_{\mathcal{V}}$ satisfies $\text{ad}_A \mathcal{D}_{\mathcal{V}} \subseteq \mathcal{D}_{\mathcal{V}} + \mathcal{B}$.

This suggests a natural way to generalize the idea of (A, B)-invariant subspaces to nonlinear systems of the form (1). Consider the system

$$\dot{x}(t) = A(x(t)) + \sum_{i=1}^{m} u_i(t) B_i(x(t)), \qquad x \in M,$$

where $A, B_1, \cdots, B_m \in V(M)$. Let $\mathcal{B}$ be the real analytic distribution on M defined by

$$\mathcal{B}(x) = \operatorname{span}\{B_1(x), \cdots, B_m(x)\}.$$

DEFINITION. An involutive real analytic distribution $\mathcal{D}$ on M is $(A, \mathcal{B})$-*invariant* if there exists an open dense submanifold M_0 of M such that $\operatorname{ad}_A \mathcal{D} \subseteq \mathcal{D} + \mathcal{B}$ and $\operatorname{ad}_{B_i} \mathcal{D} \subseteq \mathcal{D}$ on M_0 for $i = 1, 2, \cdots, m$.

Remark 1. For nonlinear systems the dimension of an $(A, \mathcal{B})$-invariant distribution $\mathcal{D}$ need not be constant (see Example 1).

Remark 2. The distribution $\mathcal{D}_{\mathcal{V}}$ on R^n which arises in the linear case satisfies the above definition. We have shown that $\operatorname{ad}_A \mathcal{D}_{\mathcal{V}} \subseteq \mathcal{D}_{\mathcal{V}} + \mathcal{B}$. To verify that $\operatorname{ad}_{B_i} \mathcal{D}_{\mathcal{V}} \subseteq \mathcal{D}_{\mathcal{V}}$ we note that for all $X \in \mathcal{D}_{\mathcal{V}}$, $X = \sum_{j=1}^{m} a_j X_j$, where $a_j \in C^w(R^n)$, $X_j(x) = v_j$, and hence with $B_i(x) = b_i$,

$$\operatorname{ad}_{b_i} X(x) = \sum_{j=1}^{m} [(b_i a_j) X_j + a_j \operatorname{ad}_{b_i} X_j].$$

Since $\operatorname{ad}_{b_j} X_j = 0$ and $X_j \in \mathcal{D}_{\mathcal{V}}$, we have $\operatorname{ad}_{b_i} \mathcal{D}_{\mathcal{V}} \in \mathcal{D}_{\mathcal{V}}$. It is clear that if $X, Y \in \mathcal{D}_{\mathcal{V}}$ then $[X, Y] \in \mathcal{D}_{\mathcal{V}}$; hence $\mathcal{D}_{\mathcal{V}}$ is involutive, and we are setting $M_0 = M = R^n$.

We conclude this section by establishing some of the basic properties of $(A, \mathcal{B})$-invariant distributions.

LEMMA 2.1. *Suppose that $\mathcal{D}_1$ and $\mathcal{D}_2$ are $(A, \mathcal{B})$-invariant distributions on M. Let $\mathcal{D}$ be the involutive distribution on M generated by $\mathcal{D}_1$ and $\mathcal{D}_2$. Then $\mathcal{D}$ is $(A, \mathcal{B})$-invariant and $\mathcal{D}_1, \mathcal{D}_2 \subseteq \mathcal{D}$.*

Proof. Let $X = X_1 + X_2$ be a real analytic vector field belonging to $\mathcal{D}_1 + \mathcal{D}_2$, where $X_1 \in \mathcal{D}_1$ and $X_2 \in \mathcal{D}_2$. Let M_1 and M_2 be open and dense submanifolds of M such that $\operatorname{ad}_{B_i} \mathcal{D}_i \subseteq \mathcal{D}_i$ and $\operatorname{ad}_A \mathcal{D}_i \subseteq \mathcal{D}_i + \mathcal{B}$ on M_i for $i = 1, 2$ and $j = 1, \cdots, m$ It follows that

$$\operatorname{ad}_{B_i} X = \operatorname{ad}_{B_i} X_1 + \operatorname{ad}_{B_i} X_2 \in \mathcal{D}_1 + \mathcal{D}_2 \quad \text{on } M_1 \cap M_2$$

for $i = 1, \cdots, m$, and that

$$\operatorname{ad}_A X = \operatorname{ad}_A X_1 + \operatorname{ad}_A X_2 \in (\mathcal{D}_1 + \mathcal{B}) + (\mathcal{D}_2 + \mathcal{B}) \quad \text{on } M_1 \cap M_2.$$

Thus $\operatorname{ad}_A (\mathcal{D}_1 + \mathcal{D}_2) \subseteq (\mathcal{D}_1 + \mathcal{D}_2) + \mathcal{B}$ and $\operatorname{ad}_{B_i} (\mathcal{D}_1 + \mathcal{D}_2) \subseteq (\mathcal{D}_1 + \mathcal{D}_2)$ for $i = 1, \cdots, m$ on $M_0 = M_1 \cap M_2$, an open and dense submanifold of M.

If $\mathcal{D}_1 + \mathcal{D}_2$ is not involutive, we construct a larger distribution $\mathcal{D}_{12}$ which includes the Lie brackets of the form $[X_1, X_2]$ where $X_1 \in \mathcal{D}_1$ and $X_2 \in \mathcal{D}_2$. A straightforward computation shows that $\operatorname{ad}_A \mathcal{D}_{12} \subseteq \mathcal{D}_{12} + \mathcal{B}$ and $\operatorname{ad}_{B_i} \mathcal{D}_{12} \subseteq \mathcal{D}_{12}$ on M_0. Repeating this procedure a finite number of times results in the required $(A, \mathcal{B})$-invariant distribution $\mathcal{D}$, and this completes the proof.

The next result plays a key role in relating $(A, \mathcal{B})$-invariant distributions to the disturbance decoupling problem considered in § 3.

LEMMA 2.2. *Let $\mathcal{D}$ be an involutive real analytic distribution on M. Then $\mathcal{D}$ is $(A, \mathcal{B})$-invariant if and only if there exists an open dense submanifold M_0 of M such that $\operatorname{ad}_{B_i} \mathcal{D} \subseteq \mathcal{D}$ on M_0 for $i = 1, \cdots, m$, and for all $x_0 \in M_0$ there exists an open neighborhood U_0 of x_0 in M_0 and functions $k_1, \cdots, k_m \in C^w(U_0)$, such that on U_0*

$$\operatorname{ad}_{(A + \sum_{i=1}^{m} k_i B_i)} \mathcal{D} \subseteq \mathcal{D}.$$

Proof. Sufficiency. Fix $x_0 \in M_0$. Suppose there exist $k_1, \cdots, k_m \in C^w(\mathcal{U}_0)$ such that

$$\mathrm{ad}_{(A+\sum_{i=1}^m k_i B_i)} \mathcal{D} \subseteq \mathcal{D},$$

where $x_0 \in \mathcal{U}_0 \subseteq M_0$ are defined as above. Choose $X \in \mathcal{D}$. Then

$$\mathrm{ad}_{(A+\sum_{i=1}^m k_i B_i)} X = \mathrm{ad}_A X + \sum_{i=1}^m (k_i \, \mathrm{ad}_{B_i} X - (Xk_i)B_i)$$

is a vector field on $\mathcal{U}_0$, and since $\mathrm{ad}_{B_i} \mathcal{D} \subseteq \mathcal{D}$ on M_0, we see that $\mathrm{ad}_A X(x) \in \mathcal{D}(x) + \mathcal{B}(x)$ for all $x \in \mathcal{U}_0$. This implies that $\mathrm{ad}_A \mathcal{D} \subseteq \mathcal{D} + \mathcal{B}$ on M_0 and hence $\mathcal{D}$ is $(A, \mathcal{B})$-invariant.

Necessity. Suppose $\mathcal{D}$ is $(A, \mathcal{B})$-invariant. Then there exists an open dense submanifold N_1 of M, and on N_1 we have $\mathrm{ad}_A \mathcal{D} \subseteq \mathcal{D} + \mathcal{B}$ and $\mathrm{ad}_{B_i} \mathcal{D} \subseteq \mathcal{D}$ for $i = 1, \cdots, m$. Let $\alpha = \max \{\dim \mathcal{D}(x) | x \in N_1\}$, and set $N_2 = \{x \in N_1 | \text{dimension } \mathcal{D}(x) = \alpha\}$. Similarly we let $\beta = \max \{\dim (\mathcal{D}(x) + \mathcal{B}(x)) | x \in N_2\}$ and pick a minimal subset $\{B_{l_1}, \cdots, B_{l_q}\}$ of $\{B_1, \cdots, B_m\}$ such that the dimension of $\mathcal{D}(x) + \mathrm{span}\{B_{l_1}(x), \cdots, B_{l_q}(x)\}$ is β for some $x \in N_2$. We now let $M_0 = \{x \in N_2 | \text{dimension } (\mathcal{D}(x) + \mathrm{span}\{B_{l_1}(x), \cdots, B_{l_q}(x)\}) = \beta\}$. It follows from the real analyticity of $\mathcal{D}$ that M_0 is an open and dense submanifold of N and hence of M. Fix $x_0 \in M_0$. Now choose an open neighborhood $\mathcal{U}_0$ of x_0 in M_0 and a coordinate map $\phi(x) = (x_1(x), \cdots, x_n(x)) : \mathcal{U}_0 \to R^n$ such that $\phi(x_0) = (0, \cdots, 0)$ and $\{\partial/\partial x_1, \cdots, \partial/\partial x_\alpha\}$ span $\mathcal{D}$ on $\mathcal{U}_0$.

That such a coordinate system exists is a consequence of the local version of Frobenius' Theorem (cf. [4]). Since $\mathrm{ad}_A \mathcal{D} \subseteq \mathcal{D} + \mathcal{B}$, and $X_i = \partial/\partial x_i \in \mathcal{D}$ on $\mathcal{U}_0$ for $i = 1, 2, \cdots, \alpha$, it follows that

$$\mathrm{ad}_A X_i = \sum_{j=1}^{\alpha} h_j^i X_j + \sum_{j=1}^{q} g_j^i B_{l_j}, \tag{3}$$

where h_j^i and g_j^i are unique functions in $C^w(\mathcal{U}_0)$ for $i = 1, \cdots, \alpha$. To complete the proof we must find $k_1, \cdots, k_m \in C^w(\mathcal{U}_0)$ with the property that on $\mathcal{U}_0$,

$$\mathrm{ad}_{(A+\sum_{j=1}^m k_j B_j)} \mathcal{D} \subseteq \mathcal{D},$$

or equivalently, for $i = 1, 2, \cdots, \alpha$,

$$\mathrm{ad}_{(A+\sum_{j=1}^m k_j B_j)} X_i \in \mathcal{D}.$$

Computing this Lie bracket we have the condition

$$\mathrm{ad}_A X_i + \sum_{j=1}^m k_j \, \mathrm{ad}_{B_j} X_i - \sum_{j=1}^m (X_i k_j) B_j \in \mathcal{D}.$$

We know from (3) that on $\mathcal{U}_0$, $\mathrm{ad}_A X_i - \sum_{j=1}^q g_j^i B_{l_j} \in \mathcal{D}$, and since $\mathrm{ad}_{B_j} X_i \in \mathcal{D}$ this condition becomes

$$\sum_{j=1}^m (X_i k_j) B_j - \sum_{j=1}^q g_j^i B_{l_j} \in \mathcal{D} \quad \text{for } i = 1, 2, \cdots, \alpha.$$

Letting certain of the k_j's be identically zero, the proof comes down to finding $k_{l_1}, \cdots, k_{l_q} \in C^w(\mathcal{U}_0)$ such that

$$\sum_{j=1}^q (X_i k_{l_j} - g_j^i) B_{l_j} \in \mathcal{D} \quad \text{for } i = 1, \cdots, \alpha.$$

Since $B_{l_1}, \cdots, B_{l_q} \notin \mathcal{D}$ it suffices to find k_{l_j}'s such that

$$X_i k_{l_j} = g_j^i \quad \text{for } j = 1, \cdots, q \text{ and } i = 1, \cdots, \alpha,$$

and the existence of the k_{l_j}'s depends on the following fact.

CLAIM.

$$X_r g_j^s = X_s g_j^r \quad \textit{for } 1 \leqq j \leqq q \textit{ and } 1 \leqq r, s \leqq \alpha. \tag{4}$$

To verify this we apply ad_{X_r} to both sides of (3). Since $\mathrm{ad}_{X_r} X_s = 0$ and $\mathrm{ad}_{B_i} X_r \in \mathcal{D}$, we find that

$$\begin{aligned} \mathrm{ad}_{X_r} \mathrm{ad}_A X_s &= \sum_{j=1}^{\alpha} (X_r h_j^s) X_j + \sum_{j=1}^{q} (X_r g_j^s) B_{l_j} + \sum_{j=1}^{q} g_j^s \, \mathrm{ad}_{X_r} B_{l_j} \\ &= \sum_{j=1}^{\alpha} p_j^s X_j + \sum_{j=1}^{q} (X_r g_j^s) B_{l_j}, \end{aligned}$$

for some $p_j^s \in C^w(\mathcal{U}_0)$. Similarly,

$$\mathrm{ad}_{X_s} \mathrm{ad}_A X_r = \sum_{j=1}^{\alpha} t_j^s X_j + \sum_{j=1}^{q} (X_s g_j^r) B_{l_j},$$

for some $t_j^s \in C^w(\mathcal{U}_0)$. Using the Jacobi identity, we have

$$\begin{aligned} \mathrm{ad}_{X_s} \mathrm{ad}_A X_r &= [[X_s, A], X_r] + [A, [X_s, X_r]] \\ &= [X_r, [A, X_s]] = \mathrm{ad}_{X_r} \mathrm{ad}_A X_s, \end{aligned}$$

and thus (4) is established.

We conclude this proof by constructing $k_{l_j} \in C^w(\mathcal{U}_0)$, with the property that $X_i k_{l_j} = g_j^i$ for $i = 1, \cdots, \alpha$ and $j = 1, \cdots, q$. For convenience we consider the case where $j = 1$ and let $f = k_{l_1}$. Using our chart $(\mathcal{U}_0, \phi)$ we can consider x_0 to be the origin 0 in R^n, and shrinking $\mathcal{U}_0$ if necessary we can treat $\mathcal{U}_0$ as an open ball in R^n centered at the origin. Thus we must find f such that

$$\begin{aligned} X_1 f &= \frac{\partial f}{\partial x_1} = g_1^1, \\ X_2 f &= \frac{\partial f}{\partial x_2} = g_1^2, \\ &\vdots \quad \vdots \quad \vdots \\ X_\alpha f &= \frac{\partial f}{\partial x_\alpha} = g_1^\alpha \text{ on } \mathcal{U}_0. \end{aligned}$$

The obvious candidate is obtained by integration, and we set $f(0, \cdots, 0, x_{\alpha+1}, \cdots, x_n) = 0$ and

$$\begin{aligned} f(x_1, \cdots, x_\alpha, x_{\alpha+1}, \cdots, x_n) = &\int_0^{x_1} g_1^1(t, 0, \cdots, 0, x_{\alpha+1}, \cdots, x_n)\, dt \\ &+ \int_0^{x_2} g_1^2(x_1, t, 0, \cdots, 0, x_{\alpha+1}, \cdots, x_n)\, dt \\ &+ \cdots \\ &+ \int_0^{x_\alpha} g_1^\alpha(x_1, \cdots, x_{\alpha-1}, t, \cdots, x_n)\, dt. \end{aligned}$$

Using (4) and the fundamental theorem of calculus it is easy to verify that $\partial f/\partial x_j = g_1^j$ for $j = 1, \cdots, \alpha$, which completes the proof.

For linear systems (2) Wonham has introduced the notion of the supremal (A, B)-invariant subspace of a given subspace $\mathscr{S}$ of R^n. If $C: R^n \to R^l$ is the output map of the linear system (2), then $\mathscr{V}_C^*$, *the supremal* (A, B)*-invariant subspace* of $\ker C$, is the subspace of R^n which is relevant to the solution of the disturbance decoupling problem [5]. We now present a generalization of this notion for nonlinear systems, and use these ideas to study the nonlinear disturbance decoupling problem in § 3.

Consider the nonlinear system (1) with state space M, and let $\mathscr{D}$ be a real analytic involutive distribution on M. An $(A, \mathscr{B})$-invariant distribution $\mathscr{D}^* \subseteq \mathscr{D}$ is called maximal if for all $(A, \mathscr{B})$-invariant distributions $\hat{\mathscr{D}} \subseteq \mathscr{D}$ we have $\hat{\mathscr{D}} \subseteq \mathscr{D}^*$.

THEOREM 2.3. *Let $\mathscr{D}$ be a real analytic involutive distribution on M. Then there exists a unique maximal $(A, \mathscr{B})$-invariant distribution $\mathscr{D}^*$ on M such that $\mathscr{D}^* \subseteq \mathscr{D}$.*

Proof. Let $\mathscr{V}^*$ denote the subset of $V(M)$ consisting of those vector fields $X \in V(M)$ which belongs to an $(A, \mathscr{B})$-invariant distribution contained in $\mathscr{D}$. That is, $X \in \mathscr{V}^*$ if and only if there exists an $(A, \mathscr{B})$-invariant distribution $\mathscr{D}_X \subseteq \mathscr{D}$ with $X \in \mathscr{D}_X$. Since $\mathscr{D}$ is involutive, Lemma 2.1 implies that $\mathscr{V}^*$ is a subspace of $V(M)$. Consider the distribution $\mathscr{D}^*: x \to \mathscr{V}^*(x)$ on M. By construction $\mathscr{D}^*$ contains every $(A, \mathscr{B})$-invariant distribution on M and is a real analytic distribution. To complete the proof we must show that $\mathscr{D}^*$ is involutive and $(A, \mathscr{B})$-invariant. Choose $X, Y \in \mathscr{D}^*$. Then locally $X = \sum_{i=1}^{p} \alpha_i(x) X_i$ and $Y = \sum_{i=1}^{q} \beta_i(x) Y_i$, where $X_i, Y_i \in \mathscr{V}^*$ by the definition of $\mathscr{D}^*$, and the functions α_i, β_i are defined locally. Since $X_1, \cdots, X_p$ are each members of $(A, \mathscr{B})$-invariant distributions on M, Lemma 2.1 implies that X belongs to an $(A, \mathscr{B})$-invariant distribution $\mathscr{D}_X$ on M, and the same results holds for Y. Invoking Lemma 2.1 again, we see that $[X, Y] \in \mathscr{D}_{[X,Y]}$, an $(A, \mathscr{B})$-invariant distribution on M, hence $\mathscr{D}^*$ is involutive. Finally, if $X \in \mathscr{D}^*$ we know that $X \in \mathscr{D}_X \subseteq \mathscr{D}^*$, where $\mathrm{ad}_A X \subseteq \mathscr{D}_X + \mathscr{B}$ and $\mathrm{ad}_{B_i} X \in \mathscr{D}_X$ on an open dense submanifold M_0 of M; thus $\mathrm{ad}_{B_i} X \in \mathscr{D}^*$, $\mathrm{ad}_A X \in \mathscr{D}^* + \mathscr{B}$ on M_0, which implies that $\mathscr{D}^*$ is $(A, \mathscr{B})$-invariant. This completes the proof.

We conclude this section by presenting the nonlinear system generalization of the supremal (A, B)-invariant subspace of $\ker C$ which is at the heart of a number of linear systems results on disturbance decoupling.

Let $C: M \to R^l$ denote the real analytic output map for the nonlinear system (1). Set $\mathscr{V}_C = \{X \in \mathscr{V}(M) | XC = 0\}$. Clearly $\mathscr{V}_C$ is a Lie subalgebra of $V(M)$, and thus the distribution

$$\mathscr{D}_C: x \to \mathscr{V}_C(x)$$

is real analytic and involutive. Using Theorem 2.3 there exists a unique maximal $(A, \mathscr{B})$-invariant distribution $\mathscr{D}_C^*$ contained in $\mathscr{D}_C$.

Thus $\mathscr{D}_C^*$ is the unique maximal $(A, \mathscr{B})$-invariant distribution $\mathscr{D}$ with the property that $XC = 0$ for all $X \in \mathscr{D}$. For linear systems the "flat" distribution $\mathscr{D}_{\mathscr{V}_C^*}: x \to \mathscr{V}_C^*$ has the property that for all $X \in \mathscr{D}_{\mathscr{V}_C^*}$, $X(x) = v \in \mathscr{V}_C^*$, and $XC = CX(x) = Cv = 0$ as $v \in \ker C$. Thus

$$\mathscr{D}_{\mathscr{V}_C^*} \subseteq \mathscr{D}_C^*$$

in the linear case.

3. Disturbance decoupling. The disturbance decoupling problem for linear systems involves changing the system dynamics by linear feedback so that a disturbance which drives the system has no effect on the output. In the nonlinear case we consider

the system

$$\begin{aligned} &\dot{x}(t) = A(x(t)) + \sum_{i=1}^{m} u_i(t)B_i(x(t)) + \sum_{i=1}^{r} v_i(t)D_i(x(t)), \qquad x \in M, \\ &y(t) = C(x(t)), \end{aligned} \tag{5}$$

where M is a real analytic manifold, $C: M \to R^l$ is real analytic, $A, B_i, D_j \in V(M)$, and u_i, v_j are admissible controls in U. Here $u = (u_1, \cdots, u_m)$ is an input vector, and $v = (v_1, \cdots, v_r)$ represents the effect of a disturbance. The definition for $(A, \mathcal{B})$-invariance for (5) is unchanged from the definition given in § 2.

DEFINITION. The system (5) is said to be *disturbance decoupled with respect to v and y* if

$$y(\cdot, x_0, u, v) = y(\cdot, x_0, u, \bar{v}),$$

for all $x_0 \in M$ and all admissible u, v, and $\bar{v}$.

DEFINITION. Given the system (5), the *Global Disturbance Decoupling Problem* (GDDP) is to find a nonlinear feedback law $u_i \to u_i + k_i(x)$, where $k_i \in C^w(M)$ for $i = 1, \cdots, m$, such that

$$\begin{aligned} &\dot{x} = (A(x) + \sum_{i=1}^{m} k_i(x)B_i(x)) + \sum_{i=1}^{m} u_iB_i(x) + \sum_{i=1}^{r} v_iD_i(x), \\ &y = C(x), \end{aligned}$$

is disturbance decoupled with respect to v and y.

The *Local Disturbance Decoupling Problem* (LDDP) is to find an open dense submanifold M_0 of M with the property that for all $x_0 \in M_0$ there exists an open neighborhood $\mathcal{U}_0$ of x_0 in M_0 and $k_1, \cdots, k_m \in C^w(\mathcal{U}_0)$, such that

$$\begin{aligned} &\dot{x} = (A(x) + \sum_{i=1}^{m} k_i(x)B_i(x)) + \sum_{i=1}^{m} u_iB_i(x) + \sum_{i=1}^{r} v_iD_i(x), \qquad x \in \mathcal{U}_0, \\ &y = C(x), \end{aligned}$$

is disturbance decoupled with respect to v and y.

For time-invariant linear systems (2) Wonham [5], [6] considers the Global Disturbance Decoupling Problem where $\sum_{i=1}^{r} v_iD_i(x) = Dv$ for a constant $n \times r$ matrix D and where $k_1(x), \cdots, k_m(x)$ are restricted to the class of linear mappings from R^n into R. He shows that this disturbance decoupling problem can be solved if and only if the range space of D is contained in the supremal (A, B)-invariant subspace $\mathcal{V}_C^*$ of the kernel of the linear output map C. This result can be stated as follows: the above disturbance decoupling problem can be solved if and only if the constant vector fields $D_1, D_2, \cdots, D_r$ described by the columns of D belong to the "flat" distribution $x \to \mathcal{V}_C^*$, denoted by $\mathcal{D}_{\mathcal{V}_C^*}$. The obvious generalization to the nonlinear case is to assert that the GDDP can be solved if and only if $D_1, \cdots, D_r \in \mathcal{D}_C^*$, where $\mathcal{D}_C^*$ is the maximal distribution described in the previous section. Example 1 shows that this obvious generalization is not valid globally, but is valid locally. This is in contrast to the linear case, where for $k_1, \cdots, k_m$ linear the GDDP and LDDP are equivalent. The following theorem shows that the role played by $\mathcal{D}_C^*$ in the LDDP is analogous to the role played by $\mathcal{V}_C^*$ in studying disturbance decoupling in the linear case.

THEOREM 3.1. *For the nonlinear system* (5) *let* $\mathcal{D}_C^*$ *denote the unique maximal* $(A, \mathcal{B})$*-invariant distribution* $\mathcal{D}$ *on* M *with* $XC = 0$ *for all* $X \in \mathcal{D}$. *Then the* LDDP *is solvable if and only if* $D_1, \cdots, D_r \in \mathcal{D}_C^*$.

We delay the proof of Theorem 3.1 to establish the following technical lemma.

LEMMA 3.2. *Suppose that* $D_1, \cdots, D_r \in V(M)$, $k_1, \cdots, k_m \in C^w(\mathcal{U})$, *where* $\mathcal{U}$ *is an open subset of* M, $\hat{A}(x) = A(x) + \sum_{i=1}^m k_i(x)B_i(x)$, *and* $C: M \to R^l$ *is the output map for the system* (5). *Let* $\mathcal{D}^{\mathcal{U}}$ *denote the* $\mathrm{ad}_{\hat{A}}$ *and* $\mathrm{ad}_{B_1}, \cdots, \mathrm{ad}_{B_n}$*-invariant involutive distribution on* $\mathcal{U}$ *generated by* $D_1, \cdots, D_r$, *and suppose that* $XC = 0$ *on* $\mathcal{U}$ *for all* $X \in \mathcal{D}^{\mathcal{U}}$. *Then there exists an* $(A, \mathcal{B})$*-invariant distribution* $\mathcal{D}$ *on* M *with the property that* $D_1, \cdots, D_r \in \mathcal{D}$ *and for all* $X \in \mathcal{D}$, $XC = 0$ *on* M.

Proof (Lemma 3.2.). Suppose $\mathcal{U} \subseteq M$ and $k_1, \cdots, k_m \in C^w(\mathcal{U})$ satisfy the hypotheses of this lemma. Then Lemma 2.2 implies that $\mathcal{D}^{\mathcal{U}}$ is an $(A, \mathcal{B})$-invariant distribution on $\mathcal{U}$. Suppose we can construct an (A, B)-invariant distribution $\mathcal{D}$ on M which extends $\mathcal{D}^{\mathcal{U}}$ to all of M and with $D_1, \cdots, D_r \in \mathcal{D}$. Then the restriction of $\mathcal{D}$ to $\mathcal{U}$ is $\mathcal{D}^{\mathcal{U}}$, which means $XC = 0$ on $\mathcal{U}$ for $X \in \mathcal{D}$, and by real analyticity $XC = 0$ on M, which will complete the proof.

Before constructing the extension $\mathcal{D}$ of $\mathcal{D}^{\mathcal{U}}$ we construct two special distributions. Shrinking $\mathcal{U}$ if necessary and relabeling $B_1, \cdots, B_m$ we can assume that $\{B_1(x), \cdots, B_q(x)\}$ are linearly independent on $\mathcal{U}$ and $\operatorname{span}\{B_1(x), \cdots, B_m(x)\} = \mathcal{B}(x) = \operatorname{span}\{B_1(x), \cdots, B_q(x)\}$ for all $x \in \mathcal{U}$. Shrinking $\mathcal{U}$ again if necessary, we can choose $E_1, \cdots, E_p \in V(M)$ such that $(dC)E_i = E_iC = 0$ for $i = 1, \cdots, p$; $\{B_1, \cdots, B_q, E_1, \cdots, E_p\}$ are linearly independent on $\mathcal{U}$; and

$$\operatorname{span}\{B_1(x), \cdots, B_q(x), E_1(x), \cdots, E_p(x)\} = \mathcal{B}(x) + \ker dC_x$$

for all $x \in \mathcal{U}$. Let $M_0 = \{x \in M | \{B_1(x), \cdots, B_q(x), E_1(x), \cdots, E_p(x)\}$ are linearly independent$\}$. Then M_0 is an open and dense submanifold of M as a consequence of the real analyticity of the vector fields. Using the fact that any real analytic manifold can be embedded in R^s for same $s > 0$ (cf. [2]) we can give M (and M_0) the structure of a real analytic Riemmanian manifold by using the standard inner product $\langle \cdot, \cdot \rangle$ on R^s to induce a Riemannian structure on M (thus $\langle \cdot, \cdot \rangle_x = \langle \cdot, \cdot \rangle$ is an inner product on $T_x(M)$ for all $x \in M$). Now $S(x) = \{\ker dC_x + \mathcal{B}(x)\}$ is a distribution on M, and on M_0

$$\mathcal{S}(x) = \operatorname{span}\{B_1(x), \cdots, B_q(x), E_1(x), \cdots, E_p(x)\}$$

is a $(q+p)$-dimensional real analytic distribution. We define the $(n-p-q)$-dimensional distribution $\mathcal{S}^{\perp}$ on M_0 by

$$\mathcal{S}^{\perp}(x) = \{v \in T_x(M) | \langle v, w \rangle_x = 0 \text{ for all } w \in \mathcal{S}(x)\}.$$

We now begin our construction of a globally defined $(A, \mathcal{B})$-invariant distribution $\mathcal{D}$ which extends $\mathcal{D}^{\mathcal{U}}$. By definition $\mathcal{D}^{\mathcal{U}}$ is the $\mathrm{ad}_{\hat{A}}$ and ad_{B_i}-invariant involutive distribution generated by $D_1, \cdots, D_r$. We have shrunk $\mathcal{U}$ and relabeled $B_1, \cdots, B_m$ if necessary, so that on $\mathcal{U}$, $\operatorname{span}\{B_1, \cdots, B_m\} = \operatorname{span}\{B_1, \cdots, B_q\}$; thus we can assume that $\hat{A} = A + \sum_{i=1}^q k_1 B_1$, and generate an $\mathrm{ad}_{\hat{A}}$ and $\mathrm{ad}_{B_1}, \cdots, \mathrm{ad}_{B_q}$-invariant distribution $\mathcal{D}^{\mathcal{U}}$. We will generate $\mathcal{D}$ in stages.

Let $\mathcal{D}_1$ denote the $\mathrm{ad}_{B_1}, \cdots, \mathrm{ad}_{B_q}$-invariant involutive distribution generated from $D_1, \cdots, D_r$. Set $\mathcal{D}_1^{\mathcal{U}} = \mathcal{D}_1|_{\mathcal{U}}$. Clearly $\mathcal{D}_1^{\mathcal{U}}$ extends to a distribution on M, but in general $\mathcal{D}_1^{\mathcal{U}}$ will not be $\mathrm{ad}_{\hat{A}}$-invariant, so we let $\mathcal{D}_2^{\mathcal{U}}$ denote the distribution on $\mathcal{U}$ spanned by the vector fields

$$\{X, \mathrm{ad}_{\hat{A}} X | X \in \mathcal{D}_1\}.$$

Since $\mathcal{D}_1$ extends $\mathcal{D}_1^{\mathcal{U}}, \mathcal{D}_2^{\mathcal{U}} \supseteq \mathcal{D}_1^{\mathcal{U}}$, and at this stage our problem is to extend $\mathcal{D}_2^{\mathcal{U}}$ to

$\mathscr{D}_2 \supseteq \mathscr{D}_1$ where $\operatorname{ad}_A \mathscr{D}_1 \supseteq \mathscr{D}_2 + \mathscr{B}$, and $\mathscr{D}_2$ is defined on M. Pick $X \in \mathscr{D}_1$. By definition,

$$\operatorname{ad}_{\hat{A}} X = \left[A + \sum_{i=1}^{q} k_i B_i, X\right] = \operatorname{ad}_A X + \sum_{i=1}^{q} (k_i \operatorname{ad}_{B_i} X - (Xk_i)B_i) \in \mathscr{D}_2^{\mathscr{U}},$$

and since $\operatorname{ad}_{B_i} X \in \mathscr{D}_1^{\mathscr{U}} \subseteq \mathscr{D}_2^{\mathscr{U}}$, we have

$$\tilde{X} = \operatorname{ad}_A X - \sum_{i=1}^{q} (Xk_i)B_i \in \mathscr{D}_2^{\mathscr{U}}.$$

Since $\mathscr{D}_2^u \supseteq \mathscr{D}^{\mathscr{U}}$ and $ZC = 0$ on $\mathscr{U}$ for all $Z \in \mathscr{D}^{\mathscr{U}}$, $dC_x(\operatorname{ad}_A X(x) - \sum_{i=1}^{q} (Xk_i)B_i(x)) = 0$ for all $x \in \mathscr{U}$, and so $\operatorname{ad}_A X(x) \in \ker dC_x + \mathscr{B}(x) = \mathscr{S}(x)$ for all $x \in \mathscr{U}$. This means that for all $Z \in \mathscr{S}$, $\langle \operatorname{ad}_A X, Z\rangle = 0$ on $\mathscr{U}$, and hence by real analyticity, on M_0. In particular, $\operatorname{ad}_A X(x) \in \mathscr{S}(x)$ for all $x \in M_0$. From the definition of $\mathscr{S}$ it follows that $\alpha_1, \cdots, \alpha_q, \cdots, \alpha_{q+p} \in C^w(M_0)$, such that

$$\operatorname{ad}_A X = \sum_{i=1}^{q} \alpha_i B_i + \sum_{i=1}^{p} \alpha_{q+i} E_i \quad \text{on } M_0. \tag{6}$$

We note that $\alpha_i|_{\mathscr{U}} = Xk_i$ for $i = 1, \cdots, q$. In vector notation, $\operatorname{ad}_A X(x) = N_x \alpha(x)$, where $\alpha = (\alpha_1, \cdots, \alpha_{p+q})$ and N_x is a matrix whose columns are $B_1, \cdots, B_q, E_1, \cdots, E_p$. Using the pseudo-inverse for N_x, we have

$$\alpha(x) = (N^*N)_x^{-1} N_x^* \operatorname{ad}_A X(x),$$

where N^* is the transpose of N. Writing $(N^*N)_x^{-1}$ as $(1/\det (N^*N)_x) \operatorname{adj} (N^*N)_x$, and letting $g(x) = \det (N^*N)_x$, we have

$$g(x)\alpha(x) = \operatorname{adj} (N^*N)_x N_x^* \operatorname{ad}_A X(x) \quad \text{on } M_0.$$

The right-hand side of this equality is defined for all $x \in M$, so that $g(x)\alpha(x)$ is defined on all of M. Note that $g(x) = 0$ if and only if $x \notin M_0$, by the definition of M_0. Going back to (6), we see that

$$g(x) \operatorname{ad}_A X(x) - \sum_{i=1}^{q} g(x)\alpha_i(x)B_i(x) = \sum_{i=1}^{p} g(x)\alpha_{q+i}(x)E_i(x) \in \ker dC_x.$$

Set $\tilde{\tilde{X}} = g(\operatorname{ad}_A X - \sum_{i=1}^{q} \alpha_i B_i) \in V(M)$. By construction $\tilde{\tilde{X}}|_{\mathscr{U}} = g\tilde{X} \in \mathscr{D}_2^{\mathscr{U}}$, and since $g \neq 0$ on $\mathscr{U}$ the distribution $\mathscr{D}_2$ on M spanned by the vector fields $\{X, \tilde{\tilde{X}} | X \in \mathscr{D}_1\}$ has the property that $\mathscr{D}_2|_{\mathscr{U}} = \mathscr{D}_2^{\mathscr{U}}$ and for $X \in \mathscr{D}_1$, $\operatorname{ad}_A X \in \mathscr{D}_2 + \mathscr{B}$ on M_0 (where $g \neq 0$). Now in general $\mathscr{D}_2$ is not ad_{B_i}-invariant. Let $\mathscr{D}_3$ denote the $\operatorname{ad}_{B_1}, \cdots, \operatorname{ad}_{B_q}$-invariant involutive distribution on M generated by $\mathscr{D}_2$, and let $\mathscr{D}_3^{\mathscr{U}} = \mathscr{D}_3|_{\mathscr{U}}$. In general $\mathscr{D}_3^{\mathscr{U}}$ will not be $\operatorname{ad}_{\hat{A}}$-invariant, so we repeat the above procedure. Shrinking $\mathscr{U}$ if necessary, we can terminate the above process so that $\mathscr{D}^{\mathscr{U}} = \mathscr{D}_\beta^{\mathscr{U}}$ for some integer $b > 0$. Then $\operatorname{ad}_A \mathscr{D}_\beta \subseteq \mathscr{D}_\beta + \mathscr{B}$ on M_0, an open dense submanifold of M, and $\operatorname{ad}_{B_i} \mathscr{D}_\beta \subseteq \mathscr{D}_\beta$ on M_0 by construction. Setting $\mathscr{D} = \mathscr{D}_\beta$ completes the proof.

Proof (Theorem 3.1). *Sufficiency.* Suppose $D_1, \cdots, D_r \in \mathscr{D}_C^*$. Since $\mathscr{D}_C^*$ is $(A, \mathscr{B})$-invariant, Lemma 2.2 shows that there exists an open dense submanifold M_0 of M such that

$$\operatorname{ad}_{B_i} \mathscr{D}_C^* \subseteq \mathscr{D}_C^* \text{ on } M_0 \quad \text{for } i = 1, \cdots, m,$$

and for all $x_0 \in M_0$ there exists an open neighborhood $\mathscr{U}_0$ of x_0 in M_0 and $k_1, \cdots, k_m \in C^w(\mathscr{U}_0)$ such that on $\mathscr{U}_0$

$$\operatorname{ad}_{(A + \sum_{i=1}^{m} k_i B_i)} \mathscr{D}_C^* \subseteq \mathscr{D}_C^*.$$

Set $\bar{A}(x) = A(x) + \sum_{i=1}^{m} k_i(x) B_i(x)$. Thus $\mathscr{D}_C^*$ is $\mathrm{ad}_{\bar{A}}$ and ad_{B_i}-invariant $(i = 1, \cdots, m)$ on $\mathscr{U}_0$, and hence $D_1, \cdots, D_r$ generate an $\mathrm{ad}_{\bar{A}}$ and $\mathrm{ad}_{B_1}, \cdots, \mathrm{ad}_{B_m}$-invariant, involutive, and real analytic distribution $\mathscr{D} \subseteq \mathscr{D}_C^*$. In particular for all $X \in \mathscr{D}$, $XC = 0$ on $\mathscr{U}_0$, since $YC = 0$ for all $Y \in \mathscr{D}_C^*$. We now show that

$$\dot{x} = \bar{A}(x) + \sum_{i=1}^{m} u_i B_i(x) + \sum_{i=1}^{r} v_i D_i(x), \qquad x \in \mathscr{U}_0,$$

$$y = C(x),$$

is disturbance decoupled with respect to v and y. Since x_0 can be any point in M_0, this will complete the proof.

Let $\bar{x} \in \mathscr{U}_0$ be any initial state, let u be any control, and v any disturbance vector. We must show that $y(\cdot, \bar{x}, u, v)$ is independent of our choice of v. For t sufficiently small, u and v are real analytic, and differentiating with respect to t we have

$$\frac{d}{dt} y(t) = y^{(1)}(t) = \bar{A}C(x(t)) + \sum_{i=1}^{m} u_i B_i C(x(t)) + \sum_{i=1}^{r} v_i D_i C(x(t)).$$

Since $D_i \in \mathscr{D}_C^*$, we have $D_i C = 0$ on $\mathscr{U}_0$, and

$$\begin{aligned} \frac{d^2}{dt^2} y(t) &= y^{(2)}(t) \\ &= \bar{A}^2 C(x(t)) + \sum_{i=1}^{m} u_i B_i \bar{A} C(x(t)) + \sum_{i=1}^{m} \dot{u}_i B_i C(x(t)) \\ &\quad + \sum_{i=1}^{m} u_i \bar{A} B_i C(x(t)) + \sum_{j=1}^{r} \sum_{i=1}^{m} u_i v_j D_j B_i C(x(t)). \end{aligned}$$

We note that $\mathrm{ad}_{B_i} D_j \in \mathscr{D}$, and thus $(B_i D_j - D_j B_i) C = 0$. Since $D_j C = 0$, we see that $D_j B_i C = 0$. Thus both $y^{(1)}(0)$ and $y^{(2)}(0)$ are independent of v. A straightforward induction argument shows that terms of the form $D_j \bar{A}^{l_1} B_{i_1}^{m_1} \cdots \bar{A}^{l_p} B_{i_p}^{m_p} C$ equal 0 because $D_j \in \mathscr{D}$, an $\mathrm{ad}_{\bar{A}}$ and ad_{B_i}-invariant distribution with $XC = 0$ for all $X \in \mathscr{D}$. A second induction shows that $y^{(3)}(0), y^{(4)}(0), \cdots$ are all independent of v. Since $y(0) = c(\bar{x})$ we have shown that the Taylor coefficients for $t \to y(t, \bar{x}, u, v)$ are independent of v. This completes one half of our proof.

Necessity. Suppose that the LDDP is solvable. To show that $D_1, \cdots, D_r \in \mathscr{D}_C^*$, we will show that $\{D_1, \cdots, D_r\}$ is contained in an $(A, \mathscr{B})$-invariant distribution $\mathscr{D}$ with $XC = 0$ for all $X \in \mathscr{D}$. Since this implies that $\mathscr{D} \subseteq \mathscr{D}_C^*$, the maximal such $(A, \mathscr{B})$-invariant distribution, the proof will be complete.

Using Lemma 3.2 we can reduce the problem to finding an open neighborhood $\mathscr{U} \subseteq M$ and $k_1, \cdots, k_m \in C^w(M)$ such that the $\mathrm{ad}_{\hat{A}}$ and ad_{B^i}-invariant involutive distribution $\mathscr{D}_0$ on $\mathscr{U}$ generated by $D_1, \cdots, D_r$ has the property that $XC = 0$ on $\mathscr{U}$ for all $x \in \mathscr{D}_0$ (here $\hat{A}(x) = A(x) + \sum_{i=1}^{m} k_i(x) B_i(x)$). Since $XC = YC = 0$ implies that $[X, Y]C = 0$, it is only necessary to check that

$$\mathrm{ad}_{\hat{A}}^{l_1} \mathrm{ad}_{B_{i_1}}^{m_1} \cdots \mathrm{ad}_{\hat{A}}^{l_p} \mathrm{ad}_{B_{i_p}}^{m_p} D_j C = 0 \quad \text{on } \mathscr{U} \tag{7}$$

for $1 \leqq i_j \leqq m$; $l_1, m_1, \cdots, l_p, m_p \geqq 0$; $p = 0, 1, 2, \cdots$; and $j = 1, \cdots, r$. It is clear that (7) is satisfied if we can verify that

$$\mathrm{ad}_{X^{c_1}}^{l_1} \cdots \mathrm{ad}_{X^{c_p}}^{l_p} D_j C = 0 \quad \text{on } \mathscr{U} \tag{8}$$

for $j = 1, \cdots, r$; $l_1, \cdots, l_p \geqq 0$; $p = 0, 1, 2, \cdots$; and $X^{c_i} = \hat{A} + \sum_{j=1}^{m} c_i^j B_j$, where $c_i = (c_i^1, \cdots, c_i^m)$ is contained in a neighborhood of the origin in R^m for $i = 1, \cdots, p$.

We are assuming that the LDDP is solvable. Thus there exists an open dense submanifold M_0 of M such that for any $\bar{x} \in M_0$ there is an open neighborhood $\bar{x} \in \mathcal{U} \subseteq M_0$ and $k_i \in C^w(\mathcal{U})$, such that if $\hat{A}(x) = A(x) + \sum_{k=1}^{m} k_i(x) B_i(x)$ then the system

$$\dot{x} = \hat{A}(x) + \sum_{i=1}^{m} u_i B_i(x) + \sum_{i=1}^{r} v_i D_i(x), \qquad x \in \mathcal{U},$$

$$y = C(x)$$

is disturbance decoupled with respect to $v = (v_1, \cdots, v_r)$ and y. Thus

$$y(\cdot, x_0, u, v) = y(\cdot, x_0, u, \tilde{v}) \tag{9}$$

for all $x_0 \in \mathcal{U}$ and for all admissible inputs u and disturbances v, $\tilde{v}$. To complete our proof it suffices to show that condition (8) holds on $\mathcal{U}$.

Pick $x_0 \in \mathcal{U}$. If we set $u_1 = \cdots = u_m = 0$, the reachable set for the system $\dot{x} = \hat{A}(x) + \sum_{i=1}^{r} v_i D_i(x)$; $x(0) = x_0$ at time t is $\mathcal{R}_t(x_0)$, which has a nonempty interior in $I(\mathcal{L}_0, \hat{A}_t \cdot x_0)$, the maximal integral manifold for the distribution generated by $\mathcal{L}_0$, the smallest Lie subalgebra of $V(M)$ containing $\{D_i, \operatorname{ad}_{\hat{A}} D_i, \operatorname{ad}_{\hat{A}}^2 D_i, \cdots\}$ (this well-known result is due to Sussmann and Jurdevic [3]). Thus we can choose an admissible control v_0 and time $t_0 > 0$ such that $x_{t_0} = x(t_0, x_0, 0, v_0)$ is in the interior of $\mathcal{R}_{t_0}(x_0)$, and so there exists $\delta > 0$ such that for $|s| < \delta$ and $j = 1, \cdots, r$,

$$(D_j)_s \cdot x_{t_0} \in \mathcal{R}_{t_0}(x_0).$$

This means that for each $s \in (-\delta, \delta)$ there exists a control v_s such that

$$x(t_0, x_0, 0, v_s) = (D_j)_s \cdot x_{t_0}. \tag{10}$$

From (9) we know that $C(x(t_0, x_0, 0, v_s)) = C(x(t_0, x_0, 0, v_0))$ for $|s| < \varepsilon$, or equivalently that

$$C((D_j)_s \cdot x_{t_0}) = C(x_{t_0}) \quad \text{for } |s| < \varepsilon.$$

Let $c_1, c_2, \cdots, c_p \in \Omega$ be vectors in R^m of length less than or equal to ε, and let

$$u(t) = \begin{cases} 0 & \text{if } 0 \leqq t \leqq t_0, \\ c_1 & \text{if } t_0 < t \leqq t_1, \\ \vdots & \\ c_p & \text{if } t_{p-1} < t \leqq t_p, \end{cases}$$

$$v(t) = \begin{cases} v_0(t) & \text{if } 0 \leqq t \leqq t_0, \\ 0 & \text{if } t_0 < t, \end{cases}$$

$$\tilde{v}(t) = \begin{cases} v_s(t) & \text{if } 0 \leqq t \leqq t_0, \\ 0 & \text{if } t_0 < t, \end{cases}$$

for some partition $0 < t_0 < t_1 < \cdots < t_p$. Thus $u, v, \tilde{v} \in \mathcal{U}$ are admissible controls and for t_0 and t_1 sufficiently small it follows from (9) that

$$C(x(t, x_0, u, v)) = C(x(t, x_0, u, \tilde{v})) \quad \text{for } 0 \leqq t \leqq t_p. \tag{11}$$

Using (10) and the fact that u is a piecewise constant control, we see that

$$x(t_p, x_0, u, v) = X_{t_p}^{c_p} \circ \cdots \circ X_{t_2}^{c_2} \circ X_{t_1}^{c_1} \cdot x_{t_0}$$

and

$$x(t_p, x_0, u, \tilde{v}) = X_{t_p}^{c_p} \circ \cdots \circ X_{t_1}^{c_1} \circ (D_j)_s \cdot x_{t_0},$$

where $X^{c_l} = \hat{A} + \sum_{i=1}^{m} c_i^l B_i \in V(M)$. In particular,

$$x(t_p, x_0, u, \tilde{v}) = X_{t_p}^{cp} \circ \cdots \circ X_{t_1}^{c1} \circ (D_j)_s \circ X_{-t_p}^{c1} \circ \cdots \circ X_{-t_p}^{cp} \circ X_{t_p}^{cp} \circ \cdots \circ X_{t_1}^{c1} \cdot x_{t_0}$$

$$= X_{t_p}^{cp} \circ \cdots \circ X_{t_1}^{c1} \circ (D_j)_s \circ X_{-t_1}^{c1} \circ \cdots \circ X_{-t_p}^{cp} \cdot x(t_p, x_0, u, v),$$

and (11) implies that

$$(12) \quad C(x(t_p, x_0, u, v)) = C(X_{t_p}^{cp} \circ \cdots \circ X_{t_1}^{c1} \circ (D_j)_s \circ X_{-t_1}^{c1} \circ \cdots \circ X_{-t_p}^{cp} \cdot x(t_p, x_0 u, v)).$$

If we vary x_0 in $\mathscr{U}$, we may find that the curves $x(\cdot, x_0, u, v)$ or $x(\cdot, x_0, u, \tilde{v})$ leave $\mathscr{U}$ when $t \in [0, t_p]$. To avoid this we choose t_p and ε sufficiently small and choose a subset $\tilde{\mathscr{U}}$ of $\mathscr{U}$ so that for all $x_0 \in \tilde{\mathscr{U}}$, $x(t, x_0 u, v)$ and $x(t, x_0, u, \tilde{v}) \in \mathscr{U}$ for $t \in [0, t_p]$. Let $\mathscr{V}_{t_1, \cdots, t_p} = \{x(t_p, x_0, u, v) | x \in \tilde{\mathscr{U}}\} \subseteq \mathscr{U}$. We now choose $\varepsilon_1 > 0$ so that the intersection of all $\mathscr{V}_{d_1, \cdots, d_p}$ with $|d_i - t_i| < \varepsilon_1$ is a nonempty open set $\mathscr{V}$. It follows from (12) that for all $x \in \mathscr{V}$, an open subset of M,

$$C(x) = C(X_{d_p}^{cp} \circ \cdots \circ X_{d_1}^{c1} \circ (D_j)_s \circ X_{-d_1}^{c1} \circ \cdots \circ X_{-d_p}^{cp} \cdot x) \quad \text{for } |d_i - t_i| < \varepsilon_1.$$

We differentiate both sides of this expression and set $s = 0$ to obtain

$$0 = dC_x(dX_{d_p}^{cp} \circ \cdots \circ dX_{d_1}^{c1} D_j(X_{-d_1}^{c1} \circ \cdots \circ X_{-d_p}^{cp} \cdot x)) \quad \text{for } x \in \mathscr{V}.$$

Taking a Taylor series expansion we have

$$dX_{d_1}^{c1} D_j(X_{-d_1}^{c1} \cdot z) = \sum_{l_1=0}^{\infty} \frac{(-d_1)^{l_1}}{l_1!} \operatorname{ad}_{X^{c_1}}^{l_1} D_j(z),$$

and repeating this p times, we obtain

$$0 = dC_x\left(\sum_{l_p=0}^{\infty} \sum_{l_{p-1}=0}^{\infty} \cdots \sum_{l_1=0}^{\infty} \frac{(-d_1)^{l_1} \cdots (-d_p)^{l_p}}{l_1! \cdots l_p!} \operatorname{ad}_{X^{c_p}}^{l_p} \cdots \operatorname{ad}_{X^{c_1}}^{l_1} D_j(x)\right)$$

for all $x \in \mathscr{V}$ and for all $|d_i - t_i| < \varepsilon_1$. Varying the d_i's we conclude that

$$dC_x \operatorname{ad}_{X^{c_p}}^{l_p} \cdots \operatorname{ad}_{X^{c_1}}^{l_1} D_j(x) = 0 \quad \text{on } \mathscr{V},$$

or

$$\operatorname{ad}_{X^{c_p}}^{l_p} \cdots \operatorname{ad}_{X^{c_1}}^{l_1} D_j C = 0 \quad \text{on } \mathscr{V},$$

for $l_1, \cdots, l_p \geqq 0$; $j = 1, \cdots, r$; and $c_1, \cdots, c_p$ sufficiently near to $0 \in R^m$ (in an ε-ball centered at 0). The above equality on $\mathscr{V}$ implies that (8) holds on $\mathscr{U}$ by real analyticity for our fixed choice of p. Varying p and repeating the above argument establishes (8) and completes the proof.

Remark. It is natural to compare Theorem 3.1 with Wonham's results [5] for the time-invariant system $\dot{x} = Ax + Bu + Dv$; $y = Cx$, or equivalently $\dot{x} = Ax + \sum_{i=1}^{m} u_i B_i + \sum_{i=1}^{r} v_i D_i$; $y = Cx$, where B_i, D_j are the columns of B and D. In [5] it is shown that there exists a linear feedback law which solves the GDDP if and only if $D_1, \cdots, D_r \in \mathscr{V}_C^*$, the supremal (A, B)-invariant subspace of $\ker C$, or equivalently $D_1, \cdots, D_r$ belong to the "flat" distribution $\mathscr{D}_*$ where $\mathscr{D}_*(x) = \mathscr{V}_C^*$. Of course this means that if $D_1, \cdots, D_r \in \mathscr{D}_*$ then the LDDP is solvable; hence $\mathscr{D}_* \subseteq \mathscr{D}_C^*$ as a consequence of Theorem 3.1. Conversely, if $D_1, \cdots, D_r \in \mathscr{D}_C^*$, then Theorem 3.1 implies that the LDDP is solvable. It is natural to conjecture that if a constant vector field $D_i \in \mathscr{D}_C^*$ then $D_i \in \mathscr{D}_*$, and hence there exists a *globally* defined and *linear* feedback law which solves the GDDP. The following theorem shows that this is the case.

THEOREM 3.3. *Let $\mathscr{D}_C^*$ denote the maximal $(A, \mathscr{B})$-invariant distribution associated with the time-invariant linear system*

$$\dot{x} = Ax + Bu, \qquad x \in R^n,$$
$$y = Cx,$$

and let $\mathscr{D}_*$ denote the "flat" distribution generated by the subspace $\mathscr{V}_C^*$. Then $\mathscr{D}_* = \mathscr{D}_C^*$. In particular, if the LDDP *is solvable for the time-invariant linear system*

$$\dot{x} = Ax + Bu + Dv, \qquad x \in R^n,$$
$$y = Cx,$$

where B is an $n \times m$ matrix and D an $n \times r$ matrix, then there exists a linear feedback law which solves the GDDP.

Proof. Clearly $\mathscr{D}_* \subset \mathscr{D}_C^*$, so it suffices to show that if $X \in \mathscr{D}_C^*$ is a real analytic vector field, then $X(x) \in \mathscr{D}_*(x) = \mathscr{V}_C^*$ for all $x \in M$. As a consequence of real analyticity, it suffices to show that $X(x) \in \mathscr{V}_C^*$ for all x in some open neighborhood in M. Using Lemma 2.2, we choose a neighborhood $U \subseteq M$ and $k_1, \cdots, k_m \in C^w(U)$, such that on U,

$$\operatorname{ad}_{(A+\sum_{i=1}^m k_i b_i)} \mathscr{D}_C^* \subseteq \mathscr{D}_C^*.$$

Here $b_1, \cdots, b_m$ are the m columns of the matrix B. Set $\hat{A}(x) = Ax + \sum_{i=1}^m k_i(x) b_i$, so that $\operatorname{ad}_{\hat{A}} \mathscr{D}_C^* \subseteq \mathscr{D}_C^*$. Since $X \in \mathscr{D}_C^*$ we see that on U, $\operatorname{ad}_{\hat{A}} X = \operatorname{ad}_A X - \sum_{i=1}^m (Xk_i) b_i + \sum_{i=1}^m k_i \operatorname{ad}_{b_i} X \in \mathscr{D}_C^*$, and $(A, \mathscr{B})$-invariance implies that $\operatorname{ad}_{b_i} X \in \mathscr{D}_C^*$. Thus

$$\operatorname{ad}_A X = \sum_{i=1}^m a_{1i} b_i + D_1, \tag{13}$$

where $a_{1i}(x) = -(Xk_i)(x) \in C^w(U)$ and $D_1 \in \mathscr{D}_C^*$. Applying $\operatorname{ad}_{\hat{A}}$ to $\operatorname{ad}_A X$, we obtain

$$\operatorname{ad}_{\hat{A}} \operatorname{ad}_A X = \left[A + \sum_{i=1} k_i b_i, \operatorname{ad}_A X \right] = \operatorname{ad}_A^2 X - \sum_{i=1}^m (\operatorname{ad}_A X k_i) b_i + \sum_{i=1}^m k_i [b_i, \operatorname{ad}_A X],$$

and from (13) we see that

$$\begin{aligned}
\operatorname{ad}_{\hat{A}} \operatorname{ad}_A X &= \operatorname{ad}_{\hat{A}} \sum_{i=1}^m a_{1i} b_i + \operatorname{ad}_{\hat{A}} D_1 \\
&= \operatorname{ad}_A \sum_{i=1}^m a_{1i} b_i - \sum_{j=1}^m \sum_{i=1}^m (a_{1i} b_i k_j) b_j + \sum_{i=1}^m k_i \left[b_i, \sum_{i=1}^m a_{1i} b_i \right] + \operatorname{ad}_{\hat{A}} D_1 \\
&= \sum_{i=1}^m (A a_{1i}) b_i + \sum_{i=1}^m a_{1i} \operatorname{ad}_A b_i - \sum_j \left\{ \sum_{i=1}^m (a_{1i} b_i k_j) \right\} b_j \\
&\qquad + \sum_{i=1}^m k_i [b_i, \operatorname{ad}_A X] - \sum_{i=1}^m k_i [b_i, D_1] + \operatorname{ad}_{\hat{A}} D_1.
\end{aligned}$$

We now use the fact that $[b_i, D_1] \in \mathscr{D}_C^*$ and $\operatorname{ad}_{\hat{A}} D_1 \in \mathscr{D}_C^*$, and set $D_2 = -\sum_{i=1}^m k_i [b_i, D_1] + \operatorname{ad}_{\hat{A}} D_1 \in \mathscr{D}_C^*$. Equating the above expressions for $\operatorname{ad}_{\hat{A}} \operatorname{ad}_A X$ we see that

$$\operatorname{ad}_A^2 X = \sum_{i=1}^m \left\{ \operatorname{ad}_A X k_i + A a_{i1} + \sum_{j=1}^m a_{1j} b_j k_i \right\} b_i + \sum_{i=1}^m a_{1i} \operatorname{ad}_A b_i + D_2,$$

or

$$\text{ad}_A^2 X = \sum_{i=1}^{m} a_{2i} b_i + \sum_{i=1}^{m} a_{1i} \text{ad}_A b_i + D_2, \tag{14}$$

where $a_{2i} = \text{ad}_A Xk_i + Aa_{1i} + \sum_{j=1}^{m} a_{1j}(b_j k_i)$. This process can be repeated by applying $\text{ad}_{\hat{A}}$ to both sides of (14), and in this way one gets a sequence of equalities of the form

$$\begin{aligned}
&\text{ad}_A X = \sum_{i=1}^{m} a_{1i} b_i + D_1, \\
&\text{ad}_A^2 X = \sum_{i=1}^{m} a_{2i} b_i + \sum_{i=1}^{m} a_{1i} \text{ad}_A b_i + D_2, \\
&\qquad \vdots \qquad\qquad\qquad\qquad \vdots \\
&\text{ad}_A^l X = \sum_{i=1}^{m} a_{li} b_i + \sum_{i=1}^{m} \{a_{l-1,i} + Aa_{l-2,i} + \cdots + A^{l-2} a_{li}\} b_i \\
&\qquad + \sum_{i=1}^{m} \{a_{l-2,i} + Aa_{l-3,i} + \cdots + A^{l-3} a_{li}\} \text{ad}_A b_i + \cdots + \sum_{i=1}^{m} a_{1i} \text{ad}_A^{l-1} b_i + D_l, \\
&\qquad \vdots \qquad\qquad\qquad\qquad \vdots \ ;
\end{aligned} \tag{15}$$

where $D_1, D_2, \cdots \in \mathscr{D}_C^*$. Since $ZC \equiv 0$ for all $Z \in \mathscr{D}_C^*$, it follows that $\text{ad}_A XC = \sum_{i=1}^{m} a_{1i} b_i C + D_1 C = \sum_{i=1}^{m} a_{1i} b_i C$, and thus $(AX - XA)C = \sum_{i=1}^{m} a_{1i} b_i C$. We note that $XC \equiv 0$ as $X \in \mathscr{D}_C^*$, so that

$$XAC + \sum_{i=1}^{m} a_{1i} b_i C = 0 \quad \text{on } U.$$

Similarly $\text{ad}_A^2 XC = (A^2 X - 2AXA + XA^2)C = -2A(XAC) + XA^2 C$, and using (15), the above expression for XAC, and $\text{ad}_A b_i = -Ab_i$, we find that $XA^2 C + \sum_{i=1}^{m} (2Aa_{1i} - a_{2i}) b_i C + \sum_{i=1}^{m} a_{1i} A b_i C = 0$. If $B_1(x) = \sum_{i=1}^{m} a_{1i}(x) b_i$ and $B_2(x) = \sum_{i=1}^{m} (2Aa_{1i}(x) - a_{2i}(x)) b_i$, then

$$XAC(x) + CB_1(x) = CAX(x) + CB_1(x) = 0,$$

and $CA^2 X(x) + CAB_1(x) + CB^2(x) = 0$ on U. It is then straightforward to show that there exist vector fields $B_3, B_4, B_5, \cdots \in B$ such that for $l = 1, 2, \cdots$,

$$CA^l X(x) + CA^{l-1} B_1(x) + \cdots + CB_l(x) = 0 \quad \text{on } U.$$

Our task is to show that for each $x \in U$, $X(x) \in \mathscr{V}_C^*$. In [5] and [6] it is shown that $\mathscr{V}_C^* = \mathscr{V}^{(r)}$, where $r = \dim \ker C$ and $\mathscr{V}^{(0)} = \ker C$, $\mathscr{V}^{(i)} = \mathscr{V}^{(i-1)} \cap A^{-1}(\mathscr{B} + \mathscr{V}^{(i-1)})$. In other words

$$\begin{aligned}
\mathscr{V}_C^* = \{v \mid &Cv = 0, CAv + CB_1 = 0, CA^2 v + CAB_1 + CB_2 = 0, \\
&\cdots, CA^r v + CA^{r-1} B_1 + \cdots + CB_r = 0 \\
&\text{for some } B_1, \cdots, B_r \text{ in the span of } \{b_1, \cdots, b_m\}\};
\end{aligned}$$

hence $X(x) \in \mathscr{V}_C^*$ for all $x \in U$ and the proof is complete.

4. Examples

Example 1. This example shows that the dimension of an $(A, \mathscr{B})$-invariant distribution $\mathscr{D}$ need not be constant. We find $\mathscr{D}_C^*$, point out an obstruction to obtaining a global real analytic feedback law to solve the GDDP, and find a local feedback law to solve the LDDP.

Consider the system

$$\dot{x} = A(x) + uB(x), \qquad x \in R^3,$$
$$y = C(x),$$

where $A(x) = A(x_1, x_2, x_3) = (x_2, 0, 0)$, $B(x) = (x_1, 0, 0)$, and $C(x) = x_1$. Here $\ker C = \{0\} \times R^2$. The distribution

$$\mathscr{D}(x) = \begin{cases} \ker C & \text{if } x_1 \neq 0, \\ \{(0,0)\} \times R & \text{if } x_1 = 0, \end{cases}$$

is clearly real analytic ($\mathscr{D}(x) = \text{span}\{X(x), Y(x)\}$ for $X(x) = (0, 0, 1)$, $Y(x) = (0, x_1, 0)$) and involutive.

CLAIM. *$\mathscr{D}$ is $(A, \mathscr{B})$-invariant.*

To show this we must find an open dense submanifold M_0 of R^3 such that $\text{ad}_A \mathscr{D} \subseteq \mathscr{D} + \mathscr{B}$ and $\text{ad}_B \mathscr{D} \subseteq \mathscr{D}$ on M_0. Set $M_0 = R^3 \sim \{x | x_1 = 0\}$. Thus $\mathscr{B} = R \times \{(0,0)\}$ on M_0 and $(\mathscr{D} + \mathscr{B})(x) = R^3$; hence $\text{ad}_A \mathscr{D} \subseteq \mathscr{D} + \mathscr{B}$ on M_0. Now $\mathscr{D} = \text{span}\{X, Y\}$, $\text{ad}_B X(x) = 0$, and $\text{ad}_B Y(x) = (0, x_1, 0) \in \mathscr{D}(x)$, so that $\text{ad}_B \mathscr{D} \subseteq \mathscr{D}$ on M_0, which shows that $\mathscr{D}$ is $(A, \mathscr{B})$-invariant.

We remark that $\dim \mathscr{D}(x)$ is 2 if $x \in M_0$ and 1 if $x \notin M_0$. Since any real analytic involutive distribution with $\text{ad}_A \mathscr{D} \subseteq \mathscr{D} + \mathscr{B}$ and $\text{ad}_B \mathscr{D} \subseteq \mathscr{D}$ on some open dense submanifold is $(A, \mathscr{B})$-invariant, it follows that $x \mapsto \ker C = \{0\} \times R^2$ is an $(A, \mathscr{B})$-invariant distribution, and thus $\mathscr{D}_C^*(x) = \ker C$. From Theorem 3.1 we know that for the system

$$\dot{x} = A(x) + uB(x) + vD(x), \qquad x \in R^3,$$
$$y = C(x),$$

where A, B, C are as above, the LDDP is solvable if and only if $D \in \mathscr{D}_C^*$. In particular, if $D(x) = (0, 1, 0)$ the LDDP is solvable. To show that the GDDP cannot be solved we must show that there exists *no* $k \in C^w(R^3)$ such that

$$(*) \qquad \begin{aligned} \dot{x} &= [A(x) + k(x)B(x)] + uB(x) + vD(x), \\ y &= C(x) \end{aligned}$$

has the property that $y(\cdot, x_0, u, v) = y(\cdot, x_0, u, \tilde{v})$ for all admissible $u, v, \tilde{v} \in U$ and for all $x_0 \in R^3$. Let $\hat{A}(x) = A(x) + k(x)B(x) = (x_2 + k(x)x_1, 0, 0)$, and let $\mathscr{L}_0$ be the Lie algebra of vector fields generated by $\{D, \text{ad}_{\hat{A}} D, \cdots\}$. Then $x \to \mathscr{L}_0(x)$ defines a real analytic involutive distribution on R^3, and from the controllability results of Sussmann and Jurdjevic [3] we know that the subset $\mathscr{R}_t(x_0) = \{x(t, x_0, 0, v) | v \in U\}$ has a nonempty interior in the integral manifold $I(\mathscr{L}_0, \hat{A}_t \cdot x_0)$. Suppose there exists $k \in C^w(R^3)$ such that $y(t, x_0, 0, v) = y(t, x_0, 0, 0)$ for all $v \in U$ or $C(x(t, x_0, 0, v)) = C(\hat{A}_t \cdot x_0)$. Then

$$\mathscr{C}(\mathscr{R}_t(x_0)) = C(\hat{A}_t \cdot x_0),$$

and hence

$$C(I(\mathscr{L}_0, \hat{A}_t \cdot x_0)) = C(\hat{A}_t \cdot x_0).$$

Thus if $X \in \mathscr{L}_0$,

$$C(X_s \circ \hat{A}_t \cdot x_0) = C(\hat{A}_t \cdot x_0) \quad \text{for all } s,$$

and differentiating with respect to s and setting $s = 0$, we find that

$$XC(\hat{A}_t \cdot x_0) = 0.$$

Since x_0 is arbitrary, $XC = 0$ for all $X \in \mathscr{L}_0$. In particular, $\operatorname{ad}_{\hat{A}} DC = 0$ on M. Now

$$\operatorname{ad}_{\hat{A}} D(x) = dD_x\hat{A}(x) - d\hat{A}_x D(x)$$

$$= \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_2 + k(x)x_1 \\ 0 \\ 0 \end{bmatrix} - \begin{bmatrix} \dfrac{\partial k}{\partial x_1} \cdot x_1 + k(x) & 1 + \dfrac{\partial k}{\partial x_2} x_1 & \dfrac{\partial k}{\partial x_3} x_1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} -1 - \dfrac{\partial k}{\partial x_2} \cdot x_1 \\ 0 \\ 0 \end{bmatrix},$$

and so

$$\operatorname{ad}_{\hat{A}} DC(x) = dC_x \operatorname{ad}_{\hat{A}} D(x) = [1 \quad 0 \quad 0] \begin{bmatrix} -1 - \dfrac{\partial k}{\partial x_2} \cdot x_1 \\ 0 \\ 0 \end{bmatrix} = -1 - \frac{\partial k}{\partial x_2} x_1.$$

The condition $\operatorname{ad}_{\hat{A}} DC = 0$ comes down to

$$\frac{\partial k}{\partial x_2} = -\frac{1}{x_1},$$

or $k(x) = -(x_2/x_1) + \alpha$ for some constant α. Since k is not defined for all $x \in R^3$ (in particular, when $x_1 = 0$, $x_2 \neq 0$) there can be *no* solution to the GDDP. On the other hand, if we take $M_0 = R^3 \sim \{x | x_1 = 0\}$ then $k(x) = -x_2/x_1$ is defined on M_0, and it is now easy to show that on M_0 the system

$$\dot{x} = [A(x) + k(x)B(x)] + uB(x) + vD(x), \qquad x \in M_0,$$
$$y = C(x)$$

is disturbance decoupled with respect to v and y. Here $A(x) + k(x)B(x) = (x_2, 0, 0) + (-x_2/x_1)(x_1, 0, 0) = (0, 0, 0)$, and the state equation is $\dot{x} = (ux_1, v, 0)$. If $x_0 = (A_1, A_2, A_3) \in M_0$, then $A_1 \neq 0$ and $x_3(t) = A_3$, $x_2(t) = \int_0^t v(\tau)\, d\tau + A_2$, $x_1 = A \exp\left(\int_0^t u(\tau)\, d\tau\right)$, and $y(t, x_0, u, v) = x_1(t) = A_1 \exp\left(\int_0^t u(\tau)\, d\tau\right)$, which is independent of v as required.

Example 2. In this example a system for which the LDDP is *not* solvable is exhibited. Consider the system

$$\dot{x} = A(x) + u_1 B_1(x) + u_2 B_2(x), \qquad x \in M,$$
$$y = C(x),$$

where $M = \{(x_1, x_2, x_3) \in R^3 | x_1 \neq 0\}$, $A(x) = (0, x_1 x_2, x_2)$, $B_1(x) = (0, x_1, 0)$, $B_2(x) = (x_1 e^{x_2}, 0, -x_3 e^{x_2})$, and $C(x) = (x_2, x_1 x_2 x_3)$, a map from M into R^2. We now show that $\mathscr{D}_C^*$ is the trivial distribution $x \to \{0\} \subseteq T_x(M)$; hence by Theorem 3.1 no disturbance can be locally decoupled from the output.

Suppose $X \in \mathscr{D}_C^*$. Our task is to show that X must be the zero vector field. By definition $XC = 0$ and $\operatorname{ad}_A X \in \mathscr{D}_C^* + \mathscr{B}$ on an open dense submanifold M_0 of M. In particular,

$$\operatorname{ad}_A X = Y + Z,$$

where $YC = 0$ ($Y \in \mathcal{D}_C^*$, hence $YC = 0$) and $Z \in \mathcal{B}$. Suppose we write X as $X(x) = (e(x), f(x), g(x))$. Then $XC = 0$ becomes

$$dC_x X(x) = \begin{bmatrix} 0 & 1 & 0 \\ x_2x_3 & x_1x_3 & x_1x_2 \end{bmatrix} \begin{bmatrix} e(x) \\ f(x) \\ g(x) \end{bmatrix}$$

$$= \begin{bmatrix} f(x) \\ x_2x_3e(x) + x_1x_3f(x) + x_1x_2g(x) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

so that $f(x) = 0$ and $x_3e(x) = -x_1g(x)$ if $x_2 \neq 0$. Thus restricting X to $M_1 = \{x \in M | x_2 \neq 0\}$, we know that $X(x) = (e(x), 0, -(x_3/x_1)e(x))$ for some $e \in C^w(M_1)$, and thus

$$\mathrm{ad}_A X(x) = \Big(x_1x_2 \frac{\partial e}{\partial x_2}(x) + x_2 \frac{\partial e}{\partial x_3}(x), -x_2e(x), -x_3x_2 \frac{\partial e}{\partial x_2}(x) - \frac{x_2}{x_1}e(x) - \frac{x_3x_2}{x_1} \frac{\partial e}{\partial x_3}(x) \Big).$$

Now $\mathrm{ad}_A X = Y + Z$ on M_1, where $YC = 0$ and $Z \in \mathcal{B}$. Thus $Z(x) = \alpha(x)B_1(x) + \beta(x)B_2(x)$ for some $\alpha, \beta \in C^w(M_1)$, and $Y(x) = (h(x), 0, -(x_3/x_1)h(x))$ for some $h \in C^w(M_1)$ from the calculations completed above. In vector notation, $\mathrm{ad}_A X = Z + Y$ becomes

$$\begin{bmatrix} x_1x_2 \dfrac{\partial e}{\partial x_2} + x_2 \dfrac{\partial e}{\partial x_3} \\ -x_2e(x) \\ -x_3x_2 \dfrac{\partial e}{\partial x_2} - \dfrac{x_2}{x_1}e(x) - \dfrac{x_3x_2}{x_1} \dfrac{\partial e}{\partial x_3} \end{bmatrix}$$

$$= \begin{bmatrix} 0 \\ \alpha(x)x_1 \\ 0 \end{bmatrix} + \begin{bmatrix} \beta(x)x_1e^{x_2} \\ 0 \\ -\beta(x)x_3e^{x_2} \end{bmatrix} + \begin{bmatrix} h(x) \\ 0 \\ -(x_3/x_1)h(x) \end{bmatrix}.$$

We note that the vector field on the right-hand side has the property that $(x_3) \times$ the first component $= (-x_1) \times$ the third component, and thus $\mathrm{ad}_A X$ has this property. In particular,

$$x_3x_1x_2 \frac{\partial e}{\partial x_2} + x_3x_2 \frac{\partial e}{\partial x_3} = x_1x_2x_3 \frac{\partial e}{\partial x_2} + x_2e(x) + x_2x_3 \frac{\partial e}{\partial x_3},$$

and so on M_1, $x_2e(x) = 0$, which implies that $e(x) = 0$ on M_1, and by real analyticity $e = 0$ on M. Thus

$$X(x) = (0, 0, 0), \quad \text{as required.}$$

Acknowledgment. The author wishes to thank the anonymous referee for his contributions to the proof of Theorem 3.3.

REFERENCES

[1] R. W. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.
[2] R. NARASIMHAN, *Analysis on Real and Complex Manifolds*, North-Holland, Amsterdam, 1968.
[3] H. SUSSMANN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), 95–116.

[4] F. WARNER, *Foundations of Differentiable Manifolds and Lie Groups*, Scott, Foresman and Co., Glenview IL, 1970.

[5] W. M. WONHAM, *Linear Multivariable Control, a Geometric Approach*, Lecture Notes in Economics and Mathematical Systems, vol. 101, Springer-Verlag, New York, 1974.

[6] W. M. WONHAM AND A. S. MORSE, *Decoupling and Pole Assignment in Linear Multivariable Systems: A Geometric Approach*, this Journal, 8 (1970), pp. 1–18.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 1, January 1981

0363-0129/81/1901-0002 $01.00/0

DETECTABILITY AND STABILIZABILITY OF TIME-VARYING DISCRETE-TIME LINEAR SYSTEMS*

B. D. O. ANDERSON† AND J. B. MOORE†

Abstract. The concepts of detectability and stabilizability are explored for time-varying systems. We study duality, invariance under feedback, an extended version of the lemma of Lyapunov, existence of stabilizing feedback laws, linear quadratic filtering and control, and the existence of approximate canonical forms.

1. Introduction. The dual concepts of observability and reachability for linear finite-dimensional time-invariant systems have found application in a variety of filtering and control problems. For example, signal model observability is a sufficient condition for an optimal minimum variance filter (Kalman filter) to exist, and signal model controllability ensures its asymptotic stability—at least in the case of linear time-invariant continuous time, finite dimensional signal models. See [1] for a leisurely exposition. Dual results hold for the linear-quadratic optimal control problem. Moreover, there have been at least two significant generalizations of these results. First, retaining the time-invariance assumption, the reachability and observability hypotheses have been weakened to stabilizability and detectability, and the results are still valid, [2], [3]. Second, time-variable problems have been considered, and with the imposition of a uniformity constraint in the now time-varying reachability and observability hypotheses, the results extend to the time-varying situation [4]. For time-varying systems one is tempted to avoid the weaker conditions of detectability and stabilizability since the technical issues raised involve nontrivial generalizations of the time-invariant results and on first glance appear formidable. However, the desirability of extending known control and filtering results to important classes of time-varying filtering and control problems is clear.

A lead has been taken in [5] with the introduction of definitions of detectability and stabilizability for discrete-time, time-varying linear systems and some applications to control and filtering problems. However an exploration of equivalent definitions and properties such as duality and invariance under feedback is not attempted in [5].

In this paper, motivated by the lead given in [5], we explore a generalization of the dual concepts of observability and controllability to linear, time-varying, discrete-time, finite-dimensional systems.

Following definitions of the concepts of detectability and stabilizability in § 2, we indicate their formal duality, and establish two simple consequences. In § 3, we show the invariance of the properties under appropriate feedback, and in § 4, we prove a significant generalization of the lemma of Lyapunov, which is useful for studying the stability of linear systems using quadratic Lyapunov functions. Section 5 considers linear quadratic problems, and we prove one of the main results of the paper: detectability and stabilizability are the key properties required to guarantee an exponentially stable Kalman filter. We also show the equivalence of the definitions of § 2 with the existence of stabilizing feedback laws of an appropriate form. Section 6 contains the concuding remarks.

Two general points can be noted. First, almost all the results are stated just for detectable pairs, rather than detectable and stabilizable pairs. Given the duality

* Received by the editors August 24, 1979, and in revised form April 16, 1980. This work was supported by the Australian Research Grants Committee.

† Department of Electrical Engineering, University of Newcastle, New South Wales, 2308, Australia.

established in the paper, there is no loss of generality. Second, the results are all stated for discrete-time systems. We elected to work with discrete-time systems rather than continuous-time systems because we were aware that, as illustrated by ideas in [6]–[9], it is often harder to get the discrete-time result than the continuous-time result. Particularly is this so when the discrete-time transition matrix can be singular. Of course, for the continuous-time case, regularity conditions must be imposed and techniques as in, for example, [12] exploited. In some cases, the continuous-time proofs are likely to be harder.

2. Detectability definitions and some implications. Consider the linear finite-dimensional state space system in discrete time

$$x_{k+1} = F_k x_k + G_k u_k, \tag{2.1a}$$

$$y_k = H_k' x_k, \tag{2.1b}$$

where x_k is the n-vector state, u_k is an input m-vector, y_k is an output p-vector, and F_k, G_k, H_k are matrices of appropriate dimension. The state transition matrix is denoted $\phi_{k,l}$ for $k \geqq l$ where $\phi_{k+1,k} = F_k$ and $\phi_{k,l} = \phi_{k,k-1}\phi_{k-1,l}$.

The detectability definition we work with is a specialization of one in [5] for finite-dimensional systems.

DEFINITION 2.1. The pair $[F_k, H_k]$ is *uniformly detectable* if there exist integers s, $t \geqq 0$ and constants d, b with $0 \leqq d < 1$, $0 < b < \infty$, such that whenever

$$\|\phi_{k+t,k}\xi\| \geqq d\|\xi\| \tag{2.2}$$

for some ξ and k, then

$$\xi' M_{k+s,k}\xi \geqq b\xi'\xi, \tag{2.3}$$

where

$$M_{k+s,k} = \sum_{i=k}^{k+s} \phi_{i,k}' H_i H_i' \phi_{i,k}. \tag{2.4}$$

Remarks. 1. For time invariant systems, detectability definitions have been given (see [3]) which require that the unstable modes of a system be observable. The above definition is a time-varying version of this notion. In fact, the definition says roughly that when a state trajectory is not fast decaying, i.e., (2.2) is satisfied, then that trajectory must be observable, i.e., (2.3) holds. Conversely, trajectories which are not observed with much output energy, i.e., those for which (2.3) fails, must be trajectories which decay, i.e., (2.2) fails. Further justification of this remark is provided by Lemma 2.2.

2. Recall (see, e.g., [4]), that if $[F_k, H_k]$ is uniform with respect to observability, the observability Gramian $M_{k+s,k}$ satisfies (for some integer s, and constants β_1, β_2)

$$0 < \beta_1 I \leqq M_{k+s,k} \leqq \beta_2 I.$$

This is clearly a sufficient condition for detectability as above. Notice, however that there are no upper bounds in the detectability definition. It is not surprising therefore that in most of the results to follow we impose upper bounds in F_k and H_k.

3. Without loss of generality, $s \geqq t$ can be assumed in the above definition, since if $s < t$, s can be replaced by t. Henceforth, we shall assume that $s \geqq t$.

The second definition we give is that of uniform stabilizability. As argued following the definition, the definition is related to that of detectability via a certain duality.

DEFINITION 2.2. The pair $[\hat{F}_k, \hat{G}_k]$ is *uniformly stabilizable* if there exist integers s, $t \geqq 0$ and constants d, b with $0 \leqq d < 1$, $0 < b < \infty$, such that whenever

$$\|\hat{\phi}_{k+1,k+1-t}\xi\| \geqq d\|\xi\| \tag{2.5}$$

for some ξ, k, then

$$\xi' \hat{Y}_{k,k-s}\xi \geqq b\xi'\xi, \tag{2.6}$$

where $\hat{\phi}_{k,l}$ is the transition matrix associated with $\hat{F}_k$ and

$$\hat{Y}_{k,k-s} = \sum_{i=k-s}^{k} \hat{\phi}_{k+1,i+1}\hat{G}_i\hat{G}_i'\hat{\phi}_{k+1,i+1}'. \tag{2.7}$$

Without loss of generality, $s \geqq t$ can be assumed.

Remark. In the time-invariant case, stabilizability is equivalent to the requirement that any uncontrollable mode be asymptotically stable (see [3]). But as the name suggests, stabilizability is also equivalent (but this must be proved) to the property that there exists a stabilizing state feedback law. The first idea is reflected in the definition above. The second will be taken up later.

The duality between detectability and stabilizability is taken up in the following lemma.

LEMMA 2.1. *Make the definitions*

$$F_k = \hat{F}'_{-k}, \qquad H_k = \hat{G}_{-k}. \tag{2.8}$$

Then

(a) $\phi_{i,k} = \hat{\phi}'_{-k+1,-i+1}$ *and* $M_{k+s,k} = Y_{-k,-k-s}$.

(b) $[\hat{F}_k, \hat{G}_k]$ *is uniformly stabilizable if and only if* $[F_k, H_k]$ *is uniformly detectable.*

(c) $x_{k+1} = F_k x_k$ *is exponentially stable if and only if* $\hat{x}_{k+1} = \hat{F}_k\hat{x}$ *is exponentially stable.*

Proof is via direct calculation. Notice that use of a dual relationship of the form $F_k = (\hat{F}'_k)^{-1}$ is not suitable, requiring as it does the existence of the inverse. Use of this second, rather unsatisfactory dual, appears to be behind many ideas of [8], [9].

We conclude the section by noting two simple consequences of the detectability definitions. The first confirms the first remark following the definition. The second will be used in a later section.

LEMMA 2.2. *With* $[F_k, H_k]$ *detectable and* F_k *bounded above, then for the system* (2.1) *with a zero input,*

$$H_k'x_k \to 0 \text{ as } k \to \infty \Rightarrow \|x_k\| \to 0 \text{ as } k \to \infty.$$

Proof (by contradiction). Assume there exists an x_0 with $H_k'x_k \to 0$ as $k \to \infty$, but with $\|x_k\| \not\to 0$ as $k \to \infty$. Now if for all $k \geqq k_1$ for some k_1, $\|\phi_{k+t,k}x_k\| < d\|x_k\|$, then $\|x_k\| \to 0$ as $k \to \infty$. Thus there exists a sequence $k_i \to \infty$ with $\|\phi_{k_i+t,k_i}x_{k_i}\| \geqq d\|x_{k_i}\|$ defining the k_i. We can further assume that $\|x_{k_i}\| \not\to 0$, for if $\|x_{k_i}\| \to 0$, as $k_1 \to \infty$, then as we now show, $\|x_k\| \to 0$ as $k \to \infty$ which is a contradiction.

Let $k \in (k_i, k_{i+1})$ but be otherwise arbitrary, and set $k = k_i + t\alpha + \beta$, with α, β integers with $\beta < t$. Then

$$x_k = \phi_{k,k_i+t\alpha}\phi_{k_i+t\alpha,k_i+t(\alpha-1)} \cdots \phi_{k_i+2t,k_i+t}x_{k_i+t}.$$

The first matrix in the product has bounded norm because F_k is bounded and $k - (k_i + t\alpha)$ is bounded.

From the definition then of the k_i,

$$\begin{aligned}\|x_k\| &\leqq \gamma_1 d^{\alpha-1}\|x_{k_i+t}\| \\ &\leqq \gamma_1\gamma_2 d^{\alpha-1}\|x_{k_i}\|,\end{aligned}$$

with γ_2 existing because of the bound on F_k. It readily follows that $\|x_{k_i}\| \to 0$ implies $\|x_k\| \to 0$ for all k. So we return to the assumption that $\|x_{k_i}\| \not\to 0$.

Define a subsequence $\{l_i\}$ of the $\{k_i\}$ such that $\|x_{l_i}\| > \gamma_3$ for all i and some $\gamma_3 > 0$. Now

$$\|\phi_{l_i+t,l_i}x_{l_i}\| \geqq d\|x_{l_i}\|,$$

and by detectability

$$\|x'_{l_i}M_{l_i+s,l_i}x_{l_i}\| \geqq b\|x_{l_i}\|^2 > b\gamma_3^2.$$

However, from the definitions of M_{l_i+s,l_i} and assumptions on H'_kx_k and u_k,

$$\|x'_{l_i}M_{l_i+s,l_i}x_{l_i}\| = \sum_{l=l_i}^{l_i+s}\|H'_lx_l\|^2 \to 0 \quad \text{as } l_i \to \infty$$

So we have a contradiction and the lemma is established. □

LEMMA 2.3. *Let $[F_k, H_k]$ be detectable. Then $[\rho F_k, H_k]$ is detectable for $1+\varepsilon \geqq \rho > 1$ and $\varepsilon > 0$ sufficiently small.*

Proof. Let a tilde denote quantities associated with $\tilde{F}_k = \rho F_k$ and H_k. Now

$$\|\tilde{\phi}_{k+t,k}\xi\| = \rho^t\|\phi_{k+t,k}\xi\|.$$

Now choose $\rho > 1$ such that $\tilde{d} = d\rho^t < 1$, where d appears in the detectability definitions for $[F_k, H_k]$. Then $\|\tilde{\phi}_{k+t,k}\xi\| \geqq \tilde{d}\|\xi\|$ if and only if $\|\phi_{k+t,k}\xi\| \geqq d\|\xi\|$. Also

$$\tilde{M}_{k+s,k} = \sum_{i=k}^{k+s}\rho^{2(i-k)}\phi'_{i,k}H_iH'_i\phi_{i,k} \geqq M_{k+s,k}.$$

Consequently, if $\xi'M_{k+s,k}\xi \geqq b\xi'\xi$, then also $\xi'\tilde{M}_{k+s,k}\xi \geqq b\xi'\xi$. Thus tying the above results together we have that whenever $\|\tilde{\phi}_{k+t,k}\xi\| \geqq \tilde{d}\|\xi\|$, then $\xi'\tilde{M}_{k+s,k}\xi \geqq \xi'\xi$ as required by the detectability definition. This establishes the lemma. □

Remarks. 1. As the proof shows, ρ can be taken as any number for which $\rho^t d < 1$, d being the quantity appearing in the uniform detectability definition associated with $[F_k, H_k]$.

2. The above lemma is a special case of a more general result which is almost as easily proved: if $[F_k, H_k]$ is detectable, there exists ε, depending on $[F_k, H_k]$, such that $[\tilde{F}_k, \tilde{H}_k]$ is detectable for all $\tilde{F}_k$, $\tilde{H}_k$ with $\|F_k - \tilde{F}_k\| < \varepsilon$, $\|H_k - \tilde{H}_k\| < \varepsilon$.

3. Invariance under feedback. With $[F_k, H_k]$ denoting an open-loop system matrix pair, there is interest in a closed-loop system matrix pair $[\bar{F}_k, H_k]$ where $\bar{F}_k = F_k - K_kH'_k$.

LEMMA 3.1. *The observability Gramians $M_{l,k}$ for the open-loop pair $[F_k, H_k]$ and $\bar{M}_{l,k}$ for the closed-loop pair $[\bar{F}_k, H_k]$, where $\bar{F}_k = F_k - K_kH'_k$, bear the following relationship:*

$$\bar{M}_{l,k} = H_{l,k}C'_kC_kH'_{l,k}, \qquad M_{l,k} = H_{l,k}H'_{l,k}, \tag{3.1}$$

where

$$H_{l,k} = [H_k \quad \phi'_{k+1,k}H_{k+1} \quad \cdots \quad \phi'_{l,k}H_l], \tag{3.2}$$

$$C_k = \begin{bmatrix} I & 0 & \cdots & 0 \\ * & I & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ * & * & \cdots & I \end{bmatrix} \tag{3.3}$$

and $*$ *denotes terms involving* F_i, K_i, H_i, *for* $i = k, k+1 \cdots l$. *Moreover, with* F_i, K_i, H_i *bounded, then for some positive constants* α_1, α_2,

$$\alpha_1 M_{l,k} \leqq \bar{M}_{l,k} \leqq \alpha_2 M_{l,k}. \tag{3.4}$$

Proof. The relationship (3.1) follows by inductive arguments, the definitions of $M_{l,k}$, and straightforward manipulations. The bounds (3.4) follow from a premultiplication by $H_{l,k}$ and postmultiplication by $H'_{l,k}$ of the inequalities,

$$0 < \alpha_1 I \leqq \lambda_{\min}(C'_k C_k) I \leqq C'_k C_k \leqq \lambda_{\max}(C'_k C_k) I \leqq \alpha_2 I < \infty.$$

The above bounds are verified as follows. First, $\lambda_{\max}(C'_k C_k) \leqq \operatorname{tr}(C'_k C_k) < \alpha_2 I$ for some $\alpha_2 < \infty$ under the boundedness assumptions, and

$$\lambda_{\min}(C'_k C_k) \geqq \frac{|C'_k C_k|}{[\lambda_{\max}(C'_k C_k)]^{n-1}} \geqq \frac{|C'_k \| C_k|}{\alpha_2^{n-1}} = \frac{1}{\alpha_2^{n-1}} = \alpha_1,$$

for some $\alpha_1 > 0$. □

LEMMA 3.2. *In the notation of the previous lemma, and with* $\bar{\phi}_{l,k}$ *the transition matrix associated with the closed loop system matrix* $\bar{F}_k = F_k - K_k H'_k$, *then for all* $l \geqq k$,

$$\bar{\phi}_{l,k} = \phi_{l,k} + [*] H'_{l,k}, \tag{3.5}$$

where $[*]$ *denotes a matrix involving* F_i, K_i, H_i *for* $i = k, k+1, \cdots l-1$.

Proof. From straightforward manipulations. □

We now have the following main result.

THEOREM 3.3. *With* F_k, H_k, *and* K_k *bounded, and with* $F_k = F_k - K_k H'_k$, *then* $[F_k, H_k]$ *is uniformly detectable if and only if* $[\bar{F}_k, H_k]$ *is uniformly detectable.*

Proof. It clearly suffices to prove that under the boundedness conditions, $[F_k, H_k]$ uniformly detectable implies $[F_k - K_k H'_k, H_k]$ uniformly detectable. Let s, t, d, b be as in the uniform detectability definition applied to $[F_k, H_k]$.

Under the boundedness assumptions, Lemma 3.2 yields

$$\begin{aligned} \|\bar{\phi}_{k+t,k}\xi\| &= \|\phi_{k+t,k}\xi + [*] H'_{k+t,k}\xi\| \\ &\leqq \|\phi_{k+t,k}\xi\| + \alpha_3 \|H'_{k+t,k}\xi\|, \end{aligned}$$

for some α_3. Let α_1 be as in Lemma 3.1, with $l = k+s$, and define

$$\bar{b} = \min\left\{\alpha_1 b, \alpha_1\left(\frac{1-d}{2\alpha_3}\right)^2\right\}, \qquad \bar{d} = d + \alpha_3\sqrt{\frac{\bar{b}}{\alpha_1}}.$$

Notice that $1 - \bar{d} > 0$. We shall show that $s, t, \bar{d}, \bar{b}$ characterize the uniform detectability of $[\bar{F}_k, H_k]$.

Suppose that $\xi' \bar{M}_{k+s,k}\xi < \bar{b}\xi'\xi$. Using (3.4), we have

$$\xi' M_{k+s,k}\xi < \frac{\bar{b}}{\alpha_1}\xi'\xi \leqq b\xi'\xi,$$

and also

$$\|H'_{k+t,k}\xi\| = (\xi' M_{k+t,k}\xi)^{1/2} \leqq (\xi' M_{k+s,k}\xi)^{1/2} < \sqrt{\frac{\bar{b}}{\alpha_1}}\|\xi\|.$$

Consequently,

$$\|\bar{\phi}_{k+t,k}\xi\| < d\|\xi\| + \alpha_3\sqrt{\frac{\bar{b}}{\alpha_1}}\|\xi\|,$$

or

$$\|\bar{\phi}_{k+t,k}\xi\| < \bar{d}\|\xi\|.$$

Equivalently, $[\bar{F}_k, H_k]$ is uniformly detectable. □

There are two useful corollaries to this theorem.

COROLLARY 3.4. *A sufficient condition for the pair* $[F_k, H_k]$ *to be uniformly detectable is that there exist a bounded gain* K_k *such that the closed-loop system* $\bar{x}_{k+1} = (F_k - K_k H'_k)\bar{x}_k$ *is exponentially stable.*

Proof. If $\bar{F}_k$ defines an exponentially stable system, $[\bar{F}_k, H_k]$ is uniformly detectable for any H_k. □

Later in the paper, we shall show that the sufficiency condition just stated is in fact also a necessity condition.

COROLLARY 3.5. *With notation as above, the following quantity is feedback invariant*:

$$\tilde{d} = \inf [d | d \in [0, 1) \text{ and } \|\phi_{k+t,k}\xi\| \geqq d\|\xi\| \text{ implies } \|\xi' M_{k+s,k}\xi\| \geqq b\|\xi\|^2 \text{ for some } b > 0].$$

Proof. Let K_k be a gain sequence, and let α_1, α_2 be constants defined in the statement of Lemma 3.1 and the proof of Lemma 3.3. Take $\varepsilon > 0$, arbitrary save that $\tilde{d} + \varepsilon < 1$. Then for $d = \tilde{d} + \varepsilon/2$, there exists $b > 0$ such that $\|\phi_{k+t,k}\xi\| \geqq d\|\xi\|$ implies $\|\xi' M_{k+s,k}\xi\| \geqq b\|\xi\|^2$.

Without loss of generality, we may replace b by $\min(b, \varepsilon^2/(4\alpha_3^2\alpha_1))$. Then with $\bar{d}$ referring to the uniform detectability definition applied to $[F_k - K_k H'_k, H_k]$, the proof of the theorem shows that we can take $\bar{d}$ with

$$\begin{aligned}
\bar{d} &= d + \frac{\alpha_3}{\sqrt{\alpha_1}}\left[\min\left\{\alpha_1 b, \alpha_1\left(\frac{1-d}{2\alpha_3}\right)^2\right\}\right]^{1/2} \\
&\leqq d + \frac{\alpha_3}{\sqrt{\alpha_1}}\sqrt{\alpha_1 b} \\
&\leqq \tilde{d} + \varepsilon/2 + \frac{\alpha_3}{\sqrt{\alpha_1}}\frac{\varepsilon}{2\alpha_3} \\
&= \tilde{d} + \varepsilon.
\end{aligned}$$

Consequently, $\inf \bar{d} \leqq \inf d = \tilde{d}$. But also we can argue that $\inf d \leqq \inf \bar{d}$, whence the feedback invariance of $\tilde{d}$. □

In case $\tilde{d} = 0$, the result simply says that uniform observability is invariant under feedback. For $\tilde{d} > 0$, the quantity has the following interpretation. Consider trajectories which are exactly, or approximately unobserved. Then they all decay at least as fast as $(\tilde{d})^{1/t}$ but they do not all decay faster. Feedback does not vary the conclusion, precisely because the trajectories are unobserved.

Finally in this section, we note that there are obvious duals of these results tied to uniform stabilizability.

4. Lemma of Lyapunov. In this section, we attempt to parallel a result relevant to estabishing stability. The continuous-time lemma of Lyapunov [10] is concerned with the matrix equation $PA + A'P = -Q$ with $Q > 0$, linking positive definiteness of P with asymptotic stability of $\dot{x} = Ax$; relaxation of $Q > 0$ is accomplished in [11], and a time-varying version using uniform complete observability ideas can be found in [12] for continuous time. A discrete-time statement applicable to the time-invariant case can be found in, e.g., [13], while a time-varying version parallel to [12] is easy to find. Here, we aim to relax the observability assumption to detectability.

The following result is comparatively straightforward to obtain. We state it as a lead-in to the more difficult time-varying result.

PROPOSITION 4.1. *Let $[F, H]$ be a detectable pair of constant matrices. Then the equation $P - F'PF = HH'$ has a unique solution $P = P' \geqq 0$ if and only if $x_{k+1} = Fx_k$ is asymptotically stable.*

Proof. We use the characterization of detectability that $Fw = \lambda w$, $H'w = 0$ and $w \neq 0$ only if $|\lambda| < 1$, ([13]). Suppose there exists $P = P' \geqq 0$. Let $Fx = \lambda x$, $x \neq 0$. Then $0 \leqq x'^* HH'x = (1 - |\lambda|^2)x'^* Px$. So $|\lambda| < 1$, or $|\lambda| \geqq 1$ and also $H'x = 0$, contradicting detectability. Conversely, if $x_{k+1} = Fx_k$ is asymptotically stable,

$$P = \sum_{i=0}^{\infty} (F')^i HH' F^i$$

is well-defined, symmetric, and nonnegative definite, and satisfies the equation $P - F'PF = HH'$. Uniqueness is easily obtained. □

To obtain a time-varying generalization, consider the sequence $\Pi_{k,N}$ defined for $k = N, N-1, N-2, \cdots$, by

$$F_k' \Pi_{k+1,N} F_k - \Pi_{k,N} = -H_k H_k' \Pi_{N,N} = 0. \tag{4.1}$$

Evidently, for $k \leqq N-1$,

$$\Pi_{k,N} = \sum_{j=k}^{N-1} \phi_{j,k}' H_j H_j' \phi_{j,k}, \tag{4.2}$$

where $\phi_{j,k}$ has the usual association with F_k. We now have a parallel to one half of Proposition 4.1.

LEMMA 4.2. *With notation as above, suppose that $x_{k+1} = F_k x_k$ is exponentially stable and F_k, H_k are bounded. Then*

$$P_k = \lim_{N \to \infty} \Pi_{k,N} \tag{4.3}$$

exists as a bounded nonnegative definite symmetric matrix, it satisfies

$$F_k' P_{k+1} F_k - P_k = -H_k H_k', \tag{4.4}$$

for all $k \geqq 0$, and $\{P_k\}$ is the unique bounded sequence to do so.

Proof. All claims are clear, except perhaps for the last. Let $\{Q_k\}$ be a second bounded sequence satisfying (4.4). Set $R_k = P_k - Q_k$. Then

$$F_k' R_{k+1} F_k - R_k = 0,$$

whence

$$\phi_{k+l,k}' R_{k+l} \phi_{k+l,k} - R_k = 0.$$

Letting $l \to \infty$, and using the exponential decay of $\phi_{k+l,k}$ and boundedness of R_{k+l}, gives $R_k = 0$. □

We now seek the converse to this result; i.e., we seek to establish exponential stability, given (4.4). One might think that $V(x_k, k) = x_k' P_k x_k$ could serve as a Lyapunov function for $x_{k+1} = F_k x_k$. After all (4.4) would then imply $V(x_{k+1}, k) - V(x_k, k) \leqq 0$. One difficulty is that $V(x_k, k)$ is not necessarily positive definite, and in fact it is easy to construct examples where it fails to be positive definite; another difficulty is that the monotone decreasing property of V along trajectories is not strict. Nevertheless we have the following result:

THEOREM 4.2 (Extended lemma of Lyapunov). *Suppose that $[F_k, H_k]$ is uniformly detectable, that F_k and H_k are bounded, that there is a bounded nonnegative definite symmetric matrix sequence P_k satisfying* (4.4) *on $[k_0, \infty)$. Then $x_{k+1} = F_k x_k$ is exponentially stable.*

Proof. We shall associate with $x_{k+1} = F_k x_k$ a "Lyapunov-like" function,

$$V(x_k, k) = x_k'(P_k + \varepsilon I)x_k, \tag{4.5}$$

for some ε still to be determined. While V may not decrease at every step, we shall show that over a larger number of steps than 1, it must strictly decrease.

Setting $V_k = V(x_k, k)$, we observe that

$$V_k - V_{k+1} = x_k' H_k H_k' x_k + \varepsilon x_k'(I - F_k' F_x)x_k.$$

Two cases arise. Let d, b be the quantities of the uniform detectability definition.

Case 1. If $\|\phi_{k+t,k} x_k\| \geqq d\|x_k\|$, then under the detectability assumption $x_x' M_{k+s,k} x_k x_k \geqq b x_k' x_k$, so that

$$\begin{aligned} V_k - V_{k+s+1} &= x_k'(M_{k+s,k} + \varepsilon)x_k - \varepsilon x_k' \phi_{k+s+1;k}' \phi_{k+s+1,k} x_k \\ &\geqq [b + \varepsilon(1-\gamma)] x_k' x_k, \end{aligned}$$

where γ is an upper bound on $\|\phi_{k+s+1,k}' \phi_{k+s+1,k}\|$, which exists by virtue of the assumption of the boundedness of F_k.

Case 2. If $\|\phi_{k+t,k} x_k\| < d\|x_k\|$, then

$$\begin{aligned} V_k - V_{k+t} &= x_k'(M_{k+t-1,k} + \varepsilon)x_k - \varepsilon x_k' \phi_{k+t,k}' \phi_{k+t,k} x_k \\ &\geqq \varepsilon(1 - d^2) x_k' x_k. \end{aligned}$$

Hence if ε is sufficiently small, there exists $\eta > 0$ such that

$$\max\{V_k - V_{k+t},\ V_k - V_{k+s+1}\} \geqq \eta\|x_k\|^2 \geqq \delta V_k,$$

where the existence of δ follows from the bound on P_k.

This inequality shows that there is a subsequence $\{V_{k_i}\}$ of $\{V_k\}$, depending on x_0 and with ith member V_{k_i} where $k_i \leqq i(s+1)$ irrespective of x_0, such that the subsequence decays exponentially fast; i.e.,

$$V_{k_i} \leqq a b^i,$$

for some $a > 0$, $0 < b < 1$. Then because $\varepsilon x_k' x_k \leqq V_k$, a subsequence $\{\|x_{k_i}\|\}$ of $\{\|x_k\|\}$, again depending on x_0, decays at least at the same rate.

We must now show that as a result, $\{\|x_k\|\}$ also decays exponentially fast. For arbitrary k, there exists a greatest k_i with

$$k_i \leqq k < k_i + s + 1.$$

Thus any x_k can be written as $x_k = \phi_{k,k_i} x_{k_i}$, where $0 \leqq k - k_i \leqq s$. Consequently

$\|\phi_{k,k_i}\|$ is bounded, and we have

$$\|x_k\| \leqq a' b^i \leqq a' b^{k_i (s+1)^{-1}} \leqq a''(b')^k,$$

where $b' = b^{(s+1)^{-1}} < 1$ and $a'' = a'/(b')^s$. This bound holds irrespective of the sequence $\{k_i\}$ induced by the particular x_0, and exhibits the required exponential convergence. □

Remark. If (4.4) holds not over $[k_0, \infty)$ but over $[k_0, k_1]$, we can still conclude the existence of constants $\alpha > 0$, $\beta \in [0, 1)$, independent of k_0, k_1, such that

$$\|\phi_{l,k}\| \leqq \alpha \beta^{l-k},$$

for $k_0 \leqq k \leqq l \leqq k_1$ by minor variation on the above argument. This remark is crucial in establishing the dual of Lemma 4.2 and Theorem 4.3, the proof of which is otherwise straightforward and is omitted. The dual will be of use in the next section.

THEOREM 4.3. *Suppose that $[\hat{F}_k, \hat{G}_k]$ is uniformly stabilizable, that $\hat{F}_k$ and $\hat{G}_k$ are bounded, and that there is a bounded nonnegative definite matrix sequence $\hat{P}_k$ satisfying*

$$\hat{P}_{k+1} = \hat{F}_k \hat{P}_k \hat{F}_k' + \hat{G}_k \hat{G}_k' \tag{4.6}$$

on $[k_0, \infty)$. Then $x_{k+1} = \hat{F}_k x$ is exponentially stable. Conversely, if $x_{k+1} = \hat{F}_k x_k$ is exponentially stable, and $\hat{F}_k$, $\hat{G}_k$ are bounded, there exists a unique bounded nonnegative definite matrix sequence $\hat{P}_k$ satisfying (4.6) *on $[k_0, \infty)$.*

5. Detectability, stabilizability and state estimation. Consider the problem of state estimation for the signal process

$$\begin{aligned} x_{k+1} &= F_k x_k + G_k w_k, \\ y_k &= H_k' x_k + v_k. \end{aligned} \tag{5.1}$$

Here, $\{w_k\}$, $\{v_k\}$ are independent, zero mean, white processes with $E[w_k w_k'] = I$, $E[v_k v_k'] = I$. (A nonunit covariance for w_k is absorbed in G_k and a nonunit covariance for v_k is absorbed by scaling y_k and H_k', so long as the covariance is nonsingular). We assume that $E[x_0 x_0'] = P_0$, $E[x_0] = m$, and $x_0, \{w_k\}, \{v_k\}$ are independent. Finally, we assume F_k, G_k, and H_k are bounded.

The main results of the section are: $[F_k, H_k]$ uniformly detectable is sufficient for the optimal (Kalman filter) error covariance to be bounded. Furthermore, if $[F_k, G_k]$ is uniformly stabilizable, the Kalman filter is exponentially stable. Uniform detectability of $[F_k, H_k]$ is sufficient for there to exist a bounded sequence K_k such that $x_{k+1} = (F_k - K_k H_k')x_k$ is exponentially stable.

LEMMA 5.1. *With notation and assumptions as above, and with $[F_k, H_k]$ uniformly detectable, there exists a state estimator producing a filtered state estimate for* (5.1) *with bounded error covariance.*

Proof. (By construction.) Let s, t, d, b have their usual meanings. By orthogonal transformation of the state coordinate basis at each time instant, we may assume that $M_{k+s,k} = M^1_{k+s+k} \dot{+} M^2_{k+s,k}$ where $M^1_{k+s,k} \geqq bI$ and $M^2_{k+s,k} < bI$. (The symbol $\dot{+}$ denotes direct sum).

Define the smoothed estimate

$$\bar{x}_{k+t|k+s} = \phi_{k+t|k} \left\{ \begin{bmatrix} (M^1_{k+s,k})^{-1} & 0 \\ 0 & 0 \end{bmatrix} \sum_k^{k+s} \phi_{i,k}' H_i y_i + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \hat{x}_{k|k+s-t} \right\}, \tag{5.2}$$

with initialization $\hat{x}_{i|i+s-t} = 0$ for $i = 0, \cdots, t-1$. The partitioning in the matrix multiplying $\hat{x}_{k|k+s-t}$ is the same as that in the matrix multiplying $\sum_k^{k+s} \phi'_{i,k} H_i y_i$. Now $y_i = H'_i \phi_{i,k} x_k + [*]$, where $[*]$ is a bounded linear combination of w_j, v_j, $j \in [k, k+s]$. Consequently,

$$\hat{x}_{k+t|k+s} = \phi_{k+t,k} \left\{ \begin{bmatrix} (M^1_{k+s,k})^{-1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} M^1_{k+s,k} & 0 \\ 0 & M^2_{k+s,k} \end{bmatrix} x_k + \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} \hat{x}_{k|k+s-t} \right\} + [*],$$

or

$$\hat{x}_{k+t|k+s} - x_{k+t} = \phi_{k+t,k} \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} (\hat{x}_{k|k+s-t} - x_k) + [*].$$

Using the detectability definition and structure of $M_{k+s,k}$, it is easily seen that

$$\left\| \phi_{k+t,k} \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} \right\| \leqq d < 1.$$

Consequently, $E[\|\hat{x}_{k+t|k+s} - x_{k+t}\|^2]$ is bounded.

Now define $\hat{x}_{k+s|k+s} = \phi_{k+s,k+t} \hat{x}_{k+t|k+s}$. Then

$$\hat{x}_{k+s|k+s} - x_{k+s} = \phi_{k+s,k+t} (\hat{x}_{k+t|k+s} - x_{k+t}) + [*],$$

and clearly the error covariance bound associated with the smoothed estimate yields a bound for the filtered estimate error covariance. □

The following is immediate by optimality.

COROLLARY 5.2. *With notation and assumptions as above and with* $[F_k, H_k]$ *uniformly detectable, the (Kalman) filter error covariance* $\Sigma_{k|k}$ *and one-step predictor error covariance* $\Sigma_{k+1|k}$ *are bounded.*

Now we couple in a stabilizability constraint to obtain exponential stability of the Kalman filter.

THEOREM 5.3. *With notation and assumptions as above, and with* $[F_k, H_k]$ *detectable and* $[F_k, G_k]$ *stabilizable, the Kalman filter is exponentially stable.*

Proof. With K_k the Kalman filter gain, and with $\Sigma_{k|k}$ and $\Sigma_{k+1|k}$ the filtered and one-step prediction error covariances, we have from [13], quoting only the equations we need,

$$K_k = \Sigma_{k|k} H_k \tag{5.3}$$

and

$$\Sigma_{k+1|k} = (F_k - K_k H'_k) \Sigma_{k|k-1} (F_k - K_k H'_k)' + [G_k K_k][G_k K_k]'. \tag{5.4}$$

The homogeneous equation associated with the one-step predictor is $x_{k+1} = (F_k - K_k H'_k) x_k$, and it is exponentially stable if and only if the filter equation is exponentially stable. This is easily checked; a formal calculation can be found in [14].

By the assumptions and Corollary 5.2, K_k is bounded as is $\Sigma_{k+1|k}$. By Theorem 4.3, exponential stability follows if the pair $[F_k - K_k H'_k]$, $[G_k K_k]$ is uniformly stabilizable.

This is easily established as follows:

$[F_k, G_k]$ is uniformly stabilizable

$\Rightarrow$

$[F_k, [G_k K_k]]$ is uniformly stabilizable (by applying the definition)

$\Rightarrow$

$[F_k + [G_k K_k]\begin{bmatrix} 0 \\ -H_k \end{bmatrix}, [G_k K_k]]$ is uniformly stabilizable

(applying invariance under feedback). □

Now we have the converse to Corollary 3.4, which together with Corollary 3.4 shows the equivalence for bounded F_k, H_k of the detectability property and the existence of an output-to-state feedback law providing exponential stability.

COROLLARY 5.4. *If $[F_k, H_k]$ is uniformly detectable (and F_k, H_k are bounded), there exists a bounded sequence K_k such that $x_{k+1} = (F_k - K_k H_k')x_k$ is exponentially stable.*

Proof. Consider the process (5.1) with $G_k = I$. Then $[F_k, G_k]$ is stabilizable, and Theorem 5.3 provides the result.

We can also consider the necessity of the detectability condition. Certainly, detectability is necessary for there to be an exponentially stable estimator of the type

$$\hat{x}_{k+1|k} = F_k \hat{x}_{k|k-1} + K_k(y_k - H_k' \hat{x}_{k|k-1});$$

(this is effectively the content of Corollary 3.4). However, we can get a slightly sharper result.

COROLLARY 5.5. *Consider the process* (5.1) *and associated assumptions, and with $[F_k, G_k]$ uniformly stabilizable. Suppose that the associated optimal filter error covariance is bounded. Then $[F_k, H_k]$ is uniformly detectable.*

Proof. In (5.4), $\Sigma_{k+1|k}$ is bounded and the pair $F_k - K_k H_k'$, $[G_k K_k]$ is uniformly stabilizable. By Theorem 4.3, $x_{k+1} = (F_k - K_k H_k')x_k$ is exponentially stable, and Corollary 3.4 then yields the result. □

Remarks. 1. The main theorem of this section appeals to almost all the important results of the preceding section. As well, it appeals to the suboptimal estimator construction of Lemma 5.1, which is not trivial and considerably more complicated than constructions which have been used in studying observable processes; see, e.g., [4], [14]. In particular, we were not able here to define an exponentially stabilizing feedback law K_k simply in terms of the observability matrix M, as can be done in the observable case, [14].

2. The corresponding regulator result of course follows by duality, though some care has to be taken because of the fact that with the interval for which the filter is studied being $[0, \infty)$, its dual is $[-\infty, 0)$, while we wish to study the regulator over $(0, \infty)$. The considerations of the remark preceding Theorem 4.3 can be applied to overcome this difficulty. An alternative approach to the regulator problem is to show that complete stabilizability implies the existence of a control yielding a bounded performance index, to conclude then that the optimal index is bounded and achievable with a linear feedback law, and to show under a detectability assumption that the closed-loop is exponentially stable. The construction of the control yielding a bounded performance index is not straightforward; the construction procedure in some way has to parallel the construction of Lemma 5.1.

Finally, in this section, we illustrate that the feedback invariant $\tilde{d}$ defines an achievable bound on how stable we can make a closed-loop system via feedback.

THEOREM 5.6. *With F_k, H_k bounded and uniformly detectable, and with $\tilde{d}$ as defined in Corollary* 3.5, *there exists a feedback law K_k such that all trajectories of $x_{k+1} = (F_k - K_k H_k')x_k$ decay at least as fast as $(\tilde{d}+\varepsilon)^k$ for arbitrary $\varepsilon > 0$.*

Proof. Define the detectability property of $[F_k, H_k]$ using $\tilde{d} + \varepsilon/2$, b, t, and s, as we are entitled to do. Choose so that $\rho^t(\tilde{d}+\varepsilon) = 1$. Then $[\rho F_k, H_k]$ is detectable, by Lemma 2.3 and the remark following the lemma. Find K_k so that $x_{k+1} = (\rho F_k - K_k^{\rho} H_k')x_k$ is exponentially stable. Then choosing $K_k = \rho^{-1} K_k^{\rho}$ ensures that $x_{k+1} = (F_k - K_k H_k')x_k$ has the desired property.

Remarks. 1. The discussion following Corollary 3.5 makes it clear that we could not obtain a feedback law K_k such that the closed-loop system trajectories decay as fast as $(\tilde{d}-\varepsilon)^{k/t}$ for some ε with $0 \leqq \varepsilon \leqq \tilde{d}$.

2. As was noted in Remark 1 following Corollary 5.5, we are unable to define a stabilizing gain sequence K_k simply in terms of the observability matrix associated with a detectable pair $[F_k, H_k]$. The construction given for the stabilizing gain sequence via Corollary 5.4 has the potential disadvantage that K_k depends on F_l, H_l for all $l \leqq k$. This is at least a "causal" dependence; when one considers the problem of constructing a stabilizing sequence $\hat{K}_k$ for a stabilizable pair $[\hat{F}_k, \hat{G}_k]$, the disadvantage is that $\hat{K}_k$ depends on $\hat{F}_l$, $\hat{G}_l$ for all $l \geqq k$. This leads one to consider whether or not there might be a sequence dependent on a finite "window" only of $[F_l, H_l]$ or $[\hat{F}_l, \hat{G}_l]$. Indeed there is. We describe the detectable situation only. Following the idea of the proof of Theorem 5.6, the Kalman filter equations are solved forward in time with F_k replaced by ρF_k, $\Sigma_{0|-1} = I$, $Q_k = I$, save that at some set of times $r_1, r_2, \cdots$, the choice of which is described below, we replace the value of $\Sigma_{r_i|r_i-1}$ predicted by the equations by $\Sigma_{r_i|r_i-1} = I$, amounting to a reinitialization. This has the effect of causing K_k^{ρ} and K_k for $k \in [r_i, r_{i+1}-1]$ to depend on F_l, H_l for $l \in [r_i, r_{i+1}-1]$, i.e., on no more than $r_{i+1} = r_i$ values of F_k, H_k. The integers r_i are chosen so that, with $\bar{\phi}_{k,l}$ the transition matrix associated with $F_k - K_k H_k'$, one has $\|\bar{\phi}_{r_{i+1}, r_i}\| < 1$. Actually, a lengthy argument will show that r_i may be taken as ir for some appropriately large integer r. The upshot is the desired exponential decay property.

6. Coordinate basis choice to display detectability. If $[F, H]$ is a time-invariant detectable pair, it is well known (see [3]) that if the coordinate basis is chosen satisfactorily, then we can have

$$F = \begin{bmatrix} F_{11} & 0 \\ F_{21} & F_{22} \end{bmatrix}, \qquad H' = [H_1' \quad 0], \tag{6.1}$$

with $[F_{11}, H_1]$ observable and $|\lambda_i(F_{22})| < 1$. We seek here a time-varying version of this result.

Assume $[F_k, H_k]$ is detectable, and d, b, s, and t have their usual meanings. We shall take b here to be very small and assume also that, for some arbitrary but fixed k,

(a) $M_{k+s+1,k}$, $M_{k+s,k}$ and $M_{k+s,k-1}$ have precisely p eigenvalues less than b, and the remaining eigenvalues much greater than b.

(b) By orthogonal changes of basis and without loss of generality or variation of the stability properties, $M_{k+s,k-1}$ and $M_{k+s+1,k}$ are diagonal, with diagonal elements taking decreasing values down the diagonal.[1]

Under these assumptions and with

$$H_{k-1}' = [H_{k-1}^{1'} \quad \overbrace{H_{k-1}^{2'}}^{p}], \qquad F_{k-1} = \begin{bmatrix} F_{k-1}^{11} & \overbrace{F_{k-1}^{12}}^{p} \\ F_{k-1}^{21} & F_{k-1}^{22} \end{bmatrix}_p,$$

[1] Note that if $\hat{x}_k = T_k x_k$ defines the orthogonal basis change, $M_{k+s,k} \to \hat{M}_{k+s,k} = T_k M_{k+s,k} T_k'$.

we assert that $\|H_{k-1}^2\|^2$ and $\|F_{k-1}^{12}\|^2$ are $O(b)$. Since b is small, this means that F_{k-1} and H_{k-1} are approximately of the form of (6.1). Furthermore, we can show that $\phi_{k+t,k}^{22} < d + O(b^{1/2})$, and $\|F_{k+t-1}^{22} F_{k+t-2}^{22} \cdots F_k^{22}\| < d + O(b^{1/2})$. This mimics the requirement in the time-invariant case that $|\lambda_i(F_{22})| < 1$. Finally, an observability result can be obtained for the pair $[F_k^{11}, H_k^1]$. In case F_k, H_k are constant, the time-invariant results are evidently recovered. Because of length restrictions, proofs are omitted.

7. Conclusions. Given the now widespread knowledge of the linear-quadratic problem and its solutions, the results of this paper are not particularly surprising. Certainly, when the ideas of the paper were being developed many of the conjectures were clear. In hindsight, there is also no real surprise in the techniques required to obtain the results. However, we must admit that many of the specific techniques, especially that of Lemma 5.1, surfaced only after exploring a number of misleading approaches and conjectures. Perhaps this accounts for the comparatively long time between the intuitive grasp of the general nature of these results and their formal derivation.

It is clear that one of the main applications of the results is to the linear-quadratic problem. However, we feel it likely that the extended lemma of Lyapunov is a result of some power, which should also find significant application. We have, for example, recently been able to use this lemma to establish that if a linear, finite-dimensional, uniformly stabilizable and detectable system is bounded-input, bounded-output stable, then the system is necessarily exponentially stable.

REFERENCES

[1] B. D. O. ANDERSON AND J. B. MOORE, *Linear Optimal Control*, Prentice-Hall, Englewood Cliffs, NJ, 1971.

[2] W. M. WONHAM, *Linear Multivariable Control, A Geometric Approach*, Lecture Notes in Economics and Mathematical Systems, vol. 101, Springer-Verlag, Berlin, 1974.

[3] H. KWAKERNAAK AND R. SIVAN, *Linear Optimal Control Systems*, Wiley-Interscience, New York, 1972.

[4] A. H. JAZWINSKI, *Stochastic Processes and Filtering Theory*, Academic Press, New York, 1970.

[5] W. W. HAGER AND L. L. HOROWITZ, *Convergence and stability properties of the discrete Riccati equation and the associated control and filtering problems*, this Journal, 14 (1976), pp. 295–312.

[6] D. L. KLEINMAN, *An easy way to stabilize a linear system*, IEEE Trans. Automat. Control, AC-15 (1970), p. 692.

[7] ———, *Stabilizing a discrete, constant, linear system with application to iterative methods for solving the Riccati equation*, IEEE Trans. Automat. Control, AC-19 (1974), pp. 252–254.

[8] W. H. KWON AND A. E. PEARSON, *A modified quadratic cost problem and feedback stabilization of a linear system*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 838–842.

[9] ———, *On feedback stabilization of time-varying discrete linear systems*, IEEE Trans. Automat. Control, AC-23 (1978), pp. 479–481.

[10] R. E. KALMAN AND J. E. BERTRAM, *Control system analysis and design by the second method of Lyapunov*, Trans. ASME Ser. D J. Basic Engineering, 82 (1960), pp. 371–400.

[11] R. E. KALMAN, *Lyapunov functions for the problem of Lur'e in automatic control*, Proc. Nat. Acad. Sci. 49 (1963), pp. 201–205.

[12] B. D. O. ANDERSON AND J. B. MOORE, *New results in linear system stability*, SIAM J. Control, 7 (1969), pp. 398–414.

[13] ———, *Optimal Filtering*, Prentice-Hall, Englewood Cliffs, NJ, 1979.

[14] J. B. MOORE AND B. D. O. ANDERSON, *Coping with singular transition matrices in estimation and control stability theory*, Internat. J. Control, 31 (1980), pp. 571–586.

[15] R. E. BELLMAN, *Introduction to Matrix Analysis*, 2nd ed., McGraw Hill, New York, 1970.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 1, January 1981

0363-0129/81/1901-0003 $01.00/0

THE EQUATION $XR+QY=\Phi$: A CHARACTERIZATION OF SOLUTIONS*

E. EMRE† AND L. M. SILVERMAN‡

Abstract. In this paper we consider the solutions of the equation $XR+QY=\Phi$. Here Q, R, Φ are given $p\times q$, $m\times t$ and $p\times t$ polynomial matrices over a field k. X and Y are $p\times m$ and $q\times t$ polynomial matrices which are unknown. Using certain recent results on the realization of matrix fraction descriptions of transfer matrices, we give a characterization (parametrization) of all possible (X, Y) which solve this equation. This also provides a system theoretic interpretation for this equation.

1. Introduction. Let Q, R, Φ be $p\times q$, $m\times t$, and $p\times t$ polynomial matrices over an arbitrary field k. In this paper, we consider the solutions of the equation $XR+QY=\Phi$ for $q\times t$ and $p\times m$ polynomial matrices Y and X over k.

This equation has been considered by several authors including Roth (1952) over fields, Gustafson (1979) over commutative rings with identity, Wolovich (1977), Bengtsson (1977), Cheng and Pearson (1978), and Kucera (1975) over polynomials. In general, it has been shown by Gustafson (1979) that over a commutative ring $\bar{R}$ with identity, $XR+QY=\Phi$ has a solution (X, Y) over $\bar{R}$ iff the matrices

$$\begin{bmatrix} Q & O \\ O & R \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} Q & \Phi \\ O & R \end{bmatrix}$$

are equivalent. This is a generalization of the result of Roth (1952) which gives the same criterion for the case where $\bar{R}$ is a field.

From a system theoretic point of view, this equation is important in the design of control systems where now the ring $\bar{R}$ is the set of polynomials over a field k, $k[x]$. For details, the reader is referred to Bengtsson (1977), Cheng and Pearson (1978), Wolovich and Ferreira (1979), Kucera (1975) and the references given there. In all of the prior work the basic tool is the invariant factor theorem for a polynomial ring $k[x]$. A direct explicit characterization of all solutions is not provided.

In § 2 of this paper, using some recent system theoretic results on the realization of matrix fraction descriptions by Fuhrmann (1977), we transform this problem (first for the case Q nonsingular) to a set of linear equations over the field k. This also provides a system theoretic interpretation and more insight. We show that a solution to the original equation exists iff a solution to the linear equations over k exists. Further, when a solution exists, all the solutions of the original equation are characterized (parametrized) in terms of all possible solutions to the equations over k.

Finally, in § 3 we consider the case where Q is a general polynomial matrix and show how the general problem can be solved using the results of § 2.

2. The case where Q is nonsingular. In what follows $k^q[z]$ denotes q-tuples of polynomials in z with coefficients in k, and $k^q(z)$ denotes q-tuples of rational functions

* Received by the editors November 1, 1979, and in revised form March 21, 1980. This research was supported in part by the National Science Foundation under Grant ENG-7908673, and in part by the U.S. Army Research Office under Grant DAA29-77-G-0225 and the U.S. Air Force under Grant AFOSR 76-3034 Mod. B through the Center for Mathematical System Theory, University of Florida, Gainesville, Florida 32611.

† Center for Mathematical System Theory, University of Florida, Gainesville, Florida 32611.

‡ Department of Electrical Engineering–Systems, University of Southern California, Los Angeles, California 90007.

in z over k. In this section we assume that $p = q$ and Q is nonsingular. k_Q denotes the k-linear space

$$k_Q := \{x \in k^p[z]: Q^{-1}x \text{ is strictly proper}\}.$$

For a $p \times m$ polynomial matrix X_1 such that $Q^{-1}X_1$ is strictly proper, the k-linear maps π, F, G, H, π_Q are defined as follows:

$$G: k^m \to k_Q, \qquad u \mapsto X_1 u \quad \text{for } u \text{ in } k^m.$$

$$\pi: k^p(z) \to k^p(z), \qquad q \mapsto \text{strictly proper part of } q.$$

$$\pi_Q: k^p[z] \to k^p[z], \qquad x \mapsto Q\pi(Q^{-1}x).$$

$$F: k_Q \to k_Q, \qquad x \mapsto \pi_Q(zx).$$

$$H: k_Q \to k^p, \qquad x \mapsto (Q^{-1}x)_{-1}.$$

Here $(Q^{-1}x)_{-1}$ is the coefficient of z^{-1} in the formal power series of $Q^{-1}x$ in z^{-1}. In terms of the above quantities we have the following results which will be needed in the sequel:

LEMMA 2.1. (Fuhrmann (1977)). *Let* $Z := Q^{-1}X_1$. *Then* $\Sigma := (F, G, H)$ *is an observable realization of* Z *with the state space* k_Q. (*We call* Σ *the* Q*-realization of* Z.)

LEMMA 2.2. *Let* X, Y *be solutions of*

$$XR + QY = \Phi. \tag{2.3}$$

Then there exist X_1, Y_1 *which are solutions of* (2.3) *such that* $Q^{-1}X_1$ *is strictly proper.*

Proof. Suppose X, Y are solutions of (2.3). Then extending the map π_Q to matrices in a natural way, define

$$X_1 := \pi_Q(X).$$

Clearly $Q^{-1}X_1$ is strictly proper. On the other hand, clearly there exists a unique polynomial matrix Q_1 such that

$$X = QQ_1 + X_1.$$

Substitution into (2.3) yields

$$X_1R + QY + QQ_1R = \Phi,$$

or with

$$Y_1 := Y + Q_1R,$$

we have

$$X_1R + QY_1 = \Phi. \qquad \square$$

Now suppose Q, R, Φ are given as before. Define

$$E(Q, R) := \{(X, Y): XR + QY = \Phi\},$$

and

$$\bar{E}(Q, R) := \{(X_1, Y_1): X_1R + QY_1 = \Phi \text{ and } Q^{-1}X_1 \text{ is strictly proper}\}.$$

In light of Lemma 2.2, it is clear that once we have a characterization of $\bar{E}(Q, R)$ we also have a characterization of $E(Q, R)$. Once we have a pair (X_1, Y_1) in $\bar{E}(Q, R)$ then $(X_1 + QQ_1, Y_1 - Q_1R)$ will be elements of $E(Q, R)$ for any $p \times p$ polynomial matrix Q_1. Also if (X, Y) is any element of $E(Q, R)$ we can obtain a unique element

(X_1, Y_1) of $\bar{E}(Q, R)$ as $(X-QQ_1, Y+Q_1R)$ for some unique polynomial matrix Q_1. That is, we have

$$E(Q, R)=\{(X_1, Y_1)+(\bar{X}, \bar{Y}): (X_1, Y_1)\in \bar{E}(Q, R) \text{ and } (\bar{X}, \bar{Y})\in \bar{H}(Q, R)\},$$

where

$$\bar{H}(Q, R):=\{(QQ_1, -Q_1R): Q_1 \text{ is an arbitrary polynomial matrix}\}.$$

Hence to obtain a characterization of $E(Q, R)$, it is sufficient to obtain a characterization of $\bar{E}(Q, R)$. For this we first define the following. Let S be a $p\times n$ polynomial matrix whose columns are a basis of k_Q.

Let F, G_1, H be the Q-realization of $Q^{-1}S$ as in Lemma 2.1. Let $\hat{F}$, $\hat{G}_1$, $\hat{H}$ be the matrix representations of F, G_1, H relative to canonical bases of K^m, K^p (note that in this case $p=n$) and columns of S as a basis for k_Q. It follows that $\hat{G}_1=I_n$, and hence

$$\hat{H}(zI-\hat{F})^{-1}=Q^{-1}S.$$

Let R be expressed as

$$R=\sum_{j=0}^{r} u_{-j}z^j,$$

where u_{-j} are $m\times p$ matrices over k. Also we can define an $n\times p$ matrix $\hat{\Phi}$ over k uniquely by

$$\pi_Q(\Phi)=S\hat{\Phi},$$

and express Φ as

$$\Phi=Q\Phi_1+S\hat{\Phi},$$

for a unique polynomial matrix Φ_1.

Finally, let $\hat{G}$ denote the (unknown) $n\times m$ matrix over k in the linear equations

$$\sum_{j=0}^{r} \hat{F}^j\hat{G}_{-j}=\hat{\Phi}. \tag{2.4}$$

THEOREM 2.5. *The following statements are equivalent:*

(i) (X_1, Y_1) *is an element of* $\bar{E}(Q, R)$.

(ii) $X_1=S\hat{G}$ *for some* $\hat{G}$ *which is a solution of* (2.4), *and* $Y_1=\Phi_1-Q_p$ *where* Q_p *is the polynomial part of* $Q^{-1}X_1R$.

Proof. 1) Suppose that $X_1R+QY_1=\Phi$ and $Q^{-1}X_1R=\Phi_1-Y_1+Q^{-1}\pi_Q(\Phi)$.

Let $\Sigma=(F, G, H)$ be the Q-realization of $Q^{-1}X_1$. Let F_1, G_1, H_1 be the matrix representations of F, G, H relative to canonical bases of k^p, k^m and the columns of S taken as a basis of k_Q. Then $F_1=\hat{F}$, $H_1=\hat{H}$, and

$$\hat{H}(zI-\hat{F})^{-1}=Q^{-1}S.$$

But then

$$\hat{H}(zI-\hat{F})^{-1}G_1=Q^{-1}SG_1=Q^{-1}X_1,$$

which implies that $X_1=SG_1$; i.e., with $\hat{G}=G_1$, $X_1=S\hat{G}$.

On the other hand,

$$Q^{-1}X_1R=\Phi_1-Y_1+Q^{-1}\pi_Q(\Phi)$$

expresses the fact that the system Σ with the input sequence consisting of the ith

columns of u_{-j}, $j = 0, \cdots, r$, reaches the state φ_i which is the ith column of $\pi_Q(\Phi)$ from the zero state. During this time it produces the output sequence consisting of the coefficient vectors of the ith column of the polynomial matrix $Q_p := \Phi_1 - Y_1$. Relative to chosen bases of k^p, k^m and k_Q, this state transition for each i, $i = 1, \cdots, p$ can be written as in (2.4). One explicit way of seeing this is equating the strictly proper part of $Q^{-1}X_1R$ to $Q^{-1}\pi_Q(\Phi)$. This yields

$$\begin{bmatrix} \hat{H} \\ \hat{H}\hat{F} \\ \vdots \end{bmatrix} [\hat{G} : \hat{F}\hat{G} : \cdots : \hat{F}^r\hat{G}] \begin{bmatrix} u_0 \\ u_{-1} \\ \vdots \\ u_{-r} \end{bmatrix} = \begin{bmatrix} \hat{H} \\ \hat{H}\hat{F} \\ \vdots \end{bmatrix} \hat{\Phi}.$$

Then, by observability of $(\hat{H}, \hat{F})$, (2.4) follows. Hence, $\hat{G}$ is a solution of (2.4). Clearly, if Q_p is the polynomial part of $Q^{-1}X_1R$ then $Y_1 = \Phi_1 - Q_p$.

2) Suppose that $X_1 := S\hat{G}$ for some solution $\hat{G}$ of (2.4). Let us now consider the system $\hat{\Sigma} = (\hat{F}, \hat{G}, \hat{H})$. Then $\hat{F}$, $\hat{G}$, $\hat{H}$ are matrix representations of F, G, H relative to canonical bases of k^p, k^m and the columns of S taken as a basis of k_Q, where $\Sigma = (F, G, H)$ is the Q-realization of $Q^{-1}X_1$. Then (2.4) expresses the fact that the input sequences consisting of ith columns of u_{-j} drive Σ from the zero state to that which is the ith column of $S\hat{\Phi} = \pi_Q(\Phi)$. In polynomial terms this can be written as

$$Q^{-1}X_1R = Q_p + Q^{-1}\pi_Q(\Phi),$$

where Q_p, the polynomial part of $Q^{-1}X_1R$, represents the outputs produced by Σ during this state transition. Now if we define Y_1 to be $\Phi_1 - Q_p$, we get

$$Q^{-1}X_1R = \Phi_1 - Y_1 + Q^{-1}\pi_Q(\Phi),$$

or

$$X_1R + QY_1 = \Phi. \qquad \square$$

Remark. By Theorem 2.5, we have transformed the problem to the problem of finding the solutions of the linear equations (2.4) in $\hat{G}$. It follows from Lemma 2.2 and Theorem 2.5 that (2.3) has a solution (X, Y) iff (2.4) has a solution $\hat{G}$. When a solution exists, one can obtain a parametrization $\hat{G}$ representing all possible solutions of (2.3) from (2.4). Then all possible X_1 are given by $S\hat{G}$ and Y_1 by $\Phi_1 - Q_p$.

Remark. An alternative approach to the problem is as follows. It is easy to see from the argument in the proof of Theorem 2.5 that X_1 is a solution iff

$$\pi_Q(X_1R) = \pi_Q(\Phi).$$

Let S be a basis matrix for k_Q. Since $Q^{-1}X_1$ is strictly proper, $X_1 = S\hat{G}$ for some (unknown) $\hat{G}$. Hence if we take the entries of $\hat{G}$ as unknowns and equate $\pi_Q(X_1R)$ to $\pi_Q(\Phi)$, we obtain the same set of linear equations given by (2.4), whose unknowns are the entries of $\hat{G}$. Then Y_1 is uniquely determined by X_1 as $\Phi_1 - Q_p$. Theorem 2.5 has all these operations built in, and is essentially the same formulation. But it yields a conceptually clearer picture as well as a system theoretic interpretation of these equations. Also it provides a more systematic way of obtaining these linear equations over k.

Remark. We should note here that construction of a basis matrix S for k_Q can be easily done as follows. Let M be a unimodular polynomial matrix such that $\bar{Q} := MQ$ is row proper (say, with ith row degree v_i). Then it is well known that for a polynomial vector x, $\bar{Q}^{-1}x$ is strictly proper iff the ith row degree of x is less than v_i. Hence, for a polynomial vector $\bar{x}$, $Q^{-1}\bar{x}$ is strictly proper iff the ith row degree of $M\bar{x}$ is less than v_i.

From this argument it follows that one choice for S is

$$S = M^{-1}\begin{bmatrix} z^{v_1-1}\cdots 1 & 0 & \cdots & 0 \\ & \ddots & & \\ 0 & & & z^{v_p-1}\cdots 1 \end{bmatrix}.$$

Remark. From the formulation given in the above remark, we see that the same approach is applicable to the case where k is a commutative ring whenever Q is row proper (i.e., the highest degree row coefficient matrix of Q is invertible over k).

Remark. The equation $RX+YQ=\Phi$ is the dual of the equation $XR+QY=\Phi$, and thus after transposition, one can apply the same results.

3. The general case. Now suppose that Q is a general $p\times q$ polynomial matrix. Then there exist unimodular polynomial matrices M_1, M_2 such that

$$M_1QM_2=\begin{bmatrix}\hat{Q} & 0\\ 0 & 0\end{bmatrix},$$

where $\hat{Q}$ is a nonsingular polynomial matrix. Define

$$\hat{X} := M_1X =: \begin{bmatrix}\hat{X}_1\\ \hat{X}_2\end{bmatrix},$$

$$\hat{Y} := M_2^{-1}Y =: \begin{bmatrix}\hat{Y}_1\\ \hat{Y}_2\end{bmatrix},$$

$$\begin{bmatrix}\tilde{\Phi}_1\\ \tilde{\Phi}_2\end{bmatrix} := M_1\Phi,$$

$$\hat{X}R+\begin{bmatrix}\hat{Q} & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}\hat{Y}_1\\ \hat{Y}_2\end{bmatrix}=\begin{bmatrix}\tilde{\Phi}_1\\ \tilde{\Phi}_2\end{bmatrix}.$$

Hence the original equation is equivalent to the equations

$$\hat{X}_2R=\tilde{\Phi}_2, \tag{3.1}$$

and

$$\hat{X}_1R+\hat{Q}\hat{Y}_1=\tilde{\Phi}_1. \tag{3.2}$$

Clearly, characterization of solutions to (3.2) can be obtained as in § 2. As for (3.1), for an explicit characterization of solutions the reader is referred to Emre (1980).

Remark. We should note here that in general M_1, M_2, $\hat{Q}$, $\hat{X}$, $\hat{Y}$, $\tilde{\Phi}_1$, $\tilde{\Phi}_2$ are not uniquely determined from the original equation. However, once M_1, M_2 are fixed and known, the solutions of the original equation are characterized by the solutions of (3.1) and (3.2). The original equation has a solution iff (3.1) and (3.2) have solutions. The solutions of the original equation can be obtained from solutions of (3.1) and (3.2), by choosing $\hat{Y}_2$ as an arbitrary polynomial matrix of suitable dimensions and using the transformations M_1 and M_2. Nonuniqueness of M_1, M_2 does not pose any problem as far as the characterization of solutions (X, Y) is concerned.

REFERENCES

G. BENGTSSON (1977), *Output regulation and internal models—a frequency domain approach*, Automatica 13, pp. 333–345.

L. CHENG AND J. B. PEARSIN, JR. (1978), *Frequency domain synthesis of multivariable linear regulators*, IEEE Trans. Automat. Control AC-23, pp. 3–15.

E. EMRE (1980), *The polynomial equation $QQ_c + RP = \Phi$ with application to dynamic feedback*, this Journal 18, pp. 611–620.

P. A. FUHRMANN (1976), *Algebraic system theory, an analyst's point of view*, J. Franklin Inst., 301, pp. 521–540.

W. H. GUSTAFSON (1979), *Roth's theorem over commutative rings*, Linear Algebra Appl. 23, pp. 245–251.

W. KUCERA (1975), *Algebraic approach to discrete stochastic control*, Kybernetika 11, pp. 114–149.

W. E. ROTH (1952), *The equations $AX - YB = C$ and $AX - XB = C$ in matrices*, Proc. Amer. Math. Soc. 3, pp. 392–396.

W. A. WOLOVICH (1977), *Skew Prime Polynomial Matrices*, Brown University Engineering Report ENG DC77-1.

W. A. WOLOVICH AND P. FERREIRA (1979), *Output regulation and tracking in linear multivariable systems*, IEEE Trans. Automat. Control AC-24, pp. 460–465.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 1, January 1981

0363-0129/81/1901-0004 $01.00/0

EXIT PROBABILITIES FOR A CLASS OF PERTURBED DEGENERATE SYSTEMS*

ONÉSIMO HERNÁNDEZ-LERMA†

Abstract. We consider a diffusion Markov process which obeys a stochastic differential equation with coefficients depending on a small parameter ε. The noise enters only in some of the components of the equation and, therefore, the process is degenerate in the sense that the backward operator associated to it is degenerate parabolic. Our problem is to estimate the probability that the process exits from a given region during a certain time interval. The method of solution is similar to one introduced by W. H. Fleming (cf. IRIA Seminars Review, 1977; Appl. Math. Optim., 4 (1978), pp. 329–346) for nondegenerate systems using techniques of stochastic control theory.

Introduction. Consider the n-dimensional process x^ε defined by

$$dx^\varepsilon = F(t, x^\varepsilon(t), y^\varepsilon(t))\,dt, \qquad s \leqq t \leqq T, \tag{0.1}$$

where ε is a positive parameter and y^ε is a diffusion Markov process in Euclidean m-space which obeys the stochastic differential equation

$$dy^\varepsilon = b(t, y^\varepsilon(t))\,dt + \sqrt{\varepsilon}\,\sigma(t, y^\varepsilon(t))\,dW, \qquad s \leqq t \leqq T. \tag{0.2}$$

Equations (0.1), (0.2) define jointly the $(n+m)$-dimensional diffusion process $(x^\varepsilon(t), y^\varepsilon(t))$, which is degenerate in the sense that its covariance matrix is nonnegative definite, or equivalently, the backward operator associated to the process is degenerate parabolic. Let τ^ε be the exit time of $x^\varepsilon(t)$ from a given bounded domain D in R^n. Our problem is to give an estimate for the exit probability

$$q^\varepsilon = P(\tau^\varepsilon \leqq T).$$

It is shown below that

$$-\varepsilon \log q^\varepsilon \to I \quad \text{as} \quad \varepsilon \to 0, \tag{0.3}$$

where I is the value function associated to certain (deterministic) control problems.

The problem stated above is sometimes called (see, e.g., [9], [13]) the exit problem. In its general form, the problem is related to the behavior of a deterministic system when it is perturbed by a random (white) noise of small intensity. Randomly perturbed dynamical systems have been extensively studied in the mathematical literature [3]–[6], [9], [13], and their applications have been explored for engineering [8], [17], random propagation problems [10], mathematical ecology [12], and other areas.

Statement (0.3) has been proved under different assumptions on the coefficients F, b, and σ [5], [6], [9], [18]. The main difference between previous results and our present case is that the backward operator of the process $(x^\varepsilon(t), y^\varepsilon(t))$ is not uniformly parabolic. This is a key point because the proof of (0.3) requires expressing q^ε as a smooth (not merely a weak) solution of a certain boundary value problem. Following a suggestion by Fleming [5], we overcome this difficulty by assuming that the backward operator is hypoelliptic [11], [14]. By a theorem of D. L. Elliot (see Clark [1]), this is equivalent to a certain form of "controllability" of the system (0.1)–(0.2). To prove (0.3) we use a method similar to that introduced by Fleming [5], [6] for nondegenerate systems using techniques of stochastic control theory.

* Received by the editors February 21, 1979, and in final revised form March 17, 1980. This research was supported in part by the Consejo Nacional de Ciencia y Tecnologia, under grant PNCB 198.

† Departamento de Matemáticas, Centro de Investigación del Instituto Politécnico Nacional, México 14, D.F., Mexico.

We begin in § 1 by describing the exit problem in precise terms. In § 2 we state the required assumptions on the coefficients of (0.1), (0.2), and the backward operator associated to the process $(x^\varepsilon(t), y^\varepsilon(t))$. In § 3 it is shown that the exit probability q^ε is a smooth solution of the boundary value problem (3.1). Making a logarithmic transformation on q^ε, we obtain in § 4 the dynamic programming equation (4.2) of certain stochastic control problems. Setting (formally) $\varepsilon = 0$ in (4.2) we obtain the dynamic programming equation of the deterministic control problem (4.7) whose value function I is the limit in (0.3). Finally, in § 5 we prove statement (0.3), using the first Ventsel-Friedlin estimate [18] and a stochastic control argument used before by Fleming [5] for nondegenerate systems.

Notation. If A is a matrix, A^* denotes its transpose and Tr (A) its trace. Given a function $v: R^n \to R^1$, v_x and v_{xx} denote the gradient and the Hessian matrix of v, respectively. Random variables are tacitly referred to a fixed underlying probability space $(\Omega, \mathcal{F}, P)$. E denotes expectation; $E_{s,x}(P_{s,x})$ denotes expectation (probability) conditional on the event $x(s) = x$.

1. Statement of the problem. We first proceed to describe the exit problem within our present context.

Consider the $(n+m)$-dimensional diffusion process $(x^\varepsilon(t), y^\varepsilon(t))$, $t \geqq 0$, satisfying the stochastic differential equation

$$\begin{aligned} dx^\varepsilon &= F(t, x^\varepsilon(t), y^\varepsilon(t))\, dt, \\ dy^\varepsilon &= b(t, y^\varepsilon(t))\, dt + \sqrt{\varepsilon}\sigma(t, y^\varepsilon(t))\, dW, \end{aligned} \tag{1.1}$$

where $x \in R^n$, $y \in R^m$, $W(t)$ is an m-dimensional standard Wiener process, and ε is a positive parameter. We can consider (1.1) as a random perturbation of the deterministic process $(x^0(t), y^0(t))$, $t \geqq 0$, which satisfies

$$\begin{aligned} dx^0 &= F(t, x^0(t), y^0(t))\, dt, \\ dy^0 &= b(t, y^0(t))\, dt. \end{aligned} \tag{1.2}$$

The functions F, b, and σ are to be made precise below (§ 2). Let D be a bounded domain in R^n, with a sufficiently smooth boundary ∂D, and let us assume that at some initial time $s \geqq 0$ the process x^ε is in D:

$$(x^\varepsilon(s), y^\varepsilon(s)) = (x, y) \in D \times R^m. \tag{1.3}$$

Let $\tau^\varepsilon = \tau^\varepsilon(s, x, y)$ denote the first exit time from D of the process $x^\varepsilon(t)$, $t \geqq s$, and define $q^\varepsilon = q^\varepsilon(s, x, y)$ as the exit probability:

$$q^\varepsilon(s, x, y) = P_{s,x,y}(\tau^\varepsilon \leqq T), \tag{1.4}$$

where $T > s$ is a (fixed) given time.

With this terminology, our problem can be stated as follows: To prove that

$$\lim_{\varepsilon \to 0} (-\varepsilon \log q^\varepsilon(s, x, y)) = I(s, x, y), \tag{1.5}$$

where $I(s, x, y)$ is the functional defined in (4.7b). We prove this statement in § 5.

2. Assumptions on the perturbed process. To simplify notation, in §§ 2 and 3 we shall drop the indexing ε. Then, in particular, equation (1.1) for the process $(x(t), y(t)) \equiv (x^\varepsilon(t), y^\varepsilon(t))$ becomes

$$\begin{aligned} dx &= F(t, x(t), y(t))\, dt, \\ dy &= b(t, y(t))\, dt + \sqrt{\varepsilon}\sigma(t, y(t))\, dW. \end{aligned} \tag{2.1}$$

Let D be a bounded domain in R^n with boundary, ∂D, of class C^2. For the (fixed) given $T>0$, let Q be the open set

$$Q=(0, T)\times D\times R^m.$$

The backward operator corresponding to the system (2.1), when applied to a function $v(s, x, y)$, can be written as

$$v_s+L^\varepsilon v\equiv v_s+\frac{\varepsilon}{2}\operatorname{Tr}(av_{yy})+F^*v_x+b^*v_y, \tag{2.2}$$

where $a(s, y)=\sigma(s, y)\sigma(s, y)^*$, and $F=F(s, x, y)$, $b=b(s, y)$ are the coefficients in (2.1).

Let us denote by $C^\infty(Q)$ the space of infinitely differentiable functions on Q, and by $C_0^\infty(Q)$ the space of functions $\phi\in C^\infty(Q)$ with compact support in Q. A locally square integrable function v on Q is said to be a "distribution solution" of the equation

$$v_s+L^\varepsilon v=0, \tag{2.3}$$

if for any "test function" $\phi\in C_0^\infty(Q)$,

$$\int_Q(-\phi_t+L^{\varepsilon *}\phi)v\,dQ=0,$$

where dQ denotes a Lebesgue measure on R^{n+m+1} and $L^{\varepsilon *}$ is the adjoint of L^ε; that is,

$$L^{\varepsilon *}\phi=\frac{\varepsilon}{2}\sum_{i,j=1}^{m}(a_{ij}\phi)_{y_iy_j}-\sum_{i=1}^{n}(F_i\phi)_{x_i}-\sum_{i=1}^{m}(b_i\phi)_{y_i}.$$

Throughout the remainder of this work we make the following assumptions.

Assumptions 2.4.

(a) The functions F, b, σ, and σ^{-1} are bounded $C^\infty(Q_0)$-functions, with bounded first derivatives, where

$$Q_0=(0, \infty)\times R^{n+m}.$$

(b) The matrix $a(s, y)=\sigma(s, y)\sigma(s, y)^*$ is strictly positive-definite (that is, there exists $c>0$ such that $y^*ay\geqq cy^*y$ for all $y\in R^m$).

(c) The backward operator given in (2.2) is hypoelliptic in Q_0 (see, e.g., Hormander [11]); this means that, in particular, if $v(s, x, y)$ is a distribution solution of (4.7) in $Q\subset Q_0$, then (after correction on a set of measure zero) v is in $C^\infty(Q)$.

(d) Let $\nu(x)$ be the outer normal to ∂D and let us write $\tau=\tau^\varepsilon(s, x, y)$. Let Γ^+ and Γ^0 denote the sets of points (t, x, y), with $x\in\partial D$, where $F(t, x, y)^*\nu(x)$ is positive and zero, respectively. It is known [16, §7] that

$$P_{s,x,y}((\tau, x(\tau), y(\tau))\in\Gamma^+\cup\Gamma^0, \tau<\infty)=1,$$

for all (s, x, y) in Q. We assume that, for all (s, x, y) in Q,

$$P_{s,x,y}((t, x(t), y(t))\in\Gamma^0 \text{ for some } t\in[s, T])=0.$$

Thus, if $\tau\leqq T$, then $(\tau, x(\tau), y(\tau))\in\Gamma^+$ almost surely.

Remark 2.5. If, for instance, the coefficients F, b, and σ satisfy assumptions (2.4a, b), and the matrix $F_y=(\partial F_i/\partial y_j)$ has rank n everywhere in Q_0, then the backward operator (2.2) is hypoelliptic (Hormander [11]). Necessary and sufficient conditions for hypoellipticity of a general linear differential operator of second order are given by Hormander [11], and Oleinik and Radkevich [14]. The hypoellipticity assumption is

related to some form of "controllability." Specifically, by a theorem of D. L. Elliot (see Clark [1]), hypoellipticity implies that the diffusion process $(x(t), y(t))$ possesses a transition probability density $p(t, (x, y), (\xi, \eta))$ which is C^∞ on $(0, \infty) \times R^{2(n+m)}$, and which satisfies the "forward" equation $p_t = L^{\varepsilon *} p$ (in the variables t, ξ, η). It also implies that $(x(t), y(t))$ has the strong Feller property. For a linear n-dimensional system, say $dx = Ax + Bdw$, where A and B are constant matrices, hypoellipticity is equivalent to the usual "controllability" criterion [7, p. 135]:

$$\text{rank } (B, AB, \cdots, A^{n-1}B) = n.$$

3. The exit probability. Let $(x(t), y(t))$, $0 \leqq t \leqq T$, be the process defined by (2.1), and assume (2.4). Consider the boundary value problem

$$\begin{aligned} v_s + L^\varepsilon v &= 0 \quad \text{in } Q = (0, T) \times D \times R^m, \\ v(s, x, y) &= 1 \quad \text{on } \Gamma_T^+, \\ &= 0 \quad \text{on } \{T\} \times D \times R^m, \end{aligned} \tag{3.1}$$

where L^ε is the operator in (2.2), and $\Gamma_T^+ = \{(s, x, y) \in \Gamma^+ : 0 < s \leqq T\}$.

By a "smooth solution" of (3.1) we mean a solution for which all the derivatives appearing in (3.1) are continuous. Let Q_1 be the set consisting of $Q \cup (\{T\} \times D \times R^m)$, together with the points $(s, x, y) \in \Gamma^+$ such that $0 < s < T$. We shall prove:

THEOREM 3.2. *The exit probability* $q(s, x, y) = P_{s,x,y}(\tau \leqq T)$ *is a smooth solution of the problem* (3.1), *and it is continuous on* Q_1.

The plan of the proof is to introduce a new (nondegenerate) process $(x^\delta(t), y(t))$, $\delta > 0$, and consider (3.1) as the limiting case when $\delta \to 0$.

Let $(x^\delta(t), y(t))$, $\delta > 0$, be the process satisfying

$$\begin{aligned} dx^\delta &= F(t, x^\delta(t), y(t))\, dt + \sqrt{\delta}\, d\tilde{W}, \\ dy &= b(t, y(t))\, dt + \sqrt{\varepsilon}\sigma(t, y(t))\, dW, \end{aligned} \tag{3.3}$$

with the same initial condition as for (2.1):

$$(x^\delta(s), y(s)) = (x(s), y(s)) = (x, y) \in D \times R^m. \tag{3.3'}$$

In (3.3), $y(t)$ is the same as in (2.1), and $\tilde{W}(t)$, $t \geqq s \geqq 0$, is a n-dimensional standard Wiener process independent of $W(t)$ and such that $\tilde{W}(s) = 0$. Let $\tau^\delta = \tau^\delta(s, x, y)$ be the first exit time from D of $x^\delta(t)$.

Define

$$\|x^\delta - x\|_t = \sup_{s \leqq r \leqq t} |x^\delta(r) - x(r)|, \qquad \theta^\delta = \tau^\delta \wedge T, \quad \theta = \tau \wedge T.$$

The proof of Theorem 3.2 is based on the following lemma.

LEMMA 3.4. *For any initial point* $(x, y) \in D \times R^m$ *and* $t > s$,

(a) $\|x^\delta - x\|_t \to 0$,

(b) $\theta^\delta \to \theta$, *and*

(c) $x^\delta(\theta^\delta) \to x(\theta)$

almost surely, as $\delta \to 0$.

Proof. First note that assumption (2.4a) implies that $F(t, x, y)$ satisfies a uniform Lipschitz condition.

Proof of (a). Writing (2.1) and (3.3) in integrated form, subtracting, and taking absolute values, we obtain

$$|x^\delta(t)-x(t)| \leqq \int_s^t |F(r, x^\delta(r), y(r))-F(r, x(r), y(r))|\, dr+\sqrt{\delta}|\tilde{W}(t)|$$

$$\leqq k\int_s^t |x^\delta(r)-x(r)|\, dr+\sqrt{\delta}|\tilde{W}(t)|,$$

so that

$$\|x^\delta-x\|_t \leqq k\int_s^t \|x^\delta-x\|_r\, dr+\sqrt{\delta}\|\tilde{W}\|_t,$$

where k is a Lipschitz constant. Therefore, by the Gronwall–Bellman inequality [7, p. 198], we have

$$\|x^\delta-x\|_t \leqq c\sqrt{\delta}\|\tilde{W}\|_t,$$

where c is a constant which depends only on k and $t-s$. Thus letting $\delta \to 0$, we obtain (a).

Proof of (b). We will show that $\theta^* \leqq \theta \leqq \theta_*$ a.s., where

$$\theta^* = \limsup_{\delta\to 0} \theta^\delta, \qquad \theta_* = \liminf_{\delta\to 0} \theta^\delta.$$

First, since D is open, it will follow from (a) that if $\theta = \tau \wedge T = T$ and $x(\theta) \in D$, then $\theta^\delta = T$ a.s. for all δ sufficiently small and, therefore, we would get (b). Similarly, if $\theta^\delta = T$ and $x^\delta(\theta^\delta) \in D$, (a) will imply (b). Thus we can assume that both $x(\theta) \in \partial D$ and $x^\delta(\theta^\delta) \in \partial D$.

Now if $x^\delta(\theta^\delta) \in \partial D$, it follows from (a) that $x(\theta_*) \in \partial D$ a.s. and, consequently, $\theta_* \geqq \theta$ a.s. To get $\theta^* \leqq \theta$, let $A_{a,\alpha}$ (for $a>0$, $\alpha>0$) be the event (in the underlying sample space) defined as follows: there exists $t \in [\theta, \theta+a]$ such that distance $(x(t), \bar{D}) \geqq \alpha$. If this holds and $\|x^\delta - x\| < \alpha$, then $\theta^\delta < \theta + a$. Thus by part (a), $\theta^* \leqq \theta + a$ on $A_{a,\alpha}$ a.s. On the other hand, $P_{s,x,y}(\bigcup_{\alpha>0} A_{a,\alpha}) = 1$ (α rational, say) by assumption (2.4d). Therefore $P_{s,x,y}(\theta^* \leqq \theta + a) = 1$. Since a is arbitrary, we get that $\theta^* \leqq \theta$ a.s.

Finally, we complete the proof of Lemma 3.4, noting that (c) is a consequence of (a) and (b). □

Proof of Theorem 3.2. By the hypoellipticity assumption (2.4c), to prove that $q(s, x, y) = P_{s,x,y}(\tau \leqq T)$ is a smooth solution (almost everywhere in Q with respect to Lebesgue measure) of (3.1), it is enough to show that:

(3.5) $\qquad q(s, x, y)$ is a distribution solution of (3.1).

Let us consider the following backward equation corresponding to the process $(x^\delta(t), y(t))$:

(3.6a) $$v_s + \tfrac{1}{2}\delta\Delta_x v + L^\varepsilon v = 0 \quad \text{in } Q,$$

where Δ_x is the Laplace operator in the x-variables, L^ε is the operator in (2.2) and $Q = (0, T) \times D \times R^m$. Define $\partial^* Q$ as $\{T\} \times D \times R^m$, together with the boundary points $(s, x, y) \in \Gamma^+$, with $0 < s < T$, and let $\psi(s, x, y)$ be a function continuous on ∂Q. By assumption (2.4b), (3.6a) is uniformly parabolic and, therefore, (see, for instance, Friedman [9, vol. I, p. 147]), the solution of (3.6a), satisfying the boundary condition

(3.6b) $$v(s, x, y) = \psi(s, x, y) \quad \text{on } \partial^* Q,$$

is

$$v(s, x, y) = E_{s,x,y}\psi(\theta^\delta, x^\delta(\theta^\delta), y(\theta^\delta)).$$

In particular, let $\psi = \psi_k (k = 1, 2, \cdots)$ be continuous bounded functions which on $\partial^* Q$ satisfy

$$\begin{aligned} \psi_k(s, x, y) &= 1 \quad \text{if } (s, x, y) \in \Gamma_T^+, \\ &= 0 \quad \text{if } (s, x, y) \in \{T\} \times D \times R^m \text{ and } d(x, \partial D) > 1/k, \\ 0 \leqq \psi_k &\leqq 1 \quad \text{if } (s, x, y) \in \{T\} \times D \times R^m \text{ and } d(x, \partial D) \leqq 1/k, \end{aligned}$$

(where d = distance), and which also satisfy

$$(*) \qquad |\psi_k - \psi_l| \to 0 \quad \text{as } k, l \to \infty$$

uniformly on compact subsets of $\bar{Q}$. Then

$$v_k^\delta(s, x, y) = E_{s,x,y}\psi_k(\theta^\delta, x^\delta(\theta^\delta), y(\theta^\delta))$$

satisfies (3.6a, b) with $\psi = \psi_k$. By Lemma 3.4, the continuity of ψ_k, and the dominated convergence theorem,

$$(**) \qquad v_k^\delta(s, x, y) \to E_{s,x,y}\psi_k(\theta, x(\theta), y(\theta)) \equiv q_k(s, x, y) \quad \text{as } \delta \to 0,$$

where $(x(t), y(t))$ is the solution of (2.1) with initial condition (3.3′), and $\theta = \min(T, \tau)$. Furthermore, $q_k(s, x, y)$ satisfies (3.6b) with $\psi = \psi_k$, and it is a distribution solution of (3.1), since

$$\int_Q (-\phi_t + L^{\varepsilon^*}\phi) q_k \, dQ = \lim_{\delta \to 0} \int_Q \left(-\phi_t + \frac{\delta}{2}\Delta_x \phi + L^{\varepsilon^*}\phi\right) v_k^\delta \, dQ = 0,$$

for any "test function" $\phi \in C_0^\infty(Q)$. Finally, (3.5) is proved in the same manner, by observing that

$$q(s, x, y) = \lim_{k \to \infty} q_k(s, x, y)$$

almost everywhere in Q. By hypoellipticity, q is a smooth solution of (3.1) (almost everywhere) in Q.

Since, in particular, q is continuous in Q, to complete the proof of Theorem 3.2 it only remains to show the continuity of q on the "accessible" boundary points of Q. Let us assume first that $z_0 = (s_0, x_0, y_0)$ is a point in Γ^+, with $s_0 < T$. Then $P_{z_0}(\tau \geqq T) = 0$. By the strong Feller property (see Remark 2.5), P_z is weakly continuous in $z = (s, x, y)$, and therefore,

$$\limsup_{\substack{z \to z_0 \\ z \in Q}} P_z(\tau \geqq T) \leqq P_{z_0}(\tau \geqq T) = 0.$$

Since $1 - q(z) = P_z(\tau > T) \leqq P_z(\tau \geqq T)$, it follows that $q(z) \to 1$ as $z \to z_0, z \in Q$. If $z = (s, x, y) \in \Gamma^+$ with $s \leqq T$, then $q(z) = 1$, and therefore, the latter limit also holds when $z \to z_0$, with $z \in Q_1$. Similarly, if $z_0 = (T, x_0, y_0)$ is a point on $\{T\} \times D \times R^m$, then $q(z_0) = P_{z_0}(\tau \leqq T) = 0$, so that

$$\limsup_{\substack{z \to z_0 \\ z \in Q}} P_z(\tau \leqq T) \leqq P_{z_0}(\tau \leqq T) = 0.$$

That is, $q(z) \to 0$ as $z \to z_0$, with $z \in Q$ (or Q_1). This completes the proof of Theorem 3.2. □

4. Associated control problems. We shall return now to the original equation (1.1) for the process $(x(t), y(t)) = (x^\varepsilon(t), y^\varepsilon(t))$. In Theorem 3.2 we saw that the exit probability $q^\varepsilon(s, x, y)$ is a smooth solution of the boundary value problem (3.1). Now a direct computation shows that the function

$$I^\varepsilon(s, x, y) = -\varepsilon \log q^\varepsilon(s, x, y)$$

satisfies the problem

$$\begin{aligned} I_s^\varepsilon + L^\varepsilon I^\varepsilon - \tfrac{1}{2} I_y^{\varepsilon *} a(s, y) I_y^\varepsilon &= 0 \quad \text{in } Q, \\ I^\varepsilon(s, x, y) &= 0 \quad \text{on } \Gamma_T^+, \\ &= \infty \quad \text{on } \{T\} \times D \times R^m; \end{aligned} \tag{4.1}$$

(the latter means, of course, that $I^\varepsilon(s, x, y) \to \infty$ as $s \to T$-, with $(x, y) \in D \times R^m$), where L^ε is the operator in (2.2).

We can write (4.1) as

$$I_s^\varepsilon + \frac{\varepsilon}{2} \operatorname{Tr}(a I_{yy}^\varepsilon) + F^* I_x^\varepsilon + H(s, y, I_y^\varepsilon) = 0, \tag{4.2}$$

where

$$H(s, y, p) = b(s, y)^* p - \tfrac{1}{2} p^* a(s, y) p. \tag{4.3}$$

Let $K(s, y, u)$ be the "dual" of $H(s, y, p)$: For $p, u \in R^m$,

$$\text{(a)} \qquad K(s, y, u) = \max_p (H(s, y, p) - p^* u), \tag{4.4}$$

$$\text{(b)} \qquad H(s, y, p) = \min_u (K(s, y, u) + p^* u).$$

Using (4.3) in the definition of K, we obtain

$$K(s, y, u) = \tfrac{1}{2}(b(s, y) - u)^* a(s, y)^{-1} (b(s, y) - u). \tag{4.5}$$

Setting (formally) $\varepsilon = 0$ in (4.2), we get the dynamic programming equation [7, Chapt. 4]

$$I_s + F(s, x, y)^* I_x + \min_u (K(s, y, u) + I_y^* u) = 0,$$

or

$$I_s + F(s, x, y)^* I_x + H(s, y, I_y) = 0, \tag{4.6}$$

for the (deterministic) control problem with system equations

$$\text{(a)} \qquad \begin{aligned} dx^0 &= F(t, x^0(t), y^0(t))\, dt, \\ dy^0 &= u(t)\, dt, \qquad (x^0(s), y^0(s)) = (x, y), \end{aligned} \tag{4.7}$$

and the value function

$$\text{(b)} \qquad I(s, x, y) = \inf_{u \in U} \int_s^\theta K(t, y^0(t), u(t))\, dt,$$

where θ is the exit time of $x^0(t)$ from D, and $U = U(s, x, y)$ is the collection of continuous functions u for which $\theta \leqq T$. We assume that U is not empty, and furthermore, that $(\theta, x^0(\theta), y^0(\theta)) \in \Gamma^+$.

Now, in terms of the function $I^\varepsilon = -\varepsilon \log q^\varepsilon$ in (4.1)–(4.2), the statement (1.5) to prove becomes

$$\lim_{\varepsilon \to 0} I^\varepsilon = I. \tag{4.8}$$

Following Fleming [5], to prove (4.8) we introduce a new (stochastic) control problem whose dynamic programming equation is precisely (4.2), for $\varepsilon > 0$. Before doing this, let us recall (see the remark about notation in the Introduction) that random variables are referred to some underlying probability space $(\Omega, \mathcal{F}, P)$. Then, if $\{\mathcal{F}_t\}$ is an increasing family of sub-σ-algebras of $\mathcal{F}$, a process $x(t)$ is said to be nonanticipative with respect to $\{\mathcal{F}_t\}$ if $x(t)$ is $\mathcal{F}_t$-measurable for each t. In (4.9) below, we assume that an increasing family of σ-algebras $\mathcal{F}_t \subset \mathcal{F}$, $0 \leqq t \leqq T$, is given and the processes $(\eta^\varepsilon(t), \xi^\varepsilon(t))$, $v(t)$ and the Wiener process $W(t)$ are nonanticipative with respect to $\{\mathcal{F}_t\}$.

(4.9) *A stochastic control problem.*

(a) system equations:

$$d\eta^\varepsilon = F(t, \eta^\varepsilon(t), \xi^\varepsilon(t))\, dt,$$

$$d\xi^\varepsilon = v(t)\, dt + \sqrt{\varepsilon}\sigma(t, \xi^\varepsilon(t))\, dW,$$

with $$(\eta^\varepsilon(s), \xi^\varepsilon(s)) = (x, y) \in D \times R^m.$$

(b) Admissible controls: the collection $U' = U'(s, x, y)$ of nonanticipative controls $v(t)$, such that

$$E \int_s^T |v(t)|^2\, dt < \infty,$$

and the corresponding assumption (2.4d) holds for $(\eta^\varepsilon, \xi^\varepsilon)$.

(c) Optimal expected system performance:

$$J^\varepsilon(s, x, y) = \min_{v \in U'} E\left\{ \int_s^{\theta^\varepsilon} K(t, \xi^\varepsilon(t), v(t))\, dt + \Phi(\theta^\varepsilon, \eta^\varepsilon(\theta^\varepsilon)) \right\},$$

where θ^ε is the minimum of T and the exit time from D of $\eta^\varepsilon(t)$, and $\Phi(s, x)$ is a bounded, nonnegative, Lipschitz function such that $\Phi = 0$ on points (s, x) for which $(s, x, y) \in \Gamma^+$.

With this notation, we have that J^ε satisfies the dynamic programming equation (4.2) on Q and the boundary condition

$$J^\varepsilon(s, x, y) = \Phi(s, x) \quad \text{on } \partial^* Q,$$

where

$$\partial^* Q = (\{T\} \times D \times R^m) \cup \Gamma_T^+.$$

To see this, let us consider the system (4.9a) with $v(t) = b(t, \xi^\varepsilon(t))$. Then the same proof as for Theorem (3.2) shows that the function

$$g^\varepsilon(s, x, y) = E_{s,x,y} \exp[-\Phi(\theta^\varepsilon, \eta^\varepsilon(\theta^\varepsilon))/\varepsilon]$$

is a smooth solution of the backward equation (3.1) in Q, is continuous in $\partial^* Q$, and satisfies the boundary condition

$$g^\varepsilon(s, x, y) = \exp[-\Phi(s, x)/\varepsilon] \quad \text{on } \partial^* Q.$$

Then, a direct calculation shows that

$$J^\varepsilon(s, x, y) = -\varepsilon \log g^\varepsilon(s, x, y)$$

is a smooth solution of (4.2) in Q and satisfies the boundary condition $J^\varepsilon = \Phi$ on $\partial^* Q$. Furthermore, from the verification theorem in dynamic programming [7, p. 159], it follows then that J^ε is indeed given as in (4.9c). In particular, we can write

$$J^\varepsilon(s, x, y) = E_{s,x,y} G^\varepsilon \quad \text{for all } (s, x, y) \in Q, \tag{4.10}$$

where

$$G^\varepsilon = \int_s^{\theta^\varepsilon} K(t, \xi^\varepsilon(t), v^\varepsilon(t))\, dt + \Phi(\theta^\varepsilon, \eta^\varepsilon(\theta^\varepsilon)); \tag{4.11}$$

$v^\varepsilon(t)$ is a nonanticipative optimal control which can be chosen in feedback form as

$$\begin{aligned} v^\varepsilon(t) &= V^\varepsilon(t, \eta^\varepsilon(t), \xi^\varepsilon(t)) \quad \text{for } s \leqq t \leqq \theta^\varepsilon, \\ &= 0 \qquad\qquad\qquad\qquad\quad \text{for } t > \theta^\varepsilon, \end{aligned} \tag{4.12}$$

with

$$\begin{aligned} v^\varepsilon(s, x, y) &= H_p(s, y, J_y^\varepsilon(s, x, y)) \\ &= b(s, y) - a(s, y) J_y^\varepsilon(s, x, y). \end{aligned}$$

Notice that, by the boundedness of b and a (assumption (2.4a)) and the definition of $J^\varepsilon = -\varepsilon \log g^\varepsilon$, it follows that v^ε is indeed an admissible control.

5. Proof of the main theorem. We are now ready to prove statement (1.5) = (4.8). We shall prove that

$$\text{(a)} \qquad \limsup_{\varepsilon \to 0} I^\varepsilon \leqq I, \tag{5.1}$$

$$\text{(b)} \qquad \liminf_{\varepsilon \to 0} I^\varepsilon \geqq I.$$

First, we need the following

LEMMA 5.2. *Let $T_1 < T$. Then for $s \leqq T_1$, there exists a constant C such that*

$$q^\varepsilon(s, x, y) \geqq e^{-C/\varepsilon},$$

or equivalently, $I^\varepsilon(s, x, y) \leqq C$, for all ε sufficiently small.

Proof. Let $(x^\varepsilon(t), y^\varepsilon(t))$ and $(x^0(t), y^0(t))$ be the processes given by (1.1) and (4.7a), respectively. By the first Ventsel–Friedlin estimate [18, Thm. 1.1], for any $\delta_1 > 0$, there exists $\varepsilon_0 > 0$ and a constant C such that

$$P(\|y^\varepsilon - y^0\|_T < \delta_1) > e^{-C/\varepsilon} \quad \text{for all } 0 < \varepsilon < \varepsilon_0.$$

(This result in Ventsel–Friedlin's paper is given in terms of the integrals defining $I(s, x, y)$ in (4.7b).) Now, writing (1.1) and (4.7a) in integrated form for $x^\varepsilon(t)$ and $x^0(t)$, we see that, for any $\delta > 0$, there exists $\delta_1 > 0$ (in fact, we can take $\delta_1 = \delta / k(T - s)$, where k is a Lipschitz constant for F) such that

$$\|y^\varepsilon - y^0\|_T < \delta_1 \Rightarrow \|x^\varepsilon - x^0\|_T < \delta.$$

On the other hand, by the definitions of U and Γ^+, we can assume that, for δ sufficiently small, the control u in (4.7a) is such that $x^0(t) \in \partial D_\delta$ from some $s < t \leqq T$, where D_δ

denotes a δ-neighborhood of D. In such a case,

$$\|x^\varepsilon - x^0\|_T < \delta \Rightarrow \tau^\varepsilon \leqq T.$$

Combining these results, we get:

$$q^\varepsilon = P(\tau^\varepsilon \leqq T) \geqq P(\|x^\varepsilon - x^0\|_T < \delta)$$
$$\geqq P(\|y^\varepsilon - y^0\|_T < \delta_1) > \exp(-C/\varepsilon),$$

for all $0 < \varepsilon < \varepsilon_0$. □

Proof of (5.1a). Consider the controlled systems $(x^0(t), y^0(t))$, $(\eta^\varepsilon(t), \xi^\varepsilon(t))$ in (4.7a) and (4.9a) with $u = v \in U$ such that $\theta < T$. Let T' be such that $\theta < T' < T$, and let T^ε be the minimum of T' and $\tau(\eta^\varepsilon)$ = exit time from D of $\eta^\varepsilon(t)$. Since $T^\varepsilon \leqq T' < T$, it follows from Theorem 3.2 and Dynkin's formula that the process $I^\varepsilon(t, \eta^\varepsilon(t), \xi^\varepsilon(t))$ satisfies

$$I^\varepsilon(s, x, y) = -E\int_s^{T^\varepsilon} \left(I_t^\varepsilon + \frac{\varepsilon}{2} \operatorname{Tr}(aI_{yy}^\varepsilon) + F * I_y^\varepsilon \right) dt + EI^\varepsilon(T^\varepsilon, \eta^\varepsilon(T^\varepsilon), \xi^\varepsilon(T^\varepsilon))$$

$$(*) \qquad = E\int_s^{T^\varepsilon} (H(t, \xi^\varepsilon(t), I_y^\varepsilon) - u(t) * I_y^\varepsilon)\, dt + EI^\varepsilon(T^\varepsilon, \eta^\varepsilon(T^\varepsilon), \xi^\varepsilon(T^\varepsilon))$$

$$\leqq E\int_s^{T^\varepsilon} K(t, \xi^\varepsilon(t), u(t))\, dt + EI^\varepsilon(T^\varepsilon, \eta^\varepsilon(T^\varepsilon), \xi^\varepsilon(T^\varepsilon)).$$

Here, we have used (4.2)–(4.3) with $b = u$, and (4.4a). On the other hand, since $u = v$, it can be seen that

(5.3) (a) $$\|\eta^\varepsilon - x^0\|_t + \|\xi^\varepsilon - y^0\|_t \to 0,$$

(b) $$T^\varepsilon \to \theta \quad (\text{since } \tau(\eta^\varepsilon) \to \theta),$$

(c) $$(\eta^\varepsilon(T^\varepsilon), \xi^\varepsilon(T^\varepsilon)) \to (x^0(\theta), y^0(\theta))$$

in probability, as $\varepsilon \to 0$. Since $I^\varepsilon = 0$ on Γ_T^+ (see (4.1)), we can obtain from Lemma 5.2 and (5.3) that

$$EI^\varepsilon(T^\varepsilon, \eta^\varepsilon(T^\varepsilon), \xi^\varepsilon(T^\varepsilon)) \to 0 \quad \text{as } \varepsilon \to 0,$$

and from (*),

$$\limsup_{\varepsilon \to 0} I^\varepsilon \leqq \int_s^\theta K(t, y^0(t), u(t))\, dt.$$

Finally, by the definition of I, we obtain (5.1a). □

Before proving (5.1b), let us consider the "deterministic analogue" ($\varepsilon = 0$) of the stochastic control problem (4.9), namely:

(5.4) (a) system equations:

$$d\eta^0 = F(t, \eta^0(t), \xi^0(t))\, dt,$$
$$d\xi^0 = v(t)\, dt,$$

with $(\eta^0(s), \xi^0(s)) = (x, y) \in D \times R^m$.

(b) admissible controls: The collection $U^0 = U^0(s, x, y)$ of continuous functions $v(t)$ such that

$$\int_s^T |v(t)|^2\, dt < \infty,$$

and the corresponding trajectory $(t, \eta^0(t), \xi^0(t))$ exits $[0, \infty) \times D \times R^m$ through a point in Γ^+, if exit occurs before or at time T.

(c) performance criterion:

$$J^0(s, x, y) = \min_{v \in U^0} \int_s^{\theta^0} K(t, \xi^0(t), v(t))\, dt + \Phi(\theta^0, \eta^0(\theta^0)),$$

where θ^0 is the minimum of T, $\tau(\eta^0)$ is the exit time from D of $\eta^0(t)$, and $\Phi(s, x)$ is the function in (4.9c). We define $J^0 = \Phi$ on $\partial^* Q$.

If we fix the control function v, the system (η^0, ξ^0) in (5.4a) is continuous in the initial state (x, y). Then it can be shown that:

(5.5) If (s, x_n, y_n) is a sequence in $Q \cup \Gamma_T^+$ such that $(s, x_n, y_n) \to (s, x, y) \in \Gamma_T^+$, then $J^0(s, x_n, y_n) \to J^0(s, x, y) = \Phi(s, x)$.

LEMMA 5.6. *For any* $(s, x, y) \in Q$,

$$\liminf_{\varepsilon \to 0} J^\varepsilon(s, x, y) \geqq J^0(s, x, y).$$

Proof. Let $v^\varepsilon(t)$ be the optimal feedback control for problem (4.9) given in (4.12), and write J^ε as in (4.10)–(4.11). We shall prove first that

$$J^0(s, x, y) \leqq G^\varepsilon + A^\varepsilon + B^\varepsilon, \tag{5.7}$$

where $A^\varepsilon, B^\varepsilon$ are defined below, so that the lemma will be proved if we can show that

$$\text{(a)} \quad \liminf_{\varepsilon \to 0} EA^\varepsilon = 0 \tag{5.8}$$

$$\text{(b)} \quad \liminf_{\varepsilon \to 0} EB^\varepsilon = 0.$$

By the assumptions (2.4a, b) on $b(s, y)$ and $a(s, y)$, and definition (4.5) of K, there exist constants c_1, c_2 with $c_1 > 0$, such that

$$K(s, y, v) \geqq c_1 |v|^2 - c_2.$$

From this and (4.11), we see that there exists a constant k_1 such that

$$\int_s^{\theta^\varepsilon} |v^\varepsilon(t)|^2\, dt > k_1 \Rightarrow J^0(s, x, y) \leqq G^\varepsilon,$$

and (5.7) follows in this case. Let us show now that

$$\int_s^{\theta^\varepsilon} |v^\varepsilon(t)|^2\, dt \leqq k_1. \tag{5.9}$$

Let $(\psi^\varepsilon(t), \phi^\varepsilon(t))$ be the process satisfying

$$d\psi^\varepsilon = F(t, \psi^\varepsilon(t), \phi^\varepsilon(t))\, dt,$$

$$d\phi^\varepsilon = v^\varepsilon(t)\, dt,$$

with $(\psi^\varepsilon(s), \phi^\varepsilon(s)) = (x, y) \in D \times R^m$. Comparing this process with (4.9a), with $v = v^\varepsilon$, we can see that, as $\varepsilon \to 0$,

$$(*) \qquad \|\eta^\varepsilon - \psi^\varepsilon\|_T + \|\xi^\varepsilon - \phi^\varepsilon\|_T \to 0$$

in probability. From this, it can be obtained that

$$(**) \qquad \liminf_{\varepsilon \to 0} (E\|\eta^\varepsilon - \psi^\varepsilon\|_T + E\|\xi^\varepsilon - \phi^\varepsilon\|_T) = 0.$$

Furthermore, by definition (5.4c) of J^0, a stochastic version of the "principle of optimality" in dynamic programming (cf. [2, p. 264]) yields:

$$J^0(s, x, y) \leqq \int_s^{\theta^\varepsilon} K(t, \phi^\varepsilon(t), v^\varepsilon(t))\, dt + J^0(\theta^\varepsilon, \psi^\varepsilon(\theta^\varepsilon), \phi^\varepsilon(\theta^\varepsilon)),$$

where θ^ε is the random time in (4.11). Adding and subtracting G^ε on the right side, and then taking absolute values, we get (5.7), where

$$A^\varepsilon = \int_s^{\theta^\varepsilon} |K(t, \phi^\varepsilon(t), v^\varepsilon(t)) - K(t, \xi^\varepsilon(t), v^\varepsilon(t))|\, dt,$$

$$B^\varepsilon = |J^0(\theta^\varepsilon, \psi^\varepsilon(\theta^\varepsilon), \phi^\varepsilon(\theta^\varepsilon)) - \Phi(\theta^\varepsilon, \eta^\varepsilon(\theta^\varepsilon))|.$$

Now, to prove (5.8a) note that by the boundedness of b, a^{-1}, and its first derivatives (assumptions (2.4)), the definition (4.5) of K, and the mean value theorem, there exists a constant C such that

$$|K(t, y, u) - K(t, y', u)| \leqq C(1 + |u|^2)|y - y'|.$$

Therefore,

$$A^\varepsilon \leqq C\|\phi^\varepsilon - \xi^\varepsilon\|_T \int_s^{\theta^\varepsilon} (1 + |v^\varepsilon(t)|^2)\, dt \leqq k_2 \|\phi^\varepsilon - \xi^\varepsilon\|_T,$$

for some constant k_2. Now (5.8a) follows from (**).

To prove (5.8b) note that, by definition (5.4c) of J_0, if $\theta^\varepsilon = T$, then $J^0 = \Phi$ and, therefore, (5.8b) follows from the Lipschitz assumption on Φ, (*), and (**). On the other hand, if $\theta^\varepsilon = \tau(\eta^\varepsilon)$, then $(\theta^\varepsilon, \eta^\varepsilon(\theta^\varepsilon), \xi^\varepsilon(\theta^\varepsilon)) \in \Gamma_T^+$, and therefore, again from (5.4c) we have:

$$B^\varepsilon = |J^0(\theta^\varepsilon, \psi^\varepsilon(\theta^\varepsilon), \phi^\varepsilon(\theta^\varepsilon)) - J^0(\theta^\varepsilon, \eta^\varepsilon(\theta^\varepsilon), \xi^\varepsilon(\theta^\varepsilon))|.$$

Now, from (*), (5.5), and the boundedness of $\Phi = J^0$ on $\partial^* D$; we get (5.8b). This completes the proof of the lemma. □

We shall now complete the proof of (5.1).

Proof of (5.1b). Let $\psi(x)$ be a nonnegative Lipschitz function such that $\psi(x) > 0$ if $x \in D$, and $\psi(x) = 0$ for $x \in \partial D$. For each $M > 0$, define

$$q_M^\varepsilon(s, x, y) = E_{s,x,y} \exp[-M\psi(x^\varepsilon(\tau^\varepsilon \wedge T))/\varepsilon],$$

and let $I_M^\varepsilon = -\varepsilon \log q_M^\varepsilon$. Then $q^\varepsilon \leqq q_M^\varepsilon$, or equivalently, $I^\varepsilon \geqq I_M^\varepsilon$.

The function $q_M^\varepsilon(s, x, y)$ satisfies (3.1) with the boundary conditions

$$q_M^\varepsilon(s, x, y) = 1 \quad \text{on } \Gamma_T^+$$

$$= \exp(-M\psi(x)/\varepsilon) \quad \text{on } \{T\} \times D \times R^m,$$

while I_M^ε satisfies (4.1) in Q with boundary conditions

$$I_M^\varepsilon(s, x, y) = 0 \quad \text{on } \Gamma_T^+$$
$$= M\psi(x) \quad \text{on } \{T\} \times D \times R^m.$$

Moreover, $J^\varepsilon = I_M^\varepsilon$ if we take Φ in (4.9c), so that

$$\Phi = I_M^\varepsilon \quad \text{on } \partial^* Q.$$

For this Φ, let $I_M^0(s, x, y)$ be the corresponding performance criterion (5.4c). Then, for fixed M and $(s, x, y) \in Q$, Lemma 5.6 implies that

$$I_M^0(s, x, y) \leqq \liminf_{\varepsilon \to 0} I_M^\varepsilon(s, x, y).$$

On the other hand, comparing the deterministic control problems (4.7) and (5.4), we see that

$$I(s, x, y) \leqq \liminf_{M \to \infty} I_M^0(s, x, y).$$

Combining the last two inequalities with the fact that $I_M^\varepsilon \leqq I^\varepsilon$, we obtain (5.1b). □

Acknowledgment. The author wishes to express his gratitude to Professor Wendell H. Fleming for many useful comments and his unlimited cooperation during the preparation of this work. Prof. Fleming and a referee pointed out several mistakes in the original manuscript. The author is also indebted to Diego B. Hernández-Castaños for reading and criticizing a previous version of this paper.

REFERENCES

[1] J. M. C. CLARK, *Stochastic differential equations on manifolds*, in Geometric Methods in System Theory, D. Q. Mayne and R. W. Brockett, eds., Reidel Publ. Co., Dordrecht, Holland, 1973.
[2] W. H. FLEMING, *Duality and a priori estimates in Markovian optimization problems*, J. Math. Anal. Appl., 16 (1966), pp. 254–279.
[3] ———, *Stochastic control for small noise intensities*, SIAM J. Control, 9 (1971), pp. 473–517.
[4] ———, *Stochastically perturbed dynamical systems*, Rocky Mountain J. Math., 4 (1974), pp. 407–433.
[5] ———, *Inclusion probability and optimal stochastic control*, IRIA Seminars Review, 1977.
[6] ———, *Exit probabilities and optimal stochastic control*, Appl. Math. Optim., 4 (1978), pp. 329–346.
[7] W. H. FLEMING AND R. W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.
[8] B. FRIEDLAND, F. E. THAU AND P. E. SARACHIK, *Stability problems in randomly-excited dynamic systems*, in Proc. Joint Autom. Control Conf., Seattle, Washington, 1966, pp. 848–861.
[9] A. FRIEDMAN, *Stochastic Differential Equations and Applications*, Vol. II, Academic Press, New York, 1976.
[10] E. GHANDOUR. *Fermat's principle in a stochastic medium*, in Differential Games and Control Theory II, E. O. Roxin, P. T. Liu, and R. L. Sternberg, eds., Marcel Dekker, New York, 1977.
[11] L. HORMANDER, *Hypoelliptic second order differential equations*, Acta Math., 119 (1968), pp. 147–171.
[12] D. LUDWIG, *Persistence of dynamical systems under random perturbations*, SIAM Rev., 17 (1975), pp. 605–639.
[13] B. J. MATKWOSKY AND Z. SCHUSS, *The exit problem for randomly perturbed dynamical systems*, SIAM J. Appl. Math., 33 (1977), pp. 365–382.
[14] O. A. OLEINIK AND E. V. RADKEVICH, *On local smoothness of generalized solutions and hypoellipticity of second order differential equations*, Russian Math. Surveys, 26 (1971), pp. 139–156.
[15] M. SCHECHTER, *Modern Methods in Partial Differential Equations*, McGraw-Hill, New York, 1976.
[16] D. STROOCK AND S. R. S. VARADHAN, *On degenerate elliptic-parabolic operators of second order and their associated diffusions*, Comm. Pure Appl. Math., 25 (1972), pp. 651–713.
[17] L. J. VAN MELLAERT AND P. DORATO, *Numerical solution of an optimal control problem with a probability criterion*, IEEE Trans. Automat. Control, AC-17 (1972), pp. 543–546.
[18] A. D. VENTSEL AND M. I. FRIEDLIN, *On small random perturbations of dynamical systems*, Russian Math. Surveys, 25 (1970), pp. 1–55.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 1, January 1981

0363-0129/81/1901-0005 $01.00/0

EXTREME POINTS AND BASIC FEASIBLE SOLUTIONS IN CONTINUOUS TIME LINEAR PROGRAMMING*

ANDRÉ F. PEROLD†

Abstract. This paper studies the extreme points arising in continuous time linear programming. The main result is for the case of constant coefficients where all so-called right analytic extreme points are characterized, analogously to the result for linear programming, in terms of certain full rank conditions. Examples of continuous time linear programs with time-varying constraints are given to show that this kind of characterization cannot hold in general.

1. Introduction. The general continuous time linear program is formulated as follows:

minimize

$$\int_0^T c(t)x(t)\,dt$$

subject to

$$B(t)x(t)+\int_0^t K(t,s)x(s)\,ds = b(t), \qquad x(t)\geqq 0, \quad t\in[0,T],$$

where $c(t)$, $b(t)$ and $B(t)$, $K(t,s)$ are given and are real vectors and matrices respectively. A special case of this is a linear optimal control problem with constraints on both the state and control variables:

minimize

$$\int_0^T \{c(t)x(t)+d(t)u(t)\}\,dt$$

subject to

$$\frac{d}{dt}x(t) = A(t)x(t)+B(t)u(t)+a(t),$$

$$0 = C(t)x(t)+D(t)u(t)+b(t),$$

$$x(t)\geqq 0, \quad u(t)\geqq 0, \quad t\in[0,T], \quad x(0) \text{ given.}$$

Continuous time linear programs are widely applicable to many real world situations, for example as intertemporal economic models of investment and planning (e.g., Bellman [2] and Dantzig [5]), and occur frequently in engineering applications (e.g., Teren [17]). They were first considered by Bellman [1], [2] in 1953 and have since been studied largely as linear programs in a function space with the emphasis on generalizing the simple but powerful results of linear programming (Dantzig [4]). The main results to date have been strong duality theorems, the first being obtained by

* Received by the editors September 12, 1979, and in final revised form April 1, 1980. This paper is part of the author's dissertation in the Department of Operations Research, Stanford University, Stanford, California. The research was supported in part by the 1975 University of the Witwatersrand Harry Hofmeyer Scholarship; the U.S. Department of Energy under Contract EY-76-0326 PA #18, the Office of Naval Research under Contract N00014-75-C-0267, and the National Science Foundation under Grants MCS76-81259A01, MCS76-20019 A01, and ENG77-06761 A01 at Stanford University.

† Graduate School of Business Administration, Harvard University, Boston, Massachusetts 02163.

Tyndall [19] in 1965, and subsequently strengthened by a number of authors, e.g., Grinold [9] and Levine and Pomerol [13]. Work has also been done on generalizing the simplex method itself to the continuous time case. In 1964 Dantzig [6] showed that the special case of the control theory formulation with no state variable constraints could be elegantly solved by an application of the Dantzig-Wolfe decomposition principle. The general case was considered first in 1954 by Lehman [12], and then pursued by Drews, Hartberger, and Segers [8] in 1970, and Teren [17] in 1977. Few clear results emerged from these papers, however, the major questions still remaining open.

In this paper we study continuous time linear programs from the point of view of examining the nature of their extreme point[1] solutions. The two fundamental results of linear programming that we wish to generalize are the following:

LEMMA 1. *If* $f: R^n \to R$ *is linear, and* f *is bounded below on* $P = \{x \in R^n : Ax = b, x \geqq 0\}$ *where* $A \in R^{m \times n}$ *and* $b \in R^m$, *then* f *attains its infimum on* P, *and moreover does so at an extreme point of* P.

LEMMA 2. $x \in P$ *is an extreme point of* P *if and only if* $A_{\cdot\beta}$ *has full column rank, where* $\beta = \{i : x_i > 0\}$.

The significance of Lemma 1 is that computational procedures can restrict their search to extreme points only. Lemma 2 in addition allows one to solve certain equations with respect to $A_{\cdot\beta}$[2] that allow one to determine an "adjacent" extreme point with an improved objective value. This is the heart of the simplex method, and forms the motivation for seeking similar results here.

Extreme points in linear programming are also called *basic feasible solutions* because of the abovementioned solvability of equations with respect to $A_{\cdot\beta}$. Both Lehman [12] and Drews, Hartberger and Segers [8] work with "basic feasible solutions" although these are vaguely defined, and moreover largely so in terms of a solvability requirement, no connection being made with extreme points. Teren [17] does likewise by imposing certain full rank conditions. These allow for only a restricted set of extreme points, none of which need be even near to optimality.

The first question we address in this paper is whether the optimum is indeed attained at an extreme point. Then we consider sufficient conditions for a feasible solution to be an extreme point, these depending on the a priori restrictions on the constraint coefficients and the class of admissible solutions, e.g., bounded measurable functions. The main and final result gives a complete characterization of extreme points in the case that $A(\cdot)$ and $K(\cdot, \cdot)$ are both constant and the admissable solutions are the space of so-called right analytic functions. This characterization is a generalization of the result for bang-bang control theory [11], and in addition yields the solvability properties required of basic feasible solutions. Examples are given to show that this kind of characterization cannot hold for general time-varying $A(\cdot)$ and $K(\cdot, \cdot)$. This result is fundamental in a sequel paper, Perold [14], where the question of moving from one extreme point to an improved one is considered.

1.1. Notation. Let $\underline{n}$ denote $\{1, 2, \cdots, n\}$.

For $A \in R^{m \times n}$ and $\alpha \subset \underline{m}$, $\beta \subset \underline{n}$, let A_α. denote the submatrix of A whose rows are indexed by α; let $A_{\cdot\beta}$ denote the submatrix of A whose columns are indexed by β.

Matrices will be denoted by upper case letters and vectors by lower case letters. The distinction between row and column vectors will be clear from the context.

$\|\cdot\|$ will denote the Euclidean norm.

---

[1] An *extreme point* of a (convex) set is one that cannot be written as a proper convex combination of two other points in the set.

[2] If $A_{\cdot\beta}$ is not square, it is usually appropriately augmented so as to be square and nonsingular.

If $I = (t', t'')$ is an open interval then $\tilde{I}$ will denote the interval $[t', t'')$.

$L_\infty^n[0, T]$ is the space of real-valued, Lebesgue measurable, essentially bounded functions from $[0, T]$ into R^n. $L_1^n[0, T]$ is the space of real-valued, Lebesgue integrable functions from $[0, T]$ into R^n. When the time interval is clear from the context these will be written as L_∞^n and L_1^n respectively.

The numbers (1), (2), $\cdots$ refer to equations, definitions, etc., in the main body of the paper. (A1), $\cdots$, (B1), $\cdots$ refer to the like in Appendices A and B.

2. Extreme points as optimal solutions. For the remainder of this paper, we shall let X denote the space of admissible functions $x:[0, T] \to R^n$. In order that the term $\int_0^t K(t, s)x(s)\,ds$ make sense, X will be assumed to be contained in L_1^n. We shall let $f: X \to R$ be the objective functional, i.e.,

$$f(x) = \int_0^T c(t)x(t)\,dt,$$

and $P(X)$ the constraint set, i.e.,

$$P(X) = \left\{ x \in X: A(t)x(t) + \int_0^t K(t, s)x(s)\,ds = b(t),\ x(t) \geqq 0,\ t \in [0, T] \right\}.$$

$P(X)$ is easily seen to be convex, and will be assumed nonempty.

The result that would be most desirable is the following: if X is a given space of "nice" functions, and if a continuous linear functional f is bounded below on $P(X)$, then f attains its infimum on $P(X)$ and moreover does so at an extreme point of $P(X)$. However, without severe preconditions on the coefficients defining the problem, the only case when this appears possible is when $X = L_\infty^n$ and $P(L_\infty^n)$ is bounded. One complicating factor is that while the notion of extreme points is purely algebraic, one seems to require heavy topological machinery merely to establish their existence.

THEOREM 3. *If the components of $c(\cdot)$, $b(\cdot)$, $B(\cdot)$, and $K(\cdot, \cdot)$ are all in L_∞, and there is an $M > 0$ such that $x \in P(L_\infty^n) \Rightarrow \|x(t)\| \leqq M$ a.e., then $f(\cdot)$ attains its infimum at an extreme point of $P(L_\infty)$.*

Proof. Let Y be the space L_1^n equipped with the weak topology. As established in [9, p. 40], $P(L_\infty^n)$ is compact in Y. Further, since $c \in L_\infty^n$, f is a continuous linear functional on Y. The proof is made complete by noting that a continuous linear functional on a compact set in a locally convex Hausdorff space attains its infimum at an extreme point of that set (see, e.g., [10, p. 74]). □

In practice we would like to have optimal solutions that are more manageable than general measurable solutions, for example piecewise analytic solutions having only finitely many breakpoints. Theorem 3 unfortunately is the best statement available, and it is still an open question whether it can be improved upon even in very special cases. Even if we know that the optimum has, say, a piecewise analytic solution, there is no guaranteee that there is a piecewise analytic extreme point solution. However, there is a motivation for continuing the study of extreme points in more useful spaces, given by the following simple result.

PROPOSITION 4. *If $f: X \to R$ is concave and x is the unique minimizer of f over $P(X)$, then x is an extreme point of $P(X)$.*

Proof. This follows immediately from the concavity of f and the definition of an extreme point. □

3. Characterizing extreme points. We begin by reviewing the proof of the characterization of extreme points in R^n given in Lemma 2. We have a given $x \in R^n$ satisfying

$$Ax = b, \qquad x \geqq 0.$$

The necessary and sufficient condition for x to be an extreme point is that $A_{\cdot\beta}$ have full column rank, where $\beta = \{i: x_i > 0\}$.

To show the sufficiency, we write $x = \lambda y + (1-\lambda)z$, for $\lambda \in (0, 1)$ and some y and z satisfying the constraints, and then show that $x = y = x$. Set $\alpha = \{i: x_i = 0\}$; then $x_\alpha = 0$, $y \geqq 0$ and $z \geqq 0$ imply $y_\alpha = z_\alpha = 0$. Therefore $A_{\cdot\beta}y_\beta = A_{\cdot\beta}z_\beta = A_{\cdot\beta}x_\beta = b$. Since $A_{\cdot\beta}$ has full column rank these equations have a unique solution, and we are done.

To show the necessity, assume that $A_{\cdot\beta}$ does not have full column rank. Then there exists a $y_\beta \neq 0$ such that $A_{\cdot\beta}y_\beta = 0$. Setting $y_\alpha = 0$ and noting that $x_\beta > 0$, we see that there is a $\theta > 0$ such that $x + \theta y \geqq 0$, $x - \theta y \geqq 0$. Hence, writing $x = \frac{1}{2}(x + \theta_y) + \frac{1}{2}(x - \theta y)$ shows that x is not an extreme point.

It is precisely these two steps that we shall try to mirror in the continuous-time case, namely

(i) being able to solve uniquely for the positive components when the remaining ones are held fixed at zero, and

(ii) being able to perturb the positive components to either side when they are not uniquely determined.

Our first result here is that uniqueness in the sense of (i) is a sufficient condition for an extreme point, although in general not a necessary condition.

LEMMA 5. *Let X be any space of admissible functions, and let $x \in P(X)$. Define Z_i, $i = 1, \cdots, n$ by $Z_i = \{t \in [0, T]: x_i(t) = 0\}$. Then x is an extreme point of $P(X)$ if x is the unique solution to*

$$\begin{aligned} &A(t)y(t) + \int_0^t K(t, s)y(s)\,ds = b(t), \qquad t \in [0, T], \\ &y_i(t) = 0, \qquad t \in Z_i, \quad i = 1, \cdots, n, \quad y \in X. \end{aligned} \tag{6}$$

Proof. The proof in R^n applies here almost identically. □

Example (A1) in Appendix A shows that the uniqueness condition of Lemma 5 need not be necessary. The main reason for this is that the restrictions $y_i(t) = 0$, $t \in Z_i$, can be redundant, as is the case in the example.

In order to obtain a "practically useful" characterization of extreme points, though, it is important that this uniqueness condition be necessary. To this end, we shall suitably restrict both the class of admissible solutions and the constraint coefficients. In the former, we shall work with solutions whose components are positive and zero over intervals of time, as opposed to more general measurable sets. Problems encountered in practice usually have solutions of this form, so that this condition will not be unduly restrictive. In the latter, we shall want to work with equations whose solutions can be determined by *solving forwards in time.* More precisely, the coefficients should be such that (6) has a unique solution if and only if

$$\begin{aligned} &A(t)y(t) + \int_0^t K(t, s)y(s)\,ds = b(t), \qquad t \in [0, \tau], \\ &y_i(t) = 0, \qquad t \in Z_i \cap [0, \tau], \quad i = 1, \cdots, n, \quad y \in X, \end{aligned} \tag{7}$$

has a unique solution for each $\tau \in [0, T]$. Without this property it becomes extremely difficult to obtain succinct uniqueness conditions on the coefficients such as full rank conditions. Examples (A2) and (A3) are cases where this property fails. In Example (A2), the slope of x at 0, and hence the whole solution, is determined only at time T by the restriction $x(T) = 0$. Example (A3) is an extreme point where x on $[0, 1)$ is determined by constraints that hold over $[1, 4)$.

Note that uniqueness in (7) for each τ is equivalent to requiring uniqueness over each interval $[\tau', \tau''] \subset [0, T]$, given the solution history $y(t)$ for $t \in [0, \tau')$. With the above scenario in mind, and assuming that we can sufficiently restrict the problem so that uniqueness becomes a necessary condition, the following restatement of the continuous-time linear program applies: choose a partition of $[0, T]$ into time intervals $\{(t_j', t_j'')\}$, and an associated partition of the variables $\{(\alpha_j, \beta_j)\}$, such that the restrictions $x_{\alpha_j}(t) = 0$, $t \in (t_j', t_j'')$, all j, uniquely determine x and yield the optimal solution. Moreover, for each j, given that x has been uniquely determined on $[0, t_j')$, the restriction $x_{\alpha_j}(t) = 0$ for $t \in [t_j', t_j'')$ serves only to determine $x_{\beta_j}(\cdot)$ on $[t_j', t_j'')$.

In the remainder of this paper we shall work with x being the space of right analytic functions, defined below.

DEFINITION 8. A function $g : [0, T] \to R$ will be called *right analytic* if for each $t \in [0, T)$, there is an $\varepsilon > 0$ and an analytic function $h : (t - \varepsilon, t + \varepsilon) \to R$ such that $g(s) = h(s)$ $\forall s \in [t, t + \varepsilon)$.

We shall let $\mathscr{A}_r$ denote the space of bounded right analytic functions on $[0, T]$. The required properties of these functions are established in Appendix B. Our motivation for choosing the class $\mathscr{A}_r$ is that it has been to date the largest class for which the local uniqueness-over-intervals result, Proposition 11 below, can be established.

The following is a key result in the subsequent analysis.

LEMMA 9. *Let $g : [0, T] \to R^n$ have right analytic components $g_i(\cdot)$, $i = 1, \cdots, n$. Then there exists a (possibly infinite) disjoint family of open intervals, $\{I_j\}$, such that $\cup \tilde{I}_j = [0, T)$, and such that for each interval I_j, each g_i satisfies*

(i) *g_i is analytic on I_j,*

(ii) *either $|g_i| > 0$ on I_j or $g_i = 0$ on I_j.*

Proof. See Appendix B. □

For a given $x \in P(\mathscr{A}_r^n)$ and its associated partition $\{I_j\}$, let

$$\alpha_j = \{i : x_i = 0 \text{ on } I_j\},$$

and

$$\beta_j = \{i : x_i > 0 \text{ on } I_j\}.$$

Let t_j' and t_j'' denote the endpoints of I_j. Then with the constraints (6), the equation for x_{β_j} on I_j becomes

$$B_{\cdot\beta_j}(t) x_{\beta_j}(t) + \int_{t_j'}^{t} K_{\cdot\beta_j}(t, s) x_{\cdot\beta_j}(s)\, ds = d^j(t), \tag{10}$$

for all $t \in [t_j', t_j'')$, where $d^j(t) = b(t) - \int_0^{t_j'} K(t, s) x(s)\, ds$.

PROPOSITION 11. *Let $x \in P(\mathscr{A}_r^n)$ and let $\{I_j\}$ be the associated partition of $[0, T]$ given by Lemma 9. If (10) has a unique right analytic solution on each interval $\tilde{I}_j$, then x is an extreme point of $P(\mathscr{A}_r^n)$.*

Proof. Suppose that x is not an extreme point. Then there exist y, $z \in P(\mathscr{A}_r)$ and $\lambda \in (0, 1)$ such that $x = \lambda y + (1 - \lambda) z$, and for some $t \in [0, T)$, $x(t) \neq y(t)$. By Lemma B8 there exist $0 \leqq r < s \leqq T$, such that $x = y$ on $[0, r)$ and $x \neq y$ on (r, s). By Lemma 9 there is an interval $I_j = (t_j', t_j'')$ of the partition such that $t_j' \leqq r < t_j''$. By analogy with the sufficiency argument presented for R^n, it is clear that on I_j, y_{β_j} satisfies

$$B_{\cdot\beta_j}(t) y_{\beta_j}(t) + \int_{t_j'}^{t} K_{\cdot\beta_j}(t, s) y_{\beta_j}(s)\, ds = e^j(t),$$

for all $t \in [t_j', t_j'']$, where

$$e^j(t) = b(t) - \int_0^{t_j'} K(t, s)y(s)\, ds.$$

However, since $y = x$ on $[0, r)$, it follows that $e^j(t) = d^j(t)$ on I_j. Thus y_{β_j} satisfies (10) on I_j. Since $y_{\alpha_j} = x_{\alpha_j} = 0$ on I_j, $y_{\beta_j} \neq x_{\beta_j}$ on $I_j \cap (r, s) \neq \varnothing$, contradicting the uniqueness hypothesis. □

Remark. In sum we have shown that being able to solve uniquely for the positive components locally is indeed a sufficient condition for a right analytic solution to be an extreme point. The important part in the proof played by the right analyticity was that when given x, $y \in P(\alpha_r)$ with $x \neq y$, we could find an earliest interval, I_j, in the partition, on which $x \neq y$. An example where we cannot find a first interval is the following. Let $x(t) = [t \sin (1/t)]^+$ (i.e., positive part), and $y(t) = t$ on $[0, 1]$. Clearly, in the partition of $[0, 1]$ induced by x, there is no first interval on which x and y differ. Note further that the theorem is false if the solution is right analytic but we choose a partition $\{I_j\}$ that does not satisfy $[0, T) = U_{j=1}^{\infty} \tilde{I}_j$. This can be seen in Example (A2) as follows:

Choose $x(t) = 1$ (not an extreme point), and choose the partition I_j with $I_j = (T/(j+1), T/j)$. Then $\bigcup_{j=1}^{\infty} [T/(j+1), T/j) = (0, T)$, and x is uniquely defined on each interval $[T/(j+1), T/j)$.

We can now proceed to find algebraic conditions on the coefficients that ensure unique solutions to equations of the type

$$D(t)x(t) + \int_{t'}^{t} L(t, s)x(s)\, ds = g(t), \qquad t \in [t', t''). \tag{12}$$

Since, by Example (A3), such conditions cannot in general also be necessary conditions for uniqueness, we shall confine ourselves to the case where the necessity has been established. This is the time-invariant case, i.e., when D and L are constants.

Before doing so, we remark briefly that in the event $D(t) = I$, (12) is a Volterra equation of the second kind, and it is well known that such equations always have unique solutions, provided that the coefficients $g_i(\cdot)$ and $L_{ij}(\cdot, \cdot)$ are in L_2. See for example [18, p. 10]. Thus if $D^{-1}(t)$ exists a.e. and $(D^{-1}g)_i$, $(D^{-1}L)_{ij} \in L_2$, we also obtain uniqueness. In general, however, $D(t)$ may be singular. This case has been studied by Dolezal [7], and in differential equation form by Silverman [16]. However, their work provides only a partial answer, and a general succinct uniqueness condition has, to our knowledge, yet to be discovered.

In the time-invariant case, (12) reads

$$Dx(t) + L \int_{t'}^{t} x(s)\, ds = g(t), \qquad t \in [t', t''). \tag{13}$$

This equation has been thoroughly studied in its more conventional differential equation form by a number of authors; see for example [3], [7], [16]. The uniqueness condition we require is the following.

LEMMA 14. *If the components of $g(\cdot)$ are analytic and x satisfies* (13), *then a necessary and sufficient condition for x to be the unique analytic solution is that there exist a scalar μ such that $\mu D + L$ has full column rank.*

Proof. See [3]. □

Remark. This full rank condition has several interesting interpretations. If D and L are square, then $\mu D + L$ has full rank iff $\det(\mu D + L)$ is not identically zero as a function

of μ.[3] On dividing by μ and setting $\varepsilon = 1/\mu$ the condition reads: $D + \varepsilon L$ has full column rank for all ε sufficiently small. If we replace (13) by its discrete time analogue using time intervals of stepsize ε, we obtain a block lower triangular coefficient matrix with each diagonal block being $D + \varepsilon L$. Clearly this block triangular matrix has full column rank iff $D + \varepsilon L$ has full column rank. Another interpretation can be made by taking Laplace transforms on both sides of (13) with dummy variable μ. The coefficient matrix of the Laplace transform of x so obtained is precisely $D + (1/\mu)L$.

In order to show that this full rank condition is also a necessary condition for a solution to be an extreme point, we require the following lemma.

LEMMA 15. *If there is no scalar μ such that $\mu D + L$ has full column rank, then for each $\tau > 0$, there exists a nontrivial analytic solution to the homogeneous equation*

$$Dx(t) + L\int_0^t x(s)\,ds = 0, \qquad t \geqq 0 \tag{16}$$

satisfying

$$\int_0^\tau x(s)\,ds = 0.$$

Proof. We use an argument similar to that given in [3, p. 418]. By assumption, for each μ there exists a nonzero (constant) vector φ_μ such that $(\mu D + L)\varphi_\mu = 0$.

If D has k columns, let $M > 2k$ be any integer and let $\mu_1, \cdots, \mu_M$ be any distinct scalars. Let G be the following $2k \times M$ matrix:

$$G = \begin{bmatrix} \varphi_{\mu_1} & \varphi_{\mu_2} & \cdots & \varphi_{\mu_M} \\ e^{\mu_1 \tau}\varphi_{\mu_1} & e^{\mu_2 \tau}\varphi_{\mu_2} & \cdots & e^{\mu_M \tau}\varphi_{\mu_M} \end{bmatrix}.$$

Since $M > 2k$, the columns of G are linearly dependent. Hence there exists a nonzero vector $\eta = (\eta_1, \cdots, \eta_M)$ such that $G_\eta = 0$.

Set

$$x(t) = \sum_{i=1}^{M} \eta_i \mu_i e^{\mu_i t}\varphi_{\mu_i}.$$

Then it is easily verified that x satisfies (16) and that $\int_0^\tau x(s)\,ds = 0$. Moreover, x is not identically zero. □

We can now obtain the main result of this paper.

THEOREM 17. (Characterization of right analytic extreme points.) *Let x be right analytic and satisfy*

$$Bx(t) + K\int_0^t x(s)\,ds = b(t), \qquad x(t) \geqq 0, \quad t \in [0, T], \tag{18}$$

where B and K are constant. Let $\{I_j\}$ be the associated partition of $[0, T]$ given by Lemma 9 and define

$$\alpha_j = \{i: x_i = 0 \text{ on } I_j\}, \qquad \beta_j = \{i: x_i > 0 \text{ on } I_j\}.$$

Then a necessary and sufficient condition for x to be an extreme point is that for each j, there exist a scalar μ such that $\mu B_{\cdot\beta_j} + K_{\cdot\beta_j}$ has full column rank.

[3] Note that $\det(\mu D + L)$ is a polynomial in μ. Hence it is zero either for all μ or for at most finitely many μ.

Proof. Sufficiency. By Lemma 14, x_{β_j} is uniquely determined on I_j. By Proposition 11, x is an extreme point.

Necessity. Suppose there is a j such that for all μ, $\mu B_{\cdot\beta_j} + K_{\cdot\beta_j}$ does not have full column rank. Since $x_{\beta_j}(t) > 0$ on I_j, there is a closed interval $[u, v] \subset I_j$ and $\varepsilon > 0$ such that $x_i(t) \geqq \varepsilon$ for $t \in [u, v]$ and $i \in \beta_j$. By Lemma 15 there exists a nonzero analytic $y_{\beta_j}(\cdot)$ satisfying

$$B_{\cdot\beta_j} y_{\beta_j}(t) + K_{\cdot\beta_j} \int_u^t y_{\beta_j}(s)\, ds = 0, \qquad t \in [u, v],$$

and

$$\int_u^v y_{\beta_j}(s)\, ds = 0.$$

Rescale so that $|y_i(t)| \leqq \varepsilon$ for all $t \in [u, v)$ and $i \in \beta_j$. By the construction of y_{β_j} in Lemma 15, this can always be done. Set $y_{\alpha_j} = 0$ on $[u, v)$ and $y = 0$ on $[0, T] \backslash [u, v)$. Then by construction y satisfies

$$By(t) + K \int_0^t y(s)\, ds = 0,$$

and

$$x(t) + y(t) \geqq 0, \qquad x(t) - y(t) \geqq 0,$$

for all $t \in [0, T]$. Hence both $x + y$ and $x - y$ satisfy (18). Since y is not identically zero, it follows that x is not an extreme point. □

4. Concluding remarks. Some conclusions are immediate from the characterization in Theorem 17:

1. Right analytic extreme points can have at most m variables positive over any interval of time, where B is $m \times n$.

2. For the control theory formulation (given in the introduction) with constant coefficients, the matrix $\mu B_{\cdot\beta_j} + K_{\beta_j}$ has the form

$$\begin{pmatrix} E - I\mu & F \\ G & H \end{pmatrix},$$

for some submatrices E, F, G, H of A, B, C, D, respectively. A sufficient condition for this matrix to have full column rank for some μ is that H have full column rank. (This is the assumption made in Teren [17].) In the event that there are no constraints on the state variables (other than the differential equation), the submatrix G above is zero. H having full column rank is then also a necessary condition. This can be interpreted as the control $u(t)$ being an extreme point of the polyhedron

$$Du(t) = -b(t), \qquad u(t) \geqq 0,$$

as in the bang-bang principle [11].

For computational purposes we would work with extreme points having only finitely many constant basis intervals. As is the case in linear programming, the index sets β_j would be appropriately enlarged so that $(\mu B_{\cdot\beta_j} + K_{\cdot\beta_j})^{-1}$ exists for some μ. This will allow the β_j (basic) variables to be represented as functions of the α_j (nonbasic) variables. In this setting, we can then work with these extreme points precisely as with the basic feasible solutions in linear programming. This is pursued further in Perold [14].

Appendix A. Examples of extreme points. The constraints are all of the form

$$B(t)x(t)+\int_0^t K(t,s)x(s)\,ds = b(t), \qquad x(t)\geqq 0, \quad t\in[0,T],$$

where $B(t)$ and $K(t,s)\in R^{m\times n}$.

(A1) An extreme point whose positive values are not uniquely determined when the remaining ones are held fixed at zero. $m=n=1$, $B(t)=t$, $b(t)=\frac{1}{3}t^3$, and $K(t,s)=-2$. Thus

$$tx(t)-2\int_0^t x(s)\,dx = \tfrac{1}{3}t^2, \qquad x(t)\geqq 0, \quad t\in[0,T].$$

One can easily show that the only solutions to this equation are of the form

$$x(t)=t^2+\alpha t,$$

for α any fixed but arbitrary scalar. The feasible (nonnegative) solutions are generated by $\alpha\geqq 0$, and the only extreme point is given by the case $\alpha=0$,

$$x(t)=t^2.$$

Observe, however, that since all feasible solutions are positive on $(0,T]$ and zero at $t=0$, the restriction $x(0)=0$ is redundant.

(A2) An extreme point whose values throughout the interval depend explicitly on T.

$$tx(t)-2\int_0^t x(s)\,ds = -t, \qquad x(t)\geqq 0, \quad t\in[0,T].$$

This differs from (A1) only in the right-hand side. Here, all solutions are of the form

$$x(t)=1+\alpha t,$$

for α any fixed but arbitrary scalar. The (only) extreme point is given by $\alpha=-1/T$,

$$x(t)=1-\frac{t}{T}.$$

In this case $x(T)=0$, $x(t)>0$ on $[0,T)$, and the restriction $x(T)=0$ does uniquely determine x.

(A3) An extreme point that is not locally uniquely defined, and has more variables positive over an interval of time than there are equations. $m=1$, $n=3$, $T=5$, $B(t)=0$ on $[0,5]$. Define $K(t,s)=[k_1(t,s)k_2(t,s)k_3(t,s)]$ on the triangle $0\leqq s\leqq t\leqq 5$ as follows:

s	t	$k_1(t,s)$	$k_2(t,s)$	$k_3(t,s)$
	$[s,1)$	0	0	0
	$[1,2)$	e^{-st}	0	0
$[0,1]$	$[2,3)$	0	e^{-st}	0
	$[3,4)$	0	0	e^{-st}
	$[4,5]$	0	0	0
$(1,5]$	$[s,5]$	0	0	0

Set

$$\bar{x}_1(t) = \bar{x}_2(t) = \bar{x}_3(t) = 1 \quad \text{on } [0, 1),$$

$$\bar{x}_1(t) = \bar{x}_2(t) = \bar{x}_3(t) = 0 \quad \text{on } [1, 5],$$

and then define $b(\cdot)$ by

$$b(t) = \int_0^t K(t, s)\bar{x}(s)\, ds, \qquad t \in [0, 5].$$

Consider now the specific values of $K(t, s)$:

(i) For $t \in [0, 1]$ the equation is trivial, i.e., $0 = 0$, so that the values of x on $[0, 1)$ are not in any way determined by the coefficients on $[0, 1)$.

(ii) For $t \in [1, 2)$, the equation reads

$$\int_0^1 e^{-st} x_1(s)\, ds = b(t).$$

The left-hand side is the Laplace transform of x_1 on the interval $[0, 1)$, evaluated at t. By the uniqueness theorem for Laplace transforms (see [20]) the above equation has a unique (a.e.) bounded solution x_1 on $[0, 1)$. By construction this solution is $x_1 = \bar{x}_1 = 1$.

(iii) For $t \in [2, 3)$ and $t \in [3, 4)$ we obtain similar equations in x_2 and x_3 respectively, and conclude that the only possible solution is $x_2 = \bar{x}_2$ and $x_3 = \bar{x}_3$ on $[0, 1)$.

The above shows that x on $[0, 1)$ is uniquely determined, independent of the choice of x on $[1, 5]$. Thus choosing $x = 0$ on $[1, 5]$ yields $x = \bar{x}$ on the whole interval, and this must be an extreme point solution.

Notice that we have one equality constraint in three nonnegative variables, and that on $[0, 1)$ all three are positive while on $[1, 5]$ all three are zero. □

Appendix B. Right analytic functions. This appendix contains the propositions about right analytic functions that are required in § 3. The main result is the proof of Lemma 9 which we restate below as Lemma B3. For completeness we review the definitions of analytic and right analytic functions.

DEFINITION B1. [e.g., 15]. Let $\Omega \subset R$ be open and $g : \Omega \to R$. Then g is said to be *analytic* on Ω if to every open interval $I \subset \Omega$ with center a, there corresponds a series $\sum_{i=0}^{\infty} c_i(t-a)^i$ which converges to $g(t)$ for all $t \in I$.

DEFINITION B2. A function $g : [0, T] \to R$ will be called *right analytic* if for each $t \in [0, T)$ there is an $\varepsilon > 0$ and an analytic function $h : (t-\varepsilon, t+\varepsilon) \to R$ such that $g(s) = h(s)$ for all $s \in [t, t+\varepsilon)$.

LEMMA B3. *Let $g : [0, T] \to R^n$ have right analytic components $g_i(\cdot)$, $i = 1, \cdots, n$. Then there exists a (possibly infinite) disjoint family of open intervals $\{I_j\}$ such that $[0, T) \doteq U_j \tilde{I}_j$, and such that for each interval I_j, each g_i satisfies*

(i) *g_i is analytic on I_j, and*

(ii) *either $|g_i| > 0$ on I_j or $g_i = 0$ on I_j.*

The proof of this result will require the following lemmas.

LEMMA B4. *Let $I \subset R$ be an open interval and $\{J_i\}$ a family of open intervals whose union is I. Let $g : I \to R$ be given. If g is analytic on each J_i then g is analytic on I.*

LEMMA B5. *Let $K \subset R$ be a compact interval, and let h be an analytic function defined on a neighborhood of K. Then either $h = 0$ on K or h has finitely many zeros in K.*

COROLLARY B6. *Let $I \subset R$ be an open interval, and let $h : I \to R$ be analytic. Let $Z(h) = \{t \in I : h(t) = 0\}$. Then either $Z(h) = I$, or $Z(h)$ has no limit point in I. In the latter case $Z(h)$ is at most countable.*

The proofs of both these lemmas may be found in [15].

LEMMA B7. *Let $\{J_\alpha\}$ be any family of open intervals, and $\{I_j\}$ be a disjoint family of open intervals such that $\cup_j I_j = \cup_\alpha J_\alpha$. Then $\cup_\alpha \tilde{J}_\alpha \subseteq \cup_j \tilde{I}_j$.*

Proof. Since the collection $\{I_j\}$ is disjoint, the connectedness of intervals implies that each J_α is contained in a unique I_j. Thus $\tilde{J}_\alpha \subseteq \tilde{I}_j$ for some j and we are done. □

Proof of Lemma B3. We shall prove the proposition for the case $n = 1$, since from this, the general case follows immediately.

By definition of g being right analytic, for each $t \in [0, T)$ there exists an $\varepsilon_t > 0$ such that g is analytic on $K_t = (t, t + \varepsilon_t) \subset [0, T)$. Let $V = \cup_t K_t$ and $W = \cup_t \tilde{K}_t$. Since V is open, there exists a disjoint family of open intervals $\{J_i\}_{i=1}^{\infty}$ whose union is V.

On each J_i we now obtain the following: By Lemma B4, g is analytic on J_i. By Corollary B6, either $g = 0$ on J_i or $g = 0$ on at most a countable sequence $\{t_1, t_2, \cdots\} \subset J_i$. In the latter case write $J_i = (t', t'')$. Since the sequence has no limit point in J_i, and since by definition g agrees with an analytic function defined on a neighborhood of t', the only possible limit point of $\{t_k\}$ is t''. Hence we may assume that

$$t' < t_1 < t_2 < \cdots < t''.$$

With this partition of J_i we can now conclude that there is a (possibly infinite) collection of disjoint open intervals $\{L_i^k\}$ such that $|g| > 0$ on L_i^k or $g = 0$ on L_i^k, and such that $\tilde{J}_i = \cup_k \tilde{L}_i^k$.

Now it is clear that $W = [0, T)$. By Lemma B7, it follows that $\cup_i \tilde{J}_i = [0, T)$. Hence $\cup_{i,k} \tilde{L}_i^k = [0, T)$. By relabeling the family $\{L_i^k\}$ as $\{I_j\}$, we obtain the desired result. □

LEMMA B8. *Let $g, h : [0, T] \to R$ be right analytic, and $t_0 \in [0, T)$ be such that $g(t_0) \neq h(t_0)$. Then there exist $0 \leqq r < s \leqq T$ such that $g = h$ on $[0, r)$ and $g \neq h$ on (r, s).*

Proof. Let $E = \{t : g(t) \neq h(t)\}$. Since $t_0 \in E$, E is nonempty. Let $r = \inf E$. Then by definition of r, $g = h$ on $[0, r)$. Since both g and h are right analytic, there is an $\varepsilon > 0$ such that g and h agree with analytic functions on $[r, r + \varepsilon]$. Again by definition of r, there is a $t \in (r, r + \varepsilon)$ such that $g(t) \neq h(t)$. By Lemma B5, $g - h$ has finitely many zeros on $[r, r + \varepsilon]$. Hence there is an $s \in (r, r + \varepsilon)$ such that $g \neq h$ on (r, s). □

Acknowledgment. The author is grateful to Professor George B. Dantzig for his guidance and inspiration during the course of this work.

REFERENCES

[1] R. D. BELLMAN, *Bottleneck problems and dynamic programming*, Proc. Nat. Acad. Sci., 39, 1953.
[2] R. E. BELLMAN, *Dynamic Programming*, Princeton University Press, Princeton, NJ, 1972.
[3] S. L. CAMPBELL, C. D. MEYER, Jr., AND N. J. ROSE, *Applications of the Drazin inverse to linear systems of differential equations with singular constant coefficients*, SIAM J. Appl. Math., 31 (1976), pp. 411–425.
[4] G. B. DANTZIG, *Linear Programming and Extensions*, Princeton University Press, Princeton, NJ, 1963.
[5] ———, *Large Scale Systems Optimization with Applications to Energy*, Technical Report SOL 77-3, April 1977, Department of Operations Research, Stanford University, Stanford, California.
[6] ———, *Linear control processes and mathematical programming*, SIAM J. Control Ser. A, 4 (1966), pp. 56–60.
[7] V. DOLEZAL, *Dynamics of Linear Systems*, Academia, Prague, 1967.
[8] W. P. DREWS, R. J. HARTBERGER AND R. B. SEGERS, *On continuous mathematical programming*, in Optimization Methods in Resource Allocation, R. W. Cottle and J. Krarup, eds., Crane, Russak and Co., New York, 1974.
[9] R. C. GRINOLD, *Continuous programming, part one: linear objectives*, J. Math. Anal. Appl., 28 (1969), pp. 32–51.
[10] R. B. HOLMES, *Geometric Functional Analysis and its Applications*, Springer-Verlag, New York, 1975.
[11] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
[12] R. S. LEHMAN, *On the Continuous Simplex Method*, RAND Research Memorandum RM-1386, Santa Monica, California, 1954.

[13] P. LEVINE AND J. C. POMEROL, *C-closed Mappings and Kuhn-Tucker Vectors in Convex Programming*, Discussion Paper 7620, Center for Operations Research and Economics, Universite Catholique de Louvain, Heverlee, Belgium, 1976.
[14] A. F. PEROLD, *On a continuous time simplex method*, in preparation.
[15] W. RUDIN, *Real and Complex Analysis*, McGraw-Hill, New York, 1974.
[16] L. M. SILVERMAN, *Inversion of multivariable linear systems*, IEEE Trans. Automat. Control, AC-14 (1969) pp. 270–276.
[17] F. TEREN, *Minimum Time Acceleration of Aircraft Turbofan Engines by Using an Algorithm Based on Nonlinear Programming*, NASA Technical Memorandum TM-73741, Lewis Research Center, Cleveland, Ohio, September 1977.
[18] F. G. TRICOMI, *Integral Equations*, Interscience, New York, 1957.
[19] W. F. TYNDALL, *A duality theorem for a class of continuous linear programming problems*, J. Soc. Indust. Appl. Math., 13 (1965), pp. 644–666.
[20] A. H. ZEMANIAN, *Distribution Theory and Transform Analysis*, McGraw-Hill, New York, 1965.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 1, January 1981

0363-0129/81/1901-0006 $01.00/0

NECESSARY CONDITIONS FOR DISTRIBUTED CONTROL PROBLEMS GOVERNED BY PARABOLIC VARIATIONAL INEQUALITIES*

VIOREL BARBU†

Abstract. Necessary conditions for optimality in distributed control problems governed by semilinear and variational parabolic inequalities are given. The optimality conditions are expressed in terms of generalized gradients, and are obtained by means of an abstract approximating control process.

1. Introduction. This paper is mainly concerned with distributed control problems with a convex cost criterion governed by parabolic semilinear equations of the form

$$\begin{aligned} &y_t + A_0 y + \beta(y) \ni Bu + f && \text{on } Q = \Omega \times]0, T[, \\ &\alpha_1 y + \alpha_2 \frac{\partial y}{\partial \nu} = 0 && \text{on } \Sigma = \Gamma \times]0, T[, \\ &y(\cdot, 0) = y_0 && \text{on } \Omega, \end{aligned} \tag{1.1}$$

where β is a maximal monotone in $R \times R$, A_0 is a second order elliptic differential operator on Ω and B is a linear symmetric continuous operator from a control space U to $L^2(\Omega)$. Here Ω is a bounded open subset of Euclidean space R^N with the boundary Γ.

In particular (1.1) represents the most convenient way to formulate a large class of parabolic variational inequalities arising in mechanics, heat transfer, and the theory of free boundary problems. Examples of this kind can be found in the book of Duvaut and Lions [7], and the survey of Lions [8].

Several results on existence, necessary conditions for optimality and approximation for control problems governed by equations of the form (1.1) have been obtained by, among others, Lions [9], Mignot [12], Yvon [19], Puel [13], and Saquez [17], [18]. The results we give here differ in certain key respects from most of the existing literature on necessary conditions. The underlying idea behind our approach consists of approximating the control problem for (1.1) by a family of smooth problems and afterwards to tend to the limit in the approximating optimality equations. This approach has already been used by the author in [2].

The plan of the paper is the following. In §§ 2 and 3 we study a general approximating process for convex control problems governed by a general class of nonlinear evolution equations in Hilbert space. The main results of this paper, Theorems 1, 2, and 3 are derived in § 4 by specializing the general theory to control problems governed by (1.1). This general approach is also used in § 5 to derive necessary conditions of optimality for control problems governed by nonlinear boundary value parabolic problems of the form

$$\begin{aligned} &y_t + A_0 y = Bu + f && \text{on } Q, \\ &\frac{\partial y}{\partial \nu} + \beta(y) \ni 0 && \text{on } \Sigma, \\ &y(\cdot, 0) = y_0 && \text{on } \Omega. \end{aligned} \tag{1.2}$$

* Received by the editors June 18, 1979, and in revised form April 11, 1980.
† Faculty of Mathematics, University of Iasi, Iasi 6600, Romania.

Based on this approach, Theorems 1 and 4 can certainly be extended to more general equations of the form (1.1) and (1.2), but we do not here attempt maximum generality nor claim to be comprehensive in any sense. The conditions which are shown in these theorems to be necessary for optimality, are detailed and made more explicit for two important cases: β locally Lipschitz and β a multivalued graph of the form (4.45).

The following notation is used. If E is a Banach space, then we shall denote by $L^p(0, T; E)$, $1 \leqq p \leqq \infty$, the space of all p-integrable E-valued functions on $[0, T]$, and by $C(0, T; E)$ the usual Banach space of all continuous functions from $[0, T]$ to E. We shall denote by $W^{1,p}(0, T; E)$ the space $\{y \in L^p(0, T; E);\ y' \in L^p(0, T; E)\}$ where the derivative y' of y is taken in the sense of vectorial distributions on $]0, T[$. Equivalently, $y \in W^{1,p}(0, T; E)$ means that $y : [0, T] \to E$ is absolutely continuous, a.e. differentiable on $]0, T[$, and

$$y(t) = y(0) + \int_0^t y'(s)\, ds \quad \text{for } t \in [0, T], \quad y' \in L^p(0, T; E). \tag{1.3}$$

In the special case when E is a space of functions we shall denote y' by the symbol y_t.

Given a lower semicontinuous convex function $\varphi : X \to \bar{R} =]-\infty, +\infty]$ we shall denote by $\partial\varphi(x) \in E'$ (the dual space of E) the set of all *subgradients* of φ at x, i.e.,

$$\partial\varphi(x) = \{x^* \in E';\ \varphi(x) \leqq \varphi(y) + (x^*, x - y) \text{ for all } y \in E\}. \tag{1.4}$$

If φ is Gâteaux differentiable at x, then $\partial\varphi(x)$ consists of a single element, namely the gradient $\nabla\varphi(x)$ of φ at x. The mapping $\partial\varphi : E \to E'$ is called the *subdifferential* of φ. We shall denote by $D(\varphi) = \{x;\ \varphi(x) < +\infty\}$ the *effective domain* of φ and by $D(\partial\varphi)$ the *domain* of $\partial\varphi$; i.e., $D(\partial\varphi) = \{x \in E;\ \partial\varphi(x) \neq \emptyset\}$. For other notation and results in convex analysis relevant to this paper we refer to the books [1], [6], [14] and to the recent survey by Rockafellar [16].

If Ω is an open subset of the Euclidean space R^N, we shall denote by $W^{k,p}(\Omega)$ and $W_0^{k,p}(\Omega)$, (k a natural number and $1 \leqq p \leqq \infty$), the usual Sobolev spaces on Ω. For $p = 2$ we shall simply write $W^{k,2}(\Omega) = H^k(\Omega)$ (respectively, $W_0^{k,2}(\Omega) = H_0^k(\Omega)$). For any real s we denote by $H^s(\Gamma)$ the corresponding Sobolev space on the boundary Γ (see [11, p. 34]). For $1 \leqq p \leqq \infty$ we denote by $W_p^{2,1}(Q)$ the space

$$W_p^{2,1}(Q) = L^p(0, T; W^{2,p}(\Omega)) \cap W^{1,p}(0, T; L^p(\Omega)),$$

where $Q = \Omega \times]0, T[$. For $p = 2$ we set $W_2^{2,1}(Q) = H^{2,1}(Q)$.

2. An abstract formulation of the problem. Let H and U be two real Hilbert spaces identified with their duals with norms denoted $|\cdot|$, $\|\cdot\|$ and with inner products $(\cdot, \cdot)$ and $\langle \cdot, \cdot \rangle$, respectively. Let B be a linear continuous operator from U to H, and let $\varphi : H \to \bar{R} =]-\infty, +\infty]$ be a lower semicontinuous, convex function, not identically $+\infty$. We set $F = \partial\varphi$.

The state equation governing our control problem is ($[0, T]$ is a finite interval)

$$\begin{aligned} &y'(t) + Fy(t) \ni Bu(t) + f(t) \quad \text{a.e. } t \in]0, T[, \\ &y(0) = y_0, \end{aligned} \tag{2.1}$$

where $y_0 \in D(\varphi)$ and $f \in L^2(0, T; H)$ are given and the *controller* $u(\cdot)$ is an element of $L^2(0, T; U)$.

It is well known (see [3], [4]) that for each $y_0 \in D(\varphi)$ and $v \in L^2(0, T; H)$ the

Cauchy problem

$$\begin{aligned} &z'(t)+Fz(t)\ni v(t) \quad \text{a.e. } t\in]0,T[,\\ &z(0)=y_0, \end{aligned} \tag{2.2}$$

has a unique solution $z=\Theta v\in W^{1,2}(0,T;H)$. Furthermore the operator Θ is Lipschitzian from $L^2(0,T;H)$ to $C(0,T;H)$.

The optimal control problem we consider here is:

(P) Minimize

$$\int_0^T L(y(t),u(t))\,dt \quad \text{in } y\in W^{1,2}(0,T;H) \text{ and } u\in L^2(0,T;U)$$

subject to (2.1).

In terms of the operator Θ defined above, problem (P) can be brought into the form

$$\min\left\{\int_0^T L(\Theta(Bu+f)(t),u(t))\,dt;\ u\in L^2(0,T;U)\right\}. \tag{2.3}$$

We proceed now to set forth the basic assumptions on L and F.

(i) *$L: H\times U\to\bar{R}$ is convex, lower semicontinuous ana $\not\equiv+\infty$. The Hamiltonian function*

$$H(y,p)=\sup\{\langle p,v\rangle-L(y,v);\ v\in U\} \tag{2.4}$$

is everywhere finite on $H\times U$.

(ii) *$F=\partial\varphi$. Every level subset $\{y\in H;\ |y|^2+\varphi(y)\leqq C\}$ is compact in H.*

(iii) *There exists a family of lower semicontinuous convex functions $\varphi^\varepsilon: H\to\bar{R}$ $(\varepsilon>0)$ such that:*

(a) *For each $y\in D(\varphi)$ one has*

$$\limsup_{\varepsilon\to 0}\varphi^\varepsilon(y)\leqq\varphi(y), \tag{2.5}$$

while $\liminf_{\varepsilon\to 0}\varphi^\varepsilon(y_\varepsilon)\geqq\varphi(y)$ for any sequence $\{y_\varepsilon\}$ convergent to y for $\varepsilon\to 0$.

(b) *There exists a lower semicontinuous, convex function $\psi: H\to\bar{R}$, $\psi\not\equiv+\infty$ such that*

$$\varphi^\varepsilon(y)\geqq\psi(y) \quad \text{for all } y\in H \text{ and } \varepsilon>0, \tag{2.6}$$

and every level subset of the form $\{y\in H;\ |y|^2+\psi(y)\leqq C\}$ is compact in H.

(c) *Let $F_\varepsilon=\partial\varphi^\varepsilon$. There exists a constant C independent of ε and λ such that*

$$(F_\varepsilon(y)-F_\lambda(z),\,y-z)\geqq -C(\varepsilon+\lambda)(1+|F_\varepsilon^0(y)|^2+|F_\lambda^0(x)|^2) \tag{2.7}$$

$$\text{for } y\in D(F_\varepsilon),\ z\in D(F_\lambda).$$

(d) *For each $\varepsilon>0$ the operator $\Theta_\varepsilon: L^2(0,T;H)\to L^2(0,T;H)$ is Gâteaux differentiable.*

Here $F_\varepsilon^0(y)=\inf\{|w|;\ w\in F_\varepsilon(y)\}$, and Θ_ε denotes the operator defined by $\Theta_\varepsilon v=z_\varepsilon$, where $z_\varepsilon\in W^{1,2}(0,T;H)$ is the solution to

$$\begin{aligned} &z_\varepsilon'+F_\varepsilon z_\varepsilon\ni v \quad \text{a.e. on }]0,T[,\\ &z_\varepsilon(0)=y_0, \end{aligned} \tag{2.8}$$

($y_0\in D(\varphi)$ is the element fixed in problem (P)). F_ε can be regarded as a *penalty operator* associated with F. Now we shall establish some technical lemmas.

LEMMA 1. *The operators* Θ_ε *and* Θ *are weakly-strongly continuous from* $L^2(0, T; H)$ *into itself. Moreover, one has*

$$|(\nabla\Theta_\varepsilon(w)v)(t)| \leqq \int_0^t |v(s)|\, ds, \qquad t \in [0, T], \tag{2.9}$$

for all $v, w \in L^2(0, T; H)$. *Here* $\nabla\Theta_\varepsilon(w) \in \mathscr{L}(L^2(0, T; H), L^2(0, T; H))$ *denotes the Gâteaux differential of* Θ_ε *at* w.

Proof. Let $\{v_n\}$ be a sequence of $L^2(0, T; H)$ which converges weakly to some element v. We set $z_n = \Theta v_n$. By (2.2) it follows that (see, e.g., [3])

$$\tfrac{1}{2}\int_0^t |z_n'|^2\, ds + \varphi(z_n(t)) \leqq \varphi(y_0) + \tfrac{1}{2}\int_0^t |v_n|^2\, ds \leqq C, \tag{2.10}$$

and

$$|z_n(t) - y_0|^2 + \int_0^t \varphi(z_n(s))\, ds \leqq C\Big(1 + \int_0^t |v_n|^2\, ds\Big), \qquad t \in [0, T]. \tag{2.11}$$

It follows by estimates (2.10) and (2.11) that $\{z_n\}$ is bounded in $W^{1,2}(0, T; H)$, and $\varphi(z_n(t)) \leqq C$ for all $t \in [0, T]$ and n. Along with Assumption (ii) and the Arzela-Ascoli theorem, the latter implies that $\{z_n\}$ is a precompact subset of $C(0, T; H)$. Hence there exists a subsequence again denoted $\{z_n\}$ and $z \in C(0, T; H)$ such that $z_n(t) \to z(t)$ uniformly on $[0, T]$. Then by a standard argument it follows that z is a solution to (2.2), i.e., $z = \Theta v$. Hence $\Theta v_n \to \Theta v$ strongly in $L^2(0, T; H)$. The weak-strong continuity of Θ_ε follows by a parallel argument.

Let w and v be given in $L^2(0, T; H)$. We have

$$\nabla\Theta_\varepsilon(w)(v) = \lim_{\lambda \to 0} (\Theta_\varepsilon(w + \lambda v) - \Theta_\varepsilon w)/\lambda,$$

and

$$(\Theta_\varepsilon(w + \lambda v) - \Theta_\varepsilon w)' + F_\varepsilon\Theta_\varepsilon(w + \lambda v) - F_\varepsilon\Theta_\varepsilon w \ni \lambda v \quad \text{on }]0, T[. \tag{2.12}$$

Multiplying (2.12) by $\Theta_\varepsilon(w + \lambda v) - \Theta_\varepsilon(w)$ and integrating over $[0, t]$, we see that

$$|\Theta_\varepsilon(w + \lambda v)(t) - \Theta_\varepsilon w(t)| \leqq \lambda \int_0^t |v(s)|\, ds, \qquad t \in [0, T],$$

which implies (2.9) as claimed.

LEMMA 2. *Let* $v \in L^2(0, T; H)$ *and* $y_0 \in D(\varphi)$ *be fixed. Then*

$$|(\Theta_\varepsilon v)(t) - (\Theta v)(t)| \leqq C\varepsilon^{1/2} \quad \textit{for } t \in [0, T], \quad \varepsilon > 0, \tag{2.13}$$

$$(\Theta_\varepsilon v)' \to (\Theta v)' \quad \textit{weakly in } L^2(0, T; H) \quad \textit{for } \varepsilon \to 0. \tag{2.14}$$

Proof. Let $z_\varepsilon = \Theta_\varepsilon v$ be the solution to (2.8). By estimates (2.10) and (2.11) (where $\varphi = \varphi^\varepsilon$ and $v_n = v$) it follows that

$$\int_0^t |z_\varepsilon'|^2\, ds + |z_\varepsilon(t)|^2 + \varphi^\varepsilon(z_\varepsilon(t)) \leqq C, \qquad t \in [0, T], \quad \varepsilon > 0, \tag{2.15}$$

(we shall denote by C several positive constants independent of ε.) To get the estimate (2.15) we have also used condition (2.6) which implies that the φ^ε uniformly majorize an affine function on H.

In particular it follows by (2.15) that $\{F_\varepsilon(z_\varepsilon)\}$ remain in a bounded subset of $L^2(0, T; H)$. Along with Assumption (iiic) this yields

$$|z_\varepsilon(t)-z_\lambda(t)|^2 \leqq C(\varepsilon+\lambda) \quad \text{for } \varepsilon, \lambda > 0, \quad t \in [0, T]. \tag{2.16}$$

Thus there exists $z^0 \in W^{1,2}(0, T; H)$ such that

$$\begin{aligned} z_\varepsilon &\to z^0 && \text{in } C(0, T; H), \\ z'_\varepsilon &\to (z^0)' && \text{weakly in } L^2(0, T; H), \\ F_\varepsilon z_\varepsilon &\to g && \text{weakly in } L^2(0, T; H). \end{aligned}$$

To conclude the proof it remains to show that

$$g(t) \in Fz^0(t) \quad \text{a.e. } t \in]0, T[.$$

By (2.8), we have

$$(z_\varepsilon(t)-z_\varepsilon(s), z_\varepsilon(t)-z)+\int_s^t (\varphi^\varepsilon(z_\varepsilon(\sigma))-\varphi^\varepsilon(z))\, d\sigma \leqq \int_s^t (v(\sigma), z_\varepsilon(\sigma)-z)\, d\sigma,$$

for all $z \in H$ and $0 < s < t \leqq T$. Letting ε tend to zero and applying Fatou's lemma we get in virtue of condition (iiia) in the assumptions

$$(z^0(t)-z^0(s), z^0(t)-z)+\int_s^t \varphi(z^0(\sigma))\, d\sigma-(t-s)\varphi(z) \leqq \int_s^t (v(\sigma), z^0(\sigma)-z)\, d\sigma,$$

and this yields

$$(z^0)'(t)+\partial\varphi(z^0(t)) \ni v(t) \quad \text{a.e. } t \in]0, T[,$$

as claimed. Hence $z^0 = \Theta v$ and by (2.16) the estimate (2.13) follows, thereby completing the proof.

3. Convergence of an approximating control process. For each $\varepsilon > 0$, denote by $L_\varepsilon : H \times U \to R$ the regularized function

$$L_\varepsilon(y, u) = \inf\left\{\frac{|y-z|^2+\|u-v\|^2}{2\varepsilon}+L(z, v);\ z \in H, v \in U\right\}. \tag{3.1}$$

It is well known (see e.g., [1, p. 107]) that L_ε is Fréchet differentiable.

Let $\delta : R^+ \to R^+$ be a continuous function satisfying

$$\lim_{\varepsilon \to 0} \delta(\varepsilon)/\varepsilon = 0, \tag{3.2}$$

and let $L^\varepsilon : H \times U \to R$ be the function defined by

$$L^\varepsilon = L_{\delta(\varepsilon)} \quad \text{for } \varepsilon > 0. \tag{3.3}$$

By $\partial L^\varepsilon(y, u) = (\partial_1 L^\varepsilon(y, u), \partial_2 L^\varepsilon(y, u)) \subset H \times U$ we shall denote the gradient of L^ε at (y, u). Let $u^* \in L^2(0, T; U)$ be an optimal control in problem (P) and let $y^* \in W^{1,2}(0, T; H)$ be the corresponding state in (1.1).

Consider the approximating control problem:

(P_ε) Minimize

$$\int_0^T (L^\varepsilon(y, u)+\tfrac{1}{2}\|u^*-u\|^2)\, dt \quad \text{in } y \in W^{1,2}(0, T; H) \text{ and } u \in L^2(0, T; U),$$

subject to

$$\begin{aligned} &y' + F_\varepsilon y \ni Bu + f \quad \text{a.e. on }]0, T[, \\ &y(0) = y_0. \end{aligned} \tag{3.4}$$

LEMMA 3. *For each* $\varepsilon > 0$, *problem* (P_ε) *has at least one solution* $(y_\varepsilon, u_\varepsilon) \in W^{1,2}(0, T; H) \times L^2(0, T; U)$.

Proof. Let $\Phi_\varepsilon : L^2(0, T; U) \to R$ be the function given by

$$\Phi_\varepsilon(u) = \int_0^T L^\varepsilon(\Theta_\varepsilon(Bu + f), u)\, dt + \tfrac{1}{2}\int_0^T \|u^* - u\|^2\, dt. \tag{3.5}$$

Inasmuch as L^ε is Lipschitzian and Θ_ε is weakly-strongly continuous, we may infer that Φ_ε is weakly lower semicontinuous on $L^2(0, T; U)$. Moreover, $\Phi_\varepsilon(u) \to +\infty$ for $\|u\|_{L^2(0,T;U)} \to +\infty$. Consequently, Φ_ε attains its infimum on $L^2(0, T; U)$ as claimed.

LEMMA 4. *For each* $\varepsilon > 0$, *there exists a function* $p_\varepsilon \in L^2(0, T; H)$ *such that*

$$p_\varepsilon = -(\nabla\Theta_\varepsilon(Bu_\varepsilon + f))^* \partial_1 L^\varepsilon(y_\varepsilon, u_\varepsilon), \tag{3.6}$$

and

$$B^* p_\varepsilon = \partial_2 L^\varepsilon(y_\varepsilon, u_\varepsilon) + u_\varepsilon - u^*. \tag{3.7}$$

Proof. Since $(y_\varepsilon, u_\varepsilon)$ is a minimum point for (P_ε) and L^ε is Fréchet differentiable, we find by a standard argument

$$\begin{aligned} \int_0^T &((\partial_1 L^\varepsilon(y_\varepsilon, u_\varepsilon), \nabla\Theta_\varepsilon(Bu_\varepsilon + f)Bv) \\ &+ \langle \partial_2 L^\varepsilon(y_\varepsilon, u_\varepsilon) + u_\varepsilon - u^*, v\rangle)\, dt = 0 \quad \text{for all } v \in L^2(0, T; U). \end{aligned} \tag{3.8}$$

Let p_ε be the function defined by (3.6) where $(\nabla\Theta_\varepsilon(\cdot))^*$ denotes the adjoint of $\nabla\Theta_\varepsilon(\cdot)$. Then (3.8) yields (3.7) as desired.

LEMMA 5. *For* $\varepsilon \to 0$ *we have*

$$y_\varepsilon \to y^* \quad \textit{strongly in } C(0, T; H), \tag{3.9}$$

$$u_\varepsilon \to u^* \quad \textit{strongly in } L^2(0, T; U), \tag{3.10}$$

$$y'_\varepsilon \to (y^*)' \quad \textit{weakly in } L^2(0, T; H). \tag{3.11}$$

Proof. It suffices to prove (3.9)–(3.11) on some subsequence. We have for all $u \in L^2(0, T; U)$, and $\varepsilon > 0$,

$$\begin{aligned} \int_0^T (L^\varepsilon(y_\varepsilon, u_\varepsilon) + \tfrac{1}{2}\|u_\varepsilon - u^*\|^2)\, dt &\leqq \int_0^T (L^\varepsilon(\Theta_\varepsilon(Bu + f), u) + \tfrac{1}{2}\|u^* - u\|^2)\, dt \\ &\leqq \int_0^T L^\varepsilon(\Theta_\varepsilon(Bu^* + f), u^*)\, dt. \end{aligned} \tag{3.12}$$

On the other hand, we have shown in Lemma 2 that

$$|\Theta_\varepsilon(Bu^* + f)(t) - \Theta(Bu^* + f)(t)| \leqq C\varepsilon^{1/2} \quad \text{for } t \in [0, T], \tag{3.13}$$

while by (3.1) it follows that

$$L^\varepsilon(\Theta_\varepsilon(Bu^* + f), u^*) \leqq L(y^*, u^*) + |\Theta_\varepsilon(Bu^* + f) - \Theta(Bu^* + f)|^2/2\delta(\varepsilon).$$

Combining the latter with (3.12) and (3.13), we see that

$$\limsup_{\varepsilon\to 0}\left(\int_0^T L^\varepsilon(y_\varepsilon, u_\varepsilon)\,dt+\tfrac{1}{2}\int_0^T \|u^*-u_\varepsilon\|^2\,dt\right)\leqq\int_0^T L(y^*, u^*)\,dt. \tag{3.14}$$

In particular, it follows that $\{u_\varepsilon\}$ is bounded in $L^2(0, T; U)$. Then multiplying both sides of (3.4) (where $y=y_\varepsilon$ and $u=u_\varepsilon$) by $F_\varepsilon y_\varepsilon$ and integrating on $[0, T]$ we get (see inequality (2.15)),

$$\int_0^T |F_\varepsilon y_\varepsilon(t)|^2\,dt+\int_0^T |y'_\varepsilon(t)|^2\,dt+\varphi^\varepsilon(y_\varepsilon(t))\leqq C \quad \text{for } t\in[0, T]. \tag{3.15}$$

Then by condition (iiib) and the Ascoli theorem it follows that the family $\{y_\varepsilon\}$ is compact in $C(0, T; H)$. Thus selecting a subsequence, if necessary, we may assume that

$$\begin{aligned} &u_\varepsilon\to u^0 && \text{weakly in } L^2(0, T; U),\\ &y_\varepsilon\to y^0 && \text{strongly in } C(0, T; H),\\ &y'_\varepsilon\to y^{0\prime} && \text{weakly in } L^2(0, T; H), \end{aligned}$$

Proceeding as in the proof of Lemma 2, we see that (y^0, u^0) satisfy (2.1). In other words $y^0=\Theta(Bu^0+f)$. On the other hand,

$$\liminf_{\varepsilon\to 0}\int_0^T L^\varepsilon(y_\varepsilon, u_\varepsilon)\,dt\geqq\int_0^T L(y^0, u^0)\,dt, \tag{3.16}$$

because the function $(y, u)\to\int_0^T L(y, u)\,dt$ is weakly lower semicontinuous on $L^2(0, T; H)\times L^2(0, T; U)$, and by (3.1) (see [1, p. 107]),

$$L^\varepsilon(y_\varepsilon, u_\varepsilon)\geqq L(z_\varepsilon, w_\varepsilon),$$

where $y_\varepsilon-z_\varepsilon\to 0$ in $L^2(0, T; H)$ and $u_\varepsilon-w_\varepsilon\to 0$ in $L^2(0, T; U)$ as $\varepsilon\to 0$. Now by (3.14) and (3.16)

$$\lim_{\varepsilon\to 0}\int_0^T \|u_\varepsilon-u^*\|^2\,dt=0.$$

Hence $u^0=u^*$, $y^0=y^*$ and all conclusions of the lemma follow.

LEMMA 6. *There exist functions $p\in L^\infty(0, T; H)$ and $q\in L^2(0, T; H)$ such that for $\varepsilon\to 0$*

$$\begin{aligned} &p_\varepsilon\to p \quad \textit{weak star in } L^\infty(0, T; H),\\ &\partial_1 L^\varepsilon(y_\varepsilon, u_\varepsilon)\to q \quad \textit{weakly in } L^1(0, T; H), \end{aligned} \tag{3.17}$$

$$(q(t), B^*p(t))\in\partial L(y^*(t), u^*(t)) \quad \textit{a.e. } t\in\,]0, T[. \tag{3.18}$$

Here $\partial L: H\times U\to H\times U$ is the subdifferential of L.

Proof. We shall argue as in the proof of Theorem 1.1 in [1, p. 222]. By the definition of ∂L^ε we have

$$\begin{aligned}(\partial_1 L^\varepsilon(y_\varepsilon, u_\varepsilon), y_\varepsilon-y^*-\varrho w)+\langle\partial_2 L^\varepsilon(y_\varepsilon, u_\varepsilon), u_\varepsilon-v_0\rangle&\\ \geqq L^\varepsilon(y_\varepsilon, u_\varepsilon)-L^\varepsilon(y^*+\varrho w, v_0) \quad &\text{for all } w\in H,\end{aligned}$$

and therefore

$$\frac{\varrho}{2}|\partial_1 L^\varepsilon(y_\varepsilon, u_\varepsilon)|\leqq\langle B^*p_\varepsilon+u^*-u_\varepsilon, u_\varepsilon-v_0\rangle+L(y^*+\varrho w, v_0) \quad \text{a.e. } t\in\,]0, T[. \tag{3.19}$$

According to Assumption (i) the Hamiltonian function H and its subdifferential ∂H are locally bounded on $H \times U$. Let $v_0(t)$ be a measurable function such that $v_0(t) \in \partial_p H(y^*(t)+\rho w, 0)$ a.e. $t \in]0, T[$. For $|w| = 1$ and ϱ sufficiently small we have $\|v_0(t)\| \leqq C$ and

$$L(y^*(t)+\varrho w, v_0(t)) = -H(y^*(t)+\varrho w, 0) \leqq C \quad \text{a.e. } t \in]0, T[.$$

Then by inequality (3.19)

$$\begin{aligned} |\partial_1 L^\varepsilon(y_\varepsilon(t), u_\varepsilon(t))| \leqq (\|B^* p_\varepsilon(t)\| + \|u_\varepsilon(t) - u^*(t)\| \\ \cdot (M + \|u_\varepsilon(t)\|) + C \quad \text{a.e. } t \in]0, T[. \end{aligned} \tag{3.20}$$

On the other hand, it follows by Lemma 1 (inequality (2.9)) that for all w and v in $L^2(0, T; H)$,

$$|((\nabla \Theta_\varepsilon(w))^* v)(t)| \leqq \int_t^T |v(s)|\, ds, \qquad t \in [0, T]. \tag{3.21}$$

The latter in conjunction with (3.6) and (3.20) implies, via Gronwall's lemma,

$$|p_\varepsilon(t)| \leqq C \quad \text{a.e. } t \in]0, T[, \quad \varepsilon > 0, \tag{3.22}$$

and therefore

$$|\partial_1 L^\varepsilon(y_\varepsilon(t), u_\varepsilon(t))| \leqq C(1 + \|u_\varepsilon(t) - u^*(t)\|)(1 + \|u^*(t)\|) \quad \text{a.e. } t \in]0, T[. \tag{3.23}$$

Then by the Dunford-Pettis criterion the set $\{\partial_1 L^\varepsilon(y_\varepsilon, u_\varepsilon)\}$ is weakly compact in $L^1(0, T; H)$ and therefore there exists a subsequence convergent to zero, again denoted $\{\varepsilon\}$, such that

$$\partial_1 L^\varepsilon(y_\varepsilon, u_\varepsilon) \to q \quad \text{weakly in } L^1(0, T; H). \tag{3.24}$$

By (3.10) and (3.23) it follows that $q \in L^2(0, T; H)$. Next, by estimate (3.22) we may assume that

$$p_\varepsilon \to p \quad \text{weak star in } L^\infty(0, T; H). \tag{3.25}$$

By (3.21) and (3.23), we see that

$$|p(t)|^2 \leqq C\left(\int_t^T \|u^*(s)\|^2\, ds + T - t\right) \quad \text{a.e. } t \in]0, T[. \tag{3.26}$$

Now from (3.7), (3.9), (3.10), (3.24), and (3.25), (3.18) follows by a standard argument (see [1, p. 236]). This completes the proof of Lemma 6.

One expects from (3.6), (3.9), (3.10), (3.24), and (3.25) that the following formula holds:

$$p = -(\partial \Theta(Bu + f))^* \partial_1 L(y^*, u^*) \quad \text{on }]0, T[, \tag{3.27}$$

where $\partial\Theta$ is the "differential" of Θ in some generalized sense. We shall see below that in some important cases one may give a precise meaning to (3.27).

4. Control problems governed by (1.1). Let Ω be a bounded and open set in R^N with a sufficiently smooth boundary Γ. Let A_0 denote the second-order elliptic symmetric operator on Ω,

$$A_0 y = -\sum_{i,j=1}^{N} (a_{ij} y_{x_i})_{x_j} + ay,$$

where $a_{ij} \in C^1(\bar{\Omega})$, $a \in L^\infty(\Omega)$, $a \geqq 0$, $a_{ij} = a_{ji}$ and there exists a positive constant ω such that

$$\sum_{i,j=1}^{N} a_{ij}(x)\xi_i\xi_j \geqq \omega|\xi|^2 \quad \text{a.e. } x \in \Omega, \quad \xi \in R^N.$$

We shall denote by $\partial/\partial\nu$ the outward normal derivative corresponding to A_0, and by $a : H^1(\Omega) \times H^1(\Omega) \to R$ the bilinear form

$$a(y, \psi) = \sum_{i,j=1}^{N} \int_\Omega (a_{ij} y_{x_i} \psi_{x_j} + a y \psi)\, dx.$$

Let U be a real Hilbert space with norm $\|\cdot\|$ and inner product $\langle \cdot, \cdot \rangle$.
Consider the following distributed control problem:

Minimize

$$\int_0^T L(y(t), u(t))\, dt \quad \text{over all } u \in L^2(0, T; U) \text{ and } y \in H^{2,1}(Q), \tag{4.1}$$

subject to

$$\begin{aligned} & y_t + A_0 y + \beta(y) \ni f + Bu && \text{on } Q, \\ & \alpha_1 y + \alpha_2 \frac{\partial y}{\partial \nu} = 0 && \text{on } \Sigma = \Gamma \times]0, T[, \\ & y(x, 0) = y_0(x) && x \in \Omega. \end{aligned} \tag{4.2}$$

Here β is a maximal monotone graph in $R \times R$ such that $0 \in \beta(0)$ and α_i, $i = 1, 2$ are positive constants satisfying: $\alpha_1^2 + \alpha_2^2 \neq 0$; B is a linear continuous operator from U to $L^2(\Omega)$; and $L : L^2(\Omega) \times U \to \bar{R}$ is a lower semicontinuous convex function on $L^2(\Omega) \times U$ satisfying Assumption (i) ($H = L^2(\Omega)$). The subscript t denotes partial differentiation with respect to t. The function $f \in L^2(Q)$ is given.

Since β is maximal monotone in $R \times R$, there exists a lower semincontinuous convex function $j : R \to \bar{R}$ such that $\partial j = \beta$. The function j is uniquely defined up to an additive constant, so that we may suppose that

$$\inf \{j(y);\ y \in R\} = j(0) = 0. \tag{4.3}$$

Problem (4.1) can be put into the form (P), where $H = L^2(\Omega)$ and $F : L^2(\Omega) \to L^2(\Omega)$ is defined by

$$Fy = Ay + \tilde{B}y \quad \text{for } y \in D(A) \cap D(\tilde{B}), \tag{4.4}$$

where

$$Ay = A_0 y \quad \text{for } y \in D(A) = \left\{ y \in H^2(\Omega);\ \alpha_1 y + \alpha_2 \frac{\partial y}{\partial \nu} = 0 \text{ on } \Gamma \right\}, \tag{4.5}$$

and

$$(\tilde{B}y)(x) = \{w \in L^2(\Omega);\ w(x) \in \beta(y(x)) \text{ a.e. } x \in \Omega\}, \tag{4.6}$$

$$D(\tilde{B}) = \{y \in L^2(\Omega);\ \exists w \in L^2(\Omega);\ w(x) \in \beta(y(x)) \text{ a.e. } x \in \Omega\}. \tag{4.7}$$

It is well known (see [3]) that $F = \partial\varphi$, where

$$\varphi(y) = \tfrac{1}{2}a(y, y) + \int_\Omega j(y)\,dx + \frac{1}{2\alpha_2}\int_\Gamma y^2\,d\sigma,$$
$$D(\varphi) = \{y \in H^1(\Omega);\ j(y) \in L^1(\Omega)\}. \tag{4.8}$$

If $\alpha_2 = 0$, then

$$\varphi(y) = \tfrac{1}{2}a(y, y) + \int_\Gamma j(y)\,dx, \qquad D(\varphi) = \{y \in H_0^1(\Omega);\ j(y) \in L^1(\Omega)\}.$$

The initial data y_0 will be chosen such that

$$y_0 \in H^1(\Omega), \quad j(y_0) \in L^1(\Omega) \quad \text{if } \alpha_2 \neq 0, \tag{4.9}$$

and

$$y_0 \in H_0^1(\Omega), \quad j(y_0) \in L^1(\Omega) \quad \text{if } \alpha_2 = 0. \tag{4.9}'$$

Let ϱ be a fixed C_0^∞ function on the real axis R such that $\int_{-\infty}^{\infty} \varrho(t)\,dt = 1$, $\varrho(t) \geqq 0$, $\varrho(t) = \varrho(-t)$ for $|t| < 1$ and $\varrho(t) = 0$ for $|t| \geqq 1$. We set $\beta_\varepsilon = \varepsilon^{-1}(1 - (1 + \varepsilon\beta)^{-1})$, and define the mollifier

$$\beta^\varepsilon(y) = \int_{-\infty}^{\infty} \beta_\varepsilon(y - \varepsilon\vartheta)\varrho(\vartheta)\,d\vartheta = \varepsilon^{-1}\int_{-\infty}^{\infty} \beta_\varepsilon(\vartheta)\varrho(\varepsilon^{-1}(y - \vartheta))\,d\vartheta. \tag{4.10}$$

Clearly β^ε is infinitely differentiable and monotone on R. Consider the family of operators $F_\varepsilon : H \to H$,

$$F_\varepsilon(y) = Ay + \beta^\varepsilon(y) \quad \text{for } y \in D(F_\varepsilon) = D(A), \quad \varepsilon > 0. \tag{4.11}$$

Obviously, $F_\varepsilon = \partial\varphi^\varepsilon$ where

$$\varphi^\varepsilon(y) = \varphi(y) + \int_\Omega j^\varepsilon(y)\,dx - \int_\Omega j(y)\,dx, \tag{4.12}$$

and

$$j^\varepsilon(y) = \int_{-\infty}^{\infty} j_\varepsilon(y - \varepsilon\vartheta)\varrho(\vartheta)\,d\vartheta, \qquad y \in R. \tag{4.13}$$

Here $j_\varepsilon(y) = \int_0^y \beta_\varepsilon(r)\,dr = \inf\{j(r) + (2\varepsilon)^{-1}|y - r|^2;\ r \in R\}$. We now show that φ^ε and F_ε satisfy Assumption (iii). Condition (iiib) is obviously satisfied with $\psi(y) = \frac{1}{2}a(y, y) + (1/2\alpha_2)\int_\Gamma y^2\,d\sigma$ for $\alpha_2 \neq 0$, and $\psi(y) = \frac{1}{2}a(y, y)$ for $\alpha_2 = 0$. We have

$$j^\varepsilon(y) \leqq j(y) + \frac{\varepsilon}{2}\int_{-\infty}^{\infty} \vartheta^2\varrho(\vartheta)\,d\vartheta \quad \text{for all } y \in R, \quad \varepsilon > 0, \tag{4.14}$$

which along with (4.12) implies (2.5). As regards the remaining condition in (iiia), we proceed as follows. Let $y_\varepsilon \to y$ strongly in $L^2(\Omega)$ for $\varepsilon \to 0$. Thus we may assume that $y_\varepsilon(x) \to y(x)$ a.e. $x \in \Omega$. By (4.13) it follows that

$$j^\varepsilon(y_\varepsilon(x)) = \int_{-\infty}^{\infty} (j(z_\varepsilon(x, \vartheta)) + (2\varepsilon)^{-1}|z_\varepsilon(x, \vartheta) - y_\varepsilon(x) - \varepsilon\vartheta|)^2\varrho(\vartheta)\,d\vartheta, \tag{4.15}$$

where $z_\varepsilon(x, \vartheta)$ is the minimum point of the function $r \to j(r) + (2\varepsilon)^{-1}|r - y_\varepsilon(x) - \varepsilon\vartheta|^2$. Suppose that $\{\int_\Omega j^\varepsilon(y_\varepsilon(x))\,dx\}$ is bounded for $\varepsilon \to 0$. Then we infer that $(z_\varepsilon(x, \vartheta) - y_\varepsilon(x) - \varepsilon\vartheta) \to 0$ a.e. $x \in \Omega$, $\vartheta \in [-1, 1]$, and therefore

$$z_\varepsilon(x, \vartheta) \to y(x) \quad \text{a.e. } x \in \Omega, \ \vartheta \in]-1, 1[. \tag{4.16}$$

Since j is lower semicontinuous, the Fatou lemma along with (4.15) yields

$$\liminf_{\varepsilon\to 0}\int_\Omega j^\varepsilon(y_\varepsilon(x))\,dx \geqq \int_\Omega\int_{-\infty}^{\infty} j(y(x))\varrho(\vartheta)\,d\vartheta\,dx = \int_\Omega j(y(x))\,dx,$$

as claimed. By Green's formula,

$$\int_\Omega (F_\varepsilon(y)-F_\lambda(z))(y-z)\,dx \geqq \int_\Omega (\beta^\varepsilon(y)-\beta^\lambda(z))(y-z)\,dx, \qquad y, z\in D(A). \tag{4.17}$$

On the other hand, we have

$$\begin{aligned}&(\beta^\varepsilon(y)-\beta^\lambda(z))(y-z)\\ &\geqq(\varepsilon-\lambda)\int_{-\infty}^{\infty}\vartheta(\beta_\varepsilon(y-\varepsilon\vartheta)-\beta_\lambda(z-\lambda\vartheta))\varrho(\vartheta)\,d\vartheta\\ &\quad+\int_{-\infty}^{\infty}(\varepsilon\beta_\varepsilon(y-\varepsilon\vartheta)-\lambda\beta_\lambda(z-\lambda\vartheta))(\beta_\varepsilon(y-\varepsilon\vartheta)-\beta_\lambda(z-\lambda\vartheta))\varrho(\vartheta)\,d\vartheta,\end{aligned} \tag{4.18}$$

because β is monotone. Next by (4.10) we see that

$$|\beta^\varepsilon(y)-\beta_\varepsilon(y)|\leqq\int_{-\infty}^{\infty}|\beta_\varepsilon(y)-\beta_\varepsilon(y-\varepsilon\vartheta)|\varrho(\vartheta)\,d\vartheta\leqq 1, \tag{4.19}$$

because β_ε is Lipschitzian with constant $1/\varepsilon$. Hence

$$|\beta_\varepsilon(y-\varepsilon\vartheta)|\leqq|\beta^\varepsilon(y)|+1 \quad \text{for } \varepsilon>0, \quad \vartheta\in[-1,1], \quad y\in R. \tag{4.20}$$

Estimates (4.19) and (4.20) inserted in (4.18) yield (2.7). Since β^ε is differentiable the operator Θ_ε, defined in § 2, is Gâteaux differentiable on $L^2(0, T; L^2(\Omega))$, and its Gâteaux differential $\nabla\Theta_\varepsilon(w)$ is given by $\nabla\Theta_\varepsilon(w)v = z$, where $z\in H^{2,1}(Q)$ is the solution to

$$\begin{aligned}&z_t+Az+\nabla\beta^\varepsilon(\Theta_\varepsilon w)z = v \quad &&\text{on } Q,\\ &z(x,0)=0 &&\text{on } \Omega.\end{aligned} \tag{4.21}$$

Since $\nabla\beta^\varepsilon(\Theta_\varepsilon w)\in L^\infty(Q)$, (4.21) has for each $v\in L^2(Q)$ a unique solution $z\in H^{2,1}(Q)\cap L^2(0, T; D(A))$.

The dual operator $(\nabla\Theta_\varepsilon(w))^*$ is given by $(\nabla\Theta_\varepsilon(w))^*q=-\zeta$, where $\zeta\in H^{2,1}(Q)\cap L^2(0, T; D(A))$ is the solution to

$$\begin{aligned}&\zeta_t-A\zeta-\nabla\beta^\varepsilon(\Theta_\varepsilon w)\zeta = q \quad &&\text{on } Q,\\ &\zeta(x,T)=0 &&\text{on } \Omega.\end{aligned} \tag{4.22}$$

Let $(y^*, u^*)\in H^{2,1}(Q)\times L^2(0, T; U)$ be any optimal pair in problem (4.1). Since Assumptions (i)–(iii) are satisfied we may apply the results established in § 3. Thus there exist sequences $\{y_\varepsilon\}\subset H^{2,1}(Q)\cap L^2(0, T; D(A))$, $\{u_\varepsilon\}\subset L^2(0, T; U)$, $\{q_\varepsilon\}\subset L^2(Q)$, $\{p_\varepsilon\}\subset H^{2,1}(Q)$, and functions $p\in L^\infty(0, T; L^2(\Omega))$, $q\in L^2(Q)$ satisfying the equations

$$(y_\varepsilon)_t+Ay_\varepsilon+\beta^\varepsilon(y_\varepsilon)=Bu_\varepsilon+f \qquad \text{on } Q, \tag{4.23}$$

$$(p_\varepsilon)_t-Ap_\varepsilon-\nabla\beta^\varepsilon(y_\varepsilon)p_\varepsilon=q_\varepsilon \qquad \text{on } Q, \tag{4.24}$$

$$y_\varepsilon(x,0)=y_0(x), \qquad p_\varepsilon(x,T)=0 \quad \text{on } \Omega, \tag{4.25}$$

$$(q(t), (B^*p)(t))\in\partial L(y^*(t), u^*(t)) \quad \text{a.e. } t\in\,]0, T[, \tag{4.26}$$

and

$$(4.26)' \qquad (q_\varepsilon(t)\, B^* p_\varepsilon(t) + u^*(t) - u_\varepsilon(t)) = \partial L^\varepsilon(y_\varepsilon(t), u_\varepsilon(t)) \quad \text{a.e. } t \in]0, T[,$$

$$(4.27) \qquad y_\varepsilon \to y^* \quad \text{strongly in } C(0, T; L^2(\Omega)),$$

$$(4.28) \qquad u_\varepsilon \to u^* \quad \text{strongly in } L^2(0, T; U),$$

$$(4.29) \qquad p_\varepsilon \to p \quad \text{weak star in } L^\infty(0, T; L^2(\Omega)),$$

$$(4.30) \qquad q_\varepsilon \to q \quad \text{weakly in } L^1(0, T; L^2(\Omega)).$$

We notice that by (3.15), $\{Ay_\varepsilon + \beta^\varepsilon(y_\varepsilon)\}$ remain in a bounded subset of $L^2(Q)$. Since β^ε is monotone we deduce, by a standard argument involving Green's formula,

$$(4.31) \qquad \|Ay_\varepsilon\|_{L^2(Q)} + \|\beta^\varepsilon(y_\varepsilon)\|_{L^2(Q)} \leqq C \quad \text{for all } \varepsilon > 0,$$

and therefore

$$(4.32) \qquad \begin{aligned} Ay_\varepsilon &\to Ay^* \quad \text{weakly in } L^2(Q), \\ \beta^\varepsilon(y_\varepsilon) &\to h \quad \text{weakly in } L^2(Q), \end{aligned}$$

where

$$h = Bu^* + f - Ay^* - y_t^*, \qquad h(x, t) \in \beta(y^*(x, t)) \quad \text{a.e. } (x, t) \in Q.$$

In particular, it follows by (4.32) that

$$y_\varepsilon \to y^* \quad \text{weakly in } H^{2,1}(Q).$$

Since $\{(y_\varepsilon)\}$ is bounded in $L^2(Q)$ and $\{y_\varepsilon\}$ in $L^2(0, T; H^2(\Omega))$, it follows that (see, e.g., [10, p. 70]) $\{y_\varepsilon\}$ is a compact subset of $L^2(0, T; H^1(\Omega))$. Hence (4.27) can be strengthened to

$$(4.27)' \qquad \begin{aligned} y_\varepsilon \to y^* \quad &\text{strongly in } C(0, T; L^2(\Omega)) \cap L^2(0, T; H^1(\Omega)), \\ &\text{and weakly in } H^{2,1}(Q). \end{aligned}$$

Now we multiply both sides of (4.24) by p_ε and integrate with respect to (x, t), to obtain, since

$$(4.33) \qquad \begin{aligned} &\int_Q A_0 p_\varepsilon p_\varepsilon \, dx\, dt \geqq C \int_Q |\text{grad } p_\varepsilon|^2 \, dx\, dt, \qquad \nabla\beta^\varepsilon(y_\varepsilon) \geqq 0, \\ &\|p_\varepsilon\|_{L^2(0,T;H^1(\Omega))} \leqq C \quad \text{for all } \varepsilon > 0. \end{aligned}$$

Next we multiply (4.24) by $\xi(p_\varepsilon)$ where $\xi(r)$ is a smooth, monotone and bounded approximation to sign r. Applying Green's formula and letting $\xi \to \text{sign } r$, we get

$$(4.34) \qquad \int_Q |\nabla\beta^\varepsilon(y_\varepsilon) p_\varepsilon| \, dx\, dt \leqq C \quad \text{for all } \varepsilon > 0.$$

Thus selecting a subsequence, if necessary, we may assume that

$$(4.35) \qquad \begin{aligned} p_\varepsilon \to p \quad &\text{weakly in } L^2(0, T; H^1(\Omega)), \\ &\text{and weak star in } L^\infty(0, T; L^2(\Omega)), \end{aligned}$$

and

$$(4.36) \qquad \nabla\beta^\varepsilon(y_\varepsilon) p_\varepsilon \to \pi_p \quad \text{weak star in } \mathcal{M}(Q),$$

where $\mathcal{M}(Q)$ denotes the space of all bounded measures on Q. Letting $\varepsilon \to 0$ in (4.24),

we see that $p \in L^\infty(0, T; L^2(\Omega)) \cap L^2(0, T; H^1(\Omega))$ is a solution (in the sense of distributions) to

$$\begin{aligned} &p_t - Ap - \pi_p = q \quad \text{on } Q, \\ &p(\cdot, T) = 0 \qquad \text{on } \Omega. \end{aligned} \tag{4.37}$$

In other words,

$$\int_Q p\varkappa_t \, dx \, dt + \int_0^T a(p, \varkappa) \, dt + \frac{\alpha_1}{\alpha_2} \int_\Sigma p\varkappa \, d\sigma \, dt + \pi_p(\varkappa) + \int_Q q\varkappa \, dx \, dt = 0 \quad \text{for all } \varkappa \in D_{\Omega,0}, \tag{4.37}'$$

with the usual modification in the case $\alpha_2 = 0$. Here $D_{\Omega,0}$ denotes the space of all infinitely differentiable functions on $\bar{Q}$ that vanish for $t \in [0, \delta]$ for some $\delta > 0$.

LEMMA 7. *$p_\varepsilon \to p$ strongly in $L^2(Q)$ on some subsequence $\varepsilon \to 0$.*

Proof. By virtue of the Sobolev embedding theorem, $H_0^s(\Omega) \subset C(\bar{\Omega})$ for $s > N/2$. Hence $L^1(\Omega) \subset H^{-s}(\Omega)$, and estimate (4.34) implies that $\{\nabla \beta^\varepsilon(y_\varepsilon) p_\varepsilon\}$ is a bounded subset of $L^1(0, T; H^{-s}(\Omega))$. Next by (4.33) we see that $\{Ap_\varepsilon\}$ is bounded in $L^1(0, T; H^{-s}(\Omega))$, and therefore $\{(p_\varepsilon)_t\}$ is bounded in $L^1(0, T; H^{-s}(\Omega))$ for $s > N/2$. Since by (4.33), $\{p_\varepsilon\}$ is bounded in $L^2(0, T; L^2(\Omega))$ we may conclude in virtue of the Arzela-Ascoli theorem that $\{p_\varepsilon\}$ is a relatively compact subset of $C(0, T; H^{-s}(\Omega))$. Hence $p \in C(0, T; H^{-s}(\Omega))$ and for $\varepsilon \to 0$,

$$\|p_\varepsilon(t) - p(t)\|_{H^{-s}(\Omega)} \to 0 \quad \text{uniformly on } [0, T]. \tag{4.38}$$

On the other hand, since $H^1(\Omega)$ is compactly embedded in $L^2(\Omega)$, for each $\eta > 0$ there is $C(\eta)$ such that (see [11, p. 102]),

$$\|p_\varepsilon(t) - p(t)\|_{L^2(\Omega)} \leqq \eta \|p_\varepsilon(t) - p(t)\|_{H^1(\Omega)} + C(\eta) \|p_\varepsilon(t) - p(t)\|_{H^{-s}(\Omega)},$$

and therefore with other constants C and $C(\eta)$ we have

$$\|p_\varepsilon - p\|_{L^2(Q)} \leqq C\eta + C(\eta) \|p_\varepsilon - p\|_{L^2(0,T;H^{-s}(\Omega))} \quad \text{for all } \eta > 0.$$

Along with (4.38) the latter implies Lemma 7, as claimed.

Summarizing, we get

THEOREM 1. *Let $(y^*, u^*) \in H^{2,1}(Q) \times L^2(0, T; U)$ be an optimal pair for problem* (4.1). *Then there exist functions $p \in L^\infty(0, T; L^2(\Omega)) \cap L^2(0, T; H^1(\Omega)) \cap C(0, T; H^{-s}(\Omega))$, $q \in L^2(Q)$ and a bounded measure $\pi_p \in \mathcal{M}(Q)$ which satisfy* (4.26) *and* (4.37). *Furthermore, (y^*, u^*, p, q, π_p) are limits in the sense of* (4.27)–(4.30) *and* (4.27)', (4.32), (4.36) *of solutions $(y_\varepsilon, u_\varepsilon, p_\varepsilon, q_\varepsilon)$ to approximating equations* (4.23)–(4.25).

Remark. If $\alpha_2 = 0$, then by (4.33) and (4.35) it follows that $p \in L^2(0, T; H_0^1(\Omega))$. It is natural to call such a function p a *coextremal* of problem (4.1) corresponding to y^*.

Now we shall discuss some consequences of Theorem 1 in two notable cases.

1°. *β is locally Lipschitzian.* According to (4.27), we may assume that $y_\varepsilon(x, t) \to y^*(x, t)$ a.e. $(x, t) \in Q$. Then by Egorov's theorem, for each $\eta > 0$ there exist a measurable subset E_η of Q and $C_\eta > 0$ such that the Lebesgue measure $m(Q \backslash E_\eta)$ of $Q \backslash E_\eta$ is $\leqq \eta$, $|y_\varepsilon(x, t)| \leqq C$ a.e. $(x, t) \in E_\eta$ and

$$y_\varepsilon(x, t) \to y^*(x, t) \quad \text{uniformly on } E_\eta \quad \text{for } \varepsilon \to 0.$$

Since $\nabla \beta^\varepsilon$ is uniformly bounded on bounded subsets, selecting a further subsequence

$\varepsilon_n \to 0$, we may assume that

$$\nabla\beta^{\varepsilon_n}(y_{\varepsilon_n}) \to \pi_\eta \quad \text{weak star in } L^\infty(E_\eta) \quad \text{for } n \to \infty. \tag{4.39}$$

Then according to Mazur's theorem, π_η is the strong limit in $L^1(E_\eta)$ of a sequence $\{\mu^m\}$ consisting of convex combinations of the $\nabla\beta^{\varepsilon_n}(y_{\varepsilon_n})$. In other words,

$$\pi_\eta(x, t) = \lim_{m\to\infty} \mu^m(x, t) \quad \text{a.e. } (x, t) \in E_\eta,$$

where μ^m are of the form

$$\mu^m(x, t) = \sum_{i\in I_m} \alpha_m^i \nabla\beta^{\varepsilon_i}(y_{\varepsilon_i}(x, t)).$$

For each m, I_m is a finite set of natural numbers of $[m, +\infty[$ and $\alpha_m^i \geqq 0$, $\sum_{i\in I_m} \alpha_m^i = 1$. Arguing as in the proof of Theorem 2 in [2] we finally find

$$\pi_\eta(x, t) \in \partial\beta(y^*(x, t)) \quad \text{a.e. } (x, t) \in E_\eta,$$

where $\partial\beta(y)$ is the *generalized gradient* of β in the sense of Clarke (see [5], [16]); i.e., $\partial\beta(y)$ is the convex hull of all elements of the form $\{\lim_{n\to\infty} \nabla\beta(y_n)\}$ where $y_n \to y$ and $\nabla\beta(y_n)$ exist. On the other hand, as seen in Lemma 7, $p_\varepsilon \to p$ strongly in $L^2(Q)$ for $\varepsilon \to 0$. Extracting a further subsequence and modifying the subset E_η, if necessary, we may assume that $p \in L^\infty(E_\eta)$ and $p_\varepsilon \to p$ uniformly on E_η for $\varepsilon \to 0$. It will be more convenient to regard the measure π_p as an element of the dual space $(L^\infty(Q))^*$ of $L^\infty(Q)$. As a matter of fact, by (4.34) and (4.36) we see that selecting a further subsequence (eventually a generalized one) we may assume that

$$\nabla\beta^\varepsilon(y_\varepsilon)p_\varepsilon \to \pi_p \quad \text{weak star in } (L^\infty(Q))^*.$$

Along with (4.39) the latter implies that $\pi_p = \pi_\eta p$ on E_η. In particular we may infer that for $\eta \neq \eta'$ one has

$$(\pi_{\eta'} - \pi_\eta)p = 0 \quad \text{a.e. on } E_\eta \cap E_{\eta'}. \tag{4.40}$$

On the other hand, estimate (4.34) yields

$$\|\pi_\eta p\|_{L^1(E_\eta)} \leqq C \quad \text{for all } \eta > 0. \tag{4.41}$$

Define on $E = \bigcup_{\eta>0} E_\eta$ the measurable function μ,

$$\mu(x, t) = \pi_\eta(x, t)p(x, t) \quad \text{for } (x, t) \in E_\eta.$$

By (4.40) and (4.41) we see that μ is well defined and belongs to $L^1(Q)$.

Let $\pi_p = (\pi_p)_a + (\pi_p)_s$ be the Lebesgue decomposition of $\pi_p \in (L^\infty(Q))^*$ into the *absolutely continuous* part $(\pi_p)_a \in L^1(Q)$ and the *singular* part $(\pi_p)_s$ (see, e.g., [15]). We have

$$\int_{E_\eta} (\pi_p)_a \varkappa \, dx \, dt + (\pi_p)_s(\varkappa) = \int_{E_\eta} \mu\varkappa \, dx \, dt,$$

for all $\varkappa \in L^\infty(Q)$ which vanish outside E_η. Next the "singularity" of $(\pi_p)_s$ implies the existence of a nondecreasing sequence of measurable sets $Q_k \subset Q$ such that $m(Q\backslash Q_k) \to 0$ for $k \to \infty$ and $(\pi_p)_s = 0$ on $L^\infty(Q_k)$. Hence

$$\int_{E\cap Q_k} ((\pi_p)_a - \mu)\varkappa \, dx \, dt = 0,$$

for all $\varkappa \in L^\infty(Q)$ which vanish outside of some $E_\eta \cap Q_k$. Since $m(Q\backslash E) = 0$, we may

infer that $\mu = (\pi_p)_a$ a.e. on Q. Remembering that $\pi_\eta(x,t) \in \partial\beta(y^*(x,t))$ a.e. $(x,t) \in Q$, we may therefore conclude that

$$(\pi_p)_a(x,t) \in \partial\beta(y^*(x,t))p(x,t) \quad \text{a.e. } (x,t) \in Q. \tag{4.42}$$

Now we shall assume that the following condition is satisfied,

$$\sup\{|wy|;\ w \in \partial\beta(y)\} \leqq C(|\beta(y)| + |y|^2 + 1), \qquad y \in R, \tag{4.43}$$

or equivalently (we recall that $\nabla\beta(y) \geqq 0$ a.e. $y \in R$),

$$\nabla\beta(y)|y| \leqq C(|\beta(y)| + |y|^2 + 1) \quad \text{a.e. } y \in R. \tag{4.43}'$$

Parenthetically we notice that every function β satisfying (4.43) is of polynomial growth at ∞. By (4.43) it follows after some computation involving (4.10) and (4.19), (4.20) that

$$|\nabla\beta^\varepsilon(y)y| \leqq C(|\beta^\varepsilon(y)| + |y|^2 + 1) \quad \text{for all } y \in R,$$

where C is a positive constant independent of ε. For each $\varepsilon > 0$ and natural number n we set

$$E_n^\varepsilon = \{(x,t) \in Q;\ |y_\varepsilon(x,t)| \leqq n\}.$$

We have

$$|\nabla\beta^\varepsilon(y_\varepsilon(x,t))| \leqq C_n \quad \text{for } (x,t) \in E_n^\varepsilon,$$

because β is locally Lipschitzian on R. Let E be an arbitrary measurable subset of Q. We have

$$\begin{aligned}
\left|\int_E p_\varepsilon(x,t)\nabla\beta^\varepsilon(y_\varepsilon(x,t))\,dx\,dt\right| &\leqq \int_{E\cap E_n^\varepsilon} |p_\varepsilon(x,t)||\nabla\beta^\varepsilon(y_\varepsilon(x,t))|\,dx\,dt \\
&\quad + \int_{E\cap C(E_n^\varepsilon)} |p_\varepsilon(x,t)||\nabla\beta^\varepsilon(y_\varepsilon(x,t))|\,dx\,dt \\
&\leqq C_n \int_E |p_\varepsilon(x,t)|\,dx\,dt + Cn^{-1}\int_{E\cap E_n^\varepsilon} |\beta_\varepsilon(y_\varepsilon)p_\varepsilon(x,t)|\,dx\,dt \\
&\quad + C\int_{E\cap E_n^\varepsilon} |y_\varepsilon(x,t)|\,dx\,dt + Cn^{-1}.
\end{aligned}$$

Since $\{\beta^\varepsilon(y_\varepsilon)\}$, $\{y_\varepsilon\}$ and $\{p_\varepsilon\}$ are bounded in $L^2(Q)$ it follows that for each $\gamma > 0$

$$\left|\int_E p_\varepsilon\nabla\beta^\varepsilon\,dx\,dt\right| \leqq Cn^{-1} + \gamma,$$

if $m(E) \leqq \zeta(\gamma)$. Since n is arbitrary we see that the family $\{\int_E p_\varepsilon\nabla\beta^\varepsilon(y_\varepsilon)\,dx\,dt\}$ is equicontinuous and so $\{p_\varepsilon\nabla\beta^\varepsilon(y_\varepsilon)\}$ is, by virtue of the Dunford-Pettis criterion, weakly compact in $L^1(Q)$. Hence $\pi_p = (\pi_p)_a \in L^1(Q)$, and the coextremal p satisfies the equation

$$\begin{aligned}
&p_t - Ap - \partial\beta(y^*)p \ni q && \text{on } Q, \\
&p(x,T) = 0 && \text{on } \Omega.
\end{aligned} \tag{4.44}$$

Since the functions q and $\pi_p = -q + p_t - Ap$ are in the space $L^1(Q)$, the solution p to (4.44) is continuous from $[0,T]$ to $L^1(\Omega)$.

We have therefore proved the following theorem.

THEOREM 2. *Under the hypotheses of Theorem* 1, *assume that* β *is locally Lipschitzian. Then the absolutely continuous part* $(\pi_p)_a$ *of the measure* π_p *satisfies* (4.42). *If in addition, condition* (4.43) *holds, then the coextremal* $p \in L^\infty(0, T; L^2(\Omega)) \cap L^2(0, T; H^1(\Omega))$ *is continuous from* $[0, T]$ *to* $L^1(\Omega)$ *and satisfies* (4.44). (*If* $\alpha_2 = 0$ *then* $p \in L^2(0, T; H_0^1(\Omega))$).

By (4.42) the coextremal p is the solution to

$$(p_t - Ap)_a \in \partial\beta(y^*)p + q \quad \text{a.e. on } Q,$$
$$p(\cdot, T) = 0 \quad \text{a.e. on } \Omega,$$

where $(p_t - Ap)_a = (\pi_p)_a$ is the absolutely continuous part of the measure $p_t - Ap$. If $\{Q_k\}$ is the sequence of measurable subsets of Q which occur in the definition of $(\pi_p)_s$, then the above equation can be equivalently written as (see (4.37)')

$$\int_Q p\varkappa_t \, dx \, dt + \int_0^T a(p, \varkappa) \, dt + \frac{\alpha_1}{\alpha_2} \int_\Sigma p\varkappa \, d\sigma \, dt + \int_Q g\varkappa \, dx \, dt + \int_Q a\varkappa \, dx \, dt = 0,$$

for all $\varkappa \in D_{\Omega,0}$ which vanish outside of some Q_k. Here $g \in L^1(Q)$ satisfies

$$g(x, t) \in \partial\beta(y^*(x, t))p \quad \text{a.e. } (x, t) \in Q.$$

2°. *A unilateral problem.* Let the graph β be defined by

$$\beta(r) = \begin{cases} 0 & \text{for } r > 0, \\ R^- & \text{for } r = 0, \\ \varnothing & \text{for } r < 0. \end{cases} \tag{4.45}$$

Then the state equation (4.2) reduces to the unilateral problem (see [3], [4]),

$$\begin{aligned} & y(x, t) \geqq 0 && \text{a.e. on } Q, \\ & y_t + A_0 y = f + Bu && \text{a.e. on } \{(x, t);\ y(x, t) > 0\}, \\ & y_t = \max\{f + Bu, 0\} && \text{a.e. on } \{(x, t);\ y(x, t) = 0\}, \\ & \alpha_1 y + \alpha_2 \frac{\partial y}{\partial \nu} = 0 && \text{a.e. on } \Sigma, \\ & y(\cdot, 0) = y_0 && \text{a.e. on } \Omega. \end{aligned} \tag{4.46}$$

According to (4.9), the initial data y_0 must satisfy

$$y_0 \in H^1(\Omega), \qquad y_0(x) \geqq 0 \quad \text{a.e. } x \in \Omega.$$

We have

$$\beta^\varepsilon(y) = \varepsilon^{-1} \int_{\varepsilon^{-1}y}^\infty (y - \varepsilon\vartheta)\varrho(\vartheta) \, d\vartheta \quad \text{for all } y \in R,$$

so that trivially,

$$\nabla\beta^\varepsilon(y) = \varepsilon^{-1} \int_{\varepsilon^{-1}y}^\infty \varrho(\vartheta) \, d\vartheta \quad \text{for all } y \in R.$$

Therefore

$$|y_\varepsilon \nabla\beta^\varepsilon(y_\varepsilon)p_\varepsilon - p_\varepsilon\beta^\varepsilon(y_\varepsilon)| = |p_\varepsilon \int_{\varepsilon^{-1}y_\varepsilon}^{\infty} \varrho(\vartheta)\,d\vartheta| \leqq \varepsilon|\nabla\beta^\varepsilon(y_\varepsilon)p_\varepsilon|. \tag{4.47}$$

Define ζ_ε and γ_ε on Q by

$$\zeta_\varepsilon(x,t) = \begin{cases} 0 & \text{if } |y_\varepsilon(x,t)| > \varepsilon, \\ 1 & \text{if } |y_\varepsilon(x,t)| \leqq \varepsilon; \end{cases}$$

$$\gamma_\varepsilon(x,t) = \begin{cases} 0 & \text{if } y_\varepsilon(x,t) > -\varepsilon, \\ 1 & \text{if } y_\varepsilon(x,t) \leqq -\varepsilon. \end{cases}$$

One has

$$p_\varepsilon\beta^\varepsilon(y_\varepsilon) = \varepsilon^{-1}p_\varepsilon\zeta_\varepsilon \int_{\varepsilon^{-1}y_\varepsilon}^{1} (y_\varepsilon - \varepsilon\vartheta)\varrho(\vartheta)\,d\vartheta + \varepsilon^{-1}p_\varepsilon y_\varepsilon\gamma_\varepsilon \quad \text{on } Q,$$

and this yields

$$|p_\varepsilon\beta^\varepsilon(y_\varepsilon)| \leqq 2\varepsilon|\nabla\beta^\varepsilon(y_\varepsilon)p_\varepsilon|(\zeta_\varepsilon + \varepsilon^{-1}|y_\varepsilon|\gamma_\varepsilon) \quad \text{a.e. on } Q. \tag{4.48}$$

Since by (4.34), $\{\nabla\beta^\varepsilon(y_\varepsilon)p_\varepsilon\}$ is bounded in $L^1(Q)$, and by virtue of (4.31) $\varepsilon^{-1}y_\varepsilon\gamma_\varepsilon = \beta^\varepsilon(y_\varepsilon)\gamma_\varepsilon$ remain in a bounded subset of $L^2(\Omega)$, there exists a sequence $\{\varepsilon\}$ convergent to zero such that

$$p_\varepsilon(x,t)\beta^\varepsilon(y_\varepsilon(x,t)) \to 0 \quad \text{a.e. } (x,t) \in Q. \tag{4.49}$$

This implies in conjunction with (4.47)

$$y_\varepsilon(x,t)\nabla\beta^\varepsilon(y_\varepsilon(x,t))p_\varepsilon(x,t) \to 0 \quad \text{a.e. } (x,t) \in Q. \tag{4.50}$$

By Lemma 7, $p_\varepsilon \to p$ in $L^2(Q)$, and by (4.32), $\beta^\varepsilon(y_\varepsilon) \to f + Bu^* - Ay^* - y_t^*$ weakly in $L^2(Q)$. This along with (4.49) implies that

$$(y_t^* + Ay^* - Bu^* - f)p = 0, \quad \text{a.e. on } Q. \tag{4.51}$$

Since $\nabla\beta^\varepsilon(y_\varepsilon)p_\varepsilon \to \pi_p$ weak star in $\mathcal{M}(Q)$, and $y_\varepsilon \to y^*$ in $C(0, T; L^2(\Omega))$ we have by (4.50) that

$$y^*(\pi_p)_a = 0 \quad \text{a.e. on } Q. \tag{4.52}$$

(This follows by the same argument as in the proof of (4.42)). We notice that the solution $y \in H^{2,1}(Q)$ to problem (4.46) satisfies the equation (see [3, II.2.4]),

$$y_t = 0 \quad \text{a.e. on } \{(x,t) \in Q;\ y(x,t) = 0\}.$$

This fact combined with (4.51) yields

$$p(Bu^* + f) = 0 \quad \text{a.e. on } \{(x,t) \in Q;\ y^*(x,t) = 0\}.$$

We have therefore proved

THEOREM 3. *Let* $(y^*, u^*) \in H^{2,1}(Q) \times L^2(0, T; U)$ *be an optimal pair for problem* (4.1) *with state equation* (4.46). *Then there exist the functions* $p \in L^\infty(0, T; L^2(\Omega)) \cap L^2(0, T; H^1(\Omega))$ *and* $q \in L^2(Q)$ *satisfying along with* y^* *and* u^* *the*

following system of equations.

(4.53) $y^* \geqq 0$ a.e. *on* Q,

(4.54) $y_t^* + A_0 y^* = f + Bu^*$ a.e. *on* $\{(x, t); y^*(x, t) > 0\}$,

(4.55) $y_t^* = 0$ a.e. *on* $\{(x, t); y^*(x, t) = 0\}$,

(4.56) $\alpha_1 y^* + \alpha_2 \dfrac{\partial y^*}{\partial \nu} = 0$ a.e. *on* Σ,

(4.57) $y^*(x, 0) = y_0(x)$ a.e. *on* Ω,

(4.58) $(p_t - A_0 p)_a = q$ a.e. *on* $\{(x, t) \in Q; y^*(x, t) > 0\}$,

(4.59) $\alpha_1 p + \alpha_2 \dfrac{\partial p}{\partial \nu} = 0$ *on* Σ; $\quad p(0, T) = 0$ *on* Ω,

(4.60) $p = 0$ a.e. *on* $\{(x, t); y^*(x, t) = 0\}$
$\cap \{(x, t); (Bu^*)(x, t) + f(x, t) \neq 0\}$,

(4.61) $(q(t), (B^* p)(t)) \in \partial L(y^*(t), u^*(t))$ a.e. *on* $]0, T[$.

If $\alpha_2 = 0$, *then* $p \in L^2(0, T; H_0^1(\Omega))$.

To give a precise meaning to (4.58), (4.59) we notice that by definition of the "singular" part $(\pi_p)_s$ of measure π_p, which for convenience will be regarded as an element of the space $(L^\infty(Q))^*$, there exists an increasing sequence $\{E_k\}$ of measurable sets satisfying $\cup_{k=1} E_k = \{(x, t) \in Q; y^*(x, t) > 0\}$ such that $(\pi_p)_s(\xi) = 0$ for any $\xi \in L^\infty(Q)$ which vanishes almost everywhere outside of some E_k. Then keeping in mind the approximating equations (4.23)–(4.26), we may write (4.58) and (4.59) as

$$\int_Q p\varkappa_t \, dx \, dt + \int_0^T a(p, \varkappa) \, dt + \frac{\alpha_1}{\alpha_2} \int_\Sigma p\varkappa \, d\sigma \, dt + \int_Q q\varkappa \, dx \, dt = 0, \tag{4.62}$$

for all $\varkappa \in D_{\Omega,0}$ which vanish outside of some E_k.

Remarks. 1°. Theorem 3 is closely related to the main result given by Saguez in [17] on control-constrained quadratic problems governed by (4.46). However there is not a large overlap and the methods are quite different.

2°. If $\{y_\varepsilon\}$ is a compact subset of $C(\bar{Q})$ (this happens for instance if the $\{y_\varepsilon\}$ remain in a bounded subset of $W_q^{2,1}(Q)$ where $q > (N+2)/2$, and in particular if $N = 1$), then $y_\varepsilon \to y^*$ uniformly on $\bar{Q}$ and by (4.48) we see that $p_\varepsilon \beta^\varepsilon(y_\varepsilon) \to 0$ strongly in $L^1(Q)$. Then (4.47) implies that $y_\varepsilon \nabla \beta^\varepsilon(y_\varepsilon) p_\varepsilon \to 0$ strongly in $L^1(Q)$ for $\varepsilon \to 0$ and therefore $\pi_p y^* = 0$ on Q. Then (4.58) becomes

$$p_t - A_0 p = q \quad \text{on } \{(x, t) \in Q; y^*(x, t) > 0\}. \tag{4.63}$$

5. Control problems with state equation (1.2). We shall study here the following control problem:

Minimize

$$\int_0^T L(y(t), u(t)) \, dt \quad \text{over all } u \in L^2(0, T; U) \text{ and } y \in H^{2,1}(Q) \tag{5.1}$$

subject to (1.2).

The function L satisfies Assumption (i) and A_0, B, U, and β satisfy the conditions stated in § 4. One assumes in addition that A_0 is coercive, i.e., $a(x) \geqq \omega_1 > 0$ a.e. $x \in \Omega$.

To rewrite problem (5.1) in the abstract form (P) we take $H = L^2(\Omega)$ and define F: $L^2(\Omega) \to L^2(\Omega)$,

$$Fy = A_0 y = -\sum_{i,j=1}^{N} (a_{ij} y_{x_i})_{x_j} + ay \quad \text{for } y \in D(F), \tag{5.2}$$

where

$$D(F) = \left\{ y \in H^2(\Omega); \frac{\partial y}{\partial \nu} + \beta(y) \ni 0 \text{ a.e. on } \Gamma \right\}. \tag{5.3}$$

It is well known (see [3]) that $F = \partial\varphi$ where

$$\varphi(y) = \tfrac{1}{2} a(y, y) + \int_\Gamma j(y)\, d\sigma \quad \text{for } y \in H^1(\Omega), \tag{5.4}$$

where $j: R \to \bar{R}$ is defined by $\partial j = \beta$. We fix $f \in L^2(Q)$ and $y_0 \in H^1(\Omega)$ such that $\varphi(y_0) < +\infty$; i.e., $j(y_0) \in L^1(\Omega)$. Define

$$F_\varepsilon y = A_0 y \quad \text{for } y \in D(F_\varepsilon), \tag{5.5}$$

where

$$D(F_\varepsilon) = \left\{ y \in H^2(\Omega); \frac{\partial y}{\partial \nu} + \beta^\varepsilon(y) = 0 \text{ a.e. on } \Gamma \right\},$$

and the β^ε are given by (4.10). Clearly $F_\varepsilon = \partial\varphi^\varepsilon$ where

$$\varphi^\varepsilon(y) = \tfrac{1}{2} a(y, y) + \int_\Gamma j^\varepsilon(y)\, d\sigma \quad \text{for } y \in H^1(\Omega), \tag{5.6}$$

and the j^ε are defined by (4.13).

Assumptions (i), (ii) and (iiib) are obviously satisfied. Observe also that condition (2.5) is implied by inequality (4.14). To verify the remaining part of condition (iiia) consider $\{y_\varepsilon\}$ strongly convergent to y in $L^2(Q)$ for $\varepsilon \to 0$. If $\lim_{\varepsilon \to 0} \inf \varphi^\varepsilon(y_\varepsilon) < +\infty$, then by (5.6) we see that the $\{y_\varepsilon\}$ remain in a bounded subset of $H^1(\Omega)$ and therefore the family of traces $\{y_0 y_\varepsilon\}$ is bounded in $H^{1/2}(\Gamma)$ (see, e.g., [11, p. 41]). Hence $\{\gamma_0 y_\varepsilon\}$ is a precompact subset of $L^2(\Gamma)$ and so by taking subsequences we may assume that $y_\varepsilon(\sigma) \to y(\sigma)$ a.e. $\sigma \in \Gamma$ (for simplicity we shall write $\gamma_0 y = y$). This implies by the same reasoning as in § 4 that

$$\liminf_{\varepsilon \to 0} \int_\Gamma j^\varepsilon(y_\varepsilon)\, d\sigma \geqq \int_\Gamma j(y)\, d\sigma,$$

and therefore $\lim_{\varepsilon \to 0} \inf \varphi^\varepsilon(y_\varepsilon) \geqq \varphi(y)$. The proof of condition (iiic) is entirely similar to that for its counterpart in problem (4.1). Let us now calculate the Gâteaux differential $\nabla\Theta_\varepsilon(w)$ of the corresponding operator $\Theta_\varepsilon: L^2(\Omega) \to L^2(\Omega)$. For each $w \in L^2(\Omega)$ and $v \in L^2(\Omega)$, $\nabla\Theta_\varepsilon(w)v = z_\varepsilon$, where $z_\varepsilon \in H^{2,1}(Q)$ is the solution to

$$\begin{aligned} &(z_\varepsilon)_t + A_0 z_\varepsilon = v && \text{on } Q, \\ &\frac{\partial z_\varepsilon}{\partial \nu} + \nabla\beta^\varepsilon(\Theta_\varepsilon w) z_\varepsilon = 0 && \text{on } \Sigma, \\ &z_\varepsilon(x, 0) = 0 && \text{on } \Omega. \end{aligned} \tag{5.7}$$

It is easily seen that the adjoint operator is given by $(\nabla\Theta_\varepsilon(w))^* q = -v_\varepsilon$ where $v_\varepsilon \in H^{2,1}(Q)$ is the solution to

$$
\begin{aligned}
&(v_\varepsilon)_t - A_0 v_\varepsilon = q && \text{on } Q,\\
&\frac{\partial v_\varepsilon}{\partial \nu} + \nabla\beta^\varepsilon(\Theta_\varepsilon w) v_\varepsilon = 0 && \text{on } \Sigma, \qquad (5.8)\\
&v_\varepsilon(x, T) = 0 && \text{on } \Omega.
\end{aligned}
$$

We are therefore in the situation described in § 2. Thus if $(y^*, u^*) \in H^{2,1}(Q) \times L^2(0, T; U)$ is an optimal pair for problem (5.1) then there exist $\{y_\varepsilon\} \subset H^{2,1}(Q)$, $\{u_\varepsilon\} \subset L^2(0, T; U)$, $\{p_\varepsilon\} \subset H^{2,1}(Q)$, and $\{q_\varepsilon\} \subset L^2(Q)$ satisfying

$$
\begin{aligned}
&(y_\varepsilon)_t + A_0 y_\varepsilon = B u_\varepsilon + f && \text{on } Q,\\
&\frac{\partial y_\varepsilon}{\partial \nu} + \beta^\varepsilon(y_\varepsilon) = 0 && \text{on } \Sigma, \qquad (5.9)\\
&y_\varepsilon(x, 0) = y_0(x) && x \in \Omega;
\end{aligned}
$$

$$
\begin{aligned}
&(p_\varepsilon)_t - A_0 p_\varepsilon = q_\varepsilon && \text{on } Q,\\
&\frac{\partial p_\varepsilon}{\partial \nu} + \nabla\beta^\varepsilon(y_\varepsilon) p_\varepsilon = 0 && \text{on } \Sigma, \qquad (5.10)\\
&p_\varepsilon(x, T) = 0;
\end{aligned}
$$

and

$$
(q_\varepsilon(t), B^* p_\varepsilon(t) + u^*(t) - u_\varepsilon(t)) \in \partial L^\varepsilon(y_\varepsilon(t), u_\varepsilon(t)) \quad \text{a.e. } t \in]0, T[. \qquad (5.11)
$$

Furthermore, according to Lemmas 5 and 6 we have

$$
\begin{aligned}
y_\varepsilon &\to y^* && \text{strongly in } C(0, T; L^2(\Omega)),\\
(y_\varepsilon)_t &\to y_t^* && \text{weakly in } L^2(Q),\\
A y_\varepsilon &\to A y^* && \text{weakly in } L^2(Q),\\
u_\varepsilon &\to u^* && \text{strongly in } L^2(0, T; U);
\end{aligned} \qquad (5.12)
$$

and

$$
\begin{aligned}
q_\varepsilon &\to q && \text{weakly in } L^1(0, T; L^2(\Omega)),\\
p_\varepsilon &\to p && \text{weak star in } L^\infty(0, T; L^2(\Omega)),
\end{aligned} \qquad (5.13)
$$

where $q \in L^2(Q)$ satisfies the equation

$$
(q(t), B^* p(t)) \in \partial L(y^*(t), u^*(t)) \quad \text{a.e. } t \in]0, T[. \qquad (5.14)
$$

On the other hand, since $\{A y_\varepsilon\}$ is bounded in $L^2(Q)$ we may infer by [3, Lemma I.10] that $\{y_\varepsilon\}$ is bounded in $L^2(0, T; H^2(\Omega))$. Since $\{(y_\varepsilon)_t\}$ is bounded in $L^2(Q)$, we conclude that $\{y_\varepsilon\}$ is compact in $L^2(0, T; H^1(\Omega))$, and therefore

$$
y_\varepsilon \to y^* \quad \text{strongly in } L^2(0, T; H^1(\Omega)). \qquad (5.15)
$$

Next, multiplying both sides of (5.10) by p_ε and $\operatorname{sgn} p_\varepsilon$ (more precisely by a C^∞-monotone approximation of sgn) and using Green's formula, one obtains

$$
\int_\Sigma |\nabla\beta^\varepsilon(y_\varepsilon) p_\varepsilon| d\sigma\, dt + \int_Q |\operatorname{grad} p_\varepsilon|^2\, dx\, dt \leqq C. \qquad (5.16)
$$

Thus by taking further subsequences, we may also assume that

$$p_\varepsilon \to p \quad \text{weakly in } L^2(0, T; H^1(\Omega)), \tag{5.17}$$

and

$$\nabla\beta^\varepsilon(y_\varepsilon)p_\varepsilon \to \gamma_p \quad \text{weak star in } \mathcal{M}(\Sigma), \tag{5.18}$$

where $\mathcal{M}(\Sigma)$ denotes the space of all bounded measures on Σ.

Letting ε tend to zero in (5.10) and using (5.13), (5.17) and (5.18), we find that $p \in L^2(0, T; H^1(\Omega)) \cap L^\infty(0, T; L^2(\Omega))$ is a weak solution to the boundary-value problem

$$\begin{aligned} &p_t - A_0 p = q && \text{on } Q, \\ &\frac{\partial p}{\partial \nu} + \gamma_p = 0 && \text{on } \Sigma, \\ &p(\cdot, T) = 0 && \text{on } \Omega; \end{aligned} \tag{5.19}$$

i.e.,

$$\int_Q p\varkappa_t \, dx \, dt + \int_0^T a(p, \varkappa) \, dt + \gamma_p(\varkappa) + \int_Q q\varkappa \, dx \, dt = 0, \tag{5.20}$$

for all $\varkappa \in D_{\Omega,0}$.

We have therefore proved the following theorem.

THEOREM 4. *Let $(y^*, u^*) \in H^{2,1}(Q) \times L^2(0, T; U)$ be an optimal pair for problem* (5.1). *Then there exist functions $p \in L^\infty(0, T; L^2(\Omega)) \cap L^2(0, T; H^1(\Omega))$, $q \in L^2(Q)$ and a measure $\gamma_p \in \mathcal{M}(\Sigma)$ satisfying along with (y^*, u^*)* (5.14) *and* (5.19). *Moreover, y^*, u^*, p, q, and γ_p are limits in the sense of* (5.12), (5.13), (5.17) *and* (5.18) *of some sequences $\{y_\varepsilon\}$, $\{u_\varepsilon\}$, $\{p_\varepsilon\}$, and $\{q_\varepsilon\}$ satisfying* (5.9), (5.10) *and* (5.11).

We have also the following analogue of Lemma 7.

LEMMA 8. *$\gamma_0 p_\varepsilon \to \gamma_0 p$ strongly in $L^2(\Sigma)$.*

Proof. Since $\{p_\varepsilon\}$ is bounded in $L^2(0, T; H^1(\Omega))$ and $\{q_\varepsilon\}$ in $L^1(0, T; L^2(\Omega))$, we see by (5.10) that $\{(p_\varepsilon)_t\}$ is bounded in $L^1(0, T; H^{-1}(\Omega))$. Then arguing as in the proof of Lemma 7 we deduce that $p_\varepsilon \to p$ in $C(0, T; H^{-1}(\Omega))$, and for each $\eta > 0$ there is a $C(\eta)$ such that

$$\|p_\varepsilon - p\|_{L^2(0,T;H^{1-\delta}(\Omega))} \leqq C\eta + C(\eta)\|p_\varepsilon - p\|_{L^2(0,T;H^{-1}(\Omega))}, \tag{5.21}$$

where $0 < \delta \leqq \frac{1}{2}$. Hence $p_\varepsilon \to p$ in $L^2(0, T; H^{1-\delta}(\Omega))$, and by the "trace" theorem ([11, p. 41]) the set $\{\gamma_0 p_\varepsilon\}$ is precompact in $L^2(0, T; H^{1/2-\delta}(\Gamma) \subset L^2(\Sigma))$ as claimed.

Theorem 5 below follows by way of nearly the same proof as for Theorem 2.

THEOREM 5. *In Theorem* 4 *assume that β is locally Lipschitzian. Then the absolutely continuous part $(\gamma_p)_a$ of the measure $\gamma_p \in \mathcal{M}(\Sigma)$ satisfies the equation*

$$(\gamma_p)_a(x, t) \in \partial\beta(y^*(x, t))p(x, t), \quad \text{a.e. } (x, t) \in \Sigma, \tag{5.22}$$

where $\partial\beta$ denotes the generalized gradient of β. If in addition, β satisfies condition (4.43), *then $\gamma_p \in L^1(\Sigma)$, and $p \in C(0, T; L^1(\Omega))$ satisfies the equation*

$$\begin{aligned} &p_t - A_0 p = q && \text{on } Q, \\ &\frac{\partial p}{\partial \nu} + \partial\beta(y^*)p \ni 0 && \text{on } \Sigma, \\ &p(x, T) = 0, && x \in \Omega, \end{aligned} \tag{5.23}$$

Finally, we shall consider the case in which the control problem (5.1) is governed by the unilateral problem

$$
\begin{aligned}
& y_t + A_0 y = f + Bu && \text{on } Q, \\
& y\frac{\partial y}{\partial \nu} = 0, \quad \frac{\partial y}{\partial \nu} \geqq 0, \quad y \geqq 0 && \text{on } \Sigma, \\
& y(x, 0) = y_0(x), && x \in \Omega,
\end{aligned} \tag{5.24}
$$

which corresponds to a maximal monotone graph β defined by (4.45). According to general theory the initial data y_0 must satisfy

$$
y_0 \in H^1(\Omega), \qquad y_0(x) \geqq 0 \quad \text{a.e. } x \in \Gamma.
$$

Variational problems of this type arise in the theory of temperature control through the boundary and theory of semipermeable walls (see [7, p. 23]).

The same reasoning which led to Theorem 3 now gives us

THEOREM 6. *Let* $(y^*, u^*) \in H^{2,1}(Q) \times L^2(0, T; U)$ *be an optimal pair for problem* (5.1) *with state equation* (5.24). *Then there exist functions* $p \in L^\infty(0, T; L^2(\Omega)) \cap L^2(0, T; H^1(\Omega))$ *and* $q \in L^2(Q)$ *which satisfy along with* (y^*, u^*) *the quasi-variational inequalities*

$$
\begin{aligned}
& y_t^* + A_0 y^* = f + Bu && \text{a.e. on } Q, \\
& y^*\frac{\partial y^*}{\partial \nu} = 0, \quad \frac{\partial y^*}{\partial \nu} \geqq 0, \quad y^* \geqq 0 && \text{a.e. on } \Sigma, \\
& y^*(x, 0) = y_0(x) && \text{a.e. } x \in \Omega;
\end{aligned} \tag{5.25}
$$

$$
\begin{aligned}
& p_t - A_0 p = q && \text{on } Q, \\
& \left(\frac{\partial p}{\partial \nu}\right)_a = 0 && \text{a.e. on } \{(x, t) \in \Sigma;\ y^*(x, t) > 0\};
\end{aligned} \tag{5.26}
$$

$$
p\frac{\partial y^*}{\partial \nu} = 0 \qquad \text{a.e. on } \Sigma, \tag{5.27}
$$

$$
\begin{aligned}
& p(x, T) = 0 && \text{a.e. } x \in \Omega; \\
& (q(t), B^* p(t)) \in \partial L(y^*(t), u^*(t)) && \text{a.e. } t \in]0, T[.
\end{aligned} \tag{5.28}
$$

Of course (5.26) is meant in the weak sense (see (4.62)).

REFERENCES

[1] V. BARBU AND TH. PRECUPANU, *Convexity and Optimization in Banach Spaces*, Sijthoff & Noordhoff-Publishing House of Romanian Academy, 1978.

[2] V. BARBU, *Necessary conditions for nonconvex distributed control problems governed by elliptic variational inequalities*, J. Math. Anal. Appl., to appear.

[3] H. BREZIS, *Problèmes unilateraux*, J. Math. Pures Appl., 51 (1972), pp. 1–164.

[4] ———, *Propriétés régularisantes de certain semi-groupes nonlinéaires*, Israel J. Math., 9 (1971), 513–534.

[5] F. H. CLARKE, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247–262.

[6] I. EKELAND AND R. TEMAM, *Analyse convexe et problèmes variationnels*, Dunod, Gauthier-Villars, Paris, 1974.

[7] G. DUVAUT AND J. L. LIONS, *Inequalities in Mechanics and Physics*, Springer-Verlag, Berlin, Heidelberg, New York, 1976.

[8] J. L. LIONS, *Partial differential inequalities*, Uspehi Mat. Nauk., 26 (1971), pp. 202–263.
[9] ———, *Various topics in the theory of optimal control of distributed systems*, in Optimal Control Theory and its Applications, Lecture Notes in Economics and Mathematical Systems 105, Springer-Verlag, New York, 1974.
[10] ———, *Quelques méthodes de resolution des problèmes aux limites nonlinéaires*, Dunod, Gauthier-Villars, Paris, 1969.
[11] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications*, Springer-Verlag, Berlin, Heidelberg, New York, 1972.
[12] F. MIGNOT, *Contrôle dans les inéquations variationnelles elliptiques*, J. Funct. Anal., 22 (1976), pp. 130–185.
[13] J. P. PUEL, *Contrôle optimal sur les solutions minimum ou maximum de certain problèmes semilineaires*, (to appear).
[14] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton NJ, 1970.
[15] ———, *Integral functional, normal integrands and measurable selections*, in Nonlinear Operators and the Calculus of Variations, J. P. Gossez, ed., Lecture Notes in Mathematics 543, Springer-Verlag, New York, 1976.
[16] ———, *La théorie des sousgradients et ses applications a l'optimisation*, Les Presses de l'Université de Montréal, Montréal, Quebec, Canada, 1978.
[17] CH. SAGUEZ, *Conditions necessaire d'optimalité pour des problèmes de contrôle optimal associes a des inéquations variationnelles* (to appear).
[18] ———, *Contrôle optimal d'inéquations variationnelles avec observation de domaines* Rapport Laboria, March, 1978.
[19] J. P. YVON, *Contrôle optimal de systemès gouvernés par des inéquations variationnelles*, Rapport Laboria, February, 1974.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 1, January 1981

0363-0129/81/1901-0007 $01.00/0

ASYMPTOTIC PROPERTIES OF STOCHASTIC APPROXIMATIONS WITH CONSTANT COEFFICIENTS*

HAROLD J. KUSHNER† AND HAI HUANG‡

Abstract. Asymptotic properties (as $n \to \infty$, and then $a \to 0$) of the Stochastic Approximation (SA) algorithm

$$(*) \qquad X_{n+1}^a = X_n^a + a h_a(X_n^a, \xi_n^a),$$

are obtained, where h_a is not necessarily additive in ξ_n^a. If $Eh_a(x, \xi_n^a) = g(x) + O(a)$ and $\dot{x} = g(x)$ is globally asymptotically stable about a solution $x_t \equiv \theta$, then the asymptotic properties of $\{(X_n^a - \theta)/\sqrt{a}\} \equiv \{U_n^a\}$ are developed. In particular, it is shown that (as $a \to 0$) a natural continuous parameter interpolation of the tail part of $\{U_n^a\}$ converges weakly to a stationary Gauss-Markov process, from which the asymptotic properties of $\{U_n^a\}$ and $\{X_n^a\}$ can be obtained for small a. The conditions on $\{\xi_n^a\}$ are reasonable from the point of view of the usual applications to adaptive systems and identification. These results seem to be the first of their type for SA's with constant coefficients. Some rate of convergence results for classical SA's are improved. Also, an application of (*) to a problem of tracking the time varying parameters of a linear system is discussed, and a limit theorem obtained. Because in the usual practical implementations of SA to problems in systems theory, the gain sequence $\{a_n\}$ does not normally go to zero (due to considerations of robustness and nonstationarities), these results are of particular importance.

1. Introduction. In [1] asymptotic properties and rates of convergence for stochastic approximations (SA) of the type

$$(1.1) \qquad X_{n+1} = X_n + a_n h(X_n, \xi_n)$$

were studied, where $\{a_n\}$ is a sequence of positive numbers tending to zero and is such that $\sum a_n = \infty$, and $\{\xi_n\}$ is a sequence of random variables. In this paper, we obtain analogous results concerning asymptotic behavior of (1.2) where $a_n = a$, a small constant. For each a, $\{\xi_n^a\}$ is a stationary sequence and f_a, g, and k_a are measurable functions, further properties of which will be given below.

$$(1.2) \qquad \begin{aligned} &X_{n+1}^a = X_n^a + a h_a(X_n^a, \xi_n^a) \equiv X_n^a + a g(X_n^a) + a f_a(X_n^a, \xi_n^a) + o(a) k_a(X_n^a, \xi_n^a), \\ &X_0^a = X_0, \quad \text{independent of } a, \qquad X_n^a \in R^r, \quad \text{Euclidean } r \text{ space.} \end{aligned}$$

Algorithms of the type (1.2) are particularly important in applications to both identification theory and adaptive systems theory; for a typical example of this problem, the results are both specialized and extended in §§ 6 and 7; in § 7 $\{\xi_n^a\}$ is allowed to be nonstationary in a way that allows us to treat the "time-varying parameter" identification problem. In engineering practice with algorithm (1.1) there is usually a constant $a > 0$ such that either $\{a_n\}$ tends to a or else that $a_n = a$, although almost all the existing analyses of (1.1), (see, e.g., [2], [3], [4]) assume $a_n \to 0$. The case (1.2) is more robust than (1.1) in the sense that it can better accommodate nonstationarities and modeling errors.

In general, little is known about the sequence $\{X_n^a\}$. Normally, $\{X_n^a\}$ does not converge w.p.1, and if $\{\xi_n^a\}$ is nonstationary $\{X_n^a\}$ may not even converge in distribution.

* Received by the editors June 11, 1979, and in revised form January 14, 1980.

† Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. This author's research was supported in part by the Air Force Office of Scientific Research under AF-AFOSR 76-3063, in part by the National Science Foundation under NSF-Eng 77-12946, and in part by the Office of Naval Research under N0014-76-C-0279-P0002.

‡ Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. This author's research was supported in part by the Air Force Office of Scientific Research under AF-AFOSR 76-3063.

Under various assumptions, (1.3) (a specialization of (1.2)) has been treated in the adaptive process literature and results such as $\overline{\lim}_n E|X_n^a| = 0$ obtained. In (1.3), B is a vector valued bilinear form and A, C are matrices (Widrow, et al. [5], Senne [6], Davisson [7]).

$$X_{n+1}^a = X_n^a + aB(X_n^a, \xi_n^a) + aC\xi_n^a + aAX_n^a. \tag{1.3}$$

Our method works under broader conditions and yields a much more complete picture of the process behavior. As in [1], [3], weak convergence methods (see note on weak convergence below) are used.

The problem of interest is roughly the following. Let θ be a globally asymptotically stable solution to $\dot{x} = g(x)$, and suppose that $Ef_a(\theta, \xi_n^a) \equiv 0$ and define $U_n^a = (X_n^a - \theta)/\sqrt{a}$. We are interested in information on the "error" $(X_n^a - \theta)$ for large n, and small a. Such information is not easy to get, but it is of importance in the design and analysis of algorithms. We proceed in a natural way by studying the distributions of U_n^a for large n and small a. Suppose that $t_n \equiv an$ and $\{n_a\}$ is any sequence which goes to ∞ fast enough as $a \to 0$ (see Theorem 1 for the "rate"), and define the piecewise constant continuous parameter process $U^a(\cdot)$ by $U_{n_a}^a(0) = U_{n_a}^a$ and $U^a(t) = U_{n_a+n}^a$ in $[na, (n+1)a)$. In the proofs, the sequence $\{n_a\}$ will always be clear from the context. The behavior of the initial part of the sequence $\{U_n^a\}$ is strongly affected by the value of X_0^a, and does not pertain to the long run behavior of the algorithm. So we study $U^a(\cdot)$ and show, in particular that this sequence converges weakly (see note below) to the stationary Gauss-Markov diffusion defined by (5.1) as $a \to 0$, where $\bar{H} = g_x(\theta)$ and R is defined below (5.1). See also the remark (i) in § 4.

The results yield stability of the process (1.2) for small a, together with the asymptotic (as $a \to 0$) error variances and correlation functions of the normalized error process $U^a(\cdot)$. It seems to us that the general approach is quite natural and straightforward and relatively easy to use. The weak convergence and stability ideas yield a lot of intuitive insight into the relations between the structure of an algorithm and its asymptotic properties. For the special adpative process case when (1.2) reduces to (1.3), the situation is simpler since we can center U_n^a about its mean value and use $n_a \equiv 0$, and we can obtain better results. See § 7.

Note on weak convergence. A sequence of random variables $\{Y_\alpha\}$ is said to be tight (equivalently, bounded in probability) if $\lim_N \sup P\{|Y_\alpha| \geqq N\} = 0$. By tightness of $\{U_n^a$, small a, $n \geqq N_a\}$ we mean that there is some $a_0 > 0$ (whose value will be unimportant) such that doubly indexed sequence is tight. We need some type of boundedness property on the tails of $\{U_n^a, n \geqq 0\}$ uniformly in a, and the above tightness property serves this need well. $D^r[0, \infty)$ denotes the space of R^r valued functions on $[0, \infty)$ which are right continuous and have left hand limits, and which is endowed with the Skorokhod topology [10]. By weak convergence of $U^a(\cdot)$ to $U(\cdot)$ in $D^r[0, \infty)$ we mean that for any real valued continuous function $F(\cdot)$ on $D^r[0, \infty)$, $EF(U^a(\cdot)) \to EF(U(\cdot))$ as $a \to 0$. This is a substantial generalization of the notion of convergence in distribution of Euclidean space random variables, and is well suited to our purposes. Billingsley [10] is an excellent reference on weak convergence theory.

Outline of the paper. In § 2, assumptions for the general problem are stated. Tightness of $\{U_n^a, n \geqq N_a$, small $a\}$ for a sequence $N_a \to \infty$ as $a \to 0$ is obtained in § 3. This tightness is crucial for it implies that the normalization of the error process by $1/\sqrt{a}$ is the appropriate one, and that the error process does not "blow up", as $n \to \infty$, and $a \to 0$. The argument is of the "Lyapunov function" type. § 4 contains some remarks concerning special cases, and on the use of the methods of this paper to extend known convergence and rate of convergence results for SA's of the type (1.1) when the SA

sequence converges in *probability* rather than (the usual assumption) almost surely. The main limit theorem, the weak convergence of $\{U^a(\cdot)\}$ to $U(\cdot)$ if $n_a \to \infty$ fast enough, is given in § 4. The proof is essentially a brute force direct construction of the limit by use of weak convergence arguments. Some of the arguments required for our proofs are similar to those used in [1], and we formulate the problem here so as to use the earlier results whenever possible. For results and techniques that might be more useful when the $\{\xi_n^a\}$ are state-dependent or the dynamics discontinuous, see [13]–[15].

2. Assumptions for § 3. Throughout the paper, K denotes an arbitrary real number (independent of x, ξ, n, a) and its value may change from usage to usage. $G_{xx}(x)$ denotes the Hessian matrix of a function G and E_n^a denotes conditioning on ξ_i^a, $i < n$.

Remarks on the assumptions. In order to study the asymptotic properties of $U^a(\cdot)$ for small a we obviously must be able to show that the "tails" of the sequences $\{X_n^a, n \geqq 0\}$ converge in some sense as $a \to 0$. This requires some stability assumptions on the "deterministic" part of (1.2), in particular that a solution $x_t = \text{constant} = \theta$ of the ODE $\dot{x} = g(x)$ is globally asymptotically stable. It seems best to deal with the stability problem by introducing a Lyapunov function $V(\cdot)$ for $\dot{x} = g(x)$. Conditions (A6)–(A7) below are often guaranteed by various forms of strong mixing conditions on $\{\xi_n^a\}$. In the usual applications to identification and adaptive systems theory [5], [8], there is an asymptotically stable A such that $g(x) = Ax$ and an affine function $f(\cdot)$ such that $f_a(x, \xi) = f(x)\xi$. Then $V(\cdot)$ is chosen to be a quadratic form and (A4)–(A5) hold, and so do (A6)–(A7) under simple conditions on $\{\xi_n^a\}$. See § 7 for more detail.

A1. *ξ_n^a is bounded by some constant K, uniformly in a, n and $Ef_a(x, \xi_n^a) = 0$, all x, a, n.*

A2. *$g(\theta) = 0$, $g(\cdot)$, $k_a(\cdot, \cdot)$ and $f_a(\cdot, \cdot)$ are measurable. The first and second partial x-derivatives of $f_a(\cdot, \xi)$ and $g(\cdot)$ are continuous for each ξ.*

A3. *There is a nonnegative three times continuous differentiable Lyapunov function $V(\cdot)$ for $\dot{x} = g(x)$ such that $V(x) \geqq 0$, $V(x) \to \infty$ as $|x| \to \infty$, $V(x) = \delta x' Q \delta x + o(|\delta x|^2)$ for some positive definite matrix Q, where $\delta x \equiv (x - \theta)$.*

A4. *For some real $\gamma > 0$, $V_x'(x)g(x) \leqq -\gamma V(x)$.*

A5. *$V_{xx}(\cdot)$ is uniformly bounded and $|f_a(x, \xi)|^2 + |k_a(x, \xi)|^2 + |g(x)|^2 \leqq K(V(x)+1)$ and $|V_x'(x)k_a(x, \xi)| \leqq K(V(x)+1)$.*

A6. $\sum_{i=n}^{\infty} a|E_n^a V_x'(x) f_a(x, \xi_i^a)| \leqq aK(V(x)+1)$.

A7. $\sum_{i=n}^{\infty} a|E_n^a(V_x'(x)f_a(x, \xi_i^a))_{xx}| \leqq aK$,

$$\sum_{i=n}^{\infty} a|E_n^a(V_x'(x)f_a(x, \xi_i^a))_x| \leqq aK(V^{1/2}(x)+1).$$

(A5) implies that f_a and g grow at most linearly in x.

3. Tightness of $\{U_n^a$, small a, $n \geqq N_a\}$. Fix $K_0 > 0$. Let N_a denote any integer such that $\exp(-(a\gamma/2)N_a) \leqq K_0 a$.

THEOREM 1. *Under* (A1)–(A7), $\{U_n^a$, *small* a, $n \geqq N_a\}$ *is tight.*

Remark. We have the $n \geqq N_a$ requirement because of the effect of the initial condition. In general, $\{U_n^a, n \geqq 0$, small $a\}$ will not be tight unless $X_0 = \theta$. So we wait (N_a steps) until (as implied by the proof of Theorem 1) the effects of the initial condition are small. For the special case (1.3), it is possible to center the sequence $\{U_n^a, n \geqq 0\}$ in such a way that $N_a \equiv 0$ can be used. See § 7.

Proof. Recall that $t_n = an$. By (A6), (3.1) is well defined.

$$V_1^a(x, t_n) = a \sum_{i=n}^{\infty} E_n^a V_x'(x) f_a(x, \xi_i^a). \tag{3.1}$$

Define V^a by

$$V^a(x, t_n) = V(x) + V_1^a(x, t_n). \tag{3.2}$$

The proof uses a "Lyapunov function approach", with Lyapunov function[1] V^a. The bounds which imply the tightness will follow from the basic inequality (3.3b). Write

$$E_n^a V^a(X_{n+1}^a, t_{n+1}) - V^a(X_n^a, t_n) = T_1 + T_2 + T_3,$$

where

$$T_1 = E_n^a V(X_{n+1}^a) - V(X_n^a),$$
$$T_2 = E_n^a V_1^a(X_n^a, t_{n+1}) - V_1^a(X_n^a, t_n),$$
$$T_3 = E_n^a V_1^a(X_{n+1}^a, t_{n+1}) - E_n^a V_1^a(X_n^a, t_{n+1}).$$

Then, truncated Taylor series expansions yield the following, where X_n^+ and X_n^{++} are random variables in the range $[X_n^a, X_{n+1}^a]$.

$$T_1 = aV_x'(X_n^a)g(X_n^a) + aV_x'(X_n^a)f_a(X_n^a, \xi_n^a)$$
$$+\frac{a^2}{2}(f(X_n^a, \xi_n^a) + g(X_n^a))' V_{xx}(X_n^+)(f_a(X_n^a, \xi_n^a) + g(X_n^a)),$$

$$T_2 = -aV_x'(X_n^a) f_a(X_n^a, \xi_n^a),$$

$$T_3 = a\sum_{i=n+1}^{\infty} E_n^a V_x'(X_{n+1}^a) f_a(X_{n+1}^a, \xi_i^a) - a\sum_{i=n+1}^{\infty} E_n^a V_x'(X_n^a) f_a(X_n^a, \xi_i^a)$$
$$= a^2 \sum_{i=n+1}^{\infty} E_n^a (V_x(X_n^a) f_a(X_n^a, \xi_i^a))_x' (f_a(X_n^a, \xi_n^a) + g(X_n^a))$$
$$+\frac{a^3}{2}\sum_{i=n+1}^{\infty} E_n^a (f_a(X_n^a, \xi_n^a)$$
$$+ g(X_n^a))'[V_x'(X_n^{++}) f_a(X_n^{++}, \xi_i^a)]_{xx}(f_a(X_n^a, \xi_n) + g(X_n^a)).$$

These expansions together with (A4)–(A7) yield (note that T_2 cancels out the second term of T_1; this is the reason for the introduction of V_1^a).

$$E_n^a V^a(X_{n+1}^a, t_{n+1}) - V^a(X_n^a, t_n) \leqq -a\gamma V(X_n^a) + a^2 K[V(X_n^a) + 1]. \tag{3.3a}$$

By (A6), $|V_1^a(x, t_n)| \leqq aK(V(x)+1)$, and by (3.3a),

$$E_n^a V^a(X_{n+1}^a, t_{n+1}) - V^a(X_n^a, t_n) \leqq -a\gamma V^a(X_n^a, t_n) + a^2 K[V^a(X_n^a, t_n) + 1]. \tag{3.3b}$$

Let a be small enough so that $a^2 K \leqq a\gamma/2$ (or, equivalently $a \leqq a_0 \equiv \gamma/2K$). Then (3.3b) yields

$$E_0^a V^a(X_n^a, t_n) \leqq \exp\left(\frac{-a\gamma n}{2}\right) V^a(X_0^a, 0) + Ka. \tag{3.4}$$

Equation (3.4) also holds for V replacing V^a, since $|V_1^a(x, n)| \leqq Ka(V(x)+1)$. Thus, by (3.4) and (A3), for any constant k_1 and $n \geqq N_a$, $a \leqq a_0$, we have

$$P\{\delta X_n^{a'} Q \delta X_n^a + o(|\delta X_n^a|^2) \geqq k_1 a\} \leqq \frac{K[\exp(-(a\gamma N_a/2))][V(X_0^a)+1] + Ka}{k_1 a} \leqq \frac{K}{k_1}. \tag{3.5}$$

[1] As will be seen, the V_1^a is used as an "averaging" device, to average out the effects of the $\{\xi_i^a\}$. If the $\{\xi_i^a\}$ are mutually independent for some a, then $V_1^a \equiv 0$.

Tightness of $\{U_n^a$, small a, $n \geqq N_a\}$ follows from (3.5) in the following way. Fix $\delta > 0$. To get the tightness it is enough to find a $k_\delta < \infty$ such that

$$P\left\{\frac{\delta X_n^{a'} Q \delta X_n^a}{a} \geqq k_\delta\right\} \leqq \delta, \quad \text{all } a \leqq a_0, \quad n \geqq N_a. \tag{3.6}$$

There is an $\varepsilon_0 > 0$ such that for $\delta x' Q \delta x \leqq \varepsilon_0$ and the $o(\cdot)$ of (A3), $|o(|\delta x|^2)| \leqq \delta x' Q \delta x/2$. For each real $k_3 > 0$, there is a $k_4(k_3) > 0$ such that $\delta x' Q \delta x \geqq k_3$ implies $V(x) \geqq k_4(k_3)$ and we can choose $k_4(\cdot)$ to be a monotonic function.

Let $n \geqq N_a$. By (3.5) (recall that K might have a different value in each usage),

$$P\left\{\frac{\delta X_n^{a'} Q \delta X_n^a}{2a} \geqq k_1\right\} \leqq \frac{K}{k_1} + P\{\delta X_n^{a'} Q \delta X_n^a \geqq \varepsilon_0\}$$

$$\leqq \frac{K}{k_1} + P\{V(X_n^a) \geqq k_4(\varepsilon_0)\} \leqq \frac{K}{k_1} + \frac{Ka}{k_4(\varepsilon_0)}.$$

Choose k_1 such that $K/k_1 = \delta/2$. If $a \leqq \bar{a} \equiv \delta k_4(\varepsilon_0)/2K$, then the right-hand side is $\leqq \delta$. If $a_0 \geqq a > \bar{a}$, note that for any $k > 0$

$$P\left\{\frac{\delta X_n^{a'} Q \delta X_n^a}{a} \geqq k\right\} \leqq P\left\{\frac{\delta X_n^{a'} Q \delta X_n^a}{\bar{a}} \geqq k\right\} \leqq P\{V(X_n^a) \geqq k_4(\bar{a}k)\}$$

$$\leqq \frac{Ka}{k_4(\bar{a}k)} \leqq \frac{Ka_0}{k_4(\bar{a}k)}.$$

Now choose k_2 such that $Ka_0/k_4(\bar{a}k_2) \leqq \delta$. Finally, define the k_δ to be used in (3.6) by $k_\delta = \max(k_1, k_2)$. Q.E.D.

4. Remarks. (i) In a practical implementation of the algorithm (1.2) a_n might not be chosen to be constant, but might be allowed to decrease to some value $a > 0$ by iteration number n_a, where n_a might be chosen such that $E|\delta X_n^a|^2 \approx Ka$, and a_n will remain at value a thereafter. But if we are interested only in the "tail" of $\{X_n^a\}$, we can often assume that the initial condition error is commensurate with the value of a (e.g., $E|\delta X_0^a|^2 \leqq Ka$). We might also be more concerned with the ability of the algorithm to track *changes* (e.g., the *changing* system parameters in the identification example (§ 7)), than with the transient errors. Then we need only look at the "errors" U_n^a for large n (say, $n \geqq N_a$) at which time transient errors due to the initial condition have been "dissipated".

(ii) *Stochastic approximation* (1.2) *with* $a_n \to 0$. Again, suppose without loss of generality that the origin is the unique asymptotically stable point of $\dot{x} = g(x)$. Let $a_n = A/(n+1)$. Then the method of Theorem 1 can be used to show tightness of $\{\delta X_n/\sqrt{a_n}, n \geqq 0\}$, without the (usually required) assumption that $\delta X_n \to 0$ w.p.1. To do this we first define $t_n = \sum_{i=0}^{n-1} a_i$ and $V_1(x, t_n) = \sum_{i=n}^{\infty} a_i E_n V_x'(x) f_a(x, \xi_i)$, where E_n denotes the expectation conditioned on ξ_i, $i < n$. Assume (A1)–(A5) and the natural analogs of (A6)–(A7), where the a under the summation is replaced by a_i and that on the right hand side is replaced by a_n. Assume $A\gamma/2 > 1$. Then $\{\delta X_n/\sqrt{a_n}, n \geqq 0\}$ can be shown to be tight. The proof closely follows the lines of the proof of Theorem 1 and we only make a few remarks concerning it. Define $V^0(x, t_n) = V(x) + V_1(x, t_n)$. Then using the method of Theorem 1, derive the inequality

$$E_n V^0(X_{n+1}, t_{n+1}) - V^0(X_n, t_n) \leqq -\gamma a_n V(X_n) + a_n^2 K[V(X_n) + 1]. \tag{4.1}$$

(4.1) holds with V^0 replacing V on the right-hand side. Effecting this replacement and

iterating (4.1) yields

$$EV^0(X_n, t_n) \leqq \left[\exp -\frac{\gamma t_n}{2}\right] V^0(X_0, 0) + K \sum_{i=0}^{n-1} \prod_{j=i+1}^{n-1} \left(1 - \frac{\gamma}{2} a_j + K a_j^2\right) a_i^2.$$

Finally, show that the above right side is bounded above by Ka_n.

This result is important because the proof of tightness of $\{\delta X_n/\sqrt{a_n}\}$ is the basic difficulty in obtaining rate of convergence results for classical stochastic approximations. If tightness of $\{\delta X_n/\sqrt{a_n}\}$ is known, then the rate of convergence proofs in [1], [3], [9] all go through with virtually no changes without using the assumption that $X_n \to \theta \equiv 0$ w.p.1, which was required in those references.

(iii) *Stochastic approximation, additive noise.* Continue with the situation in the last paragraph, but let $f(x, \xi) = \xi$, the classical Robbins-Monro case. Then (A6)–(A7) are particularly simple. There are adaptations to the Kiefer-Wolfowitz case, where $c_i = C/(i+1)^\gamma$, $a_i = A/(i+1)^\alpha$, $2\gamma < \alpha$, $\gamma > 0$, and $\{c_i\}$ is the finite difference coefficient sequence. Then the normalizing sequence is $\{\sqrt{a_n/c_n}\}$ rather than $\{\sqrt{a_n}\}$.

5. The limit theorem. Let $\{n_a\}$ denote a sequence of integers such that $n_a \geqq N_a$, where N_a is defined in § 3. Define $Q_a = n_a - N_a$. If $\{n_a\}$ is not specified further, it is an arbitrary sequence satisfying the definition. For each $a > 0$, define $U^0(\cdot)$ by $U^0(0) = U^a_{n_a}$ and for each integer i, $U^a(t) = U^a_{i+n_a}$ in $[ia, ia+a)$. We will show that $U^a(\cdot)$ converges weakly in $D^r[0, \infty)$ to the solution to the Gauss-Markov process $U(\cdot)$ defined by

$$dU = \bar{H}U dt + R^{1/2} dB, \quad U(0) = \text{weak limit of } \{U^a(0)\}, \tag{5.1}$$

where $\bar{H} = g_x(\theta)$, $B(\theta)$ is a standard Wiener process and R is defined by (see (A9) below)

$$R = \lim_{a \to 0} \sum_{-\infty}^{\infty} R^a(i),$$

where

$$R^a(i) = E f_a(\theta, \xi_j^a) f_a'(\theta, \xi_{j+i}^a).$$

Also, as asserted by the theorem, if $aQ_a \to \infty$ as $a \to 0$ then the weak limit of $\{U^a(\cdot)\}$ is the *stationary* solution to (5.1).

The proof is simplified by the following consideration. Suppose that $U^a(\cdot)$ does *not* converge weakly to $U(\cdot)$ in $D^r[0, \infty)$. Then there is a sequence $\{a_k\}$ of positive numbers which goes to zero *as fast as we wish* and a $T < \infty$ such that $U^{a_k}(\cdot)$ does *not* converge weakly to $U(\cdot)$ in $D^r[0, T]$. Thus, it suffices to show convergence of $U^{a_k}(\cdot)$ to $U(\cdot)$ in $D^r[0, T]$ for an arbitrary T, and for a sequence $\{a_k\}$ which goes to zero fast enough but is otherwise arbitrary. We will set the problem up similarly to the way it was set up in [1], so that the results of that reference can be used whenever possible.

The following assumptions are required (analogous to (A3a) of [1]). After stating the conditions, we comment on their reasonableness. Define $m_a(t) = \max\{i; ai \leqq t\}$.

A8. *There is a* $T_1 > 0$ *such that* ($f_{a,x}$ *is the gradient of* f_a)

$$P\left\{\max_{0 \leqq t \leqq T_1} a \left| \sum_{i=m_a(t_N)}^{m_a(t_N+t)-1} f_{a,x}(\theta, \xi_i^a) \right| \geqq \varepsilon \right\} \equiv \bar{k}_a(\varepsilon) \to 0$$

as $a \to 0$, *for each* $\varepsilon > 0$, *uniformly in* N.

A9. *Define* $h_j^a = f_a(\theta, \xi_j^a)$. *Then* $\{h_j^a\}$ *is stationary for each* a. *Define* $R^a(i) =$

$Eh_j^a(h_{j+i}^a)'$. Then $R^a = \sum_{-\infty}^{\infty} R^a(i)$ is absolutely summable and the sum converges[2] *uniformly in a. There is a matrix R such that $R^a \to R$ as $a \to 0$.*

A10. *Define $\rho_1^a(i) = \sup_{j,l \geqq 0} E^{1/2}|E_j^a h_{j+i}^a h_{j+i+l}^{a'} - R^a(l)|^2$. Then $\sum_{i=0}^{\infty} (\rho_1^a(i))^{1/2} < \infty$, where the sum converges uniformly in a.*

A11. *Define $\rho_2^a(i) = \sup_{k \geqq 0} E^{1/2}|E_k^a h_{k+i}^a|^2$, $i \geqq 0$. Then $\sum_{i=0}^{\infty} (\rho_2^a(i))^{1/2} < \infty$, where the sum converges uniformly in a.*

A12. $|f_{a,xx}(x, \xi)| + |g_{xx}(x)| \leqq K$.

Remarks on (A8)–(A12). The conditions do not seem to be particularly strong. Except for the boundedness of $\{\xi_n^a\}$, they are basically the conditions used in [1], adapted to the present case.

Let $\{\xi_n^a\}$ be a ϕ-mixing process in the sense of [10] with $\sum \phi_i^{1/4} < \infty$, where ϕ_i does not depend on a. Then (A9)–(A11) hold. Since use will be made of results from [1] we note that condition (A3b) of [1] always holds if the noise $\{\xi_n\}$ there is bounded (set $\tau = 0$ there). The main use of (A9)–(A11) is in showing that a certain sequence of interpolated sums converges weakly to a Wiener process. Since they were used in [1] for a closely related problem, we use them here in order to simplify the proof.

Define $k_j^a = f_x(\theta, \xi_j^a)$, let there be an $\bar{R}_i$ such that $|Ek_j^a k_{j+i}^{a'}| \leqq \bar{R}_i$, for all j and (small) $a > 0$, and define $\bar{R} = \sum_i \bar{R}_i$. Then by a Mensov-Rademacher type estimate ([3, p. 98]), there is a K (depending on $\bar{R}$) such that for each $T_1 > 0$

$$a^2 E \max_{t \leqq T_1} \left| \sum_{i=m_a(t_N)}^{m_a(t_N+t)-1} k_i^a \right|^2 \leqq Ka^2(m_a(t_N + T_1) - m_a(t_N)) \log_2^2 4[m_a(t_N + T_1) - m_a(t_N)]$$

$$\leqq T_1 Ka \log_2^2 \frac{4T_1}{a} \leqq K_1 a \log_2^2 a,$$

which implies (A8). Other examples satisfying (A8) appear in [3].

THEOREM 2. *Assume* (A1)–(A12). *Then $\{U^a(\cdot)\}$ converges weakly in $D'[0, \infty)$ to the $U(\cdot)$ of* (5.1). *If $aQ_a \to \infty$ as $a \to 0$, then $U(0)$ has the stationary distribution of $U(t)$.*

Proof. Fix $T > 0$. Let $\{\varepsilon_i\}$ and a_k denote sequences of positive numbers such that $\sum_i \varepsilon_i < \infty$, $a_k \to 0$, and (see (A8))

$$\sum_k \bar{k}_{a_k}(\varepsilon_k) < \infty. \tag{5.2}$$

If (A8) holds for some T_1 then it holds for all T_1, so we can suppose that $T_1 = T$. By the discussion at the beginning of the section it is enough to prove the theorem for $\{U^{a_k}(\cdot)\}$. We suppose w.1.o.g. that $U_{n_k}^{a_k}$ converges weakly to a random variable $U(0)$.

Part 1. Define $\sqrt{a} f_a(\theta, \xi_j^a) \equiv \delta W_j^a$, and $W_{N,n}^a \equiv \sum_{j=N}^{N+n-1} \delta W_j^a$ and let $W_N^a(\cdot)$ denote the function on $[0, T]$ which equals $W_{N,n}^a$ on $[an, an + a)$. By a truncated Taylor series expansion

$$\begin{aligned}\delta X_{n+1}^a &= [I + ag_x(\theta) + af_{ax}(\theta, \xi_n^a)]\delta X_n^a \\ &\quad + af_a(\theta, \xi_n^a) + aB_1(G(X_n^+), \delta X_n^a)\delta X_n^a + o(a)k_a(X_n^a, \xi_n^a) \\ &\equiv [I + aH_n^a]\delta X_n^a + af_a(\theta, \xi_n^a) + a\gamma_n^a,\end{aligned}$$

where $B_1(G(X_n^+), \delta X_n^a)$ is a matrix valued bilinear form in $G(X_n^+)$ and δX_n^a, and the elements of $G(X_n^+)$ are components of the second derivatives of $f_a(x, \xi_n^a) + g(x)$ evaluated at some point in the interval $[\theta, X_n^a]$. H_n^a is defined in the obvious manner.

[2] By $\sum_1^{\infty} q_i^a$ converging uniformly in a, we mean $\sum_n^{\infty} |q_i^a| \to 0$ uniformly in a as $n \to \infty$.

Thus

$$U^a_{n+1} = [I + aH^a_n]U^a_n + \delta W^a_n + \sqrt{a}\gamma^a_n. \tag{5.3}$$

Define $\Gamma^a_{N,n} \equiv \sum_{j=N}^{N+n-1} \sqrt{a}\gamma^a_n$ and let $\Gamma^a_N(\cdot)$ denote the function on $[0, T]$ which equals $\Gamma^a_{N,n}$ on $[an, an + a)$.

Define the function $C^\alpha_\beta(a)$ by $C^N_{N+1}(a) = I$ and for $n \geqq l+1$,

$$C^{N+n}_{N+l+1}(a) = \sum_{j=N+l+1}^{N+n} [I + aH^a_j] = (I + aH^a_{N+n}) \cdots (I + aH^a_{N+l+1}).$$

By iterating (5.3) and doing a summation by parts, we get (5.4), just as (3.6) of [1] was obtained.

$$\begin{aligned} U^a_{N+n+1} &= C^{N+n}_N(a)U^a_N + C^{N+n}_{N+1}(W^a_{N,n+1} + \Gamma^a_{N,n+1}) \\ &\quad - \sum_{l=1}^{n} aC^{N+n}_{N+l+1}(a)H^a_{N+l}[(W^a_{N,n+1} - W^a_{N,l}) + (\Gamma^a_{N,n+1} - \Gamma^a_{N,l})]. \end{aligned} \tag{5.4}$$

From this point on, the proof is basically a brute force construction of the limit process, by showing that each term of (5.4) converges weakly to a limit such that the resulting sum is a representation of $U(\cdot)$.

Part 2. We now claim that

$$C^{m_a(t_N+t+s)}_{m_a(t_N+s)}(a_k) \to \exp \bar{H}t \quad \text{on } [0, T] \tag{5.5}$$

uniformly w.p.1, as $k \to \infty$, for *any fixed N or sequence* $N \to \infty$ as $k \to \infty$. The limit result (5.5) follows from [1, Lemma 2], when we make the following identification of our $\{a_k\}$ with the $\{a_k\}$ in [1, Lemma 2]. To avoid confusion write the $\{a_k\}$ of [1] as $\{\bar{a}_k\}$. Then set the first $[T/a_1]$ of the $\{\bar{a}_n\}$ equal to our a_1, the next $[T/a_2]$ of the $\{\bar{a}_n\}$ equal to our a_2, etc. Then (A8), (5.2) and the Borel-Cantelli Lemma imply [1, (A2)], hence also [1, Lemma 2] and (5.5).

Part 3. Write $n_{a_k} \equiv n_k$. As in [1, Theorem 2], (A9)–(A11) imply that $W_{n_k}(\cdot)$ converges weakly to a Wiener process $W(\cdot)$ with infinitesimal covariance R; i.e., $W(t) = R^{1/2}B(t)$, where $B(\cdot)$ is a standard Wiener process. In Part 4 below it is shown that

$$\{\Gamma^{a_k}_{n_k}(\cdot)\} \text{ converges weakly to the zero process as } k \to \infty. \tag{5.6}$$

Assuming (5.6), the proof is completed via the arguments of [1, Theorem 2, Part 3], as follows. The argument in [1, Theorem 2, Part 3], can be used to show that when a and N in (5.4) are replaced by a_k and n_k resp., the resulting $H^a_{n_k+l}$ in (5.4) can be replaced by $g_x(\theta)$ without affecting the limit. With this replacement and the convergence of $\{U^a_{n_k}\}$ and $\{W^a_{n_k}(\cdot), \Gamma^a_{n_k}(\cdot)\}$, (5.1) follows from (5.4) and (5.5), since the limit of (5.4) is simply a particular integral representation of $U(\cdot)$. The stationarity argument is in Part 5.

Part 4. *Proof of* (5.6). Let M_k denote $[T/a_k]$ and $n_k = n_{a_k}$. In view of the properties of $\{\gamma^a_i\}$, (5.6) holds if (5.7) does:

$$P\left\{\sqrt{a_k} \sum_{i=n_k}^{n_k+M_k-1} |X^{a_k}_i|^2 \geqq \varepsilon\right\} \to 0 \quad \text{as } k \to \infty, \quad \text{each } \varepsilon > 0. \tag{5.7}$$

If $V(\cdot)$, the Lyapunov function of Theorem 1, is quadratic in δx, then $E|\delta X^{a_k}_{a_k+i}|^2 \leqq Ka_k$ by Theorem 1, and (5.7) holds by an application of Chebyshev's inequality. We now prove it in the general case. The subscript k on a_k will be dropped.

Recall from Theorem 1 that there is a K such that for $n \geqq N_k$ and the V^a of Theorem 1,

$$
\begin{aligned}
&EV(X_n^a) \leqq Ka, \qquad |EV^a(X_n^a, t_n)| \leqq Ka,\\
&V^a(x, t_n) \geqq -Ka,\\
&E_n^a V(X_{n+1}^a, t_{n+1}) \leqq (1-\gamma a + Ka^2)V^a(X_n^a, t_n) + a^2 K.
\end{aligned} \tag{5.8}
$$

Let K be fixed at its above value henceforth in this proof, and let $n \geqq n_k$ and $n - n_k \leqq T/a$ and let a be small enough such that

$$
-\gamma a + Ka^2 < 0, \qquad Ka < 1, \quad a < 1.
$$

Define the random variables L^a by

$$
L^a(X_n^a, n) = V^a(X_n^a, t_n) + aK + (T - a(n - N_a))a^{3/4}K.
$$

Then $L^a(X_n^a, n) \geqq 0$. By (5.8), we have

$$
\begin{aligned}
E_n^a L^a(X_{n+1}^a, n+1) \leqq{}& (1-\gamma a + Ka^2)V^a(X_n^a, t_n)\\
&+ aK + (T + aN_a - (n+1)a)Ka^{3/4} + Ka^2,
\end{aligned} \tag{5.9a}
$$

and, consequently,

$$
\begin{aligned}
E_n^a L^a(X_{n+1}^a, n+1) - L^a(X_n^a, n)&\\
\leqq (-\gamma a + Ka^2)V^a(X_n^a, t_n) + Ka^2 - Ka^{7/4} &\leqq 0.
\end{aligned} \tag{5.9b}
$$

Thus $\{L^a(X_n^a, n)\}$ is a nonnegative supermartingale for each small a. Thus, there is a real K_1 such that

$$
P\{\sup_{n_k \leqq i \leqq n_k + M_k - 1} L^a(X_i^a, i) \geqq a^{5/8}\} \leqq \frac{EL^a(X_{n_k}^a, n_k)}{a^{5/8}} \leqq \frac{K(a + a^{3/4})}{a^{5/8}} = O(a^{1/8}). \tag{5.10}
$$

There is a $K_2 < \infty$ such that if $L^a(X_n^a, n) \leqq a^{5/8}$, then $V(X_n^a) \leqq K_a a^{5/8}$. We can suppose that a is small enough so that $V(x) \leqq K_2 a^{5/8}$ implies that $V(x) \geqq \delta x' Q \delta x/2$, and $V^a(x, n) \geqq x'Qx/2 - aK$. Then, for small a and $L^a(X_n^a, n) \leqq a^{5/8}$,

$$
0 \leqq L^a(X_n^a, n) = O(a^{3/4}) + \delta_n, \tag{5.11}
$$

where $\delta_n \leqq (\delta X_n^a)' Q \delta X_n^a/2$. Equation (5.7) follows from (5.10) and (5.11), since there is a real K_0 such that with probability $1 - O(a^{1/8})$,

$$
\sqrt{a} \sum_{i=n_k}^{n_k + M_k - 1} \frac{(\delta X_i^a)' Q \delta X_i^a}{2} \leqq \sqrt{a}\left(\frac{T}{a}\right)(K_0 a^{5/8}) = O(a^{1/8}).
$$

Part 5. *Stationarity.* Let $\mathcal{U}$ denote all possible weak limits of the set $\{U_n^a, a \text{ small}, n \geqq N_a\}$ introduced in § 3. Then $\mathcal{U}$ is tight. We have proved that as $a \to 0$, $U^a(\cdot)$ converges to a solution of (5.1) for each sequence $\{n_a\}$ if $n_a \geqq N_a$. It can be shown (although we have not done so here) that $U(0)$ is independent of $B(\cdot)$ in (5.1). The measures of the limits of different subsequences of $\{U^a(\cdot)\}$ might be different only because the measures of the limiting initial condition might be different. Fix a convergent subsequence for initial times $n_a \equiv N_a$ and denote the limit by $U_0(\cdot)$. Then there is a $B(\cdot)$ such that for any $t_0 > 0$

$$
U_0(t_0) = e^{\bar{H}t_0} U_0(0) + \int_0^{t_0} e^{\bar{H}(t_0 - s)} R^{1/2}\, dB_s. \tag{5.12}
$$

Now, let $n_a = N_a + Q_a$, where $Q_a a \to t_0$ and let $U^a[t_0, \cdot)$ denote the part of the above defined $U^a(\cdot)$ on $[t_0, \infty)$. Then $U^a[t_0, \cdot)$ converges weakly to a limit $U_{t_0}(\cdot)$ whose initial condition is given by (5.12). Due to the tightness of $\mathcal{U}$, $U_{t_0}(0) = U_0(t_0)$ converges in distribution to the stationary initial condition of (5.1) as $t_0 \to \infty$, uniformly in the initial condition $U_0(0)$. This implies the stationarity of $U(0)$ if $aQ_a \to \infty$. A similar argument shows that this stationary $U(0)$ is independent of the Brownian motion used to represent $U(\cdot)$. Q.E.D.

6. Adaptive systems–examples. We will describe very briefly two of the more important systems which fall into our framework. Let $\{u_n, \mu_n\}$ and $\{y_n\}$ denote the input and output sequences, resp., of the linear system

$$y_n = -[c_1 y_{n-1} + \cdots + c_k y_{n-k}] + [b_0 u_n + \cdots + b_l u_{n-l}] + \mu_n. \tag{6.1}$$

Suppose that the system is asymptotically stable when $u_n \equiv 0$, $\mu_n \equiv 0$. Define

$$\psi_n = (-y_{n-1}, \cdots, -y_{n-k}, u_n, \cdots, u_{n-l})', \qquad \theta = (c_1, \cdots, c_k, b_0, \cdots, b_l)',$$

and let $\{\mu_n\}$ be a *zero mean* random sequence which is *independent* of the *zero mean* sequence $\{\mu_n\}$. A common algorithm for estimating θ is

$$X^a_{n+1} = X^a_n + a_n[y_n - (X^a_n)'\psi_n]\psi_n, \tag{6.2}$$

where X^a_n is the nth estimate of θ. Under various conditions (including $a_n \to 0$) $X^a_n \to \theta$ w.p.1 [2], [3]. In practice, due to extraneous noise, robustness considerations or model uncertainties, it is common for either $a_n \downarrow a > 0$ or $a_n \equiv a$, a constant, perhaps a matrix. The case where θ varies with time and $a_n \equiv a$ is dealt with in detail in the next section.

Next, consider a similar algorithm which is very useful in adaptive communications theory. Let $\{S_{ni}\}$, $i = 1, 2$ and $\{N_{ni}\}$, $i = 1, 2$, represent stationary signal and noise sequences, resp. $\{S_{n1}\}$ and $\{S_{n2}\}$ ($\{N_{n1}\}$ and $\{N_{n2}\}$, resp.) are related in the sense that they are signal (noise, resp.) processes appearing at the inputs to different antennas, but are from the same transmitting source. Let $y_n = S_{n1} + N_{n1}$ and $u_n = S_{n2} + N_{n2}$ denote the actual inputs to the two antennas. Let k be a fixed integer and set $\psi_n = (u_n, \cdots, u_{n-k})'$. In practice it is desired to find the weight vector $\bar{X}$ which is the minimizing X in the expression $E[y_n - X'\psi_n]^2$. The reasons for this, together with some interesting examples, appear in [5]. Essentially, under frequently occurring conditions the signal to noise power ratio of the sequence $\{y_n - \bar{X}'\psi_n\}$ is much greater than that of the sequence $\{y_n\}$.

The algorithm (6.2) is often used to calculate the optimum X recursively, when $a_n \equiv a$. But, in this context, (6.2) is not well understood. Usually, it is proved only that EX^a_n converges. Exceptions to this are the work of Davisson [7] (with m-dependent stationary Gaussian sequences as inputs) and Senne [6] (where the stationary inputs satisfy a type of mixing condition), where it is proved that $\overline{\lim}_n E|X^a_n - \bar{X}|^2 \to 0$ as $a \to 0$. The method of § 7 exploits the technique of the last section in order to get a more complete picture in the general case where a is small and the processes are nonstationary, an important case which actually justifies the use of the adaptive algorithm, but which has not yet been dealt with in the literature.

7. The nonstationary identification problem (6.1). In this section, the parameter θ in (6.1) is allowed to vary with time, and we let θ^a_n denote its value at time n. Since the variations in $\{\theta^a_n\}$ affect the statistics of $\{\psi^a_n\}$, the identification problem is more complicated than the adaptive communications problem, and we consider only the former case. Assume that $\{u_n, \mu_n\}$ are bounded. Nonstationarities due to the θ^a_n variations are more difficult to treat than the effects of nonstationary $\{u_n, \mu_n\}$. In order to concentrate on the more important effects and minimize the notation, we assume that

$\{u_n, \mu_n\}$ is stationary. Also $\{\mu_n\}$ is assumed to be zero mean, and independent of $\{u_n\}$ and we let $Eu_n \equiv 0$.

We now model the time variations. Let $\theta(\cdot)$ denote a uniformly continuous R^{l+k+1} valued function on $[0, \infty)$, with values in a *bounded set S*. Suppose that the parameter [3] θ_n^a takes the value $\theta(an)$. To see the reasonableness of the model note that the rate of change of the θ_n^a *must go to zero* in some sense as $a \to 0$, for otherwise tracking would not be possible. $\theta(\cdot)$ could be a random process, but no generality is gained by that, since we treat one sample function at a time anyway. The uniform continuity condition is used to assure that the $\{y_n\}$ sequence has a certain stability property on $[0, \infty)$. We want to avoid $\theta(\cdot)$ getting "wilder and wilder" as $t \to \infty$.

We could allow $\{\theta_n^a\}$ to be a random sequence for each a. Even then, its rate of change must still be proportional to a in some sense (or to a fractional power of a; but then the u_n, μ_n terms play no role in the limit as $a \to 0$). In any case, we want an (limit) equation which yields the limit of the behavior of the normalized interpolation of the error $(X_n^a - \theta_n^a)$ process in *terms of the limit of the parameter process*, so that the precise relationship can be seen. Our scheme is a natural way to get this.

The main object is to get some information on the properties of $\{U_n^a\}$ when a is small. We might be interested, for example, in an approximation to the distribution of some continuous function of $\{U_n^a, na \leqq T\}$. To get this, it makes sense to parametrize the problem so that we can get a limit result (as $a \to 0$) which will serve as the approximation to the $\{U_n^a\}$, and from which the approximation to the distributions of functions can be obtained (particularly if the convergence is in the sense of weak convergence). If we allow $a \to 0$ without simultaneously slowing down the rate of variation of θ_n^a, then obviously no limit result is possible, in general. Thus, to even discuss the behavior for small a, we must allow the θ_n^a to depend on a. As mentioned above, there are several ways in which this can be done. Our choice allows a relatively simple exhibition of the structure that the limit would have in a wide variety of cases (where, perhaps, $\theta(\cdot)$ might be a limit in some sense of the sequence of parameter variation functions $\theta^a(\cdot)$, where $\theta^a(t) = \theta_n^a$ on $[an, an + a)$). The problem is formulated and some terms are defined in Subsection 7.1. Subsection 7.2 obtains estimates concerning the dependence of the output sequence y_n on θ_n, where $\theta_n \equiv \theta$, a constant. Subsections 7.3 and 7.4 obtain a limit theorem for the interpolation of a deterministic centering sequence $\{\bar{Y}_n\}$, and tightness of $\{U_n^a\} = \{(X_n^a - \theta_n^a - \bar{Y}_n)/\sqrt{a}\}$, resp. In Subsection 7.5, the $C_m^n(a)$ (defined above (5.4)) are approximated by an exponential function and in Subsection 7.6, Theorem 5 gives the appropriate Wiener process limits and the convergence theorem for $\{U^a(\cdot)\}$. The large amount of detail is due to the awkward way that the time variations in $\{\theta_n^a\}$ affect the statistics of the y_n process.

7.1. Formulation of the problem. Let $\{y_n(\theta), \psi_n(\theta)\}$ denote the output and output-input sequence when $\theta_n^a \equiv \theta$, for all n; then $\psi_n(\theta) = \{-y_{n+1}(\theta), \cdots, -y_{n-k}(\theta), u_n, \cdots, u_{n-l}\}$. By (B1) below, these sequences are second order stationary for each $\theta \in S$. Define $R(\theta) = E\psi_i(\theta)\psi_i'(\theta)$, $R_t = R(\theta(t))$ and $\tilde{R}_n^a = E\psi_n^a\psi_n^{a'}$, the true covariance. Set $Y_n^a = X_n^a - \theta_n^a$, $\delta\theta_n^a = \theta(an + a) - \theta(an)$, $\beta_n^a = [\tilde{R}_n^a - \psi_n^a\psi_n^{a'}]$, $\beta_n(\theta) = [R(\theta) - \psi_n(\theta)\psi_n'(\theta)]$, $\tilde{F}_i^a = E\mu_i\psi_i^a$, $\gamma_n^a = \mu_n\psi_n^a - \tilde{F}_i^a$, and $F(\theta) = E\mu_i\psi_i(\theta)$. The superscript a will normally be omitted on Y_n^a, X_n^a, ψ_n^a and $\tilde{Y}_n^a$, $\bar{Y}_n^a$ for notational convenience. We have

$$X_{n+1} = X_n + a[(\theta_n^a)'\psi_n + \mu_n - X_n'\psi_n]\psi_n,$$

$$Y_{n+1} = Y_n - \delta\theta_n^a - a\tilde{R}_n^a Y_n + a\beta_n^a Y_n + a\mu_n\psi_n.$$

[3] The parameter θ_n^a is the value of $(c_1, \cdots, c_k, b_0, \cdots, b_l)'$ at time n. Then the c_i, b_j are components of (hence functions of) θ_n^a at time n.

Define the sequence $\{\bar{Y}_n\}$ by

$$\bar{Y}_{n+1} = \bar{Y}_n - \delta\theta_n^a - a\tilde{R}_n^a \bar{Y}_n + a\tilde{F}_n^a, \qquad \bar{Y}_0 = Y_0, \tag{7.1a}$$

and define $\tilde{Y}_n = Y_n - \bar{Y}_n$. Then

$$\tilde{Y}_{n+1} = \tilde{Y}_n - a\tilde{R}_n^a \tilde{Y}_n + a\beta_n^a(\tilde{Y}_n + \bar{Y}_n) + a\gamma_n^a, \qquad \tilde{Y}_0 = 0. \tag{7.1b}$$

$\bar{Y}_n$ is the "noiseless" part of Y_n, and contains the effects of the initial conditions. It is most convenient to work with the form $Y_n = \tilde{Y}_n + \bar{Y}_n$. Finally, define $\{U_n^a\} = \{\tilde{Y}_n/\sqrt{a}\}$, the sequence with whose convergence we will ultimately deal. We will not require that $n \geqq N_a$.

In order to exploit the stability properties of (6.1), it is convenient to work with (6.1) in *state variable form*. To set this up, define $\bar{u}_n = (u_n, \cdots, u_{n-l})'$ and $Z_n = (y_{n-k}, \cdots, y_{n-1})'$. Recall that, by definition, $\theta_n^a = \theta(an) =$ value of $\{c_1, \cdots, c_k, b_0, \cdots, b_l\}'$ at time n, a $(k+l+1)$ vector. For any S valued parameter θ, we define

$$A(\theta) = \begin{bmatrix} 0 & 1 & & 0 \\ & & \ddots & \\ & & & 1 \\ -c_k(\theta), & \cdots, & & -c_1(\theta) \end{bmatrix}, \quad B(\theta) = \begin{bmatrix} & & 0 & \\ & & \vdots & \\ & & 0 & \\ b_0(\theta), & \cdots, & & b_l(\theta) \end{bmatrix}, \quad C = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix},$$

$$D = [0, \cdots, 0, 1].$$

We define $A_i = A(\theta_i^a)$, $B_i = B(\theta_i^a)$. Then $Z_{i+1} = A_i Z_i + B_i \bar{u}_i + C\mu_i$, $y_i = DZ_{i+1}$. Define $Z_i(\theta) = \{y_{n-k}(\theta), \cdots, y_{n-1}(\theta)\}$. Then

$$Z_{i+1}(\theta) = A(\theta)Z_i(\theta) + B(\theta)\bar{u}_i + C\mu_i, \qquad y_i(\theta) = DZ_{i+1}(\theta). \tag{7.2}$$

Write E_n for the expectation conditioned on μ_i, $\bar{u}_i$, $i < n$. The following additional assumptions are required.

(B1) $|A^n(\theta)| \to 0$ *as* $n \to \infty$, *uniformly for* $\theta \in S$.

(B2) $\sum_{i=n}^{\infty} |E_n(\psi_i\psi_i' - \tilde{R}_i^a)| = \sum_{i=n}^{\infty} |E_n\beta_n^a|$ *bounded uniformly in* n, ω.

(B3) *There is a* $q_1 > 0$ *such that* $R(\theta) - q_1 I$ *is nonnegative definite, for all* $\theta \in S$.

(B4) $\sum_{i=n}^{\infty} |E_n(\psi_i\mu_i - \tilde{F}_i^a)| = \sum_{i=n}^{\infty} |E_n\gamma_i^a|$ *bounded uniformly in* n, ω.

(B2) and (B4) are not restrictive. Under (B1), they hold under a ϕ-mixing condition (see [10] for the definition) on $\{u_n, \mu_n\}$ with $\sum \phi_i^{1/2} < \infty$. (B4) holds if the $\{\mu_i\}$ are mutually independent, *and independent of the* $\{u_i\}$.

7.2. Some preparatory estimates. Since the statistics of the $\psi_i(\theta)$ are easier to get than those of the ψ_i, we show that ψ_i can be well approximated by $\psi_i(\theta_i^a)$, uniformly in i, for small a. Let $P(\theta)$ denote the unique symmetric positive definite Lyapunov matrix satisfying $A'(\theta)P(\theta)A(\theta) - P(\theta) = -I$. By (B1), there are $0 < \rho_1 < \rho_2 < \infty$ such that

$$\rho_1 I \leqq P(\theta) \leqq \rho_2 I \quad \text{for all } \theta \in S. \tag{7.3}$$

We now obtain a series of results concerning the closeness[4] of $R(\theta_n^a)$ to $\tilde{R}_n^a$ and $Z_{n+1}(\theta_n^a)$ to Z_{n+1}. Recall that $Z_i(\theta_n^a)$ is the value obtained from (7.2), when the θ in (7.2) is held fixed at θ_n^a for all i (i.e., $A(\theta) = A_n$, $B(\theta) = B_n$). We can write

$$Z_{n+1} = \sum_{j=-\infty}^{n} [A_n \cdots A_{j+1}](B_j\bar{u}_j + C\mu_j). \tag{7.4}$$

[4] Recall that $R(\theta_n^a)$ is the covariance $E\psi_i(\theta_n^a)\psi_i'(\theta_n^a)$; i.e., the parameter θ is held fixed at $\theta \equiv \theta_n^a$.

For small a, the sum in (7.4) converges uniformly by virtue of the stability assumption and its convergence (7.3). Indeed, by (7.3) and the fact that $|A(\theta_{n+1}^a)-A(\theta_n^a)|\to 0$ uniformly in n as $a\to 0$, there are $a_0>0$, $\varepsilon>0$, $K_1<\infty$, such that (see also [11] for a similar estimate)

$$|A_n\cdots A_{j+1}|\leqq K_1(1-\varepsilon)^{n-j},\quad \text{all } n, j,\quad \text{for } a\leqq a_0. \tag{7.5a}$$

By (B1), we can suppose that K_1, ε, are chosen such that (7.5b) also holds.

$$|A_n^j|\leqq K_1(1-\varepsilon)^j. \tag{7.5b}$$

The following approximation result is the basis of much of the rest of the development.

LEMMA 1. *Under the stability assumption* (B1),

$$\sup_n |Z_{n+1}-Z_{n+1}(\theta_n^a)|\to 0 \quad \text{as } a\to 0.$$

Proof. Let M_a denote an integer whose specific value will be selected below. The variations in the B_j cause no problem in the proof and neither does $\{C\mu_j\}$. So, in order to simplify the proof set $B_j\equiv B$, a constant, and $C=0$; i.e., the b_i components of $\theta(\cdot)$ are constant. Then

$$\begin{aligned}|Z_{n+1}-Z_{n+1}(\theta_n^a)|\leqq &\left|\sum_{j=-\infty}^{n-M_a}(A_n\cdots A_{n-M_a+1})A_{n-M_a}\cdots A_{j+1}B\bar{\mu}_j\right|\\ &+\sum_{j=-\infty}^{n-M_a}|(A_n)^{M_a}(A_n)^{n-M_a-j}B\bar{u}_j|+\sum_{j=n-M_a+1}^{n}|A_n\cdots A_{j+1}-A_n^{n-j}||B\bar{u}_j|.\end{aligned} \tag{7.6}$$

By (7.5), there is a real K (not depending on a or M_a) such that the first two terms of (7.6) are each bounded in norm by $K(1-\varepsilon)^{M_a}/\varepsilon$. We will next get a bound on the third term. In (7.6), the value of the time parameter n plays no special role and it is enough for us to show that (7.7) tends to zero uniformly in $A_0=A(\theta(0))$ as $a\to 0$:

$$\sum_{j=0}^{M_a-1}|A_0\cdots A_j-A_0^{j+1}|. \tag{7.7}$$

In (7.7) A_i takes the form $A_i=A_0+\delta_i$, and all that we assume on δ_i is that there is a real K_0 such that $|\delta_i|\leqq K_0|\theta(ia)-\theta(0)|$. It is convenient to work in a matrix norm $|\cdot|_0$ which might depend on $\theta(0)$ but where there are real K_2, K_3, independent of $\theta(0)$, such that $|\cdot|\leqq K_3|\cdot|_0$, $|\cdot|_0\leqq K_2|\cdot|$. In particular define $|A|_0^2=\sup_{|x|=1}x'A'P(\theta(0))Ax$. Then $|A_0|_0<1$. By (7.3), K_2, K_3 exist and we can also suppose that $|\delta_i|_0\leqq K_2|\theta(ia)-\theta(0)|$. For the jth term of (7.7) in the $|\cdot|_0$ norm, we have

$$\begin{aligned}&|A_0(A_0+\delta_1)\cdots(A_0+\delta_j)-A_0^{j+1}|_0\\ &\qquad\leqq |A_0|_0^j\sum_{i=1}^{j}|\delta_i|_0+|A_0|_0^{j-1}\sum_{i_2>i_1}|\delta_{i_1}\delta_{i_2}|_0+\cdots+|A_0|_0|\delta_1\cdots\delta_j|_0.\end{aligned}$$

Let $\Delta=K_2\sup_{s\leqq M_a a}|\theta(s)-\theta(0)|$. Then a crude upper bound on the above is

$$|A_0|_0^{j+1}\left\{\binom{j}{1}\left(\frac{\Delta}{|A_0|_0}\right)+\binom{j}{2}\left(\frac{\Delta}{|A_0|_0}\right)^2+\cdots+\binom{j}{j}\left(\frac{\Delta}{|A_0|_0}\right)^j\right\}\leqq |A_0|_0^{j+1}\left\{\left(1+\frac{\Delta}{|A_0|_0}\right)^j-1\right\},$$

and (7.7) satisfies

$$|(7.7)| \leqq K_3 \sum_{j=0}^{M_a-1} |A_0|_0^{j+1} \left\{ \left(1+\frac{\Delta}{|A_0|_0}\right)^j - 1 \right\}. \tag{7.8}$$

Now choose $M_a \to \infty$ as $a \to 0$ in such a way that aM_a (hence Δ) goes to zero. Then, since $\sup_{\theta(0)} |A(\theta(0))|_0 < 1$, the right side of (7.8) and $(1-\varepsilon)^{M_a}$ both tend to zero uniformly in $\theta(0) \in S$, as $a \to 0$. Q.E.D.

Similar proofs yield the following corollaries.

COROLLARY 1. *Assume* (B1). *Let $M_a a \to 0$ and $M_a \to \infty$ as $a \to 0$ and let Θ denote the set $\{\theta(u)\colon na - M_a a < u \leqq na + M_a a\}$. Then*

$$\sup_{\theta \in \Theta} |Z_{n+1}(\theta_n^a) - Z_{n+1}(\theta)| \to 0, \tag{7.9}$$

as $a \to 0$, uniformly in n.

COROLLARY 2. *Assume* (B1), (B3). *Then $|R(\theta_n^a) - \tilde{R}_n^a| \to 0$ uniformly in n as $a \to 0$. Also there are $K < \infty$, $\varepsilon_0 > 0$ such that for $n > j$*

$$|(I - aR(\theta_n^a))(I - aR(\theta_{n-1}^a)) \cdots (I - aR(\theta_{j+1}^a))| \leqq K(1 - a\varepsilon_0)^{n-j}, \tag{7.10}$$

for all n, j and small a. The function $R_t = R(\theta(t))$ is continuous. The function $F(\cdot)$ is continuous and $\tilde{F}_n^a \to F(\theta(t))$, uniformly in t, as $a \to 0$, $n \to \infty$ if an is held equal to t.

Proof. The second assertion is a consequence of the continuity of $\theta(\cdot)$, and (B3) and (7.9). The rest are consequences of Lemma 1, and Corollary 1 and the details are omitted.

7.3. A limit theorem for $\{\bar{Y}_i\}$. We next turn to the treatment of the deterministic sequence $\{\bar{Y}_i\}$. Let $\Phi(t, s)$, $t > s$, denote the fundamental solution of the linear equation $\dot{x} = -R_t x$, and let $\bar{Y}^a(\cdot)$ denote the piecewise constant function on $[0, \infty)$ with values $\bar{Y}^a(t) = \bar{Y}_n$ on $[an, an + a)$, $n \geqq 0$.

LEMMA 2. *Assume* (B1), (B3). *Then $\{\bar{Y}_n\}$ is uniformly bounded. If the $\tilde{R}_j^a$ in* (7.12) *are replaced by $R(\theta_j^a)$, then the difference between $\bar{Y}_{n+1}$ and the new right hand side converges to zero uniformly in n, as $a \to 0$. As $a \to 0$, $\bar{Y}^a(\cdot)$ converges uniformly on bounded intervals to the function $\bar{Y}(\cdot)$ defined by*

$$\bar{Y}(t) = \Phi(t, 0)\bar{Y}(0) - \int_0^t \Phi(t, s)\, d\theta_s + \int_0^t \Phi(t, s) F(\theta(s))\, ds$$

$$= \Phi(t, 0)\bar{Y}(0) - \Phi(t, 0)(\theta(t) - \theta(0)) \tag{7.11a}$$

$$- \int_0^t \Phi(t, s) R_s(\theta(t) - \theta(s))\, ds + \int_0^t \Phi(t, s) F(\theta(s))\, ds,$$

which is the unique solution to the equation

$$d\bar{Y}(t) = -R_t \bar{Y}(t)\, dt - d\theta(t) + F(\theta(t))\, dt. \tag{7.11b}$$

Proof. For the first assertion we write the solution to (7.1a) in the form (using a

summation by parts to get the second equation)

$$\bar{Y}_{n+1} = \prod_{i=0}^{n} (I - a\tilde{R}_i^a)\bar{Y}_0 - \sum_{i=0}^{n} \prod_{j=i+1}^{n} (I - a\tilde{R}_j^a)\delta\theta_i^a + \sum_{i=0}^{n} \prod_{j=i+1}^{n} (I - a\tilde{R}_j^a)a\tilde{F}_i^a$$

$$= \prod_{i=0}^{n} (I - a\tilde{R}_i^a)\bar{Y}_0 - \prod_{i=1}^{n} (I - a\tilde{R}_i^a)[\theta(an + a) - \theta(0)] \tag{7.12}$$

$$- \sum_{i=1}^{n} a \prod_{j=i+1}^{n} (I - a\tilde{R}_j^a)\tilde{R}_i^a[\theta(na + a) - \theta(ia)] + \sum_{i=0}^{n} \prod_{j=i+1}^{n} (I - a\tilde{R}_j^a)a\tilde{F}_i^a.$$

Now use Corollary 2 together with the boundedness of $\theta(\cdot)$.

The second assertion follows from Corollary 2. The last assertion then follows by letting $a \to 0$, $n \to \infty$, $an = t$ in (7.12) and noting that for $t > s$

$$\prod_{i=m_a(s)}^{m_a(t)} (I - aR(\theta_i^a)) \to \Phi(t, s), \qquad t \geqq s,$$

uniformly on bounded s, t intervals. Q.E.D.

7.4. Tightness of $\{U_n^a\}$. With the preparatory results available, we proceed to the main result, by following the pattern of development in Theorem 1.

THEOREM 3. *Under* (B1)–(B4), $\{U_n^a, n \geqq 0,$ *small* $a\}$ *is tight. In particular* (*since* $\tilde{Y}_0 = \tilde{U}_0^a = 0$), $E|\tilde{Y}_n|^2 \leqq Ka$.

Proof. The proof is quite similar to that of Theorem 1 and we only remark on the basic setup. The Lyapunov functions of Theorem 1 will be $V(\tilde{y}) = \tilde{y}'\tilde{y} = |\tilde{y}|^2$,

$$V_1^a(\tilde{y}, t_n) = 2\tilde{y}' \sum_{i=n}^{\infty} E_n\beta_i^a\tilde{y} + 2\tilde{y}' \sum_{i=n}^{\infty} E_n\beta_i^a\bar{Y}_i + 2\tilde{y}' \sum_{i=n}^{\infty} E_n\gamma_n^a,$$

$$V^a(\tilde{y}, t_n) = V(\tilde{y}) + aV_1^a(\tilde{y}, t_n).$$

By virtue of (B2) and (B4), the sums are uniformly bounded and, as required by Theorem 1,

$$|V_1^a(\tilde{y}, t_n)| \leqq K(V(\tilde{y}) + 1). \tag{7.13}$$

Now, applying the mechanisms of the proof of Theorem 1 and using the boundedness of $\{|\bar{Y}_i|\}$ yields

$$E_nV^a(\tilde{Y}_{n+1}, t_{n+1}) - V^a(\tilde{Y}_n, t_n) \leqq -a\tilde{Y}_n'\tilde{R}_n^a\tilde{Y}_n + Ka^2(1 + |\tilde{Y}_n|^2). \tag{7.14}$$

Since $\tilde{R}_n^a$ is positive definite, uniformly in small a (Corollary 2 and (B3)), there is a $\gamma > 0$ such that $\tilde{Y}_n'\tilde{R}_n^a\tilde{Y}_n \geqq \gamma V(\tilde{Y}_n)$ and the method of Theorem 1 (together with the uniform positive definiteness of $\tilde{R}_n^a$) yields the desired tightness. Q.E.D.

7.5. Approximating $C_l^n(a)$ by an exponential. Recall the function $C_l^n(a)$ introduced below (5.3). Here $-\psi_n\psi_n'$ is the H_n^a of (5.3). We can write

$$U_{n+1}^a = [I + aH_n^a]U_n^a + \sqrt{a}\beta_n^a\bar{Y}_n + \sqrt{a}\gamma_n^a. \tag{7.15}$$

The estimate (7.16) is needed in the proof of the next theorem. By (B2), (B4), (the limits

of the sums are $m_a(s)$, $m_a(t)-1$),

$$E\left|\sum_i \beta_i^a\right|^2 \leqq 2E\sum_i |\beta_i^a|\left|E_{i+1}\sum_{j\geqq i}\beta_j^a\right| \leqq \frac{K(t-s)}{a}.$$

By this estimate and Chebyshev's inequality there is a real K such that

$$P\left\{a\left|\sum_{i=m_a(s)}^{m_a(t)-1}\beta_i^a\right| \geqq \varepsilon\right\} \leqq \frac{Ka(t-s)}{\varepsilon^2}. \tag{7.16}$$

THEOREM 4. *Under* (B1)–(B4)

$$C_{m_a(s)}^{m_a(t)}(a) \to \Phi(t, s)$$

uniformly on bounded s, t intervals if $a \to 0$ *fast enough*; *in particular, through any sequence* $\{a_k\}$ *where* $\sum_k a_k < \infty$.

Proof. The proof is very similar to that of Theorem 2, Part 2. First, fix $t \leqq T$, let M denote an integer, and divide $[0, t]$ into M intervals, each of width δ. Suppose (without loss of generality) that $N = \delta/a$ is an integer and $\delta < 1$. The constants K below do not depend on a, δ or on $t \leqq T$, and their values may change from usage to usage. We have

$$\left|C_0^{N-1}(a) - \left(I + a\sum_{j=0}^{N-1} H_j^a\right)\right| \leqq a^2 \sum_{i_2 > i_1} |H_{i_2}^a H_{i_1}^a| + \cdots + a^N |H_{N-1}^a \cdots H_0^a| \leqq K\delta^2. \tag{7.17}$$

(7.17) holds (with the same K) when 0 and $N-1$ are replaced by $iN-N$ and iN, resp., for any $i > 0$. Hence,

$$\left|C_0^{m_a(t)}(a) - \left(I + a\sum_{j=NM-M}^{NM-1} H_j^a\right)\cdots\left(I + a\sum_{j=0}^{N-1} H_j^a\right)\right| \leqq K\delta. \tag{7.18}$$

Let $\{a_k\}$ satisfy $\sum a_k < \infty$. Next, we want to show that

$$\left|\left(I + a\sum_{j=NM-M}^{NM-1} H_j^a\right)\cdots\left(I + a\sum_{j=0}^{N-1} H_j^a\right) - \left(I - a\sum_{j=NM-M}^{NM-1} \tilde{R}_j^a\right)\cdots\left(I - a\sum_{j=0}^{N-1} \tilde{R}_j^a\right)\right| \to 0 \tag{7.19}$$

uniformly for $t \in \{i\delta : i \leqq T/\delta\}$, w.p.1, as $a \to 0$ through the sequence $\{a_k\}$, for each fixed $\delta > 0$. Owing to the fact that both products $(I + a\sum_{i=m_a(\tau)}^{m_a(\tau+u)-1} H_j^a)$ and $C_{m_a(\tau)}^{m_a(\tau+u)}(a)$ can be made arbitrarily close to the identity by letting u and a be small, (7.19) implies that

$$\left|C_0^{m_a(t)}(a) - \prod_{j=0}^{m_a(t)-1}(I - a\tilde{R}_j^a)\right| \to 0$$

uniformly in $t \leqq T$, w.p.1, as $a \to 0$ through $\{a_k\}$. To get (7.19), we use the estimate

$$|(7.19)| \leqq Ka\sum_{i=1}^{M}\left|\sum_{j=iN-N}^{iN-1}\beta_j^a\right| \equiv E_0(a), \tag{7.20}$$

and, by using (7.16) with $\delta = (t-s)$ and $M = T/\delta$,

$$P\{E_0(a) \geqq \varepsilon\} \leqq \sum_{i=1}^{M} P\left\{a\left|\sum_{j=iN-N}^{iN-1}\beta_j^a\right| \geqq \frac{\varepsilon\delta}{T}\right\}$$

$$\leqq \left(\frac{T}{\delta}\right)K\delta a\frac{T^2}{\varepsilon^2\delta^2} = K\left(\frac{T^3}{\delta^2}\right)a.$$

Thus $\sum_k P\{E_0(a_k) \geqq \varepsilon\} < \infty$ and the Borel-Cantelli Lemma and (7.20) imply (7.19).

Finally, use the fact that by Corollary 2, (7.19) remains true when $\tilde{R}_j^a$ is replaced by $R(\theta(aj)) \equiv R(\theta_j^a)$ and the fact that

$$\prod_{j=0}^{NM-1} (I - aR(\theta(aj))) \to \Phi(t, 0)$$

uniformly in $[0, T]$, as $a \to 0$, to complete the proof. Q.E.D.

7.6. The Wiener process and the limit theorem for $\{U^n(\cdot)\}$. Define $U^a(\cdot)$ as in § 4 (but with $n_a \equiv 0$) and define W_n^a and Γ_n^a by

$$W_n^a = \sqrt{a} \sum_{i=0}^{n-1} \beta_i^a \bar{Y}_i, \qquad \Gamma_n^a = \sqrt{a} \sum_{i=0}^{n-1} \gamma_i^a, \tag{7.21}$$

and let $W^a(\cdot)$ and $\Gamma^a(\cdot)$ be the continuous parameter processes with values W_n^a and Γ_n^a, resp., on $[an, an + a)$. By solving (7.15) and doing a partial summation, we get (5.4), but where U_N^a is replaced by 0 and all N's are deleted. The limit result is given in Theorem 5 under the additional assumptions:

(B5) $\{\mu_i\}$ *is a sequence of bounded independent and identically distributed random variables with* $E\mu_i^2 = \sigma_\mu^2$ *and* $E\mu_i = 0$. *Also* $\{\mu_i\}$ *is independent of* $\{u_i\}$.

(B6) $\{u_i\}$ *is a bounded* ϕ*-mixing process* [10] *with mixing rate* $\{\phi_i\}$ *satisfying* $\sum \phi_i^{1/2} < \infty$.

Remark on (B5)–(B6). They are stronger than necessary. (B5) is used because otherwise $F(\theta) \neq 0$ and it seems pointless to get a limit theorem for $U^a(\cdot)$, when the $\bar{Y}(\cdot)$ itself is biased by $F(\theta(\cdot))$. Also, (B5)–(B6) imply (B2) and also that (B4) is zero.

THEOREM 5. *Assume* (B1), (B3), (B5), (B6). *Then* $\{W^a(\cdot), \Gamma^a(\cdot)\}$ *converges in* $D^{2(k+l+1)}[0, \infty)$ *weakly to a Wiener process* $(W(\cdot)\Gamma(\cdot))$, *whose covariances are*

$$\text{Cov } \Gamma(t) = \sigma_\mu^2 \int_0^t R(\theta(v))\, dv, \tag{7.22a}$$

$$\text{Cov } W(t) = \sum_{l=-\infty}^{\infty} \int_0^t E\beta_0(\theta(v))\bar{Y}(v)\bar{Y}'(v)\beta_l'(\theta(v))\, dv, \tag{7.22b}$$

$$E\Gamma(t)W'(t) = \sum_{l=1}^{\infty} \int_0^t E\mu_0\psi_0(\theta(v))\bar{Y}'(v)\beta_l'(\theta(v))\, dv. \tag{7.22c}$$

$\{U^a(\cdot)\}$ *converges weakly in* $D^{k+l+1}[0, \infty)$ *to the diffusion* $U(\cdot)$ *given by*

$$dU = -R_t U dt + dW + d\Gamma. \tag{7.23}$$

Remarks. (7.22a–c) are well defined. The sequence $\{\mu_i\}$ and, for each θ, the sequence $\{\psi_n(\theta), \beta_n(\theta)\}$, is a stationary process so the subscripts 0 and l in (7.22b, c) could be i and $i + l$, resp., for any i. These expressions are calculated by first calculating the asymptotic moments of $\{\psi_n(\theta)\}$ needed in (7.22) for each θ. These are continuous functions of θ, so (7.22) makes sense. Note that the differential of the covariance at t depends only on the parameter $\theta(t)$, which is the hoped for form. Compare (7.22) to the R below (5.1). They are equivalent if we use $f_a(\theta, \xi_j^a) = \gamma_j^a + \beta_j^a \bar{Y}_j$ and neither the parameters nor $\bar{Y}_j$ vary with time.

The exact values of the covariances are complicated and one would not normally want to calculate them—even for some known "test" variation $\theta(\cdot)$. Theorem 5 gives the structure of the limit and indicates how the variances depend on the unknown function. This, in itself, is useful.

Proof. Once the assertions concerning convergence to the Wiener process are shown, the proof is completed as indicated below (5.6) for Theorem 2. Only the

assertions concerning the Wiener processes will be proved. The proof of those assertions is based on the proof of similar assertions in Theorem 2 and in [1, Theorem 2]. The main changes are due to the nonstationarity, which requires altering (A9)–(A11) (resp., (A6)–(A8) of [1]).

In our nonstationary and bounded $\{u_n, \mu_n\}$ case, (A10) and (A11) should be replaced by: *Let* $h_i^a = \gamma_i^a$ or β_i^a *and define*

$$\rho_1^a(i) = \sup_{j,l} |E_j h_{j+i}^a h_{j+i+l}^{a'} - E h_{j+i}^a h_{j+i+l}^{a'}|, \qquad l \geqq 0, i \geqq 0, \tag{7.24a}$$

$$\rho_2^a(i) = \sup_k |E_k h_{k+i}^a|, \qquad i > 0. \tag{7.24b}$$

Then

$$\sum (\rho_1^a(i))^{1/2} + \sum (\rho_2^a(i))^{1/2} < \infty, \tag{7.24c}$$

where the sums converge uniformly in a.

By the independence of $\{\mu_i\}$, (7.24) is obvious for $h_i^a = \gamma_i^a$. There are linear $F_n^a(\cdot)$ with uniformly (in n, q) bounded coefficients and $\{\varepsilon_n^q\}$ satisfying $|\varepsilon_n^q| \leqq K(1-\varepsilon)^q$ such that $y_n = F_n^q(u_n, \cdots, u_{n-q}, \mu_n, \cdots, \mu_{n-q}) + \varepsilon_n^q$. From this representation and (B5), (B6), we can readily show (7.24) for $h_i^a = \beta_i^a$. The property (7.24) was used in [1, Parts 1, 2 of proof of Theorem 2], to show that[5] $\sum_{m(t_N)}^{m(t_N+t)-1} \sqrt{a_i} h_i$ was tight and converged weakly to a continuous martingale, and that $|\sum_{m(t_n)}^{m(t_N+t)-1} \sqrt{a_i} h_i|^2$ is uniformly integrable in N. The same proof can be used when $a_i \equiv a$. Thus, $\{W^a(\cdot), \Gamma^a(\cdot)\}$ are tight in $D^{2(k+l+1)}[0, \infty)$ and all weak limits are continuous martingales and $\{|W^a(t)|^2, |\Gamma^a(t)|^2,$ small $a\}$ is uniformly integrable for each t.

Choose and fix a convergent subsequence and index it by n, and let $W(\cdot), \Gamma(\cdot)$ denote the limit. As we will see, the limit will not depend on the subsequence. Let q be an arbitrary integer, and s_i, $i \leqq q$, t, s arbitrary numbers except that $s_i < t < t+s$, and let $g(\cdot)$ be a bounded continuous function. Let E_t denote $E_{m_a(t)}$. By the weak convergence and uniform integrability,

$$\begin{aligned} &Eg(W^a(s_i), \Gamma^a(s_i), i \leqq q) E_t[\Gamma^a(t+s) - \Gamma^a(t)][\Gamma^a(t+s) - \Gamma^a(t)]' \\ &\to Eg(W(s_i), \Gamma(s_i), i \leqq q)[\Gamma(t+s) - \Gamma(t)][\Gamma(t+s) - \Gamma(t)]'. \end{aligned} \tag{7.25}$$

Evaluating the $E_t[\ ]$ term and using the independence of the $\{\mu_i\}$, yields (the limits of the sums below are $m_a(t)$, $m_a(t+s)-1$)

$$E_t[\Gamma^a(t+s) - \Gamma^a(t)][\Gamma^a(t+s) - \Gamma^a(t)]' = a E_t \sum \gamma_i^a (\gamma_i^a)' = a \sum \sigma_\mu^2 E_t \psi_i (\psi_i)'. \tag{7.26}$$

Since $\lim |E_t \psi_i \psi_i' - \tilde{R}_i^a| \to 0$ as $|i - m_a(t)| \to \infty$ by (B5), (B6), the limit of the right side is the limit of $\sum \sigma_\mu^2 \tilde{R}_i^a$, which (in turn) is the limit of a $\sum \sigma_\mu^2 R(\theta_i^a)$ which (in turn) equals $\int_t^{t+s} \sigma_\mu^2 R(\theta(v))\, dv$. Due to the arbitrariness of s_i, q, g, s, t, we have that

$$E\{[\Gamma(t+s) - \Gamma(t)][\Gamma(t+s) - \Gamma(t)]' | \Gamma(v), W(v), v \leqq t\} = \int_t^{t+s} \sigma_\mu^2 R(\theta(v))\, dv,$$

hence that the right side of (7.22a) is the quadratic covariation of $\Gamma(\cdot)$ [12]. Since it is absolutely continuous and nonrandom, $\Gamma(\cdot)$ is a Wiener process [12]. Similarly, if the right sides of (7.22b, c) are the cross quadratic covariation of $W(\cdot), \Gamma(\cdot)$, then $(W(\cdot), \Gamma(\cdot))$ is the asserted Wiener process, and the proof will be completed.

[5] In [1], $m(t) = \max\{n: \sum_0^n a_i \leqq t\}$ and $a_i \to 0$ as $i \to \infty$ and $\sum a_i = \infty$; also the superscript a was not used or needed. But the proof can also be used for our case, since only (7.24c) was used. Recall that $m_a(t) = \max\{n: an \leqq t\}$.

We now do a similar calculation for $W^n(\cdot)$. We need only show that (the limits of the sums are $m_a(t)$, $m_a(t+s)-1$ unless otherwise written)

$$aE_t \sum_i \beta_i^a \bar{Y}_i \sum_j \bar{Y}_i' \beta_i^{a'} \tag{7.27}$$

converges to the integral in (7.22b) with limits $(t, t+s)$ instead of $(0, t)$. Equation (7.26) equals (use the convention $\sum_c^b = 0$ if $b < c$)

$$\sum_{l \geqq 0} \sum_{i=m_a(t)}^{m_a(t+s)-l-1} aE_t \beta_i^a \bar{Y}_i \bar{Y}_{i+l}' \beta_{i+l}^{a'} + \sum_{l<0} \sum_{i=m_a(t)+|l|}^{m_a(t+s)-1} aE_t \beta_i^a \bar{Y}_i \bar{Y}_{i+l}' \beta_{i+l}^{a'}. \tag{7.28}$$

For all i, $i+l$ in the range of the above sums, the ϕ-mixing implies that

$$|E_t \beta_i^a \bar{Y}_i \bar{Y}_{i+l}' \beta_{i+l}^a| \leqq K\phi_{|l|}^{1/2}.$$

Since $\sum_{|l| \geqq L} \sum_{i=m_a(t)}^{m_a(t+s)} a\phi_{|l|}^{1/2} \to 0$ as $L \to \infty$, we may evaluate the limit of (7.28) by evaluating the limit of the inner sums individually as $a \to 0$, and then summing over l. By the same argument which we used for $\Gamma^a(\cdot)$ below (7.26), the limit of the lth inner sum is the same as the limit when E_t is replaced by E. Furthermore, by Lemma 1 and its corollaries, β_i^a can be replaced by $\beta_i(\theta_i^a)$ without altering the limit. Upon making these replacements, we see that lth inner sum converges to the lth integral in (7.22b) with limits $(t, t+s)$ instead of $(0, t)$. By the argument used in connection with $\Gamma(\cdot)$, this implies that $W(\cdot)$ is a Wiener process with the asserted covariance.

We need only show that the cross-quadratic covariance between $\Gamma(\cdot)$ and $W(\cdot)$ is (7.22c). The proof of this is the same as that just given for $W(\cdot)$ above. The sum is $\sum_1^\infty$ rather than $\sum_{-\infty}^\infty$, since μ_n is independent of y_i, $i < n$, and of ψ_i, β_i^a, $i \leqq n$. Q.E.D.

REFERENCES

[1] H. J. KUSHNER AND HAI HUANG, *Rates of convergence for stochastic approximation type algorithms*, this Journal, 17 (1979), pp. 607–617.

[2] L. LJUNG, *Analysis of recursive stochastic algorithms*, IEEE Trans. on Automatic Control, AC-22 (1977), pp. 551–575.

[3] H. J. KUSHNER AND D. S. CLARK, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Appl. Math. Sci. Series, no. 26, Springer-Verlag, Berlin, 1978.

[4] M. T. WASAN, *Stochastic Approximation*, Cambridge Univ. Press, Cambridge, 1969.

[5] B. WIDROW et al., *Stationary and nonstationary learning characteristics of the LMS adaptive filter*, Proc. IEEE, 64 (1976), pp. 1151–1162.

[6] K. SENNE, *Adaptive linear discrete-time estimation*, Stanford Univ. Rept. SEL 68-090, June 1968.

[7] J. K. KIM AND L. D. DAVISSON, *Adaptive linear estimation for stationary M-dependent processes*, IEEE Trans. on Information Theory, IT-21 (1975), pp. 23–31.

[8] L. LJUNG, *On positive real transfer functions and the convergence of some recursive schemes*, IEEE Trans. on Automatic Control, AC-22 (1977), pp. 539–550.

[9] H. J. KUSHNER, *Rates of convergence for sequential Monte-Carlo optimization methods*, this Journal, 16 (1978), pp. 150–168.

[10] P. BILLINGSLEY, *Convergence of Probability Measures*, Wiley, New York, 1968.

[11] C. DESOER, *Slowly varying system $\dot{x} = A(t)x$*, IEEE Trans. on Automatic Control, AC-14 (1969), p. 780.

[12] W. WONG, *Stochastic Processes in Information and Dynamical Systems*, McGraw-Hill, New York, 1971.

[13] H. J. KUSHNER, *An averaging method for stochastic approximations with constant parameters; small parameter values*. Proc., 1980 Joint Automatic Control Conference, San Francisco.

[14] H. J. KUSHNER AND HAI HUANG, *Averaging methods for the asymptotic analysis of learning and adaptive systems with small adjustment rate*, submitted to this Journal.

[15] ———, *On the weak convergence of a sequence of general stochastic difference equations to a diffusion*, SIAM J. Appl. Math., to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 1, January 1981

0363-0129/81/1901-0008 $01.00/0

A NOTE ON THE BOUNDARY STABILIZATION OF THE WAVE EQUATION*

GOONG CHEN†

Abstract. We study the energy decay rates of the wave equation in a domain where boundary damping is present. We generalize the geometrical conditions obtained earlier in (J. Math. Pures Appl., 58 (1979), pp. 249–273) by using some more general multipliers of Strauss (Comm. Pure Appl. Math., 28 (1975), pp. 265–278). The interaction between distributed damping and boundary damping is discussed. A regulator problem is also formally discussed by the synthesis method.

Introduction. The wave equation with boundary dissipation

$$\text{(WE)}\quad \begin{cases} w_{tt}(x,t)-\Delta w(x,t)=0, & x\in\Omega, \text{ bounded, open in } \mathbb{R}^n, \quad (0.1)\\ w|_{\Gamma_0}=0, & (0.2)\\ (w_t+\alpha_1 w_n+\alpha_2 w)|_{\Gamma_1}=0, & (0.3)\\ \alpha_1>0, \quad \alpha_2\geqq 0, \quad \Gamma_0\dot{\cup}\Gamma_1=\Gamma=\partial\Omega, & \\ w(x,0)=w_0(x), \quad w_t(x,0)=v_0(x) & (0.4) \end{cases}$$

has been studied in [3], [5], [6], [9], etc. The dissipative boundary condition (0.3) arises naturally from some acoustical and optical problems. It is important in the control theory of partial differential equations because the wave equation (0.1) *can be stabilized* by (0.3) [9], [11]. Moreover, if the domain $(\Omega, \Gamma_0, \Gamma_1)$ satisfies some geometrical conditions [3], the wave equation becomes *uniformly exponentially stabilizable*, and Dirichlet, Neumann or Robin type *boundary feedback controllers* on Γ_1 can be constructed from (0.3).

In § 1, we use a new energy identity of Strauss [12] to prove a uniform stabilizability result. It is shown that if the domain $(\Omega, \Gamma_0, \Gamma_1)$ satisfies the "D-conditions", then the energy of (WE) decays uniformly exponentially. This generalizes the "scsssd" geometrical conditions the author obtained earlier in [3].

In § 2, we discuss some interactions between boundary damping (0.3) and distributed positive and negative dampings for a domain which satisfies the D-conditions.

In § 3, we discuss a regulator problem associated with the boundary control of the wave equation. We show that the solution of a regulator problem is unlikely to satisfy a boundary condition like (0.3). The argument presented there is only formal because of the difficulty in the regularity of solutions.

1. Energy decay of the dissipative wave equation in a domain satisfying D-conditions. Following [3], we define the spaces

$$\mathcal{H}_1\equiv\{(w,v)\in H^1(\Omega)\oplus L^2(\Omega)\,|\,w|_{\Gamma_0}=0\},$$

$$\mathcal{H}_2\equiv D(A)=\{(w,v)\in H^2(\Omega)\oplus H^1(\Omega)\,|\,w|_{\Gamma_0}=v|_{\Gamma_0}=0,\ (v+\alpha_1 w_n+\alpha_2 w)|_{\Gamma_1}=0\},$$

$$A\equiv\begin{bmatrix}0 & 1\\ \Delta & 0\end{bmatrix} \text{ is the wave operator.}$$

Denote $u_i=\partial u/\partial x_i$, $\nabla u=(u_1,\cdots,u_n)$. Let $n=(n_1(x),\cdots,n_N(x))$ denote the unit outward normal field on Γ. We say that $(\Omega, \Gamma_0, \Gamma_1)$ satisfies the "D-conditions" ("D" for

* Received by the editors December 31, 1979, and in final form April 25, 1980.

† Department of Mathematics, Pennsylvania State University, University Park, Pennsylvania 16802.
This work was supported in part by the National Science Foundation under grant MCS 7822830.

decay) if there exists a vector field $l=(l_i(x))\in C^4(\mathbb{R}^N, \mathbb{R}^N)$ *with compact support* such that

(1.1) (1) $$l\cdot n=\sum l_j n_j \leqq 0 \quad \text{a.e. on } \Gamma_0;$$

(1.2) (2) $$l\cdot n=\sum l_j n_j \geqq \gamma_1>0 \quad \text{a.e. on } \Gamma_1, \quad (\gamma_1 \text{ some positive number}),$$

(3) with $l_{ij}\equiv \partial l_i/\partial x_j$, $l_{ijk}\equiv \partial^2 l_i/(\partial x_j \partial x_k)$, etc., the matrix $[l_{ij}(x)-\frac{1}{2}\delta_{ij}]$ is strictly positive definite on Ω, i.e.,

(1.3) $$\sum_{i,j} l_{ij}(x)\xi_i\xi_j \geqq \eta \sum \xi_k^2, \qquad \eta>\tfrac{1}{2} \quad \forall\, x\in\Omega;$$

(4) there exists $\gamma_2>0$ small enough such that

(1.4) $$\left|\sum_{i,j} l_{iijj}\right| \leqq \gamma_2 \quad \text{a.e. on } \Omega,$$

(1.5) $$\left|\sum_{i,j} l_{iij} n_j\right| \leqq \gamma_2 \quad \text{a.e. on } \Gamma_1.$$

The smallness of γ_2 will be clear in the subsequent discussion.

Remarks. (i) Comparing our assumptions with those made by Strauss in [12, p. 265, (i), (ii), (iii)], we understand that here what is important is the behavior of the multipliers $(l_i(x))$ on and near $\bar{\Omega}$; their behavior on the infinite exterior region outside some sphere becomes completely immaterial. A question naturally arises: how much of the technique developed by Morawetz, Ralston, and Strauss [7], which is useful in proving exterior energy decay theorems, can be used in solving the interior decay problem?

(ii) One can argue as in [12, Thm. 3] that there does not exist a closed polygon in $\mathbb{R}^N-\bar{\Omega}$ whose vertices lie on Γ_0 and whose corners make equal angles with n and lie in a normal plane.

(iii) It is easy to see that if we define $\tilde{l}_i(x)=x_i$ or $\tilde{l}_i(x)=x_i+\varepsilon x_i/r(\varepsilon>0, \text{small})$ and let

$$l_i(x)=\theta(x)\tilde{l}_i(x),$$

where

$$\theta(x)=\begin{cases}1 & \text{if } \|x\|\leqq R,\\ 0 & \text{if } \|x\|\geqq 2R,\end{cases} \qquad R \text{ large enough},$$

then (1.1), (1.2), (1.3), (1.4), and (1.5) are all satisfied. So a "scsssd" [3] satisfies the D-conditions.

THEOREM 1. *Let* $(\Omega, \Gamma_0, \Gamma_1)$ *satisfy the D-conditions. The exponential decay of energy*

$$E(t)\equiv \int_\Omega [|\nabla w(x,t)|^2+(w_t(x,t))^2]\,dx \leqq Me^{-\beta t}E(0)$$

holds for some $M, \beta>0$ *for the solution* $w(x,t)$ *of* (WE) *uniformly for all initial states* $(w_0, v_0)\in \mathscr{H}_1$.

Proof. According to the argument in [3], we need only consider the case $\alpha_1\equiv\alpha>0$, $\alpha_2=0$ in (0.3); i.e.,

(0.3)′ $$w_t+\alpha w_n=0 \quad \text{on } \Gamma_1.$$

It is straightforward to verify that

$$
\begin{aligned}
0 &= (w_{tt} - \Delta w) \cdot 2\{(1-\mu)tw_t + \textstyle\sum l_i w_i + \frac{1}{2}(\sum l_{ii} - 1)w\} \\
&= \frac{\partial}{\partial t}\{(1-\mu)t(w_t^2 + |\nabla w|^2) + [2\textstyle\sum l_i w_i + (\sum l_{ii} - 1)w]w_t\} \\
&\quad - \operatorname{div}\{2(1-\mu)tw_t \nabla w + (w_t^2 - |\nabla w|^2)l \\
&\qquad + [2\textstyle\sum l_i w_i + (\sum l_{ii} - 1)w]\nabla w - \frac{1}{2}w^2(\sum_i l_{iij})_{j=1}^n\} \\
&\quad + \mu w_t^2 + 2\textstyle\sum\left[l_{ij} - \left(1 - \frac{\mu}{2}\right)\delta_{ij}\right] w_i w_j - \frac{1}{2}\left(\sum_{i,j} l_{iijj}\right) w^2.
\end{aligned} \tag{1.6}
$$

In the above, the main multipliers we use are due to Strauss [12]. Define

$$
Q(t) \equiv \int_\Omega \{(1-\mu)t(w_t^2 + |\nabla w|^2) + [2\textstyle\sum l_i w_i + (\sum l_{ii} - 1)w]w_t\}\, dx.
$$

Integrating (1.6) over Ω, we obtain

$$
\begin{aligned}
\frac{d}{dt}Q(t) &= \int_{\Gamma_0} (l \cdot n) w_n^2 \, d\sigma + 2(1-\mu)t \int_{\Gamma_1} w_t w_n \, d\sigma + \int_{\Gamma_1} (l \cdot n) w_t^2 \, d\sigma \\
&\quad - \int_{\Gamma_1} (l \cdot n)|\nabla w|^2 \, d\sigma + 2\int_{\Gamma_1} (l \cdot \nabla w) w_n \, d\sigma + 2\int_{\Gamma_1} (\textstyle\sum l_{ii} - 1) w w_n \, d\sigma \\
&\quad - \frac{1}{2}\int_{\Gamma_1} \left(\sum_{i,j} l_{iij} n_j\right) w^2 \, d\sigma - \mu \int_\Omega w_t^2 \, dx \\
&\quad - 2\int_\Omega \sum\left[l_{ij}\left(1 - \frac{\mu}{2}\right)\delta_{ij}\right] w_i w_j \, dx + \frac{1}{2}\int_\Omega \left(\sum_{i,j} l_{iijj}\right) w^2 \, dx \\
&\equiv T_1 + T_2 + T_3 + T_4 + T_5 + T_6 + T_7 + T_8 + T_9 + T_{10}.
\end{aligned} \tag{1.7}
$$

For any μ satisfying $0 < \mu < 1$, there exists $t_0 > 0$ such that

$$
\frac{(1-\mu)t}{2} E(t) \leqq Q(t) \leqq tE(t), \qquad t \geqq t_0. \tag{1.8}
$$

We choose

$$
\mu = \begin{cases} 1 - (\eta - \frac{1}{2}) & \text{if } \eta < \frac{3}{2}, \quad (1.9) \\ \frac{1}{2} & \text{if } \eta \geqq \frac{3}{2}; \quad (1.10) \end{cases}
$$

then

$$
T_9 \leqq -(\eta - \tfrac{1}{2}) \int_\Omega |\nabla w|^2 \, dx \leqq 0. \tag{1.11}
$$

Now, we want to show that (1.7) is nonpositive for all $t \geqq t_1$ for some $t_1 > 0$.

$T_1 \leqq 0$ due to (1.1),
$T_2 \leqq 0$ due to (0.3),
T_3 can be absorbed into T_2 for all t large enough,
$T_4 \leqq 0$ due to (1.2),
T_5 can be absorbed into T_2 and T_4 for all t large enough,
T_6 can be absorbed into T_2 and (1.11) by the trace theorem for all t large enough,
T_7 can be absorbed into (1.11) by (1.5) if γ_2 is small,

$T_8 \leqq 0$,

T_{10} can be absorbed into (1.11) if γ_2 is small by Poincaré's inequality.

Therefore $dQ/dt \leqq 0$ for t sufficiently large. Thus $Q(t)$ is nonincreasing for all t large enough and by (1.8) the exponential decay follows. □

After Theorem 1, many corollaries such as exact boundary controllability, observability and parabolic controllability can be immediately obtained for domains satisfying D-conditions. We refer to [3] for the details.

2. Distributed damping and boundary damping. We consider a wave equation with distributed viscous damping

$$w_{tt} + 2\gamma w_t - \Delta w = 0, \qquad \gamma > 0. \tag{2.1}$$

With energy-conserving boundary conditions, one can show that the energy decay rate of (2.1) is $Me^{-2\gamma t}$ if γ is small [1]. Consider a situation where an energy-conserving boundary condition is replaced by

$$w|_{\Gamma_0} = 0, \qquad (w_t + \alpha w_n)|_{\Gamma_1} = 0, \qquad \alpha > 0; \tag{2.2}$$

i.e., boundary damping is also present on the Γ_1 part of the boundary. In general, one cannot expect any faster decay rate than $Me^{-2\gamma t}$ if $\partial\Omega$ is not well-dented. If $\partial\Omega$ is well-dented in a certain sense, then it is not too surprising that the decay rate can be improved.

THEOREM 2. *Let* $(\Omega, \Gamma_0, \Gamma_1)$ *satisfy the D-conditions. Assume that* γ, $\alpha\gamma$ *are positive and small. Then the solution* $w(x, t)$ *of* (2.1), (2.2), (0.4) *satisfies the following energy decay rate for some* $M > 0$:

$$E(t) \leqq Me^{-2\gamma t}t^{-1}E(0), \qquad t > 0 \tag{2.3}$$

uniformly for all initial states $(w_0, v_0) \in \mathscr{H}_1$.

Proof. We choose suitable multipliers:

$$\begin{aligned} 0 &= (w_{tt} + 2\gamma w_t - \Delta w) \cdot 2e^{2\gamma t}\{(1-\mu)tw_t + \textstyle\sum l_i w_i + [\gamma(1-\mu)t + \frac{1}{2}(\sum l_{ii} - 1)]w\} \\ &= \frac{\partial}{\partial t} e^{2\gamma t}\{(1-\mu)t(w_t^2 + |\nabla w|^2) + 2w_t \textstyle\sum l_i w_i + [2\gamma(1-\mu)t + (\sum l_{ii} - 1)]w_t w - \gamma w^2\} \\ &\quad - \operatorname{div} e^{2\gamma t}\Big\{2(1-\mu)tw_t\nabla w + (w_t^2 - |\nabla w|^2)l + 2(\textstyle\sum l_i w_i)\nabla w \\ &\qquad + [2\gamma(1-\mu)t + (\sum l_{ii} - 1)]w\nabla w - \tfrac{1}{2}w^2\Big(\sum_i l_{iij}\Big)_{j=1}^{N}\Big\} \\ &\quad + e^{2\gamma t}\{\mu w_t^2 + 2\sum_{i,j}\Big[l_{ij} - \Big(1 - \frac{\mu}{2}\Big)\delta_{ij}\Big]w_i w_j + (2\gamma^2 - \tfrac{1}{2}\textstyle\sum l_{iijj})w^2 + 2\gamma\mu w w_t\}. \end{aligned} \tag{2.4}$$

Define

$$\bar{Q}(t) \equiv e^{2\gamma t}\{(1-\mu)t(w_t^2 + |\nabla w|^2) + 2w_t \textstyle\sum l_i w_i + [2\gamma(1-\mu)t + (\sum l_{ii} - 1)]w_t w - \gamma w^2\}.$$

Integrating (2.4) over Ω, we obtain

$$\frac{d}{dt}\bar{Q}(t) = \sum_{i=1}^{11} \bar{T}_i,$$

where

$$\bar{T}_i \equiv T_i \cdot e^{2\gamma t} \quad \text{as in (1.7)}, \qquad i = 1, 2, 3, 4, 5, 7, 8, 9,$$

$$\bar{T}_6 \equiv e^{2\gamma t} \int_{\Gamma_1} [2\gamma(1-\mu)t + (\sum l_{ii} - 1)] w w_n \, d\sigma,$$

$$\bar{T}_{10} \equiv -e^{2\gamma t} \int_{\Omega} (2\gamma^2 - \tfrac{1}{2} \sum l_{iijj}) w^2 \, dx,$$

$$\bar{T}_{11} \equiv -2\gamma\mu \, e^{2\gamma t} \int_{\Omega} w w_t \, dx.$$

We then choose μ as in (1.9), (1.10). The counterpart of (1.8) in this case is

$$\frac{(1-\mu)t}{2} e^{2\gamma t} E(t) \leqq \bar{Q}(t) \leqq t e^{2\gamma t} E(t), \qquad t \geqq t_0. \tag{2.5}$$

The estimates can be made in the same fashion as in the proof of Theorem 1, provided that γ and $\alpha\gamma$ are small enough. (The additional term $\bar{T}_{11}$ can be absorbed into $\bar{T}_8 + \bar{T}_9$). Therefore, we again have

$$\frac{d}{dt} \bar{Q}(t) \leqq 0 \quad \text{for all } t \geqq t_1,$$

for some $t_1 > 0$. In view of (2.5), we conclude (2.3). □

Now, consider another situation. Suppose there is the presence of "negative damping" in Ω and also (positive) boundary damping on the Γ_1 part of the boundary $\partial\Omega$. The equation becomes

$$\begin{cases} w_{tt} - 2\gamma w_t - \Delta w = 0, & \gamma > 0, \\ \text{boundary conditions (2.2)}, \\ \text{initial conditions (0.4)}. \end{cases} \tag{2.6}$$

From

$$\frac{d}{dt}[\text{energy of this system}] = \frac{d}{dt} \int_{\Omega} [(w_t)^2 + |\nabla w|^2] \, dx$$

$$= 2\gamma \int_{\Omega} w_t^2 \, dx - \alpha^{-1} \int_{\Gamma_1} w_t^2 \, d\sigma,$$

we see that some energy is generated on Ω but dissipated on Γ_1. What can we say about the energy of this system? The following theorem somewhat says that in this "competition", boundary damping seems to "win".

THEOREM 3. *Let* $(\Omega, \Gamma_0, \Gamma_1)$ *satisfy the D-conditions and assume that* γ *is positive small. Then the energy of the equation* (2.6) *decays uniformly exponentially.*

Proof. The infinitesimal generator associated with the (2.6) is

$$A_1 = \begin{bmatrix} 0 & I \\ \Delta & 2\gamma \end{bmatrix}.$$

It is a bounded perturbation of the operator A by

$$B = \begin{bmatrix} 0 & 0 \\ 0 & 2\gamma \end{bmatrix}.$$

Since A is known to generate a semigroup $S(t)$ satisfying

$$\|S(t)\| \leqq Me^{-\beta t}, \qquad t \geqq 0,$$

by [8, Thm. 3.1.1, p. 80], the semigroup $S_1(t)$ generated by A_1 satisfies

$$\|S_1(t)\| \leqq Me^{(-\beta+M\|B\|)t} \leqq Me^{(-\beta+2\gamma M)t}.$$

Take $\gamma \leqq (\beta/3M)$; then the proof is complete. □

The proof above also applies to the case

$$B = \begin{bmatrix} 0 & 0 \\ \gamma_1 & \gamma_2 \end{bmatrix}, \qquad \gamma_1, \gamma_2 > 0$$

in (2.7). As γ_1 and γ_2 become large, more and more eigenvalues of $A+B$ cross the imaginary axis into the right halfplane. But one cannot expect $A+B$ to generate a semigroup $S_1(t)$ with exponential growth rate by letting γ_1 and γ_2 be arbitrarily large, since A is known to have an infinite point spectrum distributed along the negative real axis ([6], [9]).

One might ask yet another question: what might happen if there is (positive) viscous damping on Ω but "negative boundary damping" on Γ_1? That is,

$$\begin{cases} w_{tt} + 2\gamma w_t - \Delta w = 0, & \gamma > 0, \\ w|_{\Gamma_0} = 0, \quad (w_t - \alpha w_n)|_{\Gamma_1} = 0, & \alpha > 0, \\ \text{initial conditions.} \end{cases}$$

Unfortunately, this problem is not well-posed in general [5]. It is especially "strongly ill-posed" when $\alpha = 1$.

3. A regulator problem. The regulator problem corresponding to a *distributed parameter* control for the wave equation is relatively simple. One can apply the work of R. Datko [4] directly and obtain exponentially stabilizing feedback control for certain cases (see [2]). This becomes a very difficult problem if the control is *boundary value control.* R. Datko's theorems are not applicable because no existence of any semigroups is guaranteed a priori. Consider

$$\inf_{u \in L^2(0,\infty;L^2(\Gamma_1))} \int_0^\infty \left[\left\langle W \begin{bmatrix} w(\cdot, t; u) \\ v(\cdot, t; u) \end{bmatrix}, \begin{bmatrix} w(\cdot, t; u) \\ v(\cdot, t; u) \end{bmatrix} \right\rangle_{\mathcal{H}_1} + \langle Nu(\cdot, t), u(\cdot, t) \rangle_{L^2(\Gamma_1)} \right] dt, \tag{3.1}$$

where (w, v) satisfies

$$\begin{cases} \dfrac{d}{dt} \begin{bmatrix} w(\cdot, t; u) \\ v(\cdot, t; u) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ \Delta & 0 \end{bmatrix} \begin{bmatrix} w(\cdot, t; u) \\ v(\cdot, t, u) \end{bmatrix}, & (3.2) \\ \begin{bmatrix} w(\cdot, 0; u) \\ v(\cdot, 0; u) \end{bmatrix} = \begin{bmatrix} w_0 \\ v_0 \end{bmatrix} \in \mathcal{H}_1, & (3.3) \\ (\beta_1 w_n + \beta_2 w)|_{\Gamma_1} = u \in L^2(0, \infty; L^2(\Gamma_1)) \quad \text{on } \Gamma_1, \quad \beta_1 > 0, \quad \beta_2 \geqq 0, & (3.4) \\ w|_{\Gamma_0} = 0, & (3.5) \end{cases}$$

and

W = a self-adjoint strictly positive bounded linear operator on $\mathcal{H}_1$,

N = a self-adjoint strictly positive bounded linear operator on $L^2(\Gamma_1)$.

The boundary conditions (3.4) ensure the regularity $(w, v) \in L^2(0, T; \mathcal{H}_1)$ for any $T > 0$.

According to Theorem 1 and follow-up exact controllability results, if $(\Omega, \Gamma_0, \Gamma_1)$ satisfies the D-conditions, then (3.1) has at least one solution $u \in L^2(0, \infty; L^2(\Gamma_1))$. Consequently, it has a unique solution $\hat{u}$.

Synthesis procedures for (3.1)–(3.5) are not justified because the adjoint state is not smooth enough. The following discussion is only *formal*. It shows that solutions to the regulator problem are unlikely to have boundary conditions like (0.3) as stabilizing boundary conditions.

The adjoint state (p_∞, q_∞) is the limit of

$$\begin{cases} \dfrac{d}{dt}\begin{bmatrix} p_T \\ q_T \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ \Delta & 0 \end{bmatrix}\begin{bmatrix} p_T \\ q_T \end{bmatrix} - W\begin{bmatrix} w(\hat{u}_T) \\ v(\hat{u}_T) \end{bmatrix}, & 0 \leqq t \leqq T, \quad (3.6) \\ \begin{bmatrix} p_T(\cdot, T) \\ q_T(\cdot, T) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \\ (\beta_1 p_n + \beta_2 p)|_{\Gamma_1} = 0, & (3.7) \\ q_T|_{\Gamma_0} = 0, & (3.8) \\ q_T = -\beta_1 N \hat{u}_T \quad \text{on } \Gamma_1, & (3.9) \end{cases}$$

as $T \to \infty$, if such a limit exists. Assuming the feedback relation

$$\begin{bmatrix} p_\infty \\ q_\infty \end{bmatrix} = Q_\infty \begin{bmatrix} w(\hat{u}) \\ v(\hat{u}) \end{bmatrix}, \tag{3.10}$$

we derive formally the Riccati equation (cf., e.g., [10])

$$Q_\infty A - A Q_\infty + W = 0. \tag{3.11}$$

For simplicity, let

$$W = \begin{bmatrix} W_1 & 0 \\ 0 & W_2 \end{bmatrix},$$

and write

$$Q_\infty = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix};$$

then from (3.11) we get

$$Q_{12}\Delta - Q_{21} = -W_1, \tag{3.12}$$

$$Q_{11} - Q_{22} = 0, \tag{3.13}$$

$$Q_{22}\Delta - \Delta Q_{11} = 0, \tag{3.14}$$

$$Q_{21} - \Delta Q_{12} = -W_2. \tag{3.15}$$

From (3.4), (3.9), (3.10), we derive

$$\begin{aligned} \beta_1 w_n + \beta_2 w = \hat{u} &= -\frac{1}{\beta_1} N^{-1} q_\infty \\ &= -\frac{1}{\beta_1} N^{-1}(Q_{21} w + Q_{22} v). \end{aligned} \tag{3.16}$$

From here we observe that even in the simplest case $N = \beta I$ (I is the identity operator on $L^2(\Gamma_1)$) and $W = I$ (the identity operator on $\mathcal{H}_1$), (3.16) produces a boundary

condition

$$Q_{22}w_t + \beta_1 w_n + (\beta_2 I + Q_{21})w = 0,$$

which can never agree with the form (0.3) because from (3.12) and (3.15), we have

$$Q_{12}\Delta - \Delta Q_{12} = -(W_1 + W_2) = -2I.$$

Thus, Q_{12} does not compute with Δ and neither does Q_{21}. Hence, Q_{21} cannot be a constant operator.

An interesting question remains open: how to construct, if possible, an exponentially stable semigroup from the regulator problem in a domain $(\Omega, \Gamma_0, \Gamma_1)$ satisfying the D-conditions?

Acknowledgments. Most of the results in this note were originally contained in [2]. The author wishes to thank the referee for the suggestion to write this separate note and to improve the earlier results. He would also like to thank Professor Walter Strauss for sending him a reprint of [12].

REFERENCES

[1] G. CHEN, *Control and stabilization for the wave equation in a bounded domain*, this Journal, 17 (1979), pp. 66–81.

[2] ———, *Control and stabilization for the wave equation in a bounded domain, Part II*, this Journal, this issue, pp. 114–122.

[3] ———, *Energy decay estimates and exact boundary value controllability for the wave equation in a bounded domain*, J. Math. Pures Appl., 58 (1979), pp. 249–273.

[4] R. DATKO, *A linear control problem in an abstract Hilbert space*, J. Differential Equations, 9 (1971), pp. 346–369.

[5] A. MAJDA, *Disappearing solutions for the dissipative wave equation*, Indiana Univ. Math. J., 24 (1975), pp. 1119–1133.

[6] ———, *The location of the spectrum for the dissipative acoustic operator*, Indiana Univ. Math. J., 25 (1976), pp. 973–987.

[7] C. S. MORAWETZ, J. V. RALSTON AND W. A. STRAUSS, *Decay of solutions of the wave equation outside non-trapping obstacles*, Comm. Pure Appl. Math., XXX (1977), pp. 447–508.

[8] A. PAZY, *Semigroup of Linear Operators and Application to Partial Differential Equations*, Lecture note #10, Department of Mathematics, University of Maryland, College Park, MD, 1974.

[9] J. P. QUINN AND D. L. RUSSELL, *Asymptotic stability and energy decay rates for solutions of hyperbolic equations with boundary damping*, Proc. Royal Soc. Edinburgh, Ser. A, 77 (1977/78), pp. 97–127.

[10] D. L. RUSSELL, *Mathematics of Finite-Dimensional Control Systems*, Marcel-Dekker, New York, 1979.

[11] M. SLEMROD, *Stabilization of boundary control systems*, J. Differential Equations, 22 (1976), pp. 402–415.

[12] W. A. STRAUSS, *Dispersal of waves vanishing on the boundary of an exterior domain*, Comm. Pure Appl. Math., 28 (1975), pp. 265–278.

[13] C. BARDOS AND G. CHEN, *Control and stabilization for the wave equation, Part III: Domain with moving boundary*, this Journal, this issue, pp. 123–138.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 1, January 1981

0363-0129/81/1901-0009 $01.00/0

CONTROL AND STABILIZATION FOR THE WAVE EQUATION IN A BOUNDED DOMAIN, PART II*

GOONG CHEN†

Abstract. The present note makes a further study on the distributed parameter control and stabilization for the wave equation in an earlier article (SIAM J. Control Optim., 17 (1979) pp. 66–81). Decay rates and control time are improved by a new stabilization scheme of combined viscous damping and compensation. A stabilizing control for a regulator problem is also derived.

Introduction. This note is a sequel to [2], where we began our study of the distributed parameter controllability and stabilizability of the wave equation in a bounded domain. Here we improve the main results of [2] as follows. We show a new stabilization scheme including both viscous damping and compensation which leads to arbitrarily large uniform energy decay rates. This is then used to establish instantaneous distributed parameter exact controllability by feedback controls. Some other properties of the feedback controls are also discussed. A natural consequence of [2] and the present note is an exponentially stabilizing feedback from the regulator problem, which leads to another stabilization scheme.

We will continue our study on the control and stabilization for the wave equation, for the case of a domain with moving boundary, in Part III of this series of papers [12].

1. Stabilization with viscous damping and compensation. Instantaneous controllability. Let Ω be a bounded, open and connected domain with boundary $\partial\Omega$ consisting of $\Gamma_0 \dot\cup \Gamma_1 \dot\cup \Gamma_2 \dot\cup \cdots \dot\cup \Gamma_n$. Consider the equations

$$\begin{cases} \dfrac{\partial^2 w}{\partial t^2}(x,t)+2\gamma_1\dfrac{\partial w}{\partial t}(x,t)+\gamma_2 w(x,t)-\Delta w(x,t)=0, & x\in\Omega; \\ w(x,0)=w_0(x), \\ \dfrac{\partial w}{\partial t}(x,0)=v_0(x); \end{cases} \tag{1.1}$$

$$w(x,t)|_{\Gamma_0}=0; \tag{1.2}$$

$$\begin{cases} \left[\alpha_i w(x,t)+\beta_i\dfrac{\partial w}{\partial n}(x,t)\right]\Big|_{\Gamma_i}=0, & i=1,\cdots,n, \\ \alpha_i\geqq 0, \quad \beta_i\geqq 0, \quad \alpha_i^2+\beta_i^2\neq 0. \end{cases} \tag{1.3}$$

The above in the form of a system is

$$\frac{d}{dt}\begin{bmatrix} w(\cdot,t) \\ v(\cdot,t)\end{bmatrix}=\begin{bmatrix} 0 & 1 \\ \Delta-\gamma_2 & -2\gamma_1\end{bmatrix}\begin{bmatrix} w(\cdot,t) \\ v(\cdot,t)\end{bmatrix}\equiv\left(A+\begin{bmatrix} 0 & 1 \\ -\gamma_2 & -2\gamma_1\end{bmatrix}\right)\begin{bmatrix} w(\cdot,t) \\ v(\cdot,t)\end{bmatrix}.$$

The underlying space is

$$\mathscr{H}_1\equiv\{(w,v)\in H^1(\Omega)\oplus H^0(\Omega)\,|\,w \text{ satisfies (1.2) in } H^{1/2}(\Gamma_0)\},$$

with A = the wave operator, and

$$D(A)=\mathscr{H}_2\equiv\{(w,v)\in H^2(\Omega)\oplus H^1(\Omega)\,|\,w \text{ satisfies (1.2), (1.3)}\}.$$

* Received by the editors June 15, 1979, and in final revised form April 25, 1980.

† Department of Mathematics, Pennsylvania State University, University Park, Pennsylvania, 16802.
This work was supported in part by the National Science Foundation under grant MCS 7822830.

The domain Ω may be a polygon, with a mixed boundary condition on each side Γ_i.

In (1.1), the term $2\gamma_1 \partial w/\partial t$ is viscous damping and the term $\gamma_2 w$ is dispersion. Because of the similarity between (1.1) and a second order linear system

$$\ddot{y} + 2\gamma_1 \dot{y} + \gamma_2 y = 0,$$

we call $\gamma_2 w$ the compensation term. When it was first introduced in [4], the authors failed to realize that the energy decay rate of (1.1)–(1.3) can be made as large as desired provided that γ_1 and γ_2 are chosen large enough.

THEOREM 1.1. *For any $k>0$, let $w(x,t)$ be the solution of* (1.1)–(1.3) *with $\gamma_1 \equiv k(3+2k)$, $\gamma_2 \equiv 4k(1+3k+4k^2)$. Then the energy of the system decays uniformly exponentially*:

$$\int_\Omega \left[\left(\frac{\partial w}{\partial t}(x,t) \right)^2 + |\nabla w(x,t)|^2 \right] dx \leqq C(k)\, e^{-2kt} \left\| \begin{bmatrix} w_0 \\ v_0 \end{bmatrix} \right\|^2_{\mathscr{H}_1} \tag{1.4}$$

where $C(k) = O(k^3)$.

Proof. Again, we use the energy method [2], [11]. We first consider those $(w_0, v_0) \in \mathscr{H}_2$. Multiplying (1.1) by $\partial w/\partial t$ and integrating by parts, we obtain

$$\frac{1}{2}\frac{d}{dt}\int \left(\frac{\partial w}{\partial t}\right)^2 dx - \sum_{i=1}^{n} \int_{\Gamma_i} \frac{\partial w}{\partial n}\frac{\partial w}{\partial t}\, d\sigma + \frac{1}{2}\frac{d}{dt}\int |\nabla w|^2\, dx + 2\gamma_1 \int \left(\frac{\partial w}{\partial t}\right)^2 dx + \frac{\gamma_2}{2}\frac{d}{dt}\int w^2\, dx = 0.$$

With the boundary conditions (1.2), (1.3) taken into consideration, the above becomes

$$\frac{1}{2}\frac{d}{dt}\left\{ \int \left[|\nabla w|^2 + \left(\frac{\partial w}{\partial t}\right)^2 + \gamma_2 w^2 \right] dx + \sum_{\beta_i \neq 0} \frac{\alpha_i}{\beta_i} \int_{\Gamma_i} w^2\, d\sigma \right\} + 2\gamma_1 \left(\frac{\partial w}{\partial t}\right)^2 dx = 0. \tag{1.5}$$

Similarly, we use λw as multiplier and obtain

$$\lambda \left\{ \frac{d}{dt}\int \frac{\partial w}{\partial t}\, w\, dx - \int \left(\frac{\partial w}{\partial t}\right)^2 dx + \int |\nabla w|^2\, dx + \sum_{\beta_i \neq 0} \frac{\alpha_i}{\beta_i} \int_{\Gamma_i} w^2\, d\sigma + \gamma_1 \frac{d}{dt}\int w^2\, dx + \gamma_2 \int w^2\, dx \right\} = 0. \tag{1.6}$$

Adding (1.5) and (1.6), we obtain

$$\frac{1}{2}\frac{d}{dt}\left\{ \int \left[\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2 + (\gamma_2 + 2\lambda\gamma_1) w^2 + 2\gamma w \frac{\partial w}{\partial t} \right] dx + \sum_{\beta_i \neq 0} \frac{\alpha_i}{\beta_i} \int_{\Gamma_i} w^2\, d\sigma \right\}$$
$$+ \left\{ \int \left[(2\gamma_1 - \lambda) \left(\frac{\partial w}{\partial t}\right)^2 + \lambda |\nabla w|^2 + \lambda\gamma_2 w^2 \right] dx + \lambda \sum_{\beta_i \neq 0} \frac{\alpha_i}{\beta_i} \int_{\Gamma_i} w^2\, d\sigma \right\} = 0.$$

We write the above simply as

$$\frac{d}{dt} P(t) + Q(t) = 0. \tag{1.7}$$

Our proof given here differs from the one in [2], [11] in that λ is not small; indeed, we choose

$$\lambda = 2k.$$

Then with $\gamma_1 = k(3+2k)$, $\gamma_2 = 4k(1+3k+4k^2)$, the inequalities

$$2\gamma_1 - \lambda \geqq k(1+\lambda), \qquad \lambda\gamma_2 \geqq k(\lambda + \gamma_2 + 2\lambda\gamma_1),$$

are satisfied so we have

$$Q(t) \geqq k\left\{ \int \left[(1+\lambda)\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2 + (\lambda + \gamma_2 + 2\lambda\gamma_1) w^2 \right] dx + \sum_{\beta_i \neq 0} \frac{\alpha_i}{\beta_i} \int_{\Gamma_i} w^2 \, d\sigma \right\}$$
$$\geqq 2kP(t).$$

Thus, from (1.7),

$$\frac{d}{dt} P(t) + 2kP(t) \leqq \frac{d}{dt} P(t) + Q(t) = 0;$$

therefore,

$$P(t) \leqq P(0)\, e^{-2kt}.$$

It is easy to verify that

$$P(t) \geqq \frac{1}{4} \int \left[\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2 \right] dx.$$

On the other hand, from Poincaré's inequality,

$$\int w^2 \, dx \leqq K \int |\nabla w|^2 \, dx$$

and the trace theorem

$$\sum_{\beta_i \neq 0} \frac{\alpha_i}{\beta_i} \int_{\Gamma_i} w^2 \, d\sigma \leqq K \int |\nabla w|^2 \, dx, \qquad (K \text{ chosen} \geqq 1),$$

we see that

$$P(t) \leqq \frac{1}{2}\left\{ \int \left[\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2 + (\gamma_2 + 2\lambda\gamma_1) w^2 + \frac{1}{2}\left(\frac{\partial w}{\partial t}\right)^2 + 2\lambda^2 w^2 \right] dx + \sum_{\beta_i \neq 0} \frac{\alpha_i}{\beta_i} \int_{\Gamma_i} w^2 \, d\sigma \right\}$$
$$\leqq \frac{1}{2}\left\{ \int \left[\frac{3}{2}\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2 + (4k + 32k^2 + 24k^3) w^2 \right] dx + \sum_{\beta_i \neq 0} \frac{\alpha_i}{\beta_i} \int_{\Gamma_i} w^2 \, d\sigma \right\}$$
$$\leqq K(2 + 4k + 32k^2 + 24k^3) \int_\Omega \left[\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2 \right] dx.$$

Hence,

$$\int \left[\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2 \right] dx \leqq 4P(t) \leqq 4P(0)\, e^{-2kt}$$
$$\leqq C(k)\, e^{-2kt} \int [v_0^2 + |\nabla w_0|^2] \, dx = C(k)\, e^{-2kt} \left\| \begin{bmatrix} w_0 \\ v_0 \end{bmatrix} \right\|_{\mathscr{H}_1},$$

where $C(k) \equiv 4K(2 + 4k + 32k^2 + 24k^3) = O(k^3)$. □

Remark. In [10], M. Slemrod generalizes a finite-dimensional stabilization scheme: he considers

$$\begin{cases} \dfrac{d}{dt}\begin{bmatrix} w(\cdot,t) \\ v(\cdot,t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ \Delta & 0 \end{bmatrix}\begin{bmatrix} w(\cdot,t) \\ v(\cdot,t) \end{bmatrix} + Bu(t), \\ \begin{bmatrix} w(\cdot,0) \\ v(\cdot,0) \end{bmatrix} = \begin{bmatrix} w_0 \\ v_0 \end{bmatrix} \quad \in \mathscr{H}_1 \equiv H_0^1(\Omega) \oplus H^0(\Omega), \end{cases}$$

$$D(A) = (H^2(\Omega) \cap H_0^1(\Omega)) \oplus H_0^1(\Omega), \qquad B: U \to \mathscr{H}_1.$$

It is shown that if B satisfies

$$\int_0^\varepsilon \left\| B^* T^*(-t) \begin{bmatrix} w_0 \\ v_0 \end{bmatrix} \right\|_U^2 dt \geqq \delta \left\| \begin{bmatrix} w_0 \\ v_0 \end{bmatrix} \right\|_{\mathscr{H}_1}, \quad \text{some } \delta > 0, \tag{1.8}$$

where $T^*(t) =$ the group generated by $A^*(= -A)$, and if we take

$$\mathbb{K} \equiv -B^* D_{\varepsilon,\lambda}^{-1}, \qquad D_{\varepsilon,\lambda} \equiv \int_0^\varepsilon e^{-2\lambda t} T(-t) BB^* T^*(-t)\, dt, \tag{1.9}$$

then the semigroup $S(t)$ generated by $A + B\mathbb{K}$ satisfies

$$\|S(t)\| \leqq M e^{-\lambda t}, \qquad (M \text{ depends on } \lambda), \tag{1.10}$$

where λ can also be made as large as desired. In particular, if

$$\begin{aligned} U &\equiv \mathscr{H}_1, \\ B &= \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \end{aligned} \tag{1.11}$$

then (1.8) becomes a special case of the observability theorem proved in [4, Thm. 2.3.4]. Since $B^* = B$, $A + (-2\gamma)BB^*$ is the operator

$$\begin{bmatrix} 0 & 1 \\ \Delta & -2\gamma \end{bmatrix},$$

which has been discussed in detail in [2], and we understand that $A + (-2\gamma)BB^*$ can not generate a semigroup with arbitrarily large decay rates. Slemrod's stabilization scheme $A + B\mathbb{K}$ with $\mathbb{K}$ given as in (1.9) obviously produces a stronger result than $A + (-2\gamma)BB^*$ does in [2].

If one looks into the proof of [10, Thm. 2.1] more carefully, using a simple eigenfunction argument, one may find that the constant M in (1.10) is of the order of magnitude $O(\sqrt{\lambda})$ as λ becomes large (with $B =$ (1.11)). That is to say, $A + B\mathbb{K}$ generates a semigroup $S(t)$ satisfying

$$\|S(t)\| \leqq C\sqrt{\lambda}\, e^{-\lambda t}, \qquad \lambda \text{ large}. \tag{1.12}$$

Our result (1.4) in Theorem 1.1 works in the same direction as Slemrod's [10, Thm. 2.1]. With $\lambda = 2k$, the decay rate (1.12) is faster than (1.4). But the stabilized equation (1.1) is easier to construct and solve than (1.9).

As an immediate consequence, we have the following theorem of "the instantaneous controllability of the wave equation". It improves an earlier result in [2].

THEOREM 1.2: *Given any $T>0$ and any initial and final states $(w_0, v_0), (w_1, v_1) \in \mathscr{H}_1$, there exists a control $f_T(x,t) \in C([0,T]; L^2(\Omega))$ such that the solution of*

$$\begin{cases} \dfrac{\partial^2 w}{\partial t^2}(x,t) - \Delta w(x,t) = f_T(x,t), & x \in \Omega, \quad t \geqq 0, \\ w(x,0) = w_0(x), \qquad \dfrac{\partial w}{\partial t}(x,0) = v_0(x), \\ \textit{boundary conditions } (1.2), (1.3), \end{cases}$$

satisfies $w(x,T) = w_1(x)$, $\partial w(x,T)/\partial t = v_1(x)$. In other words, the wave equation can be controlled as fast as we wish by distributed parameter controls.

Proof. In order that this paper be sufficiently self-contained, we reproduce Russell's arguments presented in [2], [8].

Let $\tilde{\eta}$ be the solution of

$$\frac{\partial^2 \tilde{\eta}}{\partial t^2} - \Delta \tilde{\eta} = -\gamma_1 \frac{\partial \tilde{\eta}}{\partial t} - \gamma_2 \tilde{\eta},$$

$$\begin{bmatrix} \tilde{\eta}_0 \\ \tilde{\zeta}_0 \end{bmatrix} \in \mathscr{H}_1, \qquad \text{(initial state)}, \tag{1.13}$$

and let γ_1, γ_2 be large enough so that Theorem 1.1 applies to give the energy decay result

$$\left\| \begin{bmatrix} \tilde{\eta}(\cdot, T) \\ \tilde{\zeta}(\cdot, T) \end{bmatrix} \right\|_{\mathscr{H}_1} \leqq C(k) e^{-kT} \left\| \begin{bmatrix} \tilde{\eta}_0 \\ \tilde{\zeta}_0 \end{bmatrix} \right\|_{\mathscr{H}_1}, \tag{1.14}$$

where $C(k)$ is $O(k^{3/2})$ according to Theorem 1.1. We choose k large enough so that

$$C(k)\, e^{-kT} < 1.$$

Let $\hat{\eta}$ be the solution of

$$\begin{aligned} &\frac{\partial^2 \hat{\eta}}{\partial t^2} - \Delta \hat{\eta} = \gamma_3 \frac{\partial \hat{\eta}}{\partial t} + \gamma_4 \hat{\eta}, \qquad \gamma_3 > 0, \quad \gamma_4 < 0, \\ &\begin{bmatrix} \hat{\eta}(\cdot, T) \\ \hat{\zeta}(\cdot, T) \end{bmatrix} = \begin{bmatrix} \tilde{\eta}(\cdot, T) \\ \tilde{\zeta}(\cdot, T) \end{bmatrix} \in \mathscr{H}_1, \quad \text{(terminal state)}. \end{aligned} \tag{1.15}$$

Reversing the time direction and choosing appropriate γ_3, γ_4, we can also apply Theorem 1.1 and have

$$\left\| \begin{bmatrix} \hat{\eta}(\cdot, 0) \\ \hat{\zeta}(\cdot, 0) \end{bmatrix} \right\|_{\mathscr{H}_1} \leqq C(k)\, e^{-kT} \left\| \begin{bmatrix} \tilde{\eta}(\cdot, T) \\ \tilde{\zeta}(\cdot, T) \end{bmatrix} \right\|_{\mathscr{H}_1}.$$

Now we define

$$\tilde{f}(x,t) \equiv -\gamma_1 \frac{\partial \tilde{\eta}}{\partial t}(x,t) - \gamma_2 \tilde{\eta}(x,t), \qquad x \in \Omega, \quad 0 \leqq t \leqq T, \tag{1.16}$$

$$\hat{f}(x,t) \equiv \gamma_3 \frac{\partial \tilde{\eta}}{\partial t}(x,t) + \gamma_4 \hat{\eta}(x,t), \tag{1.17}$$

and let $\eta \equiv \tilde{\eta} - \hat{\eta}$, $f \equiv \tilde{f} - \hat{f}$. Then η is the solution of

$$\frac{\partial^2 \eta}{\partial t^2} - \Delta \eta = f,$$

with initial state $(\eta(\cdot,0), \zeta(\cdot,0)) = (\tilde{\eta}_0 - \hat{\eta}_0, \tilde{\zeta}_0 - \hat{\zeta}_0)$ and terminal state $(\eta(\cdot,T), \zeta(\cdot,T)) = (0,0)$. Since $(\hat{\eta}_0, \hat{\zeta}_0)$ depends linearly on $(\tilde{\eta}_0, \tilde{\zeta}_0)$, we can write

$$(\hat{\eta}_0, \hat{\zeta}_0) = F(\tilde{\eta}_0, \tilde{\zeta}_0), \qquad F: \mathcal{H}_1 \to \mathcal{H}_1, \qquad \|F\| \leqq [C(k)\, e^{-kT}]^2 < 1 \tag{1.18}$$

so

$$\begin{bmatrix} \eta_0 \\ \zeta_0 \end{bmatrix} = \begin{bmatrix} \eta(\cdot,0) \\ \zeta(\cdot,0) \end{bmatrix} = (I-F)\begin{bmatrix} \tilde{\eta}_0 \\ \tilde{\zeta}_0 \end{bmatrix}.$$

The above relation is always solvable because $\|F\| < 1$. Letting $(\tilde{\eta}_0, \tilde{\zeta}_0) = (I-F)^{-1}(\eta_0, \zeta_0) = (I-F)^{-1}(w_0, v_0)$, we have solved the controllability problem for any $T > 0$. □

COROLLARY 1.3. *The control $f_T(x,t)$ in Theorem* 1.2 *satisfies the following properties*:

(i) *The controlled orbit* $\{(w(\cdot,t), v(\cdot,t)) | 0 \leqq t \leqq T\}$ *is compact in* $\mathcal{H}_1$.

(ii)
$$\lim_{T \to 0+} \|f_T(\cdot,T)\|_{L^2(\Omega)} = 0, \tag{1.19}$$

$$\limsup_{T \downarrow 0\ [0,T]} \|f_T(\cdot,t)\|_{L^2(\Omega)} = \infty. \tag{1.20}$$

Proof. (i) Returning to the proof of Theorem 1.2, we see that for $(\eta_0, \zeta_0) \in D(A)$ (in (1.13)), the trajectory of $\{(\tilde{\eta}(\cdot,t), \tilde{\zeta}(\cdot,t)) | 0 \leqq t \leqq T\}$ is compact in $\mathcal{H}_1$ because $D(A) = \mathcal{H}_2$ is compact in $\mathcal{H}_1$. This property remains valid for $(\tilde{\eta}_0, \tilde{\zeta}_0) \in \mathcal{H}_1$ by the diagonal process (cf. [5]). The same reasoning shows that $\{(\hat{\eta}(\cdot,t), \hat{\zeta}(\cdot,t)) | 0 \leqq t \leqq T\}$ is also compact in $\mathcal{H}_1$. Hence $\{(w(\cdot,t), v(\cdot,t)) | 0 \leqq t \leqq T\}$, which is equal to the difference of two compact trajectories, is compact.

(ii) Since f is equal to $\tilde{f} - \hat{f}$, with (1.16), (1.17) we have

$$\|f(\cdot,T)\|_{L^2(\Omega)} \leqq \|\tilde{f}(\cdot,T)\| + \|\hat{f}(\cdot,T)\|$$

$$\Rightarrow \text{by (1.14), (1.15), (1.16), (1.17) and Poincaré's inequality}$$

$$\leqq MC(k)\, e^{-kT} \left\| \begin{bmatrix} \tilde{\eta}_0 \\ \tilde{\zeta}_0 \end{bmatrix} \right\|_{\mathcal{H}_1}, \qquad (M \text{ independent of } k)$$

$$\leqq MC(k)\, e^{-kt} \|(I-F)^{-1}\|_{\mathcal{L}(\mathcal{H}_1, \mathcal{H}_1)} \left\| \begin{bmatrix} w_0 \\ v_0 \end{bmatrix} \right\|_{\mathcal{H}_1}.$$

Now, we choose $\gamma_1, \gamma_2, \gamma_3, \gamma_4$ in the same fashion as in the proof of Theorem 1.1, we have $C(k) = O(k^{3/2})$ so

$$\lim_{k \to \infty} C(k)\, e^{-kT} \|(I-F)^{-1}\| = 0.$$

Hence (1.19) is proved.

Equation (1.20) is obvious since as the control time becomes smaller the gain must become larger in order to reach x_1 from x_0. □

Remark. From the corollary above, one easily sees that if a constraint

$$\operatorname{ess\,sup}_{0 \leqq t \leqq T} \|f(\cdot,t)\| \leqq C$$

is imposed on f, then one cannot obtain instantaneous controllability. Indeed, under this constraint there will be a unique time-optimal, bang-bang control which satisfies the controllability condition [7].

COROLLARY 1.4. *For any two states* $(w_0, v_0), (w_1, v_1) \in \mathscr{H}_1$ *and any* $T > 0$, *there is a unique distributed parameter* $f \in C([0, T]; L^2(\Omega))$ *which is optimal in* $L^2(\Omega \times [0, T])$ *and steers the system from* (w_0, v_0) *to* (w_1, v_1) *at* $t = T$.

Proof. Let $\{\phi_j\}$ be the set of complete eigenfunctions of $(-\Delta)$ in $L^2(\Omega)$ corresponding to boundary conditions (1.2)–(1.3). Let $\{\lambda_j\}$ be the associated positive eigenvalues (λ_j may have multiplicity $\geqq 1$). Let $\omega_{\pm j} = \pm\sqrt{-\lambda_j}$. Then the controllability problem

$$\begin{cases} \dfrac{\partial^2 w}{\partial t^2}(x, t) - \Delta w(x, t) = f(x, t), & 0 \leqq t \leqq T, \quad x \in \Omega, \\ w(x, 0) = w_0(x), \qquad \dfrac{\partial w}{\partial t}(x, 0) = v_0(x), \\ w(x, T) = w_1(x), \qquad \dfrac{\partial w}{\partial t}(x, T) = v_1(x), \end{cases}$$

is equivalent to the following abstract moment problem:

$$\begin{aligned} &\int_0^T \int_\Omega f(x, t)\, e^{\omega_j t} \phi_j(x)\, dx\, dt \\ &\qquad = \int_\Omega [v_1(x)\, e^{\omega_j T} - w_1(x) \omega_j\, e^{\omega_j T} - v_0(x) - w_0(x)] \phi_j(x)\, dx, \end{aligned} \tag{1.21}$$

for all $j = 1, 2, \cdots$. From the controllability Theorem 1.2, such a moment problem is always solvable with $f \in C([0, T]; L^2(\Omega))$. But $C([0, T]; L^2(\Omega))$ can be naturally embedded in $L^2(\Omega \times [0, T])$, $f \in L^2(\Omega \times [0, T])$. Let

$$M \equiv \text{the closed linear span of } \{e^{\omega_{\pm j} t} \phi_j(x) | j = 1, 2, \cdots\} \text{ in } L^2(\Omega[0, T]).$$

Then

$$L^2(\Omega \times [0, T]) = M^\perp \oplus M,$$

and

$$f = f_1 + f_2, \quad f_1 \in M^\perp, \quad f_2 \in M, \qquad f_1, f_2 \text{ unique.}$$

Then f_2 is a solution of the abstract moment problem (1.21), and $\|f_2\|_{L^2(\Omega \times [0,T])}$ is minimal [9]. The fact that $f_2 \in C([0, T]; L^2(\Omega))$ is nontrivial. One must use a generalization of a finite-dimensional argument in [1]. The detail is omitted. □

2. A regulator problem for the wave equation. We write the control system in Theorem 1.2 as

$$\begin{cases} \dfrac{d}{dt}\begin{bmatrix} w(\cdot, t; f) \\ v(\cdot, t; f) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ \Delta & 0 \end{bmatrix} \begin{bmatrix} w(\cdot, t; f) \\ v(\cdot, t; f) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} f(\cdot, t) \\ \qquad\qquad \equiv A \begin{bmatrix} w(\cdot, t; f) \\ v(\cdot, t; f) \end{bmatrix} + Bf(\cdot, t), \qquad B \equiv \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \\ \begin{bmatrix} w(\cdot, 0, f) \\ v(\cdot, 0, f) \end{bmatrix} = \begin{bmatrix} w_0 \\ v_0 \end{bmatrix} \in \mathscr{H}_1. \end{cases} \tag{2.1}$$

Consider

$$(2.2)\qquad \inf_{f\in L^2(0,\infty;L^2(\Omega))} \int_0^\infty \left[\left\langle W\begin{bmatrix} w(\cdot,t;f)\\ v(\cdot,t;f)\end{bmatrix}, \begin{bmatrix} w(\cdot,t;f)\\ v(\cdot,t;f)\end{bmatrix}\right\rangle_{\mathscr{H}_1} + \langle Uf(\cdot,t), f(\cdot,t)\rangle_{L^2(\Omega)}\right] dt,$$

where

W = a self-adjoint nonnegative linear transformation on $\mathscr{H}_1$, satisfying some observability condition,

U = a self-adjoint strictly positive linear transformation on $L^2(\Omega)$.

Since the system (2.1) is exponentially stabilizable with a feedback

$$(2.3)\qquad f \equiv [0, -\gamma]\begin{bmatrix} w\\ v\end{bmatrix} \equiv \mathbb{K}\begin{bmatrix} w\\ v\end{bmatrix},$$

i.e., $A+B\mathbb{K}$ generates a semigroup with an exponential decay rate, the regulator problem (2.2) has a unique solution $\hat{f}$ in $L^2(0,\infty;L^2(\Omega))$.

THEOREM 2.1. *There exists a self-adjoint linear operator K_∞ on $\mathscr{H}_1$ such that the unique optimal control $\hat{f}$ minimizing* (2.2) *is obtained by use of the feedback relation*

$$\hat{f}(\cdot,t) = -U^{-1}B^*K_\infty\begin{bmatrix} w(\cdot,t)\\ v(\cdot,t)\end{bmatrix}.$$

Furthermore, $A-BU^{-1}B^*K_\infty$ generates a semigroup with an exponential decay rate provided that W is strictly positive.

Proof. The existence of K_∞ is clear from the proofs in [6]. It follows from [6, Corollary 3.1] that $A-BU^{-1}B^*K_\infty$ generates a semigroup exponentially stable since hypothesis A in [6] is always satisfied by use of (2.3). □

Remarks.

(i) This stabilization scheme is different from those mentioned in § 1.

(ii) In the proof of Theorem 2.1, we avoid using Ricatti's equations for feedback synthesis due to difficulties from regularity of solutions.

Acknowledgment. The author would like to thank Professor D. L. Russell for very helpful discussions. He also thanks the referee for very constructive criticism.

REFERENCES

[1] R. W. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.

[2] G. CHEN, *Control and stabilization for the wave equation in a bounded domain*, this Journal, 17 (1979), pp. 66–81.

[3] ———, *Energy decay estimates and exact boundary value controllability for the wave equation in a bounded domain*, J. Math. Pures Appl., 58 (1979), pp. 249–273.

[4] G. CHEN AND R. S. MILLMAN, *Control theory for the wave equation in compact Riemannian manifolds*, Funkcial. Ekvac., to appear.

[5] C. DAFERMOS, *Uniform processes and semicontinuous Lyapunoff functionals*, J. Differential Equations, 4 (1968), pp. 57–65.

[6] R. DATKO, *A linear control problem in an abstract Hilbert space*, J. Differential Equations, 9 (1971), pp. 346–359.

[7] H. O. FATTORINI, *The time optimal problem for distributed control of systems described by the wave equation*, in Control Theory of Systems Governed by Partial Differential Equations, Aziz, Wingate, Balas, eds., Academic Press, New York–San Francisco–London, 1977.

[8] D. L. RUSSELL, *Exact boundary value controllability theorems for wave and heat processes in star-complemented regions*, in Differential Games and Control Theory, Roxin, Liu, Sternberg, eds., Marcel-Dekker, New York, 1974.

[9] T. I. SEIDMAN AND W. C. CHEWNING, *A convergent scheme for the boundary control of the heat equation*, this Journal, 15 (1977), pp. 64–72.

[10] M. SLEMROD, *A note on complete controllability and stabilizability for linear control systems in Hilbert space*, this Journal, 12 (1974), pp. 500–508.

[11] W. A. STRAUSS, *The Energy Method in Nonlinear Partial Differential Equations*, Instituto de Mathematica Pura e Applicada, Brazil, 1969.

[12] C. BARDOS AND G. CHEN, *Control and stabilization for the wave equation, Part III : Domain with moving boundary*, this Journal, this issue, pp. 123–138.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 1, January 1981

0363-0129/81/1901-0010 $01.00/0

CONTROL AND STABILIZATION FOR THE WAVE EQUATION, PART III: DOMAIN WITH MOVING BOUNDARY*

CLAUDE BARDOS† AND GOONG CHEN‡

Abstract. We use energy invariants to study the growth and decay estimates for solutions of the wave equation in a domain with moving boundary. Sufficient conditions are formulated which insure the exact (distributed-parameter) controllability of the wave equation.

Introduction. In three earlier papers [1], [2], [3], we have studied controllability, stabilizability and observability theory for the wave equation in a bounded domain $\Omega \subseteq \mathbb{R}^N$. Ω was a fixed domain with the passage of time. In practical situations, many processes evolve in domains whose boundary has moving parts. A simple model, e.g., is a heat process in a combustion chamber where a piston is attached. Part of the boundary moves with the motion of the piston. Partial differential equations in domains with moving boundary have been studied in [4], [5], [6], [7], etc; see also the references therein.

In this paper, we will be concerned with the distributed-parameter controllability and stabilizability problems of the wave equation in a domain with moving boundary. In particular, the domain is expanding following certain rules. As far as we know, the present paper is the first attempt to resolve the questions above. We prove that (1) the wave equation is stabilizable with the introduction of viscous damping and compensation, and (2) the wave equation is exactly controllable with distributed-parameter controllers.

We first note that the energy of the wave equation increases as the domain expands and decreases as the domain contracts. Therefore, no backward stabilizability [10] is expected in an expanding domain as one reverses the sense of time. This causes some complications to the study of controllability when one tries to apply Russell's complete stabilizability method [10]. Here we devise a scheme with high compensation, so that, when combined with the energy method of Morawetz, it can provide us with an upper bound for the energy growth during time reversal.

Basically, our assumptions are as follows. (A) The space dimension $N \neq 2$. (B) The domain rests still until $t = T_0$. (C) After T_0, the domain expands, but every point on the domain lies within the distance θt, $0 < \theta < 1$, of the origin at $t \geqq T_0$. (D) A geometrical condition on the boundary which in the stationary case reduces to the star-shapedness condition. Condition (A) is very restrictive. Condition (B) is probably redundant. Both (B) and (D) ensure that the energy of the wave equation will not change too drastically in the process of boundary deformation.

In § 1, we start with some basic notation and the statement of the controllability problem. The existence, uniqueness and continuity of solutions of wave equations is stated without proof.

In § 2, we derive the growth and decay estimates for the wave equation (without control) in an expanding domain. It is important in itself as well as preparatory for the material in § 3.

* Received by the editors June 15, 1979 and in final revised form April 25, 1980. A preliminary version of this paper has been published by INRIA, Res. Rep. #12, Rocquencourt, France, March 1980.

† Département de mathématiques, Université Paris-Nord, Paris, France.

‡ IRIA-LABORIA, Rocquencourt, France. Now at Department of Mathematics, Pennsylvania State University, University Park, Pennsylvania, 16802. This work was supported in part by the National Science Foundation under grant MCS 7822830.

In § 3, the distributed-parameter exact controllability theorem is given. We give an example of an expanding sphere in § 4.

There are many control problems related to evolution equations in domains with moving boundary which are yet to be studied:

(1) Boundary-value control for the wave equation.
(2) Distributed-parameter and boundary-value controls for the heat equation.
(3) Control problems for evolution equations in a domain with periodic moving boundary.

We hope we can treat them in the future.

1. Notation. Statement of the problem. Points in the space $\mathbb{R}^N$ are denoted as $x=(x_1, x_2, \cdots, x_N)$. Also $r^2=|x|^2=\sum x_i^2$, $\nabla=(\partial/\partial x_1, \cdots, \partial/\partial x_N)$, $\nabla_{x,t}=(\nabla, \partial/\partial t)$ and $\Delta=\sum \partial^2/\partial x_i^2$. For $t\geqq 0$, we postulate bounded open sets $\Omega(t)$. Let

$$Q(t_1, t_2)\equiv \bigcup_{t=t_1}^{t_2} \Omega(t)\times\{t\}, \qquad \Sigma(t_1, t_2)=\bigcup_{t=t_1}^{t_2} \partial\Omega(t)\times\{t\},$$

denote the space-time domain and the lateral surface from t_1 to t_2. In case $t_1=0$, we simply denote $Q(0, t_2)$ and $\Sigma(0, t_2)$ as $Q(t_2)$ and $\Sigma(t_2)$, respectively. We assume that $\Sigma(t)$ is piecewise smooth for all $t>0$. Let $\nu=(\nu_1, \cdots, \nu_N, \nu_t)=(\nu_x, \nu_t)$ be the unit outward normal at (x, t) on Σ. Throughout this paper, we assume

(H_0) (time likeness of Σ) $\quad |\nu_t|<|\nu_x|$ on $\Sigma(T)$, for any $T>0$.

From [6], we understand that (H_0) holds if and only if each point on the boundary moves in the normal direction at a speed less than one.

If u is a smooth function satisfying $u=0$ on Σ, then all the tangential derivatives of u are also vanishing on Σ. So,

$$\nabla_{x,t}u=\frac{\partial u}{\partial \nu}\nu,$$

$$\frac{\partial u}{\partial t}=\frac{\partial u}{\partial \nu}\nu_t, \quad \nabla u=\frac{\partial u}{\partial \nu}\nu_x, \quad \frac{\partial u}{\partial r}=\frac{\partial u}{\partial \nu}\nu_r \quad \left(\text{here } \nu_r=\nu_x\cdot\frac{x}{r}\right). \tag{1.1}$$

The above remains valid in the sense of distributions if u has a well-defined trace.

For each $t\geqq 0$, $H_0^1(\Omega(t))\oplus H^0(\Omega(t))$ denotes the Hilbert space equipped with the norm

$$\|(w, v)\|^2=\int_{\Omega(t)} [|\nabla w|^2+v^2]\,dx\equiv E_1(w, v)=\text{the energy of the state } (w, v),$$

or an equivalent norm,

$$\|(w, v)\|_\gamma^2=\int_{\Omega(t)} [|\nabla w|^2+\gamma w^2+v^2]\,dx, \qquad \gamma\geqq 0,$$

for any $(w, v)\in H_0^1(\Omega(t))\oplus H^0(\Omega(t))$.

We are now in a position to pose the *Exact Controllability Problem* (ECP). Let

$$\text{(CS)} \quad \begin{cases} \dfrac{\partial^2 w}{\partial t^2}(x, t)-\Delta w(x, t)=f(x, t), & (x, t)\in Q(T), \\ \begin{bmatrix} w(x, 0) \\ \dfrac{\partial w}{\partial t}(x,0) \end{bmatrix}=\begin{bmatrix} w_0(x) \\ v_0(x) \end{bmatrix} \in H_0^1(\Omega(0))\oplus H^0(\Omega(0)), \\ w|_{\Sigma(T)}=0 \end{cases}$$

be a given distributed parameter control system governed by the wave equation. For any initial state $(w_0, v_0) \in H_0^1(\Omega(0)) \oplus H^0(\Omega(0))$ and any preassigned state $(w_T, v_T) \in H_0^1(\Omega(T)) \oplus H^0(\Omega(T))$, find an admissible control $f \in L^2(Q(T))$ such that the solution $w(x, t)$ satisfies the preassigned terminal state (w_T, v_T) at $t = T$.

The theorem of existence, uniqueness and continuity of solutions is given below.

THEOREM 1.1. *Consider the equations,*

$$\begin{cases} \dfrac{\partial^2 w}{\partial t^2}(x, t) - \Delta w(x, t) = f(x, t), & (x, t) \in Q(T), \quad f \in L^2(Q(T)) \\ w(x, 0) = w_0(x) & \in H_0^1(\Omega(0)), \\ \dfrac{\partial w}{\partial t}(x, 0) = v_0(x) & \in H^0(\Omega(0)), \\ w|_{\Sigma(T)} = 0, \end{cases}$$

in $Q(T)$. *Under the assumption* (H_0), *the solution* $w(x, t)$ *exists and is unique, such that*

$$\left(w, \frac{\partial w}{\partial t}\right) \in C^0([0, T]; H_0^1(\Omega(t)) \oplus H^0(\Omega(t))).$$

Furthermore, if $(w_0, v_0) \in [H_0^1(\Omega(0)) \cap H^2(\Omega(0))] \oplus H_0^1(\Omega(0))$, *then*

$$\left(w, \frac{\partial w}{\partial t}\right) \in C^1([0, T]; H_0^1(\Omega(t)) \oplus H^0(\Omega(t))) \cap C^0([0, T]; [H_0^1(\Omega(t)) \cap H^2(\Omega(t))] \oplus H_0^1(\Omega(t))).$$

The proof can be done by a local change of coordinates [4] and continuation by a priori estimates, and then use of theorems in [8, Chapt. 3] to prove continuity.

2. Growth and decay estimates for the wave equation in an expanding domain. In the sequel, we will need the following assumptions from time to time.

(H_1) The space is not two-dimensional; i.e., $N \neq 2$.

(H_2) The domain $\Omega(0)$ rests still until $t = T_0$ for some $T_0 > 0$; i.e.,

$$\Omega(t) \equiv \Omega(0) \quad \text{for } 0 \leqq t \leqq T_0.$$

(H_3) For $t \geqq T_0$, the domain is expanding; i.e.,

$$\Omega(t_1) \subseteq \Omega(t_2) \quad \text{for } T_0 \leqq t_1 \leqq t_2,$$

and for any $x \in \Omega(t)$,

$$r = |x| \leqq \theta t, \qquad 0 < \theta < 1,$$

for some θ.

In the proofs that follow, it is not difficult to see that (H_2) can actually be relaxed to the following:

(H_2') There exists $T_0 > 0$ such that for the domain $\Omega(t)$ undergoing boundary deformations during $0 \leqq t \leqq T$, the condition

$$K_1 E_1(w_0, v_0) \leqq E_1(w(\cdot, T_0), v(\cdot, T_0)) \leqq K_2 E_1(w_0, v_0)$$

is satisfied for some $K_1, K_2 > 0$, uniformly for all $(w_0, v_0) \in H_0^1(\Omega(0)) \oplus H^0(\Omega(0))$, where $(w(\cdot, t), v(\cdot, t))$ is the solution to (WE) in Theorem 2.1.

Our first theorem, which is independent of the preceding hypotheses (H_1)–(H_3), indicates that the energy of the wave grows and decays as the boundary shrinks and expands.

THEOREM 2.1. *Let*

$$\text{(WE)} \qquad \begin{cases} \dfrac{\partial^2 w}{\partial t^2} - \Delta w = 0 \quad \text{in } Q(T), \\ w(x,0) = w_0(x) \quad \in H_0^1(\Omega(0)), \\ \dfrac{\partial w}{\partial t}(x,0) = v_0(x) \quad \in H^0(\Omega(0)), \\ w|_{\Sigma(T)} = 0, \end{cases}$$

be a wave equation in $Q(T)$. Then the energy $E_1(t)$ is nonincreasing if $\Omega(t)$ is expanding, and $E_1(t)$ is nondecreasing if $\Omega(t)$ is contracting.

Proof. Let the motion of a point on $\partial\Omega(t)$ be given by $x = x(t)$. Then $(dx/dt, 1)$ is tangent to $\Sigma(T)$ at $(x(t), t)$. Thus

$$\nu_x \cdot \frac{dx}{dt} + \nu_t = 0. \tag{2.1}$$

Since the boundary $\partial\Omega(t)$ of $\Omega(t)$ is expanding outward, the component of dx/dt in the direction of ν_x must be nonnegative; thus $\nu_x \cdot dx/dt$ is nonnegative. Hence, ν_t is nonpositive.

Now writing

$$0 = \frac{\partial w}{\partial t}\left(\frac{\partial^2 w}{\partial t^2} - \Delta w\right) = \frac{\partial}{\partial t}\left[\frac{1}{2}\left(\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2\right)\right] - \operatorname{div}\left(\frac{\partial w}{\partial t}\nabla w\right), \tag{2.2}$$

and integrating over $Q(t)$, $0 \leqq t \leqq T$, we obtain

$$\begin{aligned} &\frac{1}{2}\int_{\Omega(t)}\left(\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2\right) dx \\ &\quad = \frac{1}{2}\int_{\Omega(0)}\left(\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2\right) dx - \frac{1}{2}\int_{\Sigma(t)}\left(\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2\right)\nu_t \, d\sigma \\ &\qquad\qquad + \int_{\Sigma(t)} \frac{\partial w}{\partial t}\nabla w \cdot \nu_x \, d\sigma \end{aligned} \tag{2.3}$$

$$= \frac{1}{2}\int_{\Omega(0)}\left(\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2\right) dx + \frac{1}{2}\int_{\Sigma(t)}\left(\frac{\partial w}{\partial \nu}\right)^2 \nu_t(|\nu_x|^2 - \nu_t^2)\, d\sigma. \tag{2.4}$$

The second term is always nonpositive because ν_t is nonpositive and (H_0). Therefore $E_1(t)$ is nonincreasing.

On the other hand, if $\partial\Omega(t)$ is contracting inward, ν_t is nonpositive. The second term of (2.4) becomes nonnegative. Thus $E_1(t)$ is nondecreasing. □

In what follows, we will use C. S. Morawetz's energy invariants [9], [12] to study the growth and decay estimates for the wave equation. They are

$$E_2(t) \equiv \int_{\Omega(t)}\left[t\left(\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2\right) + 2r\frac{\partial w}{\partial r}\frac{\partial w}{\partial t} + (N-1)w\frac{\partial w}{\partial t}\right] dx,$$

$$E_3(t) \equiv \int_{\Omega(t)}\left[(r^2+t^2)\left(\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2\right) + 4tr\frac{\partial w}{\partial t}\frac{\partial w}{\partial r} + 2(N-1)tw\frac{\partial w}{\partial t} - (N-1)w^2\right].$$

LEMMA 2.2. *For any domain* $\Omega \subseteq \mathbb{R}^N$, $N \neq 2$, *we have*

$$\int_\Omega \left[\sum (N-1)\frac{x_i}{r^2}\frac{\partial w}{\partial x_i}w + \frac{(N-1)^2}{4}\frac{w^2}{r^2}\right] dx = -\frac{(N-1)(N-3)}{4}\int_\Omega \frac{w^2}{r^2}\,dx,$$

$$\int_\Omega (r^2+t^2)\left[\sum (N-1)\frac{x_i}{r^2}\frac{\partial w}{\partial x_i}w + \frac{(N-1)^2}{4}\frac{w^2}{r^2}\right] dx$$

$$= -(N-1)\int_\Omega w^2\,dx - \frac{(N-1)(N-3)}{4}\int_\Omega \frac{(r^2+t^2)}{r^2}w^2\,dx,$$

provided that $w = 0$ *on* $\partial\Omega$.

Proof. The computations are straightforward. One need only note that the singularity at $r = 0$ does not make any contribution for $N \geqq 3$. □

THEOREM 2.3. *Assume* (H_1)–(H_3). *Let* $w(x, t)$ *be the solution of* (WE) *in* $Q(T)$, $T \geqq T_0$.

(i) *If* $t\nu_t + r\nu_r = 0$ *on* $\partial\Omega(t)$ *for* $T_0 \leqq t \leqq T$, *then* $E_2(t)$ *is conserved during* $[T_0, T]$. *We have the energy decay*

$$E_1(t) \leqq \frac{(1+\theta)T_0}{1-\theta}\frac{1}{t}E_1(0), \qquad t > 0. \tag{2.5}$$

(ii) *If*

$$t\nu_t + r\nu_r \leqq 0 \quad \text{on } \partial\Omega(t), \qquad t \geqq T_0 > 0, \tag{H4}$$

holds, then $E_2(t)$ *is nonincreasing for* $t \geqq T_0$, *and* (2.5) *remains valid.*

(iii) *If* $t\nu_t + r\nu_r \geqq 0$ *on* $\partial\Omega(t)$ *for* $t \geqq T_0 \geqq 0$, *then* $E_2(t)$ *is nondecreasing during* $[T_0, T]$ *and*

$$E_1(t) \geqq \frac{(1+\theta)T_0}{(1-\theta)}\frac{1}{t}E_1(0), \qquad t \geqq T_0. \tag{2.6}$$

Equivalently, there exists $K > 0$ *independent of* (w_0, v_0) *such that*

$$E_1(t) \geqq \frac{1+\theta}{1-\theta}\frac{KT_0}{1+t}E_1(0), \qquad t \geqq 0. \tag{2.6'}$$

Proof. We know that

$$0 = 2\left(\frac{\partial^2 w}{\partial t^2} - \Delta w\right)\left(t\frac{\partial w}{\partial t} + r\frac{\partial w}{\partial r} + \frac{N-1}{2}w\right)$$

$$= \frac{\partial}{\partial t}\left[t\left(\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2\right) + 2r\frac{\partial w}{\partial r}\frac{\partial w}{\partial t} + (N-1)w\frac{\partial w}{\partial t}\right]$$

$$- \operatorname{div}\left[2t\frac{\partial w}{\partial t}\nabla w + 2(x\cdot\nabla w)\nabla w + \left(\frac{\partial w}{\partial t}\right)^2 x + (N-1)w\,\nabla w - |\nabla w|^2 x\right].$$

Integrating the above over $Q(T_0, t)$, we obtain

$$E_2(t) = E_2(T_0) + \int_{\Sigma(T_0,t)} \left\{ -\left[t\left(\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2\right) + 2r\frac{\partial w}{\partial r}\frac{\partial w}{\partial t} + (N-1)w\frac{\partial w}{\partial t}\right]\nu_t \right.$$

$$+ \left[2t\frac{\partial w}{\partial t}\nabla w + 2(x\cdot\nabla w)\nabla w + \left(\frac{\partial w}{\partial t}\right)^2 x \right.$$

$$\left.\left. + (N-1)w\nabla w - |\nabla w|^2 x\right]\cdot\nu_x\right\} d\sigma. \tag{2.7}$$

Using the relations (1.1), we can simplify the above boundary integral to

$$\int_{\Sigma(T_0,t)} \left(\frac{\partial w}{\partial \nu}\right)^2 (|\nu_x|^2 - \nu_t^2)[t\nu_t + r\nu_r]\, d\sigma. \tag{2.8}$$

From here one easily sees that the conservation of $E_2(t)$ is proved. Now, define

$$\lambda_i \equiv \frac{\partial w}{\partial x_i} + \frac{N-1}{2}\frac{x_i}{r^2} w. \tag{2.9}$$

Then,

$$E_2(t) = \int_{\Omega(t)} \left\{ t\left[\left(\frac{\partial w}{\partial t}\right)^2 + \sum_i \left(\lambda_i - \frac{N-1}{2}\frac{x_i}{r^2} w\right)^2\right] + 2r\frac{\partial w}{\partial t}\frac{\partial w}{\partial r} + (N-1)w\frac{\partial w}{\partial t}\right\} dx$$

$\Rightarrow$ (integration by parts once and simplification, using Lemma 2.2)

$$\Rightarrow = \int_{\Omega(t)} \left[t\left(\left(\frac{\partial w}{\partial t}\right)^2 + \sum \lambda_i^2\right) + 2\frac{\partial w}{\partial t}\sum x_i\lambda_i\right] dx + \frac{t(N-1)(N-3)}{4}\int_{\Omega(t)} \frac{w^2}{r^2}\, dx.$$

By (H_3), $r \leqq \theta t$ on $\Omega(t)$, we have

$$\left|2\frac{\partial w}{\partial t}\sum x_i\lambda_i\right| \leqq 2\left|\frac{\partial w}{\partial t}\right| r\left(\sum \lambda_i^2\right)^{1/2} \leqq r\left(\left(\frac{\partial w}{\partial t}\right)^2 + \sum \lambda_i^2\right)$$

$$\leqq \theta t\left(\frac{\partial w^2}{\partial t} + \sum \lambda_i^2\right).$$

Therefore,

$$\begin{aligned} &(1+\theta)t\int_{\Omega(t)} \left(\left(\frac{\partial w}{\partial t}\right)^2 + \sum \lambda_i^2\right) dx + \frac{t(N-1)(N-3)}{4}\int_{\Omega(t)} \frac{w^2}{r^2}\, dx \\ &\geqq E_2(t) \geqq (1-\theta)t \int_{\Omega(t)} \left(\frac{\partial w^2}{\partial t} + \sum \lambda_i^2\right) dx + \frac{t(N-1)(N-3)}{4}\int_{\Omega(t)} \frac{w^2}{r^2}\, dx. \end{aligned} \tag{2.10}$$

Substituting (2.9) back, using Lemma 2.2 and simplifying, we get

$$\begin{aligned} &(1+\theta)t\int_{\Omega(t)} \left(\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2\right) dx - \frac{(N-1)(N-3)}{4}\theta t\int_{\Omega(t)} \frac{w^2}{r^2}\, dx \\ &\geqq E_2(t) \geqq (1-\theta)t\int_{\Omega(t)} \left(\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2\right) dx + \frac{(N-1)(N-3)}{4}\theta t\int_{\Omega(t)} \frac{w^2}{r^2}\, dx. \end{aligned} \tag{2.11}$$

Combining the above inequalities with (2.7) and (2.8), we conclude that, for $t \geqq T_0$, if $t\nu_t + r\nu_r \leqq 0$ on $\Sigma(T_0, t)$, then $E_2(t) \leqq E_2(T_0)$, so

$$(1+\theta)T_0\int_{\Omega(T_0)} \left(\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2\right) dx \geqq (1-\theta)t\int_{\Omega(t)} \left(\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2\right) dx.$$

Hence (2.5) is proved because $E_1(T_0) = E_1(0)$.

If $t\nu_t + r\nu_r \geqq 0$ on $\Sigma(T_0, t)$, then $E_2(t) \geqq E_2(T_0)$; thus

$$(1-\theta)T_0\int_{\Omega(T_0)} \left(\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2\right) dx \leqq (1+\theta)t\int_{\Omega(t)} \left(\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2\right) dx.$$

So (2.6) is proved. □

Remark. The assumption (H_4) is similar to the "pulse illumination" condition of Cooper and Strauss [6]. For example, after T_0 if $x \in \partial\Omega(t)$ moves according to the law $|x| = \theta t$, $0 < \theta < 1$, then (H_4) reduces to $\nu_t + \theta\nu_r \leqq 0$. An example illustrating various situations in Theorem 2.3 can be found in § 4.

As for $E_3(t)$, we have the following theorem.

THEOREM 2.4. *Assume* (H_0)–(H_3). *Let* $w(x, t)$ *be the solution of* (WE) *in* $Q(T)$, $T \geqq T_0$. *If for some* T_1, $0 < T_0 \leqq T_1 \leqq T$, *the condition* (*compare* [5, *p.* 141])

$$(H5) \qquad (r^2 + t^2)\nu_t + 2tr\nu_r \geqq 0, \qquad x \in \partial\Omega(t), \quad t \geqq T_1$$

is satisfied, then $E_3(t)$ *is increasing during* $[T_1, T]$ *and*

$$E_1(t) \geqq \frac{1}{1+\theta^2}\left(\frac{1-\theta}{1+\theta}\right)^2 \frac{T_1^2}{t^2} E_1(0), \qquad t \geqq T_1 \tag{2.12}$$

or, equivalently,

$$E_1(t) \geqq \frac{1}{1+\theta^2}\left(\frac{1-\theta}{1+\theta}\right)^2 \frac{KT_1^2}{(1+t)^2} E_1(0), \qquad t > 0, \quad \text{for some } K > 0. \tag{2.12'}$$

Remarks.

(i) (H_5) is satisfied provided that $t\nu_t + r\nu_r \geqq 0$ is satisfied for $t \geqq T_1$, because we have, with (H_3),

$$\begin{aligned}(r^2 + t^2)\nu_t + 2tr\nu_r &\geqq (1+\theta^2)t^2\nu_t + 2tr\nu_r \\ &\geqq 2t^2\nu_t + 2tr\nu_r = 2t(t\nu_t + r\nu_r) \geqq 0.\end{aligned}$$

(ii) The estimate (2.12) says that under (H_5), $E_1(t)$ cannot decay with a rate faster than $1/t^2$. This estimate does not seem to be too useful from the energy decay point of view, because we already know that $E_1(t) \leqq E_1(0)$. However, such an estimate is very important in the controllability study in § 3. Compare (3.5).

(iii) Assuming both (H_4) and (H_5), we know from Theorems 2.3 and 2.4 that the energy of the wave equation decays with a rate between $1/t$ and $1/t^2$.

(iv) For a sphere expanding with a uniform speed $\theta < 1$, it is impossible to have

$$(r^2 + t^2)\nu_t + 2tr\nu_r \leqq 0 \quad \text{on } \partial\Omega(t),$$

for all $t \geqq T_1$. Therefore, one in general cannot expect a result like (2.5) from E_3.

Proof. We follow the same line of argument as in the proof of the preceding theorem. It is known that

$$\begin{aligned}0 &= 2\left(\frac{\partial^2 w}{\partial t^2} - \Delta w\right)\left[(r^2 + t^2)\frac{\partial w}{\partial t} + 2tr\frac{\partial w}{\partial r} + (N-1)tw\right] \\ &= \frac{\partial}{\partial t}\left[(r^2 + t^2)\left(\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2\right) + 4tr\frac{\partial w}{\partial t}\frac{\partial w}{\partial r} + 2(N-1)tw\frac{\partial w}{\partial t} - (N-1)w^2\right] \\ &\quad - 2\,\mathrm{div}\left[(r^2 + t^2)\frac{\partial w}{\partial t}\nabla w + t\left(\frac{\partial w}{\partial t}\right)^2 x + 2t(x \cdot \nabla w)\nabla w - t|\nabla w|^2 x + (N-1)tw\nabla w\right].\end{aligned}$$

Integrating over $Q(T_1, t)$, we obtain

$$E_3(t) = E_3(T_1) + \int_{\Sigma(T_1,\,t)} \Big\{ 2\Big[(r^2+t^2)\frac{\partial w}{\partial t}\nabla w + t\Big(\frac{\partial w}{\partial t}\Big)^2 x + 2t(x\cdot\nabla w)\nabla w - t|\nabla w|^2 x + (N-1)tw\nabla w\Big]\cdot \nu_x - \Big[(r^2+t^2)\Big(\Big(\frac{\partial w}{\partial t}\Big)^2 + |\nabla w|^2\Big) + 4tr\frac{\partial w}{\partial t}\frac{\partial w}{\partial r} + 2(N-1)tw\frac{\partial w}{\partial t} - (N-1)w^2\Big]\nu_t\Big\}\, d\sigma. \tag{2.13}$$

Using (1.1), again we simplify the above boundary integral and get

$$E_3(t) = E_3(T_1) + \int_{\Sigma(T_1,t)} \frac{\partial w^2}{\partial \nu}(|\nu_x|^2 - \nu_t^2)[(r^2+t^2)\nu_t + 2tr\nu_r]\, d\sigma. \tag{2.14}$$

Now we use (2.9) and integrate by parts; we get

$$E_3(t) = \int_{\Omega(t)} \Big\{ (r^2+t^2)\Big[\frac{\partial w^2}{\partial t} + \sum \lambda_i^2\Big] + 4t\frac{\partial w}{\partial t}\sum x_i\lambda_i \Big\}\, dx + \frac{(N-1)(N-3)}{4}\int_{\Omega(t)} \frac{r^2+t^2}{r^2} w^2\, dx.$$

For $t \geqq T_0$, $r \leqq \theta t$ on $\Omega(t)$, so

$$2tr \leqq \frac{2\theta}{1+\theta^2}(r^2+t^2);$$

thus

$$4t\frac{\partial w}{\partial t}\sum x_i\lambda_i \leqq 4tr\frac{\partial w}{\partial t}(\textstyle\sum \lambda_i^2)^{1/2} \leqq 2tr\Big(\Big(\frac{\partial w}{\partial t}\Big)^2 + \sum \lambda_i^2\Big) \leqq \frac{2\theta}{1+\theta^2}(r^2+t^2)\Big(\Big(\frac{\partial w}{\partial t}\Big)^2 + \sum \lambda_i^2\Big).$$

Therefore

$$\Big(1+\frac{2\theta}{1+\theta^2}\Big)\int_{\Omega(t)} (r^2+t^2)\Big(\Big(\frac{\partial w}{\partial t}\Big)^2 + \sum \lambda_i^2\Big)\, dx + \frac{(N-1)(N-3)}{4}\int_{\Omega(t)} \frac{r^2+t^2}{r^2} w^2\, dx$$

$$\geqq E_3(t) \geqq \Big(1-\frac{2\theta}{1+\theta^2}\Big)\int_{\Omega(t)} (r^2+t^2)\Big(\Big(\frac{\partial w}{\partial t}\Big)^2 + \sum \lambda_i^2\Big)\, dx + \frac{(N-1)(N-3)}{4}\int_{\Omega(t)} \frac{r^2+t^2}{r^2} w^2\, dx$$

$$\geqq \frac{(1-\theta)^2}{1+\theta^2} t^2 \Big\{ \int_{\Omega(t)} \Big(\Big(\frac{\partial w}{\partial t}\Big)^2 + \sum \lambda_i^2\Big)\, dx + \frac{(N-1)(N-3)}{4}\int_{\Omega(t)} \frac{w^2}{r^2}\, dx\Big\}.$$

Substituting (2.9) back and using Lemma 2.2, we obtain

$$\frac{(1+\theta)^2}{1+\theta^2}\int_{\Omega(t)} (r^2+t^2)\Big(\Big(\frac{\partial w}{\partial t}\Big)^2 + |\nabla w|^2\Big)\, dx - (N-1)\frac{(1+\theta^2}{1+\theta^2}\int_{\Omega(t)} w^2\, dx - \frac{2\theta}{1+\theta^2}\frac{(N-1)(N-3)}{4}\int_{\Omega(t)} \frac{r^2+t^2}{r^2} w^2\, dx$$

$$\geqq E_3(t) \geqq \frac{(1-\theta)^2}{1+\theta^2} t^2 \int_{\Omega(t)} \Big(\Big(\frac{\partial w}{\partial t}\Big)^2 + |\nabla w|^2\Big)\, dx. \tag{2.15}$$

Now,

$$r^2 + t^2 \leqq (1+\theta^2)t^2, \tag{2.16}$$

and combining (H_5), (2.14), (2.15) and (2.16), we obtain

$$(1+\theta)^2 t^2 \int_{\Omega(t)} \left(\left(\frac{\partial w}{\partial t} \right)^2 + |\nabla w|^2 \right) dx \geqq E_3(t) \geqq E_3(T_1)$$

$$\geqq \frac{(1-\theta)^2}{1+\theta^2} T_1^2 \int_{\Omega(T_1)} \left(\left(\frac{\partial w}{\partial t} \right)^2 + |\nabla w|^2 \right) dx$$

$$= \frac{(1-\theta)^2}{1+\theta^2} T_1^2 \int_{\Omega(0)} \left(\left(\frac{\partial w}{\partial t} \right)^2 + |\nabla w|^2 \right) dx.$$

So (2.12) is proved. □

For the case $N = 2$, the function w^2/r^2 is not integrable in general due to the singularity at $x = 0$. In order to be able to derive some energy estimate, we must assume that each $\Omega(t)$ is of annular shape.

THEOREM 2.5. *Assume that for each $t \geqq 0$, $\Omega(t)$ is an annular region and its boundary $\partial\Omega(t)$ consists of two parts $\Gamma_0 \times \{t\}$ and Γ_t. The interior part Γ_0 is star-shaped with respect to the origin. Γ_0 is fixed, but the exterior part Γ_t may be moving. Assume* (H_0), (H_2), (H_3) and (H_4). *Then $E_2(t)$ is nonincreasing $(t \geqq T_0)$ and for $t \geqq T_0$,*

$$(2.17) \qquad E_1(t) - \frac{1}{4} \frac{\theta}{1-\theta} \int_{\Omega(t)} \frac{w^2}{r^2} dx \leqq \frac{T_0}{t} \left\{ \frac{1+\theta}{1-\theta} E_1(T_0) + \frac{1}{4} \frac{\theta}{1-\theta} \int_{\Omega(T_0)} \frac{w^2}{r^2} dx \right\}.$$

Proof. It is the same as that of Theorem 2.3 except for some modifications. We replace (2.7) and (2.8) by

$$E_2(t) = E_2(T_0) + \text{boundary integral over the exterior lateral surface of } Q(T_0, t)$$

$$+ 2(t - T_0) \int_{\Gamma_0} (x \cdot \nu_x) \left(\frac{\partial w}{\partial \nu_x} \right)^2 d\sigma_x.$$

The last additional term is always nonpositive. So $E_2(t)$ is nonincreasing. Now the proof of (2.17) follows immediately from (2.11). □

3. Stabilizability and exact controllability. Our first theorem is a stabilizability theorem. Here we obtain "forward" decay estimates by using high damping and compensation. It is an analogue of [2, Thm. 4.1].

THEOREM 3.1. *Let $w(x, t)$ be the solution of*

$$(E_+) \quad \begin{cases} \dfrac{\partial^2 w}{\partial t^2}(x, 0) + 2\gamma_1 \dfrac{\partial w}{\partial t}(x, t) - \Delta w(x, t) + \gamma_2 w(x, t) = 0, & (x, t) \in Q(T), \\ w(x, 0) = w_0(x) \quad \in H_0^1(\Omega(0)), \\ \dfrac{\partial w}{\partial t}(x, 0) = v_0(x) \quad \in H^0(\Omega(0)), \\ w(x, t)|_{\Sigma(t)} = 0. \end{cases}$$

Assume that the domain is expanding; i.e., $\Omega(t_1) \subseteq \Omega(t_2)$ for $0 \leqq t_1 \leqq t_2 \leqq T$. For any $\varepsilon > 0$, if $\gamma_1 \equiv \lambda \equiv 1/2\varepsilon$, $\gamma_2 \equiv 1/\varepsilon^3$, then

$$\int_{\Omega(t)} \left[\left(\frac{\partial w}{\partial t} \right)^2 + |\nabla w|^2 + \gamma_2 w^2 \right] dx \leqq \frac{1+\varepsilon}{1-\varepsilon+2\lambda T} \int_{\Omega(0)} [v_0^2 + |\nabla w_0|^2 + \gamma_2 w_0^2] dx.$$

Proof. We use $\partial w/\partial t$ as multiplier and obtain

$$
\begin{aligned}
0 &= \frac{\partial w}{\partial t}\left(\frac{\partial^2 w}{\partial t^2} - \Delta w + 2\gamma_1 \frac{\partial w}{\partial t} + \gamma_2 w\right) \\
&= \frac{\partial}{\partial t}\left[\frac{1}{2}\left(\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2 + \gamma_2 w^2\right)\right] - \operatorname{div}\left(\frac{\partial w}{\partial t}\nabla w\right) + 2\gamma_1 \frac{\partial w^2}{\partial t}.
\end{aligned} \tag{3.1}
$$

Integrating over $Q(t)$, we obtain

$$
\begin{aligned}
\frac{1}{2}\int_{\Omega(t)} \left(\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2 + \gamma_2 w^2\right) dx &= \frac{1}{2}\int_{\Omega(0)} (v_0^2 + |\nabla w_0|^2 + \gamma_2 w_0^2)\, dx \\
&+ \int_{\Sigma(t)} \left[\left(\frac{\partial w}{\partial t}\nabla w\right)\cdot \nu_x - \frac{1}{2}\left(\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2 + \gamma_2 w^2\right)\nu_t\right] d\sigma - 2\gamma_1 \iint_{Q(t)} \left(\frac{\partial w}{\partial t}\right)^2 dx\, dt.
\end{aligned}
$$

The boundary integral is the same as that in (2.2); therefore it is nonpositive. The third term on the right is always nonpositive. Therefore we conclude that

$$
\int_{\Omega(t)} \left(\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2 + \gamma_2 w^2\right) dx \tag{3.2}
$$

is nonincreasing as a function of t.

Now, using λw as multiplier, we get

$$
\begin{aligned}
0 &= \lambda w\left(\frac{\partial^2 w}{\partial t^2} - \Delta w + 2\gamma_1 \frac{\partial w}{\partial t} + \gamma_2 w\right) \\
&= \lambda\left\{\frac{\partial}{\partial t}\left[w\frac{\partial w}{\partial t} + \gamma_2 w^2\right] - \operatorname{div}(w\,\nabla w) + |\nabla w|^2 - \frac{\partial w^2}{\partial t} + \gamma_2 w^2\right\}
\end{aligned} \tag{3.3}
$$

$\Rightarrow$ ((3.1) + (3.3) and integration over $Q(T)$)

$$
\begin{aligned}
&\int_{\Omega(T)} \left[\frac{1}{2}\left(\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2\right) + \lambda w\frac{\partial w}{\partial t} + \left(\frac{\gamma_2}{2} + \lambda\gamma_1\right)w^2\right] dx \\
&= \int_{\Omega(0)} \left[\frac{1}{2}\left(\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2\right) + \lambda w\frac{\partial w}{\partial t} + \left(\frac{\gamma_2}{2} + \lambda\gamma_1\right)w^2\right] dx \\
&\quad + \int_{\Omega(T)} (\text{negative term})\, d\sigma - \int\int_{Q(T)} \left[\lambda|\nabla w|^2 + (2\gamma_1 - \lambda)\left(\frac{\partial w}{\partial t}\right)^2 + \lambda\gamma_2 w^2\right] dx\, dt.
\end{aligned}
$$

Hence, for every $\varepsilon > 0$, we have

$$
\begin{aligned}
&\int_0^T \int_{\Omega(t)} \left[\lambda|\nabla w|^2 + (2\gamma_1 - \lambda)\left(\frac{\partial w}{\partial t}\right)^2 + \lambda\gamma_2 w^2\right] dx\, dt \\
&\quad + \int_{\Omega(T)} \left\{\frac{1}{2}\left[(1-\varepsilon)\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2\right] + \left[\frac{\gamma_2}{2} + \lambda\left(\gamma_1 - \frac{1}{2\varepsilon}\right)\right]w^2\right\} dx \\
&\quad \leqq \int_{\Omega(0)} \left\{\frac{1}{2}\left[(1+\varepsilon)\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2\right] + \left[\frac{\gamma_2}{2} + \lambda\left(\gamma_1 + \frac{1}{2\varepsilon}\right)\right]w^2\right\} dx.
\end{aligned}
$$

We choose $\lambda = \gamma_1 = 1/2\varepsilon$, $\gamma_2 = 1/\varepsilon^3$ and use the decreasing property of (3.2) to get

$$\begin{aligned}\left(\lambda T + \frac{1-\varepsilon}{2}\right)\int_{\Omega(T)}\left[\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2 + \gamma_2 w^2\right] dx \\ \leqq \frac{1+\varepsilon}{2}\int_{\Omega(0)}\left[\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2 + \gamma_2 w^2\right] dx.\end{aligned} \tag{3.4}$$

So the proof is complete. □

When we reverse the sense of time, an expanding domain becomes a contracting one. Therefore by Theorem 2.1 we cannot expect "backward" decay as we had in [2], [3], [10]. Instead, we use a scheme of high compensation in keeping with the compensation we made in Theorem 3.1. Under the assumptions in § 2, we derive the following backward growth estimate.

LEMMA 3.2. *Assume that* (H_0)–(H_3) *and* (H_5) *hold. Let* $w(x, t)$ *be the solution of the backward equation,*

$$\text{(E}_-\text{)} \qquad \begin{cases} \dfrac{\partial^2 w}{\partial t^2}(x,t) - \Delta w(x,t) + \gamma w(x,t) = 0, & (x,t) \in Q(T), \quad \gamma \geqq 0, \\ w(x,T) = w_T(x) & \in H_0^1(\Omega(T)), \\ \dfrac{\partial w}{\partial t}(x,T) = v_T(x) & \in H^0(\Omega(T)), \\ w|_{\Sigma(T)} = 0. \end{cases}$$

Then the following inequality holds:

$$\begin{aligned}\int_{\Omega(T)}\left[\left(\frac{\partial w}{\partial t}\right)^2(x,T) + |\nabla w(x,T)|^2 + \gamma w^2(x,T)\right] dx \\ \geqq \frac{1}{1+\theta^2}\left(\frac{1-\theta}{1+\theta}\right)^2\frac{T_1^2}{T^2}\int_{\Omega(0)}\left[\left(\frac{\partial w}{\partial t}\right)^2(x,0) + |\nabla w(x,0)|^2 + \gamma w^2(x,0)\right] dx.\end{aligned} \tag{3.5}$$

Proof. It is straightforward to verify that

$$\begin{aligned}0 &= 2\left(\frac{\partial^2 w}{\partial t^2} - \Delta w + \gamma w\right)\left[(r^2+t^2)\frac{\partial w}{\partial t} + 2tr\frac{\partial w}{\partial r} + 2(N-1)tw\right] \\ &= \frac{\partial}{\partial t}\left[(r^2+t^2)\left(\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2 + \gamma w^2\right) + 4tr\frac{\partial w}{\partial t}\frac{\partial w}{\partial r} + 2(N-1)tw\frac{\partial w}{\partial t} - (N-1)\dot{w}^2\right] \\ &\quad - 2\,\mathrm{div}\left[(r^2+t^2)\frac{\partial w}{\partial t}\nabla w + \left(\frac{\partial w}{\partial t}\right)^2 x + 2t(x\cdot\nabla w)w - t|\nabla w|^2 x + (N-1)tw\nabla w \right. \\ &\qquad \left. - \gamma t w^2 x\right] - 4\gamma t w^2.\end{aligned}$$

Integrating over $Q(T_1, T)$, we obtain

$$\begin{aligned}&\int_{\Omega(T)}\left[(r^2+T^2)\left(\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2 + \gamma w^2\right) + 4tr\frac{\partial w}{\partial t}\frac{\partial w}{\partial r} + 2(N-1)Tw\frac{\partial w}{\partial t} - (N-1)w^2\right] dx \\ &\quad = \int_{\Omega(T_1)}\left[(r^2+T_1^2)\left(\left(\frac{\partial w}{\partial t}\right)^2 + |\nabla w|^2 + \gamma w^2\right)\right. \\ &\qquad \left. + 4T_1 r\frac{\partial w}{\partial t}\frac{\partial w}{\partial r} + 2(N-1)T_1 w\frac{\partial w}{\partial t} - (N-1)w^2\right] dx \\ &\qquad + \text{integral over the boundary } \Sigma(T_1, T) + \iint_{Q(T_1,T)} 4\gamma t w^2\, dx\, dt.\end{aligned} \tag{3.6}$$

The above boundary integral is the same as the one in (2.14) with $t = T$, so it is nonnegative by (H_5). The very last integral over $Q(T_1, T)$ is always nonnegative.

Again using λ_i as defined in (2.9), with (H_3), we obtain

$$\left(1+\frac{2\theta}{1+\theta^2}\right)\int_{\Omega(T)}(r^2+T^2)\left(\left(\frac{\partial w}{\partial t}\right)^2+\sum\lambda_i^2+\gamma w^2\right)dx+\frac{(N-1)(N-3)}{4}\int_{\Omega(T)}\frac{r^2+T^2}{r^2}w^2\,dx$$

$$\geqq\left(1-\frac{2\theta}{1+\theta^2}\right)\int_{\Omega(T_1)}(r^2+T_1^2)\left(\left(\frac{\partial w}{\partial t}\right)^2+\sum\lambda_i^2-\gamma w^2\right)dx$$

$$+\frac{(N-1)(N-3)}{4}\int_{Q(T_1)}\frac{r^2+T_1^2}{r^2}w^2\,dx$$

$$\geqq\frac{(1-\theta)^2}{1+\theta^2}T_1^2\left[\int_{\Omega(T_1)}\left(\left(\frac{\partial w}{\partial t}\right)^2+\sum\lambda_i^2+w^2\right)dx+\frac{(N-1)(N-3)}{4}\int_{\Omega(T_1)}\frac{w^2}{r^2}\,dx\right].$$

The rest of the proof follows from the same type of argument as in Theorem 2.4 and will not be reproduced here. □

LEMMA 3.3. *Under the same assumptions as Theorem* 3.1 *and Lemma* 3.2, *respectively,*

(i) *Let* $w(x, t)$ *be the solution of* (E_+) *with initial state* (w_0, v_0). *Let* $\Phi_+(t)$: $H_0^1(\Omega(0))\oplus H^0(\Omega(0))\to H_0^1(\Omega(t))\oplus H^0(\Omega(t))$ *be the linear transformation defined by*

$$\Phi_+(t)((w_0, v_0))\equiv\left(w(\cdot\,, t), \frac{\partial w}{\partial t}(\cdot\,, t)\right).$$

Then $\Phi_+(t)$ *is continuous for each* $t\in[0, T]$.

(ii) *Let* $w(x, t)$ *be the solution of* (E_-) *with terminal state* (w_T, v_T). *Define* $\Phi_-(t)$: $H_0^1(\Omega(T))\oplus H^0(\Omega(T))\to H_0^1(\Omega(t))\oplus H^0(\Omega(t))$ *to be the linear transformation*

$$\Phi_-(t)((w_T, v_T))\equiv\left(w(\cdot\,, t), \frac{\partial w}{\partial t}(\cdot\,, t)\right).$$

Then $\Phi_-(t)$ *is also continuous for each* $t\in[0, T]$.

Proof. We need only to prove (i), since (ii) will follow in a similar manner. Let $(w_0^1, v_0^1), (w_0^2, v_0^2)\in H_0^1(\Omega(0))\oplus H^0(\Omega(0))$. Let $w(x, t)$ be the solution of (E_+) with initial state $(w_0^1-w_0^2, v_0^1-v_0^2)$. From (3.4) with T replaced by t, we have

$$\left(\lambda t+\frac{1-\varepsilon}{2}\right)\int_{\Omega(t)}\left[\left(\frac{\partial w}{\partial t}\right)^2+|\nabla w|^2+\gamma_2 w^2\right]dx\leqq\frac{1+\varepsilon}{2}\int_{\Omega(0)}\left[\left(\frac{\partial w}{\partial t}\right)^2+|\nabla w|^2+\gamma_2 w^2\right]dx.$$

Thus,

$$\left(\lambda t+\frac{1-\varepsilon}{2}\right)\|\Phi_+(t)(w_0^1-w_0^2, v_0^1-v_0^2)\|_{\gamma_2}^2\leqq\frac{1+\varepsilon}{2}\|(w_0^1-w_0^2, v_0^1-v_0^2)\|_{\gamma_2}^2.$$

The continuity of $\Phi_+(t)$ is clear. □

THEOREM 3.4. (Exact controllability of the wave equation in an expanding domain). *Assume* (H_0)–(H_3) *and* (H_5). *Assume furthermore that* $T_1 = T_0$ *in* (H_5). *Let* $T > 0$. *For any given initial state* $(w_0, v_0)\in H_0^1(\Omega(0))\oplus H^0(\Omega(0))$ *and any prescribed final state* $(w_1, v_1)\in H_0^1(\Omega(T))\oplus H^0(\Omega(T))$, *there is a control* $f\in L^2(Q(T))$ *which solves the* (ECP).

Proof. By ([2, Thm. 4.2]), we need only consider the case $T\geqq T_0$. We first consider the case of "controllability to zero", namely, $(w_1, v_1) = (0, 0)$. Let $\tilde{w}(x, t)$ be the solution

of

$$\begin{cases} \dfrac{\partial^2 \tilde{w}}{\partial t^2}(x,t) - \Delta \tilde{w}(x,t) = -2\gamma_1 \dfrac{\partial \tilde{w}}{\partial t}(x,t) - \gamma_2 \tilde{w}(x,t), & (x,t) \in Q(T), \\ (\tilde{w}(x,0), \tilde{v}(x,0)) = (p(x), q(x)) \in H_0^1(\Omega(0)) \oplus H^0(\Omega(0)), & \gamma_1 > 0, \quad \gamma_2 > 0. \end{cases}$$

The terminal state of $(\tilde{w}(x,t), \tilde{v}(x,t))$ at $t = T$ is $(\tilde{w}(x,T), \tilde{v}(x,T))$, which is in $H_0^1(\Omega(T)) \oplus H^0(\Omega(T))$. Let $\bar{w}(x,t)$ denote the solution of

$$\begin{cases} \dfrac{\partial^2 \bar{w}}{\partial t^2}(x,t) - \Delta \bar{w}(x,t) = -\gamma_2 \bar{w}(x,t), & (x,t) \in Q(T), \\ (\bar{w}(x,T), \bar{v}(x,T)) = -(\tilde{w}(x,T), \tilde{v}(x,T)) \in H_0^1(\Omega(T)) \oplus H^0(\Omega(T)). \end{cases}$$

Define

$$w(x,t) \equiv \tilde{w}(x,t) + \bar{w}(x,t), \qquad (x,t) \in Q(T),$$

$$f(x,t) \equiv -2\gamma_1 \frac{\partial \tilde{w}}{\partial t} - \gamma_2 [\tilde{w}(x,t) + \bar{w}(x,t)].$$

Then $w(x,t)$ satisfies the equation

$$\frac{\partial^2 w}{\partial t^2}(x,t) - \Delta w(x,t) = f(x,t), \qquad (x,t) \in Q(T),$$

with the terminal state $(w(x,T), v(x,T)) = (0,0)$. The initial state is

$$\begin{aligned} (w(x,0), v(x,0)) &= (\tilde{w}(x,0), \tilde{v}(x,0)) + (\bar{w}(x,0), \bar{v}(x,0)) \\ &= (p(x), q(x)) + \Phi_-(0)\Phi_+(T)[-(p(x), q(x))] \\ &= [I - \Phi_-(0)\Phi_+(T)](p(x), q(x)). \end{aligned} \tag{3.7}$$

Now, choose $\gamma_1 = \lambda = 1/2\varepsilon$, $\gamma_2 = 1/\varepsilon^3$. By Theorem 3.1 and Lemma 3.2, we deduce that

$$\|\Phi_-(0)\Phi_+(T)\|^2 \leqq (1+\theta^2)\left(\frac{1+\theta}{1-\theta}\right)^2 \frac{T^2}{T_0^2} \frac{1+\varepsilon}{1-\varepsilon+2\lambda T}. \tag{3.8}$$

Here the operator norm is relative to $\|\cdot\|_\gamma$ on $H_0^1(\Omega(0)) \oplus H^0(\Omega)$ and $H_0^1(\Omega(T)) \oplus H^0(\Omega(T))$. We choose ε so small that the right-hand side of (3.6) is smaller than 1. Therefore $I - \Phi_-(0)\Phi_+(T)$ is an invertible linear transformation from $H_0^1(\Omega(0)) \oplus H^0(\Omega(0))$ into itself. We choose

$$(p,q) = [I - \Phi_-(0)\Phi_+(T)]^{-1}(w_0, v_0).$$

Then f steers the system from (w_0, v_0) to $(0,0)$ at $t = T$.

For an arbitrarily prescribed final state $(w_1, v_1) \in H_0^1(\Omega(T)) \oplus H^0(\Omega(T))$, we first let $w^-(x,t)$ be the solution of

$$\frac{\partial^2 w^-}{\partial t^2}(x,t) - \Delta w^-(x,t) = 0, \qquad (x,t) \in Q(T),$$

$$\begin{bmatrix} w^-(x,T) \\ v^-(x,T) \end{bmatrix} = \begin{bmatrix} w_1(x) \\ v_1(x) \end{bmatrix}, \qquad \text{(terminal condition)},$$

and next let $w^+(x, t)$ be the solution of

$$\frac{\partial^2 w^+}{\partial t^2}(x, t) - \Delta w^+(x, t) = f(x, t), \qquad (x, t) \in Q(T),$$

$$\begin{bmatrix} w^+(x, 0) \\ v^+(x, 0) \end{bmatrix} = \begin{bmatrix} w_0 - w^-(x, 0) \\ v_0 - v^-(x, 0) \end{bmatrix},$$

where f is a control which steers the system from $(w_0 - w^-(x, 0), v_0 - v^-(x, 0))$ to $(0, 0)$ at $t = T$. Then $w \equiv w^+ + w^-$ satisfies the equation

$$\frac{\partial^2 w}{\partial t^2}(x, t) - \Delta w(x, t) = f(x, t),$$

with f steering the system from (w_0, v_0) to (w_1, v_1). □

Remark. The theorem above slightly generalizes Russell's "controllability via stabilizability" principle [10], in the sense that we do not need to have both forward and backward decay. What is most important is to have $\|\Phi_-(0)\Phi_+(T)\| < 1$, thereby ensuring the invertibility of $I - \Phi_-(0)\Phi_+(T)$.

In concluding this section, we would like to quote the following comment from the referee. In reading [1], [2], and the present paper, one gets the impression that uniform stabilizability and exact distributed controllability are opposite sides of the same coin. While the stabilizability results certainly lead to controllability results, the latter can be obtained independently of any energy estimates in a very simple way. For example, the results of Theorem 3.4 may be obtained as follows: assume only (H_0) (this is a much weaker hypothesis than appears in Theorem 3.4), and let $T > 0$. Let u be the unique solution to (WE) guaranteed by Theorem 2.1. Let $\alpha_T \in C^\infty(\mathbb{R})$ such that $\alpha_T(0) = 1$, $\alpha_T'(0) = \alpha_T(T) = \alpha_T'(T) = 0$. Set $w(x, t) = \alpha(t)u(x, t)$ and $f \equiv \alpha_T'' u + 2\alpha_T' u_t$. Then $f \in L^2(Q(T))$ is an admissible control, w satisfies (CS) and $w(x, T) = v(x, T) = 0$ in $\Omega(T)$. One can even choose a control satisfying, e.g., $\|f\|_{L^2(Q(T))} \leqq C$, where $C > 0$ is given a priori, by choosing α_T appropriately. Of course in this case $T > 0$ cannot be arbitrary but must be greater than some positive number depending on C and u. This little trick can obviously be extended to many other distributed control systems. The main advantage of the control schemes in § 3 is that they can be more easily implemented as feedback controls.

4. Example. Let $\Omega(0)$ be a sphere with radius r_0 in $\mathbb{R}^N$, $N \neq 2$. Let $Q(T) = \bigcup_{0 \leqq t \leqq T} \Omega(t) \times \{t\}$, where

$$\Omega(t) = \begin{cases} \Omega(0), & 0 \leqq t \leqq T_0, \\ \{z \in \mathbb{R}^N \mid z = [|x| + \theta(t - T_0)]x/|x|,\ x \neq 0,\ x \in \Omega(0),\ z = 0 \text{ if } x = 0\}, \end{cases}$$

$$t \geqq T_0,\ 0 < \theta < 1.$$

In other words, after time T_0, the boundary of $\Omega(t)$ expands radially outward with a velocity $\theta < 1$. Let

$$\frac{\partial^2 w}{\partial t^2}(x, t) - \Delta_x w(x, t) = 0 \tag{4.1}$$

be the wave equation in $Q(T)$. One can, by making the global change of variable

$$y \equiv \frac{1}{1 + \theta(t - T_0)} x, \qquad t \geqq T_0,$$

$$u(y, t) \equiv w(x, t) = w([1 + \theta(t - T_0)]y, t),$$

derive a time-dependent hyperbolic equation

$$\frac{\partial^2}{\partial t^2}u(y,t)-\frac{2\theta}{1+\theta(t-T_0)}\sum_{i=1}^{N} y_i\frac{\partial^2}{\partial y_i\partial t}u(y,t)+\frac{\theta^2}{[1+\theta(t-T_0)]^2}$$
$$\times\left[2\sum_i y_i\frac{\partial}{\partial y_i}u(y,t)+\sum_{i,j} y_iy_j\frac{\partial^2}{\partial y_i\partial y_j}u(y,t)\right]-\frac{1}{[1+\theta(t-T_0)]^2}\Delta_y u(y,t)=0, \tag{4.2}$$

on the *fixed* domain $\Omega(0)$ for $t\geqq T_0$. It is easy to check that the symbol of the principal part of (4.2), which is

$$p(t,y,\xi)=\xi_0^2-\frac{2\theta}{1+\theta(t-T_0)}\sum_i y_i\xi_0\xi_i+\frac{\theta^2}{[1+\theta(t-T_0)]^2}\sum_{i,j} y_iy_j\xi_i\xi_j-\frac{1}{[1+\theta(t-T_0)]^2}\sum_i y_i^2,$$

always has two distinct real roots ξ_0 for $p(y,\xi)=0$ with any given $(t, y, \xi_1, \cdots, \xi_n)$. Thus the strict hyperbolicity is conserved.

On the boundary of the truncated cone $Q(T_0, T)$, let the motion of a point on $\partial\Omega(t)$ be given by $x(t)$. Then,

$$\frac{dx(t)}{dt}=\theta\frac{x}{|x|}=\theta\frac{\nu_x}{|\nu_x|}.$$

Since

$$\left(\frac{dx(t)}{dt},1\right)\cdot(\nu_x,\nu_t)=0 \quad \text{on } \Sigma(T_0,T),$$

we have

$$\left\{\begin{aligned}&\theta|\nu_x|+\nu_t=0,\\&|\nu_x|^2+\nu_t^2=1,\end{aligned}\right\}\Rightarrow \quad |\nu_x|=\frac{1}{\sqrt{1+\theta^2}}, \qquad t=\frac{\theta}{\sqrt{1+\theta^2}}.$$

Returning to Theorem 2.3, one easily verifies the following:

(i) If $T_0=r_0/\theta$, then $r=\theta t$ and $t\nu_t+r\nu_r=0$ on $\partial\Omega(t)$ for $t\geqq T_0$. Hence, after T_0, $E_2(t)$ is conserved. $E_1(t)$ decays with a rate $1/t$ and $E_3(t)$ grows with a rate t.

(ii) If $T_0>r_0/\theta$, then $r<\theta t$ and $t\nu_t+r\nu_r<0$ on $\partial\Omega(t)$ for $t\geqq T_0$. Hence, $E_2(t)$ is decreasing after T_0, $E_1(t)$ decays with a rate $1/t$, $E_3(t)$ is increasing after certain $T_1>0$.

(iii) If $T_0<r_0/\theta$, then $t\nu_t+r\nu_r>0$ on $\partial\Omega(t)$ for $t\geqq T_0$. Therefore, $E_2(t)$ is increasing after T_0.

Suppose T_0 satisfies the condition that

$$r_0\leqq T_0\leqq\frac{\theta}{1-\sqrt{1-\theta^2}}r_0; \tag{4.3}$$

then

$$\frac{1-\sqrt{1-\theta^2}}{\theta}t\leqq\theta t+(r_0-\theta T_0)=|x(t)|\leqq\theta t \quad \text{on } \partial\Omega(t), \quad t\geqq T_0.$$

Thus,

$$\frac{\nu_t}{\nu_r}+\frac{2tr}{r^2+t^2}=-\theta+\frac{2tr}{r^2+t^2}\geqq 0, \quad \text{for all } t\geqq T_0,$$

so (H_5) is satisfied with $T_1=T_0$.

Combining (4.3) and (ii) above, we see that if

$$r_0/\theta \leqq T_0 \leqq \theta r_0/[1-\sqrt{1-\theta^2}],$$

then the assumptions (H_0)–(H_5) are all satisfied. By Theorem 3.3, the wave equation is exactly controllable for any $T>0$.

Acknowledgment. The authors acknowledge the influence of the work of Tartar [12] on this paper. The second author would like to thank Professor Jeffrey Rauch for some extremely helpful discussions. He is especially indebted to Professor J. L. Lions for inviting him to IRIA-LABORIA, where this paper was completed. We dedicate this paper to him.

REFERENCES

[1] G. CHEN, *Control and stabilization for the wave equation in a bounded domain*, this Journal, 17 (1979), pp. 66–81.

[2] ———, *Control and stabilization for the wave equation in a bounded domain, Part II*, this Journal, this issue, pp. 114–122.

[3] ———, *Energy decay estimates and exact boundary value controllability for the wave equation in a bounded domain*, J. Math. Pures Appl., 58 (1979), pp. 249–273.

[4] J. COOPER AND C. BARDOS, *A nonlinear wave equation in a time-dependent domain*, J. Math. Anal. Appl., 42 (1973), pp. 29–60.

[5] J. COOPER, *Local decay of solutions of the wave equation in the exterior of a moving body*, J. Math. Anal. Appl., 49 (1975), pp. 130–153.

[6] J. COOPER AND W. A. STRAUSS, *Energy boundedness and decay of waves reflecting off a moving obstacle*, Indiana Univ. Math. J., 25 (1976), pp. 671–690.

[7] J. L. LIONS, *Une remarque sur les problèmes d'évolution nonlinéares dans les domaines non cylindriques*, Rev. Roumaine Math. Pures Appl., 9 (1964), pp. 11–18.

[8] J. L. LIONS AND E. MAGENES, *Problèmes aux limites non-homogénés et applications*, vol. I, Dunod, Paris, 1968.

[9] C. S. MORAWETZ, *Energy Identities for the Wave Equation*, Courant Inst. Math. Sci. Res. Rep. ≠IMM 346, New York University, New York, 1966.

[10] D. L. RUSSELL, *Exact boundary value controllability theorems for wave and heat processes in star-complemented regions*, in Differential Games and Control Theory, Marcel-Dekker, New York, 1974.

[11] W. A. STRAUSS, *Decay of solutions of hyperbolic equations with localized nonlinear terms*, Symposia Mathematica VII, Ist. Naz. Alta Mat., Rome, 1971, pp. 339–355.

[12] L. TARTAR, *Lecture Notes in Partial Differential Equations* (unpublished), University of Wisconsin, Madison, WI, 1974.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 1, January 1981

0363-0129/81/1901-0011 $01.00/0

THE INFINITE TIME QUADRATIC CONTROL PROBLEM FOR LINEAR SYSTEMS WITH STATE AND CONTROL DELAYS: AN EVOLUTION EQUATION APPROACH*

R. B. VINTER† AND R. H. KWONG‡

Abstract. We show how a linear differential delay equation with delays in the control can be reformulated as an evolution equation with bounded input operator. As a simple application, we solve an infinite time quadratic cost control problem for the delay differential equation. We also point out how our results, even when specialized to the case where delays in the control disappear, extend results previously published in the literature.

1. Introduction. We show that a linear delay differential equation with delays in the control may be reformulated as an evolution equation on $\mathbb{R}^n \times L^2$,

$$\frac{d\tilde{x}(t)}{dt} = \mathcal{A}\tilde{x}(t) + \mathcal{B}u(t),$$

in which $\mathcal{B}$ is a bounded linear operator. We proceed to obtain a solution, making the usual stabilizability and detectability assumptions, to an infinite time quadratic cost control problem for the delay differential equation. This is achieved, very simply, by applying well known results on the quadratic cost control of evolution equations with bounded input operators. It is shown that the optimal control may be expressed in feedback form through a bounded linear operator which is the unique nonnegative, self-adjoint solution to an "algebraic Riccati equation", and that the closed loop system is stable. The class of problems considered is fairly general, but excludes "point delays" in the controls.

Quadratic control problems for linear delay differential equations with delay in the control have been studied by a number of authors [12], [13], [15]. Reference [15] alone is concerned with the infinite time problem. (In fact the case when only point delays are present in the control is treated, but the methods adapt to our problem.) In this paper however the equations characterizing the optimal feedback operator are not shown to have a unique solution, and the methods used are much more complicated than the ones given here.

The idea of introducing an evolution equation to remove delays from the control is not new. One approach has already been studied by Ichikawa [12], and applied to finite time quadratic cost control problems involving delays in the control. This approach is more general than ours since it admits point delays in the controls but, as regards our application, it suffers from a number of disadvantages.

Most importantly, the evolution equation employed by Ichikawa involves an input operator $\mathcal{B}$ which has range space *larger* than the state space of the evolution equation and mild solutions are defined not in a conventional sense, but using "extension by continuity" arguments; that is the case even when point delays in the control are absent. For such evolution equations a quadratic cost control theory is given in [12] for the finite time problem (subject to certain technical conditions holding), but not for the infinite time problem. By contrast our formulation introduces an evolution equation of a

* Received by the editors January 24, 1979, and in revised form March 13, 1980.

† Department of Computing and Control, Imperial College of Science and Technology, London SW7 2BZ, England.

‡ Department of Electrical Engineering, University of Toronto, Toronto, Canada M5S 1A4.

standard form. This enables us to exploit simple, general results which are available for the study of the infinite time problem.

A further disadvantage of Ichikawa's approach (as applied to our problem) is that the state space employed, $\mathbb{R}^n \times L^2 \times L^2$, is larger than our state space, $\mathbb{R}^n \times L^2$. The presence of the extra coordinate in the state results in a larger number of coupled equations giving the kernels of the optimal feedback operator, than would occur if our evolution equation were used to solve the finite time problem. The solution given in [12] is therefore more complicated than ours.

The core of this paper is the reformulation of a linear differential equation with delay in the control as an evolution equation. Instead of using the usual infinitesimal generator associated with differential delay equations (see, for example, [21]) we use its adjoint. Delfour–Lee–Manitius [9] made use of this adjoint operator in their study of "state space reduction" for the operator Riccati equation; the novelty here is that we draw attention to its significance in treatment of delays in the control. We examine one application in which new results are simply obtained, and point out how our results, even when specialized to the case where delays in the control disappear, extend previously known results. We hope that the reformulation will have other applications in the study of filtering and stochastic control problems, and in the study of structural properties of systems with delays in the control.

2. Some remarks on notation. All spaces are real.

Spaces of functions on $[-b, 0]$, where b is a positive number, will occur frequently and, for simplicity, the domain $[-b, 0]$ is often suppressed in our notation. Thus $L^2(-b, 0; \mathbb{R}^n)$ is written L^2, etc. In such cases the range space of the functions is determined by context.

$\mathscr{L}(X, Y)$ denotes the space of bounded, linear operators mapping the Hilbert space X into the Hilbert space Y. $\mathscr{L}(X, X)$ is written $\mathscr{L}(X)$.

The domain of a map $\mathscr{A}$ is written $\mathscr{D}\{\mathscr{A}\}$.

The adjoint of a densely defined linear operator G from one Hilbert space to another is written G^*.

The transpose of a matrix M is written M′.

χ_A denotes the indicator function for the set A.

$W^{1,2}(I, \mathbb{R}^\alpha)$ is the space of absolutely continuous $\mathbb{R}^\alpha$-valued functions on the compact interval I, with square integrable derivatives.

Given an element $h \in \mathbb{R}^n \times L^2$, $h^0 \in \mathbb{R}^n$, $h^1(\cdot) \in L^2$ will denote the two coordinates of h, thus $h = (h^0, h^1(\cdot))$. The inner product on $\mathbb{R}^n \times L^2$ will be denoted by $\langle\langle \cdot, \cdot \rangle\rangle$. All the other inner products will be denoted by $\langle \cdot, \cdot \rangle$, and the underlying space is understood from context.

3. The delay differential equation. Let b be a positive number. The $\mathbb{R}^n$-valued function $\mathscr{L}(\cdot)$ which carries $\mathbb{R}^n$-valued functions on $[-b, 0]$ into R^n is defined as follows:

$$\mathscr{L}(h(\cdot)) = \sum_{i=0}^{k} A_i h(\theta_i) + \int_{-b}^{0} A_{01}(\theta) h(\theta)\, d\theta.$$

Now suppose that the function $r(\cdot)$ has domain $[-b, \infty)$. Then, for each $t \in [0, \infty)$, the function $r_t(\cdot)$ with domain $[-b, 0]$ is defined as $r_t(\theta) = r(t + \theta)$. We shall be interested in the delay differential equation:

$$\frac{dx(t)}{dt} = \mathscr{L}(x_t(\cdot)) + B_0 u(t) + \int_{-b}^{0} B_{01}(\theta) u_t(\theta)\, d\theta, \tag{3.1}$$

$$x(0) = \xi^0, \qquad x(\theta) = \xi^1(\theta), \qquad u(\theta) = \eta(\theta), \qquad \theta \in [-b, 0]. \tag{3.2}$$

In the foregoing, k is a positive integer and $-b \leqq \theta_k \leqq \theta_{k-1} \leqq \cdots < \theta_0 = 0$. The A_i's are $n \times n$ matrices. B_0 is an $n \times m$ matrix. $A_{01}(\cdot)$, $B_{01}(\cdot)$ are functions on $[-b, 0]$ taking values respectively as $n \times n$ and $n \times m$ matrices; $A_{01}(\cdot)$ is measurable and essentially bounded and $B_{01}(\cdot)$ is square integrable.

$(\xi^0, \xi^1(\cdot), \eta(\cdot))$ is an element in $\mathbb{R}^n \times L^2(-b, 0; \mathbb{R}^n) \times L^2(-b, 0; \mathbb{R}^m)$, while $u(\cdot)$ is an element in $L^2_{\text{loc}}(0, \infty; \mathbb{R}^m)$.

A solution to (3.1), (3.2) is a function $x(\cdot)$ on $[-b, \infty)$, which is locally absolutely continuous on $[0, \infty)$ and satisfies (3.1) a.e. on $[0, \infty)$, $x(0) = \xi^0$, $x(\theta) = \xi^1(\theta)$, $-b \leqq \theta \leqq 0$, a.e.

The "inhomogeneous term" $B_0 u(t) + \int_{-b}^{0} B_{01}(\theta) u_t(\theta)\, d\theta$ is locally square integrable under our assumption; it follows from standard results [7] that there is a unique solution to (3.1), (3.2).

4. An evolution equation. We define the linear operator $\mathcal{A}$ on $\mathbb{R}^n \times L^2(-b, 0; \mathbb{R}^n)$ as follows:

Let $\chi_i(\theta) = \chi_{[\theta_i, 0]}(\theta)$. $\mathcal{D}(\mathcal{A})$ denotes the collection of all $(h^0, h^1(\cdot)) \in \mathbb{R}^n \times L^2$ for which there exists an element $z \in W^{1,2}$ with $z(-b) = 0$ such that

$$z(\theta) = h^1(\theta) - \sum_{i=1}^{k} A_i h^0 \cdot \chi_i(\theta) \quad \text{a.e. } \theta \in [-b, 0].$$

The operator $\mathcal{A}$ with domain $\mathcal{D}(\mathcal{A})$ is now defined as

$$[\mathcal{A}(h^0, h^1(\cdot))]^0 = \sum_{i=0}^{k} A_i h^0 + z(0),$$

$$[\mathcal{A}(h^0, h^1(\cdot))]^1(\theta) = A_{01}(\theta) h^0 - \frac{dz(\theta)}{d\theta}, \qquad -b \leqq \theta \leqq 0 \quad \text{a.e.}$$

where

$$z(\theta) = h^1(\theta) - \sum_{i=1}^{k} A_i h^0 \chi_i(\theta).$$

$\mathcal{A}$ is the infinitesimal generator of a strongly continuous semigroup on $\mathbb{R}^n \times L^2$ since it is the adjoint of an infinitesimal generator (see, for example, [4, pp. 45–52]). Indeed $\mathcal{A}$ is the adjoint of $\mathcal{A}^*$, where $\mathcal{A}^*$ is defined as follows:

$$\begin{aligned} &\mathcal{D}\{\mathcal{A}^*\} = \{(h(0), h(\cdot)) \mid h(\cdot) \in W^{1,2}\}, \\ &\mathcal{A}^*(h^0, h^1(\cdot)) = \left((\mathcal{L}'(h^1(\cdot)), \frac{dh^1}{d\theta}(\cdot)\right), \end{aligned} \tag{4.1}$$

(see, for example, [21]). In (4.1),

$$\mathcal{L}'(h^1(\cdot)) = \sum_{i=0}^{k} A_i' h^1(\theta_i) + \int_{-b}^{0} A_{01}'(\theta) h^1(\theta)\, d\theta.$$

$\mathcal{A}^*$ will be recognized as the infinitesimal generator of the semigroup of operators associated with evolution of solution segments for the delay differential equation

$$\frac{dx(t)}{dt} = \mathcal{L}'(x_t(\cdot)), \tag{4.2}$$

with initial data in $\mathbb{R}^n \times L^2$ [21].

We denote by $\{T(t)|t \geqq 0\}$ the semigroup generated by $\mathcal{A}$.
Define the bounded linear map $\mathcal{B}: \mathbb{R}^m \to \mathbb{R}^n \times L^2$ by

$$\mathcal{B}u = (B_0 u, B_{01}(\cdot)u), \qquad u \in \mathbb{R}^m.$$

We shall be interested in the mild solution of the evolution equation

$$\frac{d\tilde{x}(t)}{dt} = \mathcal{A}\tilde{x}(t) + \mathcal{B}u(t), \tag{4.3}$$

$$\tilde{x}(0) = \tilde{\xi}, \tag{4.4}$$

for $\tilde{\xi} \in \mathbb{R}^n \times L^2$ and $u(\cdot) \in L^2_{loc}(0, \infty; \mathbb{R}^m)$, where by the mild solution of (4.3), (4.4) we mean the continuous function $\tilde{x}(\cdot): [0, \infty) \to \mathbb{R}^n \times L^2$ given by the variation of constants formula

$$\tilde{x}(t) = T(t)\tilde{\xi} + \int_0^t T(t-s)\mathcal{B}u(s)\, ds.$$

5. Equivalence. We now relate the solution of the delay differential equation (3.1), (3.2) to the mild solution of the evolution equation (4.3), (4.4) when the initial condition on the evolution equation is appropriately chosen. For this purpose we introduce the continuous, linear map $M(\cdot, \cdot, \cdot): \mathbb{R}^n \times L^2(-b, 0; \mathbb{R}^n) \times L^2(-b, 0; \mathbb{R}^m) \to \mathbb{R}^n \times L^2(-b, 0; \mathbb{R}^n)$:

$$M(h^0, h^1(\cdot), v(\cdot)) = (h^0, m(\cdot)),$$

where

$$m(\theta) = \sum_{i=1}^{k} A_i h^1(\theta_i - \theta)\chi_i(\theta) + \int_{-b}^{\theta} A_{01}(\alpha)h^1(\alpha - \theta)\, d\alpha + \int_{-b}^{\theta} B_{01}(\alpha)v(\alpha - \theta)\, d\alpha,$$

(recall that $\chi_i(\theta) = \chi_{[\theta_i, 0]}(\theta)$).

The range of M is written $\mathcal{M}$.

We remark that our operator M can be expressed as

$$M(h^0, h^1, v) = F(h^0, h^1) + (0, Kv),$$

for $(h^0, h^1, v) \in \mathbb{R}^n \times L^2 \times L^2$, where F is the "F-operator" introduced in Delfour–Manitius [9] and K is given by

$$(Kv)(\theta) = \int_{-b}^{\theta} B_{01}(\alpha)v(\alpha - \theta)\, d\alpha, \qquad -b \leqq \theta \leqq 0.$$

The properties of the F-operator are studied further in [9], [11] and [18].

THEOREM 5.1. *Suppose that $x(t)$, $t \geqq -b$, is the solution to* (3.1), (3.2) *and that $\tilde{x}(t)$, $t \geqq 0$, is the mild solution to* (4.3), (4.4) *for initial data $\tilde{\xi}(\cdot) = M(\xi^0, \xi^1(\cdot), \eta(\cdot))$. Then, for $t \geqq 0$,*

$$\tilde{x}(t) = M(x(t), x_t(\cdot), u_t(\cdot)). \tag{5.1}$$

It is clear from Theorem 5.1 that, for initial data in $\mathcal{M}$, the mild solution of the evolution equation (4.3), (4.4) evolves in $\mathcal{M}$. This is true also for arbitrary initial conditions, provided that we wait for the delay interval to elapse:

THEOREM 5.2. *Let $\tilde{x}(t)$, $t \geqq 0$, be the mild solution to* (4.3), (4.4) *for arbitrary initial data $\tilde{\xi}(\cdot) \in \mathbb{R}^n \times L^2$. Then for $t \geqq b$,*

$$\tilde{x}(t) = M(\tilde{x}^0(t), \tilde{x}^0_t(\cdot), u_t(\cdot)), \tag{5.2}$$

where $\tilde{x}^0(t)$ is the first component of $\tilde{x}(t)$.

6. Proof of the theorems.

Proof of Theorems 5.1 *and* 5.2. To begin with, suppose that the initial data $\tilde{\xi}$ in (4.4) lie in $\mathscr{D}\{\mathscr{A}\}$ and that $u(\cdot)$ is a continuously differentiable function. With these assumptions it is known that the mild solution $\tilde{x}(t)$, $t \geqq 0$, to (4.3), (4.4) is a strong solution in the sense that $\tilde{x}(t) \in \mathscr{D}\{\mathscr{A}\}$, $t \geqq 0$, $\tilde{x}(\cdot)$ is everywhere strongly differentiable and satisfies (4.3) for each t.

We define $z(t; \cdot) \in \mathrm{L}^2$ in terms of $\tilde{x}(t)$ as follows:

$$z(t; \theta) = \tilde{x}^1(t; \theta) - \sum_{i=1}^{k} \mathrm{A}_i \tilde{x}^0(t) \cdot \chi_i(\theta), \tag{6.1}$$

where we have used $\tilde{x}^1(t; \theta)$ to denote $[\tilde{x}^1(t)](\theta)$. $z(t; \cdot) \in \{\phi \in W^{1,2} : \phi(-b) = 0\}$ since $\tilde{x}(t) \in \mathscr{D}\{\mathscr{A}\}$ for all $t \geqq 0$, and because $\tilde{x}(\cdot)$ is strongly differentiable, the same is true of $\{z(t; \cdot) | t \geqq 0\}$.

Equation (4.3) may now be written as

$$\frac{d}{dt}\tilde{x}^0(t) = \sum_{i=0}^{k} \mathrm{A}_i \tilde{x}^0(t) + z(t; 0) + \mathrm{B}_0 u(t), \tag{6.2}$$

and

$$\frac{d}{dt}\tilde{x}^1(t; \theta) = -\frac{d}{d\theta} z(t; \theta) + \mathrm{A}_{01}(\theta)\tilde{x}^0(t) + \mathrm{B}_{01}(\theta)u(t). \tag{6.3}$$

Noting (6.1) we may write (6.3) as an equation for $z(t; \theta)$:

$$\frac{dz(t; \theta)}{dt} = -\frac{dz(t; \theta)}{d\theta} + \mathrm{A}_{01}(\theta)\tilde{x}^0(t) - \sum_{i=1}^{k} \mathrm{A}_i \frac{d\tilde{x}^0(t)}{dt} \cdot \chi_i(\theta) + \mathrm{B}_{01}(\theta)u(t).$$

Now $z(t; \cdot)$ evolves in $\{\phi \in W^{1,2} | \phi(-b) = 0\}$ which is the domain of the infinitesimal generator, $-(d/d\theta)(\cdot)$, of the semigroup of truncated right shifts on L^2, $\{\Phi(t) : t \geqq 0\}$:

$$(\Phi(t)g(\cdot))(\theta) = \begin{cases} g(\theta - t), & -b \leqq \theta - t \leqq 0, \\ 0, & \text{otherwise}, \end{cases} \tag{6.4}$$

$-b \leqq \theta \leqq 0$, for all $g(\cdot) \in \mathrm{L}^2$.

We may view $z(t; \cdot)$ therefore as a strong solution to an evolution equation; it is then also a mild solution and we have the representation given by the variation of constants formula

$$z(t; \theta) = \Phi(t)z(0; \theta) + \int_0^t \Phi(t - s)\left[-\sum_{i=1}^{k} \mathrm{A}_i \frac{d\tilde{x}^0(s)}{ds} \cdot \chi_i(\theta) + \mathrm{A}_{01}(\theta)\tilde{x}^0(s) + \mathrm{B}_{01}(\theta)u(s) \right] ds.$$

Using (6.4) then, and some routine manipulations, we obtain

$$\begin{aligned} z(t; \theta) = {} & z(0, \theta - t) \cdot \chi_{[-b, 0]}(\theta - t) \\ & + \sum_{i=1}^{k} \mathrm{A}_i(-\tilde{x}^0(t) \cdot \chi_i(\theta) + \tilde{x}^0(0)\chi_i(\theta - t) + \tilde{x}^0(\theta_i - \theta + t) \cdot \chi_{[\theta_i, \theta_i + t]}(\theta)) \\ & + \int_{\max\{-b, \theta - t\}}^{\theta} (\mathrm{A}_{01}(\alpha)\tilde{x}^0(t + \alpha - \theta) + \mathrm{B}_{01}(\alpha)u(t + \alpha - \theta))\, d\alpha. \end{aligned} \tag{6.5}$$

Now suppose that $t \geqq b$. Then it follows from (6.1) and (6.5) that

$$\tilde{x}^1(t;\theta) = \sum_{i=1}^{k} A_i \tilde{x}^0(t+\theta_i-\theta)\cdot\chi_i(\theta) + \int_{-b}^{\theta} (A_{01}(\alpha)\tilde{x}^0(t+\alpha-\theta) + B_{01}(\alpha)u(t+\alpha-\theta))\,d\alpha,$$

which may be written as (5.2). Thus the conclusions of Theorem 5.2 hold when $\tilde{\xi}$ and $u(\cdot)$ are "regular".

Now suppose that the initial data $(\xi^0, \xi^1(\cdot), \eta(\cdot))$ in (3.2) belong to the subspace Q,

$$Q = \{(h(0), h(\cdot), v(\cdot)) \in \mathbb{R}^n \times L^2 \times L^2 | h(\cdot), v(\cdot) \in W^{1,2}\},$$

and take $\tilde{\xi} = M(\xi^0, \xi^1(\cdot), \eta(\cdot))$. We have that $\tilde{\xi} \in \mathscr{D}(\mathscr{A})$; indeed, this property is known [11, Theorem 3.1] when $B_{01}(\cdot) = 0$, and the argument there adapts to the general situation in an obvious manner.

For this choice of initial data we have, by (6.1),

$$z(0;\theta-t) = \sum_{i=1}^{k} A_i(\xi^1(t+\theta_i-\theta)\cdot\chi_i(\theta-t) - \tilde{x}^0(0)\cdot\chi_i(\theta-t)) + \int_{-b}^{\theta-t} (A_{01}(\alpha)\xi^1(t+\alpha-\theta) + B_{01}(\alpha)\eta(t+\alpha-\theta))\,d\alpha,$$

$-b \leqq \theta - t \leqq 0$, $t \geqq 0$. This expression for $z(0;\cdot)$ may be substituted back into (6.5) to give, for $t \geqq 0$,

$$\begin{aligned} z(t;\theta) = \sum_{i=1}^{k} A_i(-\tilde{x}^0(t)\chi_i(\theta) + \xi^1(t+\theta_i-\theta)\cdot\chi_i(\theta-t) + \tilde{x}^0(t+\theta_i-\theta)\cdot\chi_{[\theta_i,\theta_i+t]}(\theta)) \\ + \int_{\max\{-b,\theta-t\}}^{\theta} A_{01}(\alpha)\tilde{x}^0(t+\alpha-\theta)\,d\alpha \\ + \int_{-b}^{\theta-t} A_{01}(\alpha)\xi^1(t+\alpha-\theta)\,d\alpha\cdot\chi_{[-b,0]}(t-\theta) \\ + \int_{\max\{-b,\theta-t\}}^{\theta} B_{01}(\alpha)u(t+\alpha-\theta)\,d\alpha \\ + \int_{-b}^{\theta-t} B_{01}(\alpha)u(t+\alpha-\theta)\,d\alpha\cdot\chi_{[-b,0]}(t-\theta), \end{aligned}$$

which may be written as

$$z(t;\theta) = -\sum_{i=1}^{k} A_i\tilde{x}^0(t)\chi_i(\theta) + \sum_{i=1}^{k} A_i\chi_i(\theta)\tilde{x}^0(t+\theta_i-\theta) + \int_{-b}^{\theta} \{A_{01}(\alpha)\tilde{x}^0(t+\alpha-\theta) + B_{01}(\alpha)u(t+\alpha-\theta)\}\,d\alpha,$$

when we take $\tilde{x}^0(\theta) = \xi^1(\theta)$, $u(\theta) = \eta(\theta)$, for $\theta \in [-b, 0]$. ($\xi^1(0) = \xi^0 = \tilde{x}^0(0)$, remember.) It follows now from (6.2) that the locally absolutely continuous function on $[-b, \infty)$ defined to be $\xi^1(t)$, $t \in [-b, 0]$ and $\tilde{x}^0(t)$, $t \in [0, \infty)$, coincides with the solution $x(t)$, $t \geqq -b$, to the delay differential equation (3.1), (3.2). Taking note of (6.1) we see that (5.1) is satisfied.

The assertions of the theorems have been shown to be true under the additional assumptions that $u(\cdot)$ and the initial data are smooth. It remains to remove these additional assumptions; this is done by using a simple extension by continuity argument.

Choose $t \geqq 0$. Let $x(t)$ be the solution at time t to (3.1), (3.2), in which the control $u(\cdot)$ is taken to be defined on $[0, t]$.

The map

$$S_1 : \mathbb{R}^n \times L^2(-b, 0; \mathbb{R}^n) \times L^2(-b, 0; \mathbb{R}^m) \times L^2(0, t; \mathbb{R}^m)$$
$$\to \mathbb{R}^n \times L^2(-b, 0; \mathbb{R}^n) \times L^2(-b, 0; \mathbb{R}^m),$$
$$S_1(\xi^0, \xi^1, \eta(\cdot), u(\cdot)) = (x(t), x_t(\cdot), u_t(\cdot))$$

is continuous. This is proved by obvious modifications of standard arguments giving continuity of solutions to linear delay differential equations in the data (see, e.g., [7]).

Now let $\tilde{x}(\cdot)$ be the mild solution on $[0, t]$ of (4.3), (4.4). Define the map

$$S_3 : \mathbb{R}^n \times L^2(-b, 0; \mathbb{R}^n) \times L^2(0, t; \mathbb{R}^m \to \mathbb{R}^n \times L^2(-b, 0; \mathbb{R}^n),$$

and also (in the case that $t \geqq b$) the map

$$S_3 : \mathbb{R}^n \times L^2(-b, 0; \mathbb{R}^n) \times L^2(0, t; \mathbb{R}^m) \to \mathbb{R}^n \times L^2(-b, 0; \mathbb{R}^n) \times L^2(-b, 0; \mathbb{R}^m),$$

by

$$S_2(\tilde{\xi}, u(\cdot)) = \tilde{x}(t),$$

and

$$S_3(\tilde{\xi}, u(\cdot)) = (\tilde{x}^0(t), \tilde{x}_t^0(\cdot), u_t(\cdot)).$$

The two maps are readily shown to be continuous.

Our conclusions so far may now be expressed in terms of S_1, S_2 and S_3, thus:

$$MS_1(\xi^0, \xi^1(\cdot), \eta(\cdot), u(\cdot)) = S_2(M(\xi^0, \xi^1(\cdot), \eta(\cdot)), u(\cdot)), \tag{6.6}$$

for all $(\xi^1, \xi^1(\cdot), \eta(\cdot), u(\cdot))$ in the dense subset $Q \times C^1(0, t; \mathbb{R}^m)$ of $\mathbb{R}^n \times L^2(-b, 0; \mathbb{R}^n) \times L^2(-b, 0; \mathbb{R}^m) \times L^2(0, t; \mathbb{R}^m)$, and (when $t \geqq b$)

$$S_2(\tilde{\xi}, u(\cdot)) = MS_3(\tilde{\xi}, u(\cdot)), \tag{6.7}$$

for all $(\tilde{\xi}, u(\cdot))$ in the dense subset $\mathscr{D}\{\mathscr{A}\} \times C^1(0, t; \mathbb{R}^m)$ of $\mathbb{R}^n \times L^2(-b, 0; \mathbb{R}^n) \times L^2(0, t; \mathbb{R}^m)$.

But the continuity of S_1, S_2, S_3 (and M as a map from $\mathbb{R}^n \times L^2(-b, 0; \mathbb{R}^n) \times L^2(-b, 0; \mathbb{R}^m)$ to $\mathbb{R}^n \times L^2(-b, 0; \mathbb{R}^n)$) then assure that (6.6), (6.7) hold for arbitrary $u(\cdot)$ and initial data $(\xi^0, \xi^1(\cdot), \eta(\cdot))$, $\tilde{\xi}$ respectively. The theorems are proved.

We remark that the extension by continuity arguments featured in the above proof has been used previously to obtain representations of solutions to linear functional differential equations (see, e.g., [2]).

7. An infinite time quadratic cost control problem. We now apply the results of the previous sections to the infinite time quadratic cost control problem associated with (3.1), (3.2).

Let C be an $r \times n$ matrix, and let the initial data $(\xi^0, \xi^1(\cdot), \eta(\cdot)) \in \mathbb{R}^n \times L^2 \times L^2$ be given. Consider the following control problem ($\mathscr{Q}$):

Minimize

$$\int_0^\infty \{x'(t) C'Cx(t) + u'(t)u(t)\}\, dt$$

over $u(\cdot) \in L^2(0, \infty; \mathbb{R}^m)$, where $x(\cdot) : [-b, \infty) \to \mathbb{R}^n$ is the solution to (3.1), (3.2). (The case where we have $u'(t)Ru(t)$ with R a positive definite matrix can of course be readily reduced to the above setup.)

We shall show that the solution to this problem is characterized in terms of the unique nonnegative, self-adjoint solution of an "operator algebraic Riccati equation" when the system

$$\begin{aligned} \frac{dx(t)}{dt} &= \mathscr{L}(x_t(\cdot)) + B_0 u(t) + \int_{-b}^{0} B_{01}(\theta) u_t(\theta)\, d\theta, \\ y(t) &= Cx(t) \end{aligned} \tag{7.1}$$

satisfies certain stabilizability and detectability assumptions. Stabilizability will be in the sense that there exists an "L^2-stabilizing control" for arbitrary initial data; detectability will mean that a certain "transposed" system has the stabilizability property.

DEFINITION 7.1. (i) The system (7.1) is *stabilizable* if, for each $(\xi^0, \xi^1(\cdot)) \in \mathbb{R}^n \times L^2$, there exists $u \in L^2(0, \infty, \mathbb{R}^m)$ such that the solution $x(\cdot)$ to (3.1) with initial data $(\xi^0, \xi^1(\cdot), 0)$ satisfies $\int_0^\infty x'(t)x(t)\, dt < \infty$.

(ii) The system (7.1) is *detectable* if the delay differential equation

$$\frac{dx(t)}{dt} = \mathscr{L}'(x_t(\cdot)) + C'u(t)$$

($\mathscr{L}'(\cdot)$ as given in § 4) is stabilizable.

Define $\mathscr{C} \in \mathscr{L}(\mathbb{R}^n \times L^2, \mathbb{R}^r)$ as

$$\mathscr{C}\tilde{x} = C\tilde{x}^0 \quad \text{for } \tilde{x} = (\tilde{x}^0, \tilde{x}^1(\cdot)).$$

Then, by Theorem 5.1, the control problem may be reformulated as follows:

Minimize

$$\int_0^\infty \{\|\mathscr{C}\tilde{x}(t)\|^2 + u'(t)u(t)\}\, dt$$

over $u \in L^2(0, \infty; \mathbb{R}^m)$, where $\tilde{x}(t)$, $t \geqq 0$, is the mild solution of (4.3), (4.4) for initial data $\tilde{\xi} = M(\xi^0, \xi^1(\cdot), \eta(\cdot))$.

We have now placed the problem in an abstract setting in which known results are "almost" directly applicable to give a solution. There remains a difficulty: we need to show that our definitions of stabilizability and detectability, chosen as natural in this context, are adequate for application of the standard theory. This difficulty is removed by the following proposition.

PROPOSITION 7.2. *Referring to Definition* 7.1 *we have that* (i) is equivalent to (i)' *and* (ii) *is equivalent to* (ii)', where:

(i)' *For each* $\tilde{\xi} \in \mathbb{R}^n \times L^2$, *there exists some* $u(\cdot) \in L^2(0, \infty; \mathbb{R}^m)$ *such that the mild solution* $\tilde{x}(\cdot)$ *to* (4.3), (4.4) *satisfies* $\int_0^\infty \|\tilde{x}(t)\|^2\, dt < \infty$.

(ii)' *For each* $\tilde{\xi} \in \mathbb{R}^n \times L^2$, *there exists* $u(\cdot) \in L^2(0, \infty, \mathbb{R}^m)$ *such that the mild solution* $\tilde{x}(\cdot)$ *to*

$$\begin{aligned} \frac{d\tilde{x}(t)}{dt} &= \mathscr{A}^*\tilde{x}(t) + \mathscr{C}^* u(t), \\ \tilde{x}(0) &= \tilde{\xi} \end{aligned}$$

satisfies $\int_0^\infty \|\tilde{x}(t)\|^2\, dt < \infty$.

This result is not quite obvious. Consider stabilizability, for instance. The abstract condition (i)' requires that we can stabilize the evolution equation (4.3), (4.4) for *all* initial data in $\mathbb{R}^n \times L^2$. This is seemingly stronger than our definition of stabilizability, in which stabilization is required for initial data only in the subset $\mathscr{M} \subset \mathbb{R}^n \times L^2$ (it is possible that $\mathscr{M}$ is not even dense).

Proof of Proposition 7.2. Suppose (i)′. Take $(\xi^0, \xi^1(\cdot), \eta(\cdot)) \in \mathbb{R}^n \times L^2 \times L^2$. Then there exists $u(\cdot) \in L^2(0, \infty; \mathbb{R}^m)$ such that the mild solution $\tilde{x}(\cdot)$ to (4.3), (4.4) with initial data $\tilde{\xi} = M(\xi^0, \xi^1(\cdot), \eta(\cdot))$ satisfies $\int_0^\infty \|\tilde{x}(t)\|^2\, dt < \infty$. But then $\int_0^\infty \|\tilde{x}^0(t)\|^2\, dt < \infty$. By the results of Theorem 5.1, $\tilde{x}^0(\cdot)$ is the solution to (3.1), (3.2). It follows that (i) is true.

Now suppose (i). Take $\tilde{\xi} \in \mathbb{R}^n \times L^2$. Notice first that, by Theorem 5.2, for $u(\cdot) \in L^2_{\text{loc}}(0, \infty; \mathbb{R}^m)$ such that $u(t) = 0$, for $t \in [0, b]$, the mild solution $\tilde{x}(\cdot)$ to (4.3), (4.4) satisfies

$$\tilde{x}(b) = M(\tilde{x}^0(b), \tilde{x}^0_b(\cdot), 0).$$

Now let $\tilde{u}(\cdot) \in L^2(b, \infty; \mathbb{R}^m)$ be such that the solution $x(t)$, $t \geqq b$, to (3.1) with initial data $(\tilde{x}^0(b), \tilde{x}^0_b(\cdot), 0)$ (specified at $t = b$) satisfies $\int_b^\infty \|x(t)\|^2\, dt < \infty$. Such a $\tilde{u}(\cdot)$ exists by (i). Define $u \in L^2(0, \infty; \mathbb{R}^m)$ as

$$u(t) = \begin{cases} 0, & 0 \leqq t \leqq b \\ \tilde{u}(t), & t \geqq b. \end{cases}$$

By the "semigroup property" and Theorem 5.1, the mild solution $\tilde{x}(t)$, $t \geqq 0$, to (4.3), (4.4) for this $u(\cdot)$ satisfies

$$\int_0^\infty \|\tilde{x}^0(t)\|^2\, dt < \infty.$$

By Theorem 5.2, however,

$$\tilde{x}(t) = M(\tilde{x}^0(t), \tilde{x}^0_t(\cdot), u_t(\cdot)), \qquad t \geqq b.$$

Using Fubini's theorem and some standard estimates we show that $\int_b^\infty \|\tilde{x}(t)\|^2\, dt < \infty$. It follows that $\int_0^\infty \|\tilde{x}(t)\|^2\, dt < \infty$. We conclude (i)′.

Recall the interpretation of $\mathcal{A}^*$ as the infinitesimal generator of the semigroup describing evolution of trajectory segments in $\mathbb{R}^n \times L^2$ of (4.2), and note that $\mathcal{C}^* \in \mathcal{L}(\mathbb{R}^r; \mathbb{R}^n \times L^2)$ is given by $\mathcal{C}^* u = (C' u, 0)$. Take $u(\cdot) \in L^2_{\text{loc}}(0, \infty; \mathbb{R}^r)$ and $(\xi^0, \xi^1(\cdot)) \in \mathbb{R}^n \times L^2$. By well-known results the mild solution $\tilde{x}(t)$, $t \geqq 0$, to

$$\frac{d\tilde{x}(t)}{dt} = \mathcal{A}^* \tilde{x}(t) + \mathcal{C}^* u(t),$$

$$\tilde{x}(0) = (\xi^0, \xi^1(\cdot)),$$

and the solution $x(t)$, $t \geqq -b$, to the delay differential equation

$$\frac{dx(t)}{dt} = \mathcal{L}'(x_t(\cdot)) + C' u(t)$$

with initial data $(\xi^0, \xi^1(\cdot))$ are related by

$$\tilde{x}(t) = (x(t), x_t(\cdot)).$$

(See, e.g., [1], [2] or [21]).

But again by Fubini's theorem and standard estimates there exists a positive number c such that, for all $z(\cdot) \in L^2(-b, \infty; \mathbb{R})$,

$$\int_0^\infty \|z_t(\cdot)\|_{L^2}\, dt \leqq c \int_{-b}^\infty |z(t)|^2\, dt.$$

It follows simply from these results that (ii) and (ii)′ are equivalent.

The following solution to the control problem is now readily deduced from [5, Chapter 4].

THEOREM 7.3. *Suppose that the system* (7.1) *is stabilizable and detectable. Then there exists a unique nonnegative, self-adjoint operator* $\mathcal{P} \in \mathcal{L}(\mathbb{R}^n \times L^2)$ *satisfying*

$$\langle\langle \mathcal{A}\tilde{x}, \mathcal{P}\tilde{y}\rangle\rangle + \langle\langle \mathcal{P}\tilde{x}, \mathcal{A}\tilde{y}\rangle\rangle + \langle\langle \tilde{x}, [\mathcal{C}^*\mathcal{C} - \mathcal{P}\mathcal{B}\mathcal{B}^*\mathcal{P}]\tilde{y}\rangle\rangle = 0 \quad \text{for all } \tilde{x}, \tilde{y} \in \mathcal{D}\{\mathcal{A}\}. \tag{7.2}$$

The unique solution to the control problem (2) *is given by* $u(\cdot) \in L^2(0, \infty; \mathbb{R}^m)$ *satisfying*

$$u(t) = -\mathcal{B}^*\mathcal{P}(\mathrm{M}(x(t), x_t(\cdot), u_t(\cdot))), \tag{7.3}$$

where $x(\cdot)$ *is the solution of* (3.1), (3.2). *We have*

$$\begin{aligned} &\langle\langle \mathrm{M}(\xi^0, \xi^1(\cdot), \eta(\cdot)), \mathcal{P}\mathrm{M}(\xi^0, \xi^1(\cdot), \eta(\cdot))\rangle\rangle \\ &\quad = \min \int_0^\infty \{x'(t)\mathrm{C}'\mathrm{C}x(t) + u'(t)u(t)\}\, dt. \end{aligned} \tag{7.4}$$

Furthermore the "closed loop system" is exponentially stable in the sense that there exist positive numbers α, ω (*which do not depend on the initial data*) *such that*

$$\|x(t)\| \leqq \alpha \cdot e^{-\omega t}\|\mathrm{M}(\xi^0, \xi^1(\cdot), \eta(\cdot))\|, \qquad t \geqq 0.$$

We mention that the hypotheses of the theorem also assure exponential stability of the closed loop system in the following stronger sense: there exist positive constants α, ω such that

$$\|\tilde{x}(t; \tilde{\xi})\| \leqq \alpha\, e^{-\omega t}\|\tilde{\xi}\|,$$

where $\tilde{x}(t; \tilde{\xi})$ is the solution of the "closed loop evolution equation"

$$\begin{aligned} \frac{d\tilde{x}(t)}{dt} &= (\mathcal{A} - \mathcal{B}\mathcal{B}^*\mathcal{P})\tilde{x}(t), \\ \tilde{x}(0) &= \tilde{\xi}. \end{aligned}$$

For the purposes of comparison with previously available results (see § 8), we observe that the standard theory [5, Chapter 4] applied to our reformulation of the control problem also gives the operator $\mathcal{P}$ of Theorem 7.3 as

$$\mathcal{P} = \operatorname*{strong\ limit}_{T \to \infty} \mathcal{P}_T(0), \tag{7.5}$$

where $\mathcal{P}_T(\cdot) : [0, T] \to \mathbb{R}^n \times L^2$ is the unique solution, in the class of strongly continuous functions such that $\langle\langle \tilde{x}, \mathcal{P}_T(\cdot)\tilde{y}\rangle\rangle$ is absolutely continuous for all $\tilde{x}, \tilde{y} \in \mathcal{D}\{\mathcal{A}\}$, of the operator differential Riccati equation

$$\begin{aligned} &-\frac{d}{dt}\langle\langle \tilde{x}, \mathcal{P}_T(t)\tilde{y}\rangle\rangle \\ &\qquad = \langle\langle \mathcal{A}\tilde{x}, \mathcal{P}_T(t)\tilde{y}\rangle\rangle + \langle\langle \mathcal{P}_T(t)\tilde{x}, \mathcal{A}\tilde{y}\rangle\rangle + \langle\langle \tilde{x}, [\mathcal{C}^*\mathcal{C} - \mathcal{P}_T(t)\mathcal{B}\mathcal{B}^*\mathcal{P}_T(t)]\tilde{y}\rangle\rangle, \\ &\qquad\qquad \text{for all } \tilde{x}, \tilde{y} \in \mathcal{D}\{\mathcal{A}\}, \\ &\mathcal{P}_T(T) = 0, \end{aligned} \tag{7.6}$$

associated with the "finite time" quadratic cost control problem on $[0, T]$. Note too that

$$\begin{aligned}(7.7)\qquad &\langle\!\langle \mathrm{M}(\xi^0, \xi^1, \eta), \mathscr{P}_T(t)\mathrm{M}(\xi^0, \xi^1, \eta)\rangle\!\rangle \\ &= \min \int_0^{T-t} \{x'(s)\mathrm{C}'\mathrm{C}x(s) + u'(s)u(s)\}\, ds.\end{aligned}$$

We have obtained a solution to the infinite time control problem in terms of the unique, nonnegative, self-adjoint solution $\mathscr{P}$ to (7.2). Detailed knowledge of the structure of $\mathscr{P}$ is therefore highly desirable. In the special case when $\mathrm{B}_{01}(\cdot) = 0$, Kwong [16] has shown:

THEOREM 7.4. *Suppose that* $\mathrm{B}_{01}(\cdot) = 0$. *Then the operator* $\mathscr{P} \in \mathscr{L}(\mathbb{R}^n \times \mathrm{L}^2)$ *of Theorem* 7.3 *may be written as*

$$\mathscr{P}(\xi^0, \xi^1(\cdot)) = \left(\mathrm{P}_0\xi^0 + \int_{-b}^0 \mathrm{P}_1(\theta)\xi^1(\theta)\, d\theta,\ \mathrm{P}_1'(\cdot)\xi^0 + \int_{-b}^0 \mathrm{P}_2(\cdot, \bar\theta)\xi^1(\bar\theta)\, d\bar\theta\right),$$

in which $\mathrm{P}_1(\cdot)$ *and* $\mathrm{P}_2(\cdot, \cdot)$ *are continuous and piecewise continuously differentiable functions of their arguments. The following differential equations characterize* P_0, $\mathrm{P}_1(\cdot)$ *and* $\mathrm{P}_2(\cdot, \cdot)$:

$$\begin{aligned}&\mathrm{P}_0\mathrm{A}_0 + \mathrm{A}_0'\mathrm{P}_0 + \sum_{i=1}^k \mathrm{P}_1(\theta_i)\mathrm{A}_i + \sum_{i=1}^k \mathrm{A}_1'\mathrm{P}_1'(\theta_i) + \mathrm{C}'\mathrm{C} - \mathrm{P}_0\mathrm{B}\mathrm{B}'\mathrm{P}_0 \\ &\qquad + \int_{-b}^0 \mathrm{P}_1(\theta)\mathrm{A}_{01}(\theta)\, d\theta + \int_{-b}^0 \mathrm{A}_{01}'(\theta)\mathrm{P}_1'(\theta)\, d\theta = 0,\end{aligned}$$

$$\begin{aligned}\frac{d}{d\theta}\mathrm{P}_1(\theta) &= -(\mathrm{A}_0' - \mathrm{P}_0\mathrm{B}\mathrm{B}')\mathrm{P}_1(\theta) - \sum_{i=1}^k \mathrm{A}_i'\mathrm{P}_2(\theta_i, \theta) \\ &\quad - \int_{-b}^0 \mathrm{A}_{01}'(\bar\theta)\mathrm{P}_2(\bar\theta, \theta)\, d\bar\theta\end{aligned}$$

with boundary condition $\mathrm{P}_1(0) = \mathrm{P}_0$,

$$\left(\frac{\partial}{\partial\theta} + \frac{\partial}{\partial\bar\theta}\right)\mathrm{P}_2(\theta, \bar\theta) = \mathrm{P}_1'(\theta)\mathrm{B}\mathrm{B}'\mathrm{P}_1(\bar\theta)$$

with boundary condition $\mathrm{P}_2(0, \theta) = \mathrm{P}_1(\theta)$, $-b \leqq \theta \leqq 0$. *Furthermore, we have the symmetry relations*

$$\mathrm{P}_0 = \mathrm{P}_0', \qquad \mathrm{P}_2(\theta, \bar\theta) = \mathrm{P}_2'(\bar\theta, \theta).$$

8. Conditions for stabilizability and detectability. Here we give simple equivalent conditions for stabilizability and detectability, the conditions under which our solution to the control problem of § 7 applies. Similar conditions have also been obtained by Olbrot [19] from a very different point of view.

First recall that the adjoint $\mathscr{A}^*$ of $\mathscr{A}$ is the infinitesimal generator associated with the transposed delay equation (4.2). Also note that the adjoints $\mathscr{B}^*$ and $\mathscr{C}^*$ of the operators $\mathscr{B}$ and $\mathscr{C}$ introduced in § 7 are given by

$$\mathscr{B}^*(h^0, h^1(\cdot)) = \mathrm{B}_0'h^0 + \int_{-b}^0 \mathrm{B}_{01}'(\theta)h^1(\theta)\, d\theta,$$

$$\mathscr{C}^*y = (\mathrm{C}'y, 0).$$

We have:

PROPOSITION 8.1. *Define*

$$\Delta(\lambda) = \sum_{i=0}^{k} A_i e^{\lambda\theta_i} + \int_{-b}^{0} A_{01}(\theta)\, e^{\lambda\theta}\, d\theta - \lambda I,$$

and define Σ *to be the set of zeros of* $\Delta(\lambda)$ *in the closed right halfplane. Then* Σ *is known to be a finite set. We have*:

(i) *The system* (7.1) *is stabilizable if and only if*

$$\text{Rank}\left[\Delta(\lambda) \quad B_0 + \int_{-b}^{0} B_{01}(\theta)\, e^{\lambda\theta}\, d\theta\right] = n \quad \text{for all } \lambda \in \Sigma. \tag{8.1}$$

(ii) *The system* (7.1) *is detectable if and only if*

$$\text{Rank}\begin{bmatrix}\Delta(\lambda)\\ C\end{bmatrix} = n \quad \text{for all } \lambda \in \Sigma. \tag{8.2}$$

Proof. Taking transposes, we see that condition (8.1) is equivalent to the condition

$$\text{Rank}\begin{bmatrix}\Delta'(\lambda)\\ B_0' + \int_{-b}^{0} B_{01}'(\theta)\, e^{\lambda\theta}\, d\theta\end{bmatrix} = n \quad \text{for all } \lambda \in \Sigma, \tag{8.3}$$

where

$$\Delta'(\lambda) = \sum_{i=0}^{k} A_i' e^{\lambda\theta_i} + \int_{-b}^{0} A_{01}'(\theta)\, e^{\lambda\theta}\, d\theta - \lambda I.$$

But it is known that (8.3) is equivalent to the condition that the transposed delay system

$$\begin{aligned}\frac{dx}{dt} &= \mathscr{L}'(x_t(\cdot)),\\ y(t) &= \mathscr{B}^*(x(t), x_t(\cdot)),\end{aligned} \tag{8.4}$$

is detectable. (Actually, the proofs in [20], [3] are concerned only with the case $B_{01}'(\cdot) = 0$, but the arguments adapt easily to the case where $B_{01}'(\cdot) \neq 0$.) This implies that there exists an operator $L \in \mathscr{L}(\mathbb{R}^m, \mathbb{R}^n \times L^2)$ such that $\mathscr{A}^* + L\mathscr{B}^*$ generates an exponentially stable semigroup on $\mathbb{R}^n \times L^2$. Since $\mathbb{R}^n \times L^2$ is reflexive, this last condition is equivalent to the condition: there exists $G \in \mathscr{L}(\mathbb{R}^n \times L^2, \mathbb{R}^m)$ such that $\mathscr{A} + \mathscr{B}G$ generates an exponentially stable semigroup (see, e.g., [22]). But such a G exists if and only if condition (i)′ of Proposition 7.2 is satisfied (this may be deduced from [6]). By Proposition 7.2, then, (8.1) is equivalent to stabilizability of the system (7.1).

The equivalence of condition (8.2) to detectability has been shown in [20], [3].

9. Comparison with previously known results, when $B_{01}(\cdot) = 0$. Since our results obviously also apply to the case where there are no delays in the control, i.e., $B_{01}(\cdot) = 0$, a natural question to ask is: how do the results presented in this paper, when specialized to the case where $B_{01}(\cdot) = 0$, compare with previously known results [8]? We shall show that our results are in a sense slightly stronger than the known results.

Suppose then that $B_{01}(\cdot) = 0$. Under the hypothesis of Theorem 7.3, the standard theory [5, Chapter 4] gives the solution of the infinite time quadratic cost control problem as

$$u(t) = -\mathscr{B}^* \tilde{\mathscr{P}}(x(t), x_t),$$

where $\tilde{\mathcal{P}} \in \mathcal{L}(\mathbb{R}^n \times L^2)$ is the unique nonnegative, self-adjoint operator satisfying

$$(9.1)\quad \langle\langle \tilde{\mathcal{A}}\tilde{x}, \tilde{\mathcal{P}}\tilde{y}\rangle\rangle + \langle\langle \tilde{\mathcal{P}}\tilde{x}, \tilde{\mathcal{A}}\tilde{y}\rangle\rangle + \langle\langle \tilde{x}, [\mathcal{C}^*\mathcal{C} - \tilde{\mathcal{P}}\mathcal{B}\mathcal{B}^*\tilde{\mathcal{P}}]\tilde{y}\rangle\rangle = 0 \quad \text{for all } \tilde{x}, \tilde{y} \in \mathcal{D}\{\tilde{\mathcal{A}}\}.$$

Equation (9.1) is the same as equation (7.2) with the exception that $\mathcal{A}$ is replaced by the linear operator $\tilde{\mathcal{A}}$ on $\mathbb{R}^n \times L^2$: $\tilde{\mathcal{A}}(\phi^0, \phi^1) = (\mathcal{L}(\phi^1), (d\phi^1/d\theta)(\theta))$, $\mathcal{D}\{\tilde{\mathcal{A}}\} = \{(\phi(0), \phi(\cdot))\colon \phi(\cdot) \in W^{1,2}\}$.

It is known that

$$(9.2)\quad \operatorname*{strong\,limit}_{T\to\infty} \tilde{\mathcal{P}}_T(0) = \tilde{\mathcal{P}},$$

where $\tilde{\mathcal{P}}_T(\cdot)\colon [0, T] \to \mathcal{L}(\mathbb{R}^n \times L^2)$ is the unique solution of the operator differential Riccati equation

$$-\frac{d}{dt}\langle\langle \tilde{x}, \tilde{\mathcal{P}}_T(t)\tilde{y}\rangle\rangle = \langle\langle \tilde{\mathcal{A}}\tilde{x}, \tilde{\mathcal{P}}_T(t)\tilde{y}\rangle\rangle + \langle\langle \tilde{\mathcal{P}}_T(t)\tilde{x}, \tilde{\mathcal{A}}\tilde{y}\rangle\rangle + \langle\langle \tilde{x}, (\mathcal{C}^*\mathcal{C} - \tilde{\mathcal{P}}_T(t)\mathcal{B}\mathcal{B}^*\tilde{\mathcal{P}}_T(t))\tilde{y}\rangle\rangle,$$

$$(9.3)\quad \text{for all } \tilde{x}, \tilde{y} \in \mathcal{D}\{\tilde{\mathcal{A}}\},$$

$$\tilde{\mathcal{P}}_T(T) = 0,$$

in the class of strongly continuous functions with the property that $\langle\langle \tilde{x}, \tilde{\mathcal{P}}_T(\cdot)\tilde{y}\rangle\rangle$ is absolutely continuous for all $\tilde{x}, \tilde{y} \in \mathcal{D}\{\tilde{\mathcal{A}}\}$. It is also known that

$$(9.4)\quad \langle\langle (\xi^0, \xi^1), \tilde{\mathcal{P}}(\xi^0, \xi^1)\rangle\rangle = \min \int_0^\infty \{x(t)C'Cx(t) + u'(t)u(t)\}\, dt,$$

$$(9.5)\quad \langle\langle (\xi^0, \xi^1), \tilde{\mathcal{P}}_T(t)(\xi^0, \xi^1)\rangle\rangle = \min \int_0^{T-t} \{x'(s)C'Cx(s) + u'(s)u(s)\}\, ds.$$

The relationships between $\mathcal{P}$ and $\tilde{\mathcal{P}}$ and between $\mathcal{P}_T(t)$ and $\tilde{\mathcal{P}}_T(t)$ are now evident from (7.4) and (9.4), (7.7) and (9.5). Since these identities apply for arbitrary initial conditions we conclude that the operators $\tilde{\mathcal{P}}_T(t)$ and $\tilde{\mathcal{P}}$ may be factored as follows:

$$(9.6)\quad \tilde{\mathcal{P}}_T(t) = M^*\mathcal{P}_T(t)M,$$

$$(9.7)\quad \tilde{\mathcal{P}} = M^*\mathcal{P}M.$$

That $\tilde{\mathcal{P}}_T(t)$ may be factored as shown has been observed in [9]. But the factorization (9.7) is apparently new, as is property (7.5), which we now interpret as stating that the middle term in the factorization of $\tilde{\mathcal{P}}_T(0)$ converges strongly to the middle term in the factorization of $\tilde{\mathcal{P}}$; this is a stronger property than (9.2) since M is not necessarily continuously invertible.

We conclude the section by pointing out that our results relating $\mathcal{P}$ and $\tilde{\mathcal{P}}$ through the factorization of $\tilde{\mathcal{P}}$ clarify the relationship between two different characterizations of the error covariance in the filtering problem for stochastic delay systems studied by Vinter [23], and Kwong and Willsky [14]. In this problem the error covariance function $P(t, \cdot, \cdot)$ is defined to be

$$P(t, \theta, \bar{\theta}) = E\{e(t+\theta|t)\, e'(t+\bar{\theta}|t)\}, \qquad -b \leqq \theta, \quad \bar{\theta} \leqq 0,$$

where $e(t+\theta|t)$ is the estimation error at $t+\theta$, given observations up to t. We define the "error covariance operator" $\Pi(t)$ on $\mathbb{R}^n \times L^2$ as

$$\Pi(t) = \begin{bmatrix} \Pi_0(t) & \Pi_1^*(t) \\ \Pi_1(t) & \Pi_2(t) \end{bmatrix},$$

with $\Pi_0(t) = P(t, 0, 0)$, $\Pi_1(t) = P(t, \cdot, 0)$, $\Pi_2(t) = P(t, \cdot, \cdot)$. Under stabilizability and detectability hypotheses, Vinter [23] has shown, using techniques of "infinite dimensional filtering", that

$$\operatorname*{strong\ limit}_{t\to\infty} \Pi(t) = \Pi,$$

where Π is the steady state error covariance operator, and Π is the unique nonnegative, symmetric solution to an operator Riccati equation very similar to (7.2). On the other hand, in the special case when $A_2, \cdots, A_k$, $A_{01}(\cdot) = 0$, Kwong and Willsky [14] have shown by more direct arguments that

$$\operatorname*{strong\ limit}_{t\to\infty} \tilde{M}^*\Pi(t)\tilde{M} = \Sigma,$$

for some operator Σ which satisfies a Riccati-like equation very similar to (9.1). Here $\tilde{M}$ is an operator of similar structure to the operator M in § 5.

It is clear now what is going on. The hypotheses ensure not only:

(i) asymptotic convergence of the solution to the operator differential Riccati equation (see (9.2)),

but the stronger property:

(ii) asymptotic convergence of the middle term in the factorization of the solution.

In the filtering context, the error covariance operator emerges as the middle term in the factorization; it is hardly surprising then that Vinter [23], by using the stronger property (ii), obtains convergence of the error covariance operator $\Pi(t)$, whereas Kwong and Willsky [14], by using the weaker property (i), obtain merely convergence of $\tilde{M}^*\Pi(t)\tilde{M}$. We may also remark that the "kernels" associated with the steady state error covariance operator Π satisfy equations very similar to those given in Theorem 7.4. They may be formally obtained from the error covariance equations for $P(t, \theta, \bar{\theta})$ given in [14] by setting all derivatives with respect to t to zero.

10. Some concluding remarks. In § 4, we produced out of the blue an evolution equation with certain useful properties, and gave no indication of its origin. Our choice was motivated by the following ideas.

It was shown in [23] how a stochastic filtering problem involving delays in the observations may be studied using an evolution equation approach. Since associated with an infinite dimensional filtering problem is an equivalent control problem in infinite dimensions, we may therefore associate the filtering problem with delays in the observations with such an equivalent control problem, which we shall call the infinite dimensional dual control problem. (Equivalence is understood in the sense that a solution to one problem gives a solution to the other.) This infinite dimensional dual control problem has "no delays in the control". Now the same filtering problem, without being first reformulated into an evolution equation framework, is also equivalent to a control problem with "delays in the control", which we term the natural dual problem. This has been shown by Lindquist [17]. It turns out that the natural dual control problem involves the delay differential equation with delays in the control discussed in § 3, while the infinite dimensional dual control problem involves an evolution equation as given in § 4. This suggests that the evolution equation is in some sense equivalent to the delay differential equation. The precise equivalence has now been established in § 5.

Our results apply only when point delays in the control are absent. When point delays are present, we may formally write the equivalent evolution equation as

$$\frac{d\tilde{x}(t)}{dt} = \mathcal{A}\tilde{x}(t) + \tilde{\mathcal{B}}u(t), \tag{10.1}$$

where $\tilde{\mathcal{B}}$ now maps into some space of distributions which strictly contains the state space $\mathbb{R}^n \times L^2$. We may proceed to define mild solutions to (10.1) as in [12] using appropriate extension arguments. The advantage of doing so is that we employ a state space, $\mathbb{R}^n \times L^2$, which is smaller than the state space of the evolution equation in [12].

REFERENCES

[1] H. T. BANKS AND J. A. BURNS, *Hereditary control problems: numerical methods based on averaging approximations*, this Journal, 16 (1978), pp. 169–208.

[2] ——— *An abstract framework for approximate solutions to optimal control problems governed by hereditary systems*, Proceedings, International Conference on Differential Equations (University of Southern California, September 1974), H. A. Antosiewicz, ed., Academic Press, New York, 1975, pp. 10–25.

[3] K. P. M. BHAT AND H. N. KOIVO, *Modal characterization of controllability and observability for time delay systems*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 292–293.

[4] P. L. BUTZER AND H. BERENS, *Semigroups of Operators and Approximation*, Springer-Verlag, New York, 1967.

[5] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite dimensional linear systems theory*, in Lecture Notes in Control and Information Sciences, Vol. 8, A. V. Balakrishnan and M. Thoma, eds., Springer, Berlin, 1978.

[6] R. DATKO, *A linear control problem in an abstract Hilbert space*, J. Differential Eq., 9 (1971), pp. 346–359.

[7] M. Ç. DELFOUR AND S. K. MITTER, *Hereditary differential systems with constant delays, II—a class of affine systems and the adjoint problem*, J. Differential Eq., 18 (1975), pp. 18–28.

[8] M. C. DELFOUR, C. MCCALLA AND S. K. MITTER, *Stability and the infinite time quadratic cost problem for linear hereditary differential systems*, SIAM J. Control, 13 (1975), pp. 48–88.

[9] M. C. DELFOUR, E. B. LEE AND A. MANITIUS, *F-reduction of the operator Riccati equation for hereditary differential systems*, Automatica, 14 (1978), pp. 385–395.

[10] M. C. DELFOUR AND A. MANITIUS, *Control systems with delays: areas of applications and present status of the linear theory*, in New Trends in Systems Analysis, A. Bensoussan and J. L. Lions, eds., Springer-Verlag, New York, 1978, pp. 420–437.

[11] ——— *The structural operator F and its role in the theory of retarded systems, I and II*, J. Math. Anal. Appl., to appear.

[12] A. ICHIKAWA, *Optimal control and filtering of evolution equations with delay in control and observation*, Report 53, Control Theory Centre, University of Warwick, Coventry, England, June 1977.

[13] H. N. KOIVO AND E. B. LEE, *Controller synthesis for linear systems with retarded state and control variables and quadratic cost*, Automatica, 8 (1972), pp. 203–208.

[14] R. H. KWONG AND A. S. WILLSKY, *Estimation and filter stability of stochastic delay systems*, this Journal, 16 (1978), pp. 660–681.

[15] R. H. KWONG, *A stability theory for the linear-quadratic-Gaussian problem for systems with delays in the state, control, and observations*, this Journal, 18 (1980), pp. 49–75.

[16] ——— *Characterization of kernel functions associated with operator algebraic Riccati equations for linear delay systems*, Systems Control report no. 7906, University of Toronto, Toronto, Canada, June 1979.

[17] A. LINDQUIST, *A theorem on duality between estimation and control for linear stochastic systems with time delay*, J. Math. Anal. Appl., 37 (1972), pp. 516–536.

[18] A. MANITIUS, *Controllability, observability and stabilizability of retarded systems*, Proceedings, 1976 IEEE Conference on Decision and Control, Clearwater, Florida, pp. 752–758.

[19] A. W. OLBROT, *Stabilizability, detectability and spectrum assignment for linear autonomous systems with general time delays*, IEEE Trans. Automat. Control, AC-23 (1978), pp. 887–890.

[20] L. PANDOLFI, *On feedback stabilization of functional differential equations*, Boll. Un. Mat. Ital., 4, 11, Supplements al fascicolo 3, Giugno, Serie IV, Vol. XI (1975), pp. 626–635.

[21] R. B. VINTER, *On the evolution of the state of linear differential delay equations in M^2: properties of the generator*, J. Inst. Math. Appl., 21 (1978), pp. 13–23.

[22] ——— *Stabilizability and semigroups with discrete generators*, J. Inst. Math. Appl., 20 (1977), pp. 371–378.

[23] ——— *Filter stability for stochastic evolution equations*, this Journal, 15 (1977), pp. 465–485.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 1, January 1981

0363-0129/81/1901-0012 $01.00/0

INVARIANT STRUCTURES OF GENERAL DYNAMICAL SYSTEMS*

TOSHIO NOMURA† AND KATSUHISA FURUTA‡

Abstract. In the geometric approach to linear multivariable control theory, A-invariant subspaces and (A,B)-invariant subspaces have key roles in the theoretical setting. However, it is not easy to generalize these concepts to more general dynamical systems because of the nontrivial interaction between input and state variables. This paper aims at providing a conceptual framework for several invariant structures of general dynamical systems in a set-theoretical approach, and at clarifying their system-theoretical significance. Roughly speaking, an invariant structure of a dynamical system is defined as an equivalence relation of the state set whose equivalence classes are carried invariantly by the state transition function. Then, it is proved that there exists a unique maximal indistinguishable structure realized by state feedback, and a simple application of this property is performed to formally characterize the disturbance localization problem.

1. Introduction. Recently, many theoretical studies have been devoted to nonlinear dynamical systems, using modern mathematical tools which try to extend ideas in linear multivariable control theory. They are concerned with minimal realization problems of nonlinear input–output functions, and with structural properties such as controllability and observability by differential geometric methods. However, as far as feedback control problems are concerned, there has not been much systematic research reported (Brockett [1]). One of the reasons for this is the lack of useful concepts to deal with them, as well as their mathematical difficulty.

In the modern feedback control theory of linear multivariable systems, several methods such as linear algebraic, abstract algebraic and geometric approaches have been used and have produced a number of successful results. Among them, the geometric approach [7] has provided several key concepts which have enabled formal characterizations of synthesis problems of linear systems to become more comprehensive. However, it is not easy to apply the geometric concepts to more general (nonlinear) dynamical systems, because the system-theoretic meanings of the concepts have not been defined specifically enough to be generalized.

The purpose of the paper is to develop several ideas relating the invariant subspaces of the geometric approach in a set-theoretic (formal) setting; the abstract version of A-invariant subspaces and (A,B)-invariant subspaces and their relation to state feedback actions are discussed. In [3], Ishijima attempted a generalization of A-invariant subspaces and (A,B)-invariant subspaces by the Lie algebraic method, but the system-theoretic significance was not made clear.

Because of the fact that the systems are not linear, we have to be concerned with all the points of the state set of the system in order to consider an invariant structure. Consequently, an invariant structure is defined as an equivalence relation (or a partition) of the state set whose equivalence classes are carried invariantly by the state transition function either for every input or for zero input. As one of the main results of the paper, we prove the existence of a unique maximal indistinguishable structure realized by state feedback, which is a generalization of the linear theory. In the last section of the paper, we consider the disturbance localization problem as a prominent use of invariant structures.

* Received by the editors May 14, 1979, and in revised form April 16, 1980. A preliminary version of this paper was presented at the Third International Symposium on Mathematical Theory of Networks and Systems, Delft, the Netherlands, 1979; see Ref. [6].

† Department of Systems Sciences, Tokyo Institute of Technology, Ohokayama, Meguro-ku, Tokyo, Japan.

‡ Department of Control Engineering, Tokyo Institute of Technology, Ohokayama, Meguro-ku, Tokyo, Japan.

In [4], Liepa and Wonham discuss the algebraic formulation of the solvability of the disturbance containment and decoupling problems in a lattice-theoretic approach. Theorem 6.2 of the present paper overlaps with one of their results (3.07), which is a characterization of the solvability of the disturbance decoupling problem by state feedback. However, the main purpose of this paper is to introduce and develop generalized notions of A-invariant subspaces and (A,B)-invariant subspaces, which the authors of [4] have not handled fully.

2. Basic notations and development. In this section, we explain basic notations of dynamical system descriptions and of partitions of a state set. Then by introducing additional properties such as a partial ordering and a binary operation between partitions, we see that a set of partitions of the state set has the structure of a lattice-ordered set. This structure is useful when we consider inclusion relations between invariant structures defined by partitions in the later section of the paper.

Let T denote a time scale which is an additive group and a well-ordered set with a smallest element 0, where a partially ordered set is said to be well-ordered if every nonempty subset of it has a smallest element.

Remark. This time set T contains at least discrete time but not real time. Intuitively, the structure of a total order is more suitable for the time scale, which implies a well-ordered set. However, we do not have more powerful mathematical tools than transfinite induction with which to prove the existence of the feedback loop systems, etc., which will be treated in the later sections.

Let A be the input alphabet with a zero element, and B be the output alphabet. Let U be an input set which is a set of time functions $A^T = \{u | u: T \to A\}$, and let X be a state set and Y be an output set which is a subset of time functions $B^T = \{y | y: T \to B\}$. Let $T_{tt'} = \{\tau | \tau \in [t, t')\}$, $T_t = \{\tau | \tau \geqq t\}$ and $T^t = \{\tau | \tau \in [0, t)\}$; adopting the same idea for $u \in U$, let $u_{tt'} = u|T_{tt'}$ and $u^t = u|T^t$, where $u|T_{tt'}$ is a restriction of u to $T_{tt'}$, and the like. The concatenation of inputs is denoted by $u_{tt'} \circ u'_{t't''}$.

Next, we explain the description of a strongly causal, time-invariant dynamical system [5] which is a pair of functions ϕ and λ. Namely, ϕ is a state transition function defined for $t, t' \in T$ and $u \in U$, by

$$\begin{aligned} &\phi_{tt'}: X \times U_{tt'} \to X, \\ &\qquad (x_0, u_{tt'}) \mapsto x(t') = \phi_{tt'}(x_0, u_{tt'}), \end{aligned} \tag{2.1}$$

and an output function λ defined by

$$\begin{aligned} &\lambda: X \to B, \\ &\qquad x(t) \mapsto y(t) = \lambda(x(t)). \end{aligned} \tag{2.2}$$

Thus ϕ is a mapping which sends a state x_0 at t to a new state $x(t')$ at t' by applying an input $u_{tt'}$. As is usual, ϕ must satisfy the semigroup property,

$$\phi_{tt''}(x_0, u_{tt'} \circ u'_{t't''}) = \phi_{t't''}(\phi_{tt'}(x_0, u_{tt'}), u'_{t't''}).$$

Since we will be concerned only with time-invariant systems, we choose the initial time to be 0. The state transition function ϕ is usually referred to by ϕ_{0t} for $t \in T$.

Let us consider a set of partitions defined on the state set X and develop additional structures for it.

DEFINITION 2.1. A *partition* of X is denoted by $S = \{S_i | i \in I\}$, where I is an index set, satisfying

(i) $X = \bigcup_{i \in I} S_i, \quad S_i \neq \varnothing$

(ii) $(\forall i, j \in I) \quad (i \neq j \Rightarrow S_i \cap S_j = \varnothing)$.

Naturally, a partition S defines a mapping which tells where each $x \in X$ belongs, by

(i) $$S: X \to \mathcal{P}(X),$$

(2.3) $$x \mapsto S(x) = S_i \quad \text{for some } i \in I,$$

where $\mathcal{P}(X)$ is the power set of X, and it also defines an equivalence relation on X as a subset of $X \times X$, by

(ii) for x_1 and x_2 in X,

(2.4) $$(x_1, x_2) \in S \quad \Leftrightarrow \quad S(x_1) = S(x_2).$$

Then we write

(2.5) $$\Sigma = \{S | S \text{ is a partition of } X\}.$$

If we define an order relation $\prec$ between partitions, Σ becomes a complete lattice-ordered set.

DEFINITION 2.2. For partitions S and T in Σ,

(2.6) $$S \prec T \quad \Leftrightarrow \quad S \subset T,$$

where $\subset$ denotes a set inclusion on $X \times X$. In other words, for $S = \{S_i | i \in I\}$ and $T = \{T_j | j \in J\}$ in Σ,

(2.6)′ $$S \prec T \quad \Leftrightarrow \quad (\forall i \in I)\,(\exists j \in J)\,(S_i \subset T_j),$$

where $\subset$ denotes a set inclusion on X.

PROPOSITION 2.1. *Σ is a complete lattice-ordered set with respect to the order relation $\prec$.*

Proof. It is obvious from Definition 2.2. □

Next, we introduce a binary operation $\vee$ for Σ which gives a concrete construction of joints of partitions.

DEFINITION 2.3. For S and T in Σ, a mapping

(2.7) $$S \vee T = R: X \to \mathcal{P}(X)$$

is defined for any $x \in X$ by

$$R_1(x) = S(x) \cup T(x),$$

$$R_i(x) = S(R_{i-1}(x)) \cup T(R_{i-1}(x)),$$

$$R(x) = \bigcup_{i>0} R_i(x),$$

where $S(R_{i-1}(x)) = \bigcup_{y \in R_{i-1}(x)} S(y)$, and so forth.

As a mapping: $X \to \mathcal{P}(X)$, $S \vee T$ is well defined. Furthermore, the following proposition guarantees that it is also a partition on X.

PROPOSITION 2.2. *The mapping $S \vee T$ in Definition 2.3 defines a partition on X.*

Proof. Since the relation $T(S(x)) \supset T(x)$ is always satisfied for any $x \in X$ from Definition 2.3, the following relations hold:

(2.8) $$X = \bigcup_{x \in X} S(x) \subset \bigcup_{x \in X} T(S(x)) \subset \bigcup_{x \in X} (S \vee T)(x) \subset X,$$

which implies that $\{(S \vee T)(x) | x \in X\}$ is a covering of X (condition i) of Definition 2.1). Therefore, in order to prove that the mapping $S \vee T = R: X \to \mathcal{P}(X)$ defines a partition, we need only show that

(2.9) $$(\forall x_1, x_2 \in X)\,(R(x_1) \cap R(x_2) \neq \varnothing \Rightarrow R(x_1) = R(x_2)),$$

which is equivalent to saying that

$$(\forall x_1, x_2 \in X)\ (x_1 \in R(x_2) \Rightarrow R(x_1) = R(x_2)). \tag{2.10}$$

Notice first that

$$R^2 = R, \tag{2.11}$$

as mappings: $X \to \mathcal{P}(X)$. In fact, for any $x \in X$,

$$R(R(x)) \supset R(x). \tag{2.12}$$

Also, if $x \in R(R(x_0))$ for x and x_0 in X, there exists $z \in R(x_0)$ such that $x \in R(z)$. According to the definition of R, there exist sets of mappings: $X \to \mathcal{P}(X)$, $R_i (i = 1, \cdots, n)$ and $R'_j (j = 1, \cdots, m)$ where R_i and R'_j are either S or T, such that $x \in R_1(\cdots(R_n(z)\cdots)$ and $z \in R'_1(\cdots(R'_m(x_0)\cdots)$. Therefore,

$$x \in R_1(\cdots(R_n(R'_1(\cdots(R'_m(x_0)\cdots))))) \subset R(x_0). \tag{2.13}$$

Hence, from (2.12) and (2.13), equality (2.11) is shown.

Let us prove (2.10) now. By applying R to both sides of $x_1 \in R(x_2)$, we obtain

$$R(x_1) \subset R(R(x_2)) = R^2(x_2) = R(x_2), \tag{2.14}$$

because of (2.11). Conversely, for any $z \in R(x_2)$ and since $x_1 \in R(x_2)$, there exist sets of mappings: $X \to \mathcal{P}(X)$, $R_i (i = 1, \cdots, n)$ and $R'_j (j = 1, \cdots, m)$ where R_i and R'_j are either S or T, such that the following hold:

$$z \in R_1(\cdots(R_n(x_2)\cdots)), \tag{2.15}$$

$$x_1 \in R'_1(\cdots(R'_m(x_2)\cdots)). \tag{2.16}$$

Since R'_j are equivalence relations, which implies that $y_1 \in R'_j(y_2)$ iff $y_2 \in R'_j(y_1)$, (2.16) yields

$$x_2 \in R'_m(\cdots(R'_1(x_1)\cdots)). \tag{2.17}$$

By combining (2.15) and (2.17), we obtain

$$z \in R_1(\cdots(R_n(R'_m(\cdots(R'_1(x_1)\cdots))))) \subset R(x_1). \tag{2.18}$$

Hence, (2.10) is proved from (2.14) and (2.18). □

The ordering relation $\prec$ and the binary operation $\vee$ are both useful for proving the existence of a (unique) maximal partition which satisfies certain additional properties. The following definition of a special kind of partition is given in order to deal with linear systems.

DEFINITION 2.4. Let X be a linear space and L be a linear subspace of X. Then a partition S is called *affine* if S consists of all affine subspaces of X whose standard linear space is L, i.e.,

$$S = \{L + a \mid a \in X\}, \tag{2.19}$$

and, for x_1 and x_2 in X,

$$(x_1, x_2) \in S \quad \Leftrightarrow \quad (x_1 - x_2) \in L. \tag{2.20}$$

3. State feedbacks and invariant structures. In this section, we shall relate the set of partitions Σ to the state transition function ϕ, the output function λ, and state feedbacks. When we are concerned with invariant structures of linear systems because of the linearity between input and state variables in the state transition function, we will

consider only subspaces of the state set and may not take into account whether external input is being applied or not. However, in the case of more general (nonlinear) systems, owing to the nontrivial interaction between input and state variables, we are compelled to deal with both instances depending on whether the external input is available or not. Thus, we introduce several definitions of invariant structures accordingly. In the last part of this section, we verify whether the concepts of invariant structures thus defined are actually generalizing those of A-invariant subspaces and (A,B)-invariant subspaces in linear systems. Interrelations of those structures in general systems are discussed in the following sections.

The following is a formal expression of applying a state feedback to a state transition function. A state feedback f is described by a mapping

$$f: X \times A_1 \to A, \tag{3.1}$$

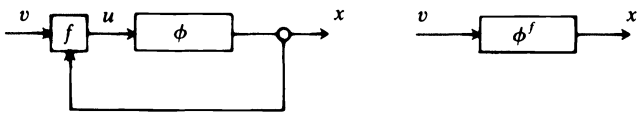
where A_1 is another input alphabet which is isomorphic to A.(Generally speaking it is not necessary to assume that A_1 is isomorphic to A. However, to make it easier, we make this assumption.) In particular, if, for any x in X, the restricted mapping

$$f(x, \cdot): A_1 \to A \tag{3.2}$$

is invertible, f is called a regular state feedback. Also, a mapping $\tilde{f}$ is a natural extension of f which is defined by

$$\begin{aligned} &\tilde{f}: X^T \times V \to U, \\ &(x, v) \mapsto u = \tilde{f}(x, v), \\ &u(t) = \tilde{f}(x, v)(t) \overset{d}{=} f(x(t), v(t)), \end{aligned} \tag{3.3}$$

where V is a set of time functions $A_1^T = \{v | v: T \to A_1\}$. For a given pair (ϕ, f), is it possible to formally define a mapping ϕ^f which is a state transition function after the state feedback f is applied to the state transition function ϕ? (see Figs. 1a and 1b). This



a FIG. 1 b

question can be answered affirmatively, since our time scale is a well-ordered set. (We don't know the answer for a more general time scale in a set-theoretical approach.) First let us summarize the theorem of transfinite induction in our setting (see [2]).

LEMMA. *Let T be a well-ordered set and $P(t)$ be a propositional function with a free variable $t \in T$. If the following statement is true for any $t \in T$, then $P(t)$ is true for all $t \in T$:*
"If $P(\tau)$ is true for any $\tau < t$, then $P(t)$ is true." □

LEMMA 3.1. *For ϕ and f as above, there exists a well-defined mapping ϕ^f which is a state transition function after f is applied to ϕ.*

Proof. Since the time scale T is well ordered, we can use transfinite induction. For any $t \in T$, $\tau \in [0, t)$, $v \in V$ and $x_0 \in X$, let us assume that $\phi^f_{0\tau}: X \times V_{0\tau} \to X$ is well defined. Let us define its solution $x^f_{0t} \in X_{0t}$ by $x^f_{0t}(\tau) = \phi^f_{0\tau}(x_0, v_{0\tau})$, and generate an input $u_{0t} \in U_{0t}$ by $u_{0t}(\tau) = f(x^f_{0t}(\tau), v_{0t}(\tau))$; that is, $u_{0t} = \tilde{f}(x^f_{0t}, v_{0t})$. Since originally ϕ is strongly causal (that is, $(\forall t)(\forall u^1, u^2 \in U)(\forall x_0)(u^1|T_{0t} = u^2|T_{0t} \Rightarrow \phi_{0t}(x_0, u^1_{0t}) =$

$\phi_{0t}(x_0, u_{0t}^2)))$ it does depend only on past input. Therefore, the existence of ϕ_{0t}^f is proved by transfinite induction; it is defined by

$$\phi_{0t}^f(x_0, v_{0t}) = \phi_{0t}(x_0, u_{0t}) = \phi_{0t}(x_0, \tilde{f}(x_{0t}^f, v_{0t})). \tag{3.4}$$

The mapping ϕ^f is a typical case of the definition of the transfinite recursion theorem. □

In the following, we give definitions of several invariant structures.

DEFINITION 3.1.

(i) A partition S is *invariant* w.r.t. ϕ (or ϕ-invariant, or simply invariant if no confusion occurs) if

$$(\forall x_1, x_2 \in X)(\forall u \in U)(\forall t \in T) \quad ((x_1, x_2) \in S \Rightarrow (\phi_{0t}(x_1, u_{0t}), \phi_{0t}(x_2, u_{0t})) \in S). \tag{3.5}$$

(ii) A partition S is *feedback invariant* w.r.t. ϕ if there exists a regular state feedback f such that S is invariant w.r.t. ϕ^f.

In the usual terms (3.5) says that, for any x_1 and x_2 in X, if x_1 and x_2 are in an equivalence class, then the pair of orbits from x_1 and x_2 for any u in U stays in the equivalence relation for any t in T. This implies that ϕ-invariant S is a congruence relation on X with respect to $\phi_{0t}(\cdot, u_{0t})$ for any t and u.

Next, by adding an output mechanism λ into consideration, let us define more structured invariant structures.

DEFINITION 3.2.

(i) A partition S is *invariant* w.r.t. (ϕ, λ) (or (ϕ, λ)-invariant) if S is invariant w.r.t. ϕ and, for any x_1 and x_2 in X, $(x_1, x_2) \in S$ implies $\lambda(x_1) = \lambda(x_2)$.

(ii) A partition S is *feedback invariant* w.r.t. (ϕ, λ) if there exists a regular state feedback f such that S is invariant w.r.t. (ϕ^f, λ).

When S is feedback invariant w.r.t. (ϕ, λ), it can also be said that S is an indistinguishable structure (by state feedback in a natural sense of terminology). Next, invariant structures which include those of Definition 3.1 are useful when closed loop systems with no external input are concerned.

DEFINITION 3.3.

(i) A partition S is *zero-invariant* w.r.t. ϕ if

$$(\forall x_1, x_2 \in X)(\forall t \in T)((x_1, x_2) \in S \Rightarrow (\phi_{0t}(x_1, 0), \phi_{0t}(x_2, 0)) \in S), \tag{3.6}$$

where $0(t)$ is the zero element of A for any t.

(ii) A partition S is *feedback zero-invariant* w.r.t. ϕ if there exists a state feedback f such that S is zero-invariant w.r.t. ϕ^f.

Remark. In the case of a feedback zero-invariant structure, since the external input is always zero, the state feedback is just a mapping: $X \to A$.

The invariant structures defined above are meant to be the generalizations and verifications of A-invariant subspaces and their relation to the state feedback. The following are expected to correspond to (A,B)-invariant subspaces.

DEFINITION 3.4.

(i) A partition S is *control invariant* w.r.t. ϕ (or control ϕ-invariant) if

$$(\forall x_1, x_2 \in X)(\forall u^1 \in U)(\exists u^2 \in U)(\forall t \in T) \tag{3.7}$$
$$((x_1, x_2) \in S \Rightarrow (\phi_{0t}(x_1, u_{0t}^1), \phi_{0t}(x_2, u_{0t}^2)) \in S).$$

(ii) A partition S is *uni-control invariant* w.r.t. ϕ if, for any $x \in X$, there exists $u^x \in U$ and, for any of (x_1, u^{x_1}) and (x_2, u^{x_2}) in a set of such pairs $\{(x, u^x)\}_{x \in X}$, $(x_1, x_2) \in S$ implies that $(\phi_{0t}(x_1, u_{0t}^{x_1}), \phi_{0t}(x_2, u_{0t}^{x_2})) \in S$ for any t.

In the usual terms, (3.7) says that, for any x_1 and x_2 in X, and u^1 in U, if x_1 and x_2 are in an equivalence class, then there exists an input u^2 in U which enables the pair of resulting orbits to stay in the equivalence relation for any t.

In the rest of this section, we consider a special case where the dynamical system (ϕ, λ) is a linear system and the partition S is affine. Therefore, ϕ is expressed by

$$\phi_{0t}(x_0, u_{0t}) = \phi_{0t}(x_0, 0) + \phi_{0t}(0, u_{0t}) = \phi^1_{0t}(x_0) + \phi^2_{0t}(u_{0t}), \tag{3.8}$$

where ϕ^1 and ϕ^2 are linear and the output function λ is linear. A state feedback f is expressed, for any t, by

$$u(t) = f(x(t), v(t)) = Fx(t) + Gv(t), \tag{3.9}$$

where F and G are matrices with $\det G \neq 0$. Let L be a linear subspace of a linear space X which is a standard linear space for S. Then, several invariant structures turn out to be mutually equivalent for linear systems as follows:

PROPOSITION 3.1. *If (ϕ, λ) is a linear system and S is affine with a standard space L, the following statements are equivalent*:

(i) *S in invariant w.r.t. ϕ.*

(ii) *S is zero-invariant w.r.t. ϕ.*

(iii) *L is A-invariant*;

$$(\overset{d}{\Leftrightarrow}(\forall x \in L)(\forall t \in T) \quad (\phi_{0t}(x, 0) = \phi^1_{0t}(x) \in L).)$$

Proof. (i)$\Rightarrow$(ii).

(i)$\Leftrightarrow$(ii). $(\forall x_1, x_2 \in X)(\forall u)(\forall t)((x_1, x_2) \in S \Rightarrow (\phi_{0t}(x_1, u_{0t}),\ \phi_{0t}(x_2, u_{0t})) \in S)$. Because of the linearity of ϕ, $\phi_{0t}(x_1, u_{0t}) - \phi_{0t}(x_2, u_{0t}) = \phi_{0t}(x_1, 0) - \phi_{0t}(x_2, 0) = \phi^1_{0t}(x_1) - \phi^1_{0t}(x_2) \in L$, which is equivalent to (ii).

(ii)$\Rightarrow$(iii). For any $x \in X$, $a \in X$ and $t \in T$, if $x \in L$, then it follows that $(\phi^1_{0t}(x + a), \phi^1_{0t}(a)) \in S$ since $(x + a, a) \in S$. Therefore, $\phi^1_{0t}(x + a) - \phi^1_{0t}(a) = \phi^1_{0t}(x) \in L$, which is (iii).

(iii)$\Rightarrow$(ii). For any x_1 and $x_2 \in X$ such that $(x_1, x_2) \in S$, there exist l_1 and l_2 such that $x_1 = l_1 + a$ and $x_2 = l_2 + a$ for some $a \in X$. Therefore, it follows that $\phi^1_{0t}(x_1) - \phi^1_{0t}(x_2) = \phi^1_{0t}(l_1 + a) - \phi^1_{0t}(l_2 + a) = \phi^1_{0t}(l_1 - l_2) \in L$, since $l_1 - l_2 \in L$, which is (ii). □

PROPOSITION 3.2. *If (ϕ, λ) is a linear system and S is affine with a standard linear space L, the following statements are equivalent*:

(i) *S is control invariant w.r.t. ϕ.*

(ii) *S is uni-control invariant w.r.t. ϕ.*

(iii) *L is (A,B)-invariant.*

$$(\overset{d}{\Leftrightarrow}(\forall x \in L)(\exists u \in U)(\forall t \in T)\ (\phi_{0t}(x, u_{0t}) \in L).)$$

Proof. (i)$\Rightarrow$(ii). This is obvious from the definitions.

(ii)$\Rightarrow$(iii). Since $(l + a, a) \in S$ for any $l \in L$ and $a \in X$, there exist u and u^1 in U such that $(\phi_{0t}(l + a, u_{0t}), \phi_{0t}(a, u^1_{0t})) \in S$ for any t. Hence, it follows that $\phi_{0t}(l + a, u_{0t}) - \phi_{0t}(a, u^1_{0t}) = \phi_{0t}(l, (u - u^1)_{0t}) \in L$. Therefore, for any $l \in L$, if we choose an input to be $u - u^1$ as above, the condition (iii) is satisfied.

(iii)$\Rightarrow$(i). For any x_1 and $x_2 \in X$ such that $(x_1, x_2) \in S$, there exist l_1 and l_2 which satisfy $x_1 = l + a$ and $x_2 = l_2 + a$. Since L is (A,B)-invariant, there exist $v^i \in U (i = 1, 2)$ such that $\phi_{0t}(l_i, v^i_{0t}) \in L$, for $i = 1, 2$. Therefore, for any x_1 and $x_2 \in X$, and any $u \in U$, if we choose an input to be $u + v^2 - v^1$, it follows that $\phi_{0t}(x_1, u_{0t}) - \phi_{0t}(x_2, (u + v^2 - v^1)_{0t}) = \phi_{0t}(l_1 + a, u_{0t}) - \phi_{0t}(l_2 + a, (u + v^2 - v^1)_{0t}) = \phi_{0t}(l_1, v^1_{0t}) - \phi_{0t}(l_2, v^2_{0t}) = l'_1 - l'_2 \in L$, where $l'_i = \phi_{0t}(l_i, v^i_{0t}) \in L$. □

Let us consider an example of linear systems which are described by

$$x(t+1) = Ax(t) + b_i u(t), \qquad t = 0, 1, \cdots, \quad i = 1, 2,$$

where $x \in R^2$, $u \in R^1$, $A = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$, $b_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $b_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. Let us take a subspace $L = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\}$ as an invariant subspace of the matrix A. Since the pair (A, b_1) is controllable, it is easy to see that, except for two trivial partitions, $\{R^2\}$ and $\{x | x \in R^2\}$, the only possible invariant structure including L is as shown in Fig. 2a, which is affine with the standard linear space L. However, since the pair (A, b_2) is not controllable, the structure of allowable invariant partitions is relaxed in the uncontrollable direction $\left\{ \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\}$, and Figs. 2b and 2c as well as Fig. 2a can be other examples of invariant structures for (A, b_2). Thus, we can see that some sort of controllability criteria are to be concerned with invariant structures, and that affine structures are not only the case for linear systems. We will not treat this kind of finer discussion here, but leave it for a later paper.

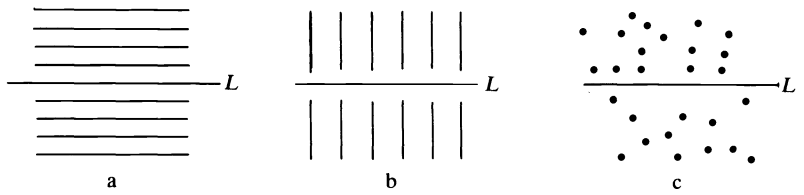


FIG. 2

4. Relations between invariant structures. I. In this section, we examine relations between feedback invariant structures and control invariant structures. Although we have not obtained conditions which specify when these two structures become equivalent as in linear systems, we prove the following.

THEOREM 4.1. *For a given dynamical system* (ϕ, λ), *there exists a unique maximal partition which is feedback invariant w.r.t.* (ϕ, λ).

Remark. In other words, this theorem guarantees the existence of a unique maximal indistinguishable structure for (ϕ, λ) attained by feedback.

In order to prove the theorem above, we need the following three lemmas, which are interesting by themselves.

LEMMA 4.1. *If a partition* S *is feedback invariant w.r.t.* ϕ, *then* S *is control invariant w.r.t.* ϕ.

Proof. Since S is feedback invariant w.r.t. ϕ, there exists a regular state feedback $f: X \times A_1 \to A$, where A_1 is isomorphic to A, and $f(x, \cdot): A_1 \to A$ is invertible for any $x \in X$. Let $x(t) = \phi_{0t}(x_1, u^1_{0t})$ for any $x_1 \in X$, $u^1 \in U$ and $t \in T$. Then, since f is regular, there exists $v_{0t} \in V_{0t}$ such that $u^1_{0t}(\tau) = f(x(\tau), v_{0t}(\tau))$ for any $\tau \in [0, t)$; that is, $\phi^f_{0t}(x_1, v_{0t}) = \phi_{0t}(x_1, u^1_{0t})$. For such v_0 let us define $u^2_{0t} \in U_{0t}$ by $u^2_{0t}(\tau) = f(\phi^f_{0\tau}(x_2, v_{0\tau}), v_{0t}(\tau))$. Then, because $(\phi^f_{0t}(x_1, v_{0t}), \phi^f_{0t}(x_2, v_{0t}) \in S$, it follows that $(\phi_{0t}(x_1, u^1_{0t}), \phi_{0t}(x_2, u^2_{0t})) \in S$. □

LEMMA 4.2. *If a partition* S *is control invariant w.r.t.* ϕ, *then there exists a state feedback* f *such that* S *is invariant w.r.t.* ϕ^f.

Remark. This lemma does not imply that the feedback f thus existing is regular (see the example at the end of this section.)

Proof. Let $S=\{S_i|i\in I\}$, and let $\{s_i\}_{i\in I}$ be a choice set of the family of sets $\{S_i\}_{i\in I}$, and $c:\{S_i\}_{i\in I}\to X$ be its choice function. Since S is control invariant, for any $x_0\in\{s_i\}_{i\in I}$ and $x\in S(x_0)$ and $u\in U$, there exists $u^1\in U$ such that $(\phi_{0t}(x_0, u_{0t}), \phi_{0t}(x, u^1_{0t}))\in S$ for any t. Let us denote this correspondence by $g_{S(x_0)}: S(x_0)\times U\to U$, $(x, u)\mapsto g_{S(x_0)}(x, u)=u_1$. Furthermore, by using the choice function c let us define a mapping $g: X\times U\to U$ by $(x, u)\mapsto g(x, u)=g_{S(c(x))}(x, u)$. Let a mapping $\mathrm{pr}: U\to A$ be a projection defined by $u\mapsto \mathrm{pr}(u)=u(0)$, and $(\mathrm{id}_X\times\mathrm{pr}): X\times U\to X\times A$ be defined by $(x, u)\mapsto (\mathrm{id}_X\times\mathrm{pr})(x, u)=(x, \mathrm{pr}(u))$. Let us define $g': X\times U\to A$ by $g'=\mathrm{pr}\circ g$, where $\circ$ means a composite mapping. Since $(\mathrm{id}_X\times\mathrm{pr})$ is surjective, and $g'(x, u)=g'(x, u')$ whenever $u(0)=u'(0)$ for any $x\in X$ and u and $u'\in U$, there exists a mapping $g_0: X\times A\to A$ which satisfies $g'=g_0\circ(\mathrm{id}_X\times\mathrm{pr})$. Meanwhile, let $h: A_1\to A$ be the isomorphism between two alphabets and let us define $(\mathrm{id}_X\times h): X\times A_1\to X\times A$ by $(x, a)\mapsto(x, h(a_1))$. At the end, let us define a state feedback $f: X\times A_1\to A$ by $f=g_0\circ(\mathrm{id}_X\times h)$. Now, we have to prove that S becomes actually ϕ^f-invariant by using the feedback f thus constructed. Let us define

$$\tau_0=\min_{\substack{(x_1,x_2)\in S\\ v\in V}}\{\tau\in T|(\phi^f_{0\tau}(x_1, v_{0\tau}), \phi^f_{0_\tau}(x_2, v_{0\tau}))\notin S\}, \tag{4.1}$$

which exists because T is well-ordered, and let the arguments be $\tilde{x}_1$, $\tilde{x}_2$ and $\tilde{v}$ which realize the minimum τ_0. If there exists an element τ_1 in $\{\tau|0\leqq\tau<\tau_0\}$ which differs from 0, then a set of arguments, $\phi_{0\tau_1}(\tilde{x}_1, \tilde{v}_{0\tau_1})$, $\phi_{0\tau_1}(\tilde{x}_2, \tilde{v}_{0\tau_1})$ and $\tilde{v}^{\tau_1}$, where $\tilde{v}^{\tau_1}$ is a shift of $\tilde{v}$ defined by $\tilde{v}^{\tau_1}(t)=\tilde{v}(t+\tau_1)$, realizes condition (4.1) with the shorter time $\tau_0-\tau_1$ which contradicts the definition of τ_0. Therefore, the set $\{\tau|0\leqq\tau<\tau_0\}$ contains only one element 0. If we set $\tilde{u}^i(0)=f(\tilde{x}_i, \tilde{v}_{0\tau_0}(0))$ for $i=1, 2$, it follows that $\tilde{x}_i(\tau_0)\overset{d}{=}\phi^f_{0_{\tau_0}}(\tilde{x}_i, \tilde{v}_{0_{\tau_0}})=\phi^f_{0_{\tau_0}}(\tilde{x}_i, \tilde{v}_{0_{\tau_0}}(0))=\phi_{0_{\tau_0}}(\tilde{x}_i, \tilde{u}^i(0))$ for $i=1, 2$, since there are no time elements in $\{\tau|0\leqq\tau<\tau\}$ except 0. However, from the construction method of f, $\tilde{u}^i(0)$ are such that $(\tilde{x}^1(\tau_0), \tilde{x}^2(\tau_0))\in S$, which violates the definition in (4.1). Therefore, there does not exist any time which satisfies (4.1). □

LEMMA 4.3. *If any two partitions S and T are feedback invariant w.r.t. ϕ (or (ϕ, λ)), then the partition $S\vee T$ is also feedback invariant w.r.t. ϕ (or (ϕ, λ)).*

Proof. The idea of the proof is to show that the feedback which can be constructed by the method of Lemma 4.2 for the partition $S\vee T$ becomes regular, by using the fact that $S\vee T$ is control invariant, and S and T are feedback invariant. Let $R=S\vee T$, and f_S and f_T be regular feedbacks for S and T, respectively; let $\{r_i\}_{i\in I}$ be a choice set of the partition R, and c be its choice function. Because of the construction method of R (i.e., Definition 2.3), for any $x_0\in\{r_i\}_{i\in I}$ and $x\in R(x_0)$ there exists an integer n such that it is possible to express

$$x\in R_1(\cdots(R_n(x_0)\cdots)) \tag{4.2}$$

for the proper choice of R_i, where $R_i(i=1, \ldots, n)$ are either S or T. Since $S^2=S$ and $T^2=T$ as mappings (see (2.11)), by subtracting those trivial repetitions, we can express (4.2) by using the minimal integer, which we use in the rest of the proof. Hence, there exist $x_i\in R(x_0)(i=1, \cdots, n-1)$ such that $x_1\in R_1(x_0)$, $x_2\in R_2(x_1)$, $\cdots$ and $x\in R_n(x_{n-1})$. Let f_i (either f_S or f_T) be a state feedback corresponding to R_i (either S or T), $(i=1, \cdots, n)$. For any $u^0\in U$ and $t\in T$, let $x^0(t)=\phi_{0t}(x_0, u^0_{0t})$, and determine $v^0\in V$ so that the equality $u^0_{0t}(\tau)=f_1(x^0(\tau), v^0_{0t}(\tau))$ $(\forall\tau\in[0, t))$ is satisfied. This is possible since $f_1(x^0(\tau), -): A_1\to A$ is invertible. Let us define u^1_{0t} by $u^1_{0t}(\tau)=f_1(\phi^{f_1}_{0\tau}(x_1, v^0_{0\tau}), v^0_{0t}(\tau))$. Then, let us denote this correspondence by $g_{1,x}: \{x\}\times U\to U$, $(x, u^0)\mapsto u^1$, and let $(j_{1,x}\times g_{1,x}): \{x\}\times U\to R(x_0)\times U$ be defined by $(x, u^0)\mapsto$

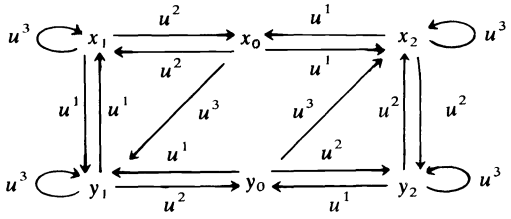
$(x_1, g_{1,x}(x, u^0))$. Furthermore, recursively, for $u^i \in U (i = 1, \cdots, n-1)$, let us define $x^i(t) = \phi_{0t}(x_i, u^i_{0t})$, and determine $v^i \in V$ so that the equality $u^i_{0t}(\tau) = f_{i+1}(x^i(\tau), v^i_{0t}(\tau))$ $(\forall \tau \in [0, t))$ is satisfied. This is possible because of the invertibility of $f_{i+1}(x^i(\tau), -) : A_1 \to A$. Let us define $u^{i+1}_{0t}(\tau) = f_{i+1}(\phi^{f_{i+1}}_{0\tau}(x_{i+1}, v^i_{0\tau}), v^i_{0t}(\tau))$. Then, let us denote this correspondence by $g_{i+1,x} : R(x_0) \times U \to U$, $(x_i, u^i) \mapsto u^{i+1}$, and let $(j_{i+1,x} \times g_{i+1,x}) : R(x_0) \times U \to R(x_0) \times U$ be defined by $(x_i, u^i) \mapsto (x_{i+1}, g_{i+1,x}(x_i, u^i))$. By collecting all these, let $g_x : \{x\} \times U \to U$ be defined by $g_x = g_{n,x} \circ (j_{n-1,x} \times g_{n-1,x}) \circ \cdots \circ (j_{1,x} \times g_{1,x})$, $(x, u^0) \mapsto g_x(x, u^0) = u^n$. By using a similar idea, it is possible to construct a mapping $g_{S(x_0)} : S(x_0) \times U \to U$, by $(x, u) \mapsto g_{S(x_0)}(x, u) \overset{d}{=} g_x(x, u)$. Furthermore, by using the choice function c, let us define a mapping $g : X \times U \to U$ by $(x, u) \mapsto g(x, u) = g_{s(c(x))}(x, u)$, and as in the proof of Lemma 4.3, let us define a mapping $g_0 : X \times A \to A$ by $(x, u^0(0)) \mapsto g_0(x, u^0(0)) = g(x, \tilde{u}^0)(0)$ for any $\tilde{u}^0$ such that $\tilde{u}^0(0) = u^0(0)$. On the other hand, let $h : A_1 \to A$ be the isomorphism between two alphabets and let us define $(\mathrm{id}_X \times h) : X \times A_1 \to X \times A$ by $(x, a_1) \mapsto (x, h(a_1))$. At the end, the state feedback $f : X \times A_1 \to A$ is defined by $f = g_0 \circ (\mathrm{id}_X \times h)$. It is possible to prove that R becomes ϕ^f-invariant for this feedback f, by using the same method as in the proof of Lemma 4.2. Therefore, in order to complete the proof, we have to show that this f is a regular feedback. In fact, since $f_i (i = 1, \cdots, n)$ are regular, $g_0(x, \cdot) : A_1 \to A$ becomes invertible for any $x \in X$. Moreover, h is isomorphic. Therefore the feedback f thus constructed is invertible for any $x \in X$. As for the feedback invariance w.r.t. (ϕ, λ), by using the expression (4.2) we can write any $x \in R(x_0)$ as $x_1 \in R_1(x_0)$, $x_2 \in R_2(x_1), \cdots$ and $x \in R_n(x_{n-1})$ for some $x_i \in X (i = 1, \cdots, n-1)$. Since S and T are feedback invariant w.r.t. (ϕ, λ), it follows that $\lambda(x_0) = \lambda(x_1), \cdots, \lambda(x_{n-1}) = \lambda(x)$; that is, $\lambda(x_0) = \lambda(x)$ for any $x \in R(x_0)$. Therefore, it is easy to see that, for any x_1 and x_2 in X, if $(x_1, x_2) \in R$, it follows that $\lambda(x_1) = \lambda(x_2)$. □

Proof of Theorem 4.1. For a given dynamical system (ϕ, λ), a set of partitions which is feedback invariant w.r.t. (ϕ, λ) is not empty, because a trivial partition $S = \{x | x \in X\}$ is always feedback invariant w.r.t. (ϕ, λ) for a zero feedback. And any chain of partitions S_i which is feedback invariant w.r.t. (ϕ, λ) is always bounded from above, that is, by $\vee_i S_i$,[1] with respect to the ordering $<$. Therefore, there exists at least one maximal partition which is feedback invariant w.r.t. (ϕ, λ), by Zorn's lemma. Moreover, it is unique, because, if M_1 and M_2 are maximal feedback invariant w.r.t. (ϕ, λ), then $M_1 \vee M_2$ is also feedback invariant w.r.t. (ϕ, λ), from Lemma 4.3, which contains M_1 and M_2. Hence there exists a unique maximal partition which is feedback invariant w.r.t. (ϕ, λ). □

Example. Let us consider an example to see the difference between control invariant structure and feedback invariant structure. Let the time scale be $T = \{0, 1, 2, \cdots\}$, the input alphabet be $A = \{u^1, u^2, u^3\}$ and the state set be $X = \{x_0, x_1, x_2, y_0, y_1, y_2\}$. The state transition function ϕ is defined in Fig. 3a. We consider a partition defined by $S = \{S_i | i = 0, 1, 2\}$ where $S_i = \{x_i, y_i\}$, $i = 0, 1, 2$. Then it is easy to see that this partition S is control ϕ-invariant. In fact, when initial states are in S_0 we may select the correspondence of inputs, $(x_0; u^1, u^2, u^3) \to (y_0; u^1, u^2, u^1)$ and $(y_0; u^1, u^2, u^3) \to (x_0; u^1, u^2, u^2)$ so that condition (3.7) is satisfied, and when initial states are in S_1 or S_2, there is an obvious correspondence of inputs. According to Lemma 4.2, there must be a feedback f so that S becomes ϕ^f-invariant. As its candidate we choose a feedback f as in Fig. 3b where another input alphabet is $A_1 = \{v^1, v^2, v^3\}$. In

[1] This holds because the set of equivalence classes on X is a complete lattice.

this case f is not regular since $f(y_0, -)\colon A_1 \to A$ is not invertible. And it is readily seen that there is no regular feedback f' which makes S be $\phi^{f'}$-invariant.



a. *The state transition function* ϕ.

	x_0	y_0	x_1	y_1	x_2	y_2
v^1	u^1	u^1	u^1	u^1	u^1	u^1
v^2	u^2	u^2	u^2	u^2	u^2	u^2
v^3	u^3	u^1	u^3	u^3	u^3	u^3

b. *The state feedback* $f\colon X \times A_1 \to A$.

FIG. 3

5. Relations between invariant structures. II. In this section, we prove the equivalence between a feedback zero-invariant structure and a uni-control invariant structure.

THEOREM 5.1. *A partition S is feedback zero-invariant w.r.t. ϕ if and only if S is uni-control invariant w.r.t. ϕ.*

Proof. It is easy to prove the "only if" part, since a set of control inputs for the uni-control invariant structure (Definition 3.4ii) can be generated by the feedback structure. The proof of the "if" part is similar to the one of Lemma 4.2. Let $E = \{(x, u^x)\}_{x \in X}$ be a set which is a subset of $X \times U$ satisfying the condition Definition 3.4ii. Let us define a state feedback $f\colon X \to A$ by $x \mapsto f(x) = u^x(0)$, where $(x, u^x) \in E$. Then S actually becomes zero-invariant w.r.t. ϕ^f by using this f. Define

$$\tau_0 = \min_{(x_1, x_2) \in S} \{\tau \in T \mid (\phi^f_{0\tau}(x_1, 0_\tau), \phi^f_{0\tau}(x_2, 0_\tau)) \in S\}, \tag{5.1}$$

and let the arguments be $\tilde{x}_1$ and $\tilde{x}_2$ which realize the minimum τ_0. Then, by the same reasoning as in the proof of Lemma 4.2, the set $\{\tau \mid 0 \leqq \tau < \tau_0\}$ contains only one element 0. From the definition of f, it follows that $\tilde{x}^i(\tau_0) \overset{d}{=} \phi^f_{0\tau_0}(\tilde{x}_i, 0_{0\tau_0}) = \phi_{0\tau_0}(\tilde{x}_i, f(\tilde{x}_i, 0(0))) = \phi_{0\tau_0}(\tilde{x}_i, u^{\tilde{x}_i}(0))$, $(i = 1, 2)$, implies that $(x^1(\tau_0), x^2(\tau_0)) \in S$, which contradicts the definition of τ_0. □

Owing to this theorem, when we consider a problem relating to zero-invariant structures, we need not deal with the exact form of the feedback, but only with a uni-control invariant structure, which is a usual situation in the linear theory. However, a set of partitions which is feedback zero-invariant cannot always be shown to be closed under the binary operation $\vee$, except for linear systems; this is unfortunate since we cannot prove the existence of a unique maximal feedback zero-invariant partition satisfying certain additional properties, although the existence of several of those maximal partitions can be guaranteed by means of Zorn's lemma. This fact causes inconvenience when we use zero-invariant structures for feedback systems with no

external input. However, we summarize this as a theorem which is comparable to Theorem 4.1.

THEOREM 5.2. *For a given dynamical system* (ϕ, λ), *there exist maximal partitions which are feedback zero-invariant w.r.t.* (ϕ, λ).

Remark. In other words, it can be said that, for (ϕ, λ), there exist maximal partitions which are uni-control invariant w.r.t. (ϕ, λ).

Proof. The proof is the same as the "former part" of Theorem 4.1. □

6. Disturbance localization. In this section, we characterize the disturbance localization problem, which is a typical example where the invariant concepts introduced in the previous sections are useful.

Let us consider an extended state transition function $\bar{\phi}$ which is defined for any t in T by

$$\bar{\phi}_{0t}: X \times U_{0t} \times W_{0t} \to X, \tag{6.1}$$

where W is a set of disturbance inputs. In particular, it satisfies

$$\bar{\phi}_{0t}(x_0, u_{0t}, 0_{0t}) \overset{d}{=} \phi_{0t}(x_0, u_{0t}).$$

In the rest of this section, we fix the initial state to be x_0.

DEFINITION 6.1.

(i) A dynamical system $(\bar{\phi}, \lambda)$ is *disturbance-localized* if

$$(\forall u \in U)(\forall w^1, w^2 \in W)(\forall t \in T) \quad (\lambda \circ \bar{\phi}_{0t}(x_0, u_{0t}, w^1_{0t}) = \lambda \circ \bar{\phi}_{0t}(x_0, u_{0t}, w^2_{0t})). \tag{6.2}$$

(ii) A dynamical system $(\bar{\phi}, \lambda)$ is *disturbance-localizable* by state feedback if there exists a regular state feedback $f: X \times A_1 \to A$ such that the dynamical system (ϕ^f, λ) is disturbance-localized.

The above Definition 6.1 says that, when a dynamical system is disturbance-localized, the response of the system is not influenced by any disturbance input w for any control input u. If we consider a special case when the disturbance is constantly zero, a feedback (ϕ, λ)-invariant structure has to exist in order to be disturbance-localizable. This observation suggests a formal characterization of the disturbance localization by state feedback.

THEOREM 6.1. *A dynamical system* $(\bar{\phi}, \lambda)$ *is disturbance-localizable by state feedback if and only if the unique maximal feedback invariant structure* $S_{\max}$ *w.r.t.* (ϕ, λ) *is disturbance invariant w.r.t.* $\bar{\phi}^f$, *where* f *is a state feedback which realizes* $S_{\max}$, *and* $S_{\max}$ *is disturbance invariant w.r.t.* $\bar{\phi}^f$ *if*

$$(\forall v \in V)(\forall w^1, w^2 \in W)(\forall t \in T) \quad ((\bar{\phi}^f_{0t}(x_0, v_{0t}, w^1_{0t}), \bar{\phi}^f_{0t}(x_0, v_{0t}, w^2_{0t})) \in S_{\max}).$$

Proof. Sufficiency is obvious from Definition 6.1. The following is the proof of necessity. Let us assume that the system is disturbance-localized by a properly chosen regular feedback f'. Let us define a subset $S_{v,t}$ of X by

$$S_{v,t} = \{x \in X | x = \bar{\phi}^{f'}_{0t}(x_0, v_{0t}, w_{0t}; w \in W\},$$

and define a family of subsets of X by

$$\bar{S} = \{S_{v,t} | v \in V, t \in T\}.$$

By S, we mean the smallest partition which is generated from $\bar{S}$. Then from the way S is defined S becomes a $(\bar{\phi}^{f'}, \lambda)$-invariant structure which is realized by a regular feedback

f'; that is, it is feedback (ϕ, λ)-invariant. In fact, any $x_1 \in S_{v,t}$ has an expression

$$x_1 = \bar{\phi}^{f'}_{0t}(x_0, v_{0t}, w^1_{0t})$$

for some $w^1 \in W$. And for any $v' \in V$ and $t' \in T$,

$$x' \overset{d}{=} \bar{\phi}^{f'}_{0t'}(x_1, v'_{0t'}, 0_{0t'})$$

$$= \bar{\phi}^{f'}_{0,t+t'}(x_0, v_{0t} \circ \sigma^t(v)_{t,t+t'}, w^1_{0t} \circ 0_{t,t+t'})$$

holds because of the semigroup property of $\bar{\phi}^{f'}$, where $\circ$ denotes the concatenation of inputs, and $\sigma^t(v)$ denotes the time shift of the input v defined by

$$\sigma^t(v)(\tau) = v(t+\tau).$$

Therefore, we get

$$x' \in S_{v_{0t} \circ \sigma^t(v'), t+t'}. \tag{6.3}$$

In general, if $(x_1, x_2) \in S$ then there exist an integer n, $y_i \in X$, $v^i \in V$ and $t_i \in T$, $i = 1, \cdots, n$, such that $y_1 = x_1$ and $y_n = x_2$, and

$$y_i \text{ and } y_{i+1} \in S_{v^i, t_i}, \qquad i = 1, \cdots, n.$$

Therefore, together with (6.3), for any $v \in V$ and $t \in T$, we get

$$\phi^{f'}_{0t}(y_i, v_{0t}) \text{ and } \phi^{f'}_{0t}(y_{i+1}, v_{0t}) \in S_{v^i_{0t'} \sigma^t(v), t_i + t}, \qquad i = 1, \cdots, n-1,$$

and hence, we can derive

$$(\phi^{f'}_{0t}(x_1, v_{0t}), \phi^{f'}_{0t}(x_2, v_{0t})) \in S.$$

Therefore S is $\bar{\phi}^{f'}$-invariant. Further, since the system $(\bar{\phi}^{f'}, \lambda)$ is disturbance-localized from the assumption, S is $(\bar{\phi}^{f'}, \lambda)$-invariant as well. On the other hand, from the definition of S, S is disturbance-invariant w.r.t. $\phi^{f'}$. Let us denote the maximal feedback invariant structure w.r.t. (ϕ, λ) by $S_{\max}$, and assume that the regular feedback f is realizing $S_{\max}$. Since $S \prec S_{\max}$ and S is disturbance invariant w.r.t. $\bar{\phi}^{f'}$, it follows that $S_{\max}$ is disturbance invariant w.r.t. $\bar{\phi}^{f}$, and hence the system is disturbance-localizable by state feedback for such $S_{\max}$ and f. □

This theorem asserts that, in order to check whether $(\bar{\phi}, \lambda)$ is disturbance-localizable, we may first find the unique maximal partition $S_{\max}$, and then, examine whether $S_{\max}$ is disturbance invariant w.r.t. $\bar{\phi}^{f}$.

In an analogous way, we can characterize the disturbance localization problem where the external input is set to be zero after the closed loop is made. Accordingly, the part of the input u in Definition 6.1 has to be altered to be 0 in the definition of the disturbance localization.

THEOREM 6.2. *A dynamical system $(\bar{\phi}, \lambda)$ is disturbance-localizable by state feedback with the external input zero if and only if there exists a maximal partition $S_{\max}$ which is feedback zero-invariant w.r.t. (ϕ, λ) such that $S_{\max}$ is disturbance zero-invariant w.r.t. $\bar{\phi}^{f}$, where f is a state feedback which realizes $S_{\max}$, and $S_{\max}$ is disturbance zero-invariant w.r.t. $\bar{\phi}^{f}$ if*

$$(\forall w^1, w^2 \in W)(\forall t \in T) \quad ((\bar{\phi}^{f}_{0t}(x_0, 0, w^1_{0t}), \bar{\phi}^{f}_{0t}(x_0, 0, w^2_{0t})) \in S_{\max}). \quad \square$$

7. Conclusion. In this paper, we introduced several invariant structures for general (nonlinear) dynamical systems in a set-theoretical setting, which generalized A-invariant subspaces and (A,B)-invariant subspaces, and developed the interrelations

between them. In particular, we proved an important property, that there exists a unique maximal indistinguishable structure by state feedback. This enabled us to formally characterize the disturbance localization problem in § 6. An analogous discussion is possible in order to formulate the model-matching problem, decoupling control, etc., where algebraic structural properties are the only concern and stability is not. If we assume more structures such a linearity, finiteness, continuity, differentiability, etc., depending on the system's attributes, we believe that more concrete discussions on control problems are possible by verifying the invariant concepts.

It is very interesting next to construct a generalized version of poles or pole-assignability in a set-theoretical plus topological setting, in order to formalize and characterize many other control problems for general dynamical systems.

REFERENCES

[1] R. W. BROCKETT, *Feedback invariants for nonlinear systems*,IFAC Symposium, Helsinki, 1978, pp. 1115–1120.

[2] P. R. HALMOS, *Naive Set Theory*, Van Nostrand, New York, 1960.

[3] S. ISHIJIMA, *The disturbance decoupling problem of nonlinear control systems*, Trans. SICE., 14 (1978), pp. 8–14 (in Japanese).

[4] P. E. LIEPA AND W. M. WONHAM, *Invariance and feedback for finite-state machines*, Proceedings 16th Allerton Conf. on Comm. Contr. and Comput., 157/164, 1978.

[5] M. D. MESAROVIC AND Y. TAKAHARA, *General Systems Theory: Mathematical Foundations*, Academic Press, New York, 1974.

[6] T. NOMURA AND K. FURUTA, *On invariant structures of nonlinear dynamical systems*, presented at the 3rd International Symposium on Mathematical Theory of Networks and Systems, Delft, the Netherlands, 1979.

[7] W. M. WONHAM, *Linear Multivariable Control—A Geometric Approach*, Lecture Notes in Economics and Mathematical Systems, Springer, New York, 1974.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 2, March 1981

0363-0129/81/1902-0001/$01.00/0

A NOTE ON ASYMPTOTICALLY EFFICIENT ESTIMATES OF PARAMETERS IN CONTINUOUS-TIME DYNAMICAL SYSTEMS*

ARUNABHA BAGCHI†

Abstract. Maximum likelihood estimation of parameters in continuous-time stochastic linear dynamical systems has recently been shown to be consistent under certain sufficient conditions. The purpose of this note is to prove that these estimates are asymptotically efficient.

1. Introduction. It has been shown recently in [1] that a maximum likelihood estimate of parameters in continuous-time linear dynamical systems yields strongly consistent estimates of the unknown system parameters. This extends corresponding results for discrete-time linear dynamical systems [2], [3]. It is known that in the discrete-time case, the maximum likelihood estimate is also asymptotically efficient [4]. The purpose of this note is to prove this result in the continuous-time situation.

2. Problem statement. Let $(\Omega, \mathcal{B}, P)$ be the basic probability space. Consider the following continuous-time linear stochastic dynamical system:

$$x(t;\omega) = \int_0^t Ax(s;\omega)\,ds + \int_0^t Bu(s)\,ds + \int_0^t F\,dW_1(s;\omega), \tag{2.1}$$

$$Y(t;\omega) = \int_0^t Cx(s;\omega)\,ds + \int_0^t G\,dW_2(s;\omega), \tag{2.2}$$

where $u(t)$ is a p-dimensional "input" function; $x(t;\omega)$ and $\gamma(t;\omega)$ are n- and m-dimensional "state" and "output" functions respectively; A, B and C are respectively $n\times n$, $n\times p$ and $m\times n$ constant but partially unknown matrices; $W_1(t;\omega)$ and $W_2(t;\omega)$ are $n\times 1$ and $m\times 1$ independent Wiener processes, and F, G are respectively $n\times n$, $m\times m$ constant but partially unknown matrices with GG^* having an inverse, where * denotes the transpose. We assume that GG^* is completely known. There is no loss of generality in assuming that $GG^* = I$, the identity matrix.

We assume that the pair (C, A) is completely observable, and that the system has reached the steady state. Working with the steady state is no restriction, since we are concerned only with asymptotic properties in the sequel.

Let θ denote the vector of all the unknown system parameters, with θ_0 denoting their true values. Let $\hat{\theta}_T(\omega)$ be a maximum likelihood estimate of θ_0 based on $Y(t;\omega)$, $0 \leqq t \leqq T$. $\hat{\theta}_T(\omega)$ is thus a minimum, in a sufficiently small neighborhood of θ_0, of the log-likelihood functional, as given in [5].

We introduce some notation to express this log-likelihood functional in a compact form. Let

$$m(\theta;t) = \int_0^t C\exp[A(t-s)]Bu(s)\,ds \tag{2.3}$$

and let $\mathcal{L}(\theta)$ be the Volterra operator

$$\mathcal{L}(\theta)f = g; \qquad g(t) = C\int_0^t \exp[(A-PC^*C)(t-s)]PC^*f(s)\,ds, \tag{2.4}$$

* Received by the editors April 17, 1979, and in revised form April 28, 1980.
† Department of Applied Mathematics, Twente University of Technology, Enschede, the Netherlands.

where P satisfies the algebraic Riccati equation

$$AP+PA^*+FF^*-PCC^*P=0. \tag{2.5}$$

We also use the notation

$$[\mathscr{L}(\theta)\,dY(\cdot\,;\omega)](t)=C\int_0^t \exp[(A-PC^*C)(t-s)]PC^*\,dY(s;\omega). \tag{2.4'}$$

Finally, for two square integrable vector-valued functions $f(\cdot)$ and $g(\cdot)$ in $[0, T]$, we use

$$[f(\cdot), g(\cdot)]=\int_0^T [f(t), g(t)]\,dt, \quad \text{with } \|f(\cdot)\|^2=[f(\cdot), f(\cdot)], \tag{2.6}$$

$$[f(\cdot), dY(\cdot\,;\omega)]=\int_0^T [f(t), dY(t;\omega)]. \tag{2.6'}$$

Then the log-likelihood functional for the problem can be expressed as

$$\begin{aligned} q(\theta;\, Y(\cdot\,;\omega);\, T)=\frac{1}{2T}\{&\|m(\theta;\cdot)+\mathscr{L}(\theta)(dY(\cdot\,;\omega)-m(\theta;\cdot))\|^2 \\ &-2[m(\theta;\cdot)+\mathscr{L}(\theta)(dY(\cdot\,;\omega)-m(\theta;\cdot)),\, dY(\cdot\,;\omega)]\}. \end{aligned} \tag{2.7}$$

Furthermore, if $\mathscr{K}(\theta)f=g$, $g(t)=C\int_0^t \exp(A(t-s))PC^*f(s)\,ds$, then $(I+\mathscr{K}(\theta))^{-1}=I-\mathscr{L}(\theta)$, and

$$dY(t;\omega)-m(\theta_0;t)\,dt=[(I+\mathscr{K}(\theta_0))\,dZ_0(\cdot\,;\omega)](t), \tag{2.8}$$

where $Z_0(\cdot\,;\omega)$, the so called "innovation process", is a Wiener process with identity covariance. These results have been proved in [5].

Let θ_i denote the ith component of θ and let $\nabla_\theta q(\theta;\, Y(\cdot\,;\omega);\, T)$ be the gradient vector with the ith component $q_i(\theta;\, Y(\cdot\,;\omega);\, T)$. Let $Q(\theta;\, Y(\cdot\,;\omega);\, T)$ be the matrix with ijth component

$$q_{ij}(\theta;\, Y(\cdot\,;\omega);\, T)=\frac{\partial^2}{\partial\theta_i\,\partial\theta_j}q(\theta;\, Y(\cdot\,;\omega);\, T).$$

We write

$$q(\theta;\, T)=Eq(\theta;\, Y(\cdot\,;\omega);\, T), \qquad q(\theta)=\lim_{T\to\infty} q(\theta;\, T),$$

and use similar notations for $\nabla_\theta q$ and Q.

These limits exist under the following regularity conditions:

Condition A. For fixed θ, assume that A is stable.

Condition B. For fixed θ, assume that the pair (C, A) is completely observable.

Condition C. We assume that the input $u(\cdot)$ is such that

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T \|u(t)\|^2\,dt<\infty \quad \text{(exists and is finite)},$$

and

$$r_u(t)=\lim_{T\to\infty}\frac{1}{T}\int_0^T u(s)u(s+t)^*\,ds$$

is a continuous function of t in every finite interval.

LEMMA 1. *$q(\theta; Y(\cdot\,; \omega); T)$ is almost surely (a.s.) differentiable with respect to (w.r.t.) θ_i, and $q_i(\theta; Y(\cdot\,; \omega); T)$ may be obtained by bringing the differentiation operation under the integral signs. A similar conclusion holds for $q_{ij}(\theta; Y(\cdot\,; \omega); T)$.*

Proof. We have only to justify this interchange of operations for the stochastic integral terms in $q(\theta; Y(\cdot\,; \omega); T)$. In fact, we have two types of stochastic integral terms appearing in q; namely,

$$\int_0^T L(t, s; \theta)\, dZ_0(s; \omega),$$

and

$$\int_0^T [M(t; \theta) \int_0^t N(s; \theta)\, dZ_0(s; \omega),\, dZ_0(t; \omega)],$$

where $L(t, s; \theta)$ is continuously differentiable with respect to t and s; $M(t; \theta)$ and $N(t; \theta)$ are continuously differentiable with respect to t; and L, M, N have partial derivatives with respect to θ existing for all orders. We begin with the first type of stochastic integral. For fixed θ,

$$(*) \quad \int_0^T L(t, s; \theta)\, dZ_0(s; \omega) = L(t, T; \theta) Z_0(T; \omega) - \int_0^T \frac{\partial L(t; s; \theta)}{\partial s} Z_0(s; \omega)\, ds \quad \text{a.s.}$$

But the right-hand side is defined a.s. in ω for all θ. The right-hand side is, therefore, a version of the stochastic integral on the left-hand side of (*), and we work with this version now. This version is clearly a.s. differentiable w.r.t. θ_i, and

$$\frac{\partial}{\partial \theta_i}\left[L(t, T; \theta) Z_0(T; \omega) - \int_0^T \frac{\partial L(t, s; \theta)}{\partial s} Z_0(s; \omega)\, ds\right]$$
$$= \frac{\partial L(t, T; \theta)}{\partial \theta_i} Z_0(T; \omega) - \int_0^T \frac{\partial}{\partial s} \frac{\partial L(t, s; \theta)}{\partial \theta_i} Z_0(s; \omega)\, ds;$$

the expression on the right is a version of the stochastic integral

$$\int_0^T \frac{\partial L(t, s; \theta)}{\partial \theta_i}\, dZ_0(s; \omega).$$

We now consider the second type of stochastic integral. For fixed θ, we use an identity for Ito integrals ([5, p. 96]):

$$\int_0^T \left[M(t; \theta) \int_0^t N(s; \theta)\, dZ_0(s; \omega),\, dZ_0(t; \omega)\right]$$
$$+ \int_0^T \left[N(t; \theta)^* \int_0^t M(s; \theta)^*\, dZ_0(s; \omega),\, dZ_0(t; \omega)\right]$$
$$(**) \qquad = \left[\int_0^T N(s; \theta)\, dZ_0(s; \omega),\, \int_0^T M(s; \theta)^*\, dZ_0(s; \omega)\right]$$
$$- \int_0^T \operatorname{Tr} N(t; \theta) M(t; \theta)\, dt.$$

The right-hand side has a version which is defined a.s. in ω for all θ, and furthermore, this version is a.s. differentiable w.r.t. θ_i. By symmetry, each stochastic integral on the left-hand side has a version that is defined a.s. in ω for all θ. These versions are also a.s.

differentiable w.r.t. θ_i. Since in the right-hand side of (**), partial differentiation may be brought under the integral sign, the same holds for the a.s. differentiable version of

$$\int_0^T \left[M(t;\theta) \int_0^t N(s;\theta)\, dZ_0(s;\omega), dZ_0(t;\omega) \right]. \qquad \square$$

If $\hat{\theta}_T(\omega)$ is an unbiased estimate of θ_0 based on $Y(t;\omega)$, $0 \leqq t \leqq T$, we can extend the Cramér–Rao inequality to show that the variance of $\hat{\theta}_T(\omega)$ must satisfy

$$E[(\hat{\theta}_T(\omega)-\theta_0)(\hat{\theta}_T(\omega)-\theta_0)^*] \geqq [TQ(\theta_0;T)]^{-1}, \tag{2.9}$$

where $A \geqq B$ means that the matrix $A-B$ is nonnegative definite. $Q(\theta_0;T)$ is thus the Fisher information matrix for the system under study. It follows from (2.9) that

$$\sqrt{Q(\theta_0;T)}\sqrt{T}(\hat{\theta}_T(\omega)-\theta_0)$$

always has zero mean and variance $\geqq I$. This motivates the following:

DEFINITION. A function $\{\hat{\theta}_T(\omega)\}$, $(T>0)$ of estimates is said to be *asymptotically efficient* if $\sqrt{Q(\theta_0)}\sqrt{T}(\hat{\theta}_T-\theta_0)$ converges in distribution as $T\to\infty$ to $N(0;I)$, where $N(0;I)$ stands for normal distribution with zero mean and identity variance.

If $Q(\theta_0)$ is positive definite, there exists a compact neighborhood $\bar{\mathcal{N}}(\theta_0)$ of θ_0 in which no other value of θ except θ_0 yields an output coinciding a.s. with the actual output $Y(t;\omega)$, $0 \leqq t \leqq T$ (see [5, pp. 210–212]). It has been shown in [1] that if $Q(\theta_0)$ is positive definite and the regularity conditions A, B, C mentioned before hold for all θ in $\bar{\mathcal{N}}(\theta_0)$, there exists a root of $\nabla_\theta q(\theta; Y(\cdot\,;\omega); T)=0$ that is strongly consistent in the sense that this root $\hat{\theta}_T(\omega)\to\theta_0$ a.s. $T\to\infty$. We prove in this article that under these conditions, $\{\hat{\theta}_T(\omega)\}$ is also asymptotically efficient.

3. Proof of asymptotic efficiency. The proof is based on two lemmas.

LEMMA 2. *Let $Q(\theta_0)$ be positive definite and conditions* A, B, C *hold for all θ in $\bar{\mathcal{N}}(\theta_0)$. Let $\bar{\theta}_T(\omega)$ be a random vector converging in probability to θ_0. Then*

$$Q(\bar{\theta}_T(\omega); Y(\cdot\,;\omega); T) \xrightarrow{\text{pr}} Q(\theta_0) \quad \text{as } T\to\infty.$$

Proof. Since $\bar{\theta}_T(\omega)$ converges in probability to θ_0, given $\varepsilon>0$ there exists $T(\varepsilon)$ such that for $T>T(\varepsilon)$,

$$P(\|\bar{\theta}_T(\omega)-\theta_0\|<\varepsilon) > 1-\varepsilon.$$

It is easy to see that $Q(\theta)$ is continuous in θ for θ in $\bar{\mathcal{N}}(\theta_0)$. Let $q_{ij}^u(\theta_0)$ and $q_{ij}^l(\theta_0)$ be the l.u.b. and g.l.b. of the ijth component $q_{ij}(\theta)$ of $Q(\theta)$ for θ in

$$\mathcal{N}_\varepsilon(\theta_0) = \{\theta \mid \|\theta-\theta_0\|<\varepsilon\} \subset \bar{\mathcal{N}}(\theta_0).$$

From [5, Thm. 8.2] we know that under the conditions of the lemma,

$$E|q_{ij}(\theta; Y(\cdot\,;\omega); T) - q_{ij}(\theta)|^2 \to 0 \quad \text{as } T\to\infty,$$

uniformly for θ in $\mathcal{N}_\varepsilon(\theta_0)$. This implies that, for an arbitrary $\varepsilon'>0$, a $T_{ij}(\varepsilon')$ exists such that for $T>T_{ij}(\varepsilon')$, and for any θ in $\mathcal{N}_\varepsilon(\theta_0)$,

$$\begin{aligned} P(q_{ij}^l(\theta_0)-\varepsilon' < q_{ij}(\theta; Y(\cdot\,;\omega); T) < q_{ij}^u(\theta_0)+\varepsilon') \\ \geqq P(q_{ij}(\theta)-\varepsilon' < q_{ij}(\theta; Y(\cdot\,;\omega); T) < q_{ij}(\theta)+\varepsilon') \\ \geqq 1-\frac{1}{\varepsilon'^2} E|q_{ij}(\theta; Y(\cdot\,;\omega); T) - q_{ij}(\theta)|^2 \\ \geqq 1-\varepsilon'. \end{aligned}$$

For any $T > \max[(T(\varepsilon), T_{ij}(\varepsilon'))]$, $1 \leqq i, j \leqq r$), for any θ in $\mathcal{N}_\varepsilon(\theta_0)$, let

$$E_T = \{\omega \,|\, \|\bar{\theta}_T(\omega) - \theta_0\| < \varepsilon, q_{ij}^1(\theta_0) - \varepsilon' < q_{ij}(\theta; Y(\cdot\,;\omega); T) < q_{ij}^u(\theta_0) + \varepsilon', 1 \leqq i, j \leqq r\}.$$

Then, $P(E_T) \geqq 1 - \varepsilon - r^2\varepsilon'$.

Any $\omega \in E_T$ satisfies

$$q_{ij}^1(\theta_0) - \varepsilon' < q_{ij}(\bar{\theta}_T(\omega); Y(\cdot\,;\omega); T) < q_{ij}^u(\theta_0) + \varepsilon', \qquad 1 \leqq i, j \leqq r. \tag{3.1}$$

Hence for $T > \max[(T(\varepsilon), T_{ij}(\varepsilon'))]$, $1 \leqq i, j \leqq r$, the probability that (3.1) is satisfied exceeds $1 - \varepsilon - r^2\varepsilon'$, $1 \leqq i, j \leqq r$. But since ε and ε' are arbitrarily small, and since $Q(\theta)$ is continuous in $\bar{\mathcal{N}}(\theta_0)$, $q_{ij}^u(\theta_0) - q_{ij}(\theta_0)$ and $q_{ij}(\theta_0) - q_{ij}^1(\theta_0)$ can be made arbitrarily small. Hence, $q_{ij}(\bar{\theta}_T(\omega); Y(\cdot\,;\omega); T)$ converges in probability to $q_{ij}(\theta_0)$, for $1 \leqq i, j \leqq r$. This proves the lemma. □

LEMMA 3. *Suppose that $Q(\theta_0)$ is positive definite and let*

$$Y_T = -(\sqrt{Q(\theta_0)})^{-1}\sqrt{T}\nabla_\theta q(\theta_0; Y(\cdot\,;\omega); T). \tag{3.2}$$

Then Y_T converges in distribution to a normal random vector with zero mean and identity variance.

Proof. By Lemma 1, we can obtain from (2.7)

$$q_i(\theta_0; Y(\cdot\,;\omega); T) = -\frac{1}{T}\int_0^T [f_i(t;\omega), dZ_0(t;\omega)];$$

here

$$f_i(t;\omega) = [(I - \mathcal{L}(\theta_0))m_i(\theta_0;\cdot) + \mathcal{L}_i(\theta_0)(I + \mathcal{K}(\theta_0))\, dZ_0(\cdot\,;\omega)](t),$$

where the suffix i appearing in the right-hand side means partial differentiation with respect to the ith component θ_i of θ, $1 \leqq i \leqq r$. In particular, $\mathcal{L}_i(\theta)$ refers to the Volterra operator

$$\mathcal{L}_i(\theta)f = g; \qquad g(t) = \int_0^t \frac{\partial}{\partial\theta_i}[C \exp\{(A - PC^*C)(t-s)\}PC^*]f(s)\, ds. \tag{3.3}$$

From the expression for $f_i(t;\omega)$, we can readily obtain, using properties of Ito integrals, that

$$\begin{aligned} E\int_0^T [f_i(t;\omega), f_j(t;\omega)]\, dt &= \int_0^T [((I - \mathcal{L}(\theta_0))m_i(\theta_0;\cdot))(t), \\ &\qquad ((I - \mathcal{L}(\theta_0))m_j(\theta_0;\cdot))(t)]\, dt \\ &\quad + [\mathcal{L}_i(\theta_0)(I + \mathcal{K}(\theta_0)), \mathcal{L}_j(\theta_0)(I + \mathcal{K}(\theta_0))] \\ &= Tq_{ij}(\theta_0; T), \end{aligned} \tag{3.4}$$

where, for Volterra operators K_1 and K_2 with kernels $K_1(t, s)$ and $K_2(t, s)$, we have used the notation

$$[K_1, K_2] = \int_0^T \int_0^t \operatorname{Tr} K_1(t, s)K_2(t, s)^*\, ds\, dt.$$

The second equality in (3.4) can be established by straightforward calculations using Lemma 1 for calculating $q_{ij}(\theta_0; Y(\cdot\,;\theta); T)$ and then using properties of Ito integrals while taking the expectation. Both the terms in the middle expression for (3.4) tend to infinity as $T \to \infty$, under the regularity conditions A, B, C at θ_0. It is easy to see from (3.4) that

$$\frac{1}{T} E \int_0^T [f_i(t;\omega), f_j(t;\omega)]\, dt \to q_{ij}(\theta_0) \quad \text{as } T \to \infty.$$

We now prove that

$$\frac{1}{T}\int_0^T [f_i(t;\omega), f_j(t;\omega)]\, dt \xrightarrow{\text{pr}} q_{ij}(\theta_0) \quad \text{as } T \to \infty. \tag{3.5}$$

To prove this assertion, note that

$$\frac{1}{T}\int_0^T [f_i(t;\omega), f_j(t;\omega)]\, dt = \frac{1}{T}\int_0^T [k_i(t), k_j(t)]\, dt + \frac{1}{T}\int_0^T [k_i(t), L_j(t;\omega)]\, dt$$
$$+ \frac{1}{T}\int_0^T [k_j(t), L_i(t;\omega)]\, dt + \frac{1}{T}\int_0^T [L_i(t;\omega), L_j(t;\omega)]\, dt,$$

where

$$k_i(t) = [(I - \mathscr{L}(\theta_0)) m_i(\theta_0; \cdot)](t)$$

and

$$L_i(t;\omega) = [\mathscr{L}_i(\theta_0)(I + \mathscr{K}(\theta_0))\, dZ_0(\cdot\,;\omega)](t).$$

Let

$$R_j(t-s) = E(L_j(t;\omega) L_j(s;\omega)^*).$$

Then

$$E\left(\frac{1}{T}\int_0^T [k_i(t), L_j(t;\omega)]\, dt\right)^2 = \frac{1}{T^2}\int_0^T \int_0^T [k_i(t), R_j(t-s) k_i(s)]\, ds\, dt$$
$$\leqq \lambda_j(T) \frac{1}{T^2} \int_0^T \|k_i(t)\|^2\, dt,$$

where $\lambda_j(T)$ is the largest eigenvalue of the operator

$$R_j(T) f = g; \qquad g(t) = \int_0^T R_j(t-s) f(s)\, ds, \quad 0 \leqq t \leqq T,$$

and is a nonnegative definite operator. Now, for an appropriate dimensional vector-

valued function k, $\|k\| = 1$,

$$\frac{1}{T}[R_j(T)k, k] = \frac{1}{T}\int_0^T \int_0^T [R_j(t-s)k(s), k(t)]\, ds\, dt$$

$$= \frac{1}{T}\int_{-\infty}^{\infty} P_j(f) \int_0^T \int_0^T [e^{i2\pi f(t-s)}k(s), k(t)]\, ds\, dt\, df$$

($P_j(f)$ is the spectral density of $R_j(T)$)

$$\leqq \frac{1}{T}\int_{-\infty}^{\infty} \|P_j(f)\| \left\| \int_0^T e^{i2\pi ft}k(t)\, dt \right\|^2 df \leqq \frac{1}{T} \sup_f \|P_j(f)\|,$$

implying that

$$\frac{\lambda_j(T)}{T} \leqq \frac{1}{T} \sup_f \|P_j(f)\| \to 0 \quad \text{as } T \to \infty.$$

Since $\lim_{T\to\infty} 1/T \int_0^T \|k_i(t)\|^2\, dt$ exists and is finite (from regularity conditions A, B, C, see [5] for details):

$$\frac{1}{T}\int_0^T [k_i(t), L_j(t;\omega)]\, dt \xrightarrow{\text{pr}} 0 \quad \text{as } T \to \infty.$$

The expected value of the square of this expression is

$$\frac{1}{T^2}\int_0^T \int_0^T \{E([L_i(t;\omega), L_j(t;\omega)][L_i(s;\omega), L_j(s;\omega)])$$

$$- E([L_i(t;\omega), L_j(t;\omega)])E([L_i(s;\omega), L_j(s;\omega)])\}\, ds\, dt.$$

Let us consider the case when L_i, L_j are one-dimensional. By rules for calculating four products of Gaussians,

$$E(L_i(t;\omega)L_j(t;\omega)L_i(s;\omega)L_j(s;\omega)) - E(L_i(t;\omega)L_j(t;\omega))E(L_i(s;\omega)L_j(s;\omega))$$

$$= E(L_i(t;\omega)L_j(s;\omega))E(L_j(t;\omega)L_i(s;\omega))$$

$$+ E(L_i(t;\omega)L_i(s;\omega))E(L_j(t;\omega)L_j(s;\omega)).$$

In the vector case, we have a finite number of such expressions. Let

$$R_{ij}(t-s) = E(L_i(t;\omega)L_j(s;\omega)^*).$$

Then

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T \int_0^T R_{ij}(t-s)R_{ij}(s-t)\, ds\, dt$$

and

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T \int_0^T R_i(t-s)R_j(t-s)\, ds\, dt$$

exist and are finite (from regularity conditions A, B, C). Hence, the result follows.

Once (3.5) is established, it follows from Taraskin [6, Thm. 5] that

$$\sqrt{T}\nabla_\theta q(\theta_0; Y(\cdot\,;\omega); T)$$

converges in distribution to a normal random vector with zero mean and covariance $Q(\theta_0)$. This proves the lemma. □

We now come to our main result.

THEOREM. *$Z_T = \sqrt{Q(\theta_0)}\sqrt{T}(\hat{\theta}_T - \theta_0)$ converges in distribution to a normal random vector with zero mean and identity variance, under the assumption that $Q(\theta_0)$ is positive definite and regularity conditions* A, B, C *hold in $\bar{\mathcal{N}}(\theta_0)$.*

Proof. Expanding $\nabla_\theta q(\theta;\, Y(\cdot\,;\omega);\, T)$ in a Taylor series about θ_0, we obtain

$$\nabla_\theta q(\theta;\, Y(\cdot\,;\omega);\, T) = \nabla_\theta q(\theta_0;\, Y(\cdot\,;\omega);\, T) + Q(\theta^*;\, Y(\cdot\,;\omega);\, T)(\theta - \theta_0),$$

where

$$\|\theta^* - \theta_0\| < \|\theta - \theta_0\|.$$

$Q(\theta_0)$ is strictly positive definite, and hence, putting $\theta = \hat{\theta}_T$ and multiplying by $(\sqrt{Q(\theta_0)})^{-1}\sqrt{T}$, we get

$$\begin{aligned} 0 = {} & (\sqrt{Q(\theta_0)})^{-1}\sqrt{T}\nabla_\theta q(\theta_0;\, Y(\cdot\,;\omega);\, T) \\ & + (\sqrt{Q(\theta_0)})^{-1}\sqrt{T}\, Q(\bar{\theta}_T;\, Y(\cdot\,;\omega);\, T)(\hat{\theta}_T - \theta_0), \end{aligned}$$

where

$$\|\bar{\theta}_T - \theta_0\| < \|\hat{\theta}_T - \theta_0\|.$$

Since $\hat{\theta}_T \to \theta_0$ a.s. and therefore in probability, as $T \to \infty$, and with $\|\bar{\theta}_T - \theta_0\| \leq \|\hat{\theta}_T - \theta_0\|$, it follows with appropriate modification from Wilks ([7, p. 104]) that $\bar{\theta}_T(\omega)$ also converges in probability to θ_0. This gives us, from Lemma 2, that

$$V_T = (\sqrt{Q(\theta_0)})^{-1} Q(\bar{\theta}_T;\, Y(\cdot\,;\omega);\, T)(\sqrt{Q}(\theta_0))^{-1} \to I$$

in probability as $T \to \infty$. Note that Z_T in the theorem may be defined by

$$V_T Z_T = Y_T,$$

where Y_T has been defined in Lemma 3. The result now follows by extension of Cramér [8, Thm. 20.6] to the vector case (Slutsky's theorem). □

4. Conclusion. We consider only the case when the observation noise covariance is completely known. When the observation noise covariance has unknown components, a modified likelihood functional suggested in [9], gives consistent estimates of the unknown parameters. The proof that the estimates are asymptotically efficient can be done using the method proposed in § 3 without any alteration.

REFERENCES

[1] A. BAGCHI, *Consistent estimates of parameters in continuous time systems*, in Analysis and Optimization of Stochastic Systems, Academic Press, New York, 1980.

[2] L. LJUNG, *On Consistency for Prediction Error Identification Methods*, Report 7405, Division of Automatic Control, Lund Institute of Technology, Sweden, March 1974.

[3] A. BAGCHI, *Consistent estimates of parameters in noisy dynamical systems*, Internat. J. Control, 26 (1977), pp. 883–900.

[4] P. E. CAINES AND L. LJUNG, *Asymptotic normality and accuracy of prediction error estimators*, Preprints Joint Automatic Control Conference (1976).

[5] A. V. BALAKRISHNAN, *Stochastic Differential Systems*, Lecture Notes in Economics and Mathematical Systems, vol. 84, Springer-Verlag, New York, 1973.

[6] A. F. TARASKIN, *Some limit theorems for stochastic integrals*, in Theory of Stochastic Processes, vol. 1, John Wiley, New York, 1974, pp. 136–151.

[7] S. WILKS, *Mathematical Statistics*, John Wiley, New York, 1962.

[8] H. CRAMÉR, *Mathematical Methods of Statistics*, Princeton University Press, Princeton, NJ, 1946.

[9] A. BAGCHI, *Continuous time systems identification with unknown noise covariance*, Automatica, 11 (1975).

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 2, March 1981

0363-0129/81/1902-0002/$01.00/0

DYNAMICAL REALIZATIONS OF FINITE VOLTERRA SERIES*

P. E. CROUCH†

Abstract. In this paper, realizations of finite Volterra series are viewed as nonlinear analytic input–output systems, with state space described by an analytic manifold. For a minimal realization guaranteed by H. J. Sussmann, the state space, which is unique up to diffeomorphism, is shown to have the homogeneous space structure of a nilmanifold, the quotient of two nilpotent Lie groups. The structure of nilmanifolds as described by A. Malcev is used to show that for these systems, the state space has a vector space structure. As a consequence of this result, it is shown that a minimal realization of a finite Volterra series can be described as a cascade of linear subsystems with polynomial link maps, in which the dimension of each linear subsystem is independent of the realization considered.

1. Introduction. In the last few years, there have been considerable advances in the theory of nonlinear input–output systems. In this paper, this theory is applied to nonlinear systems admitting input–output maps described by finite Volterra series, in order to identify the natural structure exhibited by these systems. It is shown that much of the structure and theory of linear systems generalizes directly to this class of nonlinear system.

The paper is divided into three main sections, 2, 3 and 4. In the second section, the systems and Volterra series considered in the paper are introduced, and some important results from the existing nonlinear systems theory are presented. The class of system considered is determined to a large extent by the manner in which the system structure is determined. In this paper, the structural identification problem is treated as a nonlinear realization problem; that is, given a class of input–output map, determine the common structure that exists between suitable realizations of these maps. To formulate this problem unambiguously, a class of system has to be identified within which realizations are essentially unique. The systems considered in this paper are therefore chosen to satisfy the hypotheses of Sussmann's existence and uniqueness theorem for minimal realizations [20]. One of the most significant restrictions that this imposes on the systems considered is the analyticity of the data, another is the completeness of certain vector fields associated with the system. A less important restriction is the linearity of the controls in the dynamics of the system. This restriction is introduced to simplify the structure of the input–output map. In particular, this enables the results of Krener and Lesiak [15] to be applied, giving a convenient coordinate-free representation of the Volterra kernels in terms of the system data.

Section 3 deals with the natural structure of the state space of a realization of a finite Volterra series, perhaps the fundamental question in a nonlinear realization problem. It is evident from many earlier works that most of the difficulty in analyzing these systems comes from fixing coordinate systems in which to express them. In this paper, the system is treated in a coordinate-free way, although the geometry of the system does allow convenient coordinate systems to be chosen. The state space is characterized in two stages. Firstly, the natural properties of the Lie algebra of the system allow the state space to be formulated as a homogeneous space of nilpotent Lie groups. The second stage involves considering the structure of these spaces, as first investigated by Malcev [17], in conjunction with the restrictions imposed due to the finiteness of the Volterra series. It is shown that the state space is homeomorphic to a Cartesian space, a far more profound result than that of its linear counterpart.

* Received by the editors June 28, 1979 and in final revised form April 29, 1980.
† Department of Engineering, University of Warwick, Coventry CV4 7AL, England.

Section 4 is concerned with finding natural coordinate systems in which to display the structure of these systems most effectively. As a consequence, it is shown that the state space has a natural vector space structure. A finer structure is also identified which shows that these systems are cascades of linear systems with polynomial link maps. The dimensions of these linear subsystems are invariant integers associated with the input–output map, which can be identified directly from the Volterra kernels. In the special case where only one term appears in the Volterra series, the minimal realization is shown to exhibit further structure with interesting consequences.

2. Preliminary results and definitions.

2.1. The class of system considered is defined by the following equations:

$$\dot{x} = f(x) + \sum_{i=1}^{m} u_i g_i(x), \qquad x(0) = x_0, \qquad x \in M^n,$$

$$y = h(x),$$

where the associated vector fields $f + \sum_{i=1}^{m} \alpha_i g_i$ for any $(\alpha_1 \cdots \alpha_m) \in \mathbb{R}^m$, are complete analytic vector fields on M^n, an n-dimensional real analytic connected manifold, and h is an $\mathbb{R}^q$-valued analytic function on M, with components h_i, $i = 1 \cdots q$. The vector with components u_i, $i = 1 \cdots m$ will be denoted by u.

Since the associated vector fields are complete, solutions to the above equations are defined on $[0, T]$ for all piecewise constant controls u on $[0, T]$ and all positive times T. Standard arguments allow us to extend this class of controls to include measurable controls on suitable intervals and taking values in a given constraint set $\Omega \subset \mathbb{R}^m$.

2.2. The input–output map described by the above system can be written as a convergent Volterra series, as shown in Brockett [2], Brockett and Gilbert [3], and Krener and Lesiak [15]. The formal series will be written as

$$y(t) = W_0(t) + \int_0^t W_1(t, \sigma_1)(u(\sigma_1))\, d\sigma_1 + \int_0^t \int_0^{\sigma_1} W_2(t, \sigma_1, \sigma_2)(u(\sigma_1))(u(\sigma_2))\, d\sigma_1\, d\sigma_2 + \cdots.$$

A series terminating with the term involving the pth kernel will be called a Volterra series of length p. Since each initial state obviously defines an input–output map, the dependence of the kernels on the initial state is included where it is significant.

The Volterra kernels $W_n(t, \sigma_1 \cdots \sigma_n, x)$ are multilinear maps for each $t, \sigma_1 \cdots \sigma_n, x \in \mathbb{R}^{n+1} \times M$,

$$W_n(t, \sigma_1 \cdots \sigma_n, x) : \mathbb{R}^m \times \cdots \times \mathbb{R}^m \to \mathbb{R}^q,$$

the components of which will be denoted by

$$W_n^{i_0 i_1 \cdots i_n}(t, \sigma_1 \cdots \sigma_n, x), \qquad 1 \leqq j_0 \leqq q, \quad 1 \leqq j_k \leqq m, \quad 1 \leqq k \leqq n.$$

For convenience, the kernel $W_n(t, \sigma_1 \cdots \sigma_n, x)$ will often be identified with a single component, since in most of the analysis, the distinction between different components involves only indices.

Let $V(M)$, $Z(M)$ and $C(M)$ denote the linear spaces of analytic vector fields, covector fields and functions on M, respectively. If $a \in V(M)$, $\tau \in C(M)$ and hence, $d\tau \in Z(M)$, then the Lie derivative of τ $[d\tau]$ by a, will be denoted by $L_a(\tau)[L_a(d\tau)]$.

The following identities are standard:

$$d\tau(a) = L_a(\tau) = a(\tau),$$
$$L_a(d\tau) = d(a(\tau)).$$

If $a \in V(M)$ is a complete vector field, its corresponding flow $\gamma_a : \mathbb{R} \times M \to M$ satisfies the differential equation

$$\frac{d}{dt}\gamma_a(t)x = a(\gamma_a(t)x), \qquad \gamma_a(0)x = x.$$

In Krener and Lesiak [15] (see also Crouch [8]), the kernels are shown to be given inductively by the equations

$$W_0^i(t, x) = h_i(\gamma_f(t)x),$$

(1) $$W_n^{i_0 i_1 \cdots i_n}(t, \sigma_1 \cdots \sigma_n, x) = \gamma_f(-\sigma_n)_* g_{i_n}(\gamma_f(\sigma_n)x)(W_{n-1}^{i_0 i_1 \cdots i_{n-1}}(t, \sigma_1 \cdots \sigma_{n-1}, \cdot))$$

Thus, by simple manipulation,

$$W_n^{i_0 i_1 \cdots i_n}(t, \sigma_1 \cdots \sigma_n, x) = g_{i_n}(\gamma_f(\sigma_n)x)(g_{i_{n-1}}(\gamma_f(\sigma_{n-1} - \sigma_n)\cdot)(\cdots h_{i_0}(\gamma_f(t - \sigma_1)\cdot)\cdots)).$$

In particular,

(2) $$W_n(t - s, \sigma_1 - s, \cdots \sigma_n - s, \gamma_f(s)x) = W_n(t, \sigma_1 \cdots \sigma_n, x).$$

2.3. In describing the main results for nonlinear control systems, we shall use the terminology of Sussmann [20] as far as possible. Thus, an analytic system as described above is minimal if and only if it is orbit minimal and observable. For analytic systems, orbit minimality is equivalent to accessibility; that is, the reachable set from any point in the state space M has nonempty interior in M.

Two states $x_0, x_1 \in M$ are said to be indistinguishable if the input–output maps they define, as initial states of a system, are identical. A system is observable if any pair of indistinguishable states x_0 and x_1 satisfy $x_0 = x_1$; for analytic systems, indistinguishability is an equivalence relation on M.

A system which has a given input–output is called a realization of that input–output map. The main theorem in Sussmann [20] guarantees the existence and uniqueness of minimal realizations of an input–output map. That is, given an analytic realization of an input–output map, there exists a minimal realization of the same input–output map, and if

(3) $$\begin{aligned} \dot{x}_i &= f_i(x_i) + \sum_{j=1}^{m} u_j g_{ji}(x_i), \qquad x_i(0) = x_i^0, \quad x_i \in M_i^{n_i}, \\ y &= h_i(x_i) \end{aligned}$$

are two minimal analytic realizations Σ_i, $i = 1, 2$ of the same input–output map, then there exists a unique analytic diffeomorphism $\Phi : M_1 \to M_2$ satisfying

(4) $$\Phi_* f_1 = f_2 \circ \Phi, \qquad \Phi_* g_{j1} = g_{j2} \circ \Phi, \qquad h_1 = h_2 \circ \Phi, \qquad \Phi(x_1^0) = x_2^0.$$

2.4. The accessibility property does not rule out an implicit time dependency in the autonomous system of § 2.1, since time-varying systems can fit the definition by a standard addition to the state equations. The property is therefore strengthened to strong accessibility; that is, from any point x in the state space M there exists a time $T(x)$, $0 < T < \infty$ such that the reachable set from x at time T, $R(T, x)$, has nonempty

interior in M. Both accessibility properties have algebraic characterizations as given in Sussman [22].

The linear space of vector fields $V(M)$ becomes a Lie algebra under the bracket operation

$$[a, b](\tau) = a(b(\tau)) - b(a(\tau)),$$

for $a, b \in V(M)$ and $\tau \in C(M)$. Let $\mathscr{L}$ denote the Lie subalgebra of vector fields generated by f and $g_1 \cdots g_m$ and let $\mathscr{S}$ denote the ideal in $\mathscr{L}$ generated by $g_1 \cdots g_m$.

The analytic system is accessible if and only if

$$T_x M = \mathscr{L}(x) = \{a(x);\ a \in \mathscr{L}\} \quad \forall x \in M$$

where $T_x M$ is the tangent space to M at x. The analytic system is strongly accessible if and only if

$$T_x M = \mathscr{S}(x) = \{a(x);\ a \in \mathscr{S}\} \quad \forall x \in M.$$

In this case, $R(T, x)$ has nonempty interior in M for all $T > 0$.

An important example of a strongly accessible system is an accessible stationary system, that is, an autonomous system in which $f(x_0) = 0$. It is easily deduced from (2) that if a system is stationary, then the Volterra kernels and series are stationary, in the sense that

$$W_n(t, \sigma_1 \cdots \sigma_n, x_0) = W_n(t - s, \sigma_1 - s, \cdots, \sigma_n - s, x_0).$$

2.5. The concept of weak observability was introduced by Krener and Hermann [14], and for analytic systems is a useful weakening of the concept of observability. A system is weakly observable if for all states $x_0 \in M$ there exists a neighborhood U of x_0, such that if $x_1 \in U$ is indistinguishable from x_0, then $x_0 = x_1$. Clearly, if a system is observable then it is weakly observable.

Let $\mathscr{H}$ denote the smallest linear subspace of $C(M)$ containing the functions h_i, $i = 1 \cdots q$ and closed under Lie differentiation by elements of $\mathscr{L}$. Thus, $\mathscr{H}$ consists of all linear combinations of the functions

$$L_{a_1}(L_{a_2}(\cdots(L_{a_n}(h_i)\cdots))), \qquad a_i \in \mathscr{L}.$$

In Krener and Hermann [14], it is shown that an accessible system is weakly observable if and only if

$$T_x M^* = d\mathscr{H}(x) = \{d\tau(x);\ \tau \in \mathscr{H}\} \quad \forall x \in M,$$

where $T_x M^*$ is the cotangent space to M at x.

Assume the systems Σ_1, Σ_2 as defined in (3) are such that Σ_2 is a minimal realization of an input–output map and Σ_1 is an accessible weakly observable realization. As in Sussmann [20] the canonical map $\pi' : M_1 \to M_1/R$ is closed and regular, where R is the equivalence relation of indistinguishability, and M_1/R inherits a minimal system with the same input–output map as Σ_2. Moreover π' satisfies relations like (4). Since minimal realizations are isomorphic, there exists an analytic map $\pi : M_1 \to M_2$ satisfying the relations (4). The following result is a slight adaptation of a theorem in Sussmann [21], using a technique due to Krener [16].

LEMMA 2.1. *M_1^n is a covering manifold of M_2^n and π is the covering projection.*

Proof. Since Σ_1 is weakly observable if $x \in M_2$, then $\pi^{-1}(x)$ is a discrete subset of M_1. Thus, it is sufficient to show that if $x \in M_2$, then there is a neighborhood U of x such that $U \times \pi^{-1}(x)$ is diffeomorphic to $\pi^{-1}(U)$.

Let $y \in \pi^{-1}(x)$. As in Krener [16], the accessibility of Σ_1 implies that there are n associated vector fields a_i with flows γ_i, and a neighborhood V of $0 \in \mathbb{R}^n$ such that the map σ,

$$(s_1 \cdots s_n) \to \gamma_1(s_1) \circ \cdots \circ \gamma_n(s_n)y,$$

is a diffeomorphism of V onto some open neighborhood S of y.

Since $\pi_* T_y M_1 = T_x M_2$, by restricting V if necessary, it can be assumed that π maps S diffeomorphically onto a neighborhood U of x.

Let $p : U \to V$ be the inverse of $\pi \circ \sigma$.

Since the vector fields a_i are complete, the following maps $\psi, \phi : V \times M_1 \to M_1$ are analytic, where

$$\phi(s_1 \cdots s_n, u) = \gamma_1(s_1) \circ \cdots \circ \gamma_n(s_n)u,$$
$$\psi(s_1 \cdots s_n, u) = \gamma_n(-s_n) \circ \cdots \circ \gamma_1(-s_1)u.$$

A simple consequence of indistinguishability shows that $\phi : V \times \pi^{-1}(x) \to \pi^{-1}(U)$, and so the following maps are analytic and inverses of each other.

$$\Phi : U \times \pi^{-1}(x) \to \pi^{-1}(U), \qquad \Sigma : \pi^{-1}(U) \to U \times \pi^{-1}(x),$$

where

$$\Phi(z, u) = \phi(p(z), u), \Sigma(v) = (\pi(v), \psi(p \circ \pi(v), v)). \qquad \text{Q.E.D.}$$

Conversely, if Σ_2 is a minimal realization of an input–output map with state space M_2, we can construct an accessible weakly observable realization of the input–output map on any covering manifold M_1 of M_2. This is done in a manner analogous to the construction of an accessible weakly observable realization, on the simply connected cover of M_2 in Krener [13]. (See also Crouch [8]). This system, called the simply connected cover of the minimal system, is used extensively in the following section. The preceding remarks yield the final result of the section.

THEOREM 2.2. *The accessible weakly observable analytic realizations of an input–output map are in one-to-one correspondence with the covering manifolds of the state space of a minimal realization.*

3. The state space.

3.1. In this section, the Lie algebra of a strongly accessible weakly observable realization of a finite Volterra series is characterized. From § 2.5, it is clear that this Lie algebra is isomorphic to the Lie algebra of a minimal realization.

In this paper, the Campbell–Baker–Hausdorff formula is used repeatedly. It states that for $a, b \in V(M)$

$$\gamma_a(-s)_* b(\gamma_a(s)x) = \sum_{i=1}^{\infty} \frac{s^i}{i!} ad^i a(b)(x),$$

where $ad^i a(b) = ad^{i-1} a([a, b])$, $ad^0 a(b) = b$, and this series converges for s in some neighborhood of $0 \in \mathbb{R}$. From the kernel structure equations (1), it is now clear that the kernel components are analytic in all their arguments.

Let $\mathcal{R}$ denote the subspace of $\mathcal{S}$ and $\mathcal{L}$ spanned by the vector fields $ad^i f(g_j)$ for $i \geqq 0$ and $1 \leqq j \leqq m$. It is clear that $adf : \mathcal{R} \to \mathcal{R}$ is a linear endomorphism, and that $\mathcal{R}$ generates the subalgebra $\mathcal{S}$ of $\mathcal{L}$. The following lemma describes some important properties of the kernels.

LEMMA 3.1. *Given a strongly accessible realization of a finite Volterra series of length p*:

a) *The Volterra series has length p when evaluated at any point in the state space M.*

b) *The kernel W_p depends only on the time parameters $t, \sigma_1 \cdots \sigma_p$, not on the state $x \in M$.*

Proof. That the Volterra series has length p, implies that $W_{p+i}(t, \sigma_1 \cdots \sigma_{p+i}, x_0) \equiv 0$ for a fixed initial condition x_0 and $i \geqq 1$. The formulas for the kernels for $i = 1$ give

$$\gamma_f(-\sigma_{p+1})_* g_j(\gamma_f(\sigma_{p+1})x_0)(W_p(t, \sigma_1 \cdots \sigma_p, \cdot)) \equiv 0.$$

Differentiating with respect to σ_{p+1} repeatedly yields

$$L_a W_p(t, \sigma_1 \cdots \sigma_p, x_0) \equiv 0,$$

where a is any vector field in $\mathscr{R}$.

Similarly, for $i \geqq 1$,

$$L_{a_1} L_{a_2} \cdots L_{a_i} W_p(t, \sigma_1 \cdots \sigma_p, x_0) \equiv 0,$$

where a_j, $1 \leqq j \leqq i$ are any vector fields in $\mathscr{R}$.

Using the identity

$$L_a L_b(\tau) - L_b L_a(\tau) = L_{[a, b]}(\tau),$$

and taking suitable linear combinations of the above equations, give

$$L_{a_1} L_{a_2} \cdots L_{a_i} W_p(t, \sigma_1 \cdots \sigma_p, x_0) \equiv 0, \qquad i \geqq 1, \tag{5}$$

where a_j, $1 \leqq j \leqq i$ are any vector fields in $\mathscr{S}$.

By strong accessibility there exist vector fields $a_1 \cdots a_n \in \mathscr{S}$ which span $T_{x_0}M$; and so if γ_i is the flow of a_i, $(s_1 \cdots s_n) \to \gamma_1(s_1) \circ \cdots \circ \gamma_n(s_n)x_0$ maps a neighborhood of $0 \in \mathbb{R}^n$ onto some neighborhood of $x_0 \in M$. It follows from (5) that all derivatives of the map $(s_1 \cdots s_n) \to W_p(t, \sigma_1 \cdots \sigma_n, \gamma_1(s_1) \circ \cdots \circ \gamma_n(s_n)x_0)$, vanish at $0 \in \mathbb{R}^n$. By the analytic dependence of W_p on the state, it follows that W_p does not depend on the state and thereby proves part b). To prove part a) the above procedure is repeated for W_{p+i}, $i \geqq 1$, to show that these kernels also do not depend on the state. Since they are identically zero at x_0, part a) follows immediately. Q.E.D.

Using the information above, the Lie algebra $\mathscr{L}$ is characterized, including a result already obtained in Brockett [2] for bilinear realizations.

THEOREM 3.2. *Given a strongly accessible weakly observable realization of a finite Volterra series of length p, $\mathscr{S}$ is a nilpotent Lie algebra, with a descending central series of length less than or equal to p, and the Lie algebra $\mathscr{L}$ is solvable.*

Proof. By Lemma 3.1 the kernels W_{p+i}, $i \geqq 1$ are identically zero on M. Therefore, proceeding as in Lemma 3.1 gives

$$L_{a_1} L_{a_2} \cdots L_{a_{p+1}}(h_k \circ \gamma_f(t)x) \equiv 0,$$

where a_j, $1 \leqq j \leqq p+1$ are vector fields in $\mathscr{S}$.

Differentiating repeatedly with respect to t and taking more linear combinations gives the following results:

$$L_{a_1} L_{a_2} \cdots L_{a_{p+1}} L_{b_1} \cdots L_{b_j}(h_k) = 0, \qquad j \geqq 0, \tag{6}$$

where a_l, $1 \leqq l \leqq p+1$ and b_j are vector fields in $\mathscr{S}$ and $\mathscr{L}$, respectively.

By definition of $\mathscr{H}$, this can be rewritten as

$$L_{a_1} L_{a_2} \cdots L_{a_{p+1}}(\tau) = 0, \tag{7}$$

where a_l, $1 \leqq l \leqq p+i$ are vector fields in $\mathcal{S}$ and τ is any function in $\mathcal{H}$.

Defining $\mathcal{S}^i = [\mathcal{S}, \mathcal{S}^{i-1}]$, $\mathcal{S}^1 = \mathcal{S}$, the descending central series of $\mathcal{S}$, yields that

$$L_a(\tau) = 0 = d\tau(a),$$

for arbitrary $a \in \mathcal{S}^{p+1}$ and $\tau \in \mathcal{H}$.

By weak observability, it follows immediately that $\mathcal{S}^{p+1} = \{0\}$. Thus, $\mathcal{S}$ is nilpotent with a descending central series of length less than or equal to p.

Defining $\mathcal{L}^{(i)} = [\mathcal{L}^{(i-1)}, \mathcal{L}^{(i-1)}]$, $\mathcal{L}^{(1)} = \mathcal{L}$, the derived series, makes it clear that $\mathcal{L}^{(i+1)} \subset \mathcal{S}^{(i)}$, so that $\mathcal{L}^{(p+2)} \subset \mathcal{S}^{(p+1)} \subset \mathcal{S}^{2^p} = \{0\}$. $\mathcal{L}$ is therefore a solvable Lie algebra. Q.E.D.

The following minimal systems have Volterra series of length p for any $p \geqq 1$, but all have a descending central series of length one.

$$\dot{x} = u, \qquad x(0) = 0, \quad x \in \mathbb{R},$$
$$y = x^p.$$

The next result shows that $\mathcal{L}$ is in fact finite dimensional. This result is implied by a result in Brockett [2] for stationary finite Volterra series, which shows that such a Volterra series has a finite dimensional analytic realization if and only if it has a finite dimensional bilinear realization. The Lie algebra of a finite dimensional bilinear realization is obviously finite dimensional.

LEMMA 3.3. *A strongly accessible weakly observable realization of a finite Volterra series has a finite dimensional Lie algebra.*

Proof. By Theorem 3.2, $\mathcal{S}$ is nilpotent with a descending central series of length at most p. Thus, if $a \in \mathcal{S}$, $ada : \mathcal{S} \to \mathcal{S}$ is a nilpotent endomorphism of $\mathcal{S}$ satisfying $ad^{p+1}a = 0$. By strong accessibility $\mathcal{S}(x_0) = T_{x_0}M$, so there exists vector fields $a_i \in \mathcal{S}$ such that $a_1(x_0) \cdots a_n(x_0)$ span $T_{x_0}M$, and there exists neighborhoods U of x_0 and V of $0 \in \mathbb{R}^n$, such that the map

$$s_1 \cdots s_n \to \gamma_{a_1}(s_1) \circ \cdots \circ \gamma_{a_n}(s_n)x_0$$

is a diffeomorphism of V onto U. Denote this map by $s \to \Phi(s)x_0$ and let $\Phi(-s)$ be the inverse of the diffeomorphism $\Phi(s)$. By the Campbell–Baker–Hausdorff formula and the nilpotence of $\mathcal{S}$,

$$\Phi(-s)_* a(\Phi(s)x_0) = \sum_{i=1}^{n} p_i(s)a_i(x_0), \qquad a \in \mathcal{S},$$

where $p_i(s)$ are polynomials in the finite number of variables $s_1 \cdots s_n$ and of order at most p in each variable. Thus on U, each $a \in \mathcal{S}$ can be written as

$$a(\Phi(s)x_0) = \sum_{i=1}^{n} p_i(s)\Phi(s)_* a_i(x_0).$$

However, the linear space of polynomials of order less than or equal to p in the finite number of variables $s_1 \cdots s_n$ is finite dimensional, so that $\mathcal{S}$ restricted to U is finite dimensional. Analyticity shows that $\mathcal{S}$ is finite dimensional on M, and $\mathcal{L}$ is finite dimensional since $\mathcal{S}$ has at most codimension 1 in $\mathcal{L}$. Q.E.D.

If the system is not strongly accessible this result does not follow, as the following example illustrates:

$$\dot{x}_0 = 1, \qquad \dot{x}_1 = u/(1 + x_0^2).$$

In the following section, these results are used to formulate state spaces as homogeneous spaces.

3.2. It is first noted that a strongly accessible weakly observable realization of a finite Volterra series has a finite dimensional Lie algebra $\mathscr{L}$, by Lemma 3.3, which by assumption is generated by a finite number of complete vector fields. A theorem from Palais [19] shows that these are sufficient conditions for $\mathscr{L}$ to consist of complete vector fields.

Consider a strongly accessible weakly observable realization of a finite Volterra series with state space M. Let G' be the group of diffeomorphisms of M generated by the flows of the (complete) vector fields in $\mathscr{L}$.

G' acts naturally on M as a transformation group:

$$\Phi: G' \times M \to M, \qquad (g, x) \to g \cdot x.$$

If G' can be given a (connected) Lie group structure, and then denoted by G such that Φ is analytic, then G is called a connected Lie transformation group. The following result from Palais [19] is central to the problem formulation. Every finite dimensional Lie algebra $\mathscr{L}$ of complete vector fields on a manifold M is isomorphic to the Lie algebra of a unique connected Lie transformation group G on M such that the flow of any $a \in \mathscr{L}$ is given by

$$(t, x) \to \exp ta' \cdot x,$$

where a' is the unique element of $L(G)$, the Lie algebra of G, corresponding to a, and $\exp : L(G) \to G$ is the exponential map. This shows that G' is the underlying space for a unique connected Lie transformation group G. $L(G)$ and $\mathscr{L}$ are henceforth identified.

Since the system is orbit minimal (analytic and accessible), G acts transitively on M. By a standard result (Hochschild [11]), M is therefore analytically diffeomorphic to the homogeneous space G/G_{x_0} where G_{x_0} is the isotropy group

$$G_{x_0} = \{g: g \in G, g \cdot x_0 = x_0\}.$$

Moreover, the action of G on M is effective ($g \cdot x = x, \forall x \in M \Rightarrow g = \text{identity}$) so that G_{x_0} is a closed Lie subgroup of G containing no nontrivial normal subgroups of G.

In fact, since the system is strongly accessible, the connected Lie subgroup N of G, corresponding to $\mathscr{S}$, also acts transitively on M, and so M can also be expressed as a homogeneous space N/N_{x_0}.

In the next section, this formulation is extended with particular reference to the simply connected cover of a minimal realization of a finite Volterra series.

3.3. Consider an observable strongly accessible realization of a finite Volterra series and its simply connected cover, with state spaces M and $\tilde{M}$, respectively, and let $\tilde{\pi}: \tilde{M} \to M$ be the covering projection. Let G and $\tilde{G}$ be the corresponding connected Lie transformation groups, with natural actions defined by the maps ϕ and $\tilde{\phi}$, respectively. Since the Lie algebras of G and $\tilde{G}$ are isomorphic to $\mathscr{L}$, the simply connected covering groups of G and $\tilde{G}$ are isomorphic and will be identified and denoted by G^*. Letting $\Pi: G^* \to G$ and $\tilde{\Pi}: G^* \to \tilde{G}$ be the covering homomorphisms, transitive actions of G^* are obtained on M and $\tilde{M}$ in the following ways:

$$G^* \times M \to M, \qquad (g^*, x) \to \phi(\Pi(g^*), x), \tag{8}$$

$$G^* \times \tilde{M} \to \tilde{M}, \qquad (g^*, \tilde{x}) \to \tilde{\phi}(\tilde{\Pi}(g^*), \tilde{x}). \tag{9}$$

These actions are related by the identity

$$\pi \circ \tilde{\phi}(\tilde{\Pi}(g^*), \tilde{x}) = \phi(\Pi(g^*), \pi(\tilde{x})). \tag{10}$$

Since G acts effectively on M, the above identity shows that $\operatorname{Ker}\tilde{\Pi} \subset \operatorname{Ker}\Pi$. Since $\operatorname{Ker}\Pi$ and $\operatorname{Ker}\tilde{\Pi}$ are closed discrete central subgroups of G^* such that

$$G \cong G^*/\operatorname{Ker}\Pi, \qquad \tilde{G} \cong G^*/\operatorname{Ker}\tilde{\Pi},$$

it follows that

$$G \cong \tilde{G}/(\operatorname{Ker}\Pi/\operatorname{Ker}\tilde{\Pi}).$$

Thus $\tilde{G}$ is a covering group of G with an associated transitive action on M given by

$$\tilde{G} \times M \to M, \qquad (\tilde{g}, x) \to \phi(\Pi'(\tilde{g}), x), \tag{11}$$

where $\Pi' : \tilde{G} \to G$ is the covering homomorphism and

$$\pi \circ \tilde{\phi}(\tilde{g}, \tilde{x}) = \phi(\Pi'(\tilde{g}), \pi(\tilde{x})). \tag{12}$$

If $x_1 \in \pi^{-1}(x_0)$ is set as an initial state for the simply connected cover, the transitive actions (8), (9) and (11) impart the following alternative homogeneous space structures for M and $\tilde{M}$;

$$M \cong G^*/G^*_{x_0}, \qquad \tilde{M} \cong G^*/G^*_{x_1}, \qquad M \cong \tilde{G}/\tilde{G}_{x_0}.$$

Since $\tilde{M}$ is simply connected $G^*_{x_1}$ is connected, (Chevalley [6, Corollary 1, p. 59]).

Identity (10) shows that $G^*_{x_0} \supset G^*_{x_1}$, and so $G^*_{x_0}/G^*_{x_1}$ is isomorphic to the fundamental group of M.

In the next section, some relations between these structures are explored.

3.4. A strongly accessible realization guarantees the existence of a vector field $a_0 \in \mathscr{S}$ such that $(a_0 + f)(x_0) = 0$, since $\mathscr{S}(x_0) = T_{x_0}M$. $\mathscr{L}$ can therefore be viewed as the semidirect product of the one-dimensional space spanned by $a_0 + f$ and $\mathscr{S}$. Correspondingly G^* can be considered as the semidirect product V^*N^*, where V^* is the one-dimensional Lie subgroup of G^* with generator $a_0 + f$ and N^* is the subgroup of G^* corresponding to $\mathscr{S}$. In fact, in Chevalley [7] it is shown that a connected Lie subgroup of a simply connected, connected solvable Lie group is closed and simply connected. By Theorem 3.2, $\mathscr{L}$ and hence, G and G^* are solvable, so that V^* and N^* are both closed and simply connected.

Noting the actions of G^* on M and $\tilde{M}$ makes it clear that both $G^*_{x_0}$ and $G^*_{x_1}$ contain V^*, so that

$$G^*_{x_0} = V^*(N^* \cap G^*_{x_0}), \qquad G^*_{x_1} = V^*(N^* \cap G^*_{x_1}).$$

It is now clear that since N^* is nilpotent M and $\tilde{M}$ are nilmanifolds, or homogeneous spaces of nilpotent Lie groups,

$$M \cong N^*/N^* \cap G^*_{x_0}, \qquad \tilde{M} \cong N^*/N^* \cap G^*_{x_1}.$$

To relate the structure of $\tilde{M}$ as a homogeneous space of both $\tilde{G}$ and G^*, the following preliminary results are needed.

LEMMA 3.4. *If D is a discrete central subgroup of G^* contained in $G^*_{x_1}$ then $D \cap N^*$ is the trivial group $\{e\}$.*

Proof. Assume to the contrary that $e \neq d \in D$ and $d \in D \cap N^*$. Since N^* is a connected and simply connected nilpotent Lie group, the center of N^* is connected and the exponential map is bijective (Hochschild [11]). Thus, there exists a vector field $a \in \mathscr{S}$ such that $\exp a = d$ and a belongs to the center of $\mathscr{S}$. Thus, the connected subgroup $N^* \cap G^*_{x_1}$ of N^* contains the central one-parameter subgroup of N^* with generator a. However, since G and hence, N, acts effectively on M, N^* acts almost effectively on $\tilde{M}$,

and so $N^* \cap G^*_{x_1}$ contains at most a discrete normal subgroup of N^*. This contradicts the fact that $a \neq 0$, and so $D \cap N^* = \{e\}$. Q.E.D.

THEOREM 3.5. *If D is a nontrivial discrete central subgroup of G^* contained in $G^*_{x_1}$, then there exists $a_0 \in \mathcal{S}$ so that $D \subset V^*$, where V^* is the one-parameter subgroup of G^* with generator $a_0 + f$.*

Proof. By Chevalley [7], if D is a discrete central subgroup of a simply connected, connected solvable Lie group G^*, then there exists a basis $a_1 \cdots a_m$ for the Lie algebra of G^*, such that every element of G^* has one and only one representation of the form

$$\exp t_1 a_1 \cdots \exp t_n a_n, \qquad t_i \in \mathbb{R},$$

and each element of D has one and only one representation of the form

$$\exp n_1 a_1 \cdots \exp n_r a_r, \qquad n_i \in \mathbb{Z}, \quad 1 \leqq i \leqq r \leqq n, \quad \text{where } [a_i, a_j] = 0,$$

for $1 \leqq i, j \leqq r$. In particular if R represents the vector group $\{\exp t_1 a_1 \cdots \exp t_r a_r, t_i \in \mathbb{R}, 1 \leqq i \leqq r\}$, then R/D is a compact Abelian subgroup of G^*/D. In the given situation each a_i can be expressed in the form $\alpha_i f + b_i$ where $b_i \in \mathcal{S}$ and $0 \neq \alpha_i \in \mathbb{R}$; otherwise $\exp n_i b_i \in D \cap N^*$, which contradicts Lemma 3.4. If $r = 1$, there is nothing left to prove, so assume $r > 1$. Since R/D is compact, the adjoint representation of R/D is semisimple. In this representation, the one-parameter subgroups V_i/D_i, where $V_i = \{\exp t_i(\alpha_i f + b_i);\ t_i \in \mathbb{R}\}$ and $D_i = \{\exp n_i(\alpha_i f + b_i);\ n_i \in \mathbb{Z}\}$, are mapped into the one-parameter groups V_i^* of automorphisms of the Lie algebra of G^*/D given by

$$V_i^* = \{\exp t\, ad(\alpha_i f + b_i);\ t \in \mathbb{R}\}.$$

Note that $ad(\alpha_i f + b_i)$ leaves $\mathcal{S}^j$ invariant for $1 \leqq j \leqq p$ (p is the length of the descending central series of $\mathcal{S}$) and induces endomorphisms on $\mathcal{S}^j/\mathcal{S}^{j+1}$, $1 \leqq j \leqq p-1$ and $\mathcal{L}/\mathcal{S}^1$ which are equal to the endormorphisms induced by $ad\alpha_i f$. It follows that since the representation of R/D is semisimple, it is equivalent to the induced representation on

$$\mathcal{L}/\mathcal{S}^1 \oplus \mathcal{S}^1/\mathcal{S}^2 \oplus \cdots \oplus \mathcal{S}^{p-2}/\mathcal{S}^{p-1} \oplus \mathcal{S}^p,$$

in which V_i^* is the group of automorphisms induced by $\{\exp t ad\alpha_i f,\ t \in \mathbb{R}\}$. In particular, $f + b_i/\alpha_i = f + c_i$, $1 \leqq i \leqq r$, where $b_i = c_i \alpha_i$ belongs to the center of $\mathcal{L}$. Since $G^*_{x_1}$ is connected, $f + b_i/\alpha_i \in L(G^*_{x_1})$, so by the linear independence of $a_1 \cdots a_r$, $0 \neq c_j - c_1 \in L(G^*_{x_1})$ for $2 \leqq j \leqq r$ belong to the center of $\mathcal{L}$. Since G^* acts almost effectively on $\tilde{M}$, this is a contradiction and $r = 1$. Q.E.D.

Since $\tilde{M}$ has the two representations

$$\tilde{M} \cong \tilde{G}/\tilde{G}_{x_1} \cong G^*/G^*_{x_1},$$

if D is the discrete central subgroup of G^* such that $G^*/D \cong \tilde{G}$, then $D \subset G^*_{x_1}$, and so by Theorem 3.5 there exists $a_0 \in \mathcal{S}$ such that $D \subset V^*$.

Hence, $\tilde{G} = \tilde{V}\tilde{N}$ where $\tilde{N} \cong N^*$ and $\tilde{V} \cong V^*/D$.

It is now possible to view M and $\tilde{M}$ as homogeneous spaces of $\tilde{N}$, the connected, simply connected nilpotent Lie transformation group of $\tilde{M}$ with Lie algebra $\mathcal{S}$. Write

$$M \cong \tilde{N}/\tilde{N}_{x_0}, \qquad \tilde{M} \cong \tilde{N}/\tilde{N}_{x_1}.$$

3.5. In this section, the structure theorems of nilmanifolds as in Malcev [17] and Matsushima [18] are applied to the above situation. It will be assumed that M is not simply connected, so that M and its simply connected cover $\tilde{M}$ do not coincide. Proceeding as in Matsushima [18], notice that $\tilde{N}_{x_1}$ is the connected component of the identity in $\tilde{N}_{x_0}$ and, in particular, normal in $\tilde{N}_{x_0}$. Let R be the normalizer of $\tilde{N}_{x_1}$ in $\tilde{N}$, so

that R is connected and $\tilde{N}_{x_0} \subset R$. Now

$$R/\tilde{N}_{x_0} \cong (R/\tilde{N}_{x_1})/(\tilde{N}_{x_0}/\tilde{N}_{x_1}),$$

where $R/\tilde{N}_{x_1} = S$ is a connected simply connected nilpotent Lie group and $\tilde{N}_{x_0}/\tilde{N}_{x_1} = D'$ is a discrete subgroup of S.

Applying Matsushima [18, Theorem 1] to S and D' shows that there exists a basis $a_1^* \cdots a_n^*$ of $L(S)$ with the properties

$$\mathrm{Sp}\{a_{i+1}^* \cdots a_n^*\} \text{ is an ideal of } \mathrm{Sp}\{a_i^* a_{i+1}^* \cdots a_n^*\},$$

and hence, every element of S may be expressed uniquely in the form

$$\exp t_1 a_1^* \cdots \exp t_n a_n^*, \qquad t_i \in \mathbb{R}.$$

There exists an integer m, $1 \leqq m \leqq n$, such that every element of D' may be expressed uniquely in the form

$$g_m^{*s_m} \cdots g_n^{*s_n}, \qquad s_i \in \mathbb{Z},$$

where $g_k^* = \exp a_k^*$, $m \leqq k \leqq n$.

Let $\mathcal{N}$ be the Lie algebra of $\tilde{N}_{x_1}$ and let g_k, $m \leqq k \leqq n$ be representatives in R for the elements g_k^*. If $a_k \in \mathcal{S}$ is such that $\exp a_k = g_k$, then $\exp(a_k + \mathcal{N}) = g_k^*$, and by the bijectiveness of the exponential map for a simply connected nilpotent Lie group, S, $a_k + \mathcal{N} = a_k^*$.

In particular, if $a_{n+1} \cdots a_{n+s}$ is a basis for $\mathcal{N}$, $\mathrm{Sp}\{a_{i+1} \cdots a_{n+s}\}$ is an ideal in $\mathrm{Sp}\{a_i a_{i+1} \cdots a_{n+s}\}$ for $m \leqq i \leqq n-1$. Let $\mathcal{U}$ be the Lie algebra $\mathrm{Sp}\{a_m \cdots a_{n+s}\}$, and U the corresponding connected simply connected Lie subgroup of $\tilde{N}$. Then every element of U can be written uniquely in the form

$$\exp t_m a_m \cdots \exp t_{n+s} a_{n+s}, \qquad t_i \in \mathbb{R}.$$

Moreover, every element in $\tilde{N}_{x_0}$ can be written uniquely in the form

$$\exp s_m a_m \cdots \exp s_n a_n \exp t_{n+1} a_{n+1} \cdots \exp t_{n+s} a_{n+s},$$

where $s_i \in \mathbb{Z}$, $t_i \in \mathbb{R}$, and $\tilde{N}_{x_1}$ is obtained by setting $s_i = 0$ for $m \leqq i \leqq n$.

Now if U is a subgroup of a connected nilpotent Lie group $\tilde{N}$, then U is properly contained in the normalizer of U in N, and if U is connected its normalizer is connected. Thus, the basis of $\mathcal{U}$ can be completed to a basis of $\mathcal{S} = \mathrm{Sp}\{a_1 \cdots a_m \cdots a_n \cdots a_{n+s}\}$ such that $\mathrm{Sp}\{a_{i+1} \cdots a_{n+s}\}$ is an ideal in $\mathrm{Sp}\{a_i a_{i+1} \cdots a_{n+s}\}$, and hence, every element of $\tilde{N}$ can be expressed uniquely in the form

$$\exp t_1 a_1 \cdots \exp t_{n+s} a_{n+s}, \qquad t_i \in \mathbb{R}.$$

It is now clear that $\tilde{M} \cong \tilde{N}/\tilde{N}_{x_1} \simeq V_1 \times V_2 \times \cdots \times V_n \cong \mathbb{R}^n$, where $V_i = \{\exp t a_i;\ t \in \mathbb{R}\}$. Thus, the state space of the simply connected cover is homeomorphic to a Cartesian space.

Moreover,

$$M \cong \tilde{N}/\tilde{N}_{x_0} \simeq V_1 \times \cdots \times V_m \times U/\tilde{N}_{x_0} \simeq \mathbb{R}^m \times U/\tilde{N}_{x_0}.$$

As in Matsushima [18, Theorem 2] and also Malcev [17], $U/\tilde{N}_{x_0}$ is a compact nilmanifold, and so the state space of the minimal system is homeomorphic to the product of a Cartesian space and a compact nilmanifold.

3.6. In this section, the preceding results are combined to characterize the structure of the state space of a strongly accessible observable realization of a finite Volterra series.

The basis for $\mathcal{S}$ constructed in § 3.5 and the corresponding vector fields on $\tilde{M}$ yield the differentiable homeomorphism $\Phi: \mathbb{R}^n \to \tilde{M}$ given by

$$\Phi(t_1 \cdots t_n) = \gamma_{a_1}(t_1) \circ \cdots \circ \gamma_{a_n}(t_n) x_1.$$

If $\mathcal{H}$ is the linear space of functions on M describing local observability for the minimal system, then $\tilde{\mathcal{H}} = \pi^* \mathcal{H} = \{\tau \circ \pi; \tau \in \mathcal{H}\}$ describes the corresponding space of functions on $\tilde{M}$ for the simply connected cover, as is easily verified from the results of § 2.5.

PROPOSITION 3.6. *The functions $\tilde{\tau} \circ \Phi$, $\tilde{\tau} \in \tilde{\mathcal{H}}$ are polynomials and there exist n functions $\tilde{\tau}_i \in \mathcal{H}$ such that $\tilde{\tau}_i \circ \Phi(0 \cdots t_i 0 \cdots 0)$ are nonconstant polynomials in t_i for $1 \leqq i \leqq n$.*

Proof. Consider the derivatives for $\sum_{i=1}^n k_i = p + m$:

$$\frac{\partial^{p+m}}{\partial t_1^{k_1} \cdots \partial t_n^{k_n}} \Big|_{t_1 = \cdots = t_n = 0} \tilde{\tau} \circ \Phi(t_1 \cdots t_n) = L_{a_n}^{k_n} L_{a_{n-1}}^{k_{n-1}} \cdots L_{a_1}^{k_1} \tilde{\tau}(x_1),$$

where $L_{a_i}^{k_i}$ represents k_i repeated Lie differentiations by a_i. However, by (7), this expression is identically zero for all $m > 0$. Since all functions $\tilde{\tau} \circ \Phi$ are analytic, it is deduced that they are polynomials.

Assume now that for all $\tilde{\tau} \in \tilde{\mathcal{H}}$

$$0 = \frac{\partial}{\partial t_i} \tilde{\tau} \circ \Phi(0, \cdots t_i \cdots 0) \big|_{t_i = 0};$$

then

$$0 = \frac{\partial}{\partial t_i} \tilde{\tau} \circ \gamma_{a_i}(t_i) x_1 \big|_{t_i = 0} = d\tilde{\tau}(a_i)(x_1).$$

However, by construction of the basis vectors $a_i \notin \mathcal{N}$ for $1 \leqq i \leqq n$, so that $a_i(x_1) \neq 0$. Thus, $0 = d\tilde{\mathcal{H}} a_i(x_1)$, which contradicts the weak observability of the simply connected cover. Thus, there exists $\tilde{\tau}_i \in \tilde{\mathcal{H}}$ such that $\tilde{\tau}_i \circ \Phi(0, \cdots t_i \cdots 0)$ is a nonconstant polynomial in t_i, since it contains a term $\alpha_i t_i$ with $\alpha_i \neq 0$. Q.E.D.

The main theorem can now be stated.

THEOREM 3.7. *A strongly accessible observable realization of a finite Volterra series has a state space which is homeomorphic to a Cartesian space.*

Proof. From § 3.5, if the minimal system has a state space which is not simply connected, then it is homeomorphic to the product of a Cartesian space and a compact nilmanifold, so it is sufficient to show that the compact component reduces to a single point.

Assume to the converse that the covering projection $\pi: \tilde{M} \to M$ is not trivial; then by (12)

$$\pi \circ \tilde{\phi}(\tilde{g}, x_1) = x_0,$$

for all $\tilde{g} \in \tilde{N}_{x_0}$. Thus,

$$\pi \circ \Phi(0 \cdots 0\, s_m \cdots s_n) = x_0,$$

for all $s_i \in \mathbb{Z}$, $m \leqq i \leqq n$.

In particular, for $\tilde{\tau} = \tau \circ \pi \in \tilde{\mathcal{H}}$,

$$\tilde{\tau} \circ \Phi(0 \cdots 0 s_m \cdots s_n) - \tau(x_0) = 0.$$

Thus, by Proposition 3.6, for all i, $m \leqq i \leqq n$,

$$\tilde{\tau}_i \circ \Phi(0 \cdots 0 t_i \cdots 0) - \tau_i(x_0)$$

are nonconstant polynomials with an infinite number of zeros $s_i \in \mathbb{Z}$. This is clearly a contradiction, showing that M is homeomorphic to a Cartesian space. Q.E.D.

COROLLARY 3.8. *A strongly accessible analytic realization of a finite Volterra series is observable iff it is weakly observable. In particular, an analytic realization of a finite Volterra series is strongly accessible and observable iff, for all* $x \in M$,

$$T_x M = \mathcal{S}(x), \qquad T_x M^* = d\mathcal{H}(x).$$

Proof. It is sufficient to show that a weakly observable strongly accessible system Σ_1 is observable. However, by Lemma 2.1, such a system has a state space M_1 which is a covering space of the state space M_2 of a minimal strongly accessible realization Σ_2. By Theorem 3.7, M_2 is simply connected so that the covering projection is a diffeomorphism, and it is concluded that Σ_2 is isomorphic to Σ_1 and hence observable. Q.E.D.

This corollary shows that just as in the linear time-invariant case, minimality is specified by a set of algebraic conditions. Of course, to apply the conditions a coordinate system has to be used. A convenient coordinate system will be introduced in the next section.

THEOREM 3.9. *The connected Lie transformation group G of a strongly accessible, observable realization of a finite Volterra series has a decomposition as the semidirect product VN, where N is the connected simply connected nilpotent Lie group with Lie algebra $\mathcal{S}$, and V is a one-parameter subgroup with generator $a_0 + f$, $a_0 \in \mathcal{S}$, such that the isotropy subgroup of G is VN', where N' is a connected subgroup of N. In particular, G has a faithful matrix representation.*

Proof. The decomposition of G follows from the fact that the simply connected cover and minimal systems are isomorphic, as proved in Theorem 3.7, and the properties of $\tilde{G}$ derived in § 3.4. That G has a faithful matrix representation follows from Hochschild [11, Theorem 3.2, p. 220], since the commutator group $G' = [G, G]$ is a closed subgroup of $N \subset G$ which contains no nontrivial compact subgroups (Hochschild [11, Theorem 2.3, p. 138]). Q.E.D.

4. System structure.

4.1. Canonical realizations. In this section, canonical coordinate charts are chosen to express minimal realizations of finite Volterra series using the theory developed in the preceding sections and a generalization of the methods of Chen [5] and Krener [13] for decomposing systems of differential equations with nilpotent Lie algebra.

The coordinate charts are constructed from the flows of vector fields selected from the following sequence of subalgebras of $\mathcal{S}$, where $\mathcal{N}$ is the subalgebra of $\mathcal{S}$ consisting of those vector fields which vanish at x_0:

$$\mathcal{S} \supset \mathcal{S}^2 + \mathcal{N} \supset \cdots \supset \mathcal{S}^p + \mathcal{N} \supset \mathcal{N}.$$

It is easily checked that each $\mathcal{S}^i + \mathcal{N}$ is a subalgebra of $\mathcal{S}$ since $[\mathcal{S}^i, \mathcal{S}^i] \subset \mathcal{S}^{i+j} \subset \mathcal{S}^i$.

It is convenient to define a sequence of subspaces of $\mathcal{S}$ by $\mathcal{R}^i = [\mathcal{R}, \mathcal{R}^{i+1}]$, $\mathcal{R}^1 = \mathcal{R}$. Obviously, $\mathcal{R}^i$ is the linear subspace of $\mathcal{S}$ spanned by brackets of length i. It is clear that

$$\mathcal{S}^i = \mathcal{R}^i + \mathcal{S}^{i+1}, \qquad \mathcal{S}^i + \mathcal{N} = \mathcal{R}^i + (\mathcal{S}^{i+1} + \mathcal{N}).$$

Thus, given a basis for $\mathcal{S}^{i+1} + \mathcal{N}$, this can be completed to a basis for $\mathcal{S}^i + \mathcal{N}$ with vector field lying in $\mathcal{R}^i$. In this way, a basis for $\mathcal{S}$ is constructed so that

$$\mathcal{S}^i + \mathcal{N} = \operatorname{Sp}\{a_{r_i} \cdots a_{r_{i+1}} \cdots a_{r_p} \cdots a_n a_{n+1} \cdots a_{n+s}\}, \tag{13}$$

where $a_{r_i} \cdots a_{r_{i+1}-1} \in \mathcal{R}^i$ and $\mathcal{N} = \operatorname{Sp}\{a_{n+1} \cdots a_{n+s}\}$.

As in § 3.6, the following map Φ provides a homeomorphism from $\mathbb{R}^n$ to the state space M of a minimal realization of a finite Volterra series, where $\gamma_i(t_i) = \gamma_{a_i}(t_i)$ and

$$\Phi(t_1 \cdots t_n) = \gamma_1(t_1) \circ \cdots \circ \gamma_n(t_n)x_0.$$

LEMMA 4.1. *$\Phi : \mathbb{R}^n \to M$ is a diffeomorphism.*

Proof. It is sufficient to show that Φ_* has full rank at each $t = (t_1 \cdots t_n) \in \mathbb{R}^n$. Note that the map $\phi(t) : M \to M$ defined by

$$\phi(t)x = \gamma_1(t_1) \circ \cdots \circ \gamma_n(t_n)x$$

is a diffeomorphism so that $\phi(t)_*^{-1} : T_{\Phi(t)}M \to T_{x_0}M$ is an isomorphism. The problem, therefore, reduces to showing that the vectors $\phi(t)_*^{-1}(\partial\Phi/\partial t_i)(t)$, $1 \leqq i \leqq n$, span $T_{x_0}M$. Now,

$$\phi(t)_*^{-1}\frac{\partial \Phi}{\partial t_i}(t) = \gamma_n(-t_n)_* \cdots \gamma_i(-t_i)_* a_i(\gamma_i(t_i) \circ \cdots \circ \gamma_n(t_n)x_0).$$

Noting that $[\mathscr{R}^i, \mathscr{R}^j] \subset \mathscr{R}^{i+j} \subset \mathscr{S}^{i+j} + \mathscr{N}$, and making use of the Campbell–Baker–Hausdorff formula, shows that this expression can be written in the form

$$\phi(t)_*^{-1}\frac{\partial \Phi}{\partial t_i}(t) = a_i(x_0) + \sum_{j=i+1}^{n} a_j(x_0)\alpha_j(t),$$

where $\alpha_j(\cdot)$ are polynomials in the components of t. By construction, $\operatorname{Sp}\{a_1(x_0) \cdots a_n(x_0)\} = T_{x_0}M$, so that it is easily concluded that the vectors $\phi(t)_*^{-1}(\partial\Phi/\partial t_i)(t)$, $1 \leqq i \leqq n$, do span $T_{x_0}M$. Q.E.D.

This result shows that in fact (Φ^{-1}, M) provides a global coordinate system for M, for each choice of basis for $\mathscr{S}$ constructed in the manner described, and hence strengthens the result provided in Theorem 3.7.

THEOREM 4.2. *The state space M of a strongly accessible observable realization of a finite Volterra series initialized at x_0 can be identified diffeomorphically with the vector space $T_{x_0}M$.*

Proof. The map $M \to T_{x_0}M$ given by the composition of the maps below is a diffeomorphism:

$$x \to \Phi^{-1}(x) = (t_1, t_2, \cdots, t_n),$$

$$(t_1, t_2, \cdots, t_n) \to (t_1 a_1(x_0), \cdots, t_n a_n(x_0)).$$ Q. E.D.

Before writing the system in terms of these coordinate charts, we make additional observations. By Theorem 3.9, there is a decomposition of $G = VN$, where V is the one-parameter subgroup with generator $f + a_0$, such that $(f + a_0)(x_0) = 0$. It is easily seen that $adf : \mathscr{S}^i \to \mathscr{S}^i$ is a linear endomorphism for each i and that also $ada : \mathscr{S}^i \to \mathscr{S}^{i+1} \subset \mathscr{S}^i$ for any $a \in \mathscr{S}$. Moreover, $ad(f + a_0) : \mathscr{N} \to \mathscr{N}$ since $\mathscr{N}$ consists of vector fields which vanish at x_0. It is concluded that

$$ad(f + a_0) : \mathscr{S}^{i+1} + \mathscr{N} \to \mathscr{S}^{i+1} + \mathscr{N}$$

is also a linear endomorphism for each i, and in particular induces representations of $ad(f + a_0)$ on

$$(\mathscr{S}^i + \mathscr{N})/(\mathscr{S}^{i+1} + \mathscr{N}) \cong \mathscr{R}^i + (\mathscr{S}^{i+1} + \mathscr{N})/(\mathscr{S}^{i+1} + \mathscr{N}) \cong \mathscr{R}^i/\mathscr{R}^i \cap (\mathscr{S}^{i+1} + \mathscr{N}).$$

THEOREM 4.3. *Any strongly accessible observable realization of a finite Volterra*

series of length p has an isomorphic realization of the form

$$
\left.\begin{aligned}
&\dot z_1 = A_1 z_1 + d_1\\
&\dot z_2 = A_2 z_2 + d_2(z_1)\\
&\quad\vdots\\
&\dot z_p = A_p z_p + d_p(z_1 \cdots z_{p-1})
\end{aligned}\right\}
+ \sum_{i+1}^{m} u_i
\begin{cases}
b_{i1}, & z_1(0)=0,\\
b_{i2}(z_1), & z_2(0)=0,\\
\quad\vdots\\
b_{ip}(z_1 \cdots z_{p-1}),
\end{cases}
\tag{14}
$$

$$
y = c(z_1 \cdots z_p)
$$

where b_{ji}, d_i and c are vector-valued polynomials in the components of the vectors $z_i \in \mathbb{R}^{n_i}$, $\sum_{i=1}^{p} n_i = n$ being the dimension of the state space, and $n_i = \dim \mathscr{R}^i/\mathscr{R}^i \cap (\mathscr{S}^{i+1} + \mathscr{N})$, depends only on the input–output map.

Proof. Using the basis for $\mathscr{S}$ previously described with the properties defined by (13), let $s_k = r_{k+1} - 1$, and define inductively

$$
v_k(t) = \gamma_{s_k}(-x_{s_k}(t)) \circ \cdots \circ \gamma_{r_k}(-x_{r_k}(t)) v_{k-1}(t), \tag{15}
$$

where $v_0(t) = x(t)$ is the solution of the system equations $\dot x = f(x) + \sum_{i=1}^{m} u_i g_i(x)$, $x(0) = x_0$. Let z_k be the vector with components $(x_{r_k} \cdots x_{s_k})$. It is proved by induction that if the vectors $z_1 \cdots z_{k-1}$ satisfy equations with the form of the first $k-1$ differential equations in (14), then v_{k-1} satisfies a differential equation of the form

$$
\begin{aligned}
&\dot v_{k-1} = (f + a_0)(v_{k-1}) + \sum_{i=r_k}^{n+s} \left[\alpha_i(z_1 \cdots z_{k-1}) + \sum_{j=1}^{m} u_j \beta_{ji}(z_1 \cdots z_{k-1})\right] a_i(v_{k-1}),\\
&\qquad v_{k-1}(0) = x_0,
\end{aligned}
\tag{16}
$$

where α_i, β_{ji} are polynomials in $z_1 \cdots z_{k-1}$.

Now $-a_0$, $g_j \in \mathscr{S}$ have the following expansions relative to the given basis for $\mathscr{S}$:

$$
-a_0 = \sum_{i=1}^{n+s} \alpha_i a_i, \qquad g_j = \sum_{i=1}^{n+s} \beta_{ji} a_i.
$$

The system equations can now be written as

$$
\dot v_0 = (f + a_0)(v_0) + \sum_{i=1}^{n+s} \left[\alpha_i + \sum_{j=1}^{m} u_j \beta_{ji}\right] a_i(v_0), \qquad v_0(0) = x_0.
$$

Setting $k = 1$ in (16) shows that the induction statement is true for $k = 1$. If it is assumed true for $k-1$, a differential equation for v_k can be developed by differentiating (15) and substituting for $\dot v_{k-1}$ from (16). By the Campbell–Baker–Hausdorff formula

$$
\begin{aligned}
\dot v_k = {} & \gamma_{s_k}(-x_{s_k})_* \cdots \gamma_{r_k}(-x_{r_k})_* \dot v_{k-1} - \dot x_{s_k} a_{s_k}(v_k)\\
& \cdots - \dot x_{r_k} \exp - x_{s_k}\, ad\, a_{s_k} (\cdots \exp - x_{r_k}\, ad\, a_{r_k}(a_{r_k}) \cdots)(v_k).
\end{aligned}
$$

Noting the identity $[\mathscr{S}^k, \mathscr{S}^j] \subset \mathscr{S}^{k+j}$ shows that this can be rewritten in the form

$$
\dot v_k = \gamma_{s_k}(-x_{s_k})_* \cdots \gamma_{r_k}(-x_{r_k})_* \dot v_{k-1} - \sum_{i=r_k}^{s_k} \dot x_i a_i(v_k) + \sum_{i=r_{k+1}}^{n+s} \eta_i(z_k, \dot z_k) a_i(v_k), \tag{17}
$$

where η_i is linear in $\dot z_k$ and polynomial in z_k

Consider the first term ζ in this expression, and substitute from (16) for $\dot v_{k-1}$ and set

$$
v_{k-1} = \gamma_{r_k}(x_{r_k}) \circ \cdots \circ \gamma_{s_k}(x_{s_k}) v_k
$$

in the resulting expression. Thus ζ can be split into two terms, the first being

$$\gamma_{s_k}(-x_{s_k})_* \cdots \gamma_{r_k}(-x_{r_k})_*(f+a_0)(\gamma_{r_k}(x_{r_k}) \circ \cdots \circ \gamma_{s_k}(x_{s_k})v_k).$$

The Campbell–Baker–Hausdorff formula shows that this term can be rewritten in the form

$$(f+a_0)(v_k) - \sum_{i=r_k}^{s_k} x_i ad(f+a_0)(a_i)(v_k) + \sum_{i=r_{k+1}}^{n+s} \eta_i'(z_k)a_i(v_k),$$

where it is noted that $ad(f+a_0): \mathscr{S}^k \to \mathscr{S}^k$ and η_i' are polynomials in z_k. Now the second term in this expression can be rewritten using the expansion

$$-ad(f+a_0)(a_i) = \sum_{j=r_k}^{s_k} A_{ji}a_j + \sum_{j=r_{k+1}}^{n+s} A_{ji}a_j,$$

where the matrix $[A_{ji}]$, $r_k \leqq j, i \leqq s_k$ is the matrix representation of $-ad(f+a_0)$ on $\mathscr{R}^k/\mathscr{R}^k \cap (\mathscr{S}^{k+1} + \mathscr{N})$ relative to the chosen basis. Finally it is seen that the first term in ζ has the form

$$(f+a_0)(v_k) + \sum_{i,\,j=r_k}^{s_k} A_{ji}x_i a_j(v_k) + \sum_{i=r_{k+1}}^{n+s} \eta_i''(z_k)a_i(v_k), \tag{18}$$

for polynomials η_i'' in z_k.

The second term in ζ involves terms of the form

$$\gamma_{s_k}(-x_{s_k})_* \cdots \gamma_{r_k}(-x_{r_k})_* a_k(\gamma_{r_k}(x_{r_k}) \circ \cdots \circ \gamma_{s_k}(x_{s_k})v_k),$$

where $a_k \in \mathscr{S}^k + \mathscr{N}$. Again using the Campbell–Baker–Hausdorff formula, it is easily deduced that this term can be rewritten in the form

$$a_k(v_k) + \sum_{i=r_{k+1}}^{n+s} \xi_i(z_k)a_i(v_k), \tag{19}$$

for polynomials ξ_i in z_k.

It is now clear from (18) and (19) that (17) can be rewritten in the form

$$\begin{aligned} \dot{v}_k = (f+a_0)(v_k) &+ \sum_{i=r_{k+1}}^{n+s} [\eta_i(z_k, \dot{z}_k) + \alpha_i'(z_1 \cdots z_k) + \sum_{j=1}^{m} u_j \beta_{ji}'(z_1 \cdots z_k)]a_i(v_k) \\ &+ \sum_{i=r_k}^{s_k} [-\dot{x}_i + \sum_{j=r_k}^{s_k} A_{ij}x_j + \alpha_i(z_1 \cdots z_{k-1}) + \sum_{j=1}^{m} u_j \beta_{ji}(z_1 \cdots z_{k-1})]a_i(v_k), \end{aligned} \tag{20}$$

where α_i' and β_{ji}' are polynomials in $z_1 \cdots z_k$. By equating coefficients of terms, on the right-hand side of this equation, involving the vector fields a_i, $r_k \leqq i \leqq s_k$, to zero, an equation with the form of the kth equation in (14) is obtained. If the expression for $\dot{z}_k$ thus obtained is substituted back into (20), the resultant equation has the form

$$\dot{v}_k = (f+a_0)(v_k) + \sum_{i=r_{k+1}}^{n+s} [\alpha_i''(z_1 \cdots z_k) + \sum_{j=1}^{m} u_j \beta_{ji}''(z_1 \cdots z_k)]a_i(v_k).$$

Setting $z_k(0) = 0$ shows that this equation has initial conditions $v_k(0) = x_0$, which completes the induction, once it is noted that α_i'' and β_i'' are polynomial in $z_1 \cdots z_k$.

To complete the proof note that by construction the differential equation for v_{p+1} obtained in the above process involves only vector fields vanishing at x_0, so that $v_{p+1}(t) \equiv x_0$. Thus, from (15),

$$x(t) = \gamma_1(x_1(t)) \circ \cdots \circ \gamma_n(x_n(t))x_0,$$

and $x(t)$ satisfies the system equations if and only if $z_1(t) \cdots z_p(t)$ satisfy a set of equations of the form given in (14). Note also that the output can now be written as $y = h(x) = c(z_1 \cdots z_n)$. However, as in Proposition 3.6, it is easily seen that c is a polynomial in the components of the vectors z_i, $1 \leqq i \leqq n$. Q.E.D.

This result applies trivially to the linear input–output map, showing that there exists a coordinate system in which any nonlinear realization of the linear input–output map becomes linear (cf. Krener [12]).

Write system (14) as

$$\begin{aligned} &\dot{z} = F(z) + \sum_{i=1}^{m} u_i G_i(z), \; z(0) = 0, \qquad z \in \mathbb{R}^n, \\ &y = H(z). \end{aligned} \tag{21}$$

Obviously, the map $\Phi : \mathbb{R}^n \to M$ of Lemma 4.1 is an isomorphism of systems satisfying the relations

$$\Phi_* F = f \circ \Phi, \quad \Phi_* G_i = g_i \circ \Phi, \quad H = h \circ \Phi(0), \quad \Phi(0) = x_0.$$

In order to characterize the general form of a minimal realization of a finite Volterra series of length p, the structure of the polynomials in (14) must be determined.

Write the vector space $\mathbb{R}^n$ as a direct product

$$\mathbb{R}^n = \bigoplus_{i=1}^{p} \mathbb{R}^{n_i},$$

where any $z \in \mathbb{R}^n$ has components $z_i \in \mathbb{R}^{n_i}$ as in Theorem 4.3. For any $\lambda > 0$, let $\delta_\lambda : \mathbb{R}^n \to \mathbb{R}^n$ be the diffeomorphism of $\mathbb{R}^n$ given by

$$\delta_\lambda(z) = (\lambda z_1, \cdots, \lambda^p z_p).$$

LEMMA 4.4. *The realization of Theorem* 4.3 *can be written as*

$$\dot{z} = \sum_{i=0}^{p} F_i(z) + \sum_{i=1}^{p} \sum_{j=1}^{m} u_j G_{ji}(z), \qquad y = \sum_{i=0}^{p} H_i(z), \qquad z(0) = 0,$$

where $\lambda^i F_i \circ \delta_\lambda = \delta_\lambda \circ F_i$, $\lambda^i G_{ji} \circ \delta_\lambda = \delta_\lambda \circ G_{ji}$, $H_i \circ \delta_\lambda = \lambda^i H_i$.

Proof. The proof consists of repeating Theorem 4.3 for a one-parameter family of bases for $\mathscr{S}$. In particular, if the basis used in Theorem 4.3 is given by

$$a_{r_1} \cdots a_{s_1} a_{r_2} \cdots a_{r_p} \cdots a_{s_p} a_{n+1} \cdots a_{n+s},$$

consider the one-parameter family of bases for $\lambda > 0$ given by

$$\lambda a_{r_1} \cdots \lambda a_{s_1} \lambda^2 a_{r_2} \cdots \lambda^p a_{r_p} \cdots \lambda^p a_{s_p} a_{n+1} \cdots a_{n+s}. \tag{22}$$

This in turn induces a one-parameter family of diffeomorphisms of M which is easily seen to correspond to δ_λ in the coordinate system given by Φ. Thus, by computing the canonical realization from the bases (22), the following systems are obtained.

$$\begin{aligned} &\dot{z}_\lambda = F_\lambda(z_\lambda) + \sum_{i=1}^{n} u_i G_{i\lambda}(z_\lambda), \qquad z_\lambda(0) = 0, \\ &y = H_\lambda(z_\lambda), \\ &F_\lambda = \delta_\lambda^{-1} \circ F \circ \delta_\lambda, \quad G_{i\lambda} = \delta_\lambda^{-1} \circ G_i \circ \delta_\lambda, \quad H_\lambda = H \circ \delta_\lambda. \end{aligned}$$

From the construction of the canonical realization, it is now readily observed that there

exist vectors F_i, G_{ij} such that

$$F_\lambda = \sum_{i=0}^{p} \frac{F_i}{\lambda^i}, \qquad G_{j\lambda} = \sum_{i=1}^{p} \frac{G_{ji}}{\lambda^i},$$

and which satisfy the statement of the lemma.

It remains to prove the decomposition for H. However,

$$\frac{\partial^{k_1+\cdots+k_n}}{\partial x_1^{k_1}\cdots\partial x_n^{k_n}} H(x_1\cdots x_n)\Big|_{x_1=\cdots=x_n=0} = 0,$$

if $k_1+\cdots+k_{s_1}+2k_{r_2}+\cdots+2k_{s_2}+\cdots+pk_{s_p} > p$ by (6) and the properties of the vector fields chosen for the basis of $\mathscr{S}$. Thus, H can be expanded in the form $H = \sum_{i=0}^{p} H_i$, where

$$H_i \circ \delta_\lambda = \lambda^i \circ H. \qquad \text{Q.E.D.}$$

This lemma can now be used to obtain the composition of the polynomials in (14) as follows:

$$d_i(z_1\cdots z_{i-1}) = \sum_{j=0}^{i-1} d_{ij}(z_1\cdots z_j) + d_{ii}(z_1\cdots z_{i-1}),$$

where

$$d_{ij}(\lambda z_1\cdots\lambda^j z_j) = \lambda^j d_{ij}(z_1\cdots z_j),$$
$$d_{ii}(\lambda z_1\cdots\lambda^{i-1} z_{i-1}) = \lambda^i d_{ii}(z_1\cdots z_{i-1}),$$

and

$$b_{ki}(z_1\cdots z_{i-1}) = \sum_{j=0}^{i-1} b_{kij}(z_1\cdots z_j),$$

where

$$b_{kij}(\lambda z_1\cdots\lambda^j z_j) = \lambda^j b_{kij}(z_1\cdots z_j).$$

Conversely, from these relationships it is easily checked that any system of equations of the type given in (14), satisfying the properties of Lemma 4.4, has a finite Volterra series of length p and a finite dimensional Lie algebra as specified in Theorem 3.2. These (not necessarily minimal) systems are said to be in canonical form, and provide the natural class of systems admitting finite Volterra series, replacing the linear system. Theorem 4.3 and Lemma 4.4 show that minimal realizations can be obtained from these while remaining in the same class of system.

4.2. Stability. As a simple application of the canonical realization developed above, the asymptotic stability of a canonical realization of a stationary finite Volterra series is investigated.

The stationary system in canonical form,

$$\dot{z} = F(z) + \sum_{i=1}^{m} u_i G_i(z), \qquad z(0) = 0, \quad F(0) = 0, \quad z \in \mathbb{R}^n,$$

is asymptotically stable at 0 if and only if for all $z \in \mathbb{R}^n$, $\lim_{t\to\infty} \gamma_F(t)z = 0$. Note that for stationary systems, the vector field a_0 in Theorem 4.3 can be set to zero and so the matrices A_i are representations of $-adF$ on $\mathscr{R}^i/\mathscr{R}^i \cap (\mathscr{S}^{i+1} + \mathscr{N})$, where of course $-adF : \mathscr{N} \to \mathscr{N}$. Moreover, the action of $-adF$ on $\mathscr{S}$ induces a matrix representation A of

$-adF$ on $\mathcal{S}/\mathcal{N}$ with respect to the same basis as used in Theorem 4.3. Now A is the lower block diagonal with diagonal blocks A_i, so that the characteristic polynomial of A is the product of the characteristic polynomials of the matrices A_i.

THEOREM 4.5. *A stationary realization of a finite Volterra series, in canonical form, is asymptotically stable if and only if the matrix representation of $-adF$ on $\mathcal{S}/\mathcal{N}$ has all its eigenvalues in the left half of the complex plane, or equivalently, the matrices A_i have all their eigenvalues in the left half of the complex plane.*

Proof. By the introductory remarks, it is sufficient to show that a canonical realization is asymptotically stable at $0 \in \mathbb{R}^n$ if and only if the matrices A_i have all their eigenvalues in the left half of the complex plane. However, this is a straightforward exercise. See Crouch [8] for details. Q.E.D.

4.3. Structural invariants. From Theorem 4.3, it is seen that the state space dimension and subsystem dimensions in a canonical realization of a finite Volterra series depend only on the input–output map. In this section, it is shown how these structural invariants can be determined directly from the Volterra kernels.

Some important subspaces of $\mathcal{H}$ are introduced and new terminology introduced. Let $\mathcal{H}^l$ be the smallest linear subspace of $C(M)$ containing the functions

$$ad^{k_i}f(g_{j_i})(\cdots(ad^{k_1}f(g_{j_1})(\overbrace{f\cdots f}^{m}(h_j(\cdot))\cdots))\cdots), \tag{23}$$

for $i \geqq l \geqq 0$ and arbitrary integers k_i and m. Notice that if $\tau \in \mathcal{H}^l$ and $a \in \mathcal{L}$, then $a(\tau) \in \mathcal{H}^l$. Moreover, $\mathcal{H}^0 = \mathcal{H}$, as is easily verified. Let $\mathcal{O}^l$ be the smallest linear subspace of $\mathcal{H}$ spanned by the functions in (23) for $i = l$. It follows that

$$\mathcal{H}^l = \mathcal{O}^l + \mathcal{H}^{l+1}.$$

The following result establishes some simple facts and relations between the subspaces $\mathcal{S}^i(x)$ and $\mathcal{H}^i(x)$.

PROPOSITION 4.6. *In a strongly accessible locally observable realization of a finite Volterra series of length p, $\mathcal{S}^i(x)$ and $d\mathcal{H}^i(x)$ are constant dimensional subspaces of $\mathcal{S}(x) = T_xM$ and $d\mathcal{H}^0(x) = T_x^*M$, respectively. Moreover, $d\mathcal{O}^{p-i}(x)\mathcal{R}^{i+j}(x) = 0$ and $d\mathcal{H}^{p-i}(x)\mathcal{S}^{i+1}(x) = 0$ on M, for $0 \leqq i \leqq p-1$ and $j > 0$.*

Proof. By orbit minimality (analyticity and strong accessibility), it is sufficient to prove that $\dim \mathcal{S}^i(x) = \dim \mathcal{S}^i(y)$, where $x = \gamma_a(t)y$ for any $a \in \mathcal{L}$ and $t \in \mathbb{R}$. Noting that $\mathcal{S}^i$ is an ideal in $\mathcal{L}$ makes it clear that the proof of Sussmann [22, Lemma 3.5] applies in this case also. Similarly, the proof that $\dim d\mathcal{H}^i(x) = \dim d\mathcal{H}^i(y)$ follows that of Hermann and Krener [12], since $L_a(d\tau) = dL_a(\tau) \subset d\mathcal{H}^i$ for arbitrary $a \in \mathcal{L}$ and $\tau \in \mathcal{H}^i$. The space of functions $d\mathcal{O}^{p-i}\mathcal{R}^{i+j}$, $j > 0$, $0 \leqq i \leqq p-1$, is spanned by functions in $\mathcal{O}^{p+j}$. Formula (6) shows that these are zero functions on M for $j > 0$. Since $d\mathcal{H}^{p-1} = d\mathcal{O}^{p-i} + d\mathcal{O}^{p-i+1} + \cdots + d\mathcal{O}^p$ and $\mathcal{S}^{i+1} = \mathcal{R}^{i+1} + \mathcal{R}^{i+2} + \cdots + \mathcal{R}^p$, it follows that $d\mathcal{H}^{p-i}(x)\mathcal{S}^{i+1}(x) = 0$ for $x \in M$, $0 \leqq i \leqq p-1$. Q.E.D.

Now define a bracket operation on the Volterra kernels by setting

$$W_k^{i_0 i_1 \cdots i_j i_{j+1} \cdots i_k}(t, \sigma_1 \cdots [\sigma_j, \sigma_{j+1}] \cdots \sigma_k)$$
$$= W_k^{i_0 i_1 \cdots i_j i_{j+1} \cdots i_k}(t_1 \sigma_1 \cdots \sigma_j \sigma_{j+1} \cdots \sigma_k) - W_k^{i_0 i_1 \cdots i_{j+1} i_j \cdots i_k}(t, \sigma_1 \cdots \sigma_{j+1} \sigma_j \cdots \sigma_k).$$

Similarly, define by induction

$$W_k^{i_0 i_1 \cdots i_j \cdots i_{j+l} \cdots i_k}(t, \sigma_1 \cdots [\sigma_j[\sigma_{j+1}[\cdots \sigma_{j+l}] \cdots]\sigma_{j+l+1} \cdots \sigma_k)$$
$$= W_k^{i_0 i_1 \cdots i_j i_{j+1} \cdots i_{j+l} \cdots i_k}(t, \sigma_1 \cdots \sigma_j[\sigma_{j+1}[\cdots \sigma_{j+l}] \cdots]\sigma_{j+l+1} \cdots \sigma_k)$$
$$- W_k^{i_0 i_1 \cdots i_{j+1} \cdots i_{j+l} i_j \cdots i_k}(t, \sigma_1 \cdots [\sigma_{j+1}[\sigma_{j+2} \cdots \sigma_{j+l}] \cdots]\sigma_j \sigma_{j+l+1} \cdots \sigma_k).$$

Moreover, any nested expression of skew symmetric brackets (obeying the Jacobi identity), can be expressed as sums of brackets of the above form; but this generalization will not be utilized here.

Considering a strongly accessible observable realization of a finite Volterra series makes it clear from the above definition and the kernel structure equations (1) that

$$\begin{aligned} &W_{k+j}^{s_0 s_1 \cdots s_j r_1 \cdots r_k}(t_0 t_1 \cdots t_j[\sigma_1[\sigma_2 \cdots \sigma_k] \cdots], x) \\ &\quad = d\tau^{s_0 s_1 \cdots s_j}(t_0 t_1 \cdots t_j) a_1^{r_1 \cdots r_k}(\sigma_1 \cdots \sigma_k)(x), \end{aligned} \tag{24}$$

for $k+j \leqq p$ and is identically zero for $k+j > p$, where $\tau \in \mathcal{O}^j$ and $a \in \mathcal{R}^k$, and the images of the analytic maps

$$t_0 t_1 \cdots t_j \to d\tau^{s_0 \cdots s_j}(t_0 \cdots t_j)(x),$$

$$\sigma_1 \cdots \sigma_k \to a^{r_1 \cdots r_k}(\sigma_1 \cdots \sigma_k)(x)$$

contain spanning sets for $d\mathcal{O}^j(x)$ and $\mathcal{R}^k(x)$, respectively. Further, considering a basis for $T_{x_0}M^n$ and a dual basis for $T_{x_0}M^{n^*}$ shows that there exist analytic vectors

$$h^{s_0 \cdots s_j}(t_0 \cdots t_j) \in \mathbb{R}^{n^*} \quad \text{and} \quad v^{r_1 \cdots r_k}(\sigma_1 \cdots \sigma_k) \in \mathbb{R}^n$$

such that

$$h^{s_0 \cdots s_j}(t_0 \cdots t_j) v^{r_1 \cdots r_k}(\sigma_1 \cdots \sigma_k) = \begin{cases} W_{k+j}^{s_0 \cdots s_j r_1 \cdots r_k}(t_0 \cdots t_j[\sigma_1[\cdots \sigma_k] \cdots], x_0), & k+j \leqq p, \\ 0, & k+j \geqq p. \end{cases}$$

By rearranging the set of indexes $s_0 \cdots s_j, r_1 \cdots r_k$ designating the components of each kernel in some fixed order, the kernels can be "factored" into products of matrices,

$$\bar{H}_j(t_0 \cdots t_j)\bar{G}_k(\sigma_1 \cdots \sigma_k) = \begin{cases} W_{k+j}(t_0 \cdots t_j[\sigma_1[\cdots \sigma_k] \cdots]), & j+k \leqq p, \\ 0, & j+k > p. \end{cases}$$

and in particular, the matrix of kernels $V_i(t, \sigma)$ given by

$$\begin{bmatrix} W_p(t_0[\sigma_1[\cdots \sigma_p] \cdots]) & W_{p-1}(t_0[\sigma_1[\cdots \sigma_{p-1}] \cdots]) & \cdots & W_i(t_0[\sigma_1[\cdots \sigma_i] \cdots]) \\ 0 & W_p(t_0 t_1[\sigma_1[\cdots \sigma_{p-1}] \cdots]) & & \\ \vdots & \vdots & & \vdots \\ 0 & 0 & & W_p(t_0 \cdots t_{p-i}[\sigma_i[\cdots \sigma_i] \cdots]) \end{bmatrix}$$

can be factored through $\mathbb{R}^n$ into the product

$$H_i(t)G_i(\sigma) = \begin{bmatrix} \bar{H}_0(t_0) \\ \bar{H}_1(t_0 t_1) \\ \bar{H}_{p-i}(t_0 \cdots t_{p-i}) \end{bmatrix} [\bar{G}_p(\sigma_1 \cdots \sigma_p)\bar{G}_{p-1}(\sigma_1 \cdots \sigma_{p-1}) \cdots \bar{G}_i(\sigma_1 \cdots \sigma_i)].$$

With these preliminary observations, the main result can be stated.

THEOREM 4.7. *Consider a strongly accessible observable realization of a finite Volterra series of length p. Let n_i be the minimal dimension through which the matrix $V_i(t, \sigma)$ can be factored analytically into the form $V_i(t, \sigma) = H_i(t)G_i(\sigma)$. Then if $N_i = n_i - n_{i+1}$, $1 \leqq i \leqq p-1$, $N_p = n_p$, N_i is the dimension of the ith subsystem in any minimal canonical realization of the Volterra series. In particular, $n_1 = n$ is the dimension of its state space.*

Proof. By the preceding observations, $n_i \leqq n$. Let $n_i' = \dim \mathcal{S}^i(x_0)$. By (24), it is evident that the matrix $V_i(t, \sigma)$ is constructed from vectors $a(x_0) \in \mathcal{R}^k(x_0) \subset \mathcal{S}^i(x_0)$, $p \geqq$

$k \geqq i$ and vectors $d\tau(x_0) \in \mathcal{O}^j(x_0)$, $k+j \leqq p$. It is therefore clear that the decomposition can take place in $\mathbb{R}^{n_i'}$. It is now shown that n_i' is the minimal dimension through which the factorization can take place. By Proposition 4.6, $\mathcal{S}^i(x_0)^* \subset d\mathcal{O}^0(x_0) + \cdots + d\mathcal{O}^{p-i}(x_0)$, and the image of the analytic maps $t_0 \cdots t_j \to d\tau^{s_0 \cdots s_j}(t_0 \cdots t_j)(x_0)$, $0 \leqq j \leqq p-i$, contains a spanning set for $d\mathcal{O}^0(x_0) + \cdots + d\mathcal{O}^{p-i}(x_0)$. Moreover, the image of the analytic maps $\sigma_1 \cdots \sigma_k \to a^{r_1 \cdots r_k}(\sigma_1 \cdots \sigma_k)(x_0)$, $p \geqq k \geqq i$, contains a spanning set for $\mathcal{S}^i(x_0)$. The minimality of n_i' is now clear, and it follows that

$$n_i = \dim \mathcal{S}^i(x_0) = \dim (\mathcal{S}^i / \mathcal{S}^i \cap \mathcal{N}).$$

However, in any canonical realization, the dimension of the ith subsystem

$$\begin{aligned} N_i &= \dim (\mathcal{S}^i + \mathcal{N})/(\mathcal{S}^{i+1} + \mathcal{N}) \\ &= \dim [(\mathcal{S}^i + \mathcal{N})/\mathcal{N}]/[(\mathcal{S}^{i+1} + \mathcal{N})/\mathcal{N}] = \dim (\mathcal{S}^i + \mathcal{N})/\mathcal{N} - \dim (\mathcal{S}^{i+1} + \mathcal{N})/\mathcal{N} \\ &= \dim \mathcal{S}^i/(\mathcal{S}^i \cap \mathcal{N}) - \dim \mathcal{S}^{i+1}/(\mathcal{S}^{i+1} \cap \mathcal{N}) = n_i - n_{i+1}. \end{aligned}$$

Clearly, $n_1 = n$ since $\mathcal{S}(x_0) = T_{x_0}M$. Q.E.D.

COROLLARY 4.8. *If $W_{k+j}(t_0 \cdots t_j[\sigma_1[\cdots \sigma_k]\cdots]) \equiv 0$ for k and j satisfying $p \geqq k \geqq i+1$ and $0 \leqq k+j \leqq p$, then only the first i subsystems in any canonical realization of the Volterra series are nontrivial.*

Proof. The dimensions n_k, $i+1 \leqq k \leqq p$ are all zero in Theorem 4.7. Q.E.D.

Example 4.7. Consider the following system written in canonical form:

$$\begin{aligned} \dot{x}_1 &= u, & x_1(0) &= 0, \\ \dot{x}_2 &= x_1^2 + x_1 + u, & x_2(0) &= 0, \\ y &= x_1^2 + x_2. \end{aligned}$$

Thus, $f = (x_1^2 + x_1)\,\partial/\partial x_2$, $g = \partial/\partial x_1 + \partial/\partial x_2$, $[f, g] = (2x_1 + 1)\partial/\partial x_2$ and $[g, [f, g]] = 2\partial/\partial x_2$, and all other brackets are zero. The system is therefore strongly accessible. Moreover, $dh = 2x_1\,dx_1 + dx_2$, $d[f, g](h) = 2dx_1$, so the system is weakly observable and hence observable by Corollary 3.8. The Volterra series for this system is given by

$$y(t) = 2\int_0^t \int_0^{\sigma_1} (t - \sigma_1 + 1)u(\sigma_1)u(\sigma_2)\,d\sigma_1\,d\sigma_2 + \int_0^t (t - \sigma_1 + 1)u(\sigma_1)\,d\sigma_1.$$

Thus,

$$W_2(t, \sigma_1, \sigma_2) = 2(t - \sigma_1 + 1), \qquad W_1(t, \sigma_1) = (t - \sigma_1 + 1)$$

and

$$\left[\begin{array}{c|c} W_2(t_0, [\sigma_1, \sigma_2]) & W_1(t_0, \sigma_1) \\ \hline 0 & W_2(t_0, t_1, \sigma_1) \end{array}\right] = \left[\begin{array}{c|c} 2(\sigma_2 - \sigma_1) & (t_0 - \sigma_1 + 1) \\ \hline 0 & 2(t_0 - t_1 + 1) \end{array}\right].$$

Hence,

$$V_2(t, \sigma) = 1.2(\sigma_2 - \sigma_1)$$

and

$$V_1(t, \sigma) = \left[\begin{array}{c|c} 1 & t_0 \\ \hline 0 & 2(t_0 - t_1 + 1) \end{array}\right] \left[\begin{array}{c|c} 2(\sigma_2 - \sigma_1) & 1 - \sigma_1 \\ \hline 0 & 1 \end{array}\right]$$

are minimal factorizations from which it is deduced that $n_2 = 1$, $n_1 = 2$, so that $n_2 = N_2 = 1$, $n_1 - n_2 = N_1 = 1$ and $n = N_2 + N_1 = 2$. This obviously agrees with the minimal system above.

Finally, it should be noted that these methods are similar to those employed in the bilinear case as studied in [1].

4.4. Homogeneous Volterra series. In this section, the special structure and properties of a canonical realization of a Volterra series described by a single Volterra kernel W_p (homogeneous Volterra series of degree p) is examined. Thus, realizations of the following input–output map are examined:

$$y(t, u) = \int_0^t \int_0^{\sigma_1} \cdots \int_0^{\sigma_{p-1}} W_p(t, \sigma_1 \cdots \sigma_p)(u(\sigma_1) \cdots u(\sigma_p))\, d\sigma_1 \cdots d\sigma_p.$$

By Lemma 3.1(b) the last kernel W_p is independent of state in a strongly accessible realization, and so by the identity (2), it is seen that the kernel W_p is stationary. It follows that there is no loss of generality in studying only stationary kernels W_p and hence, stationary minima l realizations.

Let $\mathcal{K}$ denote the subspace of $d\mathcal{H}$ consisting of one of the forms which vanishes at x_0 (cf. $\mathcal{N}$ is the subspace of $\mathcal{S}$ consisting of vector fields which vanish at x_0).

PROPOSITION 4.9. *In a strongly accessible observable realization of a homogeneous Volterra series of degree p, the pairing between the spaces* $d\mathcal{O}^{p-k}(x_0), \mathcal{R}^k(x_0)$ *is nondegenerate,* $1 \leqq k \leqq p$ *and*

$$\mathcal{R}^k \cap (\mathcal{S}^{k+1} + \mathcal{N}) = \mathcal{R}^k \cap \mathcal{N}, \qquad d\mathcal{O}^k \cap (d\mathcal{H}^{k+1} + \mathcal{K}) = d\mathcal{O}^k \cap \mathcal{K}.$$

In particular, $T_{x_0}M^* = \oplus_{i=0}^{p-1} d\mathcal{O}^i(x_0)$, $T_{x_0}M = \oplus_{i=1}^{p} \mathcal{R}^i(x_0)$, *are internal direct sums.*

Proof. As in the proof of Theorem 4.7, the functions comprising $W_p(t_0, t_1 \cdots t_{p-k}[\sigma_1[\cdots \sigma_k]\cdots])$ have the form

$$d\tau(t_0 t_1 \cdots t_{p-k}) a(\sigma_1 \cdots \sigma_k)(x_0),$$

where $d\tau \in d\mathcal{O}^{p-k}$, $a \in \mathcal{R}^k$, and since there is only one kernel, it follows that $d\mathcal{O}^{p-k}(x_0)\mathcal{R}^{k+j}(x_0) = 0$, for $j \neq 0$. Since the realization is strongly accessible and weakly observable, it follows that the pairing between the spaces $d\mathcal{O}^{p-k}(x_0)$ and $\mathcal{R}^k(x_0)$ must be nondegenerate.

Now if $a \in \mathcal{R}^k \cap (\mathcal{S}^{k+1} + \mathcal{N})$, then $d\mathcal{O}^{p-k}(x_0)a(x_0) = 0$, so by the nondegeneracy of the pairing between the spaces $d\mathcal{O}^{p-k}(x_0)$ and $\mathcal{R}^k(x_0)$, it follows that $a(x_0) = 0$. Thus, $\mathcal{R}^k \cap (\mathcal{S}^{k+1} + \mathcal{N}) \subset \mathcal{R}^k \cap \mathcal{N}$. Similarly, $d\mathcal{O}^k \cap (d\mathcal{H}^{k-1} + \mathcal{K}) \subset d\mathcal{O}^k \cap \mathcal{K}$. Now, $\mathcal{R}^k \cap (\mathcal{S}^{k+1} + \mathcal{N}) \subset \mathcal{R}^k \cap \mathcal{N} \Rightarrow \mathcal{R}^k \cap \mathcal{S}^{k+1} \subset \mathcal{N}$. Thus, $\mathcal{R}^k \cap \mathcal{R}^j \subset \mathcal{N}$ for $k \neq j$ and $\mathcal{R}^k(x_0) \cap \mathcal{R}^j(x_0) = \{0\}$ for $k \neq j$. The decomposition of $T_{x_0}M^*$ follows similarly. Q.E.D.

Consider a strongly accessible observable realization of a homogeneous Volterra series of degree p. Then the map

$$a^i(x_0) \to \lambda^i a^i(x_0)$$

for $a^i \in \mathcal{R}^i$ induces a linear isomorphism l of $T_{x_0}M$ for $\lambda \neq 0$, by the preceding proposition. Consider the following system:

$$\begin{aligned} \dot{x}_\lambda &= f(x_\lambda) + \sum_{i=1}^{m} u_i \lambda g_i(x_\lambda),\ x_\lambda \in M,\ x_\lambda(0) = x_0, \\ y &= h(x_\lambda). \end{aligned} \tag{25}$$

For $\lambda = 1$ the original dynamics are obtained. Clearly, $l(f(x_0)) = l(0) = 0$ since the system is stationary, and $l(g_i(x_0)) = \lambda g_i(x_0)$. Moreover, $l([a_1, a_2](x_0)) = [\tilde{a}_1, \tilde{a}_2](x_0)$, where if $a_1, a_2 \in \mathcal{L}$, $\tilde{a}_1, \tilde{a}_2$ are the corresponding vector fields in the Lie algebra

generated by f and $\lambda g_1 \cdots \lambda g_m$. These are the precise conditions on l under which Krener [12, Theorem 1] holds. Since the state space M is simply connected, it is concluded that there exists a diffeomorphism ϕ_λ of M such that $\phi_\lambda(x_1(t)) = x_\lambda(t)$, where $x_\lambda(t)$ is a solution curve of the above equation. In the next theorem, it is shown that a basis of $\mathscr{L}$ can be chosen so that the diffeomorphism ϕ_λ is in fact given by δ_λ in the coordinates of the canonical realization determined by the chosen basis.

THEOREM 4.10. *Given a realizable stationary homogeneous Volterra series of degree p, there exists a minimal canonical realization obtained by setting $\lambda = 1$ in the equations*

$$\dot{z}_\lambda = F(z_\lambda) + \sum_{i=1}^{m} u_i \lambda G_i(z_\lambda), \qquad z_\lambda(0) = 0, \quad F(0) = 0, \quad z_\lambda \in \mathbb{R}^n,$$

$$y = H(z_\lambda),$$

which satisfy

$$z_\lambda = \delta_\lambda(z_1), \quad \delta_\lambda \circ F = F \circ \delta_\lambda, \quad \delta_\lambda \circ G_i = \lambda G_i \circ \delta_\lambda, \quad \lambda^p H = H \circ \delta_\lambda.$$

Proof. The proof proceeds by constructing a minimal canonical realization as in Theorem 4.3, using some additional observations based on Proposition 4.9. In particular, $\mathscr{R}^k \cap (\mathscr{S}^{k+1} + \mathscr{N}) = \mathscr{N} \cap \mathscr{R}^k$ implies that $\mathscr{S}^k + \mathscr{N}/\mathscr{S}^{k+1} + \mathscr{N} \cong \mathscr{R}^k/\mathscr{N} \cap \mathscr{R}^k$. Thus, any basis for $\mathscr{R}^i \cap \mathscr{N} = \mathrm{Sp}\{b_{r_i'} \cdots b_{s_i}\}$ can be completed to a basis for $\mathscr{R}^i = \mathrm{Sp}\{a_{r_i} \cdots a_{s_i} b_{r_i'} \cdots b_{s_i}\}$, such that the vector fields $a_{r_i} \cdots a_{s_i}$ complete any basis of $\mathscr{S}^{i+1} + \mathscr{N}$ to a basis for $\mathscr{S}^i + \mathscr{N}$. Of course, in general $(\mathscr{R}^i \cap \mathscr{N}) \cap (\mathscr{R}^j \cap \mathscr{N}) \neq \{0\}$, $i \neq j$, so that the vector fields $b_i \cdots b_n$ are not linearly independent. However, $\mathscr{R}^i \cap (\mathscr{S}^{i+1} + \mathscr{N}) = \mathscr{N} \cap \mathscr{R}^i$, so that $(\mathscr{R}^i + \mathscr{S}^{i+1}) \cap \mathscr{N} = \mathscr{R}^i \cap \mathscr{N} + \mathscr{S}^{i+1} \cap \mathscr{N}$, which shows that $\mathscr{N} = \mathrm{Sp}\{b_1 \cdots b_n\}$.

The canonical realization is now constructed from the vector fields

$$a_1 \cdots a_{s_1} a_{r_2} \cdots a_{s_i} \cdots a_n b_1 \cdots b_{s'_1} b_{s'_2} \cdots b_n.$$

Although these vector fields which span $\mathscr{S}$ are not linearly independent, the linear relations are confined to $\mathscr{N}$. Thus it is readily observed that the system construction employed in Theorem 4.3 proceeds by using the vector fields $a_{r_i} \cdots a_{s_i} b_{r'_i} \cdots b_{s'_i}$ as a basis for $\mathscr{R}^i$ and computing the constants in the canonical realization from the identities $[f, \mathscr{R}^i] \subset \mathscr{R}^i$; $[\mathscr{R}^i, \mathscr{R}^j] \subset \mathscr{R}^{i+j}$.

Moreover, if the construction is repeated for the system (25) using the vector fields

$$\lambda a_1 \cdots \lambda a_{s_1} \lambda^2 a_{r_2} \cdots \lambda^p a_{s_p} \lambda b_1 \cdots \lambda b_{s'_1} \lambda^2 b_{r'_2} \cdots \lambda^p b_{s'_p},$$

it is observed that the constants in the canonical realization are independent of λ. Therefore, as in Lemma 4.4, the following systems in canonical form are obtained for $\lambda > 0$.

$$\dot{z}_\lambda = F(z_\lambda) + \sum_{i=1}^{m} u_i G_i(z_\lambda), \qquad z_\lambda(0) = 0, \quad z_\lambda \in \mathbb{R}^n,$$

$$y = H_\lambda(z_\lambda),$$

where $F = \delta_\lambda^{-1} \circ F \circ \delta_\lambda$, $G_i = \lambda \delta_\lambda^{-1} \circ G_i \circ \delta_\lambda$ and $H_\lambda = H \circ \delta_\lambda$.

This proves the first two identities of the theorem. The third identity follows simply from the homogeneity of the Volterra series, i.e., $y(t, \lambda u) = \lambda^p y(t, u)$. Q.E.D.

Notice that the relations given in this result pick out the leading terms in the expansions of Lemma 4.4. It is interesting to compare these results and those of § 4.1 with existing literature on the structure of nilpotent Lie groups and algebras (see for example Goodman [10]).

As an application of the above result, a topological property of the reachable set for these systems is identified. In general, given a system, define $R(T, K)$ to be the set of states reachable from the initial state at time T by controls u_i satisfying $|u_i(t)| \leqq K$, $t \in [0, T]$, $i = 1 \cdots m$.

COROLLARY 4.1.1. *A minimal realization of a stationary homogeneous Volterra series has a contractible reachable set $R(T, K)$ for all $T, K > 0$.*

Proof. Since contractibility is a homeomorphism invariant, it is sufficient to consider the reachable set of the canonical realization constructed in Theorem 4.10. Here, the solutions $z_\lambda(t)$ satisfy $z_\lambda(t) = \delta_\lambda(z_1(t))$ for controls u_i satisfying $|u_i(t)| \leqq K$, $t \in [0, T]$, $i = 1 \cdots m$. In particular, $|\lambda u_i(t)| \leqq K$ for $0 \leqq \lambda \leqq 1$. Thus, if $z \in R(T, K)$, $\delta_\lambda(z)$ also belongs to $R(T, K)$. Consider the map $C : [0, 1] \times \mathbb{R}^n \to \mathbb{R}^n$ given by $(\lambda, z) \to \delta_\lambda(z)$. By the above, C restricts to a smooth map

$$C : [0, 1] \times R(T, K) \to R(T, K).$$

It is therefore evident that C defines a smooth contraction of $R(T, K)$ to $\{0\} \in R(T, K)$. Q.E.D.

The special structure exhibited by the dynamics of a canonical realization of a homogeneous Volterra series is due to the relations $\mathscr{R}^k \cap (\mathscr{S}^{k+1} + \mathscr{N}) = \mathscr{N} \cap \mathscr{R}^k$, $1 \leqq k \leqq p$. It is possible to construct a strongly accessible canonical realization with this special structure for any realizable finite Volterra series, by lifting the dynamics to a suitable Lie group.

Given a strongly accessible observable realization of a finite Volterra series with Lie algebra $\mathscr{L}$ generated by f and g_i, $i = 1 \cdots m$. Construct an abstract Lie algebra $\mathscr{L}'$ from $\mathscr{L}$ by discarding the linear relations between elements of distinct subspaces $\mathscr{R}^i$. (This does not affect the validity of the Jacobi relation.) The map $l : \mathscr{L}' \to \mathscr{L}$ given on generators by $f' \to f$, $g_i' \to g_i$, is a homomorphism, and since $\mathscr{L}$ is finite dimensional, $\mathscr{R}^i$ is finite dimensional and so $\mathscr{L}'$ is also finite dimensional. By Hochschild [11, Theorem 1.1, p. 133], there is a simply connected Lie group G' whose algebra is given by $\mathscr{L}'$. The map l therefore extends to a homomorphism l' of G' onto G, the connected Lie transformation group of the minimal system. An accessible system with the same input–output map can now be given as

$$\dot{x}' = f'(x') + \sum_{i=1}^{m} u_i g_i'(x'), \qquad x'(0) = e, \quad x' \in G',$$

$$y = h \circ \phi(l'(x'), x_0) = h'(x'),$$

where ϕ is the natural action of G on M. By noting that $h'(\exp t(f' + a_0')) = h'(e)$, where $l(a_0') = a_0 \in \mathscr{S}$ is the vector field selected in Theorem 4.3, and for this system $\mathscr{N} = 0$, $\mathscr{R}^i \cap \mathscr{R}^j = \{0\}$, $i \neq j$, the procedure of Theorem 4.10 can be followed to obtain the required strongly accessible canonical realization.

The failure of a minimal system to satisfy the relations $\mathscr{R}^k \cap (\mathscr{S}^{k+1} + \mathscr{N}) = \mathscr{R}^k \cap \mathscr{N}$, can therefore be viewed as the result of a lack of observability in a strongly accessible system which does. Notice that this realization has the structure of that constructed in Gilbert [9]. This section is concluded by giving an example of this lifting procedure.

Example 4.12. Consider the following system in canonical form:

$$\begin{aligned}
\dot{z}_1 &= u, & z_1(0) &= 0,\\
\dot{z}_2 &= z_1 + u, & z_2(0) &= 0,\\
\dot{z}_3 &= z_1^2, & z_3(0) &= 0,\\
y &= z_3 + z_2 + z_1^2.
\end{aligned}$$

Notice that the dynamics of this system has the special form of Theorem 4.10, since $\delta_\lambda(z) = (\lambda z_1, \lambda z_2, \lambda^2 z_3)$ satisfies

$$(\lambda \dot{z}_1) = \lambda u,$$
$$(\lambda \dot{z}_2) = (\lambda z_1) + (\lambda u),$$
$$(\lambda^2 \dot{z}_3) = (\lambda z_1)^2.$$

For this system $f' = z_1 \partial/\partial z_2 + z_1^2 \partial/\partial z_3$, $g' = \partial/\partial z_1 + \partial/\partial z_2$, $[f', g'] = \partial/\partial z_2 + 2z_1 \partial/\partial z_3$, $[g', [f', g']] = 2\partial/\partial z_3$, and all other brackets are zero, so that the system is strongly accessible. However, this system is obviously not observable. Note also that $\mathscr{R}^1 \cap \mathscr{R}^2 = \{0\}$ and $\mathscr{N} = \{0\}$, so $\mathscr{R}^1 \cap (\mathscr{R}^2 + \mathscr{N}) = \mathscr{N} \cap \mathscr{R}^1$.

A minimal system with the same input–output map is given by the system in Example 4.7. In this system,

$$[f, g] - \frac{1}{2}[g, [f, g]] = 2x_1 \frac{\partial}{\partial x_2},$$

so that $\mathscr{N} \neq \{0\}$, $\mathscr{R}^1 \cap \mathscr{N} = \{0\}$, and the condition $\mathscr{R}^1 \cap (\mathscr{R}^2 + \mathscr{N}) = \mathscr{N} \cap \mathscr{R}^1$ is not satisfied.

5. Conclusions. The work presented here not only answers a basic question concerning the natural state space for realizations of finite Volterra series, but also provides a unifying approach for the study of these systems.

In the last section of the paper, some of the more interesting features of the theory of canonical realizations have been developed, emphasizing the manner in which they generalize the linear theory.

Some obvious omissions include a generalization of the rank conditions in the linear theory, for the minimality of canonical systems and a specific representation for the isomorphism between two minimal canonical realizations of the same input–output map. Another omission is the study of minimal realizations of finite Volterra series which are not strongly accessible, or in other words, systems which are intrinsically time varying. Some work on this has been presented in Crouch [8].

In § 4.3, an algorithm is developed to compute the important structural invariants in a canonical realization of a finite Volterra series.

In future papers, an algorithm will be given which constructs a minimal canonical realization from the Volterra kernels, expanding on similar constructions in the bilinear cases [1], [2].

Finally, it is evident that many of the techniques employed in this paper can be applied to systems with solvable Lie algebra (see Crouch [8]).

Acknowledgment. The author would like to thank Professor R. W. Brockett for his encouragement and helpful suggestions while the material for this paper was being prepared.

REFERENCES

[1] P. D'ALESSANDRO, A. ISIDORI AND A. RUBERTI, *Realization and structure theory of bilinear systems*, this Journal, 12 (1974), pp. 517–535.

[2] R. W. BROCKETT, *Volterra series and geometric control theory*, Automatica, 12 (1976), pp. 167–176.

[3] R. W. BROCKETT AND E. G. GILBERT, *An addendum to Volterra series and geometric control theory*, Automatica, 12 (1976), p. 635.

[4] R. W. BROCKETT, *On the Algebraic Structure of Bilinear Systems*, in Theory and Applications of Variable Structure Systems, R. R. Mohler and A. Ruberti, eds., Academic Press, New York, 1972, pp. 153–168.

[5] K. T. CHEN, *Decomposition of differential equations*, Ann. Math., 146 (1962), pp. 263–278.
[6] C. CHEVALLEY, *The Theory of Lie Groups*, Princeton Mathematical Series No. 8, Princeton University Press, Princeton, NJ, 1946.
[7] ———, *On the topological structure of solvable groups*, Ann. Math., 42 (1941), pp. 668–675.
[8] P. E. CROUCH, *Dynamical Realizations of Finite Volterra Series*, Ph.D. Thesis, Harvard University, Cambridge, MA, 1977.
[9] E. G. GILBERT, *Functional expansions for nonlinear differential systems*, IEEE Trans. Automat. Control, 22 (1977), pp. 900–921.
[10] R. W. GOODMANN, *Nilpotent Lie groups: Structure and Applications to Analysis*, Lecture Notes in Mathematics, 562, Springer-Verlag, New York, 1976.
[11] G. HOCHSCHILD, *The Structure of Lie groups*, Holden-Day, New York, 1965.
[12] A. J. KRENER, *On the equivalence of control systems and the linearization of nonlinear systems*, SIAM J. Control, 11 (1973), pp. 670–676.
[13] ———, *A decomposition theory for differentiable systems*, this Journal, 15 (1977), pp. 813–829.
[14] A. J. KRENER AND R. HERMANN, *Nonlinear observability and controllability*, IEEE Trans. Automat. Control, 22 (1977), 728–740.
[15] A. J. KRENER AND C. M. LESIAK, *The existence and uniqueness of Volterra series for nonlinear systems*, IEEE Trans. Automat. Control, 23 (1978), pp. 1090–1095.
[16] A. J. KRENER, *A generalization of Chow's theorem and the bang–bang theorem for nonlinear control problems*, SIAM J. Control, 12 (1974), pp. 43–52.
[17] A. MALCEV, *On a class of homogeneous space*, Izv. Akad. Nauk SSSR., 13 (1949), pp. 9–32; AMS Translation Series No. 39, 1951.
[18] Y. MATSUSHIMA, *On the discrete subgroups and homogeneous spaces of nilpotent Lie groups*, Nagoya Math. J., 12 (1951), pp. 95–100.
[19] R. S. PALAIS, *A global formulation of the Lie theory of transitive groups*, Mem. AMS, 22 (1957).
[20] H. J. SUSSMANN, *Existence and uniqueness of minimal realizations of nonlinear systems*, Math. Systems Theory, 10 (1977), pp. 263–284.
[21] ———, *A generalization of the closed subgroup theorem to quotients of arbitrary manifolds*, Differential Geom., 10 (1976), pp. 151–166.
[22] H. J. SUSSMANN AND V. JURDEVIC, *Controllability of nonlinear systems*, Differential Equations, 12 (1972), pp. 313–329.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 2, March 1981

0363-0129/81/1902-0003/$01.00/0

PERTURBATIONS OF NONLINEAR CONTROLLABLE SYSTEMS*

KEVIN A. GRASSE†

Abstract. In the class of nonlinear, nonautonomous control systems we consider the property of controllability to a compact set on a fixed time interval, and we give a sufficient condition for this property to be preserved under small perturbations of the control system. Our results are formulated in terms of control vector fields on a differentiable manifold.

1. Introduction. In this paper, we will be concerned with the effect of small perturbations on a certain controllability property of nonlinear control systems. More specifically, let $f: \mathbb{R}\times\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}^n$ be a mapping satisfying appropriate regularity conditions (e.g., f is C^1) and consider the control system

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

where the controls are taken from some "admissible" subclass of the set of measurable mappings of $\mathbb{R}$ into $\mathbb{R}^m$. We can pose the following general problem. If the control system (1) has a certain controllability property, then under what conditions does the perturbed control system

$$\dot{x} = f(t, x, w) + g(t, x, w),$$

for sufficiently "small" mappings g, also share this controllability property?

The following two definitions will serve to illustrate typical controllability properties. Let $[t_0, t_1]$ be a compact interval with $t_1 > t_0$. The control system (1) is said to be *strongly controllable from* $(t_0, x_0)\in\mathbb{R}\times\mathbb{R}^n$ *to a subset* $C\subseteq\mathbb{R}^n$ *at time* t_1, if for every x in C there exist an admissible control u and an absolutely continuous solution φ of the differential equation $\dot{x} = f(t, x, u(t))$ satisfying $\varphi(t_0) = x_0$ and $\varphi(t_1) = x$. The control system (1) is said to be *completely controllable* on $[t_0, t_1]$ if for every x in $\mathbb{R}^n$ it is strongly controllable from (t_0, x) to all of $\mathbb{R}^n$ at time t_1. One can also formulate weaker notions of controllability by, for example, relaxing the requirement that all points be reached at the same time.

There are two reasons why it is important to obtain results that guarantee the persistence of controllability properties under small perturbations of the control system. First, any such result can be interpreted as giving information about the well-posedness (or structural stability) of the controllability property in question, and such information has obvious physical significance. Second, if the results are formulated in a sufficiently constructive manner, then one can produce new examples of control systems having a given controllability property via perturbations of systems already known to have the property. Since our results and methods of proof are rather nonconstructive, we offer the first reason as the primary motivation for our work.

This type of problem has received a considerable amount of attention, and we will briefly mention some of the relevant literature. Nonlinear perturbations of linear control systems,

$$\dot{x} = A(t)x + B(t)w, \tag{2}$$

have been studied by Aronsson [3], Dauer [7] and Lukes [16]. The essence of their results is that the complete controllability of (2) on the interval $[t_0, t_1]$ implies the

* Received by the editors November 8, 1979.

† Department of Mathematics, University of Illinois, Urbana, Illinois 61801. Current address: Department of Mathematics, University of Oklahoma, Norman, Oklahoma 73019.

complete controllability of the perturbed system

$$\dot{x} = A(t)x + B(t)w + g(t, x, w),$$

on the interval $[t_0, t_1]$, provided that the mapping g is globally bounded or satisfies a "less-than-linear" growth condition in its last two variables. Dauer [8], [9], [10] has obtained similar results for nonlinear perturbations of nonlinear control systems having any one of the following three forms:

$$\dot{x} = A(t)x + h(t, w), \tag{3}$$

$$\dot{x} = A(t, x, w)x + B(t, x, w)w, \tag{4}$$

$$\dot{x} = h(t, x) + B(t, x)w. \tag{5}$$

It should be pointed out that the results pertaining to the systems (4) and (5) require some restrictive assumptions on the right-hand sides of the differential equations defining the control systems. We will not elaborate on these assumptions here. In [5], Brunovsky and Lobry consider time-invariant control systems of the form $\dot{x} = H(x)w$, where H is a smooth $(n \times m)$-matrix-valued function of x, and they show that, for a large class of such systems, variable-endtime controllability to a compact set is preserved under sufficiently small nonlinear perturbations. Finally, in a more geometric context, Sussmann [17] has proved that the set of k-tuples ($k \geqq 2$) of C^1 vector fields on a differentiable manifold that are globally controllable (in variable endtime) is an open subset of the set of all k-tuples of C^1 vector fields in the fine C^1 topology. In other words, controllability of a finite system of vector fields is preserved under sufficiently small C^1 perturbations. A finite set of vector fields on $\mathbb{R}^n$, when viewed as a control system, corresponds to the system (1), where the controls are taken to be piecewise constant with values in some fixed, finite subset of $\mathbb{R}^m$.

Our subject of study here is the property of strong controllability to a compact set in nonlinear, nonautonomous control systems. The main theorem of this paper (see § 4) gives a sufficient condition for this property to be stable under small perturbations. This result is in much the same spirit as a related result of Brunovsky and Lobry [5, Prop. III–6]. We note that some of our methods of proof are adaptations of their techniques, which in turn can be traced back to the paper of Aronsson [3].

We have decided to adopt a geometric point of view and phrase our results in terms of control vector fields on a differentiable manifold. The paper of Brockett [4] contains several examples which provide some justification and motivation for allowing the state space to be a manifold.

The remainder of this paper is organized as follows. In § 2, we recall some definitions and outline the basic properties of a control vector field. Section 3 contains a topological covering theorem which is a modest extension of a lemma of Brunovsky and Lobry [5, Lemma I–1]. This covering theorem is an essential tool in the proof of our main theorem in § 4. The main theorem is formulated with a certain technical hypothesis (the existence of "normal values") and we discuss this hypothesis in greater detail in § 5. Finally, in § 6, we illustrate our results with two examples.

2. Preliminaries. Our purpose in this section is to define the flow of a control vector field and establish its basic properties. For reasons of spatial economy, we will omit the proofs of the results contained in this section and direct the reader to [11, Chap. III] for the details. Some of these results can actually be regarded as standard and the others can be proved by routine, although sometimes lengthy, arguments. We begin by introducing some notation and definitions.

The term *measure* will always refer to Lebesgue measure on the real line $\mathbb{R}$. We reserve the letter J to stand for an interval in $\mathbb{R}$ which is not necessarily open or closed and contains more than one point. The notation $L(\mathbb{R}^p; \mathbb{R}^q)$ denotes the vector space of linear mappings of $\mathbb{R}^p$ into $\mathbb{R}^q$, and we equip $L(\mathbb{R}^p; \mathbb{R}^q)$ with the usual operator norm. Let V be an open subset of $\mathbb{R}^n$ and let $f: J \times V \times \mathbb{R}^m \to \mathbb{R}^n$ be a mapping such that for every t in J the mapping $(x, w) \mapsto f(t, x, w)$ is of class C^1. We let $D_2 f: J \times V \times \mathbb{R}^m \to L(\mathbb{R}^n; \mathbb{R}^n)$ and $D_3 f: J \times V \times \mathbb{R}^m \to L(\mathbb{R}^m; \mathbb{R}^n)$ denote the partial derivatives of f with respect to its second and third variables, respectively. The notation

$$\bar{D} f: J \times V \times \mathbb{R}^m \to L(\mathbb{R}^n \times \mathbb{R}^m; \mathbb{R}^n)$$

will stand for the partial derivative of f with respect to the pair of variables (x, w). The higher partial derivatives of f with respect to (x, w) are denoted by $\bar{D}^i f$ $(i > 1)$; by convention, $\bar{D}^0 f = f$ and $\bar{D}^1 f = \bar{D} f$.

DEFINITION 2.1. Let J and V be as above. A mapping $f: J \times V \times \mathbb{R}^m \to \mathbb{R}^n$ is said to be *quasi-C^r* $(1 \leqq r < \infty)$ if the following conditions are satisfied:

(i) for every t in J the mapping $(x, w) \mapsto f(t, x, w)$ is of class C^r;

(ii) for each $i = 0, 1, \cdots, r$ the mapping $\bar{D}^i f$ is locally bounded on $J \times V \times \mathbb{R}^m$;

(iii) for every (x, w) in $V \times \mathbb{R}^m$ and each $i = 0, 1, \cdots, r$ the mapping $t \mapsto \bar{D}^i f(t, x, w)$ is measurable.

The mapping f is said to be *quasi-C^∞* if it is quasi-C^r for each positive integer r.

Remark 2.2. If $f: J \times V \times \mathbb{R}^m \to \mathbb{R}^n$ is a quasi-C^r mapping, then the mean-value theorem and the local boundedness of $\bar{D} f$ imply that f satisfies a local Lipschitz condition in the variables (x, w). From this fact, one can further conclude that the mapping f satisfies a Lipschitz condition in the variables (x, w) on every compact subset of $J \times V \times \mathbb{R}^m$.

In order to formulate our results in a coordinate-free manner, we will assume that the state space of the control system is a differentiable manifold. The following paragraph contains some relevant conventions and definitions.

Differentiable manifolds 2.3.

(i) Unless stated otherwise, M will denote a finite-dimensional, second-countable, Hausdorff, differentiable manifold of class C^k with $k \geqq 2$; in particular, M is a metrizable topological space. We let TM denote the tangent bundle, $\pi: TM \to M$ the canonical projection, and $T_x M = \pi^{-1}(x)$ the tangent space to M at $x \in M$. Recall that TM is a C^{k-1} manifold and π is a C^{k-1} submersion.

(ii) Let M be n-dimensional and let (φ, U) be a coordinate chart of M. Since the mapping φ is a diffeomorphism of U onto an open subset of $\mathbb{R}^n$, for x in U the differential $d\varphi_x: T_x M \to T_{\varphi(x)} \mathbb{R}^n$ is a linear isomorphism. We will make the usual identification $T_{\varphi(x)} \mathbb{R}^n \cong \mathbb{R}^n$ and hence, regard $d\varphi_x$ as a linear isomorphism of $T_x M$ onto $\mathbb{R}^n$.

(iii) Let $\sigma: J \to M$ be an absolutely continuous mapping; i.e., for every chart (φ, U) of M the mapping $\varphi \circ \sigma$ is absolutely continuous (in the usual sense) on every compact subinterval of J where it is defined. Let $t_0 \in J$ and let (φ, U) be a chart of M such that $\sigma(t_0) \in U$. If the derivative $\dot{(\widehat{\varphi \circ \sigma})}(t_0)$ exists, then we define $\dot{\sigma}(t_0)$ in $T_{\sigma(t_0)} M$ by

$$\dot{\sigma}(t_0) = (d\varphi_{\sigma(t_0)})^{-1}(\dot{(\widehat{\varphi \circ \sigma})}(t_0)).$$

This definition is independent of the choice of the chart whose domain contains $\sigma(t_0)$. If $\sigma: J \to M$ is absolutely continuous, then it is clear that $\dot{\sigma}(t)$ exists a.e. for $t \in J$.

DEFINITION 2.4. (compare [2, Chap. 2]). Let J be an interval and let M be an n-dimensional C^k manifold. A mapping $\xi: J \times M \times \mathbb{R}^m \to TM$ is called a *quasi-C^r control vector field* on M (with control space $\mathbb{R}^m$), $1 \leqq r \leqq k-1$, if the following two

conditions are satisfied:

(i) for every (t, x, w) in $J \times M \times \mathbb{R}^m$, we have $(\pi \circ \xi)(t, x, w) = x$;

(ii) for every chart (φ, U) of M, the mapping

$$\xi_U : J \times \varphi(U) \times \mathbb{R}^m \to \mathbb{R}^n,$$

defined by

$$\xi_U(t, y, w) = d\varphi_{\varphi^{-1}(y)}(\xi(t, \varphi^{-1}(y), w)),$$

is a quasi-C^r mapping. The mapping ξ_U will be referred to as the *local representative* of ξ with respect to the chart (φ, U).

Remarks 2.5.

(i) If J is an open interval and $\xi : J \times M \times \mathbb{R}^m \to TM$ is a C^r mapping (in the usual sense) satisfying 2.4(i), then ξ is a quasi-C^r control vector field on M.

(ii) An important special case of Definition 2.4 occurs when $M = V$, an open subset of $\mathbb{R}^n$. By virtue of condition 2.4.(i) and the fact that $TV = V \times \mathbb{R}^n$, we can define a quasi-$C^r$ mapping $f : J \times V \times \mathbb{R}^m \to \mathbb{R}^n$ by the relation

$$\xi(t, x, w) = (x, f(t, x, w))$$

for (t, x, w) in $J \times V \times \mathbb{R}^m$. Alternatively, f is the local representative of ξ with respect to the chart (id_V, V). In this case one usually refers to the *control system* $f : J \times V \times \mathbb{R}^m \to \mathbb{R}^n$ on V rather than the control vector field $\xi : J \times V \times \mathbb{R}^m \to TV$.

Let $L_\infty^m(J)$ denote the set of (equivalence classes of) essentially bounded, measurable mappings of J into $\mathbb{R}^m$. Recall that $L_\infty^m(J)$ is a Banach space when equipped with the essential-supremum norm. We denote this norm by $\|\cdot\|_\infty$. Elements of $L_\infty^m(J)$ will be referred to as *controls*.

DEFINITION 2.6. Let M be an n-dimensional C^k manifold and let $\xi : J \times M \times \mathbb{R}^m \to TM$ be a quasi-C^r control vector field on M $(1 \leqq r \leqq k-1)$. If $(s, x) \in J \times M$ and $u \in L_\infty^m(J)$, then a *response of ξ with initial condition (s, x) corresponding to the control u* is a mapping $\sigma : I \to M$ such that:

(i) I is a subinterval of J containing s and σ is absolutely continuous;

(ii) $\sigma(s) = x$;

(iii) $\dot{\sigma}(t) = \xi(t, \sigma(t), u(t))$ a.e. for $t \in I$.

We have built enough regularity into the definition of a quasi-C^r control vector field to ensure the existence and uniqueness of responses for a given choice of initial condition and control. This is stated formally in the following standard theorem.

THEOREM 2.7. *Let J be an interval, let M be an n-dimensional C^k manifold, and let $\xi : J \times M \times \mathbb{R}^m \to TM$ be a quasi-C^r control vector field on M $(1 \leqq r \leqq k-1)$. Then for every (s, x) in $J \times M$ and every u in $L_\infty^m(J)$, there exist an interval $J(s, x, u)$ containing s, contained in and open in J, and a unique response*

$$\mu_{(s,x,u)} : J(s, x, u) \to M,$$

of ξ with initial condition (s, x) corresponding to the control u, which has the following maximality property. If $\sigma : I \to M$ is any response of ξ with initial condition (s, x) corresponding to the control u, then $I \subseteq J(s, x, u)$ and $\mu_{(s,x,u)}|_I = \sigma$.

DEFINITION 2.8. Let $\xi : J \times M \times \mathbb{R}^m \to TM$ be a quasi-C^r control vector field on M. Denote by $\mathscr{D}(\xi)$ the subset of $J \times J \times M \times L_\infty^m(J)$ given by

$$\mathscr{D}(\xi) = \{(t, s, x, u) \in J \times J \times M \times L_\infty^m(J) \mid t \in J(s, x, u)\},$$

and define a mapping $\mu : \mathscr{D}(\xi) \to M$ by

$$\mu(t, s, x, u) = \mu_{(s,x,u)}(t).$$

The mapping μ is called the *global flow* of ξ and the set $\mathcal{D}(\xi)$ is called the *domain of definition* of the flow.

We now summarize the properties of the global flow μ that are required for our applications.

THEOREM 2.9. *Let J be an interval, let M be an n-dimensional C^k manifold and let $\xi: J \times M \times \mathbb{R}^m \to TM$ be a quasi-C^r control vector field on M $(1 \leqq r \leqq k-1)$. The global flow $\mu: \mathcal{D}(\xi) \to M$ of ξ has the following properties:*

(i) *$\mathcal{D}(\xi)$ is an open subset of $J \times J \times M \times L^m_\infty(J)$;*

(ii) *μ is continuous;*

(iii) *for every (t, s) in $J \times J$ the set $\mathcal{D}_{(t,s)}(\xi)$ defined by*

$$\mathcal{D}_{(t,s)}(\xi) = \{(x, u) \in M \times L^m_\infty(J) \mid (t, s, x, u) \in \mathcal{D}(\xi)\}$$

is an open subset of $M \times L^m_\infty(J)$;

(iv) *if $\mathcal{D}_{(t,s)}(\xi)$ is nonempty, then the mapping $(x, u) \mapsto \mu(t, s, x, u)$ of $\mathcal{D}_{(t,s)}(\xi)$ into M is of class C^r.*

Remark 2.10. The most difficult part of the preceding theorem is the proof of the differentiable dependence of the flow on the control variable (in the sense of the Fréchet derivative). In the situation where M is an open subset of $\mathbb{R}^n$, the C^1 dependence of the flow on the control variable is a result of Lee and Markus [15, pp. 379–380]. A detailed proof of the C^r case, in the context of control vector fields, can be found in [11, §§ 3.3 and 3.4].

We conclude this section with several useful results concerning control systems defined on open subsets of $\mathbb{R}^n$. In what follows, J denotes an interval, V denotes an open subset of $\mathbb{R}^n$ and $f: J \times V \times \mathbb{R}^m \to \mathbb{R}^n$ denotes a quasi-C^r mapping $(1 \leqq r \leqq \infty)$, viewed as a control system on V (see Remark 2.5(ii)). We let $\mu: \mathcal{D}(f) \to V$ denote the global flow of f.

PROPOSITION 2.11. *For every (s_0, x_0, u_0) in $J \times V \times L^m_\infty(J)$ and every compact interval I containing s_0 and contained in $J(s_0, x_0, u_0)$, there exist an open neighborhood $V_0 \times \mathcal{N}_0$ of (x_0, u_0) in $V \times L^m_\infty(J)$ and a positive constant A such that*

$$I \times \{s_0\} \times V_0 \times \mathcal{N}_0 \subseteq \mathcal{D}(f),$$

and

$$\|\mu(t, s_0, x, u) - \mu(t, s_0, \bar{x}, \bar{u})\| \leqq A \cdot \max\{\|x - \bar{x}\|, \|u - \bar{u}\|_\infty\},$$

for every t in I and (x, u), $(\bar{x}, \bar{u})$ in $V_0 \times \mathcal{N}_0$ ($\|\cdot\|$ denotes any convenient norm on $\mathbb{R}^n$).

PROPOSITION 2.12. *Fix a point (t_0, s_0, x_0, u_0) in $\mathcal{D}(f)$ and let $\bar{D}\mu(t_0, s_0, x_0, u_0)$ denote the partial derivative of μ with respect to its last two variables at the point (t_0, s_0, x_0, u_0). Then for every (h, v) in $\mathbb{R}^n \times L^m_\infty(J)$, we have*

$$\bar{D}\mu(t_0, s_0, x_0, u_0)(h, v) = \lambda(t_0, s_0, x_0, u_0, h, v),$$

where the mapping $t \mapsto \lambda(t, s_0, x_0, u_0, h, v)$ of $J(s_0, x_0, u_0)$ into $\mathbb{R}^n$ is the unique absolutely continuous solution of the linear differential equation

$$\dot{x} = D_2 f(t, \mu(t, s_0, x_0, u_0), u_0(t))x + D_3 f(t, \mu(t, s_0, x_0, u_0), u_0(t))v(t) \tag{6}$$

satisfying the initial condition (s_0, h).

COROLLARY 2.13. *For every v in $L^m_\infty(J)$, we have*

$$D_4\mu(t_0, s_0, x_0, u_0)v = \psi(t_0, s_0, x_0, u_0, v),$$

where the mapping $t \mapsto \psi(t, s_0, x_0, u_0, v)$ of $J(s_0, x_0, u_0)$ into $\mathbb{R}^n$ is the unique absolutely continuous solution of the linear differential equation (6) *satisfying the initial condition $(s_0, 0)$.*

3. A topological covering theorem. The main result of this section is a topological covering theorem, which is a straightforward extension of a lemma of Brunovsky and Lobry [5, Lemma I-1] to mappings taking values in a differentiable manifold (as opposed to $\mathbb{R}^n$). To provide a means of verifying the hypotheses of the covering theorem in certain situations, we will also introduce the notion of a normal value of a differentiable mapping.

Our first lemma is essentially standard, but we will include the proof as it is very short. The notation $B(0, \alpha)$ denotes the open ball in $\mathbb{R}^n$ centered at 0 and of radius α, while $\bar{B}(0, \alpha)$ denotes its closure.

LEMMA 3.1. *Let X be a topological space, let $h: X \to \mathbb{R}^n$ be a continuous mapping, and let $s: B(0, 3) \to X$ be a continuous mapping such that $(h \circ s)(y) = y$, for every y in $B(0, 3)$ (in particular, $h(X) \supseteq B(0, 3)$). If $\tilde{h}: s(\bar{B}(0, 2)) \to \mathbb{R}^n$ is a continuous mapping satisfying $\|\tilde{h}(x) - h(x)\| \leqq 1$ for every x in $s(\bar{B}(0, 2))$, then $\tilde{h}(s(\bar{B}(0, 2))) \supseteq B(0, 1)$.*

Proof. Fix $\bar{y}$ in $B(0, 1)$ and define a mapping $H: \bar{B}(0, 2) \to \mathbb{R}^n$ by

$$H(y) = \bar{y} + y - \tilde{h}(s(y)).$$

Clearly, H is continuous and $H(\bar{B}(0, 2)) \subseteq \bar{B}(0, 2)$, since for y in $\bar{B}(0, 2)$ we have

$$\begin{aligned} \|H(y)\| &\leqq \|\bar{y}\| + \|y - \tilde{h}(s(y))\| \\ &= \|\bar{y}\| + \|h(s(y)) - \tilde{h}(s(y))\| < 2. \end{aligned}$$

By the Brouwer Fixed-Point Theorem, there exists z in $\bar{B}(0, 2)$ such that $H(z) = z$. This implies $\bar{y} = \tilde{h}(s(z)) \in \tilde{h}(s(\bar{B}(0, 2)))$ and, because $\bar{y} \in B(0, 1)$ was arbitrary, the proof is complete. □

LEMMA 3.2. *Let X be a topological space, let M be a finite-dimensional manifold, and let d be a metric on M compatible with the manifold topology. Let $h: X \to M$ be a continuous mapping and suppose that for some y in M, there exist an open neighborhood V of y and a continuous mapping $s: V \to X$ satisfying $(h \circ s)(z) = z$, for every z in V. Then there exist open neighborhoods V_1 and V_2 of y and an $\varepsilon > 0$ such that:*

(i) *$V_1 \subseteq V_2 \subseteq \bar{V}_2 \subseteq V$ and $\bar{V}_2$ is compact;*

(ii) *if $\tilde{h}: s(\bar{V}_2) \to M$ is a continuous mapping satisfying $d(\tilde{h}(x), h(x)) \leqq \varepsilon$ for every x in $s(\bar{V}_2)$, then $\tilde{h}(s(\bar{V}_2)) \supseteq V_1$.*

Proof. Once a suitable choice of coordinate chart is made at the point y, the proof follows in a routine manner from Lemma 3.1. We omit the details. □

Our proof of the following theorem is similar to the proof of the Brunovsky-Lobry lemma.

THEOREM 3.3. *Let X be a topological space, let M be a finite-dimensional manifold, and let d be a metric on M compatible with the manifold topology. Let $h: X \to M$ be a continuous mapping and let C be a compact subset of M such that $C \subseteq h(X)$ and h has a continuous local right inverse at every point of C. Then there exist a compact subset K of X and an $\varepsilon > 0$ such that if $\tilde{h}: K \to M$ is any continuous mapping satisfying $d(\tilde{h}(x), h(x)) \leqq \varepsilon$ for every x in K, then $C \subseteq \tilde{h}(K)$.*

Proof. By hypothesis, for every y in C there exist an open neighborhood V_y of y and a continuous mapping $s_y: V_y \to X$ such that $(h \circ s_y)(z) = z$, for every z in V_y. Using Lemma 3.2, we obtain, for every y in C, open neighborhoods U_y and W_y of y and an $\varepsilon_y > 0$ satisfying the following conditions:

(i) $U_y \subseteq W_y \subseteq \bar{W}_y \subseteq V_y$ and $\bar{W}_y$ are compact;

(ii) if $\tilde{h}: s(\bar{W}_y) \to M$ is a continuous mapping such that $d(\tilde{h}(x), h(x)) \leqq \varepsilon_y$ for every x in $s(\bar{W}_y)$, then $\tilde{h}(s(\bar{W}_y)) \supseteq U_y$.

The family $\{U_y | y \in C\}$ is an open cover of the compact set C, so we can extract a finite

subcover $\{U_{y_1}, \cdots, U_{y_k}\}$. It is easy to see that the set

$$K = \bigcup_{i=1}^{k} s_{y_i}(\bar{W}_{y_i}),$$

and the positive real number

$$\varepsilon = \min\{\varepsilon_{y_1}, \cdots, \varepsilon_{y_k}\},$$

satisfy our requirements. □

In order to be able to apply Theorem 3.3, we must have a reasonable way of determining when a mapping has a continuous local right inverse at a point in its image. The following proposition gives one such criterion, based on the inverse-mapping theorem.

PROPOSITION 3.4. *Let X and Y be Banach manifolds of class C^k $(1 \leqq k \leqq \infty)$ and let $h: X \to Y$ be a C^k mapping. Assume that for some x_0 in X the differential $dh_{x_0}: T_{x_0}X \to T_{h(x_0)}Y$ is surjective and its kernel splits in $T_{x_0}X$. Then there exist an open neighborhood V of $h(x_0)$ and a C^k mapping $s: V \to X$ such that $(s \circ h)(x_0) = x_0$ and $(h \circ s)(y) = y$ for every y in V.*

Proof. This is a straightforward consequence of the inverse-mapping theorem. See [14, p. 17] for details. □

DEFINITION 3.5. Let X and Y be Banach manifolds of class C^1 and let $h: X \to Y$ be a C^1 mapping. A point y_0 in $h(X)$ is called a *normal value* of h if there exists at least one x_0 in $h^{-1}(y_0)$ such that the differential dh_{x_0} is a split-surjective linear mapping (contrast this with the notion of a regular value, where it is required that dh_x be a split-surjective linear mapping *for every* x in $h^{-1}(y_0)$).

Remark 3.6. When Y is a finite-dimensional manifold, we can drop the splitting assumption in Proposition 3.4 and Definition 3.5, since in that case the kernel of dh_{x_0} is closed and of finite codimension, and hence always splits (see, e.g., [13, p. 186]).

For later reference, we state the following immediate consequence of Theorem 3.3 and Proposition 3.4.

THEOREM 3.7. *Let E be a Banach space, let M be a finite-dimensional C^1 manifold, and let d be a metric on M compatible with the manifold topology. Let U be an open subset of E and let $h: U \to M$ be a C^1 mapping. Suppose that C is a compact subset of $h(U)$ such that every point of C is a normal value of h. Then there exist a compact subset K of U and an $\varepsilon > 0$ such that if $\tilde{h}: K \to M$ is any continuous mapping satisfying $d(\tilde{h}(x), h(x)) \leqq \varepsilon$ for every x in K, then $C \subseteq \tilde{h}(K)$.*

4. The stability of controllability to a compact set under small perturbations. This section contains our main theorem concerning perturbations of control systems that are strongly controllable to a compact set. Before stating this theorem, we will need to set some notation, review a few definitions, and establish two preliminary technical propositions.

Notation 4.1. Let J be an interval, let M be an n-dimensional C^k manifold $(k \geqq 2)$, and let $\xi: J \times M \times \mathbb{R}^m \to TM$ be a quasi-C^1 control vector field on M with global flow $\mu: \mathcal{D}(\xi) \to M$. For each (t, s, x) in $J \times J \times M$ let $\mathcal{D}_{(t,s,x)}(\xi)$ denote the subset of $L_\infty^m(J)$ defined by

$$\mathcal{D}_{(t,s,x)}(\xi) = \{u \in L_\infty^m(J) \mid (t, s, x, u) \in \mathcal{D}(\xi)\},$$

and, if $\mathcal{D}_{(t,s,x)}(\xi)$ is nonempty, let $\mu_{(t,s,x)}: \mathcal{D}_{(t,s,x)}(\xi) \to M$ denote the mapping defined by

$$\mu_{(t,s,x)}(u) = \mu(t, s, x, u).$$

Remark 4.2. It is a consequence of Theorem 2.9 that the set $\mathscr{D}_{(t,s,x)}(\xi)$ is open in $L_\infty^m(J)$ and the mapping $\mu_{(t,s,x)}$ is of class C^1.

DEFINITION 4.3. Let $\xi: J \times M \times \mathbb{R}^m \to TM$ be a quasi-C^1 control vector field on M and let $(t, t_0, x_0) \in J \times J \times M$ with $t \geqq t_0$. The *attainable set of ξ from (t_0, x_0) at time t* is defined by

$$\mathscr{A}_\xi(t_0, x_0; t) = \{\mu(t, t_0, x_0, u) | u \in \mathscr{D}_{(t,t_0,x_0)}(\xi)\}.$$

Equivalently, $\mathscr{A}_\xi(t_0, x_0; t)$ is the image of the mapping $\mu_{(t,t_0,x_0)}$. Observe that we have $\mathscr{A}_\xi(t_0, x_0; t_0) = \{x_0\}$.

DEFINITION 4.4. Let $\xi: J \times M \times \mathbb{R}^m \to TM$ be a quasi-C^1 control vector field on M and let $(t_1, t_0, x_0) \in J \times J \times M$ with $t_1 > t_0$. The control vector field ξ is said to be *strongly controllable from (t_0, x_0) to a subset $C \subseteq M$ at time t_1* if $C \subseteq \mathscr{A}_\xi(t_0, x_0; t_1)$. When $\mathscr{A}_\xi(t_0, x_0; t_1) = M$, we say that ξ is *strongly globally controllable from (t_0, x_0) at time t_1*.

Our first proposition will require some additional notation. If V is an open subset of $\mathbb{R}^n$, C is a nonempty compact subset of $J \times V \times \mathbb{R}^m$, and $g: J \times V \times \mathbb{R}^m \to \mathbb{R}^n$ is a locally bounded mapping, then $\|g\|_C$ denotes the nonnegative real number defined by

$$\|g\|_C = \sup \{\|g(t, x, w)\| \,|\, (t, x, w) \in C\}.$$

We let $B(z, \alpha)$ denote the open ball centered at z and of radius α in any metric space under consideration. If A and B are subsets of a metric space with metric d, then the distance from A to B is defined by

$$\operatorname{dist}[A, B] = \inf \{d(a, b) | (a, b) \in A \times B\}.$$

PROPOSITION 4.5. *Let $f: J \times V \times \mathbb{R}^m \to \mathbb{R}^n$ be a quasi-C^1 mapping (viewed as a control system on V), let $\mu: \mathscr{D}(f) \to V$ denote the global flow of f, and fix a point (t_1, t_0, x_0, u_0) in $\mathscr{D}(f)$ with $t_1 > t_0$. Then for every $\varepsilon > 0$ there exist a $\delta > 0$ and a compact subset C of $J \times V \times \mathbb{R}^m$ satisfying the following condition. If $g: J \times V \times \mathbb{R}^m \to \mathbb{R}^n$ is a quasi-C^1 mapping, then $\|g\|_C \leqq \delta$ implies that*

$$[t_0, t_1] \times \{t_0\} \times B(x_0, \delta) \times B(u_0, \delta) \subseteq \mathscr{D}(f + g),$$

and

$$\|\tilde{\mu}(t, t_0, x, u) - \mu(t, t_0, \bar{x}, \bar{u})\| < \varepsilon,$$

for every t in $[t_0, t_1]$ and (x, u), $(\bar{x}, \bar{u})$ in $B(x_0, \delta) \times B(u_0, \delta)$, where $\tilde{\mu}: \mathscr{D}(f + g) \to V$ is the global flow of $f + g$.

Proof. By Proposition 2.11, there exist a $\delta_1 > 0$ and a positive constant A such that

$$[t_0, t_1] \times \{t_0\} \times B(x_0, \delta_1) \times B(u_0, \delta_1) \subseteq \mathscr{D}(f),$$

and

$$\|\mu(t, t_0, x, u) - \mu(t, t_0, \bar{x}, \bar{u})\| \leqq A \cdot \max\{\|x - \bar{x}\|, \|u - \bar{u}\|_\infty\}, \tag{7}$$

for every t in $[t_0, t_1]$ and (x, u), $(\bar{x}, \bar{u})$ in $B(x_0, \delta_1) \times B(u_0, \delta_1)$.

The set $K_0 = \{\mu(t, t_0, x_0, u_0) | t \in [t_0, t_1]\}$ is a compact subset of the open set V. Let $\varepsilon_1 > 0$ be sufficiently small that $\varepsilon_1 \leqq \varepsilon$ and $\bar{V}_{\varepsilon_1} \subseteq V$, where V_{ε_1} is the open set defined by

$$V_{\varepsilon_1} = \{x \in R^n | \operatorname{dist}[x, K_0] < \varepsilon_1\}.$$

Define a compact subset D_{δ_1} of $\mathbb{R}^m$ by

$$D_{\delta_1} = \{w \in \mathbb{R}^m | \operatorname{dist}[w, \overline{u_0(J)}] \leqq \delta_1\},$$

(we are tacitly identifying the equivalence class u_0 with a class representative that is pointwise bounded in norm by $\|u_0\|_\infty$ everywhere on J). Then $C = [t_0, t_1] \times \bar{V}_{\varepsilon_1} \times D_{\delta_1}$ is a compact subset of $J \times V \times \mathbb{R}^m$. By Remark 2.2, there is a positive constant L such that

$$\|f(t, x, w) - f(t, \bar{x}, \bar{w})\| \leqq L \cdot \max\{\|x - \bar{x}\|, \|w - \bar{w}\|\}, \tag{8}$$

for every (t, x, w), $(t, \bar{x}, \bar{w})$ in C. Furthermore, if we set

$$\delta_2 = \min\{\delta_1, \varepsilon_1/4A\},$$

then the choice of δ_1 and the Lipschitz condition (7) imply that

$$[t_0, t_1] \times \{t_0\} \times B(x_0, \delta_2) \times B(u_0, \delta_2) \subseteq \mathscr{D}(f),$$

and

$$\|\mu(t, t_0, x, u) - \mu(t, t_0, \bar{x}, \bar{u})\| \leqq \varepsilon_1/2, \tag{9}$$

for every t in $[t_0, t_1]$ and (x, u), $(\bar{x}, \bar{u})$ in $B(x_0, \delta_2) \times B(u_0, \delta_2)$. In particular, if we put $(\bar{x}, \bar{u}) = (x_0, u_0)$ in (9), then we infer that $\mu(t, t_0, x, u) \in V_{\varepsilon_1}$ for every (t, x, u) in $[t_0, t_1] \times B(x_0, \delta_2) \times B(u_0, \delta_2)$.

Let $g: J \times V \times \mathbb{R}^m \to \mathbb{R}^n$ be a quasi-C^1 mapping and let $\tilde{\mu}: \mathscr{D}(f+g) \to V$ denote the global flow of the control system $f+g$. We claim that if

$$\|g\|_C \leqq \frac{\varepsilon_1}{4(t_1 - t_0)} e^{-L(t_1 - t_0)}, \tag{10}$$

then

$$[t_0, t_1] \times \{t_0\} \times B(x_0, \delta_2) \times B(u_0, \delta_2) \subseteq \mathscr{D}(f+g), \tag{11}$$

and

$$\|\tilde{\mu}(t, t_0, x, u) - \mu(t, t_0, \bar{x}, \bar{u})\| < \varepsilon_1, \tag{12}$$

for every t in $[t_0, t_1]$ and (x, u), $(\bar{x}, \bar{u})$ in $B(x_0, \delta_2) \times B(u_0, \delta_2)$.

To establish the claim, we will first show that for every (x, u) in $B(x_0, \delta_2) \times B(u_0, \delta_2)$ the response $t \mapsto \tilde{\mu}(t, t_0, x, u)$ is defined on $[t_0, t_1]$ and maps $[t_0, t_1]$ into V_{ε_1}. Suppose that this is not the case. By a standard extension theorem in ordinary differential equations [18, p. 187], there exists t^* in $(t_0, t_1]$ such that:

(a) $t \mapsto \tilde{\mu}(t, t_0, x, u)$ is defined on $[t_0, t^*)$;

(b) $\tilde{\mu}(t, t_0, x, u) \in V_{\varepsilon_1}$ for every t in $[t_0, t^*)$;

(c) $\lim_{t \to t^{*-}} \tilde{\mu}(t, t_0, x, u)$ exists and is an element of $\partial V_{\varepsilon_1}$, the topological boundary of V_{ε_1}.

Inequality (10) and the Lipschitz condition (8) yield the estimate

$$\begin{aligned}
&\|\tilde{\mu}(t, t_0, x, u) - \mu(t, t_0, x, u)\| \\
&\quad = \Big\| x + \int_{t_0}^{t} [f(s, \tilde{\mu}(s, t_0, x, u), u(s)) + g(s, \tilde{\mu}(s, t_0, x, u), u(s))]\, ds \\
&\qquad - x - \int_{t_0}^{t} f(s, \mu(s, t_0, x, u), u(s))\, ds \Big\| \qquad (13) \\
&\quad \leqq \int_{t_0}^{t} \frac{\varepsilon_1 e^{-L(t_1 - t_0)}}{4(t_1 - t_0)}\, ds + \int_{t_0}^{t} \|f(s, \tilde{\mu}(s, t_0, x, u), u(s)) - f(s, \mu(s, t_0, x, u), u(s))\|\, ds \\
&\quad \leqq \frac{\varepsilon_1 e^{-L(t_1 - t_0)}}{4} + \int_{t_0}^{t} L\|\tilde{\mu}(s, t_0, x, u) - \mu(s, t_0, x, u)\|\, ds,
\end{aligned}$$

for every t in $[t_0, t^*)$. Applying the Gronwall inequality to (13), we obtain

$$\|\tilde{\mu}(t, t_0, x, u) - \mu(t, t_0, x, u)\| \leqq \varepsilon_1/4,$$

for every t in $[t_0, t^*)$. In combination with inequality (9), this implies that

$$\|\tilde{\mu}(t, t_0, x, u) - \mu(t, t_0, \bar{x}, \bar{u})\| \leqq 3\varepsilon_1/4, \tag{14}$$

for every t in $[t_0, t^*)$ and (x, u), $(\bar{x}, \bar{u})$ in $B(x_0, \delta_2) \times B(u_0, \delta_2)$. However, inequality (14) with $(\bar{x}, \bar{u}) = (x_0, u_0)$ is clearly inconsistent with statement (c) above. Hence, for every (x, u) in $B(x_0, \delta_2) \times B(u_0, \delta_2)$ the response $t \mapsto \tilde{\mu}(t, t_0, x, u)$ is defined on $[t_0, t_1]$ and maps $[t_0, t_1]$ into the set V_{ε_1}. Consequently, inequality (14) holds for all t in $[t_0, t_1]$. That establishes the truth of relations (11) and (12). The smaller of the two numbers δ_2 and $\varepsilon_1 e^{-L(t_1 - t_0)}/4(t_1 - t_0)$ will serve as the desired δ. □

DEFINITION 4.6. Let M be an n-dimensional manifold of class C^1. A *Finsler structure* on the tangent bundle TM is a continuous mapping $\omega: TM \to \mathbb{R}$ such that for every x in M the restriction $\omega|_{T_xM}: T_xM \to \mathbb{R}$ is a norm on the n-dimensional vector space T_xM. For a tangent vector v in TM, we will often use the notation $\|v\|_\omega$ in place of $\omega(v)$.

Remark 4.7. By using a partition-of-unity argument, one can easily show that any finite-dimensional C^1 manifold admits a Finsler structure on its tangent bundle.

The next proposition is the global version of Proposition 4.5. In this proposition we will employ the following notation. Let $\eta: J \times M \times \mathbb{R}^m \to TM$ be a quasi-C^1 control vector field on M and let $\omega: TM \to \mathbb{R}$ be a Finsler structure on TM. If C is a nonempty compact subset of $J \times M \times \mathbb{R}^m$, then $\|\eta\|_C$ denotes the nonnegative real number given by

$$\|\eta\|_C = \sup \{\|\eta(t, x, w)\|_\omega | (t, x, w) \in C\}.$$

PROPOSITION 4.8. *Let M be an n-dimensional manifold of class C^k $(k \geqq 2)$ with compatible metric d, let $\omega: TM \to \mathbb{R}$ be a Finsler structure on TM and let $\xi: J \times M \times \mathbb{R}^m \to TM$ be a quasi-C^1 control vector field on M. Denote by $\mu: \mathcal{D}(\xi) \to M$ the global flow of ξ and fix a point (t_1, t_0, x_0, u_0) in $\mathcal{D}(\xi)$ with $t_1 > t_0$. Then for every $\varepsilon > 0$ there exist a $\delta > 0$ and a compact subset C of $J \times M \times \mathbb{R}^m$ satisfying the following condition. If $\eta: J \times M \times \mathbb{R}^m \to TM$ is a quasi-C^1 control vector field on M satisfying $\|\eta\|_C \leqq \delta$, then we have*

$$[t_0, t_1] \times \{t_0\} \times B(x_0, \delta) \times B(u_0, \delta) \subseteq \mathcal{D}(\xi + \eta),$$

and

$$d(\tilde{\mu}(t, t_0, x, u), \mu(t, t_0, \bar{x}, \bar{u}) < \varepsilon,$$

for every t in $[t_0, t_1]$ and (x, u), $(\bar{x}, \bar{u})$ in $B(x_0, \delta) \times B(u_0, \delta)$, where $\tilde{\mu}: \mathcal{D}(\xi + \eta) \to M$ is the global flow of $\xi + \eta$.

Proof. Since the set $\{\mu(t, t_0, x_0, u_0) | t \in [t_0, t_1]\}$ is a compact subset of M, there exists a partition $\{s_0, s_1, \cdots, s_p\}$ of the interval $[t_0, t_1]$ with

$$t_0 = s_0 < s_1 < \cdots < s_p = t_1,$$

such that for each $i = 1, \cdots, p$ the set

$$\{\mu(t, t_0, x_0, u_0) | t \in [s_{i-1}, s_i]\}$$

is contained in the domain U_i of a coordinate chart (φ_i, U_i) of M. We set $x_i = \mu(s_i, s_0, x_0, u_0)$ for each $i = 1, \cdots, p$.

For $j = 1, \cdots, p$ consider the following statement.

(j) For every $\varepsilon > 0$ there exist a $\delta > 0$ and a compact set $C \subseteq J \times M \times \mathbb{R}^m$ such that if $\eta: J \times M \times \mathbb{R}^m \to TM$ is a quasi-C^1 control vector field on M satisfying

$\|\eta\|_C \leqq \delta$, then

$$[s_0, s_j]\times\{s_0\}\times B(x_0, \delta)\times B(u_0, \delta)\subseteq \mathscr{D}(\xi+\eta),$$

and

$$d(\tilde{\mu}(t, s_0, x, u), \mu(t, s_0, \bar{x}, \bar{u}))<\varepsilon,$$

for every t in $[s_0, s_j]$ and (x, u), $(\bar{x}, \bar{u})$ in $B(x_0, \delta)\times B(u_0, \delta)$.

We will show that statement (j) holds for $j=1, \cdots, p$ by induction on j. Because statement (p) is precisely the assertion of the proposition, this will complete the proof.

By applying Proposition 4.5 to the local representative ξ_{U_1} of ξ, we see that statement (j) holds with $j=1$. Assume that statement (k) holds for some k such that $1\leqq k\leqq p-1$. We must show that statement $(k+1)$ holds.

Let $\varepsilon>0$ be given. Applying Proposition 4.5 to the local representative $\xi_{U_{k+1}}$, we obtain a $\delta_{k+1}>0$ and a compact set C_{k+1} contained in $J\times U_{k+1}\times\mathbb{R}^m$ such that if $\|\eta\|_{C_{k+1}}\leqq\delta_{k+1}$, then

$$[s_k, s_{k+1}]\times\{s_k\}\times B(x_k, \delta_{k+1})\times B(u_0, \delta_{k+1})\subseteq \mathscr{D}(\xi+\eta), \tag{15}$$

and

$$d(\tilde{\mu}(t, s_k, x, u), \mu(t, s_k, \bar{x}, \bar{u}))<\varepsilon, \tag{16}$$

for every t in $[s_k, s_{k+1}]$ and (x, u), $(\bar{x}, \bar{u})$ in $B(x_k, \delta_{k+1})\times B(u_0, \delta_{k+1})$. Observe that we necessarily have $\delta_{k+1}\leqq\varepsilon$. Because statement (k) holds, there exist a $\delta_k>0$ and a compact set C_k contained in $J\times M\times\mathbb{R}^m$ such that if $\|\eta\|_{C_k}\leqq\delta_k$, then

$$[s_0, s_k]\times\{s_0\}\times B(x_0, \delta_k)\times B(u_0, \delta_k)\subseteq \mathscr{D}(\xi+\eta), \tag{17}$$

and

$$d(\tilde{\mu}(t, s_0, x, u), \mu(t, s_0, \bar{x}, \bar{u}))<\delta_{k+1}, \tag{18}$$

for every t in $[s_0, s_k]$ and (x, u), $(\bar{x}, \bar{u})$ in $B(x_0, \delta_k)\times B(u_0, \delta_k)$. Observe that we necessarily have $\delta_k\leqq\delta_{k+1}$.

Since μ is continuous, there exists a $\delta>0$ such that $\delta\leqq\delta_k$ and

$$\mu(s_k, s_0, x, u)\in B(x_k, \delta_{k+1}), \tag{19}$$

for every (x, u) in $B(x_0, \delta)\times B(u_0, \delta)$. Set $C=C_k\cup C_{k+1}$. Clearly, C is a compact subset of $J\times M\times\mathbb{R}^m$. Let $\eta: J\times M\times\mathbb{R}^m\to TM$ be a quasi-C^1 control vector field on M satisfying $\|\eta\|_C\leqq\delta$. From (17) and (18), we obtain

$$[s_0, s_k]\times\{s_0\}\times B(x_0, \delta)\times B(u_0, \delta)\subseteq \mathscr{D}(\xi+\eta), \tag{20}$$

and

$$d(\tilde{\mu}(t, s_0, x, u), \mu(t, s_0, \bar{x}, \bar{u}))<\delta_{k+1}\leqq\varepsilon, \tag{21}$$

for every t in $[s_0, s_k]$ and (x, u), $(\bar{x}, \bar{u})$ in $B(x_0, \delta)\times B(u_0, \delta)$. In particular, for (x, u) in $B(x_0, \delta)\times B(u_0, \delta)$ we have

$$\begin{aligned} d(\tilde{\mu}(s_k, s_0, x, u), \mu(s_k, s_0, x_0, u_0)) &= d(\tilde{\mu}(s_k, s_0, x, u), x_k)\\ &\leqq \delta_{k+1}, \end{aligned} \tag{22}$$

which by (15) implies that

$$(t, s_k, \tilde{\mu}(s_k, s_0, x, u), u)\in \mathscr{D}(\xi+\eta), \tag{23}$$

for every t in $[s_k, s_{k+1}]$. Relation (23) and the fact that

$$J(s_k, \tilde{\mu}(s_k, s_0, x, u), u) = J(s_0, x, u)$$

yield the inclusion

$$[s_k, s_{k+1}] \times \{s_0\} \times B(x_0, \delta) \times B(u_0, \delta) \subseteq \mathcal{D}(\xi + \eta). \tag{24}$$

Using (19), (22), (16), and the transitivity property of the flow, we obtain

$$\begin{aligned} d(\tilde{\mu}(t, s_0, x, u), \mu(t, s_0, \bar{x}, \bar{u})) &= d(\tilde{\mu}(t, s_k, \tilde{\mu}(s_k, s_0, x, u), u), \mu(t, s_k, \mu(s_k, s_0, \bar{x}, \bar{u}), \bar{u})) \\ &< \varepsilon, \end{aligned} \tag{25}$$

for every t in $[s_k, s_{k+1}]$ and (x, u), $(\bar{x}, \bar{u})$ in $B(x_0, \delta) \times B(u_0, \delta)$. Statement $(k+1)$ is now a consequence of (20), (21), (24) and (25). This completes the induction and the proof. □

We are now ready to state and prove our main theorem. This theorem requires an assumption concerning the existence of normal values of the mapping $\mu_{(t_1, t_0, x_0)}$, and such an assumption is bound to appear rather ad hoc at this point. We will make some attempt to justify the reasonableness of this assumption in the next section.

THEOREM 4.9. *Let M be an n-dimensional manifold of class C^k $(k \geqq 2)$ with compatible metric d, let $\omega: TM \to \mathbb{R}$ be a Finsler structure on TM, and let $\xi: J \times M \times \mathbb{R}^m \to TM$ be a quasi-C^1 control vector field on M. Denote by $\mu: \mathcal{D}(\xi) \to M$ the global flow of ξ and let $(t_1, t_0, x_0) \in J \times J \times M$ be such that $t_1 > t_0$ and the open subset $\mathcal{D}_{(t_1, t_0, x_0)}(\xi)$ of $L_\infty^m(J)$ is nonempty. Suppose that C is a compact subset of $\mathcal{A}_\xi(t_0, x_0; t_1)$ and every point of C is a normal value of the mapping*

$$\mu_{(t_1, t_0, x_0)}: \mathcal{D}_{(t_1, t_0, x_0)}(\xi) \to M.$$

Then there exist a $\delta > 0$ and a compact subset K of $J \times M \times \mathbb{R}^m$ such that if $\eta: J \times M \times \mathbb{R}^m \to TM$ is a quasi-C^1 control vector field on M satisfying $\|\eta\|_K \leqq \delta$, then $C \subseteq \mathcal{A}_{\xi+\eta}(t_0, x_0; t_1)$.

Proof. Recall that the mapping $\mu_{(t_1, t_0, x_0)}$ is of class C^1 and its image is precisely the attainable set $\mathcal{A}_\xi(t_0, x_0; t_1)$. By Theorem 3.7, there exist an $\varepsilon > 0$ and a compact subset F of $\mathcal{D}_{(t_1, t_0, x_0)}(\xi)$ such that if $\tilde{h}: F \to M$ is a continuous mapping satisfying

$$d(\tilde{h}(u), \mu_{(t_1, t_0, x_0)}(u)) \leqq \varepsilon,$$

for every u in F, then $C \subseteq \tilde{h}(F)$.

For every u in F we can apply Proposition 4.8 to obtain a $\delta_u > 0$ and a compact subset K_u of $J \times M \times \mathbb{R}^m$ such that if $\eta: J \times M \times \mathbb{R}^m \to TM$ is a quasi-C^1 control vector field on M satisfying $\|\eta\|_{K_u} \leqq \delta_u$, then

$$(t_1, t_0, x_0, v) \in \mathcal{D}(\xi + \eta),$$

and

$$d(\tilde{\mu}(t_1, t_0, x_0, v), \mu(t_1, t_0, x_0, v)) < \varepsilon,$$

for every v in $B(u, \delta_u)$, where $\tilde{\mu}$ is the global flow of $\xi + \eta$. The collection $\{B(u, \delta_u) | u \in F\}$ is an open cover of the compact set F, so we can extract a finite subcover

$$\{B(u_1, \delta_{u_1}), \cdots, B(u_k, \delta_{u_k})\}.$$

Let $K = \bigcup_{i=1}^k K_{u_i}$ and $\delta = \min\{\delta_{u_1}, \cdots, \delta_{u_k}\}$. Clearly, K is a compact subset of $J \times M \times \mathbb{R}^m$. Let $\eta: J \times M \times \mathbb{R}^m \to TM$ be a quasi-C^1 control vector field on M satisfying $\|\eta\|_K \leqq \delta$. If u is an arbitrary element of F, then $u \in B(u_i, \delta_{u_i})$ for some index i in

$\{1, \cdots, k\}$. Since

$$\|\eta\|_{K_{u_i}} \leqq \|\eta\|_K \leqq \delta \leqq \delta_{u_i},$$

we infer that

$$(t_1, t_0, x_0, u) \in \mathscr{D}(\xi + \eta)$$

and

$$d(\tilde{\mu}_{(t_1,t_0,x_0)}(u), \mu_{(t_1,t_0,x_0)}(u)) < \varepsilon.$$

By the choice of the number ε and the set F, we obtain

$$C \subseteq \tilde{\mu}_{(t_1,t_0,x_0)}(F) \subseteq \mathscr{A}_{\xi+\eta}(t_0, x_0; t_1).$$

This completes the proof. □

5. The existence of normal values. The hypothesis of Theorem 4.9 included the assumption that every point of the compact target set C is a normal value of the mapping $\mu_{(t_1,t_0,x_0)}$. In this section, we will make an attempt to justify this assumption by giving sufficient conditions for the existence of normal values in certain situations. We begin with some general properties of normal values.

LEMMA 5.1. *Let E and G be Banach spaces and let $L(E; G)$ denote the set of continuous linear mappings of E into G, viewed as a Banach space under the usual operator norm. The set of split-surjective, continuous linear mappings of E into G is an open subset of $L(E; G)$.*

Proof. See [1, p. 42]. □

PROPOSITION 5.2. *Let X and Y be Banach manifolds of class C^1, let $h: X \to Y$ be a C^1 mapping, and let $y_0 \in Y$ be a normal value of h. Then there exists an open neighborhood V_0 of y_0 such that $V_0 \subseteq h(X)$ and every point of V_0 is a normal value of h.*

Proof. By hypothesis, there exists x_0 in X such that $h(x_0) = y_0$ and the differential $dh_{x_0}: T_{x_0}X \to T_{y_0}Y$ is split surjective. Using Lemma 5.1, we obtain the existence of an open neighborhood U_0 of x_0 such that the differential $dh_x: T_xX \to T_{h(x)}Y$ is split surjective for every x in U_0. In particular, every point of the set $V_0 = h(U_0)$ is a normal value of h. Moreover, V_0 is open in Y because the mapping $h|_{U_0}$, being a submersion, is an open mapping. □

COROLLARY 5.3. *The set of normal values of a C^1 mapping $h: X \to Y$ is an open subset of Y.*

We now take up the question of the existence of normal values in the context of control systems defined on open subsets of $\mathbb{R}^n$. For the remainder of this section, J denotes an interval, V denotes an open subset of $\mathbb{R}^n$, and $f: J \times V \times \mathbb{R}^m \to \mathbb{R}^n$ denotes a quasi-C^1 mapping, viewed as a control system on V. Let $\mu: \mathscr{D}(f) \to V$ denote the global flow of f and fix a point (t_1, t_0, x_0) in $J \times J \times V$ such that $t_1 > t_0$ and the open subset

$$\mathscr{D}_{(t_1,t_0,x_0)}(f) = \{u \in L_\infty^m(J) | (t_1, t_0, x_0, u) \in \mathscr{D}(f)\}$$

of $L_\infty^m(J)$, is nonempty.

Our study of normal values of the mapping

$$\mu_{(t_1,t_0,x_0)}: \mathscr{D}_{(t_1,t_0,x_0)}(f) \to V$$

will be approached via the technique of linearization of a nonlinear control system along a given response. For this reason, it will be convenient to review a basic fact from linear control theory.

PROPOSITION 5.4. *Let $A: J \to L(\mathbb{R}^n; \mathbb{R}^n)$ and $B: J \to L(\mathbb{R}^m; \mathbb{R}^n)$ be mappings that are measurable and locally bounded and let $\Psi: J \times J \to L(\mathbb{R}^n; \mathbb{R}^n)$ denote the fundamental*

matrix solution of the linear homogeneous differential equation $\dot{x} = A(t)x$. *Then for every* (t_1, t_0, y_0) *in* $J \times J \times \mathbb{R}^n$ *with* $t_1 > t_0$, *the linear control system*

$$\dot{x} = A(t)x + B(t)w \tag{26}$$

(*with controls in* $L_\infty^m(J)$) *is strongly globally controllable from* (t_0, y_0) *at time* t_1 *if and only if for* z *in* $\mathbb{R}^n$ *the relation*

$$B(t)^T \Psi(t_0, t)^T z = 0 \quad \text{a.e. for } t \in [t_0, t_1],$$

implies that $z = 0$, *where the superscript* T *denotes the matrix transpose.*

Proof. See [6, p. 99]. □

Remarks 5.5. (i) Since the criterion in Proposition 5.4 does not involve the point y_0, we can conclude that the linear control system (26) is strongly globally controllable from (t_0, y_0) at time t_1 for some y_0 in $\mathbb{R}^n$ if and only if it is completely controllable on $[t_0, t_1]$ (see § 1).

(ii) If the linear control system (26) is completely controllable on $[t_0, t_1]$, then it is completely controllable on every interval $[t_0, t]$ such that $t \geqq t_1$ and $t \in J$. This is a direct consequence of the preceding remark and Proposition 5.4.

DEFINITION 5.6. Let $f: J \times V \times \mathbb{R}^m \to \mathbb{R}^n$ be a quasi-C^1 control system on V and let $(t_0, x_0, u_0) \in J \times V \times L_\infty^m(J)$. For t in the interval $J(t_0, x_0, u_0)$, the linear control system

$$\dot{x} = D_2 f(t, \mu(t, t_0, x_0, u_0), u_0(t))x + D_3 f(t, \mu(t, t_0, x_0, u_0), u_0(t))w$$

is called the *linear variational control system of* f *along the response* $t \mapsto \mu(t, t_0, x_0, u_0)$.

We will now formulate a necessary and sufficient condition for the derivative $D\mu_{(t_1, t_0, x_0)}(u_0)$ to be a surjective linear mapping, in which case the point $\mu(t_1, t_0, x_0, u_0)$ is a normal value of the mapping $\mu_{(t_1, t_0, x_0)}$.

THEOREM 5.7. *For* u_0 *in* $\mathscr{D}_{(t_1, t_0, x_0)}(f)$, *the derivative*

$$D\mu_{(t_1, t_0, x_0)}(u_0): L_\infty^m(J) \to \mathbb{R}^n$$

is a surjective linear mapping if and only if the linear variational control system of f *along the response* $t \mapsto \mu(t, t_0, x_0, u_0)$ *is completely controllable on* $[t_0, t_1]$.

Proof. By Remark 5.5(i), it suffices to show that $D\mu_{(t_1, t_0, x_0)}(u_0)$ is surjective if and only if the linear variational control system of f along the response $t \mapsto \mu(t, t_0, x_0, u_0)$ is strongly globally controllable from $(t_0, 0) \in J \times \mathbb{R}^n$ at time t_1. For t in $[t_0, t_1]$ and u in $L_\infty^m(J)$, let $t \mapsto \psi(t, t_0, x_0, u_0, u)$ denote the unique absolutely continuous mapping satisfying $\psi(t_0, t_0, x_0, u_0, u) = 0$ and

$$\frac{\partial}{\partial t}\psi(t, t_0, x_0, u_0, u) = D_2 f(t, \mu(t, t_0, x_0, u_0), u_0(t))\psi(t, t_0, x_0, u_0, u) + D_3 f(t, \mu(t, t_0, x_0, u_0), u_0(t))u(t),$$

a.e. for $t \in [t_0, t_1]$. It is evident that the linear variational control system of f along the response $t \mapsto \mu(t, t_0, x_0, u_0)$ is strongly globally controllable from $(t_0, 0)$ at time t_1 if and only if the mapping $u \mapsto \psi(t_1, t_0, x_0, u_0, u)$ is surjective. However, by Corollary 2.13, we have the formula

$$\psi(t_1, t_0, x_0, u_0, u) = D_4\mu(t_1, t_0, x_0, u_0)(u) = D\mu_{(t_1, t_0, x_0)}(u_0)(u).$$

Therefore, the mapping $u \mapsto \psi(t_1, t_0, x_0, u_0, u)$ is surjective if and only if the derivative $D\mu_{(t_1, t_0, x_0)}(u_0)$ is surjective. This completes the proof. □

Theorem 5.7 is of interest because it reduces the determination of the surjectivity of the derivative $D\mu_{(t_1, t_0, x_0)}(u_0)$ to the question of the complete controllability of a linear

control system, and much is known about the latter (see, e.g., [6] and [12]). The major drawback of Theorem 5.7 is that it is not a computable criterion in the sense that one must know the response $t \mapsto \mu(t, t_0, x_0, u_0)$ in order to determine the linear variational control system along that response. However, this does not preclude the use of Theorem 5.7 as a tool in deriving computable criteria for the surjectivity of the derivative of $\mu_{(t_1,t_0,x_0)}$, and we will present two results in this direction.

THEOREM 5.8. *Let $f: J \times V \times \mathbb{R}^m \to \mathbb{R}^n$ be a quasi-C^1 control system on V with global flow $\mu: \mathcal{D}(f) \to V$, and fix a point (t_0, x_0) in $J \times V$. Suppose that for some open neighborhood $J_0 \times V_0$ of (t_0, x_0) in $J \times V$ the partial derivative $D_3 f(t, x, w)$ has rank n for every (t, x, w) in $J_0 \times V_0 \times \mathbb{R}^m$ (in particular, $m \geqq n$). Then, for every t_1 in J such that $t_1 > t_0$ and $\mathcal{D}_{(t_1,t_0,x_0)}(f)$ is nonempty, the derivative $D\mu_{(t_1,t_0,x_0)}(u)$ is a surjective linear mapping for every u in $\mathcal{D}_{(t_1,t_0,x_0)}(f)$.*

Proof. Fix a control u in $\mathcal{D}_{(t_1,t_0,x_0)}(f)$ and let

$$\Psi: J(t_0, x_0, u) \times J(t_0, x_0, u) \to L(\mathbb{R}^n; \mathbb{R}^n)$$

denote the fundamental matrix solution of the linear homogeneous differential equation

$$\dot{x} = D_2 f(t, \mu(t, t_0, x_0, u), u(t))x.$$

For every (t, x, w) in $J_0 \times V_0 \times \mathbb{R}^m$, we have the following implications:

$$\begin{aligned} D_3 f(t, x, w) \text{ has rank } n &\Rightarrow D_3 f(t, x, w): \mathbb{R}^m \to \mathbb{R}^n \text{ is surjective} \\ &\Rightarrow D_3 f(t, x, w)^T: \mathbb{R}^n \to \mathbb{R}^m \text{ is injective.} \end{aligned}$$

To show that $D\mu_{(t_1,t_0,x_0)}(u)$ is surjective, it suffices, by Proposition 5.4 and Theorem 5.7, to show that the relation

$$(*) \qquad D_3 f(t, \mu(t, t_0, x_0, u), u(t))^T \Psi(t_0, t)^T z = 0,$$

for almost every t in $[t_0, t_1]$, implies that $z = 0$. But if relation $(*)$ holds, then there exists $\bar{t}$ in $[t_0, t_1]$ such that $(\bar{t}, \mu(\bar{t}, t_0, x_0, u)) \in J_0 \times V_0$ and

$$D_3 f(\bar{t}, \mu(\bar{t}, t_0, x_0, u), u(\bar{t}))^T \Psi(t_0, \bar{t})^T z = 0.$$

Because $D_3 f(\bar{t}, \mu(\bar{t}, t_0, x_0, u), u(\bar{t}))^T$ is injective and $\Psi(t_0, \bar{t})^T$ is invertible, we conclude that $z = 0$. □

The rank condition on the partial derivative $D_3 f(t, x, w)$ in Theorem 5.8 is quite strong, and this severely limits the applicability of the theorem. Consequently, we will present one final result in this section which ensures the existence of normal values on an open dense subset of the attainable set under a much weaker hypothesis. We preface this result with a remark concerning linear autonomous control systems.

Remark 5.9. Let $A \in L(\mathbb{R}^n; \mathbb{R}^n)$ and $B \in L(\mathbb{R}^m; \mathbb{R}^n)$. Recall that the linear autonomous control system $\dot{x} = Ax + Bw$ is said to be *completely controllable* if it is completely controllable on every compact interval $[t_0, t_1]$ with $t_1 > t_0$. A well-known necessary and sufficient condition for the complete controllability of the system $\dot{x} = Ax + Bw$ (due to Kalman) is that

$$\operatorname{rank} [B, AB, \cdots, A^{n-1}B] = n.$$

THEOREM 5.10. *Let $f: V \times \mathbb{R}^m \to \mathbb{R}^n$ be a C^1 autonomous control system with global flow $\mu: \mathcal{D}(f) \to V$, and let $(x_0, w_0) \in V \times \mathbb{R}^m$ be such that:*

(i) $f(x_0, w_0) = 0$;

(ii) *the linear autonomous control system*

$$\dot{x} = D_1 f(x_0, w_0)x + D_2 f(x_0, w_0)w$$

is completely controllable.

Let t_0, s, and t_1 be real numbers satisfying $t_0 < s < t_1$. Then we have

$$\mathscr{A}_f(t_0, x_0; s) \subseteq \mathscr{A}_f(t_0, x_0; t_1),$$

and every point of $\mathscr{A}_f(t_0, x_0; s)$ is a normal value of the mapping $\mu_{(t_1, t_0, x_0)}$.

Proof. Let y_0 be an arbitrary element of $\mathscr{A}_f(t_0, x_0; s)$, which we fix for the remainder of the proof. Then we have $y_0 = \mu(s, t_0, x_0, u_0)$ for some control u_0 in $\mathscr{D}_{(s,t_0,x_0)}(f)$. Identify w_0 with the constant control $v: \mathbb{R} \to \mathbb{R}^m$ given by $v(t) = w_0$ for every t in $\mathbb{R}$. Condition (i) implies that $\mu(t, t_0, x_0, w_0) = x_0$ for every t in $\mathbb{R}$.

Let u_1 be the control defined by

$$u_1(t) = \begin{cases} w_0 & \text{if } t \in [t_0, t_0 + t_1 - s], \\ u_0(t + s - t_1) & \text{otherwise.} \end{cases}$$

Because f is autonomous and $\mu(t, t_0, x_0, w_0) = x_0$ for all t in $\mathbb{R}$, the response $t \mapsto \mu(t, t_0, x_0, u_1)$ is seen to satisfy

$$\mu(t, t_0, x_0, u_1) = \begin{cases} x_0 & \text{if } t \in [t_0, t_0 + t_1 - s], \\ \mu(t + s - t_1, t_0, x_0, u_0) & \text{if } t \in [t_0 + t_1 - s, t_1], \end{cases}$$

and

$$\mu(t_1, t_0, x_0, u_1) = \mu(s, t_0, x_0, u_0) = y_0.$$

In particular, we have $y_0 \in \mathscr{A}_f(t_0, x_0; t_1)$, and this establishes the inclusion

$$\mathscr{A}_f(t_0, x_0; s) \subseteq \mathscr{A}_f(t_0, x_0; t_1).$$

The linear variational control system of f along the response $t \mapsto \mu(t, t_0, x_0, u_1)$ is given by

$$(*) \qquad \dot{x} = D_1 f(\mu(t, t_0, x_0, u_1), u_1(t))x + D_2 f(\mu(t, t_0, x_0, u_1), u_1(t))w.$$

For t in $[t_0, t_0 + t_1 - s]$ this assumes the form

$$\dot{x} = D_1 f(x_0, w_0)x + D_2 f(x_0, w_0)w.$$

Hence, condition (ii) implies that the linear control system (*) is completely controllable on the interval $[t_0, t_0 + t_1 - s]$. By remark 5.5(ii), the linear control system (*) is also completely controllable on the interval $[t_0, t_1]$. Therefore, we infer from Theorem 5.7 that the point $y_0 = \mu(t_1, t_0, x_0, u_1)$ is a normal value of the mapping $\mu_{(t_1, t_0, x_0)}$. This completes the proof. □

COROLLARY 5.11. *Let t_0, s and t_1 be real numbers satisfying $t_0 < s < t_1$. Then $\mathscr{A}_f(t_0, x_0; s)$ is contained in the interior of $\mathscr{A}_f(t_0, x_0; t_1)$ relative to V.*

Proof. This follows from the theorem and Corollary 5.3. □

COROLLARY 5.12. *For $t_0 < t_1$ the set of normal values of the mapping $\mu_{(t_1, t_0, x_0)}$ is open in V and dense in the attainable set $\mathscr{A}_f(t_0, x_0; t_1)$. In particular, $\mathscr{A}_f(t_0, x_0; t_1)$ is contained in the closure of its interior.*

Proof. The set of normal values of $\mu_{(t_1, t_0, x_0)}$ is open in V by Corollary 5.3, so it suffices to show that this set is dense in $\mathscr{A}_f(t_0, x_0; t_1)$. Let $y_0 = \mu(t_1, t_0, x_0, u_0)$ be an arbitrary point of $\mathscr{A}_f(t_0, x_0; t_1)$ and let V_0 be an arbitrary open neighborhood of y_0 in V. Since μ is continuous, there exists a real number s such that $t_0 < s < t_1$ and

$\mu(s, t_0, x_0, u_0) \in V_0$. By the theorem, $\mu(s, t_0, x_0, u_0)$ is a normal value of the mapping $\mu_{(t_1,t_0,x_0)}$. □

6. Examples.

(i) Let $GL(n, \mathbb{R})$ denote the open subset of invertible elements in $L(\mathbb{R}^n; \mathbb{R}^n)$ and let $A: J \to L(\mathbb{R}^n; \mathbb{R}^n)$ and $B: J \to GL(n, \mathbb{R})$ be mappings that are measurable and locally bounded. Define a control system,

$$f: J \times GL(n, \mathbb{R}) \times L(\mathbb{R}^n; \mathbb{R}^n) \to L(\mathbb{R}^n; \mathbb{R}^n),$$

by

$$f(t, Z, W) = (A(t) + B(t)W)Z.$$

For every (t_0, Z_0, W_0) in $J \times GL(n, \mathbb{R}) \times L(\mathbb{R}^n; \mathbb{R}^n)$ the partial derivative $D_3 f(t_0, Z_0, W_0)$ is given by the formula

$$D_3 f(t_0, Z_0, W_0)W = B(t_0)WZ_0.$$

It follows that $D_3 f(t_0, Z_0, W_0)$ is a linear isomorphism of $L(\mathbb{R}^n; \mathbb{R}^n)$ onto itself, since the inverse of $D_3 f(t_0, Z_0, W_0)$ is readily seen to be the mapping $W \mapsto B(t_0)^{-1} W Z_0^{-1}$. Hence, for every (t_1, t_0, Z_0) in $J \times J \times GL(n, \mathbb{R})$ with $t_1 > t_0$, Theorem 5.8 implies that every point of the attainable set $\mathscr{A}_f(t_0, Z_0; t_1)$ is a normal value of the mapping $\mu_{(t_1,t_0,Z_0)}$. We can therefore apply Theorem 4.9 to compact subsets of $\mathscr{A}_f(t_0, Z_0; t_1)$. Furthermore, by Corollary 5.3, the attainable set $\mathscr{A}_f(t_0, Z_0; t_1)$ is an open subset of $L(\mathbb{R}^n; \mathbb{R}^n)$ (a more detailed analysis actually shows that $\mathscr{A}_f(t_0, Z_0; t_1)$ is the connected component of $GL(n, \mathbb{R})$ containing Z_0).

Control systems of this type can arise in the study of linear optimal systems with quadratic cost criteria (see, e.g., [15, Chap. 3]). For physical reasons, a control $U: J \to L(\mathbb{R}^n; \mathbb{R}^n)$ is sometimes called a gain matrix.

(ii) Consider the control system $f: \mathbb{R}^2 \setminus \{0\} \times \mathbb{R}^2 \to \mathbb{R}^2$ defined by

$$f(x_1, x_2, w_1, w_2) = (x_1 w_1 - x_2 w_2, x_1 w_2 + x_2 w_1),$$

or, in more traditional notation,

$$\dot{x}_1 = x_1 u_1(t) - x_2 u_2(t),$$
$$\dot{x}_2 = x_1 u_2(t) + x_2 u_1(t).$$

A routine verification shows that the global flow μ of f is explicitly given by the formula

$$\mu(t, s, x_1, x_2, u_1, u_2) = \exp\left(\int_s^t u_1(\tau)\, d\tau\right) \begin{pmatrix} \cos \int_s^t u_2(\tau)\, d\tau & -\sin \int_s^t u_2(\tau)\, d\tau \\ \sin \int_s^t u_2(\tau)\, d\tau & \cos \int_s^t u_2(\tau)\, d\tau \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

From this formula we see that, for every compact interval $[t_0, t_1]$ with $t_1 > t_0$ and every (x_1, x_2) in $\mathbb{R}^2 \setminus \{0\}$, the control system f is strongly controllable from $(t_0, (x_1, x_2))$ to $\mathbb{R}^2 \setminus \{0\}$ at time t_1. Moreover, if $D_w f$ denotes the partial derivative of f with respect to the pair of variables (w_1, w_2), then we have

$$D_w f(x_1, x_2, w_1, w_2) = \begin{pmatrix} x_1 & -x_2 \\ x_2 & x_1 \end{pmatrix},$$

which is an invertible 2×2 matrix for every (x_1, x_2) in $\mathbb{R}^2 \setminus \{0\}$. By Theorem 5.8, the

derivative

$$D\mu_{(t_1,t_0,(x_1,x_2))}(u_1, u_2)$$

is a surjective linear mapping for every choice of the control (u_1, u_2). Hence, Theorem 4.9 can be applied to compact subsets of $\mathbb{R}^2\backslash\{0\}$.

Acknowledgment. The results of this paper are a portion of the author's doctoral dissertation written under the supervision of Professor Felix Albrecht at the University of Illinois at Urbana-Champaign.

REFERENCES

[1] R. ABRAHAM AND J. ROBBIN, *Transversal Mappings and Flows*, W. A. Benjamin, New York, 1967.
[2] F. ALBRECHT, *Topics in Control Theory*, Springer-Verlag, New York, 1968.
[3] G. ARONSSON, *Global controllability and bang-bang steering of certain nonlinear systems*, this Journal, 11 (1973), pp. 607–619.
[4] R. W. BROCKETT, *System theory on group manifolds and coset spaces*, this Journal, 10 (1972), pp. 265–284.
[5] P. BRUNOVSKY AND C. LOBRY, *Contrôlabilité bang-bang, contrôlabilité différentiable, et perturbation des systèmes non linéares*, Ann. Mat. Pura Appl., 105 Ser. 4 (1975), pp. 95–119.
[6] R. CONTI, *Linear Differential Equations and Control*, Academic Press, New York, 1976.
[7] J. P. DAUER, *A note on bounded perturbations of controllable systems*, J. Math. Anal. Appl., 42 (1973), pp. 221–225.
[8] ———, *Bounded perturbations of controllable systems, II*, J. Math. Anal. Appl., 48 (1974), pp. 61–69.
[9] ———, *Nonlinear perturbations of quasi-linear control systems*, J. Math. Anal. Appl., 54 (1976), pp. 717–725.
[10] ———, *Controllability of perturbed nonlinear systems*, Lincei-Rend. Sc. fis. mat. e nat., 63 (1977), pp. 345–350.
[11] K. A. GRASSE, *Controllability and Accessibility in Nonlinear Control Systems*, Thesis, University of Illinois, 1979.
[12] J. HERMES AND J. P. LASALLE, *Functional Analysis and Time Optimal Control*, Academic Press, New York, 1969.
[13] S. LANG, *Analysis II*, Addison-Wesley, Reading, MA, 1969.
[14] ———, *Differential Manifolds*, Addison-Wesley, Reading, MA, 1972.
[15] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
[16] D. L. LUKES, *Global controllability of nonlinear systems*, this Journal, 10 (1972), pp. 112–126; Erratum, 11 (1973), p. 186.
[17] H. SUSSMANN, *Some properties of vector field systems that are not altered by small perturbations*, J. Differential Equations, 20 (1976), pp. 292–315.
[18] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 2, March 1981

0363-0129/81/1902-0004/$01.00/0

ON THE ADJOINT PROCESS FOR OPTIMAL CONTROL OF DIFFUSION PROCESSES*

U. G. HAUSSMANN†

Abstract. For the Markovian control problem

$$\min_u E\left\{\int_0^T l(t, x, u)\,dt + c(x(T))\right\},$$

$$dx = f(t, x, u)\,dt + \sigma(t, x)\,dw,$$

it is shown that the adjoint process appearing in the maximum principle has the form

$$p(t, x) = -E_{tx}\left\{\frac{\partial c}{\partial x}(x(T))\Phi(T, t) + \int_t^T \frac{\partial l}{\partial x}(s, x(s), \hat{u}(s, x(s))\Phi(s, t)\,ds\right\},$$

where $\hat{u}$ is the optimal feedback control, E_{tx} denotes conditional expectation, and Φ is the fundamental matrix solution of

$$dy = \frac{\partial f}{\partial x}(t, x(t), \hat{u}(t, x(t)))y\,dt + \sum_k \frac{\partial \sigma^k}{\partial x}(t, x(t))y\,dw_k.$$

Here, σ^k is the kth column of σ and w_k is the kth component of the Brownian motion w. It is also shown that $p(t, x(t))$ satisfies (* denotes transpose)

$$dp^* = \left\{-\frac{\partial f^*}{\partial x}(t, x(t), \hat{u}(t, x(t)))p^* + \sum_k \frac{\partial \sigma^k}{\partial x}(t, x(t))V_{xx}(t, x(t))\sigma^k(t, x(t)) + \frac{\partial l^*}{\partial x}(t, x(t), \hat{u}(t, x(t)))\right\}dt$$
$$- V_{xx}(t, x(t))\sigma(t, x(t))\,dw,$$

$$p(T) = -\frac{\partial c}{\partial x}(x, (T)),$$

where V_{xx} is the Hessian of the value function V.

1. Introduction.

Consider the completely observable stochastic control problem

$$\min_u E\left\{\int_0^\tau l(t, x, u)\,dt + c_0(\tau, x(\tau))\right\}, \tag{1.1}$$

subject to

$$dx = f(t, x, u)\,dt + \sigma(t, x)\,dw, \qquad x(0) = x_0, \tag{1.2}$$

$$Ec_i(\tau, x(\tau)) = 0, \qquad i = 1, \cdots, k_0, \tag{1.3}$$

where E stands for expectation, τ is the first exit time of the process $\{t, x(t): t \geqq 0\}$ from an open set $(0, T) \times G \subset \mathbb{R}^{d+1}$, $(d+1)$-dimensional Euclidean space, $x_0 \in G$, w is a standard Brownian motion on $(\Omega, \mathcal{F}, P)$ with values in $\mathbb{R}^d$, l, f are defined on $[0, T] \times \mathbb{R}^d \times U$, where $U \subset \mathbb{R}^m$ is the set of control points, and where c maps $[0, T] \times \mathbb{R}^d$ into $\mathbb{R}^{k_0+1}$. The minimization in (1.1) is taken over the *admissible* controls $\mathcal{U}$, i.e., all Borel functions mapping $[0, T] \times G$ into U.

To be more precise, with σ and ϕ Lipschitz continuous in x, σ bounded, and $|\phi(t, x)|^2 \leqq K(1 + |x|^2)$, let x be the unique solution of

$$dx = \phi(t, x)\,dt + \sigma(t, x)\,dw, \qquad x(0) = x_0. \tag{1.4}$$

* Received by the editors September 19, 1979, and in final revised form June 16, 1980. This research was supported by the Canadian Natural Sciences and Engineering Research Council under Grant number A8051.

† Department of Mathematics, 2075 Wesbrook Mall, The University of British Columbia, Vancouver, B.C., Canada V6T 1W5.

If we assume $f(t, x, u) = \phi(t, x) + \sigma(t, x)\theta(t, x, u)$, $|\theta(t, x, u)|^2 \leqq K(1 + |x|^2)$, it follows from Girsanov's theorem that for any $u \in \mathcal{U}$ there is a probability measure P^u and a process $w^u(\cdot)$ which is a Brownian motion under P^u, such that

$$dx = f(t, x, u)\, dt + \sigma(t, x)\, dw^u. \tag{1.2}'$$

For $u \in \mathcal{U}$, we define

$$J(t, x, u) = E^u_{tx}\left\{\int_t^\tau l(s, x(s), u(s, x(s)))\, ds + c_0(\tau, x(\tau))\right\},$$

where E^u_{tx} is the conditional expectation under P^u, given $x(t) = x$. Now the problem (1.1)–(1.3) can be reformulated precisely as

$$\min_{u \in \mathcal{U}} J(0, x_0, u), \tag{1.1}'$$

subject to

$$E^u_{0x_0} c_i(\tau, x(\tau)) = 0, \qquad i = 1, \cdots, k_0. \tag{1.3}'$$

This problem can be attacked either by using the maximum principle developed by the author in [7], [8], or (in the case $k_0 = 0$) using dynamic programming as presented by Bismut [1] or Fleming and Rishel [5], to conclude that any optimal control must satisfy

$$\hat{u}(t, x) \in \arg\max_{v \in U} \tilde{H}(t, x, v, p(t, x)) \quad \text{a.e.}, \tag{1.5}$$

where $\tilde{H}(t, x, v, p) = pf(t, x, v) - l(t, x, v)$ and where p is the adjoint process (in row vector form). Let us write $\hat{P}$ for $P^{\hat{u}}$ and $\hat{E}$ for $E^{\hat{u}}$, $\hat{w}$ for $w^{\hat{u}}$.

In § 2, we shall use the maximum principle to reduce the problem with $k_0 \neq 0$ to one where $k_0 = 0$, i.e., $c = c_0$, so that dynamic programming applies. Let us now assume $k_0 = 0$.

In order to apply the maximum principle, one must have at hand a reasonable characterization of p. If we define the value function

$$V(t, x) \equiv \inf_{u \in \mathcal{U}} J(t, x, u),$$

then by the principle of optimality $V(t, x) = J(t, x, \hat{u})$. Assuming smoothness, then from [5, Chapt. VI], it follows that

$$dV(t, x(t)) = -l(t, x(t), \hat{u}(t, x(t)))\, dt + V_x(t, x(t))\sigma(t, x(t))\, d\hat{w}.$$

On the other hand, from the maximum principle, if

$$M(t) = \hat{E}\left\{\int_0^\tau l(s, x, \hat{u}(s, x))\, ds + c(\tau, x(\tau)) \,|\, x(\rho), \rho \leqq t\right\}, \tag{1.6}$$

then

$$dM = \chi(t, \omega)\, d\hat{w},$$

where $\hat{E}(\chi(t, \omega)|x(t) = x\} = -p(t, x)\sigma(t, x)$ so that $p = -V_x \equiv -(\partial V/\partial x_1, \partial V/\partial x_2, \cdots, \partial V/\partial x_d)$. These results are not satisfactory because one would like to characterize p independently of V since it is often difficult to solve for V.

For the case of nonanticipative controls $u(t, \omega)$ and $G = \mathbb{R}^d$, Kushner [12] derived a maximum principle with adjoint process $\tilde{p}$ given by ($\hat{x}$ is the optimal Ito solution, $\hat{u}$ the

corresponding control)

$$p(t) = -E\Big\{ c_x(T, \hat{x}(T))\Phi(T, t) + \int_t^T l_x(t, \hat{x}(s), \hat{u}(s))\Phi(s, t)\, ds \,|\, \mathcal{G}_t \Big\}, \tag{1.7}$$

where $\Phi(t, t_0)$ is the fundamental matrix solution of

$$dy = f_x(t, \hat{x}(t), \hat{u}(t))y\, dt + \sum_{k=1}^{d} \sigma_x^k(t, \hat{x}(t))y\, d\hat{w}_k, \tag{1.8}$$

with $\Phi(t_0, t_0) = I$, the identity. Here, σ^k is the kth column of σ and $\{\mathcal{G}_t\}$ is the σ-algebra generated by $w(s)$, $s \leqq t$. This work does not extend to the closed loop situation, but the author has shown [8], [9], [10] that if $\hat{u}$ is smooth, then p (in the closed loop case with $G = \mathbb{R}^d$) is given by

$$p(t, x) = -\hat{E}_{tx}\Big\{ c_x(T, x(T))\tilde{\Phi}(T, t) + \int_t^T (l_x + l_u\hat{u}_x)\tilde{\Phi}(s, t)\, ds \Big\}, \tag{1.9}$$

where $\tilde{\Phi}$ is the fundamental matrix solution of

$$dy = (f_x + f_u\hat{u}_x)y\, dt + \sum_k \sigma_x^k y\, d\hat{w}_k. \tag{1.10}$$

In this work, we delete the smoothness requirement on $\hat{u}$, and we show that we can take

$$p(t, x) = -\hat{E}_{tx}\Big\{ c_x(T, x(T))\Phi(T, t) + \int_t^T l_x(s, x(s), \hat{u}(s, x(s)))\Phi(s, t)\, ds \Big\}, \tag{1.11}$$

where Φ is defined by (1.8), but with $\hat{u}(t)$ replaced by the feedback control $\hat{u}(t, x(t))$.

Long ago, Fleming [3] noted the difficulty with the smoothness assumption inherent in (1.9), (1.10), and showed that if σ is constant, if $U = \mathbb{R}^m$, and if $\hat{u}(t, x)$ is smooth in x, then one can replace (1.9), (1.10) by (1.8), (1.11). Our approach for the case $\tau = T$ is still based on partial differential equations and dynamic programming, but assumes only minimal restriction as required for (1.8), (1.11) to make sense. We smooth the problem and show that the limit of the value function of these smooth problems is a generalized solution of the Hamilton-Jacobi-Bellman equation (H-J-B equation). Hence, for the unconstrained problem, our version of dynamic programming gives an existence result under hypotheses weaker than those in [5] but stronger than those in [1]. However, unlike [1], we can show that $p = -V_x$.

After the reduction in § 2 to the case $k_0 = 0$, we develop the dynamic programming results in § 3. In § 4, we then establish (1.11) for the unbounded case ($\tau = T$), and in § 5, we treat the case of G bounded when the H-J-B equation is strongly parabolic. Here, we use a general approach suggested by Davis [2] to derive a representation for V_x.

We remark, as pointed out by Bismut, that p as defined by (1.11) does satisfy an equation. If $\Psi(t, t_0) \equiv \Phi(t, t_0)^{-1}$, if $\bar{p}(t) \equiv -c_x(T, x(T))\Phi(T, t) - \int_t^T l_x(s, x, \hat{u})\Phi(s, t)\, ds$, and if $\hat{E}\{\bar{p}(0) \,|\, x(s), s \leqq t\} = \hat{E}\bar{p}(0) + \sum_{k=1}^{d} \int_0^t \psi^k\, d\hat{w}_k$, then

$$\begin{aligned} p(t, x) &= \hat{E}_{tx}\{\bar{p}(t)\} = \hat{E}\{\bar{p}(t) \,|\, x(s), s \leqq t\} \\ &= \hat{E}\Big\{ \bar{p}(T)\Phi(T, 0) - \int_t^T l_x(s)\Phi(s, 0)\, ds \,|\, x(s), s \leqq t \Big\}\Psi(t, 0) \\ &= \hat{E}\{\bar{p}(0) \,|\, x(s), s \leqq t\}\Psi(t, 0) + \int_0^t l_x(s)\Phi(s, 0)\, ds\, \Psi(t, 0). \end{aligned}$$

Since

$$d\Psi = -\Psi\Big[f_x - \sum_k \sigma_x^k \sigma_x^k\Big]\,dt - \Psi \sum_k \sigma_x^k\, d\hat{w}_k,$$

then

$$dp = \Big\{-p\Big[f_x - \sum_k (\sigma_x^k)^2\Big] + l_x - \sum_k \psi^k \Psi(t,0)\sigma_x^k\Big\}\,dt + \sum_k \{-p\sigma_x^k + \psi^k \Psi(t,0)\}\,d\hat{w}_k. \tag{1.12}$$

We shall identify the unknown functions ψ^k in terms of V at the end of § 5. The results of this work have been applied to solve some simple control problems (cf. [11]).

Finally, we point out that the constraints (1.3) could just as well be taken as inequality constraints or as constraints in integral form rather than terminal constraints. The results do not change.

2. Reduction of the problem. We shall first reduce the problem to the case when l is independent of u, since the maximum principle [7] requires the cost functional to depend only on x, not u. Then, using the Lagrange multiplier feature of the maximum principle, we shall observe that an optimal $\hat{u}$ solves a new unconstrained problem.

For the purposes of this section, we need only make the assumptions A_1–A_8 of [7]; however, instead we make the following stronger assumptions since they will be needed later.

(H_1) $\sigma(\cdot,\cdot)$, is bounded on $Q = \overline{(0,T)\times G}$.

(H_2) $f(t,x,u) = \phi(t,x) + \sigma(t,x)\theta(t,s,u)$, $l(t,x,u) = l_1(t,x) + l_2(t,x,u)$ where ϕ, θ, l_2 satisfy a linear growth condition in x, i.e.,

$$|\phi(t,x)|^2 + |\theta(t,x,u)|^2 + |l_2(t,x,u)|^2 \leqq K(1+|x|^2), \qquad 0 \leqq t \leqq T, \quad u \in U,$$

and where θ, l_2 are continuous in u for each $(t,x) \in \bar{Q}$.

(H_3) $|c(t,x)|^2 + |l_1(t,x)|^2 \leqq K(1+|x|^{2p_0})$ for some $p_0 < \infty$.

In addition all functions are Borel measurable.

Let us now consider the problem

$$\min_{u \in \mathcal{U}} \tilde{J}(0, \tilde{x}_0, u), \tag{2.1}$$

subject to

$$\tilde{E}^{*}\tilde{c}_i(\tau, \tilde{x}(\tau)) = 0, \qquad i = 1, \cdots, k_0, \tag{2.2}$$

where (* denotes transpose) $\tilde{x}^* = (x^*, \tilde{x}_{d+1})$, $\tilde{x}_0^* = (x_0^*, 0)$, $\tilde{J}(t,\tilde{x},u) = \tilde{E}^u_{t,\tilde{x}}\{\tilde{c}_0(\tau,\tilde{x}(\tau)) + \int_t^\tau \tilde{l}(t',\tilde{x}(t'))\,dt'\}$, $\tilde{c}_i(t,\tilde{x}) = c_i(t,x)$ if $i > 0$, $\tilde{c}_i(t,\tilde{x}) = c_0(t,x) + \tilde{x}_{d+1}$ if $i = 0$, $\tilde{l}(t,\tilde{x}) = l_1(t,x)$, and where $\tilde{E}^u_{t,\tilde{x}}$ is conditional expectation under $\tilde{P}^u$, and $\tilde{x}$ satisfies

$$d\tilde{x} = \tilde{f}(t,\tilde{x},u)\,dt + \tilde{\sigma}(t,\tilde{x})\,d\tilde{w}, \tag{2.3}$$

with $\tilde{w}^* = (w^*, w_0)$, w_0 a scalar Brownian motion independent of w, (Ω is enlarged to $\tilde{\Omega}$), $\tilde{f}(t,\tilde{x},u)^* = (f(t,x,u)^*, l_2(t,x,u))$, $\tilde{\sigma}(t,x) = \begin{pmatrix} \sigma(t,x) & 0 \\ 0 & 1 \end{pmatrix}$. τ is as before but u, the admissible controls, are still measurable functions of $x^* = (\tilde{x}_1, \cdots, \tilde{x}_d)$, not of $\tilde{x}_{d+1}$; i.e., the problem is partially observable with $u = u(t,x)$. It is well known that (2.1), (2.2) is equivalent to (1.1)′, (1.3)′; i.e., both problems have the same optimal controls and values (cf. [19, p. 452]).

According to the maximum principle [7], [8] (a slight variation of the proof given in [7] allows the removal of the hypothesis that f be continuous in u *uniformly* in t, and

extends the result to the Markovian case), if $\hat{u}$ is optimal, then there is a Lebesgue null set N, and a vector $\alpha \in \mathbb{R}^{k_0+1}$ such that for $t \notin N$ ($\hat{\tilde{E}}$ is expectation with respect to $\hat{\tilde{P}}$, the optimal measure on $\tilde{\Omega}$)

$$\hat{u}(t,x) \in \arg\max_{v \in U} \hat{\tilde{E}}\{\alpha^* \tilde{\chi}(t)\tilde{\sigma}(t,\tilde{x})^{-1}\tilde{f}(t,\tilde{x},v) \,|\, x(t)=x\} \quad \text{with probability 1,} \tag{2.4}$$

where we use the pseudoinverse if a matrix is not invertible, and

$$\tilde{L} \equiv \tilde{c}(\tau,\tilde{x}(\tau)) + \int_0^\tau \tilde{l}(t,\tilde{x}(t))\,dt\,e_0 = \hat{\tilde{E}}\tilde{L} + \int_0^T \tilde{\chi}(t)\,d\tilde{w}(t),$$

$$e_0^* = (1,0,\cdots,0).$$

But if [cf. the martingale representation theorem]

$$L \equiv c(\tau,x(\tau)) + \int_0^\tau l\,dt\,e_0 = \hat{E}L + \int_0^T \chi(t)\,d\hat{w}(t), \tag{2.5}$$

then $\tilde{L} = L + w_0(\tau)e_0$ and for $t \leqq \tau$,

$$\tilde{\chi}(t) = (\chi(t), e_0), \tag{2.6}$$

because $\tilde{c}_0 = \tilde{x}_{d+1} + c_0$, and $d\tilde{x}_{d+1} = l_2\,dt + dw_0$. But (2.6) implies that $\tilde{\chi}(t)$ is $\{x(s): s \leqq t\}$-measurable. Hence, (2.4) becomes, for $t \leqq \tau$, $\alpha^* = (\alpha_0, \alpha_1, \cdots, \alpha_{k_0})$,

$$\hat{u}(t,x) \in \arg\max_{v \in U} \{\alpha^* \bar{\chi}(t,x)\sigma(t,x)^{-1}f(t,x,v) + \alpha_0 l(t,x,v)\}, \tag{2.7}$$

where $\bar{\chi}(t,x) = \hat{E}\{\chi(t) \,|\, x(t)=x\}$ with χ defined by (2.5). Thus, we are back to a completely observable d-dimensional problem with $p = \alpha^*\bar{\chi}\sigma^{-1}$. From (2.5) and the Markov property of $x(\cdot)$, it follows that

$$dV(t,x(t)) = -l(t,x(t),u(t,x(t)))\,dt + \chi(t)\,d\hat{w},$$

so that $\chi(t)$ is $x(t)$-measurable; i.e., $\chi(t,\omega) = \bar{\chi}(t,x(t,\omega))$ with probability 1.

We shall now exploit the Lagrange multiplier feature of the maximum principle. This is best done in some generality, so let us consider a general problem. Define $\mathcal{U}$, x, P^u as in § 1 so that (1.4) and (1.2)′ hold. Let L_0, L_1, L_2 be $\mathbb{R}$, $\mathbb{R}^{k_1}$, $\mathbb{R}^{k_2}$-valued functions defined on the set of trajectories of x, i.e., the space of continuous functions, and assume that

$$|L(x)| \leqq K(1 + \sup\{|x(t)|^{p_0}: 0 \leqq t \leqq T\}),$$

where $L^* = (L_0, L_1^*, L_2^*)$ and K, p_0 are finite constants. Now the problem is

$$\min\{E^u L_0(x): E^u L_1(x) \leqq 0,\ E^u L_2(x) = 0,\ u \in \mathcal{U}\}. \tag{2.8}$$

Let us first show that, subject to constraint qualifications, the maximum principle conditions are sufficient for optimality.

LEMMA 2.1. *If there exist $\hat{u} \in \mathcal{U}$, $\alpha \neq 0$, with $\alpha_0 \leqq 0$, $\alpha_1 \leqq 0$, such that $\hat{E}L_1 \leqq 0$, $\alpha_1 \cdot \hat{E}L_1 = 0$, $\hat{E}L_2 = 0$, $\alpha^*\chi\sigma^{-1}f^u \leqq \alpha^*\chi\sigma^{-1}f^{\hat{u}}$ a.e. $(dt \times d\hat{P})$ for all $u \in U$ where χ is defined by*

$$L = \hat{E}L + \int_0^T \chi\,d\hat{w},$$

then $\hat{u}$ is optimal provided the following constraint qualifications hold:

i) *If $L_2 \not\equiv 0$, then $\{E^u L_2: u \in \mathcal{U}\}$ contains a neighborhood of 0.*

ii) *If $L_1 \not\equiv 0$, there exists $u \in \mathcal{U}$ such that $E^u L_2 = 0$ and $E^u L_1$ lies in the open negative orthant.*

Proof. The hypotheses imply that $\alpha \cdot E^u L \leqq \alpha \cdot \hat{E}L = \alpha_0 \hat{E} L_0$ (cf. [7, (3.4)]). Since $\alpha \cdot E^u L = \alpha_0 E^u L_0 + \alpha_1 \cdot E^u L_1 + \alpha_2 \cdot E^u L_2$, then

$$\alpha_0 E^u L_0 \leqq \alpha_0 \hat{E} L_0;$$

i.e., $\hat{u}$ is optimal if $\alpha_0 \neq 0$. But if $\alpha_0 = 0$, then (ii) implies that $\alpha_1 = 0$ and now (i) implies that $\alpha_2 = 0$, contradicting $\alpha \neq 0$. Thus, the constraint qualifications eliminate the abnormal case $\alpha_0 = 0$, and the proof is complete. Note if $L_i \equiv 0$ then α_i, $i = 1, 2$ should be taken as 0 in the statement of the lemma. This fact yields

COROLLARY 2.2. *For the unconstrained completely observable problem (i.e.,* $\min \{E^u L_0(x): u \in \mathcal{U}\}$*), the maximum principle is sufficient for optimality.*

COROLLARY 2.3. *If* $\hat{u}$, α *are as in the lemma and* $\hat{u}$, u' *optimal, then* $\alpha^* \chi' \sigma^{-1} f^u \leqq \alpha^* \chi' \sigma^{-1} f^{u'}$ *a.e. for all* $u \in U$, *where* χ' *is given by*

$$L = E'L + \int_0^T \chi' \, dw'. \tag{2.9}$$

Proof. The corollary says that the multiplier α' corresponding to u' can be taken as α. As in the lemma,

$$\alpha \cdot E^u L \leqq \alpha \cdot \hat{E}L = \alpha_0 \hat{E} L_0 = \alpha_0 E' L_0 \leqq \alpha \cdot E'L, \tag{2.10}$$

the last equality follows since $\hat{u}$, u' are both optimal, hence must have the same cost. Now (2.9) yields

$$E^u(\alpha \cdot L) = E'(\alpha \cdot L) + E' \int_0^T \beta^u \alpha^* \chi' \sigma^{-1} [f^u - f^{u'}] \, dt, \tag{2.11}$$

where β^u is the Radon-Nikodým derivative of P^u with respect to P'. Then (2.10), (2.11) and the maximum principle (cf. [7, (3.5)]) imply the result.

Suppose now that $\hat{u}$ is optimal for (2.8) and that α is given by the maximum principle. Then (2.10) implies that $\hat{u}$ is optimal for the unconstrained problem [$\bar{\alpha} = -\alpha$],

$$\min \{E^u \bar{\alpha}^* L(x): u \in \mathcal{U}\}, \tag{2.12}$$

and also that if u' is also optimal for (2.8) then u' also solves (2.12). There is, however, one further point to consider. For (2.8), the adjoint process is defined by

$$\begin{aligned} p(t, x) &= \alpha^* \hat{E}\{\chi(t) \,|\, x(t) = x\} \sigma(t, x)^{-1} \\ &= -\hat{E}\{\bar{\alpha}^* \chi(t) \,|\, x(t) = x\} \sigma(t, x)^{-1}, \end{aligned}$$

so it is also the adjoint for (2.12). It is this latter which we shall actually compute. But suppose in (2.12) we use u' as an optimal solution, to obtain $p'(t, x) = -E'\{\bar{\alpha}^* \chi'(t) \,|\, x(t) = x\} \sigma(t, x)^{-1}$. Is $p' = p$?

With $u = u'$ in (2.10), we have $\alpha \cdot E'L = \alpha \cdot \hat{E}L$ and hence,

$$\begin{aligned} \alpha \cdot L &= \hat{E}(\alpha \cdot L) + \int_0^T \alpha \cdot \chi \, d\hat{w} \\ &= E'(\alpha \cdot L) + \int_0^T \alpha \cdot \chi' \, dw' \\ &= E'(\alpha \cdot L) + \int_0^T \alpha \cdot \chi' \, d\hat{w} - \int_0^T \alpha \cdot \chi' \sigma^{-1} (f' - \hat{f}) \, dt, \end{aligned}$$

or

$$\int_0^T \alpha \cdot (\chi - \chi')\, d\hat{w} + \int_0^T \alpha \cdot \chi' \sigma^{-1}(f' - \hat{f})\, dt = 0.$$

The nonnegativity of the second integral (Corollary 2.3) implies that $\alpha \cdot \chi = \alpha \cdot \chi'$ a.e. $(dt \times d\hat{P})$. Hence, two different optimal controls give rise to the *same* adjoint process.

If we now apply these results to our original problem (1.1)′, (1.3)′, then we can conclude: if u is optimal, then it is optimal [with $c^* = (c_0, c_1, \cdots, c_{k_0})$] for

$$\min_{u \in \mathcal{U}} E^u_{0x_0}\left\{-\alpha^* c(\tau, x(\tau)) - \alpha_0 \int_0^\tau l(s, x(s), u(s, x(s)))\, ds\right\}. \tag{2.13}$$

This follows from (2.7) and Corollary 2.2. Moreover, the adjoint process is independent of which optimal control is chosen.

3. The value function. Let $G = \mathbb{R}^d$ so $\tau = T$, write $c(x)$ for $c(T, x)$, and let $k_0 = 0$, i.e., no constraints. We assume that

$$f(t, x, u) = \phi(t, x) + \begin{pmatrix} 0 \\ g(t, x, u) \end{pmatrix},$$

$$\sigma(t, x) = \begin{pmatrix} 0 & 0 \\ 0 & \bar{\sigma}(t, x) \end{pmatrix},$$

where g has $d - \nu$ components and $\bar{\sigma}$ is a $(d-\nu)\times(d-\nu)$ matrix. We make the following assumptions:

(A_1) U is compact.

(A_2) $\bar{\sigma}, \bar{\sigma}^{-1}$ are bounded on $\bar{Q} = [0, T] \times \mathbb{R}^d$;
$\bar{\sigma}$ is continuous on $\bar{Q}$;
$\bar{\sigma}_x, \phi_x, g_x$ exist for each (t, u) and are bounded on $\bar{Q} \times U$;

$$|f(t, x, u)|^2 \leqq K(1 + |x|^2);$$

g is continuous on $\bar{Q} \times U$.

(A_3) $l_2(t, x, u)$ is continuous and $(l_1)_x$, $(l_2)_x$, c_x exist;

$$|l_1(t, x)|^2 + |l_2(t, x, u)|^2 + |(l_1)_x(t, x)|^2 + |(l_2)_x(t, x, u)|^2 + |c(x)|^2 + |c_x(x)|^2 \leqq K(1 + |x|^{2q_0}), \qquad q_0 < \infty.$$

(A_4) The system (1.4) has a transition density $\tilde{p}(s, x, t, y)$ such that for some $\mu > 1$,

$$\int_{s'}^T \int_{\mathbb{R}^d} \tilde{p}(s, x; t, y)^\mu\, dy\, dt < \infty,$$

for all $s' > s$.

All functions are assumed Borel measurable and all x derivatives are assumed continuous in x.

Observe that A_4 is satisfied if $\phi(t, x) = A(t)x$, $\sigma(t, x) = \sigma(t)$ and (A, σ) is completely controllable, [5, Thm. 9.1], or if $\sigma = \bar{\sigma}$ and σ, ϕ smooth, [6, Chapt. 6, Thms. 4.5, 5.4], or if $\sigma = \bar{\sigma}$ and σ is uniformly continuous, ϕ bounded, [14, Thm. 8.1].

In [15], [16] Fleming proves that the H-J-B equation has a generalized solution for the cases $G = \mathbb{R}^n$, G bounded, respectively, under the added hypotheses that f, l are bounded and continuous (although the x derivatives need not be continuous in x) and $\sigma(t, x) = \sigma(t, x_{\nu+1}, \cdots, x_d)$. Then in [17], Rishel shows that this generalized solution is the value function in the bounded case. We use a method developed in [5] under much

stronger smoothness assumptions to show that the value functional is a generalized solution of the H-J-B equation. This result and the approximation used in the proof will then be used in the next section to relate p to V_x.

We now define the *nonanticipative controls* on $[s, T]$, i.e., $\tilde{\mathcal{U}}(s)$. u is in $\tilde{\mathcal{U}}(s)$ if u is a U-valued stochastic process defined on a probability space $(\Omega^u, \mathcal{F}^u, P^u)$, adapted to a family of increasing σ-algebras $\{\mathcal{F}_t^u\}$, such that there exists a Brownian motion $(w_t^u, \mathcal{F}_t^u)$. Hence, if $u \in \tilde{\mathcal{U}}(s)$ then

$$dx = f(t, x, u(t, \omega))\, dt + \sigma(t, x)\, dw^u, \qquad x(s) = x_s, \tag{3.1}$$

has a unique solution x^u on $[s, T]$, because of A_2. For $u \in \tilde{\mathcal{U}}(t)$, we set

$$\tilde{J}(t, x, u) = E_{t,x}^u\left\{c(x^u(T)) + \int_t^T l(s, x^u(s), u(s))\, ds\right\},$$

$$\tilde{V}(t, x) = \inf\{\tilde{J}(t, x, u) : u \in \tilde{\mathcal{U}}(t)\}.$$

We shall need the following result. Let y^1, y^2 be two processes on $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$ satisfying

$$dy^i = a^i(t, y^i, \omega)\, dt + b^i(t, y^i)\, d\tilde{w}, \qquad y^i(s) = y^i,$$

and let

$$I^i(s, y^i) = \tilde{E}\{\gamma^i(y^i(T)) + \int_s^T \delta^i(t, y^i(t), \omega)\, dt\},$$

where $|a^i(t, y, \omega)|^2 + |b^i(t, y, \omega)|^2 \leqq K(1 + |y|^2)$, $|a_y^i| + |b_y^i| \leqq K$, and $|\gamma_y^1(y)|^2 + |\delta_y^1(t, y, \omega)|^2 \leqq K(1 + |y|^{2q_0})$.

LEMMA 3.1.

$$\begin{aligned}
&|I^1(s, y^1) - I^2(s, y^2)| \\
&\leqq K_0(1 + |y^1|^{2q_0} + |y^2|^{2q_0})^{1/2}\Big\{|y^1 - y^2|^2 + \int_s^T \tilde{E}[|a^1(t, y^2(t), \omega) - a^2(t, y^2(t), \omega)|^2 \\
&\qquad + |b^1(t, y^2(t)) - b^2(t, y^2(t))|^2]\, dt\Big\}^{1/2} \\
&\quad + \tilde{E}|\gamma^1(y^2(T)) - \gamma^2(y^2(T))| + \int_s^T \tilde{E}|\delta^1(t, y^2(t), \omega) - \delta^2(t, y^2(t), \omega)|\, dt.
\end{aligned}$$

Proof.

$$\begin{aligned}
|I^1 - I^2| &\leqq \tilde{E}|\gamma_y^1(\rho_1 y^1(T) + (1 - \rho_1) y^2(T))|\, |y^1(T) - y^2(T)| \\
&\quad + \int_s^T \tilde{E}|\delta_y^1(\rho_2(t) y^1(t) + (1 - \rho_2(t)) y^2(t))|\, |y^1(t) - y^2(t)|\, dt \\
&\quad + \tilde{E}|\gamma^1(y^2(T)) - \gamma^2(y^2(T))| + \int_s^T \tilde{E}|\delta^1(t, y^2(t), \omega) - \delta^2(t, y^2(t), \omega)|\, dt,
\end{aligned}$$

by the mean value theorem. But the growth condition on γ_y^1 implies that the first term on the right is bounded by

$$B = \{K_1(1 + \tilde{E}|y^1(T)|^{2q_0} + \tilde{E}|y^2(T)|^{2q_0})\tilde{E}|y^1(T) - y^2(T)|^2\}^{1/2}.$$

The linear growth condition on a^i, b^i implies that $\tilde{E}|y^i(t)|^{2q_0} \leqq K_2(1 + |y^i|^{2q_0})$.

Finally,

$$y^1(t)-y^2(t)=(y^1-y^2)+\int_s^t (a^1(y^1(t'))-a^1(y^2(t')))\,dt'$$
$$+\int_s^t (b^1(y^1(t'))-b^1(y^2(t')))\,d\tilde{w}$$
$$+\int_s^t (a^1(y^2(t'))-a^2(y^2(t')))\,dt'+\int_s^t (b^1(y^2(t'))-b^2(y^2(t')))\,d\tilde{w},$$

so that the uniform bound on a^1_y, b^1_y and Gronwall's inequality implies

$$\tilde{E}|y^1(t)-y^2(t)|^2$$
$$\leqq K_3\left\{|y^1-y^2|^2+\int_s^T \tilde{E}[|a^1(y^2(t'))-a^2(y^2(t'))|^2+|b^1(y^2(t'))-b^2(y^2(t'))|^2]\,dt'\right\}.$$

Hence,

$$B\leqq\left\{K_4(1+|y^1|^{2q_0}+|y^2|^{2q_0})\left(|y^1-y^2|^2+\int_s^T \tilde{E}[|a^1(y^2(t))-a^2(y^2(t))|^2\right.\right.$$
$$\left.\left.+|b^1(y^2(t))-b^2(y^2(t))|^2]\,dt\right)\right\}^{1/2}.$$

The second term is bounded similarly to obtain the result.

COROLLARY 3.2. $|\tilde{V}(s,y')-\tilde{V}(s,y)|\leqq K_0(1+|y|^{2q_0}+|y'|^{2q_0})^{1/2}|y'-y|$.

Proof. For each $u\in\tilde{\mathcal{U}}(s)$, we have, from the lemma,

$$|\tilde{J}(s,y',u)-\tilde{J}(s,y,u)|\leqq K_0(1+|y|^{2q_0}+|y'|^{2q_0})^{1/2}|y'-y|,$$

so the result follows.

COROLLARY 3.3. $|\tilde{V}(s,y)|\leqq K_5(1+|y|^{2q_0+2})^{1/2}$.

Proof. Take $a^2=b^2=y^2=\gamma^2=\delta^2=0$, so $y^2(t)\equiv 0$, $I^2=0$, with $a^1(t,y,\omega)=f(t,y,u(t,\omega))$, $b^1=\sigma$, $y^1=y$, $\gamma^1=c$, $\delta^1(t,y,\omega)=l(t,y,u(t,\omega))$, $\tilde{J}(s,y,u)=I^1(s,y^1)$. Thus,

$$|\tilde{J}(s,y,u)|\leqq K_0(1+|y|^{2q_0})^{1/2}\left(|y|^2+\int_s^T |f(t,0,u)|^2+|\sigma(t,0)|^2\,dt\right)^{1/2}$$
$$+E^u|c(0)|+\int_s^T E^u|l(t,0,u)|dt$$
$$\leqq K_0(1+|y|^{2q_0})^{1/2}(|y|^2+2(T-s)K)^{1/2}+K^{1/2}+(T-s)K^{1/2},$$

by (A_2) and (A_3). The result follows.

Next we approximate the original problem by a nondegenerate smooth one. Extend $\bar{\sigma}$ onto $\mathbb{R}\times\mathbb{R}^d$ as a continuous function with support on a set bounded in t and $\bar{\sigma}_x$ bounded on $\mathbb{R}\times\mathbb{R}^d$. Set

$$\bar{\sigma}_n(t,x)=\int_{\mathbb{R}^d}\int_{-\infty}^{\infty}\bar{\sigma}(t-s,x-y)\beta^n(s,y)\,ds\,dy,$$

where β^n is a C^∞ function, nonnegative, with support in $S_{1/n}=\{(t,x):|t|\leqq 1/n,\ |x|\leqq 1/n\}$ and $\iint\beta^n\,ds\,dy=1$. It follows that $\bar{\sigma}_n$ is smooth, $\bar{\sigma}_n\to\bar{\sigma}$ uniformly on compact

sets, and $\bar{\sigma}_n$, $(\bar{\sigma}_n)_x$ are bounded uniformly in n. Since $(\bar{\sigma})^{-1}$ is bounded on Q, then on Q $|\det \bar{\sigma}| \geqq \rho > 0$, so $|\det \bar{\sigma}_m| \geqq \rho/2$ on $S_n \cap Q$ for $m \geqq m_n$. On $S_n \cap Q$ define $\bar{\sigma}^n = \bar{\sigma}_{n \vee m_n}$ and extend it to $[0, T] \times \mathbb{R}^d$ to be smooth, invertible, and bounded along with $\bar{\sigma}^n_x$ and $(\bar{\sigma}^n)^{-1}$ uniformly in n. Since $(a^{-1})_{x_i} = -a^{-1} a_{x_i} a^{-1}$, then $(\bar{\sigma}^n)^{-1}_x$ is bounded. Now set

$$\sigma^n(t, x) = \begin{pmatrix} \sqrt{2/n}\, I & 0 \\ 0 & \bar{\sigma}^n(t, x) \end{pmatrix}.$$

Next extend ϕ to be zero off $[0, T]$, and set

$$\phi^n(t, x) = \int_{\mathbb{R}^d} \int_{-\infty}^{\infty} \phi(t - s, x - y) \beta^n(s, y)\, ds\, dy.$$

Then for any $q < \infty$, any A compact,

$$\iint_A |\phi^n(t, x) - \phi(t, x)|^q\, dt\, dx \to 0.$$

Moreover, ϕ^n is smooth, satisfies a linear growth condition uniformly in n, and $(\phi^n)_x$ is bounded uniformly in n. Next, extend g as a continuous function on $\mathbb{R} \times \mathbb{R}^d \times U$, zero for t not in some compact set, and define

$$g^n(t, x, u) = \int_{\mathbb{R}^d} \int_{-\infty}^{\infty} g(t - s, x - y, u) \beta^n(s, y)\, ds\, dy$$

on $S_n \times U$. Extend g^n to $[0, T] \times \mathbb{R}^d \times U$ as a smooth bounded function such that g^n_x is bounded uniformly in n and $|g^n(t, x, u)|^2 \leqq K(1 + |x|^2)$. Then, $g^n \to g$ uniformly on compact sets.

It follows now that σ^n, $\phi^n \in C^{1,2}(\bar{Q})$; $\theta^n \equiv (\sigma^n)^{-1} \begin{pmatrix} 0 \\ g^n \end{pmatrix} \in C^{1,1}(\bar{Q} \times U)$; and σ^n, $(\sigma^n)^{-1}$, $(\sigma^n)_x$, $(\phi^n)_x$, θ^n, θ^n_x are bounded. c^n and l^n are defined in the same way as ϕ^n. Then c^n and l^n_2 are smooth, and along with c^n_x, $(l^n_1)_x$ they satisfy a polynomial growth condition. Moreover, on compact sets, $l^n_1 \to l_1$ and $c^n \to c$ in L_q (just as was the case for ϕ^n). Also $l^n_2 \to l_2$ uniformly on compact sets (since l_2 is continuous).

It now follows from [5, §§ VI.4/6] that the "n" problem has an optimal feedback solution $u^n(t, x)$, and that

$$J^n(s, x, u^n) = V^n(s, x) = \tilde{V}^n(s, x) = \tilde{J}^n(s, x, u^{n,x}),$$

where $u^{n,x}(t, \omega) = u^n(t, x^n(t, \omega))$ and x^n is the unique solution of

$$dx = \phi^n(t, x)\, dt + \sigma^n(t, x)\, dw, \qquad x(s) = x, \tag{3.2}$$

on $(\Omega, \mathcal{F}, P)$. Then $\tilde{V}^n$ satisfies

$$\begin{aligned} 0 = {} & V^n_s + \frac{1}{n} \sum_{i=1}^{\nu} V^n_{x_i x_i} \\ & + \sum_{ij=\nu+1}^{d} \frac{1}{2} [\bar{\sigma}^n (\bar{\sigma}^n)^*]_{ij} V^n_{x_i x_j} + V^n_x \phi^n + \min_{u \in U} \left[\sum_{i=\nu+1}^{d} g^n_i V^n_{x_i} + l^n \right], \end{aligned} \tag{3.3}$$

$$V^n(T, x) = c^n(x).$$

We will next proceed to show that $\tilde{V}^n \to \tilde{V}$ and $\tilde{V}$ is a (generalized) solution of (3.3) in the limit as $n \to \infty$.

LEMMA 3.4. (cf. [5, Lemma 8.1, p. 179]):

(a) $|\tilde{V}^n_x(t, y)| \leqq M(1 + |y|^{2q_0})^{1/2}$.

(b) *For any bounded set* $A \subset Q$ *and* $1 < \lambda < \infty$, *there exists* $M_{A,\lambda}$ *such that*

$$\int_A \left[|\tilde{V}_s^n|^\lambda + \sum_{i,j=+1}^{n} |\tilde{V}_{x_i x_j}^n|^\lambda \right] dy\, ds \leqq M_{A,\lambda}.$$

Proof. (a) follows from Corollary 3.2. Since $\tilde{V}_x^n$, f^n, l^n are uniformly bounded on bounded sets as are $\tilde{V}^n$ (cf. Corollary 3.3), then the proof of [5, Lemma 8.1(b)] establishes (b). (Note that $\{\bar{\sigma}^n(\bar{\sigma}^n)^*\}$ are equicontinuous and elliptic uniformly in n so that the a priori estimates are uniform in n.)

We need the following auxiliary result. For $u \in \tilde{\mathcal{U}}(s)$ let $x^u(t)$ be the solution of (3.1) for $t \geqq s$, $x^u(s) = y$ (fixed), and $x(t)$ be the solution of (1.4) for $t \geqq s$, $x(t) = y$, $w = w^u$. For θ square integrable in (t, ω), adapted, set

$$\zeta_s^t(\theta(\cdot)) \equiv \exp\left\{ \int_s^t \theta(t') \cdot dw^u - \frac{1}{2} \int_s^t |\theta(t')|^2\, dt' \right\},$$

and set $[\theta^u(t)]^* \equiv (0, [\bar{\sigma}^{-1}(t, x(t))g(t, x(t), u(t, \omega))]^*)$.

LEMMA 3.5. *If* ψ *is measurable and* $|\psi(t, x)|^2 \leqq K(1 + |x|^{2p})$, *then*

$$E^u \psi(t, x^u(t)) = E^u\{\zeta_s^t(\theta^u(\cdot))\psi(t, x(t))\},$$

and there exists $\lambda_0 > 1$ *such that for* $1 < \lambda < \lambda_0$

$$\sup\{E^u[\zeta_s^t(\theta^u(\cdot))^\lambda] : u \in \tilde{\mathcal{U}}(s),\ t \geqq s\} < \infty.$$

Proof.

$$dx^u = \phi(t, x^u)\, dt + \sigma(t, x^u)\, d\bar{w}^u,$$

where

$$d\bar{w}^u = dw^u + \bar{\theta}^u(t)\, dt$$

is a Brownian motion under $\bar{P}^u$ and $\bar{\theta}^u$ is defined like θ^u but with $x(t)$ replaced by $x^u(t)$. But by law uniqueness $(x^u, \bar{w}^u, \bar{P}^u) = (x, w^u, P^u)$ so that

$$E^u \psi(t, x^u(t)) = \bar{E}^u \left[\exp\left\{ \int_s^t \bar{\theta}^u(t') \cdot d\bar{w}^u - \frac{1}{2} \int_s^t |\bar{\theta}^u(t')|^2\, dt' \right\} \psi(t, x^u(t)) \right]$$

$$= E^u[\zeta_s^t(\theta^u(\cdot))\psi(t, x(t))].$$

Moreover, ($\mathcal{F}_t$ is the σ-algebra generated by the past of x),

$$E^u \zeta_s^t (\theta^u)^\lambda = E^u\{(E^u \zeta_s^T(\theta^u) | \mathcal{F}_t)^\lambda\}$$

$$\leqq E^u\{\zeta_s^T(\theta^u)^\lambda\}$$

is bounded uniformly in u (cf. [7, Corollary 4.2]).

LEMMA 3.6. *If* A *is a bounded subset of* Q, *then*

(i) $\tilde{V}^n \to \tilde{V}$ *uniformly on* $\bar{A}$,

(ii) $\tilde{V}_s^n$, $\tilde{V}_x^n$, $\tilde{V}_{x_i x_j}^n$, $i, j = \nu + 1, \cdots, d$, *converge weakly to* $\tilde{V}_s$, $\tilde{V}_x$, $\tilde{V}_{x_i x_j}$ *respectively in* $L_\lambda(A)$ *for* $1 < \lambda < \infty$,

(iii) *for* $i > \nu$, $\tilde{V}_{x_i}^n \to \tilde{V}_{x_i}$ *pointwise in* Q, *and* $\tilde{V}_{x_i}$ *is continuous in* $(t, x_{\nu+1}, \cdots, x_d)$ *for* $(x_1, \cdots, x_\nu)$ *fixed.*

Proof. Without loss of generality, assume $A = [0, T] \times A_0$. First we establish $\tilde{V}^n \to \tilde{V}$ pointwise. For $u \in \tilde{\mathcal{U}}(s)$

$$\tilde{J}^n(s, y, u) = E^u\left\{ c^n(\tilde{x}^n(T)) + \int_s^T l^n(t, \tilde{x}^n(t), u(t, \omega))\, dt \right\},$$

where $\tilde{x}^n(s) = y$ and, for $t \geqq s$,

$$d\tilde{x}^n = f^n(t, \tilde{x}^n(t), u(t))\, dt + \sigma^n(t, \tilde{x}^n(t))\, dw^u.$$

We now apply Lemma 3.1 with $y^1 = y^2 = y$, $a^1(t, x, \omega) = f^n(t, x, u(t, \omega))$, $b^1 = \sigma^n$, $\delta^1 = l^n$, $\gamma^1 = c^n$, and $a^2(t, x, \omega) = f(t, x, u(t, \omega))$, $b^2 = \sigma$, $\delta^2 = l$, $\gamma^2 = c$, $\tilde{w} = w^u$. Thus,

$$\begin{aligned}
&|\tilde{J}^n(s, y, u) - \tilde{J}(s, y, u)| \\
&\leqq K_6(1 + |y|^{2q_0})^{1/2} \Big\{ \int_s^T E^u[|f^n(t, x^u(t), u(t)) - f(t, x^u(t), u(t))|^2 \\
&\qquad + |\sigma^n(t, x^u(t)) - \sigma(t, x^u(t))|^2]\, dt \Big\}^{1/2} \\
&\qquad + E^u|c^n(x^u(T)) - c(x^u(T))| \\
&\qquad + \int_s^T E^u|l^n(t, x^u(t), u(t)) - l(t, x^u(t), u(t))|\, dt.
\end{aligned}$$

But

$$\lim_{n\to\infty} \sup_u \int_s^T E^u|l_1^n(t, x^u(t)) - l_1(t, x^u(t))|^2\, dt = \lim_{s' \downarrow s} \lim_{n\to\infty} \sup_u \int_{s'}^T E^u|l_1^n - l_1|^2\, dt,$$

since $E^u|l_1^n(t, x^u(t)) - l_1(t, x^u(t))|^2 \leqq 2K(1 + E^u|x^u(t)|^{2q_0})$ is bounded uniformly in t, u by (A_3), Lemma 3.5, and the fact that $\sup\{E^u|x(t)|^q : 0 \leqq t \leqq T\} < \infty$, for all $q < \infty$. Now with $B = \{x : |x| \leqq N\}$, $N < \infty$, we have from Lemma 3.5

$$\begin{aligned}
&\int_{s'}^T E^u|l_1^n(t, x^u(t)) - l_1(t, x^u(t))|^2\, dt \\
&\quad = \int_{s'}^T E^u|l_1^n(t, x(t)) - l_1(t, x(t))|^2 \zeta_s^t(\theta^u)\, dt \\
&\quad = \Big(\int_{s'}^T \int_B + \int_{s'}^T \int_{B^c} \Big) |l_1^n(t, x) - l_1(t, x)|^2 E^u\{\zeta_s^t(\theta^u) | x(t) = x\} \tilde{p}(t, x; s, y)\, dx\, dt \\
&\quad \leqq \|l_1^n - l_1\|^2_{2q, [s', T]\times B} \|E^u[\zeta_s^{\cdot}(\theta^u) | x(\cdot) = \cdot\,] \tilde{p}(\cdot, \cdot; s, y)\|_{q', [s', T]\times B} \\
&\qquad + \Big(\int_{s'}^T E^u|l_1^n(t, x(t)) - l_1(t, x(t))|^{2q} 1_{\{|x(t)| > N\}}\, dt \Big)^{1/q} \Big(\int_s^T E^u(\zeta_s^t(\theta^u)^{q'}\, dt) \Big)^{1/q'}
\end{aligned}$$

where 1_A is the characteristic function of A.

Since $l_1^n \to l_1$ in $L_{2q}([s', T]\times B)$ for any $q < \infty$, since $l_1^n(t, \cdot)$ is polynomially bounded uniformly in n and t, and since

$$\begin{aligned}
&\int_{s'}^T \int_{\mathbb{R}^d} E^u[\zeta_s^t(\theta^u) | x(t) = x]^{q'} \tilde{p}(t, x; s, y)^{q'}\, dx \\
&\leqq \|\tilde{p}(\cdot, \cdot; s, y)\|^{\mu(q'-1)/(\mu-1)}_{\mu, [s', T]\times\mathbb{R}^d} \Big\{ \int_{s'}^T \int_{\mathbb{R}^d} E^u[\zeta_s^t(\theta^u) | x(t) = x]^{q'(\mu-1)/(\mu-q')} \\
&\qquad \tilde{p}(t, x; s, y)\, dx\, dt \Big\}^{(\mu-q')/(\mu-1)}
\end{aligned}$$

is bounded uniformly in u by (A_4), Jensen's inequality and Lemma 3.5, for $1 < q' < \mu$, q' small enough, then $\lim_{n\to\infty} \int_{s'}^T E^u|l_1^n - l_1|^2\, dt = 0$ uniformly in $u \in \tilde{\mathcal{U}}(s)$, and so $\lim_{n\to\infty} \int_s^T E^u|l_1^n - l_1|^2\, dt = 0$ uniformly in u.

Next, consider

$$E^u|c^n(x^u(T))-c(x^u(T))|$$
$$=E^u\zeta_s^T(\theta^u)|c^n(x(T))-c(x(T))|$$
$$\leqq[E^u(E^u\{\zeta_s^T(\theta^u)|x(T)=x\})^{q'}]^{1/q'}\left[\int_{\mathbb{R}^d}|c^n(x)-c(x)|^q\tilde{p}(T,x;s,y)\,dx\right]^{1/q}.$$

As above, the first factor on the right side is bounded uniformly in u. Since $c^n\to c$ in $L_q(B)$, then $c^n\to c$ in measure on B. Since the density $\tilde{p}$ exists, then $|c^n(x(T))-c(x(T))|1_{\{\omega:|x(T)|\leqq N\}}\to 0$ in probability for each $N<\infty$. Moreover, c^n, c have a polynomial bound uniformly in n, $E^u|x(T)|^{q_0}<\infty$ for all $q_0<\infty$, and $P^u\{|x(T)|>N\}\to 0$ as $N\to\infty$, so that $\int_{|x|>N}|c^n(x)-c(x)|^q\tilde{p}(T,x;s,y)\,dx\to 0$ as $N\to\infty$ uniformly in n. Hence, $\lim_{n\to\infty}E^u|c^n(x^u(T))-c(x^u(T))|=0$ uniformly in u.

Now consider

$$\int_s^T E^u|l_2^n(t,x^u(t),u(t))-l_2(t,x^u(t),u(t))|\,dt,$$
$$\leqq E^u\left(\int_s^{\tau_N}+\int_{\tau_N}^T\right)|l_2^n(t,x^u(t),u(t))-l_2(t,x^u(t),u(t))|\,dt,$$

where $\tau_N=\min(T,\inf\{t\geqq s:|x^u(t)|\geqq N\})$. By continuity of l_2, $l_2^n\to l_2$ uniformly on compact sets, so for each $N<\infty$, $E^u\int_s^{\tau_N}|l_2^n-l_2|\,dt\to 0$ uniformly in u. Moreover, l_2^n, l_2 satisfy a polynomial growth condition uniformly in t, u, n and

$$E^u\int_{\tau_N}^T|l_2^n-l_2|^2\,dt\leqq\left(\int_s^T 2K(1+E^u|x^u(t)|^{2q_0})dt\right)^{1/p}\left(E(T-\tau_N)\right)^{1/p_0},$$

where q_0 comes from the growth condition and $p^{-1}+p_0^{-1}=1$. But

$$E^u(T-\tau_N)\leqq TP^u\left\{\sup_{s\leqq t\leqq T}|x^u(t)|\geqq N\right\}$$
$$=T\int_{\{\omega:\sup|x(t)|\geqq N\}}\zeta_s^T(\theta^u)\,dP^u$$
$$\leqq TE^u\{\zeta_s^T(\theta^u)^\lambda\}^{1/\lambda}P^u\left\{\omega:\sup_{s\leqq t\leqq T}|x(t)|\geqq N\right\}^{1/\lambda'}$$
$$\to 0$$

uniformly in u, as $N\to\infty$, for $1<\lambda<\lambda_0$, $\lambda^{-1}+(\lambda')^{-1}=1$. Also, $E^u|x^u(t)|^{2p}$ is bounded uniformly in t, u. Thus, $\lim_{n\to\infty}\int_s^T E^u|l_2^n-l_2|\,dt=0$ uniformly in u.

The terms involving σ and f are treated similarly. This proves that $\tilde{V}^n\to\tilde{V}$ pointwise on Q. To establish the uniform convergence on $\bar{A}$, observe that Lemma 3.4(a) implies $|\tilde{V}^n(s,y)-\tilde{V}^n(s,y')|\leqq M_A|y-y'|$ and similarly, for $\tilde{V}$, on $\bar{A}$. Given $\varepsilon>0$, there exist points $\{y^i\}_{i=1}^M$ such that the spheres of radius $\varepsilon/(3M_A)$ centered at y^i cover $\bar{A}_0$. Hence,

$$(*)\qquad |\tilde{V}^n(s,y)-\tilde{V}(s,y)|<2\frac{\varepsilon}{3}+\max_{1\leqq i\leqq M}|\tilde{V}^n(s,y^i)-\tilde{V}(s,y^i)|;$$

i.e., $\tilde{V}^n(s,\cdot)\to\tilde{V}(s,\cdot)$ uniformly on compact sets for each s fixed. Following [5, Chapt.

VI, § 8], fix y^i and set $\psi^n(t, z) = \tilde{V}^n(t, y)$ where

$$\begin{aligned} z_j &= \sqrt{n}(y_j - y_j^i), \qquad j = 1, \cdots, \nu, \\ &= y_j, \qquad\qquad\quad j = \nu+1, \cdots, d. \end{aligned} \tag{3.4}$$

Now the a priori estimates of [5, Chapt. VI, Lemmas 8.1, 8.2] and the Ascoli theorem show that ψ^n, $\psi^n_{z_i}$ converge uniformly on any compact set. Hence, $\psi^n(s, z) \to \tilde{V}(s, \tilde{y}^i)$ uniformly in s where $\tilde{y}_j^i = y_j^i$, $j = 1, \cdots, \nu$, $\tilde{y}_j^i = z_j$, $j = \nu+1, \cdots, d$. Thus, for n sufficiently large,

$$\max_{1 \leqq i \leqq M} |\tilde{V}^n(s, y^i) - \tilde{V}(s, y^i)| = \max_{1 \leqq i \leqq M} |\psi^n(s, 0, \cdots, 0, y^i_{\nu+1}, \cdots, y^i_d) - \tilde{V}(s, y^i)|$$

$$< \frac{\varepsilon}{3},$$

for all $s \in [0, T]$. Substituting this into (*) shows that $\tilde{V}^n \to \tilde{V}$ uniformly on $\bar{A}$.

Observe that for $i > \nu$, $\psi^n_{z_i} = \tilde{V}^n_{y_i}$, so that (iii) also holds.

Finally, (i) and the weak sequential compactness implied by Lemma 3.4 imply (ii). This proves the lemma.

Remark. If we write $\psi^n(t, z, y^0) = \tilde{V}^n(t, y)$, where z is defined as in (3.4) but with y^i replaced by y^0, then the above proof (and the proof of [5, Chapt. VI, Lemma 8.2]) actually shows that for $i > \nu$, y^0 fixed, $\psi^u_{z_i}(t, z, y^0) \to V_{y_i}(t, \tilde{y})$ uniformly for (t, z) in a compact set, where $\tilde{y}_i = y_i^0$, $i \leqq \nu$, $\tilde{y}_i = z_i$, $i > \nu$.

We say that $W(t, x)$ is a *generalized solution* of

$$\begin{aligned} W_t(t, x) + \frac{1}{2} \sum_{i,j=\nu+1}^{d} (\bar{\sigma}(t, x)\bar{\sigma}^*(t, x))_{ij} W_{x_i x_j} + W_x(t, x)\phi(t, x) + l_1(t, x) \\ + \min_{u \in U} \left[\sum_{i=\nu+1}^{d} g_i(t, x, u) W_{x_i}(t, x) + l_2(t, x, u) \right] = 0, \end{aligned} \tag{3.5}$$

if W is continuous on $\bar{Q}$, satisfies a polynomial growth condition, and for any bounded $A \subset Q$, W_t, W_{x_i}, $i = 1, \cdots, d$, $W_{x_i x_j}$, $i, j = \nu+1, \cdots, d$, are in $L_\lambda(A)$ for some $\lambda > \mu/(\mu - 1)$, and (3.5) is satisfied almost everywhere on Q.

THEOREM 3.7. *Assume* A_1–A_4. *Then* $\tilde{V}$ *is a generalized solution of* (3.5).

Proof. By Lemma 3.6(i), $\tilde{V}$ is continuous on $[0, T] \times \{x: |x| \leqq N\}$ for all $N < \infty$. Hence, it is continuous on $\bar{Q}$. By Corollary 3.3, it satisfies a polynomial growth condition. By Lemma 3.6(ii), the generalized partial derivatives are suitably bounded, i.e., in L_λ.

It remains to show that $\tilde{V}$ satisfies (3.5) a.e. If ψ is a smooth function of compact support A in Q, then (3.3) yields

$$\begin{aligned} 0 = &-\frac{1}{n} \int\int_A \sum_{i=1}^{\nu} \tilde{V}^n_{y_i} \psi_{y_i} \, dy \, ds \\ &+ \int\int_A \left[\tilde{V}^n_s + \frac{1}{2} \sum_{ij=\nu+1}^{d} [\bar{\sigma}^n(\bar{\sigma}^n)^*]_{ij} \tilde{V}^n_{x_i x_j} + \tilde{V}^n_x \phi^n + l_1^n \right] \psi \, dy \, ds \\ &+ \int\int_A \min_{u \in U} \left[\sum_{i=\nu+1}^{d} g_i^n \tilde{V}^n_{x_i} + l_2^n \right] \psi \, dy \, ds. \end{aligned} \tag{3.6}$$

Now let $n \to \infty$. Lemma 3.4(a) implies that the first term goes to 0. Set $a^n \equiv \bar{\sigma}^n(\bar{\sigma}^n)^*$.

Then

$$\iint_A a_{ij}^n \tilde{V}_{x_i x_j}^n \psi = \iint_A (a_{ij}^n - a_{ij}) \tilde{V}_{x_i x_j}^n \psi + \iint a_{ij} \tilde{V}_{x_i x_j}^n \psi.$$

But $\|a_{ij}^n - a_{ij}\|_{\lambda', A} \to 0$ $[\lambda^{-1} + (\lambda')^{-1} = 1]$ because $\bar{\sigma}^n \to \bar{\sigma}$ in $L_q(A)$ for all $q < \infty$, and $\|\tilde{V}_{x_i x_j}^n\|_{\lambda, A}$ is uniformly bounded by weak convergence, so

$$\iint_A a_{ij}^n \tilde{V}_{x_i x_j}^n \psi \to \iint_A a_{ij} \tilde{V}_{x_i x_j} \psi.$$

Similar arguments show that the second integral in (3.6) converges to the same expression without n.

Moreover,

$$g_i^n(s, y, u) \tilde{V}_{x_i}^n(s, y) = [g_i^n(s, y, u) - g_i(s, y, u)] \tilde{V}_{x_i}^n(s, y) + g_i(s, y, u) \tilde{V}_{x_i}^n(s, y).$$

Since $g^n \to g$ uniformly on compact sets and $\tilde{V}_x^n$ and g are bounded on bounded sets, then $g_i^n \tilde{V}_{x_i}^n \to g_i \tilde{V}_{x_i}$ for each (s, y) uniformly in u (Lemma 3.6(iii)). l_2^n is treated similarly, and so the last integral in (3.6) also converges appropriately; i.e.,

$$0 = \iint_Q \left[\tilde{V}_s + \frac{1}{2} \sum_{ij=\nu+1}^{d} a_{ij} \tilde{V}_{x_i x_j} + \tilde{V}_x \phi + l_1 + \min_{u \in U} (\tilde{V}_x f_2 + l_2) \right] \psi \, ds \, dy.$$

Hence, $\tilde{V}$ satisfies (3.5) a.e.

COROLLARY 3.8. *Assume* A_1–A_4. *Then there exists an optimal feedback control for the (unconstrained) problem.*

Proof. cf. [5, Chapt. VI, Thm. 8.3].

COROLLARY 3.9. $\tilde{V} = V$.

Proof. cf. [5, Chapt. IV, Thm. 8.1 and Corollary 4.2].

4. The adjoint process for $G = \mathbb{R}^d$. Assume now that we are given an optimal control $\hat{u} \in \mathcal{U}$ so that

$$\begin{aligned} dx &= \phi(t, x)\, dt + \sigma(t, x)\, dw \\ &= f(t, x, \hat{u}(t, x))dt + \sigma(t, x)\, d\hat{w}, \end{aligned}$$

where $\hat{w}$ is a standard Brownian motion on $(\Omega, \mathcal{F}, \hat{P})$. We shall first show that, cf. (1.5), $p(t, x) = -\tilde{V}_x(t, x)$; then we shall give the required representation of $\tilde{V}_x$ $(= V_x)$. Again, we treat first the reduced case when $k_0 = 0$. With x as in (1.4), set $\hat{u}(t) \equiv \hat{u}(t, x(t, \omega))$, and let $\bar{x}^n$ be the unique solution of

$$\bar{x}^n(t) = x_0 + \int_0^t f(s, x(s), \hat{u}(s))\, ds + \int_0^t \sigma^n(s, \bar{x}^n(s))\, d\hat{w}.$$

Then,

$$\begin{aligned} &c^n(\bar{x}^n(T)) \\ &= \tilde{V}^n(0, x_0) + \int_0^T \Big\{ \tilde{V}_t^n(t, \bar{x}^n(t)) + \frac{1}{n} \sum_{i=1}^{\nu} \tilde{V}_{x_i x_i}^n(t, \bar{x}^n(t)) \\ &\qquad + \frac{1}{2} \sum_{ij=\nu+1}^{d} (\bar{\sigma}^n(t, \bar{x}^n(t)) \bar{\sigma}^n(t, \bar{x}^n(t))^*)_{ij} \tilde{V}_{x_i x_j}^n(t, \bar{x}^n(t)) \\ &\qquad + \tilde{V}_x^n(t, \bar{x}^n(t)) f(t, x(t), \hat{u}(t)) \Big\}\, dt \\ &\qquad + \sum_{i=1}^{\nu} \sqrt{\frac{2}{n}} \int_0^T \tilde{V}_{x_i}^n(t, \bar{x}^n(t))\, d\hat{w}_i \end{aligned}$$

$$
\begin{aligned}
&+\sum_{ij=\nu+1}^{d}\int_0^T \tilde{V}^n_{x_i}(t,\bar{x}^n(t))\bar{\sigma}^n(t,\bar{x}^n(t))_{ij}\,d\hat{w}_j\\
&\geqq -\int_0^T l^n(t,\bar{x}^n(t),\hat{u}(t))\,dt+\tilde{V}^n(0,x_0)\\
&\quad+\int_0^T \tilde{V}^n_x(t,\bar{x}^n(t))[f(t,x(t),\hat{u}(t))-f^n(t,\bar{x}^n(t),\hat{u}(t))]\,dt\\
&\quad+\sum_{i=1}^{\nu}\sqrt{\frac{2}{n}}\int_0^T \tilde{V}^n_{x_i}(t,\bar{x}^n(t))\,d\hat{w}_i\\
&\quad+\sum_{ij=\nu+1}^{d}\int_0^T \tilde{V}^n_{x_i}(t,\bar{x}^n(t))\bar{\sigma}^n(t,\bar{x}^n(t))_{ij}\,d\hat{w}_j.
\end{aligned}
\tag{4.1}
$$

Standard arguments using Gronwall's inequality and the convergence of $\int_0^T \hat{E}|\sigma^n(s,x(s))-\sigma(s,x(s))|^\lambda\,ds$ to zero (cf. Lemma 3.6) show that $\hat{E}\sup_t|\bar{x}^n(t)-x(t)|^\lambda\to 0$ for $\lambda<\infty$. This and the polynomial growth condition on c^n_x (uniform in n) imply

$$
\begin{aligned}
\hat{E}|c^n(\bar{x}^n(T))-c(x(T))|&\leqq\{K_{q_0}(1+\hat{E}|x(T)|^{2q_0}+\hat{E}|\bar{x}^n(t)-x(t)|^{2q_0})\hat{E}|\bar{x}^n(T)-x(T)|^2\}^{1/2}\\
&\quad+\hat{E}|c^n(x(T))-c(x(T))|\to 0,
\end{aligned}
$$

as in Lemma 3.6. Similarly,

$$
\hat{E}\int_0^T |l^n(t,\bar{x}^n(t),\hat{u}(t))-l(t,x(t),\hat{u}(t))|\,dt\to 0.
$$

Moreover, $\tilde{V}^n_x$ is polynomially bounded uniformly in n so $E\int_0^T|\tilde{V}^n_x(f-f^n)|\,dt\to 0$, $\sqrt{(2/n)}E|\int_0^T \tilde{V}^n_{x_i}\,d\hat{w}_i|\to 0$. Finally, observe that

$$
\bar{x}^n_i(t)=x_i(t)+\sqrt{\frac{2}{n}}\hat{w}_i(t),\qquad i\leqq\nu,
$$

so that if we define the process $x^0(t)\in\mathbb{R}^\nu$ by $x^0_i(t)=x_i(t)$, and if we set $\tilde{y}^n_i(t)=x_i(t)$, $z^n_i(t)=\sqrt{2}\,\hat{w}_i(t)$, $i\leqq\nu$, $\tilde{y}^n_i(t)=z^n_i(t)=\bar{x}^n_i(t)$, $i>\nu$, then

$$
\tilde{V}^n(t,\bar{x}^n(t))=\psi^n(t,z^n(t),x^0(t)),
$$

where $\tilde{V}^n$, ψ^n are related as in the proof of Lemma 3.6. But by the remark following Lemma 3.6, for $i>\nu$, $\psi^n_{z_i}(t,z,y^0)\to\tilde{V}_{x_i}(t,\tilde{y})$ uniformly on compact subsets for y^0 fixed. Since $\sup_t|\bar{x}^n(t)-x(t)|\to 0$ in probability, then $\hat{P}\{\omega:\sup_{0\leqq t\leqq T}|\hat{w}(t)|>N$, $\sup_{0\leqq t\leqq T}|\bar{x}^n(t)|>N\}<\varepsilon$, for all $N>N_\varepsilon$, $n>n_\varepsilon$ (but n_ε independent of N). Hence,

$$
\tilde{V}^n_{x_i}(t,\bar{x}^n(t))-\tilde{V}_{x_i}(t,\tilde{y}^n(t))\to 0
$$

in measure (t,ω). Again Lemma 3.4(a), the boundedness and Lipschitz continuity of $\bar{\sigma}^n$, and the above convergence in measure imply that

$$
\begin{aligned}
\int_0^T \tilde{V}^n_{x_i}(t,\bar{x}^n(t))\bar{\sigma}^n_{ij}\,d\hat{w}_j&=\int_0^T[\tilde{V}^n_{x_i}(t,\bar{x}^n(t))-\tilde{V}_{x_i}(t,\tilde{y}^n(t))]\bar{\sigma}^n_{ij}\,d\hat{w}_j\\
&\quad+\int_0^T \tilde{V}_{x_i}(t,\tilde{y}^n(t))\bar{\sigma}^n_{ij}\,d\hat{w}_j\\
&\to\int_0^T \tilde{V}_{x_i}(t,x(t))\bar{\sigma}_{ij}\,d\hat{w}_j,
\end{aligned}
$$

in the mean, since (by Lemma 3.6(iii)) $\tilde{V}_{x_i}$ is continuous in x_j, $j > \nu$. Thus, we have (recall (1.6) and following) from (4.1)

$$-\int_0^T (p(t, x(t)) + \tilde{V}_x(t, x(t)))\sigma(t, x(t))\, d\hat{w} \geqq 0,$$

whence follows

LEMMA 4.1. *Assume* A_1–A_4. *Then*

$$p_i(t, x) = -\tilde{V}_{x_i}(t, x), \qquad i > \nu. \tag{4.2}$$

Note that since x is independent of $w_1, \cdots, w_\nu$, then $p_i \equiv 0$ for $i \leqq \nu$.

Let us write $u^x(s, \omega) \equiv \hat{u}(s, x(s, \omega; x))$ where $x(s, \omega; x)$, $s \geqq t$, is the unique solution of

$$dx = \phi(s, x)\, ds + \sigma(s, x)\, dw, \qquad x(t) = x.$$

Then $u^x \in \tilde{\mathcal{U}}(t)$ and $\tilde{V}(t, x) = \tilde{J}(t, x, u^x)$, cf. Corollary 3.9. Thus, we are interested in $\tilde{J}_x(t, x, u^x)$. For $u \in \tilde{\mathcal{U}}(t)$, x fixed, we let $x^u(s)$, $s \geqq t$, be the solution of (3.1) with $x^u(t) = x$ on $(\Omega^u, \mathcal{F}^u, P^u)$. Then we have the following characterization of the gradient.

LEMMA 4.2. *For* $u \in \tilde{\mathcal{U}}(t)$

$$\tilde{J}_x(t, x, u) = E^u\left\{c_x[x^u(T)]\Phi^u(T, t) + \int_t^T l_x(s, x^u(s), u(s))\Phi^u(s, t)\, ds\right\}, \tag{4.3}$$

where Φ^u *is the fundamental matrix solution of the linearized version of* (3.1).

Proof. For $z \in \mathbb{R}^d$, let $x^{u,z}(\cdot)$ be the solution of (3.1) with $x^{u,z}(t) = x + z$. From well-known estimates (cf. [12, pp. 559–561]), it follows that $E^u|x^{u,z}(s) - x^u(s)|^2 = O(|z|^2)$ and $E^u|x^{u,z}(s) - x^u(s) - \Phi^u(s, t)z|^2 = o(|z|^2)$ uniformly in s. Hence,

$$\begin{aligned}
&\tilde{J}(t, x+z, u) - \tilde{J}(t, x, u) \\
&= E^u\Big\{c_x[x^u(T)]\Phi^u(T, t)z + \int_t^T l_x(s, x^u(s), u(s))\Phi^u(s, t)z\, ds \\
&\qquad + c_x[x^u(T)][x^{u,z}(T) - x^u(T) - \Phi^u(T, t)z] \\
&\qquad + [c_x((1-\phi_1(\omega))x^u(T) + \phi_1(\omega)x^{u,z}(T)) - c_x(x^u(T))][x^{u,z}(T) - x^u(T)] \\
&\qquad + \int_t^T l_x(s, x^u(s), u(s))[x^{u,z}(s) - x^u(s) - \Phi^u(s, t)z] \\
&\qquad\qquad + [l_x(s, (1-\phi_2(\omega))x^u(s) + \phi_2(\omega)x^{u,z}, u(s)) - l_x(s, x^u(s), u(s))] \\
&\qquad\qquad\qquad \cdot [x^{u,z}(s) - x^u(s)]\, ds\Big\} \\
&= E^u\left\{c_x[x^u(T)]\Phi^u(T, t) + \int_t^T l_x(s, x^u(s), u(s))\Phi^u(s, t)\, ds\right\}z + o(|z|),
\end{aligned}$$

by the continuity and polynomial growth of c_x and l_x. Here, $0 \leqq \phi_i(\omega) \leqq 1$. This proves (4.3).

Define Φ by

$$d\Phi(t, t') = f_x(t, x(t), \hat{u}(t, x(t)))\Phi(t, t')\, dt + \sum_{k=1}^d \sigma_x^k(t, x(t))\Phi(t, t')\, d\hat{w}_k, \tag{4.4}$$

with $\Phi(t', t') = I$, σ^k being the kth column of σ.

THEOREM 4.3. *Assume* A_1–A_4. *Then*

$$-p_i(t,x) = V_{x_i}(t,x)$$

$$= \hat{E}_{tx}\left\{ c_x(x(T))\Phi(T,t) + \int_t^T l_x(s,x(s),\hat{u}(s,x(s)))\Phi(s,t)\,ds \right\}_i, \qquad i > \nu. \tag{4.5}$$

Proof. Since $\tilde{J}(t,y,u^x) - V(t,y) \geqq 0$, with equality at $y = x$, and since $\tilde{J}_{x_i}$, V_{x_i} exist, then

$$\tilde{J}_{x_i}(t,x,u^x) = V_{x_i}(t,x), \qquad i > \nu.$$

The result now follows from (4.3) since $u^x(s,\omega) = \hat{u}(s,x(s,\omega))$ with $P^{u^x} = \hat{P}|_{x(t)=x}$, so that $\Phi^{u^x} = \Phi$.

If we strengthen A_1–A_4 by demanding that l_2 satisfy a linear growth condition, then the results of § 2 apply and so (4.5) is also valid for the constrained case, i.e., $k_0 \geqq 1$, so that (c_x is now a matrix)

$$p_i(t,x) = \hat{E}_{tx}\left\{ \alpha^* c_x(x(T))\Phi(T,t) + \alpha_0 \int_t^T l_x(s,x(s),\hat{u}(s,x(s)))\Phi(s,t)\,ds \right\}_i, \qquad i > \nu.$$

5. Finite domains. Let G be bounded with boundary ∂G of class C^2, and set $Q = (0,T)\times G$. Again, we begin by considering the unconstrained case, i.e., $k_0 = 0$. If one proceeds as in the previous section, some difficulties arise in discussing the dependence of τ on the initial point x or $x+z$ in Lemma 4.2. Instead, we shall restrict the discussion to the smooth nondegenerate case where we can apply an idea due to Davis [2].

(A_5) σ is in $C^{1,2}(\bar{Q})$; i.e., the partial derivatives σ_t, σ_{xx} exist and are continuous on $\bar{Q}$; $\sigma(t,x)^{-1}$ exists on $\bar{Q}$.

(A_6) f and l are in $C^{1,1}(\bar{Q}\times U)$; i.e., f_t, f_x, l_t, l_x exist and are continuous on $\bar{Q}\times U$ as are f, l.

(A_7) $c(T,x)$ is in $C^2(G)$, and $c(t,x) = \tilde{c}(t,x)$ on $[0,T]\times\partial G$, where $\tilde{c}$ is a function in $C^{1,2}(\bar{Q})$.

Observe that A_1, A_5–A_7 imply that V is in $C^{1,2}(Q)\cap C^{0,1}(\bar{Q})$, [5], and A_4 is satisfied with $\phi \equiv 0$. Let us write $\partial^* Q$ for $\{T\}\times G \cup [0,T]\times\partial G$.

We now apply the argument of Davis to obtain a representation of V_x. As we shall need to find $d\{V_x\Phi\}$, we shall need an expression for dV_x, i.e., for $V_{xt} + \mathscr{L}V_x$ where $\mathscr{L}$ is the differential generator of x under $\hat{P}$; i.e.,

$$\mathscr{L}V = \frac{1}{2}\sum_{i,j=1}^{d} a_{ij}(t,x)V_{x_ix_j} + \sum_{i=1}^{d} f_i(t,x,\hat{u}(t,x))V_{x_i},$$

where $a_{ij}(t,x) = [\sigma(t,x)\sigma(t,x)^*]_{ij}$. Since we have

$$V_t + \mathscr{L}V + \hat{l} = 0, \qquad V(t,x) = c(t,x) \quad \text{for } (t,x)\in\partial^* Q, \tag{5.1}$$

we might just differentiate with respect to x, except that we do not know that V_{tx}, V_{xxx} exist! Also, to apply Ito's lemma, we would require V_x to be $C^{1,2}$, and in fact it is only in $W_\lambda^{1,2}(Q)$, $\lambda < \infty$; i.e., V_{xt} and V_{xxx} are in $L_\lambda(Q)$. Thus, an approximation argument due to Rishel [5] will be needed.

Let us write $\hat{\phi}(t,x)$ for $\phi(t,x,\hat{u}(t,x))$ for any function ϕ, but note that

$$\hat{\phi}_x(t,x) \equiv \phi_x(t,x,\hat{u}(t,x)) \neq \frac{\partial\hat{\phi}}{\partial x}(t,x) \equiv \phi_x(t,x,\hat{u}(t,x)) + \phi_u(t,x,\hat{u}(t,x))\hat{u}_x(t,x).$$

Now set $H(t,x,p) = \max_{u\in U}\tilde{H}(t,x,u,p)$, cf. (1.5). The following lemma is crucial.

LEMMA 5.1. *Assume* A_1, A_5–A_7. *Then H is differentiable almost everywhere and, for any $k = 1, 2, \cdots, d$,*

$$\frac{\partial}{\partial x_k} H(t, x, -V_x(t, x)) = -\sum_i V_{x_i x_k}(t, x)\hat{f}_i(t, x) - V_x \hat{f}_{x_k}(t, x) - \hat{l}_{x_k}(t, x) \quad a.e. \tag{5.2}$$

Proof. Since V is in $C^{1,2}(Q) \cap C^{0,1}(\bar{Q})$ then as in [4, (2.6)] for $K \subset Q$, K compact, there exists C_K such that if $x, \tilde{x} \subset K$, then

$$|H(t, x, -V_x(t, x)) - H(t, \tilde{x}, -V_x(t, \tilde{x}))| \leqq C_K |x - \tilde{x}|.$$

Hence, if we set $h(t, x) \equiv H(t, x, -V_x(t, x))$ then $\partial h(t, x)/\partial x_k$ exists for $x_k \notin N_k$, a Lebesgue null set depending on t, $x_1, \cdots, x_{k-1}, x_{k+1}, \cdots, x_d$. Moreover, $\{(t, x)\colon \partial h(t, x)/\partial x_k \text{ exists}\}$ is measurable, cf. [18, p. 294, ex. i], and so by Fubini's theorem $\partial h/\partial x_k$ exists a.e. For such (t, x), (5.2) follows as in [4, Lemma 2.1]. The importance of this result is that because we are using the optimal control $\hat{u}$, then no derivatives with respect to u need be taken.

Next we differentiate (5.1) to obtain an equation for $V_{x_k} \equiv V^k$ as a *Schwartz distribution*; moreover, this equation will have a unique solution. Then we shall show that the equation even has a solution in $W_\lambda^{1,2}(Q)$ for any $\lambda < \infty$, so that this solution must be V^k, i.e., $V^k \in W_\lambda^{1,2}(Q)$.

LEMMA 5.2. *Assume* A_1, A_5–A_7 *and that $G' \subset G$ is compact, $\phi \in C^2$, the support of $\phi(t, \cdot)$ is contained in G', and $\phi(0, x) = 0$. Then*

$$\begin{aligned} &\int_0^T \int_G -V^k \phi_t - \sum_i \left(\frac{1}{2}\sum_j a_{ij} V^k_{x_j}\right)\phi_{x_i} \\ &\quad -\left\{\frac{1}{2}\sum_{ij}[(a_{ij})_{x_i} V^k_{x_j} - (a_{ij})_{x_k} V^i_{x_j}] - \sum_i [\hat{f}_i V^k_{x_i} + (\hat{f}_{x_k})_i V^i] - \hat{l}_{x_k}\right\}\phi \, dx \, dt \\ &\qquad + \int_G c_{x_k}(T, x)\phi(T, x) \, dx = 0. \end{aligned} \tag{5.3}$$

Proof.

$$\begin{aligned} \int_0^T \frac{\partial}{\partial t} \int_G V\phi_{x_k} \, dx \, dt &= \int_G V\phi_{x_k} \, dx \Big|_{t=0}^{t=T} \\ &= -\int_G V_{x_k}\phi \, dx \Big|_0^T + \int_G \frac{\partial}{\partial x_k}(V\phi) \, dx \Big|_0^T \\ &= -\int_G c_{x_k}(T, x)\phi(T, x) \, dx, \end{aligned}$$

since $\phi(0, x) = 0$ and $\phi(t, x) = 0$ for $x \in \partial G$. But the first integral is also equal to

$$\begin{aligned} \int_0^T \frac{\partial}{\partial t} \int_{G'} V\phi_{x_k} \, dx \, dt &= \int_0^T \int_{G'} \frac{\partial}{\partial t}(V\phi_{x_k}) \, dx \, dt \\ &= \int_0^T \int_{G'} V_t\phi_{x_k} + V\phi_{x_k t} \, dx \, dt \\ &= \int_0^T \int_{G'} V_t\phi_{x_k} - V^k\phi_t \, dx \, dt. \end{aligned}$$

Since, on Q,

$$V_t + \frac{1}{2}\sum_{ij} a_{ij} V_{x_i x_j} + \sum_i \hat{f}_i V_{x_i} + \hat{l} = 0,$$

then

$$\int_0^T \int_G -V^k \phi_t - \left\{\frac{1}{2}\sum_{ij} a_{ij} V_{x_i x_j} + \sum_i \hat{f}_i V_{x_i} + \hat{l}\right\} \phi_{x_k} \, dx \, dt + \int_G c_{x_k}(T)\phi(T) \, dx = 0.$$

However,

$$\begin{aligned} \int_G a_{ij} V_{x_i x_j} \phi_{x_k} \, dx &= -\int_G V_{x_j}[(a_{ij}\phi)_{x_k} - (a_{ij})_{x_k}\phi]_{x_i} \, dx \\ &= \int_G V^k_{x_j}(a_{ij}\phi)_{x_i} - V^i_{x_j}(a_{ij})_{x_k}\phi \, dx \\ &= \int_G a_{ij} V^k_{x_j} \phi_{x_i} + [(a_{ij})_{x_i} V^k_{x_j} - (a_{ij})_{x_k} V^i_{x_j}]\phi \, dx, \end{aligned}$$

and so (5.3) follows from the above and (5.2).

COROLLARY 5.3. *V^k is the unique solution in $V_2(Q)$ of*

$$\begin{aligned} &\psi_t + \sum_i \left(\frac{1}{2}\sum_j a_{ij}\psi_{x_j}\right)_{x_i} - \frac{1}{2}\sum_{ij}(a_{ij})_{x_i}\psi_{x_j} + \Psi = 0, \\ &\psi(t, x) = V_{x_k}(t, x) \quad \text{for } (t, x) \in \partial^* Q, \end{aligned} \tag{5.4}$$

where

$$\Psi(t, x) = \frac{1}{2}\sum_{ij}(a_{ij})_{x_k} V_{x_i x_j} + \sum_i \hat{f}_i V_{x_i x_k} + \hat{f}_{x_k} \cdot V_x + \hat{l}_{x_k}.$$

Note that $V_2(Q)$ consists of those elements in $W_2^{0,1}(Q)$ for which $\operatorname{ess\,sup}_t (\int_G |\psi(t, x)|^2 \, dx)^{1/2} + (\iint_Q |\psi_x(t, x)|^2 \, dx \, dt)^{1/2} < \infty$, where $W_2^{0,1}(Q)$ is the Sobolev space of functions ψ in $L_2(Q)$ for which $\psi_x \in L_2(Q)$; cf. [13] for more details.

Proof. From (5.3), it follows that V^k satisfies (5.4). Since $V \in C^{1,2}(Q)$ and $V_x \in C(\bar{Q})$ then $V^k \in V_2(Q)$. $C(\bar{Q})$ is the space of continuous functions on $\bar{Q}$. Now [13, Chapt. 3, Thm. 3.1] gives the uniqueness.

LEMMA 5.4. *$V^k \in W_q^{1,2}(Q')$ for any $q < \infty$, where $Q' = G' \times (\delta, T - \delta)$ with G' open, $\bar{G}' \subset G$, $\delta > 0$.*

Proof. On Q', consider

$$u_t + \frac{1}{2}\sum_{ij} a_{ij} u_{x_i x_j} + \Psi = 0, \quad u = V_{x_k} \quad \text{on } \partial^* Q'. \tag{5.5}$$

According to the results in [5, Appendix E], $V_{x_k x_i}$ is Hölder continuous on $\partial^* Q'$, in x with parameter μ and in t with parameter $\mu/2$, where $\mu = 1 - (n+2)/\lambda$ and $\lambda > n + 2$ is arbitrary. It now follows from [13, Chapt. 4, Thm. 9.1], that (5.5) has a unique solution in $W_q^{1,2}(Q')$ for $q < \lambda/(n+2)$. If $q > n + 2$, i.e., $\lambda > (n+2)^2$, then u and u_x are in $C(\bar{Q}')$. Hence, $u \in V_2(Q')$ and so by the last corollary (restricted to Q') it follows that $u = V^k \in W_q^{1,2}(Q')$.

Using an approximation argument from [5] together with the ideas in [2], we have

THEOREM 5.5. *Assume* A_1, A_5–A_7. *Then*

$$
\begin{aligned}
&-p(t,x) = V_x(t,x) \\
&\quad = \hat{E}_{tx}\left\{ c_x(\tau, x(\tau))\Phi(\tau,t) + \int_t^\tau l_x(s, x(s), \hat{u}(s, x(s)))\Phi(s,t)\, ds \right\}.
\end{aligned} \tag{5.6}
$$

Proof. Since V is $C^{1,2}(Q)$, Ito's lemma implies $p = -V_x$. V^k can be approximated by C^∞ functions $\tilde{\psi}^m$ such that $\tilde{\psi}^m$, $\tilde{\psi}^m_x$ converge uniformly to V^k, V^k_x, and $\tilde{\psi}^m_t$, $\tilde{\psi}^m_{xx}$ converge in $L_q(Q')$ to V^k_t, V^k_{xx}. Then

$$
\begin{aligned}
d(\Phi_{kl}\tilde{\psi}^m) &= \sum_i (\hat{f}_{x_i})_k \Phi_{il}\tilde{\psi}^m\, dt + \sum_{ij} (\sigma_{kj})_{x_i}\Phi_{il}\tilde{\psi}^m\, d\hat{w}_j \\
&\quad + \Phi_{kl}\left\{ \tilde{\psi}^m_t + \frac{1}{2}\sum_{ij} a_{ij}\tilde{\psi}^m_{x_ix_j} + \sum_i \hat{f}_i\tilde{\psi}^m_{x_i} \right\} dt \\
&\quad + \Phi_{kl}\sum_{ij} \tilde{\psi}^m_{x_i}\sigma_{ij}\, d\hat{w}_j + \sum_{ijn} (\sigma_{kj})_{x_i}\Phi_{il}\tilde{\psi}^m_{x_n}\sigma_{nj}\, dt,
\end{aligned}
$$

and the boundedness of σ, σ_x, $\tilde{\psi}^m$, $\tilde{\psi}^m_x$ implies

$$
\begin{aligned}
\hat{E}_{tx} d(\Phi_{kl}\tilde{\psi}^m) = \hat{E}_{tx}\Big\{ &\Phi_{kl}\left[\tilde{\psi}^m_t + \frac{1}{2}\sum_{ij} a_{ij}\tilde{\psi}^m_{x_ix_j} + \sum_i \hat{f}_i\tilde{\psi}^m_{x_i} \right] \\
&+ \sum_i \Phi_{il}\left[(\hat{f}_{x_i})_k\tilde{\psi}^m + \sum_{jn} (\sigma_{kj})_{x_i}\sigma_{nj}\tilde{\psi}^m_{x_n} \right] \Big\}\, dt.
\end{aligned}
$$

By the same argument as in the proof of [5, Lemma V.11.2], it follows that for $t' > t > 0$, $\tau' = \inf\{s > t : x(s) \notin Q'\}$,

$$
\begin{aligned}
&\hat{E}_{tx}\{\Phi_{kl}(t' \wedge \tau', t) V^k(t' \wedge \tau', x(t' \wedge \tau'))\} \\
&\quad = \hat{E}_{tx}\Big\{ \Phi_{kl}(\tau', t) V^k(\tau', x(\tau')) + \int_{t \wedge \tau'}^{\tau'} -\Phi_{kl}\left[V^k_t + \frac{1}{2}\sum_{ij} a_{ij}V^k_{x_ix_j} + \sum_i \hat{f}_i V^k_{x_i} \right] \\
&\qquad\qquad - \sum_i \Phi_{il}\left[(\hat{f}_{x_i})_k V^k + \sum_{jn} (\sigma_{kj})_{x_i}\sigma_{nj} V^k_{x_n} \right] ds \Big\}.
\end{aligned}
$$

Now let $t' \to t$, use (5.5) and sum on k, to obtain

$$
\begin{aligned}
&V_{x_l}(t,x) \\
&\quad = \hat{E}_{tx}\Big\{ [V_x(\tau', x(\tau'))\Phi(\tau', t)]_l + \int_t^{\tau'} \sum_k \left[\frac{1}{2}\sum_{ij} (a_{ij})_{x_k} V_{x_ix_j} + V_x\hat{f}_{x_k} + \hat{l}_{x_k} \right]\Phi_{kl} \\
&\qquad\qquad - \sum_{ik} \left[(\hat{f}_{x_i})_k V_{x_k} + \sum_{jn} (\sigma_{kj})_{x_i}\sigma_{nj} V_{x_kx_n} \right]\Phi_{il}\, ds \Big\} \\
&\quad = \hat{E}_{tx}\left\{ V_x(\tau', x(\tau'))\Phi(\tau', t) + \int_t^{\tau'} \hat{l}_x(s, x(s))\Phi(s,t)\, ds \right\}_l.
\end{aligned}
$$

(5.6) is obtained in the limit as $Q' \uparrow Q$, and so the proof is complete.

It now follows from (2.8) that, in general ($k_0 \neq 0$), we have

$$p(t,x)=\hat{E}_{tx}\Big\{\alpha^* c_x(\tau,x(\tau))\Phi(\tau,t)+\alpha_0\int_t^\tau l_x(s,x(s),\hat{u}(s,x(s)))\Phi(s,t)\,ds\Big\}, \tag{5.7}$$

where α is given by the maximum principle, so $\alpha \neq 0$ and $\alpha_0 \leqq 0$.

Remark. We can now obtain the forward equation satisfied by p at least in the nondegenerate case, cf. (1.12), under the hypotheses A_1–A_3, A_5–A_7. In the proof of Theorem 5.5 we only looked at the drift term for $d(\Phi_{kl}\psi^m)$, but the same analysis shows that (with $t=0$, $x=x_0$) (since $\tilde{\psi}_x^m \to V_x^k$ uniformly)

$$V_{x_l}(0,x_0)=\Big\{c_x(\tau,x(\tau))\Phi(\tau,0)+\int_0^\tau \hat{l}_x(s,x(s))\Phi(s,0)\,ds\Big\}_l$$
$$-\sum_j\int_0^\tau\Big\{\sum_{ik}(\sigma_{kj})_{x_i}\Phi_{il}V_{x_k}+\Phi_{kl}V_{x_kx_i}\sigma_{ij}\Big\}\,d\hat{w}_j$$

in the space $L_1(\hat{P})$, or

$$V_x(0,x_0)=c_x(\tau,x(\tau))\Phi(\tau,0)+\int_0^\tau \hat{l}_x(s,x(s))\Phi(s,0)\,ds$$
$$-\sum_j\int_0^\tau (V_x\sigma^j)_x(s,x(s))\Phi(s,0)\,d\hat{w}_j$$
$$=-\bar{p}(0)-\sum_j\int_0^\tau (V_x\sigma^j)_x\Phi(s,0)\,d\hat{w}_j$$
$$=-E\bar{p}(0)-\sum_j\int_0^\tau[\psi^j(s)+(V_x\sigma^j)_x\Phi(s,0)]\,d\hat{w}_j.$$

The unbounded case ($\tau = T$) follows by using the domains $Q_l=(0,T)\times\{x:|x|\leqq l\}$ with boundary condition $V(\tau,x(\tau))$ and then letting $l\to\infty$. Hence, we have

$$\psi^j(s)=-(V_x\sigma^j)_x(s,x(s))\Phi(s,0) \quad \text{with probability 1.}$$

But $p(s)=-V_x(s,x(s))$ and

$$\psi^j(s)\Psi(s,0)=-(V_x\sigma^j)_x(s,x(s)),$$

so that (1.12) written in column form gives

$$dp^*=-\Big\{\hat{f}_x^*(t,x(t))p^*-\sum_k[\sigma_x^k(t,x(t))]^*V_{xx}(t,x(t))\sigma^k(t,x(t))-\hat{l}_x(t,x(t))^*\Big\}\,dt$$
$$-V_{xx}(t,x(t))\sigma(t,x(t))\,d\hat{w}, \tag{5.8}$$
$$p(\tau)=-c_x(\tau,x(\tau)).$$

Note that if we think of p in feedback form, i.e., $p(t,x)=-V_x(t,x)$, then we might write $V_{xx}(t,x(t))$ as $-p_x(t,x(t))$. If σ is independent of x, then the drift term alone gives the deterministic adjoint equation: $\dot{p}^*=-\hat{f}_x^*p^*+\hat{l}_x^*$. Finally, if $\sigma=0$, i.e., the problem is deterministic, then (5.8) reduces to the deterministic adjoint equation. Hence, it seems very likely that (5.8) also holds in the degenerate case. Note that in the definition of V and in (5.8), we simply replace c and l by $-\alpha^*c$ and $-\alpha_0 l$ (although α is unknown) in case constraints are present, i.e., $k_0 \neq 0$.

REFERENCES

[1] J. M. BISMUT, *Théorie probabiliste du contrôle des diffusions*, Mem. Amer. Math. Soc., 4 (1976), No. 167.

[2] M. H. A. DAVIS, *Functionals of diffusion processes as stochastic integrals*, Math. Proc. Cambridge Phil. Soc., 87 (1980), pp. 157–166.

[3] W. H. FLEMING, *Optimal control of partially observable diffusion processes*, this Journal, 6 (1968), pp. 194–214.

[4] ———, *Stochastic control for small noise intensities*, this Journal, 9 (1971), pp. 473–517.

[5] W. H. FLEMING AND R. W. RISHEL, *Deterministic and Stochastic Control*, Springer-Verlag, New York, 1975.

[6] A. FRIEDMAN, *Stochastic Differential Equations*, Academic Press, New York, 1975.

[7] U. G. HAUSSMANN, *General necessary conditions for optimal control of stochastic systems*, Math. Programming Stud., 6 (1976), pp. 30–48.

[8] ———, *On the stochastic maximum principle*, this Journal, 16 (1978), pp. 236–251.

[9] ———, *Functionals of Ito processes as stochastic integrals*, this Journal, 16 (1978), pp. 252–269.

[10] ———, *On the integral representation of functionals of Ito processes*, Stochastics, 3 (1979), pp. 17–27.

[11] ———, *Some examples of optimal stochastic controls*, SIAM Rev., 23 (1981), to appear.

[12] H. J. KUSHNER, *Necessary conditions for continuous parameter stochastic optimization problems*, this Journal, 10 (1972), pp. 550–565.

[13] O. A. LADYSHENSKAYA, V. A. SOLONNIKOV AND N. N. URAL'SEVA, *Linear and Quasilinear Equations of Parabolic Type*, American Mathematical Society, Providence, RI, 1968.

[14] D. W. STROOCK AND S. R. S. VARADHAN, *Diffusion processes with continuous coefficients*, Comm. Pure Appl. Math., 22 (1969), pp. 345–400, 479–530.

[15] W. H. FLEMING, *The Cauchy problem for degenerate parabolic equations*, J. Math. Mech., 13 (1964), pp. 987–1008.

[16] ———, *Duality and a priori estimates in Markovian optimization problems*, J. Math. Anal. Appl., 16 (1966), pp. 254–279.

[17] R. W. RISHEL, *Weak solutions of a partial differential equation of dynamic programming*, this Journal, 9 (1971), pp. 519–528.

[18] M. E. MUNROE, *Introduction to Measure and Integration*, Addison-Wesley, Reading, MA, 1953.

[19] V. E. BENEŠ, *Existence of optimal stochastic controls*, this Journal, 9 (1971), pp. 446–472.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 2 March 1981

0363-0129/81/1902-0005/$01.00/0

NONLINEAR FILTERING FORMULAS FOR DISCRETE-TIME OBSERVATIONS*

YOSHIKI TAKEUCHI† AND HAJIME AKASHI†

Abstract. This paper presents two types of nonlinear filtering formulas in the form of differential equations for the case where the signal is a continuous-time and the observation is a discrete-time process. The observation is corrupted by additive Gaussian white noise. The method of solution is based on Girsanov's measure transformation technique and a family of probability measures is introduced which is indexed by the continuous-time parameter. By computing the time evolution of these measures, the conditional expectation of a functional of the signal, given the observations, with respect to the original measure is smoothly updated. The obtained formulas are recursive with respect to the observation sequence whereas the well-known Bayes' formula is nonrecursive in the general case considered.

1. Introduction and summary. This paper is concerned with the continuous-discrete nonlinear filtering problem within the additive white Gaussian noise framework: Let x_t, $t \in [0, T]$ be a continuous-time signal process which takes values in a complete separable metric space S. Suppose that at each $t_j \in [0, T]$, $1 \leqq j \leqq N$, we have an observation $z_j \in \mathbb{R}^m$ related with x_t, $t \leqq T$ such that

$$z_j = h_j + v_j, \tag{1}$$

where h_j is a certain functional of x_s, $s \leqq t_j$ and z_k, $k \leqq j-1$; and v_j, $j \leqq N$ is a zero-mean white Gaussian noise sequence such that

$$Ev_j v_k' = \delta_{jk} R(j), \tag{2}$$

and for all j, v_j is independent of x_s, $s \leqq t_j$ and z_k, $k \leqq j-1$. At each time $t \in [t_j, t_{j+1})$, we wish to compute the minimal variance filtered estimate of $f(x_t)$ based on z_k, $k \leqq j$, i.e., the conditional expectation $E[f(x_t)|z_k, k \leqq j]$, for a suitable real-valued function f on S.

For the corresponding continuous-time problem, quite general formulas are applicable [1]–[7]. In particular, Fujisaki et al. [1] derived a stochastic differential equation for the filter which can be applied in most cases. As is pointed out in [1], a generalized Bayes' formula [9] is useful in applications only when t is fixed, because the estimate at a future time can not be computed without using all the past data. The advantage of the filtering formula in the form of a stochastic differential equation lies in the recursive structure with respect to the observed data.

Although the continuous-time problems are of much theoretical interest, discrete-time observations are more convenient in many applications because of digital computer implementations. For the continuous-discrete case described above, it is also desirable to obtain a formula which is recursive in the observation z_j. By applying the classical Bayes' formula, Jazwinski [8] obtained such a recursive formula. However, his result is not applicable unless x_t is a Markov process in $\mathbb{R}^n$, h_j is simply a function of x_{t_j}, i.e., $h_j = h_j(x_{t_j})$, and v_j, $j \leqq N$ is completely independent of x_t, $t \leqq T$.

This paper presents a new method of updating the conditional expectation from $E[f(x_{t_j})|z_k, k \leqq j-1]$ or $E[f(x_{t_{j-1}})|z_k, k \leqq j-1]$ to $E[f(x_{t_j})|z_k, k \leqq j]$ via differential equations which are applicable for the general class of continuous-discrete nonlinear filtering problems. First, a family of probability measures: $\hat{P}_t$, $t \in [0, T]$ is introduced by

* Received by the editors January 2, 1980, and in revised form June 2, 1980.
† Department of Precision Mechanics, Faculty of Engineering, Kyoto University, Kyoto 606, Japan.

a measure transformation on the original probability measure P. Roughly speaking, the measure $\hat{P}_t$ for $t \in [t_{j-1}, t_j]$ is defined in such a way that the conditional expectation $E^t[f(x_{t_j})|z_k, k \leqq j]$ with respect to $\hat{P}_t$ is the estimate of $f(x_{t_j})$ under the hypothesis that z_k, $k \leqq j-1$ are given by (1) and the jth observation z_j is replaced by

$$z_j^t = \int_{t_{j-1}}^{t} u_s \, ds + v_j, \tag{3}$$

where u_t, $t \in [0, T]$ is a measurable process with the property

$$h_j = \int_{t_{j-1}}^{t_j} u_s \, ds, \qquad 1 \leqq j \leqq N, \tag{4}$$

(see Definition 2 and Lemma 2 for details). Defining $\hat{P}_t$ in this way, we have $E^{t_{j-1}}[f(x_{t_j})|z_k, k \leqq j] = E[f(x_{t_j})|z_k, k \leqq j-1]$ since $z_j^{t_{j-1}}$ is composed of no signal ($z_j^{t_{j-1}}$ is the value of z_j^t at $t = t_{j-1}$). On the other hand, we also have $E^{t_j}[f(x_{t_j})|z_k, k \leqq j] = E[f(x_{t_j})|z_k, k \leqq j]$ since $z_j^{t_j} = z_j$. Consequently, by computing the evolution of $E^t[f(x_{t_j})|z_k, k \leqq j]$ with respect to t, a formula is obtained by which the estimate is updated from $E[f(x_{t_j})|z_k, k \leqq j-1]$ to $E[f(x_{t_j})|z_k, k \leqq j]$ (see (34) in Theorem 1). Similarly, by computing the evolution of $E^t[f(x_t)|z_k, k \leqq j]$, we have another formula which directly updates the estimate from $E[f(x_{t_{j-1}})|z_k, k \leqq j-1]$ to $E[f(x_{t_j})|z_k, k \leqq j]$ ((45) in Theorem 2).

This paper is organized as follows. In § 2, notation, definitions and technical conditions are given to formulate the filtering problem precisely. Section 3 is devoted to presenting preliminary lemmas concerned with the measure transformation and the properties of the transformed measures. These results are applied in § 4 to derive two types of nonlinear filtering formulas (Theorems 1 and 2). Two examples are given in § 5 for the better understanding of the results.

2. Notation and definitions.

General notation. Throughout this paper, column vectors are denoted by lower case letters and matrices are denoted by capital letters. The identity matrix of any dimension is I. The prime denotes the transpose of a vector or a matrix. The Euclidean norm is denoted by $|\cdot|$. The trace of a square matrix A is tr $[A]$, and if A is nonsingular, A^{-1} denotes the inverse matrix of A. The triplet $(\Omega, \mathcal{F}, P)$ is a complete probability space where Ω is a sample space with elementary events ω, $\mathcal{F}$ is a σ-algebra of the subsets of Ω, and P is a probability measure. E and $E[\cdot|\mathcal{G}]$, $\mathcal{G} \subset \mathcal{F}$ denote respectively the expectation and the conditional expectation, given $\mathcal{G}$, with respect to P. $\sigma\{\cdot\}$ is the minimal sub-σ-algebra of $\mathcal{F}$ with respect to which the family of $\mathcal{F}$-measurable random variables $\{\cdot\}$ is measurable. If $\mathcal{F}_1$ and $\mathcal{F}_2$ are sub-σ-algebras of $\mathcal{F}$, then $\mathcal{F}_1 \vee \mathcal{F}_2$ denotes the minimal σ-algebra containing both $\mathcal{F}_1$ and $\mathcal{F}_2$. From now on, it is assumed that each sub-σ-algebra contains all null sets in $\mathcal{F}$.

Signal process. Let x_t, $0 \leqq t \leqq T$ be a stochastic process on $(\Omega, \mathcal{F}, P)$ which describes the signal or the state process of interest and takes values in a complete separable metric space S. (In particular, S could be the n-dimensional Euclidean space $\mathbb{R}^n$.) The only major assumption on x_t, $t \leqq T$ is that the space $\mathbb{D}^*(\mathcal{A})$ (see Definition 1 below) is nonempty.

Observations. Let t_j, $1 \leqq j \leqq N$ be a finite set in [0, T] such that $0 < t_1 < \cdots < t_N = T$. At each discrete time point t_j, we have the observation z_j given by (1). For convenience,

let $t_0 = 0$ and $z_0 = 0$. Define σ-algebras:

$$\mathscr{Z}_j = \sigma\{z_k;\ k \leqq j\}, \qquad 0 \leqq j \leqq N,$$

$$\mathscr{G}_j = \sigma\{x_s;\ s \leqq t_{j+1}\} \vee \sigma\{z_k;\ k \leqq j\}, \qquad 0 \leqq j \leqq N-1,$$

$$\mathscr{G}_N = \sigma\{x_s;\ s \leqq T\} \vee \sigma\{z_k;\ k \leqq N\},$$

$$\mathscr{F}_t = \sigma\{x_s;\ s \leqq t\} \vee \sigma\{z_k;\ k \leqq j-1\}, \qquad t \in [t_{j-1}, t_j), \quad 1 \leqq j \leqq N,$$

$$\mathscr{F}_T = \mathscr{G}_N.$$

For the observation (1), we will assume the following conditions:

(C–1) $h_j, j \leqq N$ is adapted to $\mathscr{G}_{j-1}$; i.e., $h_j(\omega) = h_j(x_s, s \leqq t_j, z_k, k \leqq j-1)$.

(C–2) For each j, $v_r, r \geqq j$ are independent of $\mathscr{G}_{j-1}$.

(C–3) For all j, $R(j)$ is nonsingular with bounded elements.

(C–4) $P\{\omega; \sum_{j=1}^{N} |R^{-1/2}(j)h_j|^2 < \infty\} = 1$.

Now, let us define a class of functions f on S.

DEFINITION 1.[1] Let f be a real-valued measurable function on S such that

$$E|f(x_t)|^2 < \infty \quad \text{for all } t \in [0, T].$$

The function f is said to belong to space $\mathbb{D}^*(\mathscr{A})$ if there exists a jointly (t, ω)-measurable real-valued function $\mathscr{A}_t f(\omega)$ adapted to $\mathscr{F}_t$ such that

$$\int_0^T E|\mathscr{A}_t f|^2\, dt < \infty,$$

and

$$\mathscr{M}_t(f) = f(x_t) - Ef(x_0) - \int_0^t \mathscr{A}_s f\, ds \tag{5}$$

is an $(\mathscr{F}_t, P)$-martingale.

It should be noted that in the special case where x_t is a Markov process, $\mathscr{A}_t$ is the generator given by

$$\mathscr{A}_t f(x_t) = \lim_{s \downarrow 0} \frac{1}{s}\{E[f(x_{t+s})|\sigma\{x_t\}] - f(x_t)\}. \tag{6}$$

From now on, it is assumed that $\mathbb{D}^*(\mathscr{A})$ is nonempty.

3. Preliminary lemmas. This section is devoted to presenting some preliminary lemmas which are necessary to obtain the main results (Theorems 1 and 2) of this paper. These lemmas are concerned with the properties of new measures introduced on $(\Omega, \mathscr{F})$.

Let us define a sequence $\zeta_j, 0 \leqq j \leqq N$ by

$$\begin{aligned} &\zeta_j = \sum_{k=1}^{j} [-h_k' R^{-1}(k) v_k - \tfrac{1}{2} h_k' R^{-1}(k) h_k], \qquad 1 \leqq j \leqq N, \\ &\zeta_0 = 0. \end{aligned} \tag{7}$$

First, let us describe Kunita's result [2] in a slightly generalized form.

[1] $\mathbb{D}^*(\mathscr{A})$ is the continuous-discrete time analogue of the class $\mathbb{D}(\mathscr{A})$ introduced by Fujisaki et al. [1] for the continuous-time case.

LEMMA 1. (Kunita [2, Lemma 2.14]). *In addition to* (C–1)–(C–4), *assume*

(C–5) $E \exp(\zeta_N) = 1$,

and

(C–6) $P\{\omega; \exp(\zeta_N) = 0\} = 0$.

Let us define $\tilde{P}$ *on* $(\Omega, \mathcal{F})$ *by*

$$\tilde{P}(A) = \int_A \exp(\zeta_N)\, dP, \qquad A \in \mathcal{F}. \tag{8}$$

Then:

(i) $\tilde{P}$ *is a probability measure on* $(\Omega, \mathcal{F})$ *and* $\tilde{P} \sim P$; *i.e.,* $\tilde{P}$ *is mutually absolutely continuous with respect to* P.

(ii) z_j, $j \leqq N$, *with respect to* $\tilde{P}$, *is a zero-mean independent Gaussian sequence with the covariance* $\tilde{E} z_j z_k' = \delta_{jk} R(j)$ (*where* $\tilde{E}$ *denotes the expectation with respect to* $\tilde{P}$).

(iii) z_k, $k \geqq j$, *with respect to* $\tilde{P}$, *are independent of* $\mathcal{G}_{j-1}$.

Proof. It is obvious from (C–5) and (8) that $\tilde{P}$ is a probability measure and that $\tilde{P} \ll P$ (i.e., $\tilde{P}$ is absolutely continuous with respect to P). Then, by (C–6), $P \ll \tilde{P}$ follows from [3, Lemma 6.8]. Hence $\tilde{P} \sim P$, and (i) is proved. The proof of (ii) and (iii) is the same as for [2, Lemma 2.14]. Since [2] is written in Japanese, the proof of Lemma 1 (ii) and (iii) is given, for convenience, in the Appendix.

Now we introduce a continuous-time process u_t, $0 \leqq t \leqq T$ which generates interpolation of the sequence h_j, $j \leqq N$.

DEFINITION 2. Let $u \equiv \{u_t; 0 \leqq t \leqq T\}$ be an $\mathbb{R}^m$-valued measurable process. The process u is said to belong to $\mathbb{I}(h)$ if

(i) for all $1 \leqq j \leqq N$,

$$h_j = \int_{t_{j-1}}^{t_j} u_s\, ds; \tag{9}$$

(ii) for all $1 \leqq j \leqq N$ and $t \in [t_{j-1}, t_j)$, u_t is $\mathcal{G}_{j-1}$-measurable;

(iii) there exists a constant $1 \leqq K \leqq \to < \infty$ such that for all $1 \leqq j \leqq N$,

$$\left\{ \sup_{t_{j-1} \leqq t < t_j} |u_t| \right\} (t_j - t_{j-1}) \leqq K |h_j| \qquad P\text{-a.s.}; \tag{10}$$

and

(iv) for all $1 \leqq j \leqq N$ and $t \in [t_{j-1}, t_j]$,

$$\left| R^{-1/2}(j) \left(\int_{t_{j-1}}^{t} u_s\, ds \right) \right| \leqq |R^{-1/2}(j) h_j| \qquad P\text{-a.s.} \tag{11}$$

The process u is said to belong to $\mathbb{I}_0(h)$ if there exist $v \in \mathbb{I}(h)$ and $t \in [0, T]$ such that

$$u_s = \begin{cases} v_s & \text{for } 0 \leqq s < t, \\ 0 & \text{for } t \leqq s \leqq T. \end{cases} \tag{12}$$

Remark 1. It is clear that $\mathbb{I}(h) \subset \mathbb{I}_0(h)$. Note that $\mathbb{I}(h)$ is nonempty; i.e., there always exists at least one process u which belongs to $\mathbb{I}(h)$ because

$$u_t = \frac{h_j}{(t_j - t_{j-1})}, \qquad t_{j-1} \leqq t < t_j, \qquad 1 \leqq j \leqq N$$

satisfies the conditions (i)–(iv).

For $u \in \mathbb{I}_0(h)$, let

$$h_j(u) = \int_{t_{j-1}}^{t_j} u_s \, ds, \qquad 1 \leqq j \leqq N, \tag{13}$$

and define $\phi_j(u)$ by

$$\begin{aligned} &\phi_j(u) = \exp\left\{ \sum_{k=1}^{j} [h_k'(u)R^{-1}(k)z_k - \tfrac{1}{2}h_k'(u)R^{-1}(k)h_k(u)] \right\}, \qquad 1 \leqq j \leqq N, \\ &\phi_0(u) = 1. \end{aligned} \tag{14}$$

Note that if $u \in \mathbb{I}(h)$, then $h_j(u) = h_j$, and hence $\phi_j(u) = \{\exp(\zeta_j)\}^{-1}$. Thus, $P = \int \phi_N(u)\, d\tilde{P}$ for $u \in \mathbb{I}(h)$.

Let $u \in \mathbb{I}(h)$ be fixed. For $t \in [0, T]$, let $\tilde{u}^t \equiv \{\tilde{u}_s^t;\ 0 \leqq s \leqq T\}$ be the process defined by

$$\tilde{u}_s^t = \begin{cases} u_s & \text{for } 0 \leqq s < t, \\ 0 & \text{for } t \leqq s \leqq T. \end{cases} \tag{15}$$

Then, $\tilde{u}^t \in \mathbb{I}_0(h)$. Note that $\phi_j(\tilde{u}^t)$ is a $(\mathcal{G}_j, \tilde{P})$-martingale with $\tilde{E}\phi_N(\tilde{u}^t) = 1$. Let us introduce a family of measures: $\hat{P}_t$, $0 \leqq t \leqq T$ by

$$\hat{P}_t(A) = \int_A \phi_N(\tilde{u}^t)\, d\tilde{P}, \qquad A \in \mathcal{F}. \tag{16}$$

Then, each $\hat{P}_t$ is clearly a probability measure on $(\Omega, \mathcal{F})$. The following lemma describes the properties of $\hat{P}_t$.

LEMMA 2. *Assume* (C-1)-(C-6). *If* $u \in \mathbb{I}(h)$, *then, for* $t \in [t_{j-1}, t_j]$, $\hat{P}_t$ *has the following properties*:

(i) $\hat{P}_t \sim \tilde{P}$ *and* $\hat{P}_t \sim P$.

(ii) *The distribution of* $\{x_s;\ s \leqq t_j,\ z_k;\ k \leqq j-1\}$ *with respect to* $\hat{P}_t$ *is equivalent to the one with respect to* P. *Furthermore,*

$$\hat{P}_t(A) = P(A) \quad \text{for all } A \in \mathcal{G}_{j-1}.$$

(iii) z_k, $k \geqq j+1$, *with respect to* $\hat{P}_t$, *is a zero-mean independent Gaussian sequence with covariance* $E^t z_k z_r' = \delta_{kr}R(k)$, $k, r \geqq j+1$, *and is independent of* $\mathcal{G}_j$ (*where* E^t *is the expectation with respect to* $\hat{P}_t$). *Furthermore, for each* $\mathcal{B}(\subset \mathcal{F})$ *which is independent of* $\mathcal{G}_j$ *with respect to* $\tilde{P}$,

$$\hat{P}_t(A) = \tilde{P}(A) \quad \text{for all } A \in \mathcal{B}.$$

(iv) *The distribution of* $z_1, z_2, \cdots, z_N$ *with respect to* $\hat{P}_t$ *is equivalent to that of* $z_1^t, z_2^t, \cdots, z_N^t$ *with respect to* P; *i.e.*,

$$E^t \exp\left(i \sum_{k=1}^{N} \xi_k' z_k \right) = E \exp\left(i \sum_{k=1}^{N} \xi_k' z_k^t \right), \qquad \xi_k \in \mathbb{R}^m, \quad i = \sqrt{-1}, \tag{17}$$

where

$$z_k^t = h_k(\tilde{u}^t) + v_k. \tag{18}$$

(v) $z_j - h_j(\tilde{u}^t)$, *with respect to* $\hat{P}_t$, *is a zero-mean Gaussian vector with covariance* $R(j)$ *and is independent of* $\mathcal{G}_{j-1}$; *i.e., for all* $A \in \mathcal{G}_{j-1}$,

$$E^t I_A \exp(i\xi_j'\{z_j - h_j(\tilde{u}^t)\}) = \hat{P}_t(A) \exp(-\tfrac{1}{2}\xi_j' R(j)\xi_j).$$

Proof. First, (i) is clear by (C-6), [3, Lemma 6.8] and Lemma 1 (i). Let $\rho_j = \exp(\zeta_j)$. Then, since $\phi_j(\tilde{u}^t)$ is a $(\mathcal{G}_j, \tilde{P})$-martingale and since $\phi_{j-1}(\tilde{u}^t) = \phi_{j-1}(u) = (\rho_{j-1})^{-1}$, we

have

$$\hat{P}_t(A) = \int_A \tilde{E}[\phi_N(\tilde{u}^t)|\mathcal{G}_{j-1}]\, d\tilde{P}$$

$$= \int_A (\rho_{j-1})^{-1}\, d\tilde{P} = P(A) \quad \text{for all } A \in \mathcal{G}_{j-1},$$

where the third equality follows from Lemma 1(i). This proves (ii).

To show (iii), note that $\phi_N(\tilde{u}^t)$ is equal to $\phi_j(\tilde{u}^t)$ and is $\mathcal{G}_j$-measurable. Then it follows from Lemma 1 (ii) and (iii) that for all $A \in \mathcal{G}_j$,

$$E^t I_A \exp\left\{ i \sum_{k=j+1}^{N} \xi_k' z_k \right\} = \tilde{E}\phi_j(\tilde{u}^t) I_A \exp\left\{ i \sum_{k=j+1}^{N} \xi_k' z_k \right\}$$

$$= (\tilde{E}\phi_j(\tilde{u}^t) I_A)\left(\tilde{E} \exp\left\{ i \sum_{k=j+1}^{N} \xi_k' z_k \right\} \right) \tag{19}$$

$$= \hat{P}_t(A) \prod_{k=j+1}^{N} \exp\{-\tfrac{1}{2}\xi_k' R(k) \xi_k\}.$$

This proves the first part of (iii). Noting that $\phi_j(\tilde{u}^t)$, with respect to $\tilde{P}$, is independent of $\mathcal{B}$, we obtain

$$\hat{P}_t(A) = \int_A \tilde{E}[\phi_j(\tilde{u}^t)|\mathcal{B}]\, d\tilde{P} = \tilde{P}(A) \quad \text{for all } A \in \mathcal{B}.$$

Hence, we have (iii).

Let us show (iv). Noting that $z_k^t = v_k$ for $k \geqq j+1$ and that v_k, $k \geqq j+1$, with respect to P, is independent of $\mathcal{G}_j$, we have

$$E \exp\left(i \sum_{k=1}^{N} \xi_k' z_k^t \right) = \mathrm{E} \exp\left(i \sum_{k=1}^{j} \xi_k' z_k^t \right) \cdot E \exp\left(i \sum_{k=j+1}^{N} \xi_k' v_k \right). \tag{20}$$

Let

$$\chi_j^t = \exp\{i\xi_j' h_j(\tilde{u}^t) - \tfrac{1}{2}\xi_j' R(j) \xi_j\}.$$

Then, since

$$E \exp\left(i \sum_{k=1}^{j} \xi_k' z_k^t \right) = EE[\exp\{i\xi_j' z_j^t\}|\mathcal{G}_{j-1}] \exp\left(i \sum_{k=1}^{j-1} \xi_k' z_k \right)$$

$$= E\chi_j^t \exp\left(i \sum_{k=1}^{j-1} \xi_k' z_k \right),$$

it can be seen from (20) that

$$E \exp\left(i \sum_{k=1}^{N} \xi_k' z_k^t \right) = E\chi_j^t \exp\left(i \sum_{k=1}^{j-1} \xi_k' z_k \right) E \exp\left(i \sum_{k=j+1}^{N} \xi_k' v_k \right). \tag{21}$$

On the other hand, it follows from (iii) that

$$E^t \exp\left(i \sum_{k=1}^{N} \xi_k' z_k \right) = E^t \exp\left(i \sum_{k=1}^{j} \xi_k' z_k \right) E^t \exp\left(i \sum_{k=j+1}^{N} \xi_k' z_k \right). \tag{22}$$

It can be seen from (19) with $A = \Omega$ that

$$E^t \exp\left(i \sum_{k=j+1}^{N} \xi_k' z_k\right) = \exp\left(-\frac{1}{2} \sum_{k=j+1}^{N} \xi_k' R(k) \xi_k\right) \tag{23}$$
$$= E \exp\left(i \sum_{k=j+1}^{N} \xi_k' v_k\right).$$

From (21), (22) and (23), we only have to show that

$$E^t \exp\left(i \sum_{k=1}^{j} \xi_k' z_k\right) = E \chi_j^t \exp\left(i \sum_{k=1}^{j-1} \xi_k' z_k\right). \tag{24}$$

Let

$$\tilde{m} = h_j(\tilde{u}^t) + iR(j)\xi_j,$$

and

$$\gamma = \exp\{\tilde{m}' R^{-1}(j) z_j - \tfrac{1}{2} \tilde{m}' R^{-1}(j) \tilde{m}\}.$$

If we note that $\phi_{j-1}(\tilde{u}^t) = \phi_{j-1}(u) = (\rho_{j-1})^{-1}$ and that $\tilde{E}[\gamma | \mathcal{G}_{j-1}] = 1$, it follows that

$$E^t \exp\left(i \sum_{k=1}^{j} \xi_k' z_k\right) = \tilde{E} \phi_j(\tilde{u}^t) \exp\left(i \sum_{k=1}^{j} \xi_k' z_k\right)$$

$$= \tilde{E} \gamma \phi_{j-1}(u) \chi_j^t \exp\left(i \sum_{k=1}^{j-1} \xi_k' z_k\right)$$

$$= \tilde{E} \tilde{E}[\gamma | \mathcal{G}_{j-1}] \phi_{j-1}(u) \chi_j^t \exp\left(i \sum_{k=1}^{j-1} \xi_k' z_k\right)$$

$$= \tilde{E} (\rho_{j-1})^{-1} \chi_j^t \exp\left(i \sum_{k=1}^{j-1} \xi_k' z_k\right)$$

$$= E \chi_j^t \exp\left(i \sum_{k=1}^{j-1} \xi_k' z_k\right).$$

Hence, we have (24), and consequently (iv).

Finally, to show (v), note that

$$\phi_j(\tilde{u}^t) \exp\left(i\xi_j'[z_j - h_j(\tilde{u}^t)]\right) = \phi_j(u) \bar{\gamma} \exp\left(-\tfrac{1}{2} \xi_j' R(j) \xi_j\right),$$

where

$$\bar{\gamma} = \exp\{\bar{m}' R^{-1}(j) v_j - \tfrac{1}{2} \bar{m}' R^{-1}(j) \bar{m}\},$$

and

$$\bar{m} = h_j(\tilde{u}^t) - h_j + iR(j)\xi_j.$$

Then, since $EI_A \bar{\gamma} = EI_A E[\bar{\gamma} | \mathcal{G}_{j-1}] = P(A) = \hat{P}_t(A)$ by virtue of assertion (ii), it follows that

$$E^t I_A \exp\left(i\xi_j'[z_j - h_j(\tilde{u}^t)]\right) = \tilde{E} I_A \phi_j(\tilde{u}^t) \exp\left(i\xi_j'[z_j - h_j(\tilde{u}^t)]\right)$$
$$= \tilde{E} I_A \phi_j(u) \bar{\gamma} \exp\left(-\tfrac{1}{2} \xi_j' R(j) \xi_j\right)$$
$$= EI_A \bar{\gamma} \exp\left(-\tfrac{1}{2} \xi_j' R(j) \xi_j\right)$$
$$= \hat{P}_t(A) \exp\left(-\tfrac{1}{2} \xi_j' R(j) \xi_j\right).$$

This completes the proof.

From now on, for simplicity of notation, we will use ϕ_j^t, h_j^t and f_t to denote respectively $\phi_j(\tilde{u}^t)$, $h_j(\tilde{u}^t)$ and $f(x_t)$.

By virtue of Lemma 2, the following lemma is immediately obtained.

LEMMA 3. *Assume* (C-1)-(C-6). *Let* $g = g(\omega)$ *be any* $\mathcal{G}_l$*-measurable random variable which is* $\hat{P}_t$*-integrable,* $t \in (t_{j-1}, t_j]$. *If* $u \in \mathbb{U}(h)$, *then*

$$E^t[g|\mathcal{Z}_k] = \frac{\tilde{E}[g\phi^t_{j\wedge(l\vee k)}|\mathcal{Z}_{(j\vee l)\wedge k}]}{\tilde{E}[\phi^t_{j\wedge(l\vee k)}|\mathcal{Z}_{(j\vee l)\wedge k}]}. \tag{25}$$

In particular, for $l = j$, $t = t_j$ *and* $k = j + r$, $0 \leqq r \leqq N - j$,

$$E^{t_j}[g|\mathcal{Z}_{j+r}] = E^{t_j}[g|\mathcal{Z}_j] = E[g|\mathcal{Z}_j]. \tag{26}$$

Remark 2. Note that if g is a $\mathcal{G}_{j-1}$-measurable random variable, then $\hat{P}_t$-integrability of g for $t \geqq t_{j-1}$ is implied by P-integrability of g. This is immediately seen from Lemma 2 (ii).

4. Nonlinear filtering formulas. This section presents our main results on the nonlinear filtering problem. The first formula is based on the following lemma.

LEMMA 4. *In addition to* (C-1)-(C-6), *assume*

$$\sum_{j=1}^{N} E|R^{-1/2}(j)h_j|^4 < \infty. \tag{C-7}^2$$

Let $g \equiv g(\omega)$ *be any real-valued* $\mathcal{G}_{j-1}$*-measurable random variable such that* $E|g|^2 < \infty$. *If* $u \in \mathbb{U}(h)$, *then if we define*

$$\hat{g}^t = E^t[g|\mathcal{Z}_j], \qquad t \in [t_{j-1}, t_j), \quad 1 \leqq j \leqq N, \tag{27}$$

$\hat{g}^t$ *satisfies*

$$\frac{d\hat{g}^t}{dt} = \widehat{\{g - \hat{g}^t\}u_t' R^{-1}(j)\{z_j - h_j^t\}}^t, \tag{28a}$$

with the conditions

$$\hat{g}^{t_{j-1}} = E[g(\omega)|\mathcal{Z}_{j-1}] \tag{28b}$$

and

$$\hat{g}^{t_j} = E[g(\omega)|\mathcal{Z}_j]. \tag{28c}$$

Proof. First, we will show that

$$\frac{d}{dt}\tilde{E}[\phi_j^t|\mathcal{Z}_j] = \tilde{E}\left[\frac{d\phi_j^t}{dt}\middle|\mathcal{Z}_j\right]. \tag{29}$$

For this purpose, it suffices to show

$$\tilde{E}\left|\frac{d\phi_j^t}{dt}\right| < \infty,$$

since $\tilde{E}|\phi_j^t| = \tilde{E}\phi_j^t = 1 < \infty$. Noting that $u \in \mathbb{U}(h)$ and applying Lemma 2 (ii) and (v), we

[2] (C-7) implies (C-4).

have

$$\tilde{E}\left|\frac{d\phi_j^t}{dt}\right| = \tilde{E}|\phi_j^t u_t' R^{-1}(j)\{z_j - h_j^t\}|$$
$$\leqq \tilde{K}\tilde{E}\phi_j^t |R^{-1/2}(j)h_j|\,|R^{-1/2}(j)\{z_j - h_j^t\}|$$
$$\leqq \tilde{K}(E^t|R^{-1/2}(j)h_j|^2)^{1/2}(E^t|R^{-1/2}(j)\{z_j - h_j^t\}|^2)^{1/2}$$
$$\leqq \tilde{K}(E|R^{-1/2}(j)h_j|^2)^{1/2}(E|R^{-1/2}(j)v_j|^2)^{1/2} < \infty.$$

where $\tilde{K} = K/(t_j - t_{j-1})$. Hence, we have (29).

Now noting that

$$\hat{g}^t = \frac{\tilde{E}[\phi_j^t g|\mathscr{Z}_j]}{\tilde{E}[\phi_j^t|\mathscr{Z}_j]},$$

we can see from (29) that

$$\frac{d\hat{g}^t}{dt} = \frac{\frac{d}{dt}\tilde{E}[\phi_j^t g|\mathscr{Z}_j]}{\tilde{E}[\phi_j^t|\mathscr{Z}_j]} - \hat{g}^t\widehat{u_t' R^{-1}(j)\{z_j - h_j^t\}}^t.$$

Then, if it is shown that

$$\frac{d}{dt}\tilde{E}[\phi_j^t g|\mathscr{Z}_j] = \tilde{E}\left[\frac{d\phi_j^t}{dt} g\middle|\mathscr{Z}_j\right], \tag{30}$$

we have (28a). For (30), it is sufficient to show that

$$\tilde{E}\left|\frac{d\phi_j^t}{dt} g\right| < \infty,$$

since $\tilde{E}|\phi_j^t g| = E^t|g| = E|g| < \infty$. If we note that $u \in \mathbb{U}(h)$ and that g and h_j are $\mathscr{G}_{j-1}$-measurable, it follows from Lemma 2 (ii) and (v) that

$$\tilde{E}\left|\frac{d\phi_j^t}{dt} g\right| = \tilde{E}\phi_j^t |g u_t' R^{-1}(j)\{z_j - h_j^t\}|$$
$$= E^t|g u_t' R^{-1}(j)\{z_j - h_j^t\}|$$
$$\leqq (E^t|g|^2)^{1/2}(E^t|u_t' R^{-1}(j)\{z_j - h_j^t\}|^2)^{1/2}$$
$$\leqq \tilde{K}(E|g|^2)^{1/2}(E^t|Qh_j|^4)^{1/4}(E^t|Q\{z_j - h_j^t\}|^4)^{1/4}$$
$$\leqq \tilde{K}(E|g|^2)^{1/2}(E|Qh_j|^4)^{1/4}(E|Qv_j|^4)^{1/4} < \infty,$$

where $Q = R^{-1/2}(j)$. Hence, we have (30) and consequently, (28a). This completes the proof.

Remark 3. By the above proof, $\hat{P}_t$-integrability of $d\hat{g}^t/dt$ has been shown.

According to Lemma 4, the estimate is updated from $E[f_{t_j}|\mathscr{Z}_{j-1}]$ to $E[f_{t_j}|\mathscr{Z}_j]$ by solving a differential equation. By making use of this fact, we have the following theorem which presents the first nonlinear filtering formula.

THEOREM 1. *Assume* (C-1)–(C-7), *Let*

$$\hat{f}(t) = \begin{cases} E[f_t|\mathscr{Z}_{j-1}] & \text{for } t \in (t_{j-1}, t_j), \quad 1 \leqq j \leqq N \\ E[f_t|\mathscr{Z}_j] & \text{for } t = t_j, \qquad\quad 0 \leqq j \leqq N. \end{cases} \tag{31}$$

If $f \in \mathbb{D}^*(\mathscr{A})$ and $u \in \mathbb{I}(h)$, then $\hat{f}(t)$ with the initial condition $\hat{f}(0) = Ef(x_0)$ is given as follows:

(i) *Between observations, i.e., for $t \in (t_{j-1}, t_j)$,*

$$\hat{f}(t) = \hat{f}(t_{j-1}) + \int_{t_{j-1}}^{t} E[\mathscr{A}_s f(\omega) | \mathscr{Z}_{j-1}]\, ds. \tag{32}$$

(ii) *At an observation at $t = t_j$,*

$$\hat{f}(t_j) = \hat{f}^{t_j}_{t_j}, \tag{33}$$

where $\hat{f}^{t_j}_{t_j}$ is determined by

$$\begin{aligned} &\frac{d\hat{f}^t_{t_j}}{dt} = \widehat{\{f_{t_j} - \hat{f}^t_{t_j}\} u_t' R^{-1}(j)\{z_j - h^t_j\}}^{t}, \\ &\hat{f}^{t_{j-1}}_{t_j} = \hat{f}(t_j -), \qquad t \in (t_{j-1}, t_j]. \end{aligned} \tag{34}$$

Proof. By applying Lemma 4 with $g = f_{t_j}$, we get (34) and (33) from (28a), (28b) and (28c). To show (32), let $\mathscr{M}_t = \mathscr{M}_t(f)$ for simplicity. Then we have

$$f_t = f_{t_{j-1}} + \int_{t_{j-1}}^{t} \mathscr{A}_s f\, ds + \mathscr{M}_t - \mathscr{M}_{t_{j-1}}, \qquad t > t_{j-1}. \tag{35}$$

Since

$$E[\mathscr{M}_t - \mathscr{M}_{t_{j-1}} | \mathscr{Z}_{j-1}] = E[E\{\mathscr{M}_t - \mathscr{M}_{t_{j-1}} | \mathscr{F}_{t_{j-1}}\} | \mathscr{Z}_{j-1}] = 0,$$

taking the conditional expectation $E[\cdot\,|\mathscr{Z}_{j-1}]$ on both sides of (35), we have (32). This completes the proof.

Remark 4. It should be noted that $E[f_s | \mathscr{Z}_{j-1}]$, $s > t_{j-1}$ has a continuous modification and hence,

$$E[f_{t_i} | \mathscr{Z}_{j-1}] = E[f_{t_j -} | \mathscr{Z}_{j-1}] = \hat{f}(t_j -).$$

Remark 5. In Theorem 1, the computation of the estimate $\hat{f}(t_j)$ is given by two steps, i.e., prediction and correction, respectively described by (32) and (34). Although $\hat{f}(t)$, for $t_{j-1} < t < t_j$, is the filtered estimate of $f_t \equiv f(x_t)$ in the sense that $\mathscr{Z}_{j-1}$ is the available data at t, it is also the predicted estimate in the sense that $\mathscr{Z}_{j-1}$ is the available data at t_{j-1}. Hence, $\hat{f}(t_j -)$ is the one-step prediction of f_{t_j} based on z_k, $k \leqq j-1$, and formula (32) is a predictor in this sense. On the other hand, formula (34) plays the role of corrector which updates the estimate from $\hat{f}(t_j -) = E[f_{t_j} | \mathscr{Z}_{j-1}]$ to $\hat{f}(t_j) = E[f_{t_j} | \mathscr{Z}_j]$.

According to Theorem 1, we must solve equations (32) and (34) over the same time interval by turns. If we are interested in computing $\hat{f}(t_j)$, $1 \leqq j \leqq N$ rather than $\hat{f}(t)$, $0 \leqq t \leqq T$, we can obtain a formula which directly updates the estimate from $\hat{f}(t_{j-1})$ to $\hat{f}(t_j)$ without using the predicted estimate but by computing the evolution of $\hat{f}^t_t$. In order to obtain this direct formula, we shall impose additional conditions (C–8) and (C–9) given below.

Let us start by describing the martingale property of $M(f)$ with respect to $\tilde{P}$.

LEMMA 5. *Assume* (C–1)–(C–6). *Let*

$$\begin{aligned} \mathscr{F}^+_t &= \sigma\{x_s;\, s \leqq t\} \vee \sigma\{z_k;\, k \leqq j\} \\ &= \mathscr{F}_t \vee \sigma\{z_j\}, \qquad t \in [t_{j-1}, t_j), \quad 1 \leqq j \leqq N. \end{aligned} \tag{36}$$

If $f \in \mathbb{D}^*(\mathscr{A})$ and if the condition

(C–8) $E\rho_N |\mathscr{M}_t(f)| < \infty$ *for all $t \in [0, T]$*

is satisfied, then $(\mathscr{M}_t(f), \mathscr{F}^+_t, \tilde{P})$ is a martingale.

Proof. Let $\mathcal{M}_t = \mathcal{M}_t(f)$ for simplicity. By (C–8), it is evident that $\mathcal{M}_t$ is integrable with respect to $\tilde{P}$ for all $t \in [0, T]$. Since $\mathcal{F}_t^+ \supset \mathcal{F}_t$ and $\mathcal{M}_t$ is adapted to $\mathcal{F}_t$, $\mathcal{M}_t$ is also adapted to $\mathcal{F}_t^+$. Now, it should be established that

$$\tilde{E}[\mathcal{M}_s - \mathcal{M}_t | \mathcal{F}_t^+] = 0 \quad \text{for all } 0 \leqq t < s \leqq T. \tag{37}$$

Note that if $t \in [t_{j-1}, t_j)$ and $s \in (t_{r-1}, t_r]$, $r \geqq j$, we have

$$\begin{aligned} &\tilde{E}[\mathcal{M}_s - \mathcal{M}_t | \mathcal{F}_t^+] \\ &= \tilde{E}[\tilde{E}\{\mathcal{M}_s - \mathcal{M}_{t_{r-1}} | \mathcal{F}_{t_{r-1}}^+\} + \sum_{k=j+1}^{r-1} \tilde{E}\{\mathcal{M}_{t_k} - \mathcal{M}_{t_{k-1}} | \mathcal{F}_{t_{k-1}}^+\} + \mathcal{M}_{t_j} - \mathcal{M}_t | \mathcal{F}_t^+]. \end{aligned} \tag{38}$$

Hence, (37) is shown if $\tilde{E}[\mathcal{M}_s - \mathcal{M}_t | \mathcal{F}_t^+] = 0$ is shown for $t_{j-1} \leqq t < s \leqq t_j$. Since z_j, with respect to $\tilde{P}$, is independent of $\mathcal{G}_{j-1} (\supset \mathcal{F}_t)$ and since $\mathcal{M}_s - \mathcal{M}_t$ is $\mathcal{G}_{j-1}$-measurable for $t_{j-1} \leqq t < s \leqq t_j$, we have

$$\begin{aligned} \tilde{E}[\mathcal{M}_s - \mathcal{M}_t | \mathcal{F}_t^+] &= \tilde{E}[\mathcal{M}_s - \mathcal{M}_t | \mathcal{F}_t] \\ &= \frac{E[\rho_{j-1}(\mathcal{M}_s - \mathcal{M}_t) | \mathcal{F}_t]}{E[\rho_{j-1} | \mathcal{F}_t]} \\ &= E[\mathcal{M}_s - \mathcal{M}_t | \mathcal{F}_t] = 0, \qquad t_{j-1} \leqq t < s \leqq t_j, \end{aligned} \tag{39}$$

where the third equality follows from $\mathcal{F}_t$-measurability of ρ_{j-1}. This completes the proof.

For $t \in [t_{j-1}, t_j)$, let

$$\tilde{\phi}_t = \tilde{E}[\phi_j^t | \mathcal{F}_t^+] \tag{40}$$

and

$$q_t = \tilde{E}\left[\frac{d\phi_j^t}{dt} \middle| \mathcal{F}_t^+\right]. \tag{41}$$

As the second step to obtain the direct formula, we will show the following lemma.

LEMMA 6. *Assume* (C–1)–(C–6). *Let*

$$\mathcal{N}_t(u) = \tilde{\phi}_t - 1 - \int_0^t q_s \, ds. \tag{42}$$

If $u \in \mathbb{I}(h)$, *then* $(\mathcal{N}_t(u), \mathcal{F}_t^+, \tilde{P})$ *is a martingale.*

Proof. For simplicity, let $\mathcal{N}_t = \mathcal{N}_t(u)$. By definition, it is clear that $\mathcal{N}_t$ is adapted to $\mathcal{F}_t^+$ and is integrable with respect to $\tilde{P}$ for all $t \in [0, T]$. Let $t \in [t_{j-1}, t_j)$ and $s \in (t_{r-1}, t_r]$, $r \geqq j$. Then, it follows from Fubini's theorem that

$$\begin{aligned} \tilde{E}[\mathcal{N}_s - \mathcal{N}_t | \mathcal{F}_t^+] &= \tilde{E}[\tilde{\phi}_s - \tilde{\phi}_t - \int_t^s q_\tau \, d\tau | \mathcal{F}_t^+] \\ &= \tilde{E}[\phi_r^s - \phi_j^t - (\phi_r^s - \phi_j^t) | \mathcal{F}_t^+] = 0. \end{aligned} \tag{43}$$

This completes the proof.

By Lemmas 5 and 6, $\mathcal{M}_t = \mathcal{M}_t(f)$ and $\mathcal{N}_t = \mathcal{N}_t(u)$ are $(\mathcal{F}_t^+, \tilde{P})$-martingales if the assumptions are fulfilled. Since $M_t \mathcal{N}_t$ is, by definition, $\tilde{P}$-integrable, there exists an $\mathcal{F}_t^+$-adapted and $\tilde{P}$-integrable process, denoted by $[\mathcal{M}, \mathcal{N}]_t$, such that $\mathcal{M}_t \mathcal{N}_t - [\mathcal{M}, \mathcal{N}]_t$ is an $(\mathcal{F}_t^+, \tilde{P})$-martingale. The process $[\mathcal{M}, \mathcal{N}]$ is $\tilde{P}$-a.s. unique (hence, it is P-a.s. and $\hat{P}_t$-a.s. unique by Lemma 2) and is called the quadratic covariation of $\mathcal{M}_t$ and $\mathcal{N}_t$ (see [11] for details).

We now show the following theorem which presents another formula for the nonlinear filtering problem.

THEOREM 2. *Let*

$$\psi_t = (\tilde{\phi}_t)^{-1}[\mathcal{M}, \mathcal{N}]_t, \qquad t \in [0, T]. \tag{44}$$

Assume (C-1)–(C-8) *and*

(C-9) $E|\psi_t|^2 < \infty$, *for all* t.

If $f \in \mathbb{D}^*(\mathcal{A})$ *and* $u \in \mathbb{U}(h)$, *then* $\hat{f}_t^t \equiv \widehat{f(x_t)}^t$ *satisfies*

$$d\hat{f}_t^t = d\hat{\psi}_t^t + \widehat{\mathcal{A}_t f}^t\, dt + \widehat{\{f_t - \hat{f}_t^t + \hat{\psi}_t^t\} u_t' R^{-1}(j)\{z_j - h_j^t\}}^t\, dt, \tag{45}$$

$$t \in (t_{j-1}, t_j], \quad 1 \leqq j \leqq N.$$

The filtered estimate $E[f_{t_j}|\mathcal{Z}_j]$ *is given by the value of* $\hat{f}_t^t$ *at* $t = t_j$, *i.e.*,

$$E[f_{t_j}|\mathcal{Z}_j] = \hat{f}_{t_j}^{t_j}, \qquad 0 \leqq j \leqq N. \tag{46}$$

Moreover, if u_t *is adapted to* $\mathcal{F}_t^+$, *then* $\psi_t \equiv 0$.

Proof. To show (45), note that

$$\phi_j^t f_t = \phi_j^t\{f_t - \mathcal{M}_t\} + \phi_j^t \mathcal{M}_t. \tag{47}$$

Let $k = j - 1$ for simplicity. Then it follows from (5) that

$$\phi_j^t\{f_t - \mathcal{M}_t\} = \phi_j^{t_k}\{f_{t_k} - \mathcal{M}_{t_k}\} + \int_{t_k}^t \phi_j^s \mathcal{A}_s f\, ds + \int_{t_k}^t \{f_s - \mathcal{M}_s\}\left(\frac{d\phi_j^s}{ds}\right) ds. \tag{48}$$

Substituting (48) into (47), we obtain

$$\begin{aligned} \phi_j^t f_t = {} & \phi_j^{t_k} f_{t_k} + \phi_j^t \mathcal{M}_t - \phi_j^{t_k} \mathcal{M}_{t_k} - \int_{t_k}^t \mathcal{M}_s\left(\frac{d\phi_j^s}{ds}\right) ds \\ & + \int_{t_k}^t \phi_j^s \mathcal{A}_s f\, ds + \int_{t_k}^t f_s\left(\frac{d\phi_j^s}{ds}\right) ds. \end{aligned} \tag{49}$$

Now, we will show that

$$\begin{aligned} \tilde{E}[\phi_j^t f_t|\mathcal{Z}_j] = {} & \tilde{E}[\phi_j^{t_k} f_{t_k}|\mathcal{Z}_j] + \tilde{E}[[\mathcal{M}, \mathcal{N}]_t - [\mathcal{M}, \mathcal{N}]_{t_k}|\mathcal{Z}_j] \\ & + \int_{t_k}^t \tilde{E}[\phi_j^s \mathcal{A}_s f|\mathcal{Z}_j]\, ds + \int_{t_k}^t \tilde{E}[\phi_j^s f_s u_s' R^{-1}(j)\{z_j - h_j^s\}|\mathcal{Z}_j]\, ds. \end{aligned} \tag{50}$$

From (49), we have (50) if it is shown that

$$\tilde{E}[\phi_j^t \mathcal{M}_t|\mathcal{Z}_j] = \tilde{E}[\phi_j^{t_k} \mathcal{M}_{t_k}|\mathcal{Z}_j] + \tilde{E}\left[\int_{t_k}^t \mathcal{M}_s\left(\frac{d\phi_j^s}{ds}\right) ds|\mathcal{Z}_j\right] + \tilde{E}[[\mathcal{M}, \mathcal{N}]_t - [\mathcal{M}, \mathcal{N}]_{t_k}|\mathcal{Z}_j]. \tag{51}$$

Noting that

$$\tilde{E}[\phi_j^t \mathcal{M}_t|\mathcal{Z}_j] = \tilde{E}[\tilde{\phi}_t \mathcal{M}_t|\mathcal{Z}_j], \tag{52}$$

and that

$$\tilde{\phi}_t = \tilde{\phi}_{t_k} + \int_{t_k}^t q_s\, ds + \mathcal{N}_t - \mathcal{N}_{t_k},$$

we have

$$\tilde{E}[\phi_j^t \mathcal{M}_t|\mathcal{Z}_j] = \tilde{E}[\tilde{\phi}_{t_k} \mathcal{M}_t|\mathcal{Z}_j] + \tilde{E}\left[\int_{t_k}^t \mathcal{M}_t q_s\, ds|\mathcal{Z}_j\right] + \tilde{E}[(\mathcal{N}_t - \mathcal{N}_{t_k})\mathcal{M}_t|\mathcal{Z}_j]. \tag{53}$$

Then, by applying Lemma 5 to (53), we get (51) immediately. Hence, we have (50).

Next, note that

$$\hat{f}_t^t = \frac{\tilde{E}[\phi_j^t f_t | \mathscr{Z}_j]}{\tilde{E}[\phi_j^t | \mathscr{Z}_j]} = \frac{\tilde{E}[\phi_j^t f_t | \mathscr{Z}_j] - \tilde{E}[[\mathscr{M}, \mathscr{N}]_t | \mathscr{Z}_j]}{\tilde{E}[\phi_j^t | \mathscr{Z}_j]} + \hat{\psi}_t^t. \tag{54}$$

For the first term of the right-hand side of (54), substituting (50) and applying (29), we have

$$\begin{aligned} \frac{\tilde{E}[\phi_j^t f_t | \mathscr{Z}_j] - \tilde{E}[[\mathscr{N}, \mathscr{M}]_t | \mathscr{Z}_j]}{\tilde{E}[\phi_j^t | \mathscr{Z}_j]} &= \hat{f}_{t_k}^{t_k} - \hat{\psi}_{t_k}^{t_k} + \int_{t_k}^t \widehat{\mathscr{A}_s f^s}\, ds \\ &\quad + \int_{t_k}^t \widehat{\{f_s - \hat{f}_s^s + \hat{\psi}_s^s\} u_s' R^{-1}(j)\{z_j - h_j^s\}^s}\, ds. \end{aligned} \tag{55}$$

Then, substituting (55) into (54), we immediately obtain (45).

Finally, if u_t is adapted to $\mathscr{F}_t^+$, then ϕ_j^t is also adapted to $\mathscr{F}_t^+$. Hence,

$$\tilde{\phi}_t = \tilde{E}[\phi_j^t | \mathscr{F}_t^+] = \phi_j^t,$$

and (42) implies that $\mathscr{N}_t \equiv 0$. Then, $[\mathscr{M}, \mathscr{N}] = 0$, and consequently $\psi_t \equiv 0$. This completes the proof.

5. Examples. So far, we have obtained two main theorems, i.e., Theorems 1 and 2, for the nonlinear filtering problem. Although we must apply Theorem 1 in the case where the estimate $\hat{f}(t)$, $0 \leqq t \leqq T$ is required, both theorems are applicable when we are concerned with computing $\hat{f}(t_j)$, $0 \leqq j \leqq N$. For the latter case, it is not easy to decide which one is more useful since, as usual in nonlinear filtering problems, those theorems do not give feasible explicit solutions. However, we may say that if we can find the process $u \in \mathbb{I}(h)$ which is adapted to $\mathscr{F}_t^+$, then (45) in Theorem 2 is proper since $\psi_t \equiv 0$. On the other hand, if u_t, $t_{j-1} \leqq t < t_j$ depends on x_s, $t < s \leqq t_j$, then it is better to apply Theorem 1 since evaluation of ψ_t is not easy in general. In this section, examples are shown where the minimal variance estimate $\hat{x}_j = E[x_{t_j} | \mathscr{Z}_j]$ is feasible.

Example 1. (*conditionally Gaussian case*). Let $S = \mathbb{R}^n$ and consider a stochastic differential equation

$$\begin{aligned} dx_t &= A(t, z)x_t\, dt + G(t)\, dw_t, \\ x_0 &= \xi^0, \qquad t \in [0, T], \end{aligned} \tag{56}$$

where $x_t \in \mathbb{R}^n$ is the state vector; w_t is a d'-dimensional standard Brownian motion process which is independent of $\{v_j;\ 1 \leqq j \leqq N\}$; ξ^0 is a Gaussian vector with mean $\hat{\xi}^0$ and covariance matrix Q^0 and is independent of $\{v_j;\ 1 \leqq j \leqq N\}$ and $\{w_t;\ 0 \leqq t \leqq T\}$. The observation $z = \{z_j;\ 1 \leqq j \leqq N\}$ is given by (1) with

$$h_j = H_1(j, z)x_{t_j} + \int_{t_{j-1}}^{t_j} H_2(s, z)x_s\, ds, \qquad 1 \leqq j \leqq N,$$

namely,

$$z_j = H_1(j, z)x_{t_j} + \int_{t_{j-1}}^{t_j} H_2(s, z)x_s\, ds + v_j, \qquad 1 \leqq j \leqq N. \tag{57}$$

We assume the following conditions.

(C–10) For all $t \in [0, T]$, $j \leqq N$ and $\kappa \in (\mathbb{R}^m)^N$, the elements of $A(t, \kappa)$, $G(t, \kappa)$, $H_1(j, \kappa)$ and $H_2(t, \kappa)$ are bounded.

(C–11) For all $\kappa \in (\mathbb{R}^m)^N$, $A(t, \kappa)$ and $H_2(t, \kappa)$, $t_{j-1} \leqq t < t_j$ are continuous in t.

(C–12) For each j and $t \in [t_{j-1}, t_j)$, $A(t, z)$, $H_2(t, z)$ and $H_1(j, z)$ are $\mathscr{Z}_{j-1}$-measurable.

The system (56) and (57) is linear in x but nonlinear in z. Note that the observation (57) is more general than the usual Kalman filter model in that the observation at $t = t_j$ depends on the past state $\{x_s; t_{j-1} \leqq s < t_j\}$. Let

$$h_j^* \equiv \frac{1}{t_j - t_{j-1}} h_j.$$

Then, in this example, we can take $u \in \mathbb{U}(h)$ as

$$u_t = h_j^* \quad \text{for } t \in [t_{j-1}, t_j), \quad 1 \leqq j \leqq N.$$

Hence,

$$h_j^t \equiv h_j(\tilde{u}^t) = (t - t_{j-1}) h_j^*.$$

Note the properties:

(i) Given $\mathscr{Z}_{j-1}$, x_{t_j}, h_j and h_j^t are jointly Gaussian with respect to P.

(ii) x_{t_j}, h_j and h_j^t are $\mathscr{G}_{j-1}$-measurable.

(iii) $P = \hat{P}_t$ on $\mathscr{G}_{j-1}$.

(iv) $z_j - h_j^t$, with respect to $\hat{P}_t$, is a Gaussian vector independent of $\mathscr{G}_{j-1}$.

Then, we notice that x_{t_j} and h_j^t, given $\mathscr{Z}_j$, are jointly Gaussian. Let

$$\eta_j = [x_{t_j}' \, h_j^{*\prime}]',$$

$$\hat{\eta}_j = E[\eta_j | \mathscr{Z}_j],$$

$$\hat{\eta}_j^t = E^t[\eta_j | \mathscr{Z}_j]$$

and

$$B_j^t = E^t[(\eta_j - \hat{\eta}_j^t)(\eta_j - \hat{\eta}_j^t)' | \mathscr{Z}_j].$$

Then, it follows from simple computations with (28), (34) and the above-mentioned Gaussian property that $\hat{\eta}_j$ is given by solving the following pair of differential equations:

$$\frac{d\hat{\eta}_j^t}{dt} = B_j^t C' R^{-1}(j)\{z_j - 2(t - t_{j-1}) C' \hat{\eta}_j^t\},$$

$$\hat{\eta}_j^{t_{j-1}} = \hat{\eta}_{j^-}, \qquad t \in [t_{j-1}, t_j), \tag{58}$$

$$\frac{dB_j^t}{dt} = -2(t - t_{j-1}) B_j^t C' R^{-1}(j) C B_j^t,$$

$$B_j^{t_{j-1}} = B_{j^-}, \qquad t \in [t_{j-1}, t_j). \tag{59}$$

In (58) and (59), we have set

$$C = [0 \;\; I], \qquad 0 : m \times n, \qquad I : m \times m,$$

$$\hat{\eta}_{j^-} = E[\eta_j | \mathscr{Z}_{j-1}],$$

and

$$B_{j^-} = E[(\eta_j - \hat{\eta}_{j^-})(\eta_j - \hat{\eta}_{j^-})' | \mathscr{Z}_{j-1}].$$

It is not difficult to see that the solution of the pair (58) and (59) is given by

$$\hat{\eta}_j^t = \hat{\eta}_{j^-} + B_j^t \tilde{C}_t' R^{-1}(j)\{z_j - C\hat{\eta}_{j^-}\}, \tag{60}$$

and

$$B_j^t = B_{j^-} + B_{j^-}\tilde{C}_t'\{\tilde{C}_t B_{j^-}\tilde{C}_t' + R(j)\}^{-1}\tilde{C}_t B_{j^-}, \tag{61}$$

where $\tilde{C}_t = (t - t_{j-1})C$. Let

$$\begin{aligned}
\hat{x}_{j^-} &= E[x_{t_j}|\mathscr{Z}_{j-1}],\\
\hat{h}_{j^-} &= E[h_j|\mathscr{Z}_{j-1}],\\
Q_j &= E[(x_{t_j} - \hat{x}_j)(x_{t_j} - \hat{x}_j)'|\mathscr{Z}_j],\\
Q_{j^-} &= E[(x_{t_j} - \hat{x}_{j^-})(x_{t_j} - \hat{x}_{j^-})'|\mathscr{Z}_{j-1}],\\
M_{j^-} &= E[(x_{t_j} - \hat{x}_{j^-})(h_j - \hat{h}_{j^-})'|\mathscr{Z}_{j-1}]
\end{aligned}$$

and

$$N_{j^-} = E[(h_j - \hat{h}_{j^-})(h_j - \hat{h}_{j^-})'|\mathscr{Z}_{j-1}].$$

Then, setting $t = t_j$ in (60) and (61) and noting that $h_j = (t_j - t_{j-1})h_j^*$, we have

$$\hat{x}_j = \hat{x}_{j^-} + M_{j^-}\{N_{j^-} + R(j)\}^{-1}\{z_j - \hat{h}_{j^-}\} \tag{62a}$$

and

$$Q_j = Q_{j^-} - M_{j^-}\{N_{j^-} + R(j)\}^{-1}M_{j^-}'. \tag{62b}$$

Also, it easily follows from (32) and (56) that

$$\begin{bmatrix} \hat{x}_{j^-} \\ \hat{h}_{j^-} \end{bmatrix} = \tilde{\Phi}^z(t_j, t_{j-1})\hat{x}_{j-1} \tag{62c}$$

and

$$\begin{aligned}
\left[\begin{array}{c|c} Q_{j^-} & M_{j^-} \\ \hline M_{j^-}' & N_{j^-} \end{array}\right] &= \tilde{\Phi}^z(t_j, t_{j-1})Q_{j-1}\tilde{\Phi}^z(t_j, t_{j-1})'\\
&\quad + \int_{t_{j-1}}^{t_j} \tilde{\Phi}^z(t_j, s)G(s)G'(s)\tilde{\Phi}^z(t_j, s)'\,ds,
\end{aligned} \tag{62d}$$

where

$$\tilde{\Phi}^z(t, s) = \begin{bmatrix} \Phi^z(t, s) \\ \Psi^z(t, s) \end{bmatrix}, \qquad t, s \in [t_{j-1}, t_j]$$

is given by

$$\frac{\partial}{\partial t}\Phi^z(t, s) = A(t, z)\Phi^z(t, s), \qquad \Phi^z(s, s) = I$$

and

$$\Psi^z(t, s) = H_1(j, z)\Phi^z(t, s) + \int_s^t H_2(\tau, z)\Phi^z(t, s)\,d\tau.$$

Thus, with $\hat{x}_0 = \hat{\xi}^0$ and $Q_0 = Q^0$, (62a)–(62d) forms a recursive nonlinear filter. In the special case where $H_2 \equiv 0$, $H_1(j, z) \equiv H_1(j)$ and $A(t, z) \equiv A(t)$, formula (62) reduces to the continuous-discrete Kalman filter derived by Jazwinski [8].

Example 2. Let $S = \mathbb{R}^n \times \{0, 1\}$ and $x_t \equiv (\alpha_t, \xi_t)$, and consider the system of equations

$$d\xi_t = A(t, z)\xi_t\,dt + G(t)\,dw_t, \qquad \xi_0 = \xi^0, \quad t \in [0, T] \tag{63}$$

and

$$z_j = \int_{t_{j-1}}^{t_j} \alpha_s H(s)\xi_s \, ds + v_j, \qquad 1 \leqq j \leqq N, \tag{64}$$

where $H(t)$ is an $m \times n$-dimensional matrix with time-dependent and bounded elements and α_t is a binary-valued process which changes from 1 to 0 at a random time $\tau(\omega)$, i.e., $\alpha_t = I_{\{\tau(\omega)<t\}}$. The random variable $\tau(\omega)$, which is assumed to be independent of ξ_t, $t \leqq T$ and v_j, $j \leqq N$, takes zero with probability π, and given that $\tau(\omega) > 0$, $\tau(\omega)$ is exponentially distributed with a parameter λ; i.e.,

$$P\{\tau(\omega) \leqq t\} = \pi + (1-\pi)(1-e^{-\lambda t}).$$

Let y_t, $0 \leqq t \leqq T$ be the continuous-time process defined by

$$dy_t = \alpha_t H(t)\xi_t \, dt + \tilde{R}(t) \, d\tilde{v}_t, \qquad y_0 = 0, \quad t \in [0, T],$$

where $\tilde{v}$ is an m-dimensional standard Brownian motion process. Then, the observation (64) can be constructed from y_t, $t \leqq T$ by setting

$$z_j = y_{t_j} - y_{t_{j-1}}$$

and

$$R(j) = \int_{t_{j-1}}^{t_j} \tilde{R}(s)\tilde{R}'(s) \, ds.$$

Let $u_t = \alpha_t H(t)\xi_t$. Then u_t is adapted to $\mathscr{F}_t^+$. Hence, Theorem 2 is applicable. Let

$$\theta_t = \{1 - \alpha_t\}\xi_t;$$

then

$$\xi_t = \alpha_t \xi_t + \theta_t.$$

Hence, we have

$$\hat{\xi}_t^t = \hat{\alpha}_t^t \bar{\xi}_t^t + \hat{\theta}_t^t, \tag{65}$$

where

$$\bar{\xi}_t^t = E^t[\xi_t | \alpha_t = 1, \mathscr{Z}_j].$$

Therefore, it suffices to compute the evolutions of $\hat{\alpha}_t^t$, $\bar{\xi}_t^t$ and $\hat{\theta}_t^t$ in order to obtain the minimal variance estimate $\hat{\alpha}_j = \hat{\alpha}_{t_j}^{t_j}$ and $\hat{\xi}_j = \hat{\xi}_{t_j}^{t_j}$. Since the operator $\mathscr{A}_t$ is given by

$$\begin{aligned}\mathscr{A}_t f(x_t) &= \mathscr{A}_t f(\xi_t, \alpha_t) \\ &= \lambda\alpha_t\{f(\xi_t, 0) - f(\xi_t, 1)\} + f_\xi'(x_t)A(t, z)\xi_t + \tfrac{1}{2}\operatorname{tr}[G' f_{\xi\xi}(x_t)G],\end{aligned}$$

it follows from (45) that

$$d\hat{\alpha}_t^t = [(e_t - \lambda)\hat{\alpha}_t^t - e_t\{\hat{\alpha}_t^t\}^2] \, dt, \qquad \hat{\alpha}_0^0 = \pi \tag{66}$$

and

$$d\hat{\theta}_t^t = [\{A(t, z) - e_t I\}\hat{\theta}_t^t + \lambda\hat{\alpha}_t^t F\bar{\kappa}_t^t] \, dt, \qquad \hat{\theta}_0^0 = \pi\hat{\xi}^0, \tag{67}$$

where

$$F=[I\,\vdots\,0],\qquad I:n\times n,\quad 0:n\times m,$$

$$\tilde{A}(t,z)=\left[\begin{array}{c:c}A(t,z) & 0\\ \hdashline H(t) & 0\end{array}\right],$$

$$\bar{\kappa}_t^t=E^t[\kappa_t|\alpha_t=1,\mathscr{Z}_j],$$

$$\kappa_t=[\xi_t'\,\vdots\,h_j^{t\prime}]',$$

$$e_t=\bar{\kappa}_t^{t\prime}\tilde{A}'(t,z)C'R^{-1}(j)\{z_j-C\bar{\kappa}_t^t\}-\mathrm{tr}\,[\tilde{A}'(t,z)C'R^{-1}(j)C\bar{B}_t^t],$$

and

$$\bar{B}_t^t=E^t[(\kappa_t-\bar{\kappa}_t^t)(\kappa_t-\bar{\kappa}_t^t)'|\alpha_t=1,\mathscr{Z}_j].$$

Now, we shall obtain the evolutions of $\bar{\kappa}_t^t$ and $\bar{B}_t^t$. Noting that $E^t[\cdot\,|\alpha_t=1,\mathscr{Z}_j]=E^t[\cdot\,|\mathscr{Z}_j]/\hat{\alpha}_t^t$ and that ξ_t and h_j^t, given $\alpha_t=1$, are Gaussian, it can be seen from Lemma 4 and Theorem 2 that

$$\begin{aligned}&d\bar{\kappa}_t^t=\{\tilde{A}-\bar{B}_t^t[C'R^{-1}C\tilde{A}+\tilde{A}'C'R^{-1}C]\}\bar{\kappa}_t^t\,dt+\bar{B}_t^t\tilde{A}'C'R^{-1}z_j\,dt,\\ &\bar{\kappa}_{t_{j-1}}^{t_{j-1}}=\bar{\kappa}_{j-1}=[\bar{\xi}_{j-1}'\,\vdots\,0]',\qquad t\in[t_{j-1},t_j],\end{aligned}\tag{68}$$

and

$$\begin{aligned}&d\bar{B}_t^t=[\tilde{A}\bar{B}_t^t+\bar{B}_t^t\tilde{A}'+\tilde{G}]\,dt-\bar{B}_t^t\{C'R^{-1}C\tilde{A}+\tilde{A}'C'R^{-1}C\}\bar{B}_t^t\,dt,\\ &\bar{B}_{t_{j-1}}^{t_{j-1}}=\bar{B}_{j-1}=\left[\begin{array}{c:c}\bar{Q}_{j-1} & 0\\ \hdashline 0 & 0\end{array}\right],\qquad t\in[t_{j-1},t_j],\end{aligned}\tag{69}$$

where

$$\bar{\xi}_{j-1}=\bar{\xi}_{t_{j-1}}^{t_{j-1}}=E[\xi_{t_{j-1}}|\alpha_{t_{j-1}}=1,\mathscr{Z}_{j-1}],$$

$$\bar{Q}_{j-1}=E[(\xi_{t_{j-1}}-\bar{\xi}_{j-1})(\xi_{t_{j-1}}-\bar{\xi}_{j-1})'|\alpha_{t_{j-1}}=1,\mathscr{Z}_{j-1}],$$

$$\tilde{G}=\tilde{G}(t)=\left[\begin{array}{c:c}G(t)G'(t) & 0\\ \hdashline 0 & 0\end{array}\right],$$

and $\tilde{A}=\tilde{A}(t,z)$.

Since $\bar{\xi}_t^t$ is an element of $\bar{\kappa}_t^t$, we can compute $\hat{\alpha}_t^t$, $\bar{\kappa}_t^t$ and $\hat{\theta}_t^t$ recursively by (66)–(69). Hence, we can compute $(\hat{\xi}_j,\hat{\alpha}_j)=(\hat{\xi}_{t_j}^{t_j},\hat{\alpha}_{t_j}^{t_j})$ by (65).

Remark 6. The solution of (68) and (69) is directly available in the following form:

$$\bar{\kappa}_t^t=\tilde{\Psi}^z(t,t_{j-1})\bar{\kappa}_{j-1}+B_t^*C'\{CB_t^*C'+R(j)\}^{-1}\{z_j-C\tilde{\Psi}^z(t,t_{j-1})\bar{\kappa}_{j-1}\},$$

$$\bar{B}_t^t=B_t^*-B_t^*C'\{CB_t^*C'+R(j)\}^{-1}CB_t^*,$$

$$B_t^*=\tilde{\Psi}^z(t,t_{j-1})\bar{B}_{j-1}\tilde{\Psi}^z(t,t_{j-1})+\int_{t_{j-1}}^t\tilde{\Psi}^z(t,s)\tilde{G}(s)\tilde{\Psi}^z(t,s)'\,ds,$$

where $\tilde{\Psi}^z$ is the transition matrix given by

$$\frac{d}{dt}\tilde{\Psi}^z(t,s)=\tilde{A}(t,z)\tilde{\Psi}^z(t,s),\qquad\tilde{\Psi}^z(s,s)=I.$$

Appendix.

Proof of Lemma 1 (ii) *and* (iii). The properties (ii) and (iii) of $\tilde{P}$ are shown in the following way.

For $\xi_j \in \mathbb{R}^m$, $1 \leqq j \leqq N$, let us define $\gamma_j = \gamma_j(\xi)$ by

$$\gamma_j = \exp\left\{ i \sum_{k=1}^{j} \xi_k'[h_k + v_k] + \frac{1}{2} \sum_{k=1}^{j} \xi_k' R(k) \xi_k \right\}, \qquad 1 \leqq j \leqq N,$$

and $\gamma_0 = 1$. Then, it follows that

$$\gamma_j \rho_j = \exp\left\{ \sum_{k=1}^{j} m_k' R^{-1}(k) v_k - \frac{1}{2} \sum_{k=1}^{j} m_k' R^{-1}(k) m_k \right\},$$

where

$$m_k = iR(k)\xi_k - h_k.$$

Clearly, $\gamma_j \rho_j$ is also a $(\mathcal{G}_j, P)$-martingale. Then, for all $l \geqq j$ and $A \in \mathcal{G}_{j-1}$, we have

$$\begin{aligned} \int_A \gamma_l \, d\tilde{P} &= \int_A \gamma_l E[\rho_N | \mathcal{G}_l] \, dP = \int_A \gamma_l \rho_l \, dP \\ &= \int_A \gamma_{j-1} \rho_{j-1} \, dP = \int_A \gamma_{j-1} \, d\tilde{P}. \end{aligned}$$

Hence, γ_j is a $(\mathcal{G}_j, \tilde{P})$-martingale, and we have

$$\int_A \gamma_l (\gamma_{j-1})^{-1} \, d\tilde{P} = \int_A \tilde{E}[\gamma_l | \mathcal{G}_{j-1}](\gamma_{j-1})^{-1} \, d\tilde{P} = \tilde{P}(A).$$

This implies that

$$\int_A \exp\left\{ i \sum_{k=j}^{l} \xi_k' z_k \right\} d\tilde{P} = \tilde{P}(A) \prod_{k=j}^{l} \exp\{-\tfrac{1}{2} \xi_k' R(k) \xi_k\}.$$

The above equation proves (ii) and (iii) simultaneously. This completes the proof.

REFERENCES

[1] M. FUJISAKI, G. KALLIANPUR AND H. KUNITA, *Stochastic differential equations for the nonlinear filtering problem*, Osaka J. Math., 9 (1972), pp. 19–40.

[2] H. KUNITA, *Estimation of Stochastic Processes*, Sangyo-tosho, Tokyo, 1976. (In Japanese.)

[3] R. S. LIPTSER AND A. N. SHIRYAYEV, *Statistics of Random Processes I: General Theory*, Springer-Verlag, New York, 1977.

[4] A. N. SHIRYAYEV, *Stochastic equations of nonlinear filtering of Markovian jump processes*, Problemy Peredači Informacii, 2 (1966), pp. 3–22.

[5] R. S. LIPTSER AND A. N. SHIRYAYEV, *Nonlinear filtering of Markov diffusion processes*, Trudy Mat. Inst. Steklov, 104 (1968), pp. 135–180.

[6] G. KALLIANPUR AND C. STRIEBEL, *Stochastic differential equations in statistical estimation problems*, in Multivariable Analysis II, Academic Press, New York, 1970.

[7] A. SEGALL, *Centralized and decentralized control schemes for Gauss-Poisson processes*, IEEE Trans. Automat. Control, AC-23 (1978), pp. 47–57.

[8] A. H. JAZWINSKI, *Stochastic Processes and Filtering Theory*, Academic Press, New York, 1970.

[9] G. KALLIANPUR AND C. STRIEBEL, *Estimation of stochastic processes: Arbitrary system process with additive white noise observation errors*, Ann. Math. Statist., 39 (1968), pp. 785–801.

[10] I. V. GIRSANOV, *On transforming a certain class of stochastic processes by absolutely continuous substitution of measures*, Theory Prob. Appl., 5 (1960), pp. 285–310.

[11] C. DOLÉANS-DADE AND P. A. MEYER, *Intégrales stochastiques par rapport aux martingales locales*, in Séminaire de Probabilités IV, Lecture Notes in Mathematics, no. 124, Springer-Verlag, New York, 1970, pp. 77–107.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 2, March 1981

0363-0129/81/1902-0006$01.00/0

SPECTRAL THEORY OF THE LINEAR-QUADRATIC OPTIMAL CONTROL PROBLEM: ANALYTIC FACTORIZATION OF RATIONAL MATRIX-VALUED FUNCTIONS*

EDMOND A. JONCKHEERE† AND LEONARD M. SILVERMAN†

Abstract. The inversion of the Toeplitz operator T_Φ, associated with the operator-valued function Φ defined on the unit circle, is well known to involve a special factorization, called analytic factorization, of the function Φ. New results and algorithms concerning this factorization are presented, in the special case where Φ is rational and matrix-valued.

1. Introduction. If Φ is a rational function from the unit circle to $\mathbb{C}^{n\times n}$, then the associated Toeplitz operator T_Φ can be represented by the semi-infinite block-matrix whose blocks are the Fourier coefficients of Φ, the negative Fourier coefficients being in the upper triangular part [7]. It can be shown [12, Theorem 2] that T_Φ is invertible if and only if Φ admits a so-called *analytic factorization*

$$\Phi = \Lambda^* P;$$

Λ and P are functions defined on the unit circle, with vanishing negative Fourier coefficients, and taking values in $\mathbb{C}^{n\times n}$; Λ^{-1} and P^{-1} exist and should be of the same type as Λ and P; a star denotes the conjugate transpose or the adjoint. If the factors exist and are known, then it can be shown that [12, Theorem 2]

$$T_\Phi^{-1} = T_{P^{-1}} T_{\Lambda^{-1}}^*.$$

This representation of the inverse in terms of the factors Λ and P is useful, because in many cases it allows the computation and the study of the inverse [12]–[17].

Although the equivalence between the inversion and the factorization problems is fairly well known, there has not been that much interest in the factorization problem itself. Practical algorithms are lacking to determine whether a function is factorable and to determine the factors if they exist; moreover, the basic properties of the factors, if they exist, are not yet known. Similarly, in the context of the inversion of Toeplitz operators, although theoretical results exist [5], invertibility tests are needed. Along that line, we, however, point out the partial result of Pattanayak [32].

In this paper, we look in detail at the analytic factorization problem, in the case where Φ is rational and matrix-valued. An algorithm to determine whether Φ is factorable, and to compute the factors if they exist, is presented. The general properties of the factors emerge. The derivation of our results relies heavily on the so-called Hilbert problem of factorization theory [16, Theorem 3.1].

The paper is organized as follows. Section 2 is concerned with spaces of analytic functions, Toeplitz operators, and the precise statement of the equivalence between the inversion problem and a factorization problem. In § 3, we distinguish two types of analytic factorizations: the above-mentioned factorization and another slightly different one (this is necessary for mathematical rigor); the associated Hilbert problem is also introduced. Section 4 is concerned with the pole and zero removal phase of the factorization; this enables the reduction of Φ to a matrix-valued function of constant

* Received by the editors March 7, 1978 and in final revised form May 12, 1980. This research was supported by the National Science Foundation under Grant ENG 76-14379 and by the Joint Services Electronics Program through AFOSR/AFSC under Contract F44620-71-C-0067.

† Department of Electrical Engineering–Systems, University of Southern California, Los Angeles, California 90007.

determinant. Section 5 is concerned with the problem of factoring a matrix-valued function of constant determinant with the Hilbert problem playing a crucial role. Section 6 briefly reviews the whole factorization algorithm. Section 7 is devoted to a particular case, the spectral factorization; we examine this classical problem in the light of our results; this enables us to point out and to rectify an error in the proof of the discrete-time spectral factorization statement [20]. Section 8 is the conclusion. The pole and zero removal phase of the factorization (§ 4) needs some elements of index theory as applied to Fredholm Toeplitz operators; the necessary background material is relegated to Appendix A.

This research is motivated by previous papers of the authors [26], [27] which are concerned with the discrete-time linear-quadratic optimal control problem, in the case where the quadratic cost to be infimized is not necessarily positive semidefinite. To grasp the significance of the factorization problem in control theory, consider the discrete-time, linear, finite-dimensional system $x(k+1)=Ax(k)+Bu(k)$; $x(k)\in\mathbb{R}^n$, $u(k)\in\mathbb{R}^n$, and A and B are real, time-invariant matrices of compatible sizes; the pair (A, B) is controllable. We further assume that A is asymptotically stable; this last assumption does not introduce any loss of generality; see [27, § II, C]. Together with the dynamical system, we define two outputs $y(k)=Cx(k)+Du(k)$ and $z(k)=Ex(k)+Fu(k)$; $y(k)$ and $z(k)$ are in $\mathbb{R}^m$; C, D, E, and F are real, time-invariant matrices of compatible sizes. Let the initial state be $x(i)=\xi$, and define the control sequence $U(i,t)=[u^*(i)\cdots u^*(t-1)]^*$. The quadratic performance index, to be infimized, is then defined as $J[\xi, U(i,t)]=\sum_{k=i}^{t-1} y^*(k)z(k)$. This performance index is not necessarily positive semidefinite. Hence, it might not be bounded from below, resulting in the optimal cost diverging towards minus infinity. More precisely, the question that arises is whether or not there exists a sequence of matrices $\{N(t-i)=N^*(t-i)\in\mathbb{R}^{n\times n} : t\geqq i\}$ such that $J[\xi, U(i,t)]\geqq\xi^*N(t-i)\xi$, for all ξ, all $U(i,t)$ and all $t\geqq i$. If such a sequence exists, we say, more simply, that "the cost is bounded from below". Although this seems a purely control theoretic problem, it has a wide range of interpretation and application [27], [34] which makes it a fundamental problem in system theory. There have been many attempts to characterize boundedness of the cost in the frequency-domain and to find a useful test to check whether or not the cost is bounded from below; see the discussions in [27, §§ I and IV]. In [27], this boundedness problem was restated in the appropriate Hilbert space setting, and it was shown that the crucial underlying mathematical issue is a Toeplitz operator having its spectrum included in $\mathbb{R}^+$. To define this Toeplitz operator, let $J(e^{i\theta})=D+C(e^{j\theta}I-A)^{-1}B$ and $K(e^{j\theta})=F+E(e^{j\theta}I-A)^{-1}B$. Let $\Phi=KJ^*$. The Toeplitz operator in question is T_Φ. *Hence, the cost is bounded from below if and only if the spectrum of T_Φ is a subset of* $\mathbb{R}^+$. It was shown in [27, § IV] and [33, § IV] that, even if the cost is not bounded from below, the spectrum of T_Φ is the union of a compact subset of the real line and at most a finite set of isolated, real eigenvalues of finite multiplicities. It follows that the frequency-domain characterization of the existence of a lower bound to the cost can be stated as follows: *The cost is bounded from below if and only if for all* $\lambda\in(-\infty, 0)$ *there exists a factorization*

$$\Phi-\lambda I=\Lambda_\lambda^* P_\lambda,$$

where Λ_λ and P_λ are defined on the unit circle, with vanishing negative Fourier coefficients, and taking values in $\mathbb{R}^{m\times m}$; moreover, Λ_λ^{-1} and P_λ^{-1} should exist and be of the same type as Λ_λ and P_λ. Since $T_{\Phi-\lambda I}=T_\Phi-\lambda I$, the set of λ's for which the above factorability condition breaks down is the spectrum of T_Φ. This spectrum is important, because its structure reflects the intrinsic properties of the control problem [27], [33], [34]. For example, it can be shown that, if the transfer matrices J and K are invertible and minimum phase, then there are no isolated eigenvalues of finite multiplicities in the

spectrum of T_Φ. Along the same line, it was shown in [27, § III] that the compact part of the spectrum contains information as to whether or not a reduction of the algebraic Riccati equation is possible. The spectrum of T_Φ is thus worth examining; this provides a new approach to linear-quadratic control, the spectral theoretic approach. To compute the spectrum of T_Φ, we have to look at the factorability of $\Phi - \lambda I$, *for all* λ*'s*. The results of this paper allow us to check the factorability of $\Phi - \lambda I$, *for a given* λ. To compute the spectrum it remains to find how the results of this paper should be used in order to handle the case where λ is variable. In [33], precisely this task is achieved.

2. Setup. Let $L^\rho(\mathbb{C}^n)(1 \leqq \rho \leqq \infty;\ n = 1, 2, \cdots)$ be the classical Lebesgue spaces, for the normalized measure $d\theta/2\pi$, of functions from the unit circle $\mathbb{T} = \{e^{j\theta}: \theta \in [0, 2\pi)\}$ to the space $\mathbb{C}^n$. It is easily seen that $L^\infty(\mathbb{C}^n) \subseteq L^2(\mathbb{C}^n)$. The closed subspace of $L^\rho(\mathbb{C}^n)$ of functions whose Fourier coefficients vanish on the strictly negative (positive) integers is the Hardy space $H^\rho(\mathbb{C}^n)(K^\rho(\mathbb{C}^n))$. Define the orthogonal projections

$$P_{H^\rho(\mathbb{C}^n)}: L^\rho(\mathbb{C}^n) \to H^\rho(\mathbb{C}^n), \qquad 1 \leqq \rho \leqq \infty,$$

$$P_{K^\rho(\mathbb{C}^n)}: L^\rho(\mathbb{C}^n) \to K^\rho(\mathbb{C}^n), \qquad 1 \leqq \rho \leqq \infty.$$

$C(\mathbb{C}^n)$, $P(\mathbb{C}^n)$, and $R(\mathbb{C}^n)$ are the sets of functions from $\mathbb{T}$ to $\mathbb{C}^n$ that are continuous, trigonometric polynomials, and ratios of trigonometric polynomials (or rational), respectively.

Let M_n be the algebra of endomorphisms of the space $\mathbb{C}^n$; this can be identified with the algebra of $n \times n$ complex matrices. Then $L^\rho(M_n)$, $H^\rho(M_n)$ and $K^\rho(M_n)$ are defined as the sets of functions Φ from $\mathbb{T}$ to M_n such that the function

$$\mathbb{T} \to \mathbb{C},$$

$$e^{j\theta} \mapsto x^* \Phi(e^{j\theta}) y,$$

be in $L^\rho(\mathbb{C})$, $H^\rho(\mathbb{C})$, and $K^\rho(\mathbb{C})$, respectively, for all x and all y in $\mathbb{C}^n$. $H^\rho(M_n)$ and $K^\rho(M_n)$ are closed subspaces of $L^\rho(M_n)$. Also, observe that $L^\infty(M_n) \subseteq L^2(M_n)$. Define the following orthogonal projections:

$$P_{H^\rho(M_n)}: L^\rho(M_n) \to H^\rho(M_n), \qquad 1 \leqq \rho \leqq \infty,$$

$$P_{K^\rho(M_n)}: L^\rho(M_n) \to K^\rho(M_n), \qquad 1 \leqq \rho \leqq \infty.$$

The norm of $\Phi \in L^\infty(M_n)$ is defined by $\operatorname{ess\,sup}\{\|\Phi(e^{j\theta})\|: \theta \in [0, 2\pi)\}$. This norm, together with pointwise algebraic operations, makes $L^\infty(M_n)$ a Banach algebra. $H^\infty(M_n)$ and $K^\infty(M_n)$ are clearly closed subalgebras of $L^\infty(M_n)$.

$C(M_n)$, $P(M_n)$, and $R(M_n)$ are the sets of functions Φ from $\mathbb{T}$ to M_n such that the function

$$\mathbb{T} \to \mathbb{C},$$

$$e^{j\theta} \mapsto x^* \Phi(e^{j\theta}) y,$$

be in $C(\mathbb{C})$, $P(\mathbb{C})$, and $R(\mathbb{C})$, respectively, for all x and all y in $\mathbb{C}^n$.

It is convenient to introduce the function

$$\chi: \mathbb{T} \to \mathbb{T}, \qquad e^{j\theta} \mapsto e^{j\theta}.$$

If $\Phi \in L^2(M_n)$, then its Fourier expansion is

$$\Phi = \sum_{k=-\infty}^{+\infty} \Phi_k \chi^k.$$

The function Φ^* is defined by

$$\Phi^* = \sum_{k=-\infty}^{+\infty} \Phi_k^* \chi^{-k},$$

where Φ_k^* is the conjugate transpose (or the adjoint) of Φ_k.

References that make ample contact with the above material are Hoffman [1], Düren [2], Helson [3] and Douglas [4, Ch. 6].

Let $\Phi \in L^\infty(M_n)$. Then the Toeplitz operator associated with Φ is by definition

$$T_\Phi: H^2(\mathbb{C}^n) \to H^2(\mathbb{C}^n), \qquad \varphi \mapsto T_\Phi \varphi = P_{H^2(\mathbb{C}^n)}(\Phi\varphi).$$

It is useful to introduce the Laurent operator associated with $\Phi \in L^\infty(M_n)$ which is defined by

$$L_\Phi: L^2(\mathbb{C}^n) \to L^2(\mathbb{C}^n), \qquad \varphi \mapsto L_\Phi \varphi = \Phi\varphi.$$

Standard references about Toeplitz operators are Douglas [4, Ch. 7], [5], Grenander and Szegö [6] and Widom [7].

As was said in the introduction, the problem of inverting the Toeplitz operator T_Φ is closely related to the problem of factoring the function Φ to which it is associated.

DEFINITION 1. Let $\Phi \in L^\infty(M_n)$. A *weak analytic factorization* for Φ is by definition

$$\Phi = \bar{\Lambda}^* \cup \bar{P}, \tag{1a}$$

where

$$\bar{\Lambda}, \bar{\Lambda}^{-1}, \bar{P}, \bar{P}^{-1} \in H^\infty(M_n), \tag{1b}$$

$$U \in L^\infty(M_n), \tag{1c}$$

$$U(e^{j\theta}) \text{ is unitary for almost every } \theta \in [0, 2\pi); \tag{1d}$$

moreover, there exists an element

$$V \in H^\infty(M_n), \tag{1e}$$

$$V^{-1} \in H^\infty(M_n), \tag{1f}$$

such that

$$\operatorname{ess\,sup}\{\|V(e^{j\theta}) - U(e^{j\theta})\|: \theta \in [0, 2\pi)\} < 1. \tag{1g}$$

PROPOSITION 1. *Let $\Phi \in L^\infty(M_n)$. Then the Toeplitz operator T_Φ is invertible if and only if Φ has a weak analytic factorization.*

Proof. Early versions of this result are available in Widom [8, Thm. I] and Pousson [9, Thm. 3.4]. This factorability criterion was subsequently developed by Pousson [10] and Devinatz [11]. The result in its definitive and general form is to be found in Rabindranathan [12, Thm. 2].

This paper is mainly concerned with the factorization of Proposition 1 in the case where Φ is in $L^\infty(M_n) \cap R(M_n)$.

The main difference between the spectral factorization and the weak analytic factorization is that, in the latter, the function Φ to be factored is not restricted to satisfy the condition $\Phi = \Phi^*$, nor is it restricted to take positive semidefinite values along the unit circle. This kind of factorization seems to have been introduced by Gohberg and Krein in a famous series of papers [13]–[15]. These papers, in the continuous-time setting, are mostly concerned with integral equations and do not look at the factorization problem itself.

A general question of terminology: Since $\Phi \in L^\infty(M_n) \cap R(M_n)$, this function can be analytically extended from $\mathbb{T}$ to $\mathbb{C}\backslash\{p_1, p_2, \cdots\}$, where $\{p_1, p_2, \cdots\}$ is the *finite* set of poles of the extension of Φ. The value taken by the extension at $z \in \mathbb{C}\backslash\{p_1, p_2, \cdots\}$ will be written $\Phi(z)$. In the sequel, we shall not explicitly specify whether Φ should be interpreted as the original function on $\mathbb{T}$ or its extension to $\mathbb{C}\backslash\{p_1, p_2, \cdots\}$; this will be implicitly specified by the context.

3. Strong analytic factorization and Hilbert problem. Conditions (1c)–(1g) in Def. 1 are somewhat troublesome. However, we shall prove in the sequel that, if $\Phi \in L^\infty(M_n) \cap R(M_n)$ and if a weak analytic factorization exists, then one has always the freedom to take $U = \mathrm{I}$, so that Conditions (1c)–(1g) become completely irrelevant. This motivates the following definition:

DEFINITION 2. A *strong analytic factorization* for $\Phi \in L^\infty(M_n)$ is by definition

$$\Phi = \Lambda^* P, \tag{2a}$$

where

$$\Lambda, P \in H^\infty(M_n), \tag{2b}$$

$$\Lambda^{-1}, P^{-1} \in H^\infty(M_n). \tag{2c}$$

Strong analytic factorability is clearly a condition much stronger and much cleaner than weak analytic factorability. Moreover, it will be shown that the strong analytic factorization belongs to a broad class of factorizations, all of them having the same underlying algebra. This algebraic nature of the factorization problem is treated in detail in McNabb and Schumitzky [16]. In algebraic terms the factorization problem is defined as follows: Let R be a ring of unit element e. Let p^+ and p^- be two projections defined on R that commute. Let $p^0 = p^+p^- = p^-p^+$. Define the additive groups $R^+ = p^+(R)$, $R^- = p^-(R)$, and $R^0 = p^0(R)$. (R^+, R^-) is said to be a *factorization structure* in R if [16, § 2]

(a) R^+, R^- are subrings of R; $e \in R^+ \cap R^-$;

(b) p^0 is a ring homomorphism of R^+ and R^- into R^0;

(c) $R^+R^- \subseteq R^+ + R^-$.

In algebraic terms, the factorization problem is posed as follows: Let $x \in R$; then does there exist a factorization $x = uv$, where $u, u^{-1} \in R^+$, and $v, v^{-1} \in R^-$? If yes, compute the factors.

Let us now prove that the strong analytic factorization belongs to the class of factorizations which can be posed within the same algebraic setting.

THEOREM 1. *$(K^\infty(M_n), H^\infty(M_n))$ is a factorization structure in $L^\infty(M_n)$.*

Proof. As said in § 2, $L^\infty(M_n)$ is a Banach algebra, and hence a ring. The unit element is I.

The projections $P_{K^\infty(M_n)}$ and $P_{H^\infty(M_n)}$ commute, because $P_{K^\infty(M_n)}P_{H^\infty(M_n)}$ and $P_{H^\infty(M_n)}P_{K^\infty(M_n)}$ are both the orthogonal projection from $L^\infty(M_n)$ onto M_n.

By definition of the projections, we have

$$K^\infty(M_n) = P_{K^\infty(M_n)}[L^\infty(M_n)],$$

$$H^\infty(M_n) = P_{H^\infty(M_n)}[L^\infty(M_n)];$$

moreover,

$$M_n = P_{K^\infty(M_n)}P_{H^\infty(M_n)}[L^\infty(M_n)].$$

Finally, the following conditions are easily verified:

(a) $K^\infty(M_n)$ and $H^\infty(M_n)$ are closed subalgebras (and hence subrings) of $L^\infty(M_n)$; $I \in K^\infty(M_n) \cap H^\infty(M_n) = M_n$;

(b) $P_{K^\infty(M_n)} P_{H^\infty(M_n)} = P_{H^\infty(M_n)} P_{K^\infty(M_n)}$ is a ring homomorphism of $K^\infty(M_n)$ and $H^\infty(M_n)$ into M_n;

(c) $K^\infty(M_n) H^\infty(M_n) \subseteq K^\infty(M_n) + H^\infty(M_n)$.

This completes the proof; see [16, § 2].

The strong analytic factorization thus belongs to the class of factorizations considered in [16]. A first consequence of this fact is the following:

THEOREM 2. *Let $\Phi \in L^\infty(M_n)$. If $\Phi = \Lambda^* P$ is a strong analytic factorization, then any other strong analytic factorization has the form $[(C^*)^{-1}\Lambda]^*(CP)$, where C is in M_n and nonsingular.*

Proof. See McNabb and Schumitzky [16, Proposition 2.1].

Any factorization problem that can be stated within the above-described algebraic framework is solvable via an associated *Hilbert problem* [16, § 3].

DEFINITION 3. Let $\Phi \in L^\infty(M_n)$. Then the *Hilbert problem* associated with the strong analytic factorization of Φ is defined as the following system of equations in (L, R):

$$P_{K^\infty(M_n)}(L^*\Phi) = I, \tag{3a}$$

$$P_{H^\infty(M_n)}(\Phi R) = I, \tag{3b}$$

$$L, R \in H^\infty(M_n), \tag{3c}$$

$$L^{-1}, R^{-1} \in H^\infty(M_n). \tag{3d}$$

THEOREM 3. *Let $\Phi \in L^\infty(M_n)$. Then Φ admits a strong analytic factorization if and only if there exists a (unique) solution to the Hilbert problem* (3a)–(3d). *Moreover, should (L, R) be the solution, then a strong analytic factorization is given by*

$$\Phi = (L^{-1})^*(R_0 R^{-1}),$$

where R_0 is the coefficient of χ^0 in the Fourier expansion of R.

Proof. It is a direct consequence of [16, Thm. 3.1 and Corollary 3.2].

As pointed out in [16], Condition (3d) on the solution of the Hilbert problem is difficult to check in practice. However, when the matrix-valued function to be factored has a constant determinant, the problem gets easier.

DEFINITION 4. Let $\Phi^c \in L^\infty(M_n)$, and let $\det \Phi^c(e^{j\theta}) = c = \text{constant} \neq 0$ for almost every $\theta \in [0, 2\pi)$. The Hilbert problem associated with the strong analytic factorization of Φ^c is the following system of matrix equations in (L^c, R^c):

$$P_{K^\infty(M_n)}(L^{c*}\Phi^c) = I, \tag{4a}$$

$$P_{H^\infty(M_n)}(\Phi^c R^c) = I, \tag{4b}$$

$$L^c, R^c \in H^\infty(M_n), \tag{4c}$$

$$(L^c)^{-1}, (R^c)^{-1} \in H^\infty(M_n). \tag{4d}$$

The following result is clearly the same as that of Theorem 3.

THEOREM 4. *Let $\Phi^c \in L^\infty(M_n)$, and let $\det \Phi^c(e^{j\theta}) = c \neq 0$ for almost every $\theta \in [0, 2\pi)$. Then Φ^c has a strong analytic factorization if and only if there exists a (unique) solution (L^c, R^c) to the Hilbert problem* (4a)–(4d).

The following result in important; it asserts that the factorability criterion of Theorem 4 can be simplified.

THEOREM 5. *Let* $\Phi^c \in L^\infty(M_n)$, *and let* $\det \Phi^c(e^{j\theta}) = c \neq 0$ *for almost every* $\theta \in [0, 2\pi)$. *If there exists a solution* (L^c, R^c) *to* (4a)–(4c), *then this solution automatically satisfies Condition* (4d).

Proof. Let (L^c, R^c) be a solution to (4a)–(4c). This system of equations can obviously be rewritten

$$L^c * \Phi^c = M^c,$$

$$\Phi^c R^c = S^c *,$$

$$M^c, S^c \in H^\infty(M_n), \qquad M_0^c = S_0^c = I,$$

$$L^c, R^c \in H^\infty(M_n).$$

It follows that

$$\det [L^c(e^{j\theta})]^* c = \det M^c(e^{j\theta}) \quad \text{for almost every } \theta \in [0, 2\pi).$$

But, since L^c and M^c are in $H^\infty(M_n)$, this equation implies

$$\det M^c(e^{j\theta}) = \det I \neq 0,$$

$$\det [L^c(e^{j\theta})]^* = \det \frac{I}{c} \neq 0 \quad \text{for almost every } \theta \in [0, 2\pi).$$

Since all the entries of L^c are in $H^\infty(\mathbb{C})$, it follows that $\operatorname{adj} L^c \in H^\infty(M_n)$. Hence, $(L^c)^{-1} = c^* \operatorname{adj} L^c$ exists and is in $H^\infty(M_n)$.

The proof of $(R^c)^{-1} \in H^\infty(M_n)$ goes similarly and is omitted.

The case of Theorem 5 is a pathological case, besides those cited in [16, Theorem 3.3], of factorizations where Condition (4d) on the solution of the Hilbert problem is irrelevant.

The approach taken in this paper, which relies heavily on the Hilbert problem, can now be explained in more detail. By premultiplication and postmultiplication of Φ by suitably constructed factors, the matrix-valued function Φ is transformed into the matrix-valued function Φ^c of constant determinant, which is such that Φ is factorable if and only if Φ^c is. Theorems 4 and 5 are then applied to Φ^c, and this yields the solution of the factorization problem.

4. Pole and zero removal. This section is concerned with the transformation of the matrix-valued function Φ into the function Φ^c which is in $P(M_n)$ and of constant determinant, if this is possible. The procedure consists in eliminating the poles of Φ and the zeros of $\det \Phi$ by premultiplication of Φ by factors that are, together with their inverses, in $K^\infty(M_n)$ and by postmultiplication of Φ by factors that are, together with their inverses, in $H^\infty(M_n)$. Hence, Φ is factorable if and only if Φ^c is; moreover, the factorization of Φ can easily be determined from that of Φ^c.

THEOREM 6 (pole removal). *Let* $\Phi \in L^\infty(M_n) \cap R(M_n)$. *Then there exist factors*

$$\Lambda^\infty, P^\infty \in H^\infty(M_n) \cap R(M_n), \tag{5a}$$

having the property

$$(\Lambda^\infty)^{-1}, (P^\infty)^{-1} \in H^\infty(M_n) \cap R(M_n), \tag{5b}$$

that reduce Φ *to a matrix-valued trigonometric polynomial*:

$$\Phi^p = (\Lambda^\infty)^* \Phi P^\infty, \tag{5c}$$

$$\Phi^p \in P(M_n). \tag{5d}$$

Moreover, T_Φ is Fredholm if and only if T_{Φ^P} is; in addition, should T_Φ be Fredholm, then $\operatorname{ind}_t(\det \Phi, 0) = \operatorname{ind}_t(\det \Phi^P, 0)$.

Proof. Consider the (i, j) entry φ_{ij} of Φ. Let it have a pole $p_k (1 < |p_k| < \infty)$ of multiplicity κ. Define the elementary factor

$$P^{\infty,ij,k} = \begin{bmatrix} 1 & 0 & & & & \\ 0 & 1 & & 0 & & \\ & & \ddots & & & \\ & & & (\chi - p_k)^\kappa & & \\ & 0 & & & \ddots & \\ & & & & & 1 \end{bmatrix},$$

where $(\chi - p_k)^\kappa$ is located on the (j, j) entry. Obviously,

$$P^{\infty,ij,k}, (P^{\infty,ij,k})^{-1} \in H^\infty(M_n) \cap R(M_n),$$

and it is clear that the (i, j) entry of the matrix-valued function

$$\Phi P^{\infty,ij,k}$$

does not have poles at p_k.

Let now the entry φ_{ij} have a pole at $p_l (0 < |p_l| < 1)$ of multiplicity λ. Define the elementary factor

$$\Lambda^{\infty,ij,l} = \begin{bmatrix} 1 & 0 & & & & \\ 0 & 1 & & 0 & & \\ & & \ddots & & & \\ & & & [\chi - (p_l^{-1})^*]^\lambda & & \\ & 0 & & & \ddots & \\ & & & & & 1 \end{bmatrix},$$

where $[\chi - (\rho_l^{-1})^*]^\lambda$ is located on the (i, i) entry. Obviously,

$$\Lambda^{\infty,ij,l}, (\Lambda^{\infty,ij,l})^{-1} \in H^\infty(M_n) \cap R(M_n).$$

Moreover, the (i, j) entry of

$$(\Lambda^{\infty,ij,l})^* \Phi$$

has no poles at p_l.

The procedure for eliminating all the poles of nonzero norm from all the entries of Φ should now be clear. Define

$$P^\infty = \prod_{i,j=1}^{n} \prod_{\substack{k \\ 1<|p_k|<\infty}} P^{\infty,ij,k},$$

$$\Lambda^\infty = \prod_{i,j=1}^{n} \prod_{\substack{l \\ 0<|p_l|<1}} \Lambda^{\infty,ij,l}.$$

Obviously, the factors so defined satisfy (5).

It remains to prove the additional claims. From (5c), we have

$$\det \Phi^p(e^{j\theta}) = \det [\Lambda^\infty(e^{j\theta})]^* \det \Phi(e^{j\theta}) \det P^\infty(e^{j\theta}).$$

By construction of the factors, we have

$$\det \Lambda^\infty(e^{j\theta}) \neq 0 \quad \forall \theta \in [0, 2\pi),$$
$$\det P^\infty(e^{j\theta}) \neq 0 \quad \forall \theta \in [0, 2\pi).$$

It then follows from Proposition A.1 that T_Φ and T_{Φ^p} are Fredholm at the same time.

By construction of the factors, it also follows that $\det P^\infty$ has no zeros in the open unit disk $\mathbb{D}$ and no poles at all; on the other hand, $\det (\Lambda^\infty)^*$ has an equal number of poles and zeros in $\mathbb{D}$. Hence, by Proposition A.2, $\text{ind}_t (\det \Phi, 0) = \text{ind}_t (\det \Phi^p, 0)$. This completes the proof.

The transformation of Theorem 6 ends up with a function $\Phi^p \in P(M_n)$. The next step is to transform the matrix-valued function Φ^p into a function Φ^c in $P(M_n)$ with $\det \Phi^c$ a nonzero constant map.

THEOREM 7 (zero removal). *Let* $\Phi^p \in P(M_n)$ *with* $\det \Phi^p(e^{j\theta}) \neq 0$, $\forall \theta \in [0, 2\pi)$, *and* $\text{ind}_t (\det \Phi^p, 0) = 0$. *Then there exist factors*

$$\Lambda^0, P^0 \in H^\infty(M_n) \cap R(M_n), \tag{6a}$$

having the property

$$(\Lambda^0)^{-1}, (P^0)^{-1} \in H^\infty(M_n) \cap R(M_n), \tag{6b}$$

that reduce Φ^p *into a matrix-valued function* Φ^c *whose determinant is a nonzero constant*:

$$\Phi^c = (\Lambda^0)^* \Phi^p P^0, \tag{6c}$$

$$\Phi^c \in P(M_n), \tag{6d}$$

$$\det \Phi^c = \text{constant} \neq 0. \tag{6e}$$

Proof. Obviously, $\det \Phi^p \in P(\mathbb{C})$. Since $\det \Phi^p(e^{j\theta}) \neq 0$, $\forall \theta \in [0, 2\pi)$, $\det \Phi^p$ has no zeros on $\mathbb{T}$. Let $\{z_k : k = 1, 2, \cdots, K\}$ be the set of zeros of $\det \Phi^p$ in $\mathbb{C} \backslash \mathbb{D}$. Similarly, let $\{z_l : l = K+1, \cdots, K+L\}$ be the set of zeros of $\det \Phi^p$ in $\mathbb{D} \backslash \{0\}$. The procedure basically consists in eliminating step by step all of these zeros from $\det \Phi^p$. It yields a set of matrix-valued maps $\Phi^0 = \Phi^p, \Phi^1, \Phi^2, \cdots, \Phi^{K+L}$, where $\det \Phi^k$ does not have any zero at $z_1, z_2, \cdots, z_{k-1}$, and z_k.

To show the recursion, assume that the zero $z_k (1 < |z_k| < \infty; k = 1, 2, \cdots, K)$ of order κ has to be eliminated from $\det \Phi^{k-1}$. Since the matrix $\Phi^{k-1}(z_k)$ is singular, there exists a vector

$$u^k = \begin{bmatrix} u_1^k \\ u_2^k \\ \vdots \\ u_n^k \end{bmatrix} \neq 0,$$

such that

$$\Phi^{k-1}(z_k) u^k = 0.$$

Choose a component of u^k different from zero; let $u_i^k \neq 0$. Define the elementary factor

$$P^{0,k,1} = \begin{bmatrix} 1 & 0 & \cdots & \dfrac{u_1^k}{\chi - z_k} & \cdots & 0 \\ 0 & 1 & \cdots & \dfrac{u_2^k}{\chi - z_k} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ 0 & 0 & \cdots & \dfrac{u_i^k}{\chi - z_k} & \cdots & 0 \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \dfrac{u_n^k}{\chi - z_k} & \cdots & 1 \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \end{matrix} \quad i\text{th column}.$$

Obviously,

$$P^{0,k,1}, (P^{0,k,1})^{-1} \in H^\infty(M_n) \cap R(M_n).$$

Moreover, we have

$$\Phi^{k-1} P^{0,k,1} \in P(M_n),$$

and $\det(\Phi^{k-1} P^{0,k,1})$ has a zero of order $\kappa - 1$ at z_k. The recursion to eliminate completely the zero z_k from $\det \Phi^{k-1}$ is now obvious. It yields a set of elementary factors

$$P^{0,k,1}, P^{0,k,2}, \cdots, P^{0,k,\kappa},$$

such that

$$P^{0,k,m}, (P^{0,k,m})^{-1} \in H^\infty(M_n) \cap R(M_n), \qquad m = 1, 2, \cdots, \kappa.$$

Define

$$P^{0,k} = P^{0,k,1} P^{0,k,2} \cdots P^{0,k,\kappa}.$$

Obviously,

$$P^{0,k}, (P^{0,k})^{-1} \in H^\infty(M_n) \cap R(M_n).$$

We have

$$\Phi^{k-1} P^{0,k} \in P(M_n),$$

and $\det(\Phi^{k-1} P^{0,k})$ does not have any zero at z_k. Hence, we can write

$$\Phi^k = \Phi^{k-1} P^{0,k}.$$

Assume now that the zero $z_l (0 < |z_l| < 1; l = K+1, \cdots, K+L)$ of multiplicity λ is to be eliminated from $\det \Phi^{l-1}$. Since the matrix $\Phi^{l-1}(z_l)$ is singular, there exists a vector

$$v^l = \begin{bmatrix} v_1^l \\ v_2^l \\ \vdots \\ v_n^l \end{bmatrix} \neq 0,$$

such that

$$(v^l)^*\Phi^{l-1}(z_l)=0.$$

Let v_j^l be a component of v^l different from zero. Define the elementary factor

$$\Lambda^{0,l,1}=\begin{bmatrix} 1 & 0 & \cdots & \dfrac{v_1^l}{\chi-(z_l^{-1})^*} & \cdots & 0 \\ 0 & 1 & \cdots & \dfrac{v_2^l}{\chi-(z_l^{-1})^*} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ 0 & 0 & \cdots & \dfrac{v_j^l}{\chi-(z_l^{-1})^*} & \cdots & 0 \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \dfrac{v_n^l}{\chi-(z_l^{-1})^*} & \cdots & 1 \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \\ \\ \\ j\text{th column} \\ \\ \\ \end{matrix}.$$

Obviously,

$$\Lambda^{0,l,1},\ (\Lambda^{0,l,1})^{-1}\in H^\infty(M_n)\cap R(M_n).$$

Moreover,

$$(\Lambda^{0,l,1})^*\Phi^{l-1}\in P(M_n),$$

and $\det[(\Lambda^{0,l,1})^*\Phi^{l-1}]$ has a zero of order $\lambda-1$ at z_l. The recursion to eliminate completely the zero at z_l is now obvious. It yields a set of factors

$$\Lambda^{0,l,1},\ \Lambda^{0,l,2},\cdots,\Lambda^{0,l,\lambda},$$

such that

$$\Lambda^{0,l,m},\ (\Lambda^{0,l,m})^{-1}\in H^\infty(M_n)\cap R(M_n),\qquad m=1,2,\cdots,\lambda.$$

Define

$$\Lambda^{0,l}=\Lambda^{0,l,1}\Lambda^{0,l,2}\cdots\Lambda^{0,l,\lambda}.$$

We have

$$(\Lambda^{0,l})^*\Phi^{l-1}\in P(M_n),$$

and $\det[(\Lambda^{0,l})^*\Phi^{l-1}]$ has no zeros at z_l. Hence, we can write

$$\Phi^l=(\Lambda^{0,l})^*\Phi^{l-1}.$$

The general procedure to eliminate all the zeros in $\mathbb{C}\backslash\{0\}$ should now be clear. Define

$$P^0=P^{0,1}P^{0,2}\cdots P^{0,K},$$
$$\Lambda^0=\Lambda^{0,K+1}\Lambda^{0,K+2}\cdots\Lambda^{0,K+L}.$$

It is claimed that these factors satisfy (6). We have

$$\Phi^{K+L}=(\Lambda^0)^*\Phi^pP^0.$$

Obviously,

$$\Phi^{K+L} \in P(M_n),$$

and $\det \Phi^{K+L}$ has no zeros in $\mathbb{C}\backslash\{0\}$.

Let us now prove that $\det \Phi^{K+L}$ is, in fact, a nonzero constant map. Since $\det \Phi^p(e^{j\theta}) \neq 0, \forall \theta \in [0, 2\pi)$, it follows from the construction of the factors Λ^0 and P^0 that $\det \Phi^{K+L}(e^{j\theta}) \neq 0, \forall \theta \in [0, 2\pi)$. Hence $\det \Phi^{K+L}$ is not identically zero. By construction of Λ^0 and P^0, $\det P^0$ has no poles in $\mathbb{D}$ and no zeros at all; on the other hand, $\det (\Lambda^0)^*$ has an equal number of poles and zeros in $\mathbb{D}$. By Proposition A.2, it then follows that

$$\operatorname{ind}_t (\det \Phi^{K+L}, 0) = \operatorname{ind}_t (\det \Phi^p, 0).$$

Hence,

$$\operatorname{ind}_t (\det \Phi^{K+L}, 0) = 0.$$

But $\det \Phi^{K+L} \in P(\mathbb{C})$, and $\det \Phi^{K+L}$ has no zeros in $\mathbb{C}\backslash\{0\}$. Hence, $\det \Phi^{K+L}$ has an equal number of zeros and poles at the origin. In other words, $\det \Phi^{K+L}$ is a nonzero constant map, and one has $\Phi^{K+L} = \Phi^c$. This completes the proof.

The situation is summarized by the following theorem:

THEOREM 8. *Let* $\Phi \in L^\infty(M_n) \cap R(M_n)$, *with* $\det \Phi(e^{j\theta}) \neq 0, \forall \theta \in [0, 2\pi)$, *and with* $\operatorname{ind}_t (\det \Phi, 0) = 0$. *Let* $\Phi^c \in P(M_n)$, *where* $\det \Phi^c$ *is a nonzero constant map, result from the application of the algorithms of Theorems* 6 *and* 7 *to* Φ. *Then* Φ *admits a strong (weak) analytic factorization if and only if* Φ^c *admits a strong (weak) analytic factorization.*

Proof. Assume Φ^c admits the strong analytic factorization $\Phi^c = (\Lambda^c)^* P^c$. It is then readily verified that a strong analytic factorization of Φ is given by $\Lambda^* P$, where $\Lambda = \Lambda^c(\Lambda^\infty \Lambda^0)^{-1}$ and $P = P^c(P^\infty P^0)^{-1}$.

Now, let Φ admit the strong analytic factorization $\Lambda^* P$; then a strong analytic factorization of Φ^c is given by $(\Lambda^c)^* P^c$, where $\Lambda^c = \Lambda\Lambda^\infty\Lambda^0$ and $P^c = PP^\infty P^0$.

The case of the weak analytic factorization is proved the same way.

It thus remains to check whether $\Phi^c \in P(M_n)$, where $\det \Phi^c$ is a nonzero constant map, is factorable, and, if a factorization exists, to determine the factors. These are the topics of the next section.

5. Factorization of a matrix-valued function of constant determinant. The solution to the problem of factoring a matrix-valued map whose determinant is a nonzero constant relies completely on Theorems 4 and 5 of § 3.

We first prove that, in the constant determinant case, weak and strong analytic factorizations are equivalent.

THEOREM 9. *Let* $\Phi^c \in P(M_n)$, *with* $\det \Phi^c(e^{j\theta}) = c \neq 0$ *for almost every* $\theta \in [0, 2\pi)$. *Then* Φ^c *has a weak analytic factorization if and only if it has a strong analytic factorization.*

Proof. If Φ^c admits a strong analytic factorization, it obviously admits a weak analytic factorization.

Conversely, if Φ^c admits a weak analytic factorization, then, by Proposition 1, the Toeplitz operator T_{Φ^c} is invertible. This guarantees the existence of a solution $R^c \in H^2(M_n)$ to the equation $P_{H^2(M_n)}(\Phi^c R^c) = I$.

Let us show that this result can be strengthened to $R^c \in H^\infty(M_n)$. We know that there exists a solution $R^c \in H^2(M_n)$ to the equation $\Phi^c R^c = S^{c*}$, for some $S^c \in H^2(M_n)$, with $S_0^c = I$. Since $\Phi^c \in P(M_n)$ and $R^c \in H^2(M_n)$, it is clear that S^c has a Fourier transform supported on a finite set of positive integers. Thus, we write $S^c =$

$I+\sum_{k=1}^{K} S_k^c \chi^k$. Thus, there exists a solution $R^c \in H^2(M_n)$ to the equation $R^c = c^{-1}(\text{adj } \Phi^c)(I+\sum_{k=1}^{K} S_k^{c*}\chi^{-k})$. Since adj $\Phi^c \in P(M_n)$, it is clear that the solution R^c has a Fourier transform supported on a finite set of positive integers. Hence, $R^c \in H^\infty(M_n)$.

Thus there exists a solution R^c to (4b), (4c). From Proposition 1, it is easily seen that, as a consequence of the invertibility of T_{Φ^c}, $T_{\Phi^{c*}}$ is also invertible. This guarantees the existence of a solution L^c to (4a), (4c). Thus there exists a solution (L^c, R^c) to the system of equations (4a)–(4c). But, by Theorem 5, this solution automatically satisfies Condition (4d). Hence, there exists a solution (L^c, R^c) to the system of matrix equations (4a)–(4d). It then follows from Theorem 4 that Φ^c admits a strong analytic factorization.

Now, we can prove that, in the rational case, weak and strong analytic factorizations are equivalent.

THEOREM 10. *Let* $\Phi \in L^\infty(M_n) \cap R(M_n)$. *Then* Φ *admits a weak analytic factorization if and only if it admits a strong analytic factorization.*

Proof. If Φ admits a strong analytic factorization, then it obviously admits a weak analytic factorization.

Conversely, if Φ has a weak analytic factorization, then, by Proposition 1, the Toeplitz operator T_Φ is invertible. By Proposition A.4, $\det \Phi(e^{j\theta}) \neq 0$, $\forall \theta \in [0, 2\pi)$, and $\text{ind}_t (\det \Phi, 0) = 0$. Hence, by Theorems 6 and 7, Φ^c exists. Since Φ has a weak analytic factorization, so has Φ^c by Theorem 8. By Theorem 9, since Φ^c has a weak analytic factorization, it has a strong analytic factorization. By Theorem 8, Φ has a strong analytic factorization.

In the remainder of this paper, we shall thus primarily be concerned with the strong analytic factorization.

We now proceed to the problem of the strong analytic factorization of Φ^c. We need some preliminaries, however. Since $\Phi^c \in P(M_n)$ its Fourier expansion takes the form

$$\Phi^c = \sum_{k=-M}^{N} \Phi_k^c \chi^k, \qquad M = 0, 1, 2, \cdots, \quad N = 0, 1, 2, \cdots. \tag{7a}$$

Define

$$\Sigma = \sum_{k=0}^{N+M} \Phi_{k-M}^c \chi^k, \tag{7b}$$

$$\Omega = \sum_{k=0}^{N+M} (\Phi_{N-k}^c)^* \chi^k. \tag{7c}$$

It is easily seen that

$$\Phi^c = \frac{\Sigma}{\chi^M}, \qquad (\Phi^c)^* = \frac{\Omega}{\chi^N}; \tag{7d}$$

moreover,

$$\Sigma, \Omega \in H^\infty(M_n) \cap P(M_n).$$

We need the following lemma:

LEMMA 1. *Let* $\Sigma, \Omega \in H^\infty(M_n)$; $L^c, R^c \in H^2(M_n)$; *and* $M = 0, 1, 2, \cdots, N = 0, 1, 2, \cdots$. *If*

$$P_{H^2(M_n)}\left(\frac{\Omega L^c}{\chi^N}\right) = I, \tag{8a}$$

(8b) $$P_{H^2(M_n)}\left(\frac{\Sigma R^c}{\chi^M}\right) = I,$$

then

$$\Omega L^c = \sum_{k=0}^{N-1} A_k \chi^k + I\chi^N, \qquad A_k \in M_n; \qquad k = 0, \cdots, N-1,$$

$$\Sigma R^c = \sum_{k=0}^{M-1} B_k \chi^k + I\chi^M, \qquad B_k \in M_n; \qquad k = 0, \cdots, M-1,$$

and conversely.

Proof. This result is easily proved by writing out the Fourier expansions.

The following theorem asserts that, if a solution to the system of equations (4a)–(4c) exists, then it has a rather simple form.

THEOREM 11. *Let* $\Phi^c \in P(M_n)$, *with* $\det \Phi^c(z) = c, \forall z \in \mathbb{C}$. *Then, if the solution* (L^c, R^c) *to the system of equations* (4a)–(4c) *exists, it has the form*

$$L^c = \sum_{k=0}^{l} L_k^c \chi^k,$$

$$R^c = \sum_{k=0}^{r} R_k^c \chi^k,$$

where

$$l = \operatorname{degr} \operatorname{adj} \Omega + (1-n)N,$$

$$r = \operatorname{degr} \operatorname{adj} \Sigma + (1-n)M.$$

(degr adj $\Omega(\Sigma)$ *is by definition the largest power of* χ *in the Fourier expansion of the adjoint of* $\Omega(\Sigma)$.)

Proof. Let (L^c, R^c) be the solution to the system of equations (4a)–(4c). From (7), it is easily seen that (8a)–(8b) is merely a rewriting of (4a)–(4b). Then, by Lemma 1, we have

$$L^c = \Omega^{-1}\left(\sum_{k=0}^{N-1} A_k \chi^k + I\chi^N\right).$$

Since $\det \Phi^c(z) = c \neq 0$, we have $\det \Omega = c^* \chi^{nN}$. Hence,

$$L^c = \frac{\operatorname{adj} \Omega}{c^* \chi^{nN}}\left(\sum_{k=0}^{N-1} A_k \chi^k + I\chi^N\right).$$

By hypothesis, $L^c \in H^\infty(M_n)$; but the denominator of L^c has a zero of order nN at $0 \in \mathbb{D}$; thus, one can choose a set $\{A_k \in M_n : k = 0, 1, \cdots, N-1\}$ such that the appropriate pole-zero cancellation occurs. It then follows that L^c has nonzero Fourier coefficients up to the order degr adj $\Omega + N - nN$, which is necessarily in $\mathbb{Z}^+$.

The proof for R^c is the same and is omitted.

We can now write out the final result.

THEOREM 12. *Let* $\Phi^c \in P(M_n)$, *with* $\det \Phi^c$ *a nonzero constant map. Then* Φ^c *admits a strong analytic factorization* $\Phi^c = (\Lambda^c)^* P^c$ *if and only if there exists a* (*unique*) *solution* $(\{L_k^c \in M_n : k = 0, 1, \cdots, l\}, \{R_k^c \in M_n : k = 0, 1, \cdots, r\})$ *to the system of matrix*

equations

$$\begin{bmatrix} \Phi_0^{c*} & \Phi_1^{c*} & \Phi_2^{c*} & \cdots & \cdot & \cdot \\ \Phi_{-1}^{c*} & \Phi_0^{c*} & \Phi_1^{c*} & \cdots & & \\ \vdots & \vdots & \vdots & & & \\ \Phi_{-M}^{c*} & \Phi_{-M+1}^{c*} & \Phi_{-M+2}^{c*} & \cdots & & \\ 0 & \Phi_{-M}^{c*} & \Phi_{-M+1}^{c*} & \cdots & & \\ 0 & 0 & \Phi_{-M}^{c*} & \cdots & & \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \Phi_{-M}^{c*} & \Phi_{-M+1}^{c*} \\ 0 & 0 & 0 & \cdots & 0 & \Phi_{-M}^{c*} \end{bmatrix} \begin{bmatrix} L_0^c \\ L_1^c \\ \\ \cdot \\ \cdot \\ \cdot \\ \\ L_{l-1}^c \\ L_l^c \end{bmatrix} = \begin{bmatrix} I \\ 0 \\ \\ \cdot \\ \cdot \\ \cdot \\ \\ 0 \\ 0 \end{bmatrix}, \tag{9a}$$

$$\begin{bmatrix} \Phi_0^c & \Phi_{-1}^c & \Phi_{-2}^c & \cdots & \cdot & \cdot \\ \Phi_1^c & \Phi_0^c & \Phi_{-1}^c & \cdots & & \\ \vdots & \vdots & \vdots & & & \\ \Phi_N^c & \Phi_{N-1}^c & \Phi_{N-2}^c & \cdots & & \\ 0 & \Phi_N^c & \Phi_{N-1}^c & \cdots & & \\ 0 & 0 & \Phi_N^c & \cdots & & \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \Phi_N^c & \Phi_{N-1}^c \\ 0 & 0 & 0 & \cdots & 0 & \Phi_N^c \end{bmatrix} \begin{bmatrix} R_0^c \\ R_1^c \\ \\ \cdot \\ \cdot \\ \cdot \\ \\ R_{r-1}^c \\ R_r^c \end{bmatrix} = \begin{bmatrix} I \\ 0 \\ \\ \cdot \\ \cdot \\ \cdot \\ \\ 0 \\ 0 \end{bmatrix}. \tag{9b}$$

Moreover, should the solution exist, the factors are then given by

$$\Lambda^c = (L^c)^{-1}, \tag{10a}$$

$$P^c = R_0^c (R^c)^{-1}, \tag{10b}$$

where

$$L^c = \sum_{k=0}^{l} L_k^c \chi^k, \tag{10c}$$

$$R^c = \sum_{k=0}^{r} R_k^c \chi^k. \tag{10d}$$

Proof. By Theorems 4 and 5, Φ^c admits a strong analytic factorization if and only if there exists a solution (L^c, R^c) to the system of matrix equations (4a)–(4c). Using (7a) and Theorem 11, it is easily seen that (9a)–(9b) is merely a rewriting of (4a)–(4b). Hence, Φ^c has a strong analytic factorization if and only if there exists a solution (L^c, R^c) to the system of matrix equations (9a)–(9b).

The additional claims follow from a general result of factorization theory; see McNabb and Schumitzky [16, Theorem 3.1 and Corollary 3.2].

The solution to the problem of factoring the matrix-valued function Φ^c of constant determinant is thus given by the *finite* system of linear matrix equations (9a)–(9b). Notice that the system of equations (3a)–(3b) of the primary Hilbert problem is, in general, an *infinite* system of linear matrix equations. The problem of the existence and the computation of the solution to (9a)–(9b) should not cause problems because the system of equations is *linear* and *finite.*

A system of linear matrix equations similar to (9a) appears in a celebrated prediction problem; see Levinson [22], Wiener [23, Appendix], and Kailath [24, § 7]. The system of equations of prediction theory can be solved recursively and efficiently via Levinson's algorithm. However, (9a) *cannot*, in general, be solved via Levinson's algorithm. The reason is that the application of Levinson's algorithm to (9a) requires the existence of a solution to the associated reverse time system of equations, which might not have a solution even if a solution to (9a) exists. A deeper reason is that strong analytic factorability is *not* invariant under time reversal [25]. To see this, observe that the function $\Phi^c = \begin{pmatrix} \chi & 1 \\ 0 & \chi^{-1} \end{pmatrix}$ is factorable; then observe that the function obtained after time reversal, namely, $\begin{pmatrix} \chi^{-1} & 1 \\ 0 & \chi \end{pmatrix}$, is not factorable. (Hint: use Theorem 12.)

6. Summary of the algorithm. We briefly summarize the algorithm to determine whether $\Phi \in L^\infty(M_n) \cap R(M_n)$ is factorable and to compute the factors if they exist.

The first step is to check whether $\det \Phi(e^{j\theta}) \neq 0$, $\forall \theta \in [0, 2\pi)$. If this condition is not verified, then Φ is not factorable (Proposition A.1). If this condition is satisfied, then check whether $\operatorname{ind}_t (\det \Phi, 0) = 0$. If no, Φ is not factorable (Proposition A.4). If yes, then go through the algorithms of Theorems 6 and 7 to compute Λ^∞, P^∞, Λ^0, P^0, and Φ^c. By Theorem 8, the problem is now to factor Φ^c. Thus, check whether there exists an appropriate solution (L^c, R^c) to the system of matrix equations (9a)–(9b). If no, Φ is not factorable (Theorem 12). If yes, compute the solution (L^c, R^c) to (9a)–(9b). Then compute Λ^c and P^c using (10). Then a strong analytic factorization of Φ is given by $\Phi = \Lambda^* P$, where $\Lambda = \Lambda^c(\Lambda^\infty \Lambda^0)^{-1}$ and $P = P^c(P^\infty P^0)^{-1}$.

Observe the following result:

THEOREM 13. *Let* $\Phi \in L^\infty(M_n) \cap R(M_n)$, *with* $\det \Phi(e^{j\theta}) \neq 0$, $\forall \theta \in [0, 2\pi)$, *and* $\operatorname{ind}_t (\det \Phi, 0) = 0$. *If there exists a solution* (L, R) *to the system of equations* (3a)–(3c), *then this solution automatically satisfies Condition* (3d).

Proof. Since $\det \Phi(e^{j\theta}) \neq 0$, $\forall \theta \in [0, 2\pi)$, and $\operatorname{ind}_t (\det \Phi, 0) = 0$, by Theorems 6 and 7, Φ can be reduced to Φ^c of constant determinant.

Let (L, R) be the solution to (3a)–(3c). Equation (3a) can be rewritten

$$L^* \Phi = M,$$

$$M \in H^\infty(M_n), \qquad M_0 = I.$$

By Theorems 6 and 7, this can be rewritten

$$L^*[(\Lambda^\infty \Lambda^0)^*]^{-1} \Phi^c (P^\infty P^0)^{-1} = M;$$

we further have

$$[(\Lambda^\infty \Lambda^0)^{-1} L]^* \Phi^c = M(P^\infty P^0).$$

This last equation is equivalent to (4a). By working out (3b) the same way, one finds that

$$((\Lambda^\infty \Lambda^0)^{-1} L[(P_0^\infty P_0^0)^{-1}]^*, (P^\infty P^0)^{-1} R[(\Lambda_0^\infty \Lambda_0^0)^{-1}]^*) = (L^c, R^c)$$

is an appropriate solution to (4a)–(4c). But, by Theorem 5, this solution automatically satisfies (4d). Hence, $(L^c)^{-1}$ and $(R^c)^{-1}$ exist and are in $H^\infty(M_n)$. It then follows that L^{-1} and R^{-1} exist and are in $H^\infty(M_n)$.

7. The discrete spectral factorization. The discrete spectral factorization is the problem of the existence and the computation of a strong analytic factorization of Φ, when it is subject to the constraints

$$\Phi = \Phi^* \in L^\infty(M_n) \cap R(M_n),$$

$$\Phi(e^{j\theta}) > 0 \quad \forall \theta \in [0, 2\pi).$$

The discrete spectral factorization is considered in Motyka and Cadzow [20]. However, this paper contains a gap; this will be proved later by a counterexample. Another way to look at the discrete spectral factorization is to start from the continuous-time spectral factorization, and then to apply to bilinear transformation in order to recover the corresponding discrete-time result; this approach was taken by Anderson et al. [21]; it, however, fails to provide a deep insight into the problem. In view of these facts, and although it is widely used, the discrete spectral factorization deserves some attention.

THEOREM 14. *Let* $\Phi = \Phi^* \in L^\infty(M_n) \cap R(M_n)$, *with* $\Phi(e^{j\theta}) > 0, \forall \theta \in [0, 2\pi)$. *Then* Φ *always admits a strong analytic factorization* $\Lambda^*\Lambda$.

Proof. Since $\Phi = \Phi^*$, and $\Phi(e^{j\theta}) > 0$ for all $\theta \in [0, 2\pi)$, by the definition of the Laurent operator, it is easily seen that the operator L_Φ is positive definite self-adjoint. Then the Toeplitz operator T_Φ is positive definite self-adjoint (to see this, it is useful to consider the infinite matrix representation of L_Φ and T_Φ). Then T_Φ does not have zero in its spectrum. Hence, T_Φ is invertible. By Proposition 1, Φ admits a weak analytic factorization; furthermore, by Theorem 10, Φ has a strong analytic factorization.

It remains to prove that any strong analytic factorization has the form $\Phi = \Lambda^*\Lambda$.

Since $\Phi(e^{j\theta}) > 0, \forall \theta \in [0, 2\pi)$, T_Φ is Fredholm by Proposition A.1. Moreover, since $\det \Phi(e^{j\theta})$ is real and strictly positive for all $\theta \in [0, 2\pi)$, by the definition of the topological index, we have $\text{ind}_t(\det \Phi, 0) = 0$. Then, by Theorems 6 and 7, Φ can be reduced to a matrix-valued function $\Phi^c = \Phi^{c*}$ of constant determinant by elementary factors that are such that $\Lambda^\infty = P^\infty$ and $\Lambda^0 = P^0$. By Theorem 8, Φ^c is then factorable. By Theorem 12, there exists a solution (L^c, R^c) to (9a)–(9b). Since $\Phi^c = \Phi^{c*}$, (9a) is the same as (9b). Hence, by Theorem 12, $\Lambda^c = P^c$. Finally, a strong analytic factorization of Φ is given by $[\Lambda^c(\Lambda^\infty\Lambda^0)^{-1}]^*[\Lambda^c(\Lambda^\infty\Lambda^0)^{-1}]$.

This result is not novel; it is contained in Saeks [17, § III, Lemma]; it is contained partially in Nagy and Foias [18, Ch. V, § 7], in Helson [3, Lecture XI], and in Rosenblum and Rovnyak [19, Theorem 3.1]. It is, however, interesting to see how this particular result can be recovered in our more general setting.

In the case of the spectral factorization, (9a), which is the same as (9b), can be solved via Levinson's algorithm. Indeed, in this case, a solution to the reverse time system of equations exists; this is related to the fact that spectral factorability is preserved under time reversal [25]. Interestingly, it can be shown that the solution of Levinson's algorithm converges to the *exact* solution of (9a) after a *finite* number of steps.

In [20], which deals with the discrete spectral factorization, the reduction of Φ to Φ^c contains several unclear features. Moreover, the proof of the factorability of Φ^c [20, § V, Step D] is definitely wrong. Indeed, the proof uses *only* the constancy of $\det \Phi^c$ and the property $\Phi^c = \Phi^{c*}$, and *not* the fact that $\Phi^c(e^{j\theta}) > 0$ for all $\theta \in [0, 2\pi)$. Thus, if this proof were correct, then the function $\Phi^c = \begin{pmatrix} 0 & \chi \\ \chi^{-1} & 0 \end{pmatrix}$ would admit a strong analytic

factorization. But the application of the criterion of Theorem 12 proves that this function is not factorable! Also, the algorithm of [20, § V, Step D] to factor Φ^c is unable to provide the factors in all cases; to show this, we invite the reader to try to factor $\Phi^c = \begin{pmatrix} 1 & \chi \\ \chi^{-1} & 2 \end{pmatrix}$ using the algorithm of [20, § V, Step D]. In view of these facts, the spectral factorization algorithms that do not use the condition $\Phi(e^{j\theta}) > 0$, $\forall \theta \in [0, 2\pi)$ should be reviewed with some care.

8. Conclusions. We have offered in this paper the basic results and algorithms concerning the analytic factorization of a rational matrix-valued function. The importance of this problem in the discrete-time linear-quadratic optimal control problem has been clearly shown in [26] and [27].

The approach that has been taken here is rather analytical. It would be interesting to set up a more algebraically oriented approach. The key idea is provided in [16]. From [16, §§ 5 and 6] and Theorem 13, it can be shown that the strong analytic factorability of Φ is equivalent to the invertibility of an imbedded element $\tau(\Phi)$, where τ is a map: $L^\infty(M_n) \cap R(M_n) \to K^\infty(M_n) \otimes H^\infty(M_n)$, a suitable product being defined on the tensor algebra. It turns out that the imbedded element is a matrix defined over a *noncommutative* ring. This shows the algebraic problem underlying the strong analytic factorization. In a further paper, we shall go into more detail through that.

Finally, it is worth mentioning that a system of linear matrix equations like (9a) has several system-theoretic interpretations [25]; we also leave these topics to a further paper.

Appendix A. Fredholmness and index theory. The transformation of Φ into the matrix-valued function Φ^c of constant determinant needs some elements of index theory as applied to Fredholm Toeplitz operators. The necessary results are briefly reviewed; for more details, see Douglas [4, Ch. 5], [5, Introduction and Lecture 1], Gohberg and Krein [15], [28], Coburn [29], [30], and Atkinson [31].

The results of this appendix are presented within the framework of the inversion of Toeplitz operators. We could equally have worked within the factorization framework; we, however, feel that the former approach is simpler.

A separable Hilbert space operator A is said to be *Fredholm* if it has closed range and if Ker (A) and Ker (A^*) are finite-dimensional. Should A be Fredholm, then its *analytical index* is defined by

$$\text{ind}_a\,(A) = \dim \text{Ker}\,(A) - \dim \text{Ker}\,(A^*).$$

The following proposition is proved in Douglas [5, Theorem 2]:

PROPOSITION A.1. *Let* $\Phi \in C(M_n)$. *Then* T_Φ *is Fredholm if and only if* $\det \Phi(e^{j\theta}) \neq 0$, $\forall \theta \in [0, 2\pi)$.

Let $\varphi \in C(\mathbb{C})$, and let $\varphi : \mathbb{T} \to \mathbb{C}\backslash\{0\}$. Then the *topological index* $\text{ind}_t\,(\varphi, 0)$ is defined as the winding number of the image of $\mathbb{T}$ under φ with respect to the origin. Should $\text{ind}_t\,(\varphi, 0) = 0$, then the map $\varphi : \mathbb{T} \to \mathbb{C}\backslash\{0\}$ is homotopic to a constant map. Notice the following easily proved result:

PROPOSITION A.2. *Let* $\varphi \in C(\mathbb{C}) \cap R(\mathbb{C})$, *and let* $\varphi : \mathbb{T} \to \mathbb{C}\backslash\{0\}$. *Write* $\varphi = \mu/\nu$, *where*

$$\mu = \sum_{k=0}^{\alpha} \mu_k \chi^k, \qquad 0 \leqq \alpha < \infty,$$

$$\nu = \sum_{k=0}^{\beta} \nu_k \chi^k, \qquad 0 \leqq \beta < \infty.$$

Then $\text{ind}_t(\varphi, 0)$ *equals the number of* (*repeated*) *zeros of* (*the extension of*) μ *inside the open unit disk* $\mathbb{D}$ *minus the number of* (*repeated*) *zeros of* (*the extension of*) ν *inside* $\mathbb{D}$.

The following result is due to Douglas [5, Lecture 1]:

PROPOSITION A.3. *Let* $\Phi \in C(M_n)$, *and let* T_Φ *be Fredholm. Then*

$$\text{ind}_a(T_\Phi) = -\text{ind}_t(\det \Phi, 0).$$

Propositions A.1 and A.3 readily yield the following:

PROPOSITION A.4. *Let* $\Phi \in C(M_n)$. *Necessary for* T_Φ *to be invertible is that* $\det \Phi(e^{j\theta}) \neq 0, \forall \theta \in [0, 2\pi)$, *and* $\text{ind}_t(\det \Phi, 0) = 0$.

This necessary condition for invertibility is not, in general, also sufficient. Indeed, a Fredholm Toeplitz operator whose associated matrix-valued function has determinant homotopic to a constant may have a nontrivial kernel [5, Lecture 1], [32] and may, hence, be noninvertible. In fact, the toughest problem in the inversion of a Toeplitz operator is to determine under what conditions a Fredholm Toeplitz operator with determinant of the associated matrix-valued function homotopic to a constant is invertible; this problem is treated, within the factorization framework, in § 5.

REFERENCES

[1] K. HOFFMAN, *Banach Spaces of Analytic Functions*, Prentice-Hall, Englewood Cliffs, NJ, 1962.

[2] P. L. DÜREN, *Theory of H^p Spaces*, Academic Press, New York, 1970.

[3] H. HELSON, *Lectures on Invariant Subspaces*, Academic Press, New York, 1964.

[4] R. G. DOUGLAS, *Banach Algebra Techniques in Operator Theory*, Academic Press, New York, 1972.

[5] R. G. DOUGLAS, *Banach Algebra Techniques in the Theory of Toeplitz Operators*, Regional Conference Series, vol. 15, American Mathematical Society, Providence, R.I., 1972.

[6] V. GRENANDER AND G. SZEGÖ, *Toeplitz Forms and their Applications*, University of California Press, Berkeley and Los Angeles, 1958.

[7] H. WIDOM, *Toeplitz matrices*, in *Studies in Real and Complex Analysis*, I. I. Hirshman, Jr., ed., Prentice-Hall, Englewood Cliffs, NJ, 1965.

[8] ———, *Inversion of Toeplitz matrices. II*, Illinois J. Math., 4 (1960), pp. 88–99.

[9] H. R. POUSSON, *Systems of Toeplitz operators on H^2*, Proc. Amer. Math. Soc., 19 (1968), pp. 603–608.

[10] ———, *Systems of Toeplitz operators on H^2. II*, Trans. Amer. Math. Soc., 113 (1968), pp. 527–536.

[11] A. DEVINATZ, *Toeplitz operators on H^2 spaces*, Trans. Amer. Math. Soc., 112 (1964), pp. 304–317.

[12] M. RABINDRANATHAN, *On the inversion of Toeplitz operators*, J. Math. Mech., 19 (1969). pp. 195–206.

[13] I. C. GOHBERG AND M. G. KREIN, *On the factorization of operators in Hilbert space*, Amer. Math. Soc. Transl. Ser., 51 (1966), pp. 155–188.

[14] M. G. KREIN, *Integral equations on the half-line with kernels depending on the difference of arguments*, Amer. Math. Soc. Transl. Ser., 22 (1962), pp. 163–288.

[15] I. C. GOHBERG AND M. G. KREIN, *Systems of integral equations with kernels depending on the difference of arguments*, Amer. Math. Soc. Transl. Ser., 14 (1960), pp. 217–288.

[16] A. MCNABB AND A. SCHUMITZKY, *Factorization of operators–I: Algebraic theory and examples*, J. Funct. Anal., 9 (1972), pp. 262–295.

[17] R. SAEKS, *The factorization problem—a survey*, Proc. IEEE, 64 (1976) pp. 90–95.

[18] B. SZ.-NAGY AND C. FOIAS, *Harmonic Analysis of Operators on a Hilbert Space*, North-Holland, Amsterdam, 1970.

[19] M. ROSENBLUM AND J. ROVNYAK, *The factorization problem for nonnegative operator-valued functions*, Bull. Amer. Math. Soc., 77 (1971), pp. 287–318.

[20] P. R. MOTYKA AND J. A. CADZOW, *The factorization of discrete-process spectral matrices*, IEEE Trans. Automat. Control, AC-12 (1967), pp. 698–707.

[21] B. D. O. ANDERSON, K. L. HITZ, AND N. D. DIEM, *Recursive algorithm for spectral factorization*, IEEE Trans. Circuits and Systems, CAS-21, (1974), pp. 742–750.

[22] N. LEVINSON, *The Wiener RMS (root mean square) error criterion in filter design and prediction*, J. Math. Phys., 25 (1947), pp. 261–278.

[23] N. WEINER, *The Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications*, John Wiley, New York, 1949.

[24] T. KAILATH, *A view of three decades of linear filtering theory*, IEEE Trans. Inform. Theory, IT-20, (1974), pp. 146–181.

[25] S. Y. KUNG, *private communication.*

[26] E. A. JONCKHEERE AND L. M. SILVERMAN, *The general discrete-time linear-quadratic control problem*, Proc. IEEE Conf. Decision and Control, New Orleans, Louisiana, December 1977, pp. 1239–1244.

[27] ———, *Spectral theory of the linear-quadratic control problem: discrete-time single-input case*, IEEE Trans. Circuits and Systems, Special Issue on Mathematical Foundation of System Theory, CAS-25, (1978), pp. 810–825.

[28] I. C. GOHBERG AND M. G. KREIN, *The basic propositions on defect numbers, root numbers, and indices of linear operators*, Amer. Math. Soc. Transl. Ser., 13 (1960), pp. 185–264.

[29] L. A. COBURN, *The C^*-algebra generated by an isometry*, Bull. Amer. Math. Soc., 73 (1967), pp. 722–726.

[30] ———, *The C^*-algebra generated by an isometry. II.* Trans. Amer. Math. Soc., 137 (1969), pp. 211–217.

[31] F. V. Atkinson, *The normal solubility of linear equations in normed spaces*, Mat. Sb, 28 (1951), pp. 3–14.

[32] S. PATTANAYAK, *On Toeplitz operators on Quarter Plane with Matrix-valued Symbols*, Ph.D. dissertation, Department of Mathematics, State University of New York, Stony Brook, 1972.

[33] E. A. JONCKHEERE AND L. M. SILVERMAN, *Spectral theory of the linear-quadratic optimal control problem: a new algorithm for spectral computations*, IEEE Trans. Automat. Control, to appear.

[34] ———, *Spectral theory of linear control and estimation problems*, International Symposium on Systems Optimization and Analysis IRIA, Paris, France, Dec. 11–13, 1978; see also Lecture Notes in Control and Information Sciences, Springer-Verlag, Berlin-Heidelberg-New York, 14 (1979), pp. 99–109.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 3, May 1981

0363-0129/81/1903-0001 $01.00/0

A DEGREE METHOD FOR FREE BOUNDARIES IN STOCHASTIC CONTROL*

IOANNIS KARATZAS† AND VÁCLAV E. BENEŠ‡

Abstract. In stochastic control problems with a bounded control set, the Bellman-Hamilton-Jacobi equation leads to two-sided free-boundary problems for the switching surfaces, expressible as an equivalent set of integral equations containing the boundary functions in a very implicit way that seems to preclude the standard method used in the Stefan problem. It is natural then to try to use the topological Leray-Schauder methods to study the properties of solutions. We apply such an approach to the sample problem: $\min_u E\int_0^T [f(x_t)+|u(x_t, t)|]\,dt$, subject to $dx_t = u(x_t, t)\,dt + dw_t$, $|u| \leqq 1$, with w_t a Wiener process. The absolute value cost $|u|$ leads to finding the boundaries of a "dead zone" in (x, t)-space that separates the zones $u = \pm 1$ for the optimal u. The a priori bounds requisite for the Leray-Schauder approach come from usual probabilistic and PDE estimates. Then the integral equations are shown to have the form (homeomorphism + compact) for which a degree theory is available. Finally, a simple homotopy shows that the free boundary is continuously differentiable; separate arguments establish its uniqueness and monotonicity.

CONTENTS

1. Introduction. We use a sample problem to describe topological methods for certain questions in stochastic control. In all these questions, the Bellman-Hamilton-Jacobi equation of dynamic programming suggests a bang-bang optimal law and so, after a transformation (similar to that used for Stefan's problem) of BHJ to an equivalent system of integral equations, a search for the optimal "switching surfaces" or "free boundaries". To the integral equations we apply Leray-Schauder methods thus: first, standard PDE and probabilistic estimates provide the requisite a priori bounds; then, a natural homotopy establishes properties of the free boundaries and the value function.

The sample problem concerns the linear control of the stochastic differential equation $dx_t = u(x_t, t)\,dt + dw_t$ over a finite time horizon T, where $\{w_t; 0 \leqq t \leqq T\}$ is a Wiener process on an underlying probability space $(\Omega, \mathbf{F}, P)$. This is sometimes called the "controlled noisy integrator," and is depicted in block diagram form in Fig. 1.

* Received by the editors November 8, 1979, and in revised form May 6, 1980.

† Department of Mathematical Statistics, Columbia University, New York, New York 10027. The results in this paper are drawn from this author's doctoral thesis, Karatzas [1979].

‡ Bell Laboratories, Murray Hill, New Jersey 07974.

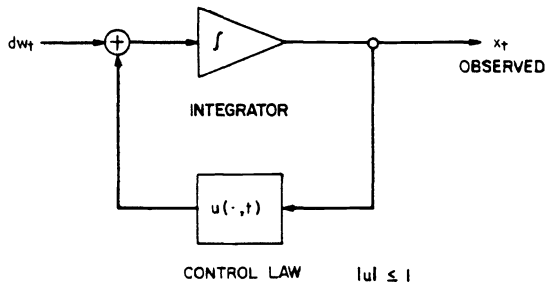


FIG. 1. *Control of noisy integrator* $dx_t = u(x_t, t)\,dt + dw_t$.

There will be two kinds of cost incurred in this problem. First, one pays $f(x_T)$ for being in the wrong place at the final time; second, one pays $|u|$ per unit time for using the control law. The control problem is to choose a law u: $\mathbf{R}\times[0, T]\to[-1, 1]$ so as to minimize the expected total cost. Intuitively speaking, the function of the controller u is to push the output to the left if the latter is too positive or to the right if it is too negative, thus keeping it as small as possible.

It has been shown that in the absence of an explicit cost of control, it costs nothing to push harder (up to the allowed limits), and pushing hard in the right direction is better than pushing only a little; thus, the physically obvious bang-bang law, $u(x, t) = -\operatorname{sgn} x$, is indeed optimal in that case (Beneš [1974], [1975], Ikeda and Watanabe [1977]).

In the case of an "expensive" or "costly" controller, however, the optimization problem must be solved essentially by balancing the costs of control against those of performance. Exerting more control than the optimal will improve the performance (i.e., make x_T smaller) but not so much as to overwhelm the cost of the additional control effort. On the other hand, exerting less control than the optimal will result in a deterioration of the performance which will counterbalance what is saved in effort. Therefore, in the problem formulated above, where there is an explicit running cost of control $|u|$, a dead zone in two dimensions suggests itself, in which one should do nothing. More specifically, one expects the optimal law to be of the form

$$\begin{aligned} u(x, T-\tau) &= -1, \qquad x > s(\tau), \\ &= 0, \qquad |x| \leqq s(\tau), \\ &= 1, \qquad x < -s(\tau), \end{aligned}$$

for some positive and "reasonable" function $\{s(\tau);\ 0 \leqq \tau \leqq T\}$ of the "time-to-go" τ.

Indeed, under some symmetry, smoothness and convexity assumptions on the terminal cost function $f(x)$ and making use of the general theory of quasilinear parabolic equations, the Girsanov theorem, the Feynman-Kac formula and the maximum principle for parabolic operators, it is possible to establish existence, uniqueness, continuity and monotonicity for such a boundary function $\{s(\tau);\ 0 \leqq \tau \leqq T\}$, for any positive time horizon T. In order to prove smoothness of $s(\tau)$, the so called "free-boundary problem" is transformed into the equivalent one of studying a system of two nonlinear integral equations for the free boundary and the value function along it; actually, a whole family of such pairs of equations indexed by a gain parameter α, $0 \leqq \alpha \leqq 1$, $|u| \leqq \alpha$ is considered. The latter is viewed as a family of continuous transformations from the Banach space of pairs of continuous functions into another Banach space; a generalized Leray-Schauder degree can be defined for these transformations,

which form a homotopy indexed by the gain parameter α. Using the degree of knowledge provided by the especially simple form of the transformation at the endpoint $\alpha = 0$ (uncontrolled case) and the invariance of topological degree under homotopy, it is proved that the free boundary function $s(\tau)$ for the original problem $\alpha = 1$ (fully controlled case) is continuously differentiable on $[0, T]$ for any $T > 0$.

2. Formulation.

2.1. The optimal control problem. Let $f(x)$ be a function satisfying assumptions A.1–A.3 below:

A.1. $f(x)$ is an even, $C^3(\mathbf{R})$, nonnegative function. Both $f(x)$ and $f'(x)$ increase monotonically to infinity on $\mathbf{R}^+$ as $x \to \infty$.

A.2. $f(x)$ is uniformly convex; i.e., there exists a positive constant k such that $f''(x) \geqq k > 0$, for all $x \in \mathbf{R}$.

A.3. $-L \leqq f''(x) \leqq 0$ on $\mathbf{R}^+$, for some $L \geqq 0$; i.e., the second derivative of the function $f(x)$ is decreasing with distance from the origin.

Let the family $\mathbf{A}$ of admissible feedback controls consist of all jointly measurable functions $u: \mathbf{R} \times [0, T] \to [-1, 1]$. The stochastic control problem is to choose a control law $u \in \mathbf{A}$ so as to minimize the expected total cost

$$J(x, T-\tau; u) = E\left[f(x_T) + \int_{T-\tau}^{T} |u(x_t, t)|\, dt\right] \tag{2.1}$$

of starting at place x, time $T - \tau$, subject to

$$dx_t = u(x_t, t)\, dt + dw_t, \qquad T - \tau \leqq t \leqq T, \tag{2.2}$$

$$x_{T-\tau} = x, \tag{2.3}$$

where $\{w_t;\ T - \tau \leqq t \leqq T\}$ is a Wiener process on an underlying probability space $(\Omega, \mathbf{F}, P)$. E denotes expectation with respect to the probability measure P.

The first question that arises is the following: in what sense is the stochastic differential equation (2.2) to be understood? Because we cannot expect the optimal u to be Lip or even continuous in x (in fact, as pointed out in the Introduction, the natural candidate for the optimal law is discontinuous at the moving cutoff points $\pm s(\tau)$) and so we cannot just resort to the classical Ito theory. The answer to this question is contained in a very important paper by Zvonkin [1974], where it is shown that a stochastic differential equation like (2.2) in one space dimension with $u(x, t)$ bounded and measurable does possess a pathwise unique strong nonanticipative solution x_t.

An alternative approach consists in constructing a solution measure to (2.2) by means of the Girsanov theorem, rather than attempting to construct the paths of a solution process. More specifically, for any $u \in \mathbf{A}$, $x \in \mathbf{R}$ and $\tau \in [0, T]$ one considers the process

$$x_t = x + w_{t-(T-\tau)}, \qquad T - \tau \leqq t \leqq T,$$

under the new measure

$$\tilde{P}(d\omega) = \exp\left[\int_{T-\tau}^{T} u(x + w_{t-T+\tau}, t)\, dw_t - \frac{1}{2}\int_{T-\tau}^{T} u^2(x + w_{t-T+\tau}, t)\, dt\right] \cdot P(d\omega).$$

Because u is bounded and measurable, $\tilde{P}$ is a probability measure and Girsanov's

theorem [1960] asserts that the process

$$\begin{aligned}\tilde{w}_t &= w_{t-T+\tau} - \int_{T-\tau}^{t} u(x + w_{\lambda-T+\tau}, \lambda)\, d\lambda \\ &= x_t - x - \int_{T-\tau}^{t} u(x_\lambda, \lambda)\, d\lambda, \qquad T-\tau \leqq t \leqq T\end{aligned} \tag{2.4}$$

is a Wiener process on the probability space $(\Omega, \mathbf{F}, \tilde{P})$. The process x_t thus constructed is a weak solution of the stochastic differential equation

$$dx_t = u(x_t, t)\, dt + d\tilde{w}_t,$$

as indicated by (2.4). Now the control problem can be viewed as an extremal problem on the choice of the best measure.

In either case, the optimization problem can be treated through the Bellman equation of dynamic programming for the value function

$$V(x, \tau) = \inf_{u \in \mathbf{A}} J(x, T-\tau; u), \tag{2.5}$$

which for the problem under consideration takes the form

$$V_\tau = \frac{1}{2} V_{xx} + \min_{|u| \leqq 1} [uV_x + |u|] = \frac{1}{2} V_{xx} + a(V_x), \qquad (x, \tau) \in \mathbf{R} \times (0, T], \tag{2.6}$$

$$V(x, 0) = f(x), \qquad x \in \mathbf{R}. \tag{2.7}$$

Equation (2.6) is a quasilinear partial differential equation of parabolic type, where the nonlinearity $a(p)$ is given by

$$\begin{aligned} a(p) &= 1 + p, && p < -1, \\ &= 0, && |p| \leqq 1, \\ &= 1 - p, && p > 1. \end{aligned} \tag{2.8}$$

A verification theorem (Fleming and Rishel [1975, p. 159]) asserts that if $V(x, \tau)$ is a solution of the Cauchy problem (2.6)–(2.7) in the space $C^{2,1}(\mathbf{R} \times [0, T])$ satisfying a polynomial growth condition in the space variable x, then $V(x, \tau) \leqq J(x, T-\tau; u)$ for any $u \in \mathbf{A}$ and any initial condition $(x, \tau) \in \mathbf{R} \times [0, T]$. On the other hand, if u^* is a control law in $\mathbf{A}$ such that

$$\begin{aligned} u^*(x, T-\tau) &= -1, && V_x(x, \tau) > 1, \\ &= 0, && |V_x(x, \tau)| \leqq 1, \\ &= 1, && V_x(x, \tau) < -1, \end{aligned} \tag{2.9}$$

for almost all $(x, \tau) \in \mathbf{R} \times [0, T]$, then u^* is optimal: $V(x, \tau) = J(x, T-\tau; u^*)$, and $(x, \tau) \in \mathbf{R} \times [0, T]$. Existence of a $C^{2,1}$ solution to the Cauchy problem (2.6)–(2.7) and of an optimal control law u^* for the problem under consideration is guaranteed by the so-called existence theorems in Fleming and Rishel [1975, pp. 166–170]. Crucial to the applicability of the above theorems are the assumptions of compactness of the action space $[-1, 1]$, of Lip continuity of the running cost $|u|$ in the control variable as well as the smoothness and growth conditions on the terminal cost function $f(x)$ (assumptions A.1–A.3).

Note. The optimal feedback control law (2.9) obtained by the verification theorem is actually optimal in the larger class **U** of nonanticipative laws with values in $[-1, 1]$. For a discussion of this point, see Davis and Varaiya [1973].

Once the existence and the functional form of the optimal law have been settled, the interesting problem is to examine whether there exists a "free" or "moving" boundary $x = s(\tau)$ in two dimensions such that, if one considers the regions

$$\begin{aligned} D(-1) &= \{(x, \tau);\ x > s(\tau),\ 0 < \tau < T\}, \\ D(0) &= \{(x, \tau);\ |x| < s(\tau),\ 0 < \tau < T\}, \\ D(1) &= \{(x, \tau);\ x < -s(\tau),\ 0 < \tau < T\}, \end{aligned} \tag{2.10}$$

then the gradient $V_x(x, t)$ of the value function $V(x, \tau)$ is equal to ± 1 on $\pm s(\tau)$ and

$$\begin{aligned} V_x(x, \tau) &> 1 \quad \text{in } D(-1), \\ V_x(x, \tau) &< -1 \quad \text{in } D(1), \\ |V_x(x, \tau)| &< 1 \quad \text{in } D(0). \end{aligned}$$

If such a boundary function $\{s(\tau);\ 0 \leqq \tau \leqq T\}$ exists, the optimal law $u^*(x, T-\tau)$ can be written in the more suggestive form

$$\begin{aligned} u^*(x, T-\tau) &= -1 \quad \text{in } D(-1), \\ &= 0 \quad \text{in } \bar{D}(0), \\ &= 1 \quad \text{in } D(1). \end{aligned} \tag{2.11}$$

In § 4 we prove the following theorem:

THEOREM 2.1. *Under assumptions* A.1–A.3 *on the terminal cost* $f(x)$, *there exists a unique solution* $V(x, \tau)$ *in* $C^{2,1}$ $(\mathbf{R} \times [0, T])$ *to the Cauchy problem of solving the Bellman equation* (2.6) *subject to the initial condition* (2.7). *The gradient* $V_x(x, \tau)$ *of this solution is, for fixed* $0 \leqq \tau \leqq T$, *an odd, strictly increasing (to infinity) function of* x, *achieving the value* $1(-1)$ *at a certain unique, finite point* $s(\tau)$ $[-s(\tau)]$. *The function* $\{s(\tau);\ 0 \leqq \tau \leqq T\}$ *is Lipschitz continuous and increasing on* $[0, T]$, *any* $T > 0$.

COROLLARY. *The optimal control law* $u^*: \mathbf{R} \times [0, T] \to [-1, 1]$ *for which the infimum of* $J(x, T-\tau; u)$ *over* **A** *is achieved is given by* (2.11), *with* $\{s(\tau);\ 0 \leqq \tau \leqq T\}$ *as in Theorem* 2.1.

2.2. The free-boundary problem. We introduce the parameter α, $0 \leqq \alpha \leqq 1$, to allow for a variation in the gain of the controller in the optimal control problem: $|u| \leqq \alpha$. We thus consider a whole family of optimization problems indexed by α, with corresponding value functions $V^{(\alpha)}(x, \tau)$, $0 \leqq \alpha \leqq 1$, satisfying the Cauchy problems for the Bellman equation

$$V_\tau^{(\alpha)} = \frac{1}{2} V_{xx}^{(\alpha)} + \min_{|u| \leqq \alpha} [u V_x^{(\alpha)} + |u|] = \frac{1}{2} V_{xx}^{(\alpha)} + \alpha a(V_x^{(\alpha)}), \qquad (x, \tau) \in \mathbf{R} \times [0, T], \tag{2.12}$$

$$V^{(\alpha)}(x, 0) = f(x), \qquad x \in \mathbf{R}, \tag{2.13}$$

where $a(p)$ is the function defined in (2.8). In this and later sections, we drop explicit dependence of the value function $V(x, \tau)$ and the free boundary $s(\tau)$ on the gain parameter α, whenever $\alpha = 1$ (original problem, fully controlled case).

The unique solution of the equation $f'(x) = 1$ is denoted by b; i.e.,

$$f'(b) = 1. \tag{2.14}$$

The "free-boundary problem" is formulated as follows: Find functions $s^{\alpha}(\tau)$ on $[0, T]$, $V^{(\alpha)}(x, \tau)$ on $\mathbf{R}^{+}\times[0, T]$ such that:

$$V_{\tau}^{(\alpha)} = \frac{1}{2} V_{xx}^{(\alpha)} \quad \text{in } D^{*}(0) = \{(x, \tau);\ 0 < x < s^{\alpha}(\tau),\ 0 < \tau < T\}, \tag{2.15}$$

$$V_{\tau}^{(\alpha)} = \frac{1}{2} V_{xx}^{(\alpha)} + \alpha - \alpha V_{x}^{(\alpha)} \quad \text{in } D(-\alpha) = \{(x, \tau);\ x > s^{\alpha}(\tau),\ 0 < \tau < T\}, \tag{2.16}$$

$$V_{x}^{(\alpha)}(0, \tau) = 0, \qquad 0 < \tau < T, \tag{2.17}$$

$$V^{(\alpha)}(x, 0) = f(x), \qquad x \in \mathbf{R}, \tag{2.18}$$

$$s^{\alpha}(0) = b, \quad \text{where } f'(b) = 1, \tag{2.19}$$

$$V_{x}^{(\alpha)}[s(\tau), \tau] = 1, \qquad 0 < \tau < T. \tag{2.20}$$

The curve $\{(x, \tau);\ x = s^{\alpha}(\tau),\ 0 \leqq \tau \leqq T\}$ is the unknown "free" or "moving" boundary, which is to be determined together with $V^{(\alpha)}(x, \tau)$.

DEFINITION 2.1. We say that $s^{\alpha}(\tau)$, $V^{(\alpha)}(x, \tau)$ form *a solution to the free-boundary problem* (2.15)–(2.20) *on* $\mathbf{R}^{+}\times[0, T]$, if

(i) $s^{\alpha}(\tau)$, $0 \leqq \tau \leqq T$, is Lipschitz continuous on $[0, T]$,

(ii) $V^{(\alpha)}(x, \tau)$ is a $C^{2,1}(\mathbf{R}^{+}\times[0, T])$ function,

(iii) the equations and initial and boundary conditions (2.15)–(2.20) are satisfied.

PROPOSITION 2.1. *The Cauchy problem of solving the Bellman equation* (2.6) *subject to initial condition* (2.7) *in* $C^{2,1}(\mathbf{R}\times[0, T])$ *is equivalent to the free-boundary problem* (2.15)–(2.20), *for* $\alpha = 1$.

Proof. If $V(x, \tau)$ is the $C^{2,1}(\mathbf{R}\times[0, T])$ solution to the Cauchy problem (2.6), (2.7) it clearly satisfies all requirements of Definition 2.1, by Theorem 2.1. On the other hand, if $s(\tau)$, $V(x, \tau)$ is a solution to the free-boundary problem (2.15)–(2.20) in the sense of Definition 2.1, $V(x, \tau)$ can be evenly extended to the whole of $\mathbf{R}\times[0, T]$; the resulting function is $C^{2,1}$ on $\mathbf{R}\times[0, T]$ and an application of the maximum principle for parabolic operators asserts that $V_x(x, \tau) > 1$ in $D(-1)$, $|V_x(x, \tau)| < 1$ in $D(0)$ and $V_x(x, \tau) < -1$ in $D(1)$. So $V(x, \tau)$ satisfies (2.6) and thus solves the Cauchy problem.

In §§ 6, 7 we prove the following theorem:

THEOREM 2.2. *Under assumptions* A.1–A.3 *on the terminal cost function* $f(x)$, *the free-boundary problem* (2.15)–(2.20) *possesses a unique solution on* $\mathbf{R}^{+}\times[0, T]$ *in the sense of Definition* 2.1, *with* $s(\tau)$ *continuously differentiable on* $[0, T]$, *any* $T > 0$.

3. Summary. In § 2, the stochastic control problem was formulated and two methods of approaching it were considered. The first method consists in obtaining as much information about the function $s^{\alpha}(\tau)$ satisfying $V_{x}^{(\alpha)}[s^{\alpha}(\tau), \tau] = 1$ as possible through a direct investigation of the properties of the value function $V^{(\alpha)}(x, \tau)$; the latter is viewed as the solution to the Cauchy problem (2.12)–(2.13) for the Bellman equation of dynamic programming. The second method regards the problem of determining the boundary curve $x = s^{\alpha}(\tau)$ as a free-boundary problem in the sense of Definition 2.1. Actually, these two lines of approach are equivalent, as was pointed out in Proposition 2.1.

The first method is undertaken in § 4, where the assertions of Theorem 2.1 are proved. Using the basic theory of quasilinear partial differential equations of parabolic type and the maximum principle for parabolic operators, we localize the boundary curve $x = s^{\alpha}(\tau)$ in the (x, τ)-plane and prove some of its properties, such as right

continuity and monotonicity (Proposition 4.2). The most important result in this section is the stochastic representation (4.44) for $V_{xx}^{(\alpha)}(x, \tau)$ (Proposition 4.5); it is established via the Feynman-Kac formula and the Girsanov theorem and relates the second space derivative of the value function to the local time spent by the solution process of (2.2) at the two branches of the free boundary $x = \pm s^{\alpha}(\tau)$. The main corollary of the representation (4.44) is the positive lower bound (4.55) on the second derivative of the value function, independent of $0 \leqq \alpha \leqq 1$, a fact of paramount importance throughout the whole paper. First, it helps in proving Lip continuity of $s^{\alpha}(\tau)$ on $[0, T]$, for any $T > 0$, and thus establishes an "a priori" bound on the growth $s'^{\alpha}(\tau)$ of the free-boundary function, independent of $0 \leqq \alpha \leqq 1$. Second, it provides the crucial step in the proof of the smoothness of $s(\tau)$, globally in time (§ 7).

We embark on the free-boundary problem approach in § 5, our main concern now being to establish continuous differentiability for the function $s(\tau)$, $0 \leqq \tau \leqq T$. Following a standard method in problems of this sort, such as the Stefan problem (e.g., Friedman [1959], Rubinštein [1967]), we transform the free-boundary problem into the equivalent one of studying the pair of integral equations (5.10), (5.17) for $[s'^{\alpha}(\tau), w'^{\alpha}(\tau)]$, $w^{\alpha}(\tau) = V^{(\alpha)}[s^{\alpha}(\tau), \tau]$ (Proposition 5.1). However, the resulting integral equations are far more difficult to deal with analytically than the corresponding ones for the Stefan problem, in the sense that they are not amenable in any natural way to a straightforward fixed point analysis; the reason for this difficulty has to be traced back to the highly implicit boundary condition $V_x[s(\tau), \tau] = 1$ (the corresponding condition for the Stefan problem is $s'(\tau) = -V_x[s(\tau), \tau]$ and the integral equations are solvable by the contraction mapping principle). The integral equations are made use of in the study of the special case $f(x) = x^2$, particularly in disproving the alleged solution $s(\tau) = \frac{1}{2} + \tau$ suggested by R. C. Davis [1968] in an unsuccessful attempt to attack this important special case; they are also used in the proof of strict monotonicity of the free boundary. Fig. 2 of § 5.3 depicts the free boundary $s(\tau)$ in the special case $f(x) = x^2$. The plot was obtained by numerically solving the quasilinear partial differential equation of dynamic programming and identifying the points where $V_x(x, \tau) = 1$.

Our method of establishing existence of a continuous solution (s', w') to the integral equations (5.10), (5.17) uses homotopy and topological degree and is carried out in §§ 6, 7. More specifically, we regard the solutions $(s'^{\alpha}, w'^{\alpha}) = (c^{\alpha}, \nu^{\alpha})$ of the integral equations, for all possible values $0 \leqq \alpha \leqq 1$, as zero-points of the transformations $\phi^{\alpha}(c, \nu)$ defined in (6.5)–(6.12) on a Banach space $\mathbf{X}_{\sigma}$ to another Banach space $\mathbf{Y}_{\sigma}$. The gain parameter α, $0 \leqq \alpha \leqq 1$, acts as a homotopy on the family of continuous transformations $\{\phi^{\alpha}(c, \nu); 0 \leqq \alpha \leqq 1\}$ from $\mathbf{X}_{\sigma}$ to $\mathbf{Y}_{\sigma}$ (Corollary 6.1). On the other hand, the operators ϕ^{α} admit, for each $0 \leqq \alpha \leqq 1$, a decomposition of the form "compact plus homeomorphism," for which a topological degree can be defined (Propositions 6.1, 6.3, 6.4). The uncontrolled case $\alpha = 0$ is, however, penetrable, in the sense that the whole operator $\phi^{0}(c, \nu)$ is a homeomorphism. Because of the "a priori" bound on $\|(s'^{\alpha}, w'^{\alpha})\|$ established in Proposition 4.7, the topological degree of the zero point in $\mathbf{Y}_{\sigma}$ with respect to the mapping ϕ^{α} and the set $G \subseteq \mathbf{X}_{\sigma}$ defined in (6.13) is invariant under the homotopy, i.e., independent of α; so $\deg[\phi^{1}, G, 0] = \deg[\phi^{0}, G, 0] = \pm 1$, from the fact that ϕ^{0} is a homeomorphism. Now the existence of a pair $(c, \nu) \in G$ satisfying $\phi^{1}(c, \nu) = 0$ at the endpoint $\alpha = 1$ is established by appealing to a basic "existence result" in degree theory. The above heuristics are substantiated in the proof of Theorem 7.1, in which we establish continuous differentiability of the free boundary for "small times" $0 \leqq \tau \leqq \sigma$, σ a sufficiently small constant. The method is then applied step by step and the solution is extended into the future, up to any finite time horizon $T > 0$ (proof of Theorem 2.2). As has already been pointed out, the

feasibility of this extension is a consequence of the positive lower bound (4.55) on the curvature of the value function.

Most of the details are carried out separately in appendices; see §§ 8, 9 and 10.

4. Preliminary results and a priori bounds. We recall the anisotropic Hölder spaces that enter into the a priori estimates of Schauder type for parabolic equations (Ladyženskaja, Solonnikov and Ural'ceva [1968, pp. 7–8]). Consider an arbitrary open, bounded and connected set Q in $\mathbf{R}$ and denote by Q_T the rectangle $Q\times(0, T)$. For any positive integer l, any $0<\beta\leqq 1$, we consider the Banach space $H_{l+\beta,(l+\beta)/2}(\bar{Q}_T)$ of functions $u(x,\tau)$ that are continuous in $\bar{Q}_T$, together with all derivatives of the form $D_t^r D_x^s$ for $2r+s\leqq l$ and have a finite norm

$$|u|_{Q_T}^{(l+\beta)} \equiv \langle u\rangle_{Q_T}^{(l+\beta)} + \sum_{j=0}^{l} \langle u\rangle_{Q_T}^{(j)}, \tag{4.1}$$

where

$$\langle u\rangle_{Q_T}^{(0)} = |u|_{Q_T}^{(0)} = \max_{Q_T} |u|, \qquad \langle u\rangle_{Q_T}^{(j)} = \sum_{(2r+s=j)} |D_t^r D_x^s u|_{Q_T}^{(0)},$$

$$\langle u\rangle_{Q_T}^{(l+\beta)} = \sum_{(2r+s=l)} \langle D_t^r D_x^s u\rangle_{x,Q_T}^{(\beta)} + \sum_{0<l+\beta-2r-s<2} \langle D_t^r D_x^s u\rangle_{t,Q_T}^{(l+\beta-2r-s)/2},$$

$$\langle u\rangle_{x,Q_T}^{(\nu)} = \sup_{(x,\tau),(x',\tau)\in Q_T} \frac{|u(x,\tau)-u(x',\tau)|}{|x-x'|^{\nu}}, \qquad 0<\nu\leqq 1,$$

$$\langle u\rangle_{t,Q_T}^{(\nu)} = \sup_{(x,\tau),(x,\tau')\in Q_T} \frac{|u(x,\tau)-u(x,\tau')|}{|\tau-\tau'|^{\nu}}, \qquad 0<\nu\leqq 1.$$

We are interested in studying the Cauchy problem

$$V_\tau^{(\alpha)} = \frac{1}{2} V_{xx}^{(\alpha)} + \alpha a(V_x^{(\alpha)}), \qquad (x,\tau)\in \mathbf{R}\times(0, T], \tag{2.12}$$

$$V^{(\alpha)}(x, 0) = f(x), \qquad x\in\mathbf{R}, \tag{2.13}$$

where $a(p)$ is the function defined in (2.8). According to Ladyženskaja et al. [1968, Theorem 8.1, p. 495], there exists a unique solution $V^{(\alpha)}(x,\tau)$ to the Cauchy problem above in the strip $\mathbf{R}_T = \mathbf{R}\times[0, T]$, for any $0\leqq\alpha\leqq 1$. $V^{(\alpha)}(x,\tau)$ satisfies (2.12) in the classical sense and furthermore belongs to the space $H_{2+\beta,1+\beta/2}(\bar{Q}_T)$ for any bounded rectangle $Q_T\subseteq\mathbf{R}_T$, with Hölder constants in (4.1) independent of α, $0\leqq\alpha\leqq 1$. Here, β is the modulus of continuity of the nonlinear function $a(p)$; in our case, $a(p)$ is Lipschitz continuous with $\beta = 1$.

We now consider the smooth (three times continuously differentiable) approximations $a_n(p)$, $n\in\mathbf{N}$ to the function $a(p)$, given by

$$a_n(p) = c_n(p-1) + c_n(-p+1), \tag{4.2}$$

where $c_n(p)$ is the sequence of functions defined below, along with their derivatives:

$$\begin{aligned} c_n(p) &= 0, && p<-\frac{1}{n},\\ &= -p, && p>\frac{1}{n},\\ &= \int_{-1/n}^{p} c_n'(u)\,du, && |p|\leqq\frac{1}{n}, \end{aligned} \tag{4.3}$$

$$
\begin{aligned}
c_n'(p) &= -\frac{1}{2} - \frac{2}{3} n^3 p^3, \qquad 0 \leqq p \leqq \frac{1}{2n} \\
&= -\frac{7}{12} - \frac{n}{2}\left(p - \frac{1}{2n}\right) - n^2\left(p - \frac{1}{2n}\right)^2 + \frac{2}{3}\left(p - \frac{1}{2n}\right)^3, \qquad \frac{1}{2n} < p \leqq \frac{1}{n}, \\
&= 1, \qquad p > \frac{1}{n}, \\
&= -1 - c_n'(-p), \qquad p < 0.
\end{aligned} \tag{4.4}
$$

Besides

$$
\begin{aligned}
c_n''(p) &= -n + 2n^3 p^2, \qquad 0 \leqq p \leqq \frac{1}{2n}, \\
&= -2n^3\left(p - \frac{1}{n}\right)^2, \qquad \frac{1}{2n} < p \leqq \frac{1}{n}, \\
&= 0, \qquad p > \frac{1}{n}, \\
&= c_n''(-p), \qquad p < 0,
\end{aligned}
$$

and

$$
\begin{aligned}
c_n'''(p) &= 4n^3 p, \qquad 0 \leqq p \leqq \frac{1}{2n}, \\
&= 4n^3\left(\frac{1}{n} - p\right), \qquad \frac{1}{2n} < p \leqq \frac{1}{n}, \\
&= 0, \quad p > \frac{1}{n}, \\
&= -c_n'''(-p), \qquad p < 0.
\end{aligned}
$$

The corresponding Cauchy problems

$$
V_\tau^{(n,\alpha)} = \frac{1}{2} V_{xx}^{(n,\alpha)} + \alpha a_n(V_x^{(n,\alpha)}), \qquad (x, \tau) \in \mathbf{R} \times (0, T], \tag{4.5}
$$

$$
V^{(n,\alpha)}(x, 0) = f(x), \qquad x \in \mathbf{R} \tag{4.6}
$$

have, for any $0 \leqq \alpha \leqq 1$, a unique $C^{2,1}$ solution on $\mathbf{R}_T$ which belongs, for each $n \in \mathbf{N}$, to the Banach space $H_{5+\beta,(5+\beta)/2}(\bar{Q}_T)$, for any bounded cylinder $Q_T \subseteq \mathbf{R}_T$ (see, for instance, Ladyženskaja et al. [1968, p. 456]). The additional smoothness of the solutions is a result of the greater smoothness of the coefficients in (4.5) compared to $a(p)$ in (2.12). We consider now the Cauchy problems for $V_x^{(n,\alpha)}$ and $V_{xx}^{(n,\alpha)}$,

$$
(V_x^{(n,\alpha)})_\tau = \frac{1}{2}(V_x^{(n,\alpha)})_{xx} + \alpha a_n'(V_x^{(n,\alpha)})(V_x^{(n,\alpha)})_x, \qquad (x, \tau) \in \mathbf{R} \times (0, T], \tag{4.7}
$$

$$
V_x^{(n,\alpha)}(x, 0) = f'(x), \qquad x \in \mathbf{R}, \tag{4.8}
$$

and

$$(V_{xx}^{(n,\alpha)})_\tau = \frac{1}{2}(V_{xx}^{(n,\alpha)})_{xx} + \alpha a_n'(V_x^{(n,\alpha)})(V_{xx}^{(n,\alpha)})_x + \alpha a_n''(V_x^{(n,\alpha)})V_{xx}^{(n,\alpha)}(V_{xx}^{(n,\alpha)}), \qquad (x,\tau)\in \mathbf{R}\times(0,T], \tag{4.9}$$

$$V_{xx}^{(n,\alpha)}(x,0) = f''(x), \qquad x\in\mathbf{R}, \tag{4.10}$$

respectively. Equations (4.7) and (4.9) hold in the classical sense and are derived from (4.5), (4.7), respectively, by differentiation. By the continuous dependence of the solutions of parabolic partial differential equations on their coefficients (stability theorems) we have that $V^{(n,\alpha)}(x,\tau)$, $V_x^{(n,\alpha)}(x,\tau)$, $V_{xx}^{(n,\alpha)}(x,\tau)$ converge as $n\to\infty$ to $V^{(\alpha)}(x,\tau)$, $V_x^{(\alpha)}(x,\tau)$, $V_{xx}^{(\alpha)}(x,\tau)$, respectively, uniformly on compact (x,τ) sets.

It is easy to check that $V^{(\alpha)}(x,\tau)$ satisfies a polynomial growth condition in x. Indeed, assumption A.3 implies in particular that $f''(x)$ is decreasing in $x\geqq 0$ and consequently,

$$0<k\leqq f''(x)\leqq K = f''(0), \quad \text{all } x\in\mathbf{R}. \tag{4.11}$$

If we use the "naive" control law $u(x,t)\equiv 0$, we immediately get from (2.1), (2.5)

$$V^{(\alpha)}(x,\tau)\leqq E(K|x+w_\tau|^2)\leqq K(x^2+\tau), \qquad (x,\tau)\in\mathbf{R}\times[0,T]. \tag{4.12}$$

Similarly, the approximating functions $V^{(n,\alpha)}(x,\tau)$ as well as their derivatives $V_x^{(n,\alpha)}(x,\tau)$, $V_{xx}^{(n,\alpha)}(x,\tau)$ satisfy polynomial growth conditions.

Consider the parabolic operator

$$\mathbf{L} = \frac{1}{2}\frac{\partial^2}{\partial x^2} + \alpha a_n'(V_x^{(n,\alpha)})\frac{\partial}{\partial x} + \alpha a_n''(V_x^{(n,\alpha)})V_{xx}^{(n,\alpha)}\cdot - \frac{\partial}{\partial\tau}. \tag{4.13}$$

From (4.9) one gets $\mathbf{L}(V_{xx}^{(n,\alpha)})=0$, and since $V_{xx}^{(n,\alpha)}(x,0)=f''(x)\geqq k>0$, $x\in\mathbf{R}$, the maximum principle for parabolic operators (Friedman [1964, p. 43]) yields

$$V_{xx}^{(n,\alpha)}(x,\tau)\geqq 0, \qquad (x,\tau)\in\mathbf{R}\times[0,T].$$

Similarly, $\mathbf{L}(V_{xx}^{(n,\alpha)}-K) = -\alpha a_n''(V_{xx}^{(n,\alpha)})V_{xx}^{(n,\alpha)}K\geqq 0$ in $\mathbf{R}\times(0,T]$,

$$V_{xx}^{(n,\alpha)}(x,0)-K = f''(x)-K\leqq 0 \quad \text{in } \mathbf{R}.$$

A second application of the maximum principle now gives

$$0\leqq V_{xx}^{(n,\alpha)}(x,\tau)\leqq K \quad \text{in } \mathbf{R}\times[0,T], \tag{4.14}$$

while a passage to the limit as $n\to\infty$ in (4.14) asserts that

$$0\leqq V_{xx}^{(\alpha)}(x,\tau)\leqq K \quad \text{in } \mathbf{R}\times[0,T], \tag{4.15}$$

for any $\alpha\in[0,1]$. As a consequence of (4.15), the gradient $V_x^{(\alpha)}(\cdot,\tau)$ is, for any $\tau\in[0,T]$, an increasing function of x.

We are interested in determining, for each $\tau\in[0,T]$, the point(s) $s^\alpha(\tau)$ for which $V_x^{(\alpha)}[s^\alpha(\tau),\tau]=1$. It behooves us, therefore, to examine more closely the gradient function $V_x^{(\alpha)}(x,\tau)$.

PROPOSITION 4.1. Monotonicity and Lip continuity of the gradient $V_x^{(\alpha)}(x,\tau)$ in τ. *For any* $0\leqq\alpha\leqq 1$, $x>0$, *the gradient* $V_x^{(\alpha)}(x,\tau)$ *of the value function*

(i) *is a decreasing function of* τ *on* $[0,T]$.

(ii) *is Lipschitz continuous in* τ *on* $[0,T]$ *with Lipschitz constant* $\alpha K + L/2$.

Proof. We introduce the parabolic operators, indexed by the integer n,

$$\mathbf{N}=\frac{1}{2}\frac{\partial^2}{\partial x^2}+\alpha a_n'(V_x^{(n,\alpha)})\frac{\partial}{\partial x}-\frac{\partial}{\partial \tau}, \tag{4.16}$$

and propose to show that

$$-\left(\alpha K+\frac{L}{2}\right)(\tau_2-\tau_1)\leqq V_x^{(\alpha)}(x,\tau_2)-V_x^{(\alpha)}(x,\tau_1)\leqq 0, \tag{4.17}$$

for any $x>0$, $0\leqq\tau_1<\tau_2\leqq T$, $0\leqq\alpha\leqq 1$,

which incorporates the two assertions of the theorem. In particular, (4.17) implies

$$f'(x)-\left(\alpha K+\frac{L}{2}\right)\tau\leqq V_x^{(\alpha)}(x,\tau)\leqq f'(x), \tag{4.18}$$

for any $x>0$, $0\leqq\tau\leqq T$, $0\leqq\alpha\leqq 1$.

First, it is observed that, because of (4.7) and assumption A.3 on f,

$$\mathbf{N}[f'(x)-V_x^{(n,\alpha)}(x,\tau)]=\mathbf{N}[f'(x)]=\frac{1}{2}f'''(x)+\alpha a_n'[V_x^{(n,\alpha)}(x,\tau)]f''(x)\leqq 0,$$

$$(x,\tau)\in\mathbf{R}^+\times(0,T].$$

Consider any rectangle $\bar{Q}_T=\{(x,\tau);\ 0\leqq x\leqq q,\ 0\leqq\tau\leqq T\}$ in $\mathbf{R}^+\times[0,T]$. The function $u(x,\tau)=f'(x)-V_x^{(n,\alpha)}(x,\tau)$ attains its maximum on $\bar{Q}_T$ at some point $P^0=(x^0,\tau^0)$. Denote by $S(P^0)$ the set of points Z in $\bar{Q}_T$ which can be connected to P^0 by a simple, continuous curve in $\bar{Q}_T$ along which the τ-coordinate is nondecreasing from Z to P^0; obviously $P=(x^0,0)\in S(P^0)$. Suppose that $u(P^0)<0$; then by $\mathbf{N}[u(x,\tau)]\leqq 0$ and by the strong maximum principle for parabolic operators (Friedman [1964, Theorem 2.1, p. 34]) $u(P)=u(P^0)<0$, a contradiction, because $u(P)=u(x^0,0)=f'(x^0)-V_x^{(n,\alpha)}(x^0,0)=0$. So $u(P^0)\geqq 0$, which implies a fortiori that $u(x,\tau)=f'(x)-V_x^{(n,\alpha)}(x,\tau)\geqq 0$ in $\bar{Q}_T$, for any such $\bar{Q}_T$. A passage to the limit as $n\to\infty$ in the above inequality yields the right-hand side of (4.18), while letting $\tau\to 0$, we obtain

$$V_{x\tau}^{(n,\alpha)}(x,0)\leqq 0 \quad \text{in } \mathbf{R}^+. \tag{4.19}$$

It is also checked that

$$\mathbf{N}[f'(x)-\left(\alpha K+\frac{L}{2}\right)\tau-V_x^{(n,\alpha)}(x,\tau)]=\frac{1}{2}f'''(x)+\alpha a_n'[V_x^{(n,\alpha)}(x,\tau)]f''(x)+\left(\alpha K+\frac{L}{2}\right)$$

$$\geqq 0, \qquad \mathbf{R}^+\times(0,T],$$

where again (4.7), (4.2), (4.4) and assumption A.3 on f have been used. Reasoning as before, we get by the strong maximum principle: $f'(x)-(\alpha K\times L/2)\tau-V_x^{(n,\alpha)}(x,\tau)\leqq 0$ in $\mathbf{R}^+\times[0,T]$. The left-hand side of (4.18) follows readily if we let $n\to\infty$ in the inequality above; if we now divide both sides of the latter by τ and then let $\tau\to 0$ we obtain

$$-\left(\alpha K+\frac{L}{2}\right)\leqq V_{x\tau}^{(n,\alpha)}(x,0) \quad \text{in } \mathbf{R}^+. \tag{4.20}$$

We now differentiate both sides of (4.7) with respect to τ and get

$$\mathbf{L}(V_{x\tau}^{(n,\alpha)})=0 \quad \text{in } \mathbf{R}\times(0,T], \tag{4.21}$$

where **L** is the parabolic operator introduced in (4.13); all partial derivatives involved in (4.21) are understood in the classical sense. A strong maximum principle argument can again be used to show that (4.19), (4.21) imply

$$V_{x\tau}^{(n,\alpha)}(x,\tau) \leqq 0 \quad \text{in } \mathbf{R}^+ \times [0,T]. \tag{4.22}$$

Similarly, we check that

$$\mathbf{L}\left[V_{x\tau}^{(n,\alpha)}(x,\tau) + \left(\alpha K + \frac{L}{2}\right)\right] = \alpha a_n''[V_x^{(n,\alpha)}(x,\tau)]\left(\alpha K + \frac{L}{2}\right)$$

$$\leqq 0 \quad \text{in } \mathbf{R}^+ \times (0,T],$$

$$V_{x\tau}^{(n,\alpha)}(x,0) + \left(\alpha K + \frac{L}{2}\right) \geqq 0 \quad \text{in } \mathbf{R}^+,$$

whence

$$-\left(\alpha K + \frac{L}{2}\right) \leqq V_{x\tau}^{(n,\alpha)}(x,\tau) \quad \text{in } \mathbf{R}^+ \times [0,T], \tag{4.23}$$

again by a maximum principle argument. From (4.22), (4.23) one gets immediately

$$-\left(\alpha K + \frac{L}{2}\right)(\tau_2 - \tau_1) \leqq V_x^{(n,\alpha)}(x,\tau_2) - V_x^{(n,\alpha)}(x,\tau_1) \leqq 0,$$

$$x > 0, \quad 0 \leqq \tau_1 < \tau_2 \leqq T,$$

and therefore (4.17) in the limit as $n \to \infty$. Q.E.D.

We define

$$\begin{aligned} s^\alpha(\tau) &= \sup\{x > 0 : V_x^{(\alpha)}(x,\tau) = 1\}, & 0 < \tau \leqq T, \\ &= b, & \tau = 0. \end{aligned} \tag{4.24}$$

PROPOSITION 4.2. Boundedness and monotonicity of $s^\alpha(\tau)$. *The function* $s^\alpha(\tau)$ *defined on* $[0,T]$ *for any* $0 \leqq \alpha \leqq 1$ *by* (4.24) *is increasing and right continuous. Moreover,*

$$b \leqq s^\alpha(\tau) \leqq m\left[1 + \left(\alpha K + \frac{L}{2}\right)\tau\right], \qquad 0 \leqq \tau \leqq T, \tag{4.25}$$

where $m = f'^{-1}$ *is the inverse function of* f' *on* $\mathbf{R}^+$.

Proof. The fact that $s^\alpha(\tau)$ is increasing on $[0,T]$ follows from Proposition 4.1, where it is asserted that $V_x^{(\alpha)}(x,\cdot)$ is decreasing on $[0,T]$, for $x > 0$. The "localization inequalities" (4.25) are a consequence of relation (4.18) and monotonicity. In order to prove right continuity of $s^\alpha(\tau)$, consider any $\tau \in [0,T)$ and a sequence $\tau_n \downarrow \tau$. The corresponding sequence $\{s^\alpha(\tau_n); n \in \mathbf{N}\}$ is bounded below and monotone decreasing, therefore, convergent to some number s^*; obviously, $s^\alpha(\tau) \leqq s^*$ since $s^\alpha(\tau) \leqq s^\alpha(\tau_n)$, for all $n \in \mathbf{N}$. But $V_x^{(\alpha)}$ is continuous, so that $V_x^{(\alpha)}[s^\alpha(\tau_n), \tau_n] \to V_x^{(\alpha)}(s^*, \tau)$ as $n \to \infty$. Therefore, $V_x^{(\alpha)}(s^*,\tau) = 1$ which implies $s^* \leqq s^\alpha(\tau)$. Consequently, $s^* = s^\alpha(\tau)$ and right continuity is established.

PROPOSITION 4.3. Monotone dependence on the gain parameter α. *For any two values* α_1, α_2, $0 \leqq \alpha_1 \leqq \alpha_2 \leqq 1$ *of the gain parameter* α,

$$V_x^{(\alpha_1)}(x,\tau) \geqq V_x^{(\alpha_2)}(x,\tau), \qquad (x,\tau) \in \mathbf{R}^+ \times [0,T], \tag{4.26}$$

$$s^{\alpha_1}(\tau) \leqq s^{\alpha_2}(\tau), \qquad 0 \leqq \tau \leqq T. \tag{4.27}$$

Remark. Inequalities (4.26), (4.27) above mean that the use of control is more efficient the larger the gain parameter α is.

Proof. Consider the Cauchy problem (4.7), (4.8) for the two values α_1, α_2 of the gain parameter α, $\alpha_1 \leqq \alpha_2$, and note that

$$(V_x^{(n,\alpha_2)} - V_x^{(n,\alpha_1)})_\tau = \frac{1}{2}(V_x^{(n,\alpha_2)} - V_x^{(n,\alpha_1)})_{xx} + \alpha_2 a_n'(V_x^{(n,\alpha_2)}) V_{xx}^{(n,\alpha_2)}$$
$$- \alpha_1 a_n'(V_x^{(n,\alpha_1)}) V_{xx}^{(n,\alpha_1)} \quad \text{in } \mathbf{R}\times(0, T],$$
$$V_x^{(n,\alpha_2)}(x, 0) - V_x^{(n,\alpha_1)}(x, 0) = 0 \quad \text{in } \mathbf{R}.$$

The difference of the last two terms on the right-hand side of the partial differential equation can be written as

$$\alpha_2 a_n'(V_x^{(n,\alpha_2)}) V_{xx}^{(n,\alpha_2)} - \alpha_1 a_n'(V_x^{(n,\alpha_1)}) V_{xx}^{(n,\alpha_1)}$$
$$= \alpha_2 a_n'(V_2^{(n,\alpha_2)}) \cdot (V_x^{(n,\alpha_2)} - V_x^{(n,\alpha_1)})_x + (\alpha_2 - \alpha_1) a_n'(V_x^{(n,\alpha_1)}) V_{xx}^{(n,\alpha_1)}$$
$$+ \alpha_2 V_{xx}^{(n,\alpha_1)} \frac{a_n'(V_x^{(n,\alpha_2)}) - a_n'(V_x^{(n,\alpha_1)})}{V_x^{(n,\alpha_2)} - V_x^{(n,\alpha_1)}} (V_x^{(n,\alpha_2)} - V_x^{(n,\alpha_1)}).$$

Define the linear parabolic operator

$$\mathbf{M} = \frac{1}{2}\frac{\partial^2}{\partial x^2} + \alpha_2 a_n'(V_x^{(n,\alpha_2)}) \frac{\partial}{\partial x} + \alpha_2 V_{xx}^{(n,\alpha_1)} \frac{a_n'(V_x^{(n,\alpha_2)}) - a_n'(V_x^{(n,\alpha_1)})}{V_x^{(n,\alpha_2)} - V_x^{(n,\alpha_1)}} - \frac{\partial}{\partial \tau},$$

so that

$$\mathbf{M}[V_x^{(n,\alpha_2)} - V_x^{(n,\alpha_1)}] = -(\alpha_2 - \alpha_1) V_{xx}^{(n,\alpha_1)} a_n'(V_x^{(n,\alpha_1)}) \geqq 0 \quad \text{in } \mathbf{R}^+(0, T],$$
$$V_x^{(n,\alpha_2)}(x, 0) - V_x^{(n,\alpha_1)}(x, 0) = 0 \quad \text{in } \mathbf{R}^+.$$

Equation (4.26) now follows by a strong maximum principle argument, similar to that used in the proof of Proposition 4.2, and a passage to the limit as $n \to \infty$. Equation (4.27) is a direct consequence of (4.26) and the definition of $s^\alpha(\tau)$. Q.E.D.

It is easily seen from (2.12) by formal differentiation and from the definition (4.24) of the moving boundary $s^\alpha(\tau)$ that the gradient $V_x^{(\alpha)}$ satisfies the linear equation

$$(V_x^{(\alpha)})_\tau = \frac{1}{2}(V_x^{(\alpha)})_{xx} + \alpha g_\alpha(x, \tau)(V_x^{(\alpha)})_x, \tag{4.28}$$

away from the free boundary $x = s^\alpha(\tau)$, where $g_\alpha(x, \tau)$ is the discontinuous "turned-around" drift

$$g_\alpha(x, \tau) = \begin{cases} -1, & x > s^\alpha(\tau), \\ 0, & |x| \leqq s^\alpha(\tau), \\ 1, & x < -s^\alpha(\tau). \end{cases} \tag{4.29}$$

This fact suggests a new approximation scheme for $V_x^{(\alpha)}$, $V_{xx}^{(\alpha)}$; we approximate the discontinuous drift $g_\alpha(x, \tau)$ by a sequence of smooth functions $g_{\alpha,m}(x, \tau)$ to be defined below in (4.34).

Consider the functions, for each $m \in \mathbf{N}$,

$$e_m(x) = 2m^2\left(\frac{1}{m} - x\right), \qquad 0 \leqq x \leqq \frac{1}{m},$$

$$= 0, \qquad x > \frac{1}{m}, \tag{4.30}$$

$$= e_m(-x), \qquad x < 0,$$

$$\bar{e}_m(x) = m^2 x\left(\frac{2}{m} - x\right), \qquad 0 \leqq x \leqq \frac{1}{m},$$

$$= 1, \qquad x > \frac{1}{m}, \tag{4.31}$$

$$= -\bar{e}_m(-x), \qquad x < 0,$$

$$\bar{\bar{e}}_m(x) = mx^2\left(1 - \frac{2}{3}mx\right), \qquad 0 \leqq x \leqq \frac{1}{m},$$

$$= x - \frac{2}{3m}, \qquad x > \frac{1}{m}, \tag{4.32}$$

$$= \bar{\bar{e}}_m(-x), \qquad x < 0,$$

and notice that $e_m(x)$, $\bar{e}_m(x)$ are the derivatives of $\bar{e}_m(x)$, $\bar{\bar{e}}_m(x)$, respectively. Now define the approximating potential and drift terms $b_{\alpha,m}(x, \tau)$, $g_{\alpha,m}(x, \tau)$ by

$$b_{\alpha,m}(x, \tau) = \left(-\frac{1}{2}\right)[e_m(x - s^\alpha(\tau)) + e_m(x + s^\alpha(\tau))], \tag{4.33}$$

$$g_{\alpha,m}(x, \tau) = \left(-\frac{1}{2}\right)[\bar{e}_m(x - s^\alpha(\tau)) + \bar{e}_m(x + s^\alpha(\tau))], \tag{4.34}$$

where

$$g_{\alpha,m}(x, \tau) = \begin{cases} \displaystyle\int_0^x b_{\alpha,m}(\xi, \tau)\, d\xi, & x > 0, \\ 0, & x = 0, \\ -g_{\alpha,m}(-x, \tau), & x < 0. \end{cases}$$

The Cauchy problem

$$(V_x^{(m,\alpha)})_\tau = \frac{1}{2}(V_x^{(m,\alpha)})_{xx} + \alpha g_{\alpha,m}(x, \tau)(V_x^{(m,\alpha)})_x \quad \text{in } \mathbf{R} \times (0, T], \tag{4.35}$$

$$V_x^{(m,\alpha)}(x, 0) = f'(x) \quad \text{in } \mathbf{R} \tag{4.36}$$

has a unique solution $V_x^{(m,\alpha)}(x, \tau)$ in the strip $\mathbf{R}_T$, which belongs to the space $H_{3+\beta,3+\beta/2}(\bar{Q}_T)$ for any bounded rectangle $Q_T \in \mathbf{R}_T$, where β is the modulus of Hölder continuity of the function $e_m(x)$: $\beta = 1$ in the present case (Ladyženskaja et al. [1968, Theorem 5.2, p. 320]). We can therefore differentiate (4.35) with respect to x and still get an equation holding in the classical sense; thus, we obtain the corresponding Cauchy problem for $V_{xx}^{(m,\alpha)}$,

$$(V_{xx}^{(m,\alpha)})_\tau = \frac{1}{2}(V_{xx}^{(m,\alpha)})_{xx} + \alpha g_{\alpha,m}(x, \tau)(V_{xx}^{(m,\alpha)})_x + \alpha b_{\alpha,m}(x, \tau)(V_{xx}^{(m,\alpha)}) \tag{4.37}$$

$$\text{in } \mathbf{R} \times (0, T],$$

$$V_{xx}^{(m,\alpha)}(x,0)=f''(x) \quad \text{in } \mathbf{R}. \tag{4.38}$$

The latter has a unique solution $V_{xx}^{(m,\alpha)}$ in $\mathbf{R}_T$ belonging to the space $H_{2+\beta,1+\beta/2}(\bar{Q}_T)$, any bounded rectangle $Q_T \subseteq \mathbf{R}_T$.

Because $g_{\alpha,m}(x,\tau)\to g_\alpha(x,\tau)$ uniformly in (x,τ), one gets by the continuous dependence of solutions of PDE's on their coefficients that $V_x^{(m,\alpha)}(x,\tau)$, $V_{xx}^{(m,\alpha)}(x,\tau)$ converge to $V_x^{(\alpha)}(x,\tau)$, $V_{xx}^{(\alpha)}(x,\tau)$ respectively, uniformly on compact (x,τ) sets, as $m\to\infty$. This fact enables one to get stochastic representations for $V_x^{(\alpha)}(x,\tau)$, $V_{xx}^{(\alpha)}(x,\tau)$ though the Feynman-Kac formula.

PROPOSITION 4.4. Representation of $V_x^{(\alpha)}(x,\tau)$. *The gradient $V_x^{(\alpha)}(x,\tau)$ of the value function admits the stochastic representation*

$$V_x^{(\alpha)}(x,\tau)=E\left[f'(x+w_\tau)\cdot\exp\left\{\alpha\int_{T-\tau}^{T} g_\alpha(x+w_{t-T+\tau},T-t)\,dw_t - \frac{\alpha^2}{2}\int_{T-\tau}^{T} g_\alpha^2(x+w_{t-T+\tau},T-t)\,dt\right\}\right] \tag{4.39}$$

where $\{w_t;\,0\leqq t\leqq T\}$ is a Wiener process on the underlying probability space $(\Omega,\mathbf{F},P)$ and E denotes expectation with respect to the probability measure P.

Proof. We introduce the notation

$$\zeta_{T-\tau}^{T}(\alpha g_\alpha)=\alpha\int_{T-\tau}^{T} g_\alpha(x+w_{t-T+\tau},T-t)\,dw_t-\frac{\alpha^2}{2}\int_{T-\tau}^{T} g_\alpha^2(x+w_{t-T+\tau},T-t)\,dt. \tag{4.40}$$

According to the Feynman-Kac theorem (Friedman [1975, p. 148] and Beneš [1974]) the unique solution to the Cauchy problem (4.35)–(4.36) satisfying a polynomial growth condition in x admits the stochastic representation

$$V_x^{(m,\alpha)}(x,\tau)=E[f'(x+w_\tau)\cdot\exp\zeta_{T-\tau}^{T}(\alpha g_{\alpha,m})], \tag{4.41}$$

subject to mild smoothness conditions, such as Lip continuity, on the drift $g_{\alpha,m}$. To show that the right-hand side of (4.41) converges as $m\to\infty$ to that of (4.39) uniformly on compact (x,τ) sets, it is sufficient to prove

$$E|\exp\zeta_{T-\tau}^{T}(\alpha g_{\alpha,m})-\exp\zeta_{T-\tau}^{T}(\alpha g_\alpha)|^2\to 0 \quad \text{as } m\to\infty, \tag{4.42}$$

uniformly on compact (x,τ) sets.

By Girsanov [1960],

$$E[\exp\zeta_{T-\tau}^{T}(\phi)]=1, \tag{4.43}$$

for any bounded, nonanticipative Wiener functional ϕ. Consequently,

$$\begin{aligned}
&E|\exp\zeta_{T-\tau}^{T}(\alpha g_{\alpha,m})-\exp\zeta_{T-\tau}^{T}(\alpha g_\alpha)|^2\\
&=E\left[\exp\left\{\alpha^2\int_{T-\tau}^{T} g_\alpha^2(x+w_{t-T+\tau},T-t)\,dt\right\}\right.\\
&\quad+\exp\left\{\alpha^2\int_{T-\tau}^{T} g_{\alpha,m}^2(x+w_{t-T+\tau},T-t)\,dt\right\}\\
&\quad\left.-2\exp\left\{\alpha^2\int_{T-\tau}^{T} g_\alpha(x+w_{t-T+\tau},T-t)g_{\alpha,m}(x+w_{t-T+\tau},T-t)\,dt\right\}\right].
\end{aligned}$$

The expression under the expectation sign on the right-hand side of the above equality

is bounded in absolute value by $\exp(4\alpha^2\tau) \leqq \exp(4T)$, and it is easy to see that

$$\int_{T-\tau}^{T} g^j_{\alpha,m}(x + w_{t-T+\tau},\, T-t)\, dt \xrightarrow{P} \int_{T-\tau}^{T} g^j_{\alpha}(x + w_{t-T+\tau},\, T-t)\, dt, \qquad j = 1, 2,$$

as $m \to \infty$, uniformly on compact (x, τ) sets. Now (4.42) follows by the bounded convergence theorem. Since $V_x^{(m,\alpha)}(x, \tau)$ converges to $V_x^{(\alpha)}(x, \tau)$, (4.39) follows from (4.41) by a passage to the limit in the latter, as $m \to \infty$.

PROPOSITION 4.5. Representation of $V_{xx}^{(\alpha)}(x, \tau)$. *$V_{xx}^{(\alpha)}(x, \tau)$ admits the stochastic representation*

$$V_{xx}^{(\alpha)}(x, \tau) = E[f''(x + w_\tau) \cdot \exp\{-\alpha\xi_\alpha(\tau, x, w) + \zeta^T_{T-\tau}(\alpha g_\alpha)\}], \tag{4.44}$$

with

$$\begin{aligned}\xi_\alpha(\tau, x, w) = {} & |x + w_\tau - b| - |x - s^\alpha(\tau)| + |x + w_\tau + b| - |x + s^\alpha(\tau)| \\ & + 2\int_{T-\tau}^{T} g_\alpha(x + w_{t-T+\tau},\, T-t)\, dw_t \\ & + 2\int_{T-\tau}^{T} \chi_{\{|x + w_{t-T+\tau}| < s^\alpha(T-t)\}}\, d\rho_\alpha(t),\end{aligned} \tag{4.45}$$

where $\{w_t;\ 0 \leqq t \leqq T\}$ is a Wiener process on the underlying probability space $(\Omega, \mathbf{F}, P)$, E denotes expectation with respect to P, $\rho_\alpha(t) = -s^\alpha(T-t)$, $T - \tau \leqq t \leqq T$, is a bounded, increasing and right continuous function and $\zeta^T_{T-\tau}(\alpha g_\alpha)$ is defined in (4.40).

Remark. The expression $\xi_\alpha(\tau, x, w)$ defined in (4.45) can be identified as the analogue of Tanaka's formula for the local time of the "solution process" $x + w_{t-T+\tau}$ at the two branches of the free boundary $\pm s^\alpha(T-t)$, $T - \tau \leqq t \leqq T$. For a definition of the local time for the Wiener process, as well as a derivation of Tanaka's formula in that case, see McKean [1962].

Proof. According to the Feynman-Kac theorem (Friedman [1975, p. 148] and Beneš [1974]) the unique solution to the Cauchy problem (4.37), (4.38) satisfying a polynomial growth condition in x admits the stochastic representation

$$\begin{aligned}V_{xx}^{(m,\alpha)}(x, \tau) = E\Bigg[& f''(x + w_\tau) \exp\left\{\alpha \int_{T-\tau}^{T} b_{\alpha,m}(x + w_{t-T+\tau},\, T-t)\, dt\right\} \\ & \cdot \exp \zeta^T_{T-\tau}(\alpha g_{\alpha,m})\Bigg]\end{aligned} \tag{4.46}$$

subject to mild smoothness assumptions on the coefficients, such as Lip continuity of the drift $g_{\alpha,m}$ and Hölder continuity of the potential term $b_{\alpha,m}$, which are satisfied in our case.

If

$$\begin{aligned}\eta_t &= x + w_{t-T+\tau} - s^\alpha(T-t) = x + w_{t-T+\tau} + \rho_\alpha(t), \\ \theta_t &= x + w_{t-T+\tau} + s^\alpha(T-t) = x + w_{t-T+\tau} - \rho_\alpha(t),\end{aligned} \qquad T - t \leqq t \leqq T;$$

an application of Ito's formula gives

$$\begin{aligned}\int_{T-\tau}^{T} b_{\alpha,m}(x + w_{t-T+\tau},\, T-t)\, dt = {} & \bar{\bar{e}}_m(\eta_{T-\tau}) - \bar{\bar{e}}_m(\eta_T) + \bar{\bar{e}}_m(\theta_{T-\tau}) - \bar{\bar{e}}_m(\theta_T) \\ & + \int_{T-\tau}^{T} \bar{e}_m(\eta_t)\, dw_t + \int_{T-\tau}^{T} \bar{e}_m(\theta_t)\, dw_t \\ & + \int_{T-\tau}^{T} \bar{e}_m(\eta_t) d\rho_\alpha(t) - \int_{T-\tau}^{T} \bar{e}_m(\theta_t) d\rho_\alpha(t).\end{aligned} \tag{4.47}$$

Now we pass to the limit as $m \to \infty$ in (4.47). Observe that $\bar{e}_m(\xi) \to \operatorname{sgn} \xi$, $\bar{\bar{e}}_m(\xi) \to |\xi|$. It is easily shown that

$$\int_{T-\tau}^{T} \bar{e}_m(\eta_t) dw_t \xrightarrow{P} \int_{T-\tau}^{T} \operatorname{sgn}(\eta_t) dw \quad \text{as } m \to \infty, \tag{4.48}$$

$$\int_{T-\tau}^{T} \bar{e}_m(\eta_t) d\rho_\alpha(t) \xrightarrow{P} \int_{T-\tau}^{T} \operatorname{sgn} \eta_t \, d\rho_\alpha(t) \quad \text{as } m \to \infty. \tag{4.49}$$

Both relations remain true if η_t is replaced by θ_t.

From (4.47), (4.48) and (4.49) one concludes that along some subsequence (m_n),

$$\begin{aligned}
&\int_{T-\tau}^{T} b_{\alpha,m_n}(x + w_{t-T+\tau}, T-t)\, dt \\
&\quad \xrightarrow[n\uparrow\infty]{\text{a.s. } (P)} |x - s^\alpha(\tau)| - |x + w_\tau - b| + |x + s^\alpha(\tau)| - |x + w_\tau + b| \\
&\quad + \int_{T-\tau}^{T} [\operatorname{sgn}(x + w_{t-T+\tau} - s^\alpha(T-t)) + \operatorname{sgn}(x + w_{t-T+\tau} + s^\alpha(T-t))]\, dw_t \\
&\quad + \int_{T-\tau}^{T} [\operatorname{sgn}(x + w_{t-T+\tau} - s^\alpha(T-t)) - \operatorname{sgn}(x + w_{t-T+\tau} + s^\alpha(T-t))] d\rho_\alpha(t).
\end{aligned} \tag{4.50}$$

But

$$\operatorname{sgn}(x + w_{t-T+\tau} - s^\alpha(T-t)) + \operatorname{sgn}(x + w_{t-T+\tau} + s^\alpha(T-t)) = -2g_\alpha(x + w_{t-T+\tau}, T-t),$$

$$\operatorname{sgn}(x + w_{t-T+\tau} - s^\alpha(T-t)) - \operatorname{sgn}(x + w_{t-T+\tau} + s^\alpha(T-t)) = -2\chi_{\{|x+w_{t-T+\tau}|<s^\alpha(T-t)\}},$$

and substitution into (4.50) yields

$$B_{\alpha,n} = \alpha \int_{T-\tau}^{T} b_{\alpha,m_n}(x + w_{t-T+\tau}, T-t)\, dt \xrightarrow[n\to\infty]{\text{a.s. } (P)} B_\alpha = -\alpha \xi_\alpha(\tau, x, w), \tag{4.51}$$

where $\xi_\alpha(\tau, x, w)$ is the entity defined in (4.45). Since $B_{\alpha,n} \leqq 0$, a.s. (P), for any $n \in \mathbf{N}$, one concludes that $B_\alpha \leqq 0$ or $\xi_\alpha(\tau, w, x) \geqq 0$, a.s. (P).

It has to be shown that the right-hand side of the representation of $V_{xx}^{(m,\alpha)}(x, \tau)$ in (4.46) converges as $m \to \infty$ to that of (4.44), namely

$$\begin{aligned}
&E[f''(x + w_\tau) \exp B_{\alpha,n} \exp \zeta_{T-\tau}^{T}(\alpha g_{\alpha,m_n})] \\
&\qquad \xrightarrow[n\uparrow\infty]{} E[f''(x + w_\tau) \exp B_\alpha \exp \zeta_{T-\tau}^{T}(\alpha g_\alpha)],
\end{aligned} \tag{4.52}$$

uniformly on compact (x, τ) sets. It is easy to prove that

$$\zeta_{T-\tau}^{T}(\alpha g_{\alpha,m_n}) \xrightarrow{P} \zeta_{T-\tau}^{T}(\alpha g_\alpha) \quad \text{as } n \to \infty, \tag{4.53}$$

and that

$$\begin{aligned}
&E|f''(x + w_\tau) \exp B_{\alpha,n} \exp \zeta_{T-\tau}^{T}(\alpha g_{\alpha,m_n}) - f''(x + w_\tau) \exp B_\alpha \exp \zeta_{T-\tau}^{T}(\alpha g_\alpha)| \\
&\leqq K \cdot E|\exp \zeta_{T-\tau}^{T}(\alpha g_{\alpha,m_n}) - \exp \zeta_{T-\tau}^{T}(\alpha g_\alpha)| \\
&\quad + K \cdot E[\exp \zeta_{T-\tau}^{T}(\alpha g_\alpha) |\exp B_{\alpha,n} - \exp B_\alpha|].
\end{aligned} \tag{4.54}$$

Evidently,

$$E[\exp \zeta^T_{T-\tau}(\alpha g_{a,m_n})] = E[\exp \zeta^T_{T-\tau}(\alpha g_\alpha)] = 1,$$

by (4.43). Lemma 6.7 in Lipster and Shiryayev [1977] asserts that $E|\exp \zeta^T_{T-\tau}(\alpha g_{\alpha,m_n}) - \exp \zeta^T_{T-\tau}(\alpha g_\alpha)|$ tends to zero as $n \to \infty$. As for the second term on the right-hand side of (4.54), observe that $\exp \zeta^T_{T-\tau}(\alpha g_\alpha)|\exp B_{\alpha,m_n} - \exp B_\alpha| \leqq 2 \exp \zeta^T_{T-\tau}(\alpha g_\alpha)$; convergence of this term to zero as $n \uparrow \infty$ follows from (4.51) and the dominated convergence theorem.

This establishes (4.52); a passage to the limit as $n \to \infty$ on both sides of (4.46) yields the desired representation (4.44).

We are now in a position to prove the following important result.

PROPOSITION 4.6. Propagation of uniform convexity. *There exists a positive function* $k(T)$, $T > 0$, *independent of* $\alpha \in [0, 1]$, *such that*

$$V^{(\alpha)}_{xx}(x, \tau) \geqq k(T) > 0, \qquad (x, \tau) \in \mathbf{R} \times [0, T]. \tag{4.55}$$

Proof. From obvious estimates, one finds

$$a\xi^a(\tau, x, w) - \zeta^T_{T-\tau}(a g_a) \leqq 2|w_\tau| + 4\left[m\left(1 + \left[K + \frac{L}{2}\right]\tau\right) - b\right] - \zeta^T_{T-\tau}(-\alpha g_\alpha), \quad \text{a.s. } (P).$$

Therefore, by virtue of the representation (4.44),

$$\begin{aligned} V^{(\alpha)}_{xx}(x, \tau) \geqq k \cdot \exp\left\{-4\left[m\left(1 + \left[K + \frac{L}{2}\right]\tau\right) - b\right]\right\} \\ \cdot E[e^{-2|w_\tau|} \exp \zeta^T_{T-\tau}(-\alpha g_\alpha)]. \end{aligned} \tag{4.56}$$

We consider a family of probability measures $\bar{P}_\alpha$ on $(\Omega, \mathbf{F})$, $0 \leqq \alpha \leqq 1$, defined by

$$\bar{P}_\alpha(d\omega) = \exp \zeta^T_{T-\tau}(-\alpha g_\alpha) \cdot P(d\omega). \tag{4.57}$$

According to Girsanov's theorem, for each $0 \leqq \alpha \leqq 1$, the process

$$\bar{w}^\alpha_{t-(T-\tau)} = w_{t-(T-\tau)} + \alpha \int_{T-\tau}^t g_\alpha(x + w_{\lambda - T + \tau}, T - \lambda)\, d\lambda, \qquad T - \tau \leqq t \leqq T \tag{4.58}$$

is a Wiener process on the space $(\Omega, \mathbf{F}, \bar{P}_\alpha)$. Obviously, from (4.58), $|w_\tau| \leqq |\bar{w}^\alpha_\tau| + \tau$, a.s. (P), because the measures P, $\bar{P}_\alpha$ are equivalent. Therefore, from (4.56),

$$V^{(\alpha)}_{xx}(x, \tau) \geqq k \cdot \exp\left\{-4\left[m\left(1 + \left[K + \frac{L}{2}\right]\tau\right) - b\right] - \tau\right\} \cdot \bar{E}_\alpha\,[\exp(-2|\bar{w}^\alpha_\tau|)]. \tag{4.59}$$

But

$$\begin{aligned} \bar{E}_\alpha[\exp(-2|\bar{w}^\alpha_\tau|)] = E[\exp(-2|w_\tau|)] = 2(2\pi)^{-1/2} e^{2\tau} \int_{2\tau^{1/2}}^\infty e^{-x^2/2}\, dx \\ \geqq \frac{2(2\pi)^{-1/2}}{\tau^{1/2} + (1+\tau)^{1/2}} \end{aligned}$$

where we have used the estimate $\int_\alpha^\infty e^{-x^2/2}\, dx \geqq 2e^{-a^2/2}/(a + (4 + a^2)^{1/2})$ due to Komatsu (Ito-McKean [1974, p. 17]). So, finally,

$$V^{(\alpha)}_{xx}(x, \tau) \geqq k(T) = \frac{k}{(2\pi)^{1/2}} \frac{\exp\left\{-4\left[m\left(1 + \left[K + \frac{L}{2}\right]T\right) - b\right] - T\right\}}{(1+T)^{1/2}},$$

$$(x, \tau) \in \mathbf{R} \times [0, T].$$

COROLLARY. *For any* $0 \leqq \alpha \leqq 1$, $s^{\alpha}(\tau)$ *is continuous on* $[0, T]$.

Proof. We already know that $V_x^{(\alpha)}(x, \tau) \to \infty$ as $x \to \infty$ (relation (4.18)). Now (4.55) gives the additional information that, for any $0 \leqq \tau \leqq T$, $V_x^{(\alpha)}(\cdot, \tau)$ is strictly increasing and therefore $s^{\alpha}(\tau)$ is unambiguously defined through

$$V_x^{(\alpha)}[s^{\alpha}(\tau), \tau] = 1, \qquad 0 \leqq \tau \leqq T. \tag{4.60}$$

Consider $0 \leqq \tau \leqq T$ and a sequence of points $\{\tau_n\}$ in $[0, T]$, $\tau_n \to \tau$. By (4.25), $\{s^{\alpha}(\tau_n)\}$ is a bounded sequence so we can extract a convergent subsequence $\{s^{\alpha}(\tau_{n_k})\}$ thereof, converging say to s^*. By continuity of $V_x^{(\alpha)}(x, \tau)$, $V_x^{(\alpha)}[s^{\alpha}(\tau_{n_k}), \tau_{n_k}] \to V_x^{(\alpha)}(s^*, \tau)$ as $k \to \infty$; hence $V_x^{(\alpha)}(s^*, \tau) = 1$. From (4.60), one gets $s^* = s^{\alpha}(\tau)$. Thus, any convergent subsequence of $\{s^{\alpha}(\tau_n)\}$ converges to $s^{\alpha}(\tau)$; the same holds true therefore for the whole sequence, which establishes continuity of $s^{\alpha}(\tau)$ on $[0, T]$.

In later sections (6, 7) it will be shown that $s^{\alpha}(\tau)$ is actually continuously differentiable on $[0, T]$. The "a priori" bounds established below will play a crucial role in the proof of this fact (proof of Theorem 2.2, § 7).

PROPOSITION 4.7. A priori bound on $\|s'^{\alpha}\|$. *The function* $s^{\alpha}(\tau)$ *is Lipschitz continuous on* $[0, T]$, *with a constant independent of* $0 \leqq \alpha \leqq 1$.

If $s^{\alpha}(\tau)$ *is also continuously differentiable on* $[0, T]$, *then for any* $0 \leqq \alpha \leqq 1$,

$$\|s'^{\alpha}\| = \sup_{0 \leqq \tau \leqq T} |s'^{\alpha}(\tau)| \leqq \frac{K + \frac{L}{2}}{k(T)}, \tag{4.61}$$

where $k(T)$ *is the positive constant in* (4.55).

Proof. Suppose $0 \leqq \tau \leqq \tau + \varepsilon \leqq T$; by definition, $V_x^{(\alpha)}[s^{\alpha}(\tau), \tau] = V_x^{(\alpha)}[s^{\alpha}(\tau + \varepsilon), \tau + \varepsilon] = 1$ and consequently,

$$\begin{aligned} 0 \leqq V_x^{(\alpha)}[s^{\alpha}(\tau + \varepsilon), \tau] - V_x^{(\alpha)}[s^{\alpha}(\tau), \tau] &= V_x^{(\alpha)}[s^{\alpha}(\tau + \varepsilon), \tau] - V_x^{(\alpha)}[s^{\alpha}(\tau + \varepsilon), \tau + \varepsilon] \\ &\leqq \left(K + \frac{L}{2}\right)\varepsilon, \end{aligned}$$

for any $0 \leqq \alpha \leqq 1$, by (4.17), Proposition 4.1.

On the other hand, there exists a number s^* between $s^{\alpha}(\tau)$ and $s^{\alpha}(\tau + \varepsilon)$ such that, by the mean value theorem,

$$V_x^{(\alpha)}[s^{\alpha}(\tau + \varepsilon), \tau] - V_x^{(\alpha)}[s^{\alpha}(\tau), \tau] = V_{xx}^{(\alpha)}(s^*, \tau)[s^{\alpha}(\tau + \varepsilon) - s^{\alpha}(\tau)],$$

and therefore,

$$0 \leqq s^{\alpha}(\tau + \varepsilon) - s^{\alpha}(\tau) \leqq \frac{\left(K + \frac{L}{2}\right)\varepsilon}{V_{xx}^{(\alpha)}(s^*, \tau)} \leqq \frac{K + \frac{L}{2}}{k(T)}\varepsilon,$$

whence Lip continuity and the a priori bound (4.61) on the derivative.

COROLLARY. *If* $s^{\alpha}(\tau)$ *is continuously differentiable on* $[0, T]$ *and* $w^{\alpha}(\tau) = V^{(\alpha)}[s^{\alpha}(\tau), \tau]$, *then*

$$\|w'^{\alpha}\| = \sup_{0 \leqq \tau \leqq T} |w'^{\alpha}(\tau)| \leqq \frac{K + \frac{L}{2}}{k(T)} + \frac{K}{2}. \tag{4.62}$$

Proof. The proof is an immediate consequence of (4.61), (4.15) and the fact that $w'^{\alpha}(\tau) = s'^{\alpha}(\tau) + \frac{1}{2}V_{xx}^{(\alpha)}[s^{\alpha}(\tau), \tau]$.

Collecting the various pieces together we can now prove Theorem 2.1.

Proof of Theorem 2.1. The Cauchy problem (2.6), (2.7) has a unique $C^{2,1}$ solution on $\mathbf{R}\times[0, T]$ which actually belongs to the space $H_{2+\beta,1+\beta/2}(\bar{Q}_T)$, $\beta = 1$ and Q_T any bounded rectangle in $\mathbf{R}_T$. Strict monotonicity and unbounded increase of the gradient $V_x^{(\alpha)}(x, \tau)$ follow from Proposition 4.6 and (4.18), respectively, and so the boundary points $s^{\alpha}(\tau)$, $0 \leqq \tau \leqq T$, are unambiguously defined through (4.60): $V_x^{(\alpha)}[s(\tau), \tau] = 1$. The boundary is monotone increasing, localized by (4.25), continuous (Corollary to Proposition 4.6) and Lip (Proposition 4.7).

5. The integral equations. In the first two paragraphs of this section, we prove the following proposition:

PROPOSITION 5.1. *The free-boundary problem* (2.15)–(2.20), *with a continously differentiable free-boundary function* $s^{\alpha}(\tau)$, *is equivalent to the problem of finding a continuous solution* $[s'^{\alpha}(\tau), w'^{\alpha}(\tau)]$ *to the pair of integral equations* (5.10) *and* (5.17) *below.*

In § 5.1 we prove the necessity and in § 5.2 the sufficiency of the pair of integral equations. The latter are used in § 5.3 to study the special case $f(x) = x^2$ and in § 5.4 to prove the strict monotonicity of the free boundary.

5.1. Analysis. Suppose that the free-boundary problem (2.15)–(2.20) has the solution $s^{\alpha}(\tau)$, $V^{(\alpha)}(x, \tau)$ in the sense of Definition 2.1, with $s^{\alpha}(\tau)$ continuously differentiable on $[0, T]$. We introduce the fundamental Gaussian kernel

$$K(x, \tau; \xi, u) = [2\pi(\tau - u)]^{-1/2} \exp\left[-\frac{1}{2}\frac{(x-\xi)^2}{\tau - u}\right], \qquad x, \xi \in \mathbf{R}, \quad 0 \leqq u < \tau, \tag{5.1}$$

along with Green's and Neumann's functions for the half-plane $x > 0$,

$$G(x, \tau; \xi, u) = K(x, \tau; \xi, u) - K(x, \tau; -\xi, u) \tag{5.2}$$

and

$$N(x, \tau; \xi, u) = K(x, \tau; \xi, u) + K(x, \tau; -\xi, u), \tag{5.3}$$

respectively. Each one of them, designated by the generic symbol M, satisfies the heat equation $M_\tau = \frac{1}{2}M_{xx}$ $(M_u + \frac{1}{2}M_{\xi\xi} = 0)$ in the forward (backward) variables. Also, $K_\xi = -K_x$, $N_\xi = -G_x$.

In the region $D^*(0)$ we use the standard method (for problems of the Stefan type—see Friedman [1959]) of integrating Green's identity

$$\frac{\partial}{\partial \xi}(NV_1^{(\alpha)} - N_\xi V^{(\alpha)}) - \frac{\partial}{\partial u}(2NV^{(\alpha)}) = 0, \tag{5.4}$$

over the domain $0 < \xi < s^{\alpha}(u)$, $0 < \varepsilon < u < \tau - \varepsilon$; letting $\varepsilon \downarrow 0$, we get the representation of $V^{(\alpha)}(x, \tau)$ in $D^*(0)$,

$$\begin{aligned} 2V^{(\alpha)}(x, \tau) = &\int_0^{\tau} N[x, \tau; s^{\alpha}(u), u]\{V_1^{(\alpha)}[s(u), u] + 2V^{(\alpha)}[s^{\alpha}(u), u]s'^{\alpha}(u)\}\, du \\ &+ \int_0^{\tau} G_x[x, \tau; s^{\alpha}(u), u]V^{(\alpha)}[s^{\alpha}(u), u]\, du + 2\int_0^{b} N(x, \tau; \xi, 0)f(\xi)\, d\xi. \end{aligned} \tag{5.5}$$

By assumption, $V_1^{(\alpha)}[s^{\alpha}(\tau), \tau] = V_x^{(\alpha)}[s^{\alpha}(\tau), \tau] = 1$; introducing the notation

$$w^{\alpha}(\tau) = V^{(\alpha)}[s^{\alpha}(\tau), \tau], \tag{5.6}$$

one gets

$$2V^{(\alpha)}(x,\tau)=\int_0^\tau N[x,\tau;s^\alpha(u),u][1+2w^\alpha(u)s'^\alpha(u)]\,du$$
$$+\int_0^\tau G_x[x,\tau;s^\alpha(u),u]w^\alpha(u)\,du+2\int_0^b N(x,\tau;\xi,0)f(\xi)\,d\xi, \tag{5.7}$$

and upon differentiating with respect to x,

$$2V_1^{(\alpha)}(x,\tau)=\int_0^\tau N_x[x,\tau;s^\alpha(u),u]\,du+2\int_0^\tau G[x,\tau;s^\alpha(u),u]w'^\alpha(u)\,du$$
$$+2\int_0^b G(x,\tau;\xi,0)f'(\xi)\,d\xi, \qquad (x,\tau)\in D^*(0). \tag{5.8}$$

The free boundary $\{s^\alpha(\tau);\,0\leqq\tau\leqq T\}$ is assumed to be continuously differentiable on $[0,T]$. For such functions (or even Lip on $[0,T]$) the "jump relation" below is valid (see Friedman [1959], Rubinštein [1967]):

$$\lim_{x\uparrow s^\alpha(\tau)}\int_0^\tau N_x[x,\tau;s^\alpha(u),u]\rho(u)\,du=\rho(\tau)+\int_0^\tau N_x[s^\alpha(\tau),\tau;s^\alpha(u),u]\rho(u)\,du, \tag{5.9}$$

$\rho(\tau)$ being any continuous function on $[0,T]$. An application of the jump relation (5.9) to the representation of the gradient in (5.8) with $\rho(\tau)=1$ yields the integral equation

$$1=\int_0^\tau N_x[s^\alpha(\tau),\tau;s^\alpha(u),u]\,du+2\int_0^\tau G[s^\alpha(\tau),\tau;s^\alpha(u),u]w'^\alpha(u)\,du$$
$$+2\int_0^b G[s^\alpha(\tau),\tau;\xi,0]\cdot f'(\xi)\,d\xi. \tag{5.10}$$

Let us now represent $V^{(\alpha)}(x,\tau)$ in $D(-\alpha)$. The function

$$\Gamma(x,\tau;\xi,u)=K(x-\alpha\tau,\tau;\xi-\alpha u,u), \qquad x,\xi\in\mathbf{R}, \qquad 0\leqq u<\tau\leqq T, \tag{5.11}$$

satisfies the equation

$$\frac{1}{2}\Gamma_{xx}-\alpha\Gamma_x-\Gamma_\tau=0\left(\frac{1}{2}\Gamma_{\xi\xi}+\alpha\Gamma_\xi+\Gamma_u=0\right), \tag{5.12}$$

in the forward (backward) variables. In $D(-\alpha)$ we integrate Green's identity

$$\frac{\partial}{\partial\xi}(\Gamma V_1^{(\alpha)}-\Gamma_\xi V^{(\alpha)}-2\alpha\Gamma V^{(\alpha)})-\frac{\partial}{\partial u}(2\Gamma V^{(\alpha)})=-2\alpha\Gamma \tag{5.13}$$

over the region $s^\alpha(u)<\xi<M$, $0<\varepsilon<u<\tau-\varepsilon$. Since $V^{(\alpha)}(x,\tau)\leqq K(x^2+\tau)$ (see (4.12)) we obtain the following representation for $V^{(\alpha)}(x,\tau)$ in $D(-\alpha)$, by letting $\varepsilon\downarrow 0$, $M\uparrow\infty$:

$$2V^{(\alpha)}(x,\tau)=-\int_0^\tau K[x-\alpha\tau,\tau;s^\alpha(u)-\alpha u,u]\{1+2w^\alpha(u)[s'^\alpha(u)-\alpha]\}\,du$$
$$-\int_0^\tau K_x[x-\alpha\tau,\tau;s^\alpha(u)-\alpha u,u]w^\alpha(u)\,du$$
$$+2\int_b^\infty K(x-\alpha\tau,\tau;\xi,0)f(\xi)\,d\xi$$
$$+2\alpha\int_0^\tau\int_{s^\alpha(u)}^\infty K(x-\alpha\tau,\tau;\xi-\alpha u,u)\,d\xi\,du. \tag{5.14}$$

Differentiation of (5.14) with respect to x gives the representation for the gradient in $D(-\alpha)$,

$$
\begin{aligned}
2V_1^{(\alpha)}(x,\tau) = &-\int_0^\tau K_x[x-\alpha\tau,\tau;s^\alpha(u)-\alpha u,u]\,du \\
&-2\int_0^\tau K[x-\alpha\tau,\tau;s^\alpha(u)-\alpha u,u][w'^\alpha(u)-\alpha]\,du \\
&+2\int_b^\infty K(x-\alpha\tau,\tau;\xi,0)f'(\xi)\,d\xi.
\end{aligned} \tag{5.15}
$$

We now let $x \downarrow s^\alpha(\tau)$ in the above expression. The "jump relation" now takes the form (see Rubinštein [1967, p. 99])

$$
\begin{aligned}
&\lim_{x\downarrow s^\alpha(\tau)} \int_0^\tau K_x[x-\alpha\tau,\tau;s^\alpha(u)-\alpha u,u]\rho(u)\,du \\
&\qquad = -\rho(\tau)+\int_0^\tau K_x[s^\alpha(\tau)-\alpha\tau,\tau;s^\alpha(u)-\alpha u,u]\rho(u)\,du,
\end{aligned} \tag{5.16}
$$

for any continuous function $\rho(\tau)$ on $[0, T]$. An application of (5.16) to the representation (5.15) for $\rho(\tau)=1$ in the limit as $x \downarrow s^\alpha(\tau)$ gives the integral equation

$$
\begin{aligned}
1 = &-\int_0^\tau K_x[s^\alpha(\tau)-\alpha\tau,\tau;s^\alpha(u)-\alpha u,u]\,du \\
&-2\int_0^\tau K[s^\alpha(\tau)-\alpha\tau,\tau;s^\alpha(u)-\alpha u,u][w'^\alpha(u)-\alpha]\,du \\
&+2\int_b^\infty K[s^\alpha(\tau)-\alpha\tau,\tau;\xi,0]f'(\xi)\,d\xi.
\end{aligned} \tag{5.17}
$$

5.2. Synthesis. To prove sufficiency of (5.10), (5.17) for the free-boundary problem, it is assumed that the system of equations possesses a solution pair $[s'^\alpha(\tau), w'^\alpha(\tau)]$ with both functions continuous on $[0, T]$, $s^\alpha(0)=b$, $w^\alpha(0)=f(b)$, and a solution to the free-boundary problem in the sense of Definition 2.1 is constructed from them. We make explicit use of the fact, established in Proposition 4.2, that the boundary function $s^\alpha(\tau)$ is increasing on $[0, T]$.

Indeed, let $[s'^\alpha(\tau), w'^\alpha(\tau)]$ be such a solution to the system of equations (5.10), (5.17) and define the function $V^{(\alpha)}(x,\tau)$ in $D^*(0)$ by

$$
\begin{aligned}
2V^{(\alpha)}(x,\tau) = &\int_0^\tau N[x,\tau;s^\alpha(u),u][1+2w^\alpha(u)s'^\alpha(u)]\,du \\
&+\int_0^\tau G_x[x,\tau;s^\alpha(u),u]w^\alpha(u)\,du \\
&+2\int_0^b N(x,\tau;\xi,0)f(\xi)\,d\xi,
\end{aligned}
$$

in accordance with (5.7). It is easily checked that $V^{(\alpha)}(x,\tau)$ thus defined satisfies the heat equation (2.15) in $D^*(0)$, along with $V^{(\alpha)}(x,0)=f(x)$, $0 \leqq x \leqq b$ (compare with

Lemma 9.1). Differentiation with respect to x yields, in accordance with (5.8),

$$2V_1^{(\alpha)}(x,\tau)=\int_0^\tau N_x[x,\tau;s^\alpha(u)\,u]\,du+2\int_0^\tau G[x,\tau;s^\alpha(u),u]w'^\alpha(u)\,du$$
$$+2\int_0^b G(x,\tau;\xi,0)f'(\xi)\,d\xi,$$

from which it is obvious that $V_1^{(\alpha)}(0,\tau)=0$, because $N_x(0,\tau;\xi,u)=G(0,\tau;\xi,u)=0$. Therefore (2.17) is also satisfied. It remains to check the conditions of Definition 2.1 along the boundary. Let $x\uparrow s^\alpha(\tau)$ in the above expression for the gradient. Using the jump relation (5.9) and the integral equation (5.10), we get

$$\begin{aligned}2V_1^{(\alpha)}[s^\alpha(\tau),\tau]&=1+\int_0^\tau N_x[s^\alpha(\tau),\tau;s^\alpha(u),u]\,du\\&\quad+2\int_0^\tau G[s^\alpha(\tau),\tau;s^\alpha(u),u]w'^\alpha(u)\,du\\&\quad+2\int_0^b G[s^\alpha(\tau),\tau;\xi,0]f'(\xi)\,d\xi\\&=2,\end{aligned}$$

whence the condition (2.20).

Now we integrate Green's identity (5.4) over the region $0<x<s^\alpha(u)$, $0<\varepsilon<u<\tau-\varepsilon$, using the initial and boundary data $V^{(\alpha)}(x,0)=f(x)$, $0\leqq x\leqq b$, $V_1^{(\alpha)}(0,\tau)=0$ and $V_1^{(\alpha)}[s^\alpha(\tau),\tau]=1$, $0\leqq\tau\leqq T$. In the limit as $\varepsilon\downarrow 0$, we get a representation analogous to (5.5) which, compared with the above definition and with $p(\tau)=V^{(\alpha)}[s^\alpha(\tau),\tau]-w^\alpha(\tau)$, yields

$$\int_0^\tau\{G_x[x,\tau;s^\alpha(u),u]+2s'^\alpha(u)N[x,\tau;s^\alpha(u),u]\}p(u)\,du=0\quad\text{in } D^*(0). \tag{5.18}$$

We aim to show that $p(\tau)=0$ on $[0,T]$. Let $x\uparrow s^\alpha(\tau)$ in (5.18) and get, from the jump relation (5.9),

$$p(\tau)+\int_0^\tau F(\tau,u)p(u)\,du=0,\qquad 0\leqq\tau\leqq T, \tag{5.19}$$

where $F(\tau,u)=G_x[s^\alpha(\tau),\tau;s^\alpha(u),u]+2s'^\alpha(u)N[s^\alpha(\tau),\tau;s^\alpha(u),u]$. If V is an upper bound on $\|s'^\alpha\|=\sup_{0\leqq\tau\leqq T}|s'^\alpha(\tau)|$, then $b\leqq s^\alpha(\tau)\leqq b+VT$, for any $0\leqq\tau\leqq T$, and

$$|F(\tau,u)|\leqq 3V[2\pi(\tau-u)]^{-1/2}+\left[2V+\frac{2b+2VT}{\tau-u}\right][2\pi(\tau-u)]^{-1/2}\exp\left[\frac{-b^2}{2(\tau-u)}\right]$$
$$\leqq(2\pi)^{-1/2}\left[3V+\frac{4(b+2VT)}{b^2}\right](\tau-u)^{-1/2}=\frac{q}{(\tau-u)^{1/2}}.$$

Applied to (5.19), the previous result gives

$$|p(\tau)|\leqq q\int_0^\tau\frac{|p(u)|}{(\tau-u)^{1/2}}\,du,\qquad 0\leqq\tau\leqq T.$$

Now Cannon and Hill [1967, Lemma 7, p. 7] guarantees that $p(\tau)=0$, $0\leqq\tau\leqq T$. Therefore $V^{(\alpha)}[s^\alpha(\tau),\tau]=w^\alpha(\tau)$, $0\leqq\tau\leqq T$. In a similar fashion, one can construct a solution $V^{(\alpha)}(x,\tau)$ of (2.16) in $D(-\alpha)$ and verify all requisite conditions.

5.3. Remarks on a special case. In the important special case of a quadratic final cost function $f(x)=x^2 (b=\frac{1}{2})$ the system of integral equations (5.10), (5.17) for the fully controlled case $\alpha=1$ takes the form

$$\begin{aligned} 1 = &\int_0^\tau N_x[s(\tau),\tau;s(u),u]\,du + 2\int_0^\tau G[s(\tau),\tau;s(u),u]w'(u)\,du \\ &+4\int_0^{1/2} G[s(\tau),\tau;\xi,0]\xi\,d\xi, \end{aligned} \tag{5.20}$$

$$\begin{aligned} 1 = &-\int_0^\tau K_x[s(\tau)-\tau,\tau;s(u)-u,u]\,du \\ &-2\int_0^\tau K[s(\tau)-\tau,\tau;s(u)-u,u][w'(u)-1]\,du \\ &+4\int_{1/2}^\infty K[s(\tau)-\tau,\tau;\xi,0]\xi\,d\xi. \end{aligned} \tag{5.21}$$

In an attempt to solve this problem, R. C. Davis [1968] came up with the answer $s(\tau)=\frac{1}{2}+\tau$ for the free-boundary function. However, the method and particularly the argument on p. 71 of the above-mentioned paper are wrong. Here, we show that the result is also wrong, namely that *for the particular choice* $s(\tau)=\frac{1}{2}+\tau$, $0\leqq\tau\leqq T$, *there exists no continuous function* $w'(\tau)$ *on* $[0,T]$ *in such a way that* (5.20), (5.21) *can be simultaneously satisfied for all* $0\leqq\tau\leqq T$.

Suppose the contrary is true; (5.21) then becomes the Volterra integral equation of the first kind: $\int_0^\tau (\tau-u)^{-1/2}[w'(u)-1]\,du = 2\tau^{1/2}$, $0\leqq\tau\leqq T$, whose solution is found, by inspection, to be $w'(\tau)=2$, $0\leqq\tau\leqq T$. Substituting this value into (5.20), we get

$$\begin{aligned} 1 = &\int_0^\tau \left[N_x\left(\frac{1}{2}+\tau,\tau;\frac{1}{2}+u,u\right)+4G\left(\frac{1}{2}+\tau,\tau;\frac{1}{2}+u,u\right)\right]du \\ &+4\int_{-1/2}^{1/2} K\left(\tau+\frac{1}{2},\tau;\xi,0\right)\xi\,d\xi. \end{aligned} \tag{5.22}$$

Noticing that

$$\int_0^\tau [2\pi(\tau-u)]^{-1/2}\exp\left(-\frac{\tau-u}{2}\right)du = 2\Phi(\tau^{1/2})-1,$$

$$\int_0^\tau [2\pi(\tau-u)]^{-1/2}\left[2+\frac{1+\tau+u}{\tau-u}\right]\exp\left\{-\frac{(1+\tau+u)^2}{2(\tau-u)}\right\}du = 2[1-\Phi(\tau^{1/2}+\tau^{-1/2})],$$

where $\phi(x)=(2\pi)^{-1/2}\exp(-x^2/2)$, $\Phi(x)=\int_{-\infty}^x \phi(\xi)\,d\xi$, we get from (5.22), after some simple algebra,

$$\begin{aligned} &4(1-\tau)\Phi(\tau^{1/2})+4(1+\tau)\Phi(\tau^{1/2}+\tau^{-1/2}) \\ &\quad = 6+2(2\pi)^{-1/2}\int_0^\tau (\tau-u)^{-1/2}\exp\left[-\frac{(1+\tau+u)^2}{2(\tau-u)}\right]du \\ &\qquad -4(2\pi)^{-1/2}\tau^{1/2}[e^{-\tau/2}-e^{-(\tau+1)^2/2\tau}] \\ &\quad \leqq 6+4\left(\frac{\tau}{2\pi}\right)^{1/2}. \end{aligned}$$

The last inequality above is a consequence of (5.22). It is not satisfied, however, for all $\tau > 0$; e.g., for the choice $\tau = 1$ the left-hand side is approximately equal to 7.82 and the right-hand side to 7.70. The contradiction implies that the function $s(\tau) = \frac{1}{2} + \tau$ cannot be the free boundary for this problem.

Figure 2 is a plot of the free-boundary function $s(\tau)$ obtained by numerically solving the Bellman equation (2.6) subject to the initial condition $V(x, 0) = x^2$, and

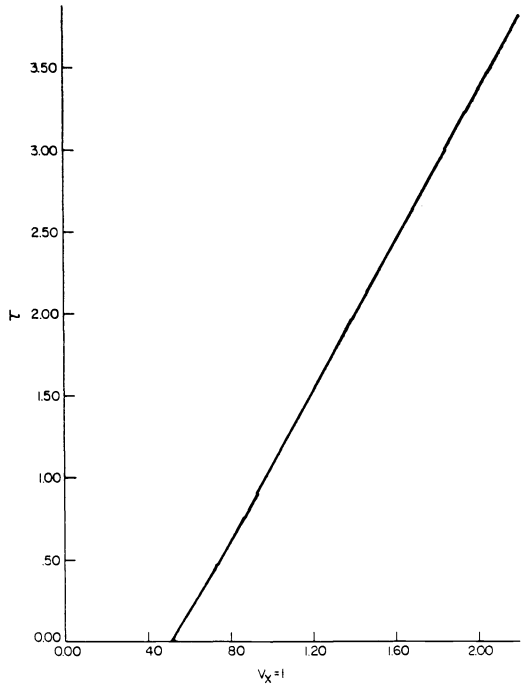


FIG. 2. *The free boundary curve in the special case of a quadratic terminal cost.*

identifying the points where $V_x(x, \tau) = 1$. The MOL1D (Methods of Lines) package for solving partial differential equations was utilized for this purpose (see Hyman [1976]). The following truncation scheme was used. Since by (4.18)

$$2x - 2\tau \leqq V_x(x, \tau) \leqq 2x,$$

it seemed appropriate to truncate the problem at a large number M, impose the (approximate) boundary condition $V_x(M, \tau) \cong 2M - \tau$ and solve the resulting initial-boundary value problem for (2.6) on the strip $0 \leqq x \leqq M$, $0 \leqq \tau \leqq T$. For sufficiently large M, the solution to the above problem would hopefully approximate the one for the Cauchy problem in the region of interest ($x = 0.5$ to $x = 0.5 + T$).

It turned out that $M = 16$ is large enough for this problem. Also, for this choice of M, the solution inside the region of interest is insensitive to changes in the boundary condition imposed at $x = M$; one gets identical plots of the free-boundary curve for boundary conditions $V_x(M, \tau) = 2M - 2\tau$, $2M - \tau$ and $2M$ at $x = M$.

The form of the curve suggests that it would be of interest to examine the asymptotic behavior of the free boundary. A linear behavior as $\tau \to \infty$ clearly suggests itself. On the other hand, it is possible to calculate explicitly the initial slope of the free boundary by using representation (4.39), Ito's theorem, Girsanov's change of measure

and a stopping time argument. It turns out that $s'(0)=\frac{1}{2}$, in striking accordance with Fig. 2.

5.4. Strict monotonicity of the free boundary. One can differentiate with respect to x in the representations (5.8) and (5.15) of the gradient of the value function and get the representations for $V_{xx}^{(\alpha)}(x,\tau)$ below:

$$V_{xx}^{(\alpha)}(x,\tau)=\int_0^\tau G_x[x,\tau;s^\alpha(u),u][w'^\alpha(u)-s'^\alpha(u)]\,du + \int_0^b N(x,\tau;\xi,0)f''(\xi)\,d\xi \quad \text{in } D^*(0), \tag{5.23}$$

$$V_{xx}^{(\alpha)}(x,\tau)=-\int_0^\tau K_x[x-\alpha\tau,\tau;s^\alpha(u)-\alpha u,u][w'^\alpha(u)-s'^\alpha(u)]\,du + \int_b^\infty K(x-\alpha\tau,\tau;\xi,0)f''(\xi)\,d\xi \quad \text{in } D(-\alpha). \tag{5.24}$$

Using the jump relations (5.9) and (5.16), along with the fact that $V_{xx}^{(\alpha)}[s^\alpha(\tau),\tau]=2[w'^\alpha(\tau)-s'^\alpha(\tau)]$, one gets the integral equations

$$w'^\alpha(\tau)-s'^\alpha(\tau)=\int_0^\tau G_x[s^\alpha(\tau),\tau;s^\alpha(u),u][w'^\alpha(u)-s'^\alpha(u)]\,du + \int_0^b N[s^\alpha(\tau),\tau;\xi,0]f''(\xi)\,d\xi, \tag{5.25}$$

$$w'^\alpha(\tau)-s'^\alpha(\tau)=-\int_0^\tau K_x[s^\alpha(\tau)-\alpha\tau,\tau;s^\alpha(u)-\alpha u,u][w'^\alpha(u)-s'^\alpha(u)]\,du + \int_b^\infty K[s^\alpha(\tau)-\alpha\tau,\tau;\xi,0]f''(\xi)\,d\xi. \tag{5.26}$$

It can be shown that the above system of equations is actually equivalent to the free-boundary problem (2.15)–(2.20), but we shall not pursue this line of approach here. Instead, we shall use these equations to prove the following result:

PROPOSITION 5.2. *The free boundary $s(\tau)$ for the original problem ($\alpha=1$) is strictly increasing on $[0,T]$.*

Proof. It has already been shown that the boundary is increasing (Proposition 4.2). In the region $D^*(0)=\{(x,\tau);0<x<s(\tau),0<\tau<T\}$ the function $\nu(x,\tau)=V_{x\tau}(x,\tau)$ satisfies

$$\nu_\tau=\frac{1}{2}\nu_{xx} \qquad \text{in } D^*(0),$$

$$\nu(0,\tau)=0, \qquad 0<\tau<T,$$

$$\nu(x,0)=\frac{1}{2}f'''(x), \qquad 0<x<b,$$

and we already know (Proposition 4.1) that $\nu(x,\tau)\leqq 0$, in $D^*(0)$. Suppose there exist two points $0\leqq\tau'<\tau''\leqq T$ such that $s(\tau)\equiv s(\tau')$, $\nu(x,\tau)=0$, for all $\tau'\leqq\tau\leqq\tau''$. Then on the line segment $x=s(\tau)$, $\tau'\leqq\tau\leqq\tau''$, $\nu(x,\tau)=0$, and by the strong maximum principle, $\nu(x,\tau)=0$ first in the rectangle $\{(x,\tau);0<x<s(\tau),\tau'<\tau<\tau''\}$ and then in the whole of

$D^*(0)$. Therefore, $s(\tau) \equiv s(\tau')$ for all $\tau \in [\tau', \tau'']$ for some $0 < \tau' < \tau'' \leqq T$, implies $s(\tau) = b$ for all $\tau \in [0, T]$.

The latter is clearly impossible if, for some $\bar{x}$, $0 < \bar{x} < b$, $f''(\bar{x}) < 0$, because then also $\nu(\bar{x}, 0) < 0$, a contradiction. It remains to be ruled out also for the case $f'''(x) = 0$, $0 \leqq x \leqq b$, which implies that $f(x)$ is quadratic on $[0, \frac{1}{2}]$ ($b = \frac{1}{2}$ for simplicity, $f(x) = x^2$, $0 \leqq x \leqq \frac{1}{2}$) and continuous with decreasing curvature on $[\frac{1}{2}, \infty)$, $f''(x) \leqq 2$, $\frac{1}{2} \leqq x < \infty$. One gets from (5.25), (5.26) with $(s, \tau) = \frac{1}{2}$

$$(5.27)\quad w'(\tau) = (2\pi)^{-1/2} \int_0^\tau (\tau - u)^{-3/2} \exp\left[\frac{-1}{2(\tau - u)}\right] w'(u)\, du + [2\Phi(\tau^{-1/2}) - 1],$$

$$(5.28)\quad w'(\tau) \leqq -(2\pi)^{-1/2} \int_0^\tau (\tau - u)^{-1/2} \exp\left[-\frac{1}{2}(\tau - u)\right] w'(u)\, du + 2[1 - \Phi(\tau^{1/2})].$$

Equation (5.27) is Volterra of the second kind in $w'(\tau)$ and can be readily solved: $w'(\tau) = 1$, $0 \leqq \tau \leqq T$, for which value the right-hand side of (5.28) becomes

$$-(2\pi)^{-1/2} \int_0^\tau \exp(-u/2) u^{-1/2}\, du + 2[1 - \Phi(\tau^{1/2})] = 3 - 4\Phi(\tau^{1/2}) < 1 = w'(\tau), \qquad 0 < \tau < T,$$

a contradiction to (5.28).

6. A homotopy of compact operators and a convex class of homeomorphisms. The integral equations (5.10) and (5.17) do not lend themselves (at least in an obvious way) to a direct analysis that might establish the existence of a fixed point (s', w') through the contraction mapping principle or Schauder's fixed-point theorem, as is the case with the Stefan problem (see, for example, Friedman [1959]). The main reason for this difficulty is the form of the free-boundary condition; in our problem, the free-boundary function is implicitly defined via the relation $V_x[s(\tau), \tau] = 1$, whereas in problems of the Stefan type the free-boundary condition is of the form $s'(\tau) = -V_x[s(\tau), \tau]$ which not only renders the corresponding integral equation(s) amenable to a straightforward fixed-point analysis, but also provides valuable and explicit information about the smoothness of the free boundary (Schaeffer [1976]).

In our effort to prove smoothness of the free boundary in the problem under consideration, we are thus led to use more sophisticated tools such as topological degree theory. More specifically, we take the continuous functions c, ν on $[0, \sigma]$, define

$$(6.1)\qquad s(\tau) = b + \int_0^\tau c(u)\, du, \qquad w(\tau) = f(b) + \int_0^\tau \nu(u)\, du, \qquad 0 \leqq \tau \leqq \sigma,$$

and consider the space $\mathbf{X}_\sigma = C_{[0,\sigma]} \times C_{[0,\sigma]}$ of pairs of functions (c, ν) normed by

$$(6.2)\qquad \|(c, \nu)\| = \|c\| + \|\nu\| = \sup_{0 \leqq \tau \leqq \sigma} |c(\tau)| + \sup_{0 \leqq \tau \leqq \sigma} |\nu(\tau)|.$$

Under this norm, $\mathbf{X}_\sigma$ is a Banach space. We also consider the linear space Z_σ, subspace of $C_{[0,\sigma]}$, consisting of all functions $f(\tau) \in C_{[0,\sigma]} \cap C^1_{(0,\sigma]}$ satisfying $|f(\tau)| \leqq C\tau^{1/2}$, $0 \leqq \tau \leqq \sigma$ for some positive constant C, $\sup_{0 < \tau \leqq \sigma} \tau^{1/2} |f'(\tau)| < \infty$ and for which $\lim_{\tau \downarrow 0} \tau^{1/2} f'(\tau)$ exists and is finite.

Let this space be normed by

$$(6.3)\qquad \|f\|_{1/2} = |\lim_{\tau \downarrow 0} \tau^{1/2} f'(\tau)| + \sup_{0 < \tau \leqq \sigma} \tau^{-1/2} |f(\tau)| + \sup_{0 < \tau \leqq \sigma} \tau^{1/2} |f'(\tau)|.$$

It is assumed that the interval $[0, \sigma]$ is "sufficiently small," namely, that (6.4) below is satisfied (see (8.1), (10.5)):

$$\sigma \leqq \min\left[\frac{b}{2V}, \frac{1}{2(1+V)^2}, \frac{2}{(1+V)^2}\ln\frac{3bK}{3bK-1}\right], \tag{6.4}$$

where $K = f''(0) = \max_{x\in R} f''(x)$, $f'(b) = 1$ and V is an upper bound on $\|(c, \nu)\|$.

We now consider the integral operators defined for each $\alpha \in [0, 1]$ on X_σ by (5.10) and (5.17), namely,

$$\phi_1(\tau; c, \nu) = k_1(\tau; c) + h_1(\tau; c, \nu), \qquad 0 \leqq \tau \leqq \sigma, \tag{6.5}$$

$$\phi_2^\alpha(\tau; c, \nu) = k_2^\alpha(\tau; c, \nu) + h_2^\alpha(\tau; c), \qquad 0 \leqq \tau \leqq \sigma, \tag{6.6}$$

where

$$k_1(\tau; c) = \int_0^\tau N_x[s(\tau), \tau; s(u), u]\, du, \tag{6.7}$$

$$\begin{aligned} k_2^\alpha(\tau; c, \nu) = &-\int_0^\tau K_x[s(\tau)-\alpha\tau, \tau; s(u)-\alpha u, u]\, du \\ &-2\int_0^\tau K[s(\tau)-\alpha\tau, \tau; s(u)-\alpha u, u][\nu(u)-\alpha]\, du, \end{aligned} \tag{6.8}$$

$$h_1(\tau; c, \nu) = \lambda(\tau; c, \nu) + \gamma(\tau; c), \tag{6.9}$$

where

$$\begin{aligned} \gamma(\tau; c) &= 2\int_0^b G[s(\tau), \tau; \xi, 0] f'(\xi)\, d\xi - 1, \\ \lambda(\tau; c, \nu) &= 2\int_0^\tau G[s(\tau), \tau; s(u), u]\nu(u)\, du, \end{aligned} \tag{6.10}$$

$$h_2^\alpha(\tau; c) = 2\int_b^\infty K[s(\tau)-\alpha\tau, \tau; \xi, 0] f'(\xi)\, d\xi - 1. \tag{6.11}$$

The operators (6.5), (6.6) can be written in a more compact form as

$$\phi^\alpha(c, \nu) = k^\alpha(c, \nu) + h^\alpha(c, \nu), \tag{6.12}$$

where

$$k^\alpha(c, \nu) = [k_1(c), k_2^\alpha(c, \nu)], \qquad h^\alpha(c, \nu) = [h_1(c, \nu), h_2^\alpha(c, \nu)],$$
$$\phi^\alpha(c, \nu) = [\phi_1(c, \nu), \phi_2^\alpha(c, \nu)]$$

denote mappings from the Banach space $\mathbf{X}_\sigma$ into the Banach space $\mathbf{Y}_\sigma = C_{[0,\sigma]}$ normed by (6.2).

It should be pointed out that if (c^α, ν^α) is a solution of the system of integral equations (5.10), (5.17) for any particular value of α, $0 \leqq \alpha \leqq 1$, then $\phi^\alpha(c^\alpha, \nu^\alpha) = 0$, i.e., (c^α, ν^α) is a 0-point of the transformation ϕ^α, and vice-versa. By Proposition 4.7 and its Corollary it is known "a priori" that all possible 0-points of the transformations ϕ^α, for any $0 \leqq \alpha \leqq 1$, lie in the set

$$G = \{(c, \nu) \in \mathbf{X}_\sigma;\ \|(c, \nu)\| < rV\} \quad \text{for some } r > 1, \tag{6.13}$$

where

$$V = V(T) = \frac{2K+L}{k(T)} + \frac{K}{2}. \tag{6.14}$$

In the remainder of this section, we study the properties of the integral operators in (6.12). It is proved (Proposition 6.1) that, for each $0 \leqq \alpha \leqq 1$, $k^{\alpha}(c, \nu)$ is a compact operator from $\mathbf{X}_{\sigma}$ into $\mathbf{Y}_{\sigma}$ and that $h^{\alpha}(c, \nu)$ is a homeomorphism from $\mathbf{X}_{\sigma}$ into

$$\mathbf{Z}_{\sigma} = Z_{\sigma} \times Z_{\sigma}$$

(Proposition 6.3). It is also proved that the family of operators $\{\phi^{\alpha}(c, \nu);\ 0 \leqq \alpha \leqq 1\}$ is jointly continuous in $[(c, \nu), \alpha]$ when viewed as a mapping from $M \times [0, 1]$ into $\mathbf{Y}_{\sigma}$, where M is an arbitrary bounded subset of $\mathbf{X}_{\sigma}$; i.e., the family $\{\phi^{\alpha}(c, \nu);\ 0 \leqq \alpha \leqq 1\}$ is a *homotopy* of continuous operators from $\mathbf{X}_{\sigma}$ into $\mathbf{Y}_{\sigma}$ (Corollary 6.1). Finally, the family of homeomorphisms $\{h^{\alpha}(c, \nu);\ 0 \leqq \alpha \leqq 1\}$ is shown to be convex in the sense of Definition 6.1 (Proposition 6.4).

For operators of the above form (compact plus homeomorphism) a *topological degree* (generalized Leray-Schauder degree) can be defined, for which all the basic properties of the Leray-Schauder degree carry through; see Browder [1976], Cronin [1964]. The degree is invariant under homotopy, equal to ± 1 for homeomorphisms and if G is open, $G \subseteq \mathbf{X}_{\sigma}$, $p \in \mathbf{Y}_{\sigma}$:

$$\deg[\phi, G, p] \neq 0 \Rightarrow \exists q \in G, \quad \text{s.t. } \phi(q) = p. \tag{6.15}$$

Proposition (6.15) is the basic "existence result" in degree theory. We use the above-mentioned properties of topological degree in § 7 (proof of Theorem 7.1). First, it is verified that the operator $\phi^{0}(c, \nu)$ is a homeomorphism (endpoint $\alpha = 0$, uncontrolled case) and therefore $\deg[\phi^{0}, G, 0] = \pm 1$, G being defined in (6.13). Second, this degree knowledge is extended to the other endpoint ($\alpha = 1$, fully controlled case) by virtue of the fact that $\{\phi^{\alpha}(c, \nu);\ 0 \leqq \alpha \leqq 1\}$ is a homotopy indexed by α: $\deg[\phi^{1}, G, 0] = \pm 1$. Finally, the basic existence result (6.15) of degree theory guarantees the existence of a 0-point of the transformation $\phi = \phi^{1}$ in G, thus proving the existence of a continuously differentiable function $s(\tau)$; $0 \leqq \tau \leqq \sigma$. Continuation of the solution into the future is then possible, mainly because of the propagation of uniform convexity (Proposition 4.6).

PROPOSITION 6.1. *For each* $0 \leqq \alpha \leqq 1$, *the operator* $k^{\alpha}(c, \nu)$: $\mathbf{X}_{\sigma} -> \mathbf{Y}_{\sigma}$ *is compact. Furthermore, the family* $\{k^{\alpha}(c, \nu);\ 0 \leqq \alpha \leqq 1\}$ *is a homotopy of compact operators from* $\mathbf{X}_{\sigma}$ *into* $\mathbf{Y}_{\sigma}$.

Proof. By the Ascoli-Arzela theorem, compactness of the operator $k_1(\tau; c)$: $C_{[0,\sigma]} -> C_{[0,\sigma]}$ is equivalent to equicontinuity and uniform boundedness of the set $k_1(M)$, for any bounded set $M \subseteq C_{[0,\sigma]}$.

Suppose V is an upper bound on $\|(c, \nu)\|$, for example, $V(T)$ in (6.14). By virtue of inequality (8.20) we have, for any $c \in C_{[0,\sigma]}$, $\|c\| \leqq V$,

$$|k_1(\tau; c)| \leqq (V + 6b^{-1})\tau^{1/2}, \qquad 0 \leqq \tau \leqq \sigma,$$

which proves uniform boundedness. On the other hand, if $0 \leqq \tau_1 < \tau_2 \leqq \sigma$,

$$k_1(\tau_2; c) - k_1(\tau_1; c) = \int_0^{\tau_1} \{N_x[s(\tau_2), \tau_2; s(u), u] - N_x[s(\tau_1), \tau_1; s(u), u]\}\, du$$

$$+ \int_{\tau_1}^{\tau_2} N_x[s(\tau_2), \tau_2; s(u), u]\, du,$$

and upon using (8.15), (8.20) we obtain for $0<\varepsilon<\frac{1}{2}$

$$|k_1(\tau_2; c)-k_1(\tau_1; c)|\leqq\left[\frac{B_1}{1/2-\varepsilon}+V+6b^{-1}\right](\tau_2-\tau_1)^{\varepsilon},$$

which proves equicontinuity. In a similar manner, we can establish compactness of the operators $k_2^{\alpha}(c, \nu)$, $0\leqq\alpha\leqq 1$. Indeed, assuming $(c, \nu)\in\mathbf{X}_{\sigma}\|(c, \nu)\|\leqq V$ and using (8.19) we get

$$|k_2^{\alpha}(\tau; c, \nu)|\leqq 3(1+V)\tau^{1/2}, \qquad 0\leqq\tau\leqq\sigma,$$

while using (8.11), (8.14) and (8.19) we can show that, for $0\leqq\tau_1<\tau_2\leqq\sigma$, $0<\varepsilon<\frac{1}{2}$;

$$|k_2^{\alpha}(\tau_2; c, \nu)-k_2^{\alpha}(\tau_1; c, \nu)|\leqq\left[\frac{3+4V}{1/2-\varepsilon}+3(1+V)\right](\tau_2-\tau_1)^{\varepsilon},$$

and thus establish uniform boundedness and equicontinuity. Consequently, the whole operator $k^{\alpha}(c, \nu)=[k_1(c), k_2^{\alpha}(c, \nu)]$ is compact. Let us now prove joint continuity of $k^{\alpha}(c, \nu)$ in $((c, \nu), \alpha)\in\mathbf{X}_{\sigma}\times[0, 1]$, uniformly with respect to (c, ν), $\|(c, \nu)\|\leqq V$. It is easily seen from (8.7) that

$$|k_1(\tau, \tilde{c})-k_1(\tau; c)|\leqq 2(1+85b^{-4})\|\tilde{c}-c\|\tau^{1/2}, \qquad 0\leqq\tau\leqq\sigma.$$

On the other hand, we have

$$|k_2^{\alpha}(\tau; \tilde{c}, \nu)-k_2^{\alpha}(\tau; c, \nu)|\leqq 4\|\tilde{c}-c\|\tau^{1/2}, \qquad 0\leqq\tau\leqq\sigma,$$

from (8.3), (8.6), (8.1),

$$|k_2^{\tilde{\alpha}}(\tau; c, \nu)-k_2^{\alpha}(\tau; c, \nu)|\leqq 6|\tilde{\alpha}-\alpha|\tau^{1/2}, \qquad 0\leqq\tau\leqq\sigma,$$

from (8.9), (8.10), (8.1), and finally,

$$|k_2^{\alpha}(\tau; c, \tilde{\nu})-k_2^{\alpha}(\tau; c, \nu)|\leqq 2\|\tilde{\nu}-\nu\|\tau^{1/2}, \qquad 0\leqq\tau\leqq\sigma.$$

Combining the last four inequalities together, one gets

$$\|k^{\tilde{\alpha}}(\tilde{c}, \tilde{\nu})-k^{\alpha}(c, \nu)\|\leqq 2(7+85b^{-4})\sigma^{1/2}[\|(\tilde{c}, \tilde{\nu})-(c, \nu)\|+|\tilde{\alpha}-\alpha|],$$

which proves joint continuity.

PROPOSITION 6.2. *The family* $\{h^{\alpha}(c, \nu);\ 0\leqq\alpha\leqq 1\}$ *is a homotopy of continuous operators from* $\mathbf{X}_{\sigma}$ *into* $\mathbf{Y}_{\sigma}$.

Proof. We have to show joint continuity of $h^{\alpha}(c, \nu)=[h_1(c, \nu), h_2^{\alpha}(c)]$ in $[(c, \nu), \alpha]\in\mathbf{X}_{\sigma}\times[0, 1]$, uniformly with respect to $(c, \nu)\in\mathbf{X}_{\sigma}$, $\|(c, \nu)\|\leqq V$. From (9.3), we get

$$|\gamma(\tau; \tilde{c})-\gamma(\tau; c)|\leqq 2\|\tilde{c}-c\|\tau^{1/2}, \qquad 0\leqq\tau\leqq\sigma,$$

and, using (8.5), (8.1),

$$|\lambda(\tau; \tilde{c}, \nu)-\lambda(\tau; c, \nu)|\leq 2(1+14b^{-4})\|\tilde{c}-c\|\tau^{1/2}, \qquad 0\leqq\tau\leqq\sigma.$$

Also,

$$|\lambda(\tau; c, \tilde{\nu})-\lambda(\tau; c, \nu)|\leqq 2\|\tilde{\nu}-\nu\|\tau^{1/2}, \qquad 0\leqq\tau\leqq\sigma.$$

Therefore, by (6.9) we obtain

$$|h_1(\tau; \tilde{c}, \tilde{\nu})-h_1(\tau; c, \nu)|\leqq 2(3+14b^{-4})\|(\tilde{c}, \tilde{\nu})-(c, \nu)\|\tau^{1/2}, \qquad 0\leqq\tau\leqq\sigma.$$

On the other hand, one gets from (10.4), (10.4′)

$$|h_2^{\tilde{\alpha}}(\tau; \tilde{c})-h_2^{\alpha}(\tau; c)|\leqq K(2+3b)(\|\tilde{c}-c\|+|\tilde{\alpha}-\alpha|)\tau^{1/2}, \qquad 0\leqq\tau\leqq\sigma.$$

Therefore,

$$\|h^{\tilde{\alpha}}(\tilde{c}, \tilde{\nu}) - h^{\alpha}(c, \nu)\| \leqq [K(2+3b)+2(3+14b^{-4})]\sigma^{1/2}[\|(\tilde{c}, \tilde{\nu})-(c, \nu)\|+|\tilde{\alpha}-\alpha|],$$

which proves joint continuity.

COROLLARY 6.1. *The family* $\{\phi^{\alpha}(c, \nu);\ 0 \leqq \alpha \leqq 1\}$ *is a homotopy of continuous operators from* $\mathbf{X}_{\sigma}$ *into* $\mathbf{Y}_{\sigma}$.

PROPOSITION 6.3. *For each* α, $0 \leqq \alpha \leqq 1$, $h^{\alpha}(c, \nu)$ *is a homeomorphism from* $\mathbf{X}_{\sigma}$ *to* $\mathbf{Z}_{\sigma}$.

Proof. First, we establish the existence of the inverse operator of $h^{\alpha}(c)$. Suppose the function $h_2^{\alpha}(\tau)$, $0 \leqq \tau \leqq \sigma$, from Z_{σ} is given; we are seeking a continuously differentiable function $s(\tau)$, $0 \leqq \tau \leqq \sigma$ such that

$$2\int_b^{\infty} K[s(\tau)-\alpha\tau, \tau; \xi, 0]f'(\xi)\,d\xi - 1 = h_2^{\alpha}(\tau), \qquad 0 \leqq \tau \leqq \sigma.$$

We put $y(\tau) = s(\tau) - \alpha\tau$, and consider the function

$$F(y, \tau) = 2\int_b^{\infty} K(y, \tau; \xi, 0)f'(\xi)\,d\xi - 1 - h_2^{\alpha}(\tau), \qquad (y, \tau) \in \mathbf{R}\times[0, \sigma].$$

The problem is to determine a continuously differentiable function $y(\tau)$, $0 \leqq \tau \leqq \sigma$, satisfying $F[y(\tau), \tau] = 0$, $0 \leqq \tau \leqq \sigma$. $F(y, \tau)$ is continuous on its domain of definition and satisfies $F(b, 0) = 0$, so that $y(0) = b$. Besides,

$$\begin{aligned} F_y(y, \tau) &= 2\int_b^{\infty} K_x(y, \tau; \xi, 0)f'(\xi)\,d\xi \\ &= 2K(y, \tau; b, 0) + 2\int_b^{\infty} K(y, \tau; \xi, 0)f''(\xi)\,d\xi > 0 \quad \text{on } \mathbf{R}\times[0, \sigma], \end{aligned}$$

since it is assumed that $f''(x) \geqq k > 0$. On the other hand, $F_{\tau}(y, \tau)$ also exists, is continuous on $(0, \sigma]$ and $\tau^{1/2}F_{\tau}(y, \tau)$ is continuous on $[0, \sigma]$. By the implicit function theorem, there exists a continuously differentiable function $y(\tau)$, $0 \leqq \tau \leqq \sigma$, satisfying $F[y(\tau), \tau] = 0$, $0 \leqq \tau \leqq \sigma$. It suffices then to take $s(\tau) = y(\tau) + \alpha\tau$, $s'(\tau) = c(\tau) = y'(\tau) + \alpha$, $0 \leqq \tau \leqq \sigma$.

Assume that $h^{\alpha}(\tau) = [h_1(\tau), h_2^{\alpha}(\tau)]$ are given from $\mathbf{Z}_{\sigma}$. Once $c(\tau) \in C_{[0,\sigma]}$ has been determined from $h_2^{\alpha}(\tau)$, $\gamma(\tau; c)$, where $0 \leqq \tau \leqq \sigma$, is also determined by (6.10) and belongs to the space Z_{σ} (Corollary 9.1). Then also, $\lambda(\tau) = h_1(\tau) - \gamma(\tau; c)$ belongs to Z_{σ}, since both $h_1(\tau)$ (by assumption) and $\gamma(\tau; c)$ do. For such a function $\lambda(\tau)$, the integral equation (9.9)

$$2\int_0^{\tau} G[s(\tau), \tau; s(u), u]\nu(u)\,du = \lambda(\tau), \qquad 0 \leqq \tau \leqq \sigma,$$

has a unique solution $\nu(\tau) \in C_{[0,\sigma]}$ (see Lemma 9.5). Therefore, starting with $h^{\alpha}(\tau) = [h_1(\tau), h_2^{\alpha}(\tau)] \in Z_{\sigma}$, we get a pair $(c, \nu) \in \mathbf{X}_{\sigma}$ such that

$$\begin{aligned} h_1(\tau) &= 2\int_0^{\tau} G[s(\tau), \tau; s(u), u]\nu(u)\,du \\ &\quad + 2\int_0^{b} G[s(\tau), \tau; \xi, 0]f'(\xi)\,d\xi - 1, \qquad 0 \leqq \tau \leqq \sigma, \end{aligned} \tag{6.16}$$

$$h_2^{\alpha}(\tau) = 2\int_b^{\infty} K[s(\tau)-\alpha\tau, \tau; \xi, 0]f'(\xi)\,d\xi - 1, \qquad 0 \leqq \tau \leqq \sigma, \tag{6.17}$$

be satisfied, with $s(\tau) = b + \int_0^\tau c(u)\,du$, $0 \leqq \tau \leqq \sigma$. Thus, the inverse operator $(h^\alpha)^{-1}$ exists.

It remains to check that $(h^\alpha)^{-1}$ is also continuous. Suppose we are given two vectors $h^\alpha = (h_1, h_2^\alpha)$, $\tilde{h}^\alpha = (\tilde{h}_1, \tilde{h}_2^\alpha)$ in $\mathbf{Z}_\sigma$. It is clear from the above discussion that there will exist two vectors (c, ν), $(\tilde{c}, \tilde{\nu})$, respectively, such that (6.16) and (6.17) be satisfied. Relation (10.6) of Lemma 10.2 implies that

$$\|\tilde{c} - c\| \leqq 6\|\tilde{h}_2^\alpha - h_2^\alpha\|_{1/2},$$

while (9.8) of Corollary 9.1 gives

$$\|\gamma(\tilde{c}) - \gamma(c)\|_{1/2} \leqq B_4\|\tilde{c} - c\|_{1/2} \leqq 6B_4\|\tilde{h}_2^\alpha - h_2^\alpha\|_{1/2}.$$

Since $\lambda = h_1 - \gamma$, we have

$$\|\tilde{\lambda} - \lambda\|_{1/2} \leqq \|\tilde{h}_1 - h_1\|_{1/2} + 6B_4\|\tilde{h}_2^\alpha - h_2^\alpha\|_{1/2}.$$

Finally, because of (9.10) of Lemma 9.5,

$$\|\tilde{\nu} - \nu\| \leqq B_5(\|\tilde{c} - c\| + \|\tilde{\lambda} - \lambda\|_{1/2}) \leqq B_7(\|\tilde{h}_1 - h_1\|_{1/2} + \|\tilde{h}_2^\alpha - h_2^\alpha\|_{1/2}),$$

where $B_7 = B_5[1 + 6(1 + B_4)]$. Therefore,

$$\|(\tilde{c}, \tilde{\nu}) - (c, \nu)\| \leqq (6 + B_7)\|(\tilde{h}_1, \tilde{h}_2^\alpha) - (h_1, h_2^\alpha)\|_{1/2};$$

i.e., the inverse operator is continuous.

DEFINITION 6.1. Let $\mathbf{X}$, $\mathbf{Y}$ be Banach spaces and $\mathbf{H}$ a class of homeomorphisms from $\mathbf{X}$ to $\mathbf{Y}$. $\mathbf{H}$ is said to be *convex*, if for each open subset G of $\mathbf{X}$ and any pair of elements $h_0, h_1 \in \mathbf{H}_G$, the restriction of the class $\mathbf{H}$ on G, the mapping h_λ of $\bar{G}$ into $\mathbf{Y}$ defined by $h_\lambda(x) = (1-\lambda)h_0(x) + \lambda h_1(x)$, $x \in \bar{G}$, $\lambda \in [0, 1]$, also belongs to $\mathbf{H}_G$.

PROPOSITION 6.4. *The family of homeomorphisms* $\{h^\alpha(c, \nu);\ 0 \leqq \alpha \leqq 1\}$ *from* $\mathbf{X}_\sigma$ *into* $\mathbf{Z}_\sigma$ *is convex in the sense of Definition* 6.1.

Proof. It is sufficient to show that $\{h_2^\alpha(c);\ 0 \leqq \alpha \leqq 1\}$ is a convex family of homeomorphisms from $C_{[0,\sigma]}$ into Z_σ. Indeed, consider $\alpha_1, \alpha_2 \in [0, 1]$ and $0 \leqq \lambda \leqq 1$. We have to determine a continuous function $c(\,\cdot\,)$ on $[0, \sigma]$, such that $s(\tau) = b + \int_0^\tau c(u)\,du$, $0 \leqq \tau \leqq \sigma$, satisfies

$$2\lambda \int_b^\infty K[s(\tau) - \alpha_1\tau, \tau; \xi, 0]f'(\xi)\,d\xi + 2(1-\lambda)\int_b^\infty K[s(\tau) - \alpha_2\tau, \tau; \xi, 0]f'(\xi)\,d\xi - 1$$
$$= \lambda h_2^{\alpha_1}(\tau) + (1-\lambda)h_2^{\alpha_2}(\tau) \triangleq h_\lambda(\tau) \in Z_\sigma.$$

Consider the function

$$F(s, \tau) = 2\lambda \int_b^\infty K(s - \alpha_1\tau, \tau; \xi, 0)f'(\xi)\,d\xi$$
$$+ 2(1-\lambda)\int_b^\infty K(s - \alpha_2\tau, \tau; \xi, 0)f'(\xi)\,d\xi - 1 - h_\lambda(\tau) \quad \text{on } \mathbf{R} \times [0, \sigma].$$

The condition above can be formulated as $F[s(\tau), \tau] = 0$, $0 \leqq \tau \leqq \sigma$. $F(s, \tau)$ is continuous on its domain of definition and satisfies $F(b, 0) = 2\lambda \cdot \frac{1}{2} + 2(1-\lambda) \cdot \frac{1}{2} - 1 = 0$, so $s(0) = b$. Observe that

$$F_s(s, \tau) = 2\lambda \cdot K(s - \alpha_1\tau, \tau; b, 0) + 2(1-\lambda) \cdot K(s - \alpha_2\tau, \tau; b, 0)$$
$$+ 2\lambda \int_b^\infty K(s - \alpha_1\tau, \tau; \xi, 0)f''(\xi)\,d\xi + 2(1-\lambda) \cdot \int_b^\infty K(s - \alpha_2\tau, \tau; \xi, 0)f''(\xi)\,d\xi$$

is positive on $\mathbf{R}\times[0,\sigma]$, since $f''(x)\geqq k>0$. Besides, $F_\tau(s,\tau)$ also exists, is continuous on $(0,\sigma]$ and $\tau^{1/2}F_\tau(s,\tau)$ is continuous on $[0,\sigma]$. By the implicit function theorem, there exists a continuously differentiable function $s(\tau)$, $0\leqq\tau\leqq\sigma$, satisfying $F[s(\tau),\tau]=0$, $0\leqq\tau\leqq\sigma$. Therefore, the inverse operator h_λ^{-1} exists. Its continuity is proved as in Proposition 6.3.

7. A method for studying the integral equations by homotopy and topological degree. In this section, we substantiate the topological degree method for proving smoothness of the free-boundary function $s(\tau)$, $0\leqq\tau\leqq T$ which was briefly outlined at the beginning of § 6. Existence of a continuous solution pair (c',w') to the integral equations (5.10), (5.17) for $\alpha=1$ is first established for small times (Theorem 7.1) and then extended into the future (proof of Theorem 2.2).

LEMMA 7.1. *In the case* $\alpha=0$ (*absence of control*) *the integral operator defined in* (6.6), *with* $s(\tau)=b+\int_0^\tau c(u)\,du$, $0\leqq\tau\leqq T$,

$$\phi_2^0(\tau;c,\nu)=-\int_0^\tau K_x[s(\tau),\tau;s(u),u]\,du-2\int_0^\tau K[s(\tau),\tau;s(u),u]\nu(u)\,du$$
$$+2\int_b^\infty K[s(\tau),\tau;\xi,0]f'(\xi)\,d\xi-1, \tag{7.1}$$

can equivalently be written as

$$\phi_2^0(\tau;c,\nu)=-\int_0^\tau N_x[s(\tau),\tau;s(u),u]\,du-2\int_0^\tau G[s(\tau),\tau;s(u),u]\nu(u)\,du$$
$$+2\int_b^\infty G[s(\tau),\tau;\xi,0]f'(\xi)\,d\xi-1. \tag{7.2}$$

Proof. Given the pair (c,ν) in $C_{[0,T]}\times C_{[0,T]}$ we construct the functions $s(\tau)=b+\int_0^\tau c(u)\,du$, $w(\tau)=f(b)+\int_0^\tau\nu(u)\,du$, $0\leqq\tau\leqq T$, and consider the solution of the heat equation: $V_2=1/2V_{11}$ in the domain $\{(x,\tau);\ x>s(\tau),\ 0<\tau<T\}$ subject to the initial and boundary conditions $V(x,0)=f(x)$, $x\geqq b$ and $V[s(\tau),\tau]=w(\tau)$, $0\leqq\tau\leqq T$. $V(x,\tau)$ is supposed to satisfy a polynomial growth condition in x. If we integrate Green's identity $(\partial/\partial\xi)(KV_1-K_\xi V)-(\partial/\partial\xi)(2KV)=0$ in the domain $\{(\xi,u);\ \varepsilon<u<\tau-\varepsilon,\ s(u)<\xi<M\}$, where $M>x>s(\tau)$, and then let $\varepsilon\downarrow0$, $M\uparrow\infty$, we get the following representation for $V(x,\tau)$, in accordance with (5.14):

$$2V(x,\tau)=-\int_0^\tau K[x,\tau;s(u),u]\{V_1[s(u),u]+2w(u)s'(u)\}\,du$$
$$+\int_0^\tau K_\xi[x,\tau;s(u),u]w(u)\,du+2\int_b^\infty K(x,\tau;\xi,0)f(\xi)\,d\xi,$$

and differentiation with respect to x gives, in accordance with (5.15),

$$2V_1(x,\tau)=-\int_0^\tau K_x[x,\tau;s(u),u]V_1[s(u),u]\,du-2\int_0^\tau K[x,\tau;s(u),u]w'(u)\,du$$
$$+2\int_b^\infty K(x,\tau;\xi,0)f'(\xi)\,d\xi.$$

Now, letting $x \downarrow s(\tau)$ we get, by the jump relation (5.16),

$$
\begin{aligned}
V_1[s(\tau), \tau] = &- \int_0^\tau K_x[s(\tau), \tau; s(u), u] V_1[s(u), u]\, du \\
&- 2 \int_0^\tau K[s(\tau), \tau; s(u), u] \nu(u)\, du \\
&+ 2 \int_b^\infty K[s(\tau), \tau; \xi, 0] f'(\xi)\, d\xi.
\end{aligned} \tag{7.3}
$$

Equivalently, we can integrate Green's identity $(\partial/\partial\xi)(NV_1 - N_\xi V) - (\partial/\partial u)(2NV) = 0$, where N is Neumann's function (5.3), and finally come up with

$$
\begin{aligned}
V_1[s(\tau), \tau] = &- \int_0^\tau N_x[s(\tau), \tau; s(u), u] V_1[s(u), u]\, du \\
&- 2 \int_0^\tau G[s(\tau), \tau; s(u), u] \nu(u)\, du \\
&+ 2 \int_b^\infty G[s(\tau), \tau; \xi, 0] f'(\xi)\, d\xi.
\end{aligned} \tag{7.4}
$$

The right-hand sides of (7.3), (7.4) are therefore equal for any possible assignment of the value of the gradient $V_1[s(u), u]$ along the curve $\xi = s(u)$, in particular for $V_1[s(u), u] = 1$.

THEOREM 7.1. Existence of a continuously differentiable free-boundary function, for small times. *For a terminal cost function $f(x)$ satisfying assumptions* A.1–A.3 *of* § 2 *and initial step σ satisfying condition* (6.4), *there exists a solution (c, ν) in $\mathbf{X}_\sigma = C_{[0,\sigma]} \times C_{[0,\sigma]}$ to the integral equations* (5.10), (5.17) *with $s(\tau) = b + \int_0^\tau c(u)\, du$, $w(\tau) = f(b) + \int_0^\tau \nu(u)\, du$, $0 \leqq \tau \leqq \sigma$, for the fully controlled case $\alpha = 1$, and hence a solution to the free-boundary problem* (2.15)–(2.20) *on $\mathbf{R}^+ \times [0, \sigma]$ with continuously differentiable $s(\tau)$, $0 \leqq \tau \leqq \sigma$.*

Proof. We have written the integral equations (5.10), (5.17) in the operator form

$$\phi^\alpha(c, \nu) \triangleq k^\alpha(c, \nu) + h^\alpha(c, \nu) = 0, \tag{6.12}$$

and reduced the problem of solving them to that of finding the 0-points $(c^\alpha, \nu^\alpha) \in \mathbf{X}_\sigma$ of the transformations ϕ^α, $0 \leqq \alpha \leqq 1$.

It has been shown that the family of mappings $\{k^\alpha(c, \nu);\ 0 \leqq \alpha \leqq 1\}$ is a family of compact operators from $\mathbf{X}_\sigma$ to $\mathbf{Y}_\sigma$, and that the family of homeomorphisms $\{h^\alpha(c, \nu);\ 0 \leqq \alpha \leqq 1\}$ from $\mathbf{X}_\sigma$ to the subspace $\mathbf{Z}_\sigma$ of $\mathbf{Y}_\sigma$ is convex in the sense of Definition 6.1 (see Propositions 6.1, 6.3, 6.4).

For operators of this form, a topological degree can be defined (Browder [1976]) which generalizes the notion of the Leray-Schauder degree (Cronin [1964]) and inherits all its basic properties. Let us denote the degree of the 0-point 0 in the Banach space $\mathbf{Y}_\sigma$ with respect to the mapping ϕ^α and the open set $G \in X_\sigma$ by

$$\deg[\phi^\alpha, G, 0]. \tag{7.5}$$

Because of the fact that family $\{\phi^\alpha(c, \nu);\ 0 \leqq \alpha \leqq 1\}$ is a homotopy of continuous transformations from $\mathbf{X}_\sigma$ into $\mathbf{Y}_\sigma$ (Corollary 6.1), the degree (7.5) is invariant under the homotopy, i.e., independent of α, $0 \leqq \alpha \leqq 1$, if we select the set G in such a way that, for any $\alpha \in [0, 1]$, all possible 0-points (c^α, ν^α) of ϕ^α, i.e., points for which $\phi^\alpha(c, \nu) = 0$, lie within G.

The selection of such a set G is made possible by the "a priori" bounds of § 4, established in Proposition 4.7 and its Corollary. It is proved there that if the boundary function s^α has a continuous derivative c^α and $\nu^\alpha = w'^\alpha$, $w^\alpha(\tau) = V^{(\alpha)}[s^\alpha(\tau), \tau]$, i.e., if (c^α, ν^α) is a possible 0-point of the transformation ϕ^α in view of Proposition 2.1, then

$$\|(c^\alpha, \nu^\alpha)\| = \|c^\alpha\| + \|\nu^\alpha\| = \sup_{0 \leqq \tau \leqq \sigma} |c^\alpha(\tau)| + \sup_{0 \leqq \tau \leqq \alpha} |\nu^\alpha(\tau)| \leqq V, \tag{7.6}$$

with

$$V = V(T) = \frac{2K + L}{k(T)} + \frac{K}{2}, \tag{6.14}$$

for any $\alpha \in [0, 1]$. This a priori bound enables us to take as set G, with respect to which the topological degree will be considered, the set

$$G = \{(c, \nu) \in \mathbf{X}_\sigma;\ \|(c, \nu)\| < rV\} \quad \text{for some } r > 1. \tag{6.13}$$

Now, since "no 0-points of ϕ^α, $0 \leqq \alpha \leqq 1$, can escape through the boundary ∂G of G," the degree $\deg[\phi^\alpha, G, 0]$ is independent of α, $0 \leqq \alpha \leqq 1$ (Cronin [1964]).

Let us calculate the degree at the endpoint $\alpha = 0$. In this case, the integral operators (6.5), (6.6) take the form (see Lemma 7.1)

$$\begin{aligned}\phi_1^0(\tau; c, \nu) = &\int_0^\tau N_x[s(\tau), \tau; s(u), u]\, du + 2\int_0^\tau G[s(\tau), \tau; s(u), u]\nu(u)\, du \\ &+ 2\int_0^b G[s(\tau), \tau; \xi, 0]f'(\xi)\, d\xi - 1.\end{aligned} \tag{6.5}'$$

$$\begin{aligned}\phi_2^{0\prime}(\tau; c, \nu) = &-\int_0^\tau N_x[s(\tau), \tau; s(u), u]\, du - 2\int_0^\tau G[s(\tau), \tau; s(u), u]\nu(u)\, du \\ &+ 2\int_b^\infty G[s(\tau), \tau; \xi, 0]f'(\xi)\, d\xi - 1,\end{aligned} \tag{6.6}'$$

or, equivalently, (6.5) together with

$$\phi_2^0(\tau; c) \triangleq \phi_1^0(\tau; c, \nu) + \phi_2^{0\prime}(\tau; c, \nu) = 2\left\{\int_{-\infty}^\infty K[s(\tau), \tau; \xi, 0]f'(\xi)\, d\xi - 1\right\}. \tag{7.7}$$

The function $s^0(\tau)$ is now obtained through $\phi_2^0(\tau; c^0) = 0$; $0 \leqq \tau \leqq \sigma$. This "decoupling" of the integral equations is not surprising, because in the uncontrolled case we are essentially solving the heat equation with initial condition $V^{(0)}(x, 0) = f(x)$; the solution is $V^{(0)}(x, \tau) = \int_{-\infty}^\infty K(x, \tau; \xi, 0)f(\xi)\, d\xi$, its gradient $V_1^{(0)}(x, \tau) = \int_{-\infty}^\infty K(x, \tau; \xi, 0)f'(\xi)\, d\xi$ and the function $s^0(\tau)$ is characterized by $\int_{-\infty}^\infty K[s^0(\tau), \tau; \xi, 0]f'(\xi)\, d\xi = 1$, $0 \leqq \tau \leqq \sigma$.

We exploit the special structure of $\phi^0 = (\phi_1^0, \phi_2^{0\prime})$; it can be seen, by a repetition of the steps in the proof of Proposition 6.3, that the whole operator ϕ^0 is a homeomorphism from $\mathbf{X}_\sigma$ into $\mathbf{Z}_\sigma \subseteq \mathbf{Y}_\sigma$. Thus, by a familiar property of the Leray-Schauder degree which also holds in this generalized context,

$$\deg[\phi^0, G, 0] = \pm 1, \tag{7.8}$$

and by invariance of degree under homotopy,

$$\deg[\phi^1, G, 0] = \deg[\phi^0, G, 0] = \pm 1. \tag{7.9}$$

The fundamental existence result of degree theory [Proposition (6.15)] guarantees the existence of a point $(c, \nu) \in G$ such that

$$\phi(c, \nu) = \phi^1(c, \nu) = 0. \tag{7.10}$$

In view of Proposition 5.1 and with the identification $(s', w') = (c, \nu)$, (7.10) means that there exists a solution to the free-boundary problem (2.15)–(2.20) in the sense of Definition 2.1 on $\mathbf{R}^+ \times [0, \sigma]$, with $s(\tau)$ continuously differentiable on $[0, \sigma]$.

We now extend the result of Theorem 7.1 to any finite time horizon $T > 0$. Use is made of the results in § 4, such as the monotonicity of the free boundary and the propagation of uniform convexity.

Proof of Theorem 2.2. It is proved in Theorem 7.1 that the free-boundary problem possesses a solution on $\mathbf{R}^+ \times [0, \sigma]$ in the sense of Definition 2.1 with $s(\tau)$ continuously differentiable on $[0, \sigma]$, provided σ is sufficiently small, namely,

$$\sigma \leqq \min\left[\frac{b}{2V}, \frac{1}{2(1+V)^2}, \frac{2}{(1+V)^2} \ln \frac{3bK}{3bK-1}\right], \tag{6.4}$$

where V is the a priori bound (6.14).

We now apply the proof of Theorem 7.1 step by step. In order to show that an extension of the solution up to any finite time horizon $T > 0$ is possible, we have to establish the following fact: If the free-boundary problem has a solution on $\mathbf{R}^+ \times [0, t]$ with $s(\tau)$ continuously differentiable on $[0, t]$, for some $0 < t < T$, there exists an $\eta > 0$, independent of t, such that the free-boundary problem has a solution on $\mathbf{R}^+ \times [0, t+\eta]$ with $s(\tau)$ continuously differentiable on $[0, t+\eta]$.

Indeed, the method used in § 5 can be applied again to show that the free-boundary problem on $\mathbf{R}^+ \times [t, T]$ is equivalent to the system of integral equations

$$\begin{aligned} 1 = &\int_t^\tau N_x[s(\tau), \tau; s(u), u]\, du + 2\int_t^\tau G[s(\tau), \tau; s(u), u]\nu(u)\, du \\ &+ 2\int_0^{s(t)} G[s(\tau), \tau; \xi, t] V_x(\xi, t)\, d\xi, \qquad t \leqq \tau \leqq T, \end{aligned} \tag{7.11}$$

$$\begin{aligned} 1 = &-\int_t^\tau K_x[s(\tau)-\tau, \tau; s(u)-u, u]\, du \\ &-2\int_t^\tau K[s(\tau)-\tau, \tau; s(u)-u, u][\nu(u)-1]\, du \\ &+2\int_{s(t)}^\infty K[s(\tau)-\tau, \tau; \xi, t] V_x(\xi, t)\, d\xi, \end{aligned} \tag{7.12}$$

in accordance with (5.10), (5.17). The function $V(x, t)$ inherits all the basic properties of $f(x)$ which enabled us to prove the existence of a solution to the integral equations (5.10), (5.17) on $[0, \sigma]$; $V(x, t)$ is even, twice continuously differentiable, with a Lipschitz continuous second derivative satisfying the uniform convexity condition, $V_{xx}(x, t) \geqq k(T) > 0$ (Proposition 4.6). Thus, if we proceed with the method of Theorem 7.1 but start from $\tau = t$ upward (instead of $\tau = 0$), then by analogy we are able to solve the free-boundary problem with continuously differentiable $s(\tau)$ on $[t, t+\eta]$, where

$$\eta \leqq \min\left[\frac{s(t)}{2V}, \frac{1}{2(1+V)^2}, \frac{2}{(1+V)^2}\ln\frac{3Ks(t)}{3Ks(t)-1}\right]. \tag{7.13}$$

In view of the results of § 4 (inequalities (4.25)) it suffices to take

$$\eta = \min\left[\frac{b}{2V}, \frac{1}{2(1+V)^2}, \frac{2}{(1+V)^2}\ln\frac{3Km\left(1+\left(K+\frac{L}{2}\right)T\right)}{3Km\left(1+\left(K+\frac{L}{2}\right)T\right)-1}\right],$$

independent of t. Therefore the process of extending the solution into the future can be carried out.

Uniqueness follows from Theorem 2.1 in view of Proposition 2.1.

8. Appendix A. In this section we collect together some basic continuity properties of the kernels encountered in the integral equations (5.10) and (5.17). We make extensive use of these properties in studying the nature of the integral operators in § 6. It should be noted that they are valid only for "small times"; more specifically, it is assumed throughout this section that $0 \leqq \tau \leqq \sigma$, where

$$\sigma \leqq \min\left[\frac{b}{2V}, \frac{1}{2(1+V)^2}\right]. \tag{8.1}$$

V is an upper bound on $\|c\| = \sup|c(\tau)|$ on $0 \leqq \tau \leqq \sigma$, $s(\tau) = b + \int_0^\tau c(u)\,du$. We use tacitly the fact that $\sigma < \frac{1}{2}$, which follows from $2(1+V)^2\sigma \leqq 1$. On the other hand, it is immediate from the inequality $2V\sigma \leqq b$ that

$$\frac{b}{2} \leqq s(\tau) \leqq \frac{3b}{2}, \qquad 0 \leqq \tau \leqq \sigma. \tag{8.2}$$

LEMMA 8.1. *For any $c, \tilde{c} \in C_{[0,\sigma]}$ with $\|c\|, \|\tilde{c}\| \leqq V$, any $0 \leqq \alpha \leqq 1, 0 \leqq u \leqq \tau \leqq \sigma$ and $s(\tau) = b + \int_0^\tau c(u)\,du$, $\tilde{s}(\tau) = b + \int_0^\tau \tilde{c}(u)\,du$, $0 \leqq \tau \leqq \sigma$, the following estimates hold:*

$$\begin{aligned}|K[\tilde{s}(\tau)-\alpha\tau, \tau; \tilde{s}(u)-\alpha u, u] - K[s(\tau)-\alpha\tau, \tau; s(u)-\alpha u, u]| \\ \leqq 2(1+V)\|\tilde{c}-c\|(\tau-u)^{1/2},\end{aligned} \tag{8.3}$$

$$|K[\tilde{s}(\tau), \tau; -\tilde{s}(u), u] - K[s(\tau), \tau; -s(u), u]| \leqq 28b^{-5}\|\tilde{c}-c\|(\tau-u)^{3/2}, \tag{8.4}$$

$$|G[\tilde{s}(\tau), \tau; \tilde{s}(u), u] - G[s(\tau), \tau; s(u), u]| \leqq 2[1+V+14b^{-5}]\|\tilde{c}-c\|(\tau-u)^{1/2}, \tag{8.5}$$

$$\begin{aligned}|K_x[\tilde{s}(\tau)-\alpha\tau, \tau; \tilde{s}(u)-\alpha u, u] - K_x[s(\tau)-\alpha\tau, \tau; s(u)-\alpha u, u]| \\ \leqq \|\tilde{c}-c\|(\tau-u)^{-1/2},\end{aligned} \tag{8.6}$$

$$|N_x[\tilde{s}(\tau), \tau; \tilde{s}(u), u] - N_x[s(\tau), \tau; s(u), u]| \leqq (1+85b^{-4})\|\tilde{c}-c\|(\tau-u)^{-1/2}. \tag{8.7}$$

Proof. The difference of the two kernels on the left-hand side of (8.3) can be written as

$$K[\tilde{s}(\tau)-\alpha\tau,\tau;\tilde{s}(u)-\alpha u,u] \cdot\left[1-\exp\left\{\frac{[\tilde{s}(\tau)-\tilde{s}(u)-\alpha(\tau-u)]^2-[s(\tau)-s(u)-\alpha(\tau-u)]^2}{2(\tau-u)}\right\}\right]. \tag{8.8}$$

The difference of the squares in (8.8) is bounded in absolute value by $|[\tilde{s}(\tau)-s(\tau)]-[\tilde{s}(u)-s(u)]| \cdot |[\tilde{s}(\tau)-\alpha\tau]+[s(\tau)-\alpha\tau]-[\tilde{s}(u)-\alpha u]-[s(u)-\alpha u]| \leqq 2(1+V)\|\tilde{c}-c\|(\tau-u)^2$. Therefore, if x denotes the expression in the second braces of (8.8), we have $|x| \leqq (1+V)\|\tilde{c}-c\|(\tau-u) \leqq 2V(1+V)\sigma < 2(1+V)^2\sigma \leqq 1$, by assumption (8.1). Note at this point that $|x| \leqq t < 1$ implies $|1-e^x| \leqq et$. Consequently, the expression in (8.8) is bounded above in absolute value by

$$(2\pi)^{-1/2}(\tau-u)^{-1/2}|1-e^x| \leqq (2\pi)^{-1/2} \cdot e(1+V)\|\tilde{c}-c\|(\tau-u)^{1/2} < 2(1+V)\|\tilde{c}-c\|(\tau-u)^{1/2}.$$

To prove (8.4), the difference of the kernels is written as $[\tilde{s}(\tau)-s(\tau)]K_x[s^*(\tau),\tau;-\tilde{s}(u),u]-[\tilde{s}(u)-s(u)] \cdot K_x[s(\tau),\tau;-s^*(u),u]$, by virtue of the mean value theorem, where $s^*(\tau)$ and $s^*(u)$ are numbers between $s(\tau)$, $\tilde{s}(\tau)$ and $s(u)$, $\tilde{s}(u)$ respectively. Note that because of (8.2)

$$|K_x[s^*(\tau),\tau;-\tilde{s}(u),u]| \leqq (2\pi)^{-1/2}(\tau-u)^{-1/2} \cdot 3b \exp\left[\frac{-b^2}{2(\tau-u)}\right] \leqq 14b^{-5}(\tau-u)^{3/2},$$

where use has been made of the fact that $x^3e^{-x} \leqq 1.4$, $x \geqq 0$. A similar estimate holds for $K_x[s(\tau),\tau;-s^*(u),u]$, whence the validity of (8.4).

Inequality (8.5) follows directly from (8.3), (8.4). As for (8.6), the difference of the kernels on the left-hand side is equal to J_1+J_2, where

$$J_1=\left(\frac{\tilde{s}(\tau)-\tilde{s}(u)}{\tau-u}-\alpha\right)(K[\tilde{s}(\tau)-\alpha\tau,\tau;\tilde{s}(u)-\alpha u,u]-K[s(\tau)-\alpha\tau,\tau;s(u)-\alpha u,u])$$

and

$$J_2=\frac{[\tilde{s}(\tau)-s(\tau)]-[\tilde{s}(u)-s(u)]}{\tau-u}K[s(\tau)-\alpha\tau,\tau;s(u)-\alpha u,u].$$

Using (8.3) and the fact that $2(1+V)^2\sigma \leqq 1$, we get the estimate

$$\begin{aligned}|J_1| &\leqq (\|\tilde{c}\|+\alpha) \cdot e(2\pi)^{-1/2}(1+V)\|\tilde{c}-c\|(\tau-u)^{1/2}\\ &\leqq (1.1)(1+V)^2\sigma\|\tilde{c}-c\|(\tau-u)^{-1/2}\\ &\leqq 0.55\|\tilde{c}-c\|(\tau-u)^{-1/2}.\end{aligned}$$

On the other hand,

$$|J_2| \leqq \|\tilde{c}-c\| \cdot (2\pi)^{-1/2}(\tau-u)^{-1/2} \leqq 0.40\|\tilde{c}-c\|(\tau-u)^{-1/2}.$$

A combination of these two estimates yields (8.6).

Finally,

$$N_x[s(\tau),\tau;s(u),u]-N_x[\tilde{s}(\tau),\tau;\tilde{s}(u),u]=J+J',$$

where

$$J=K_x[s(\tau),\tau;s(u),u]-K_x[\tilde{s}(\tau),\tau;\tilde{s}(u),u],$$

and J' is J with $[s(u), \tilde{s}(u)]$ replaced by $[-s(u), -\tilde{s}(u)]$. An estimate for J is already available in (8.6). To estimate J' we write it as $J'_1 + J'_2$, with

$$J'_1 = \frac{\tilde{s}(\tau)+\tilde{s}(u)}{\tau-u}\{K[\tilde{s}(\tau), \tau; -\tilde{s}(u), u] - K[s(\tau), \tau; -s(u), u]\},$$

$$J'_2 = \frac{[\tilde{s}(\tau)-s(\tau)]+[\tilde{s}(u)-s(u)]}{\tau-u} K[s(\tau), \tau; -s(u), u].$$

Using (8.2) and (8.4) one gets the estimates

$$|J'_1| \leqq \frac{3b}{\tau-u} \cdot 28b^{-5}\|\tilde{c}-c\|(\tau-u)^{3/2} = 84b^{-4}\|\tilde{c}-c\|(\tau-u)^{1/2}$$

and

$$|J'_2| \leqq \|\tilde{c}-c\|(2\pi)^{-1/2}(\tau-u)^{-3/2} \exp\left[\frac{-b^2}{2(\tau-u)}\right] \leqq b^{-4}\|\tilde{c}-c\|(\tau-u)^{1/2},$$

where the inequality $x^2 e^{-x} \leqq 0.6$, $x \geqq 0$ has been used. A combination of the estimates for J'_1, J'_2, J gives (8.7).

COROLLARY 8.1. *For any $c \in C_{[0,\sigma]}$ with $\|c\| \leqq V$, any $\alpha, \tilde{a} \in [0, 1]$, $0 \leqq u \leqq \tau < \sigma$, and $s(\tau) = b + \int_0^\tau c(u)\, du$, $0 \leqq \tau \leqq \sigma$, we have*

$$|K[s(\tau)-\tilde{a}\tau, \tau; s(u)-\tilde{\alpha}u, u] - K[s(\tau)-\alpha\tau, \tau; s(u)-\alpha u, u]| \leqq 2(1+V)|\tilde{\alpha}-\alpha|(\tau-u)^{1/2}, \tag{8.9}$$

$$|K_x[s(\tau)-\tilde{\alpha}\tau, \tau; s(u)-\tilde{\alpha}u, u] - K_x[s(\tau)-\alpha\tau, \tau; s(u)-\alpha u, u]| \leqq |\tilde{\alpha}-\alpha|(\tau-u)^{-1/2}. \tag{8.10}$$

LEMMA 8.2. *For any $c \in C_{[0,\sigma]}$; $\|c\| \leqq V$, $0 < \varepsilon < \frac{1}{2}$, $0 \leqq \alpha \leqq 1$ and any $0 \leqq u \leqq \tau_1 < \tau_2 \leqq \sigma$, we have*

$$|K[s(\tau_2)-\alpha\tau_2, \tau_2; s(u)-\alpha u, u] - K[s(\tau_1)-\alpha\tau_1, \tau_1; s(u)-\alpha u, u]| \leqq (\tau_2-\tau_1)^{\varepsilon}(\tau_1-u)^{-1/2-\varepsilon}, \tag{8.11}$$

$$|K[s(\tau_2), \tau_2; -s(u), u] - K[s(\tau_1), \tau_1; -s(u), u]| \leqq B_0(\tau_2-\tau_1)(\tau_2-u)^{1/2}, \tag{8.12}$$

$$|G[s(\tau_2), \tau_2; s(u), u] - G[s(\tau_1), \tau_1; s(u), u]| \leqq (1+B_0)(\tau_2-\tau_1)^{\varepsilon}(\tau_1-u)^{-1/2-\varepsilon}, \tag{8.13}$$

$$|K_x[s(\tau_2)-\alpha\tau_2, \tau_2; s(u)-\alpha u, u] - K_x[s(\tau_1)-\alpha\tau_1, \tau_1; s(u)-u, u]| \leqq (1+2V)(\tau_2-\tau_1)^{\varepsilon}(\tau_1-u)^{-1/2-\varepsilon}, \tag{8.14}$$

$$|N_x[s(\tau_2), \tau_2; s(u), u] - N_x[s(\tau_1), \tau_1; s(u), u]| \leqq B_1(\tau_2-\tau_1)^{\varepsilon}(\tau_1-u)^{-1/2-\varepsilon}, \tag{8.15}$$

where $B_0 = (4b^3V + 3(9b^2+1))/b^6$ and B_1 is the constant introduced in (8.18).

Proof. We first prove (8.11); the difference of the two kernels can be written as $J_1 + J_2$, where

$$J_1 = K[s(\tau_2)-\alpha\tau_2, \tau_2; s(u)-\alpha u, u] - K[s(\tau_1)-\alpha\tau_1, \tau_2; s(u)-\alpha u, u]$$

$$J_2 = K[s(\tau_1)-\alpha\tau_1, \tau_2; s(u)-\alpha u, u] - K[s(\tau_1)-\alpha\tau_1, \tau_1,; s(u)-\alpha u, u].$$

Besides,

$$J_1 = K[s(\tau_2)-\alpha\tau_2, \tau_2; s(u)-\alpha u, u] \cdot \left[1 - \exp\left\{-\frac{[s(\tau_1)-\alpha\tau_1-s(u)+\alpha u]^2 - [s(\tau_2)-\alpha\tau_2-s(u)+\alpha u]^2}{2(\tau_2-u)}\right\}\right]. \tag{8.16}$$

The difference of the squares in (8.16) is bounded in absolute value by

$$|s(\tau_2)-s(\tau_1)-\alpha(\tau_2-\tau_1)|\cdot|s(\tau_2)-s(u)-\alpha(\tau_2-u)+s(\tau_1)-s(u)-\alpha(\tau_1-u)|$$
$$\leqq 2(1+V)^2(\tau_2-\tau_1)(\tau_2-u),$$

so that, if x denotes the expression inside the inner brackets in (8.16), then evidently $|x|\leqq(1+V)^2(\tau_2-\tau_1)\leqq(1+V)^2\sigma<1$ and upon using again the fact that $|1-e^x|\leqq et$, $|x|\leqq t<1$ we have, $|J_1|\leqq(2\pi)^{-1/2}e(1+V)^2\cdot(\tau_2-\tau_1)(\tau_2-u)^{-1/2}$. But

$$(\tau_2-\tau_1)(\tau_2-u)^{-1/2}\leqq\sigma(\tau_2-\tau_1)^{\varepsilon}(\tau_2-u)^{-1/2-\varepsilon}, \tag{8.17}$$

which by virtue of $2(1+V)^2\sigma\leqq 1$ gives the final estimate $|J_1|\leqq(0.55)(\tau_2-\tau_1)^{\varepsilon}(\tau_1-u)^{-1/2-\varepsilon}$ On the other hand, $(2\pi)^{1/2}\cdot|J_2|\leqq(\tau_1-u)^{-1/2}-(\tau_2-u)^{-1/2}\leqq(\tau_2-\tau_1)^{\varepsilon}(\tau_1-u)^{-1/2-\varepsilon}$. Combining the estimates for J_1, J_2, one easily gets (8.11).

To prove (8.12) we write the difference of the kernels by virtue of the mean value theorem as I_1+I_2, with $I_1=[s(\tau_2)-s(\tau_1)]K_x[s^*,\tau_2;-s(u),u]$, $I_2=(\tau_2-\tau_1)K_\tau[s(\tau_1),\tau^*;-s(u),u]$ where τ^* and s^* are numbers between τ_1, τ_2 and $s(\tau_1)$, $s(\tau_2)$, respectively. Note that, because of (8.2),

$$|I_1|\leqq V(\tau_2-\tau_1)\cdot\frac{3b}{(2\pi)^{1/2}(\tau_2-u)^{3/2}}\exp\left[\frac{-b^2}{2(\tau_2-u)^2}\right]$$
$$\leqq 8(2\pi)^{-1/2}b^{-3}V(\tau_2-\tau_1)(\tau_2-u)^{1/2}.$$

On the other hand,

$$K_\tau(x,\tau;\xi,u)=\frac{(x-\xi)^2-(\tau-u)}{2(\tau-u)^2}K(x,\tau;\xi,u),$$

so that

$$|I_2|\leqq\frac{\tau_2-\tau_1}{2(2\pi)^{1/2}}\frac{9b^2+1}{(\tau^*-u)^{5/2}}\exp\left[\frac{-b^2}{2(\tau^*-u)}\right]\leqq\frac{6(9b^2+1)}{b^6(2\pi)^{1/2}}(\tau_2-\tau_1)(\tau^*-u)^{1/2}.$$

The estimates of I_1, I_2 together give (8.12). Equation (8.13) follows directly from (8.11), (8.12) and 8.(17).

We are now in a position to prove (8.14); we put

$$K_x[s(\tau_2)-\alpha\tau_2,\tau_2;s(u)-\alpha u,u]-K_x[s(\tau_1)-\alpha\tau_1,\tau_1;s(u)-\alpha u,u]=I_1+I_2-I_3.$$

where I_1, I_2, I_3 along with their estimates are given by

$$I_1=\frac{s(\tau_2)-s(\tau_1)}{\tau_2-u}K[s(\tau_2)-\alpha\tau_2,\tau_2;s(u)-\alpha u,u],$$

$$|I_1|\leqq\frac{V}{(2\pi)^{1/2}}\frac{\tau_2-\tau_1}{(\tau_2-u)^{3/2}}\leqq\frac{V}{(2\pi)^{1/2}}\frac{(\tau_2-\tau_1)^{\varepsilon}}{(\tau_1-u)^{1/2+\varepsilon}};$$

$$I_2=\left[\frac{s(\tau_1)-s(u)}{\tau_2-u}-\alpha\right]\{K[s(\tau_2)-\alpha\tau_2,\tau_2;s(u)-\alpha u,u]$$
$$-K[s(\tau_1)-\alpha\tau_1,\tau_1;s(u)-\alpha u,u]\},$$

and, by (8.11),

$$|I_2| \leqq (1+V)(\tau_2-\tau_1)^{\varepsilon}(\tau_1-u)^{-1/2-\varepsilon},$$

$$I_3 = \frac{\tau_2-\tau_1}{\tau_2-u}\,\frac{s(\tau_1)-s(u)}{\tau_1-u}\,K[s(\tau_1)-\alpha\tau_1,\ \tau_1;\ s(u)-\alpha u,\ u],$$

$$|I_3| \leqq \frac{V}{(2\pi)^{1/2}}\,\frac{\tau_2-\tau_1}{(\tau_1-u)^{1/2}(\tau_2-u)} \leqq \frac{V}{(2\pi)^{1/2}}\,\frac{(\tau_2-\tau_1)^{\varepsilon}}{(\tau_1-u)^{1/2+\varepsilon}}.$$

Combining the three estimates above, one gets (8.14).

Finally (8.15) is proved. The difference of the two kernels is $I+I'$, where $I = K_x[s(\tau_1), \tau_1; s(u), u] - K_x[s(\tau_2), \tau_2; s(u), u]$ and I' is I with $s(u)$ replaced by $-s(u)$. An estimate of I is given by (8.14); to estimate I', the latter is decomposed as $I' = I_1' + I_2' - I_3'$, where I_1', I_2' and I_3' are as I_1, I_2, I_3 above with $\alpha = 0$ and $s(u)$ replaced by $-s(u)$. We estimate I_1' by the same quantity as for I_1, while for I_2', I_3' we have the upper bounds

$$|I_2'| \leqq \frac{3b}{\tau_2-u}\cdot B_0(\tau_2-\tau_1)(\tau_2-u)^{1/2} \leqq 3bB_0\cdot\frac{(\tau_2-\tau_1)^{\varepsilon}}{(\tau_2-u)^{1/2+\varepsilon}},$$

$$|I_3'| \leqq \frac{3b(\tau_2-\tau_1)}{(2\pi)^{1/2}(\tau_1-u)^{5/2}}\exp\left[\frac{-b^2}{2(\tau_1-u)}\right] \leqq 3b^{-3}(\tau_2-\tau_1)(\tau_1-u)^{-1/2} \leqq 3b^{-3}\,\frac{(\tau_2-\tau_1)^{\varepsilon}}{(\tau_1-u)^{1/2+\varepsilon}},$$

where (8.2), (8.12), (8.17) and the inequality $x^2e^{-x} \leqq 0.6$, $x \geqq 0$ have been used. Combining all estimates together and introducing the constant

$$B_1 = 1+3V+3b^{-3}+3b^{-5}[4b^3V+3(9b^2+1)], \tag{8.18}$$

one verifies the validity of (8.15). The proof of the lemma is complete.

LEMMA 8.3. *For any $c \in C_{[0,\sigma]}$, $\|c\| \leqq V$ and $0 \leqq \alpha \leqq 1$, $s(\tau) = b + \int_0^\tau c(u)\,du$, $0 \leqq \tau \leqq \sigma$, we have*

$$|K_x[s(\tau)-\alpha\tau,\ \tau;\ s(u)-\alpha u,\ u]| \leqq \frac{1}{2}(1+V)(\tau-u)^{1/2}, \qquad 0 \leqq u < \tau \leqq \sigma, \tag{8.19}$$

$$|N_x[s(\tau), \tau; s(u), u]| \leqq \frac{1}{2}(V+6b^{-1})(\tau-u)^{1/2}, \qquad 0 \leqq u < \tau \leqq \sigma. \tag{8.20}$$

9. Appendix B. In this section we discuss some properties of the integral operator

$$\gamma(\tau; c) = 2\int_0^b G[s(\tau), \tau; \xi, 0]f'(\xi)\,d\xi - 1, \qquad 0 \leqq \tau \leqq \sigma, \tag{9.1}$$

where $s(\tau) = b + \int_0^\tau c(u)\,du$, $c \in C_{[0,\sigma]}$, $\|c\| = \sup_{0\leqq\tau\leqq\sigma}|c(\tau)| \leqq V$ and σ satisfies (8.1). We also establish the solvability of the Volterra integral equation (9.9) below.

LEMMA 9.1. *Under the above assumptions,*

$$|\gamma(\tau; c)| \leqq \left(2V+K+\frac{1}{2b}\right)\tau^{1/2}, \qquad 0 \leqq \tau \leqq \sigma. \tag{9.2}$$

As an immediate consequence, $\gamma(0; c) = \lim_{\tau\downarrow 0}\gamma(\tau; c) = 0$.

Proof. Because $f'(b) = 1$, we have $\frac{1}{2}\gamma(\tau; c) = I_1 + I_2 - I_3$, with

$$I_1 = \int_{-b}^{b}\{K[s(\tau), \tau; \xi, 0] - K(b, \tau; \xi, 0)\}f'(\xi)\,d\xi,$$

$$I_2 = \int_{-b}^{b} K(b, \tau; \xi, 0)[f'(\xi)-f'(b)]\,d\xi, \qquad I_3 = \int_{-\infty}^{-b} K(b, \tau; \xi, 0)\,d\xi.$$

The following estimates hold:

$$I_3 = 1 - \Phi(2b\tau^{-1/2}) \leqq (2\pi)^{-1/2} \frac{\tau^{1/2}}{2b},$$

$$|I_2| \leqq K \int_{-b}^{b} (b-\xi) K(b, \tau; \xi, 0)\, d\xi \leqq K(2\pi)^{-1/2} \tau^{1/2},$$

$$|I_1| = \frac{|s(\tau)-b|}{\tau} \int_{-b}^{b} |s^* - \xi| K(s^*, \tau; \xi, 0)\, d\xi \leqq 2V(2\pi)^{-1/2} \tau^{1/2},$$

where s^* is a number between b and $s(\tau)$. Equation (9.2) follows from these three estimates.

LEMMA 9.2. Continuity of $\gamma(\tau; c)$ in c. *For any* $c, \tilde{c} \in C_{[0,\sigma]}$ *such that* $\|c\|, \|\tilde{c}\| \leqq V$,

$$|\gamma(\tau; \tilde{c}) - \gamma(\tau; c)| \leqq 2|\tilde{s}(\tau) - s(\tau)| \tau^{-1/2} \leqq 2\|\tilde{c} - c\| \tau^{1/2}, \qquad 0 \leqq \tau \leqq \sigma. \tag{9.3}$$

Proof. Consider c, $\tilde{c}$ as above, along with $s(\tau) = b + \int_0^\tau c(u)\, du$, $\tilde{s}(\tau) = b + \int_0^\tau \tilde{c}(u)\, du$. By the mean value theorem, there exists a number s^* between $s(\tau)$ and $\tilde{s}(\tau)$ such that

$$\begin{aligned} \gamma(\tau; \tilde{c}) - \gamma(\tau; c) &= [\tilde{s}(\tau) - s(\tau)] \int_{-b}^{b} K_x(s^*, \tau; \xi, 0) f'(\xi)\, d\xi \\ &= \frac{\tilde{s}(\tau) - s(\tau)}{\tau} \int_{-b}^{b} (\xi - s^*) K(s^*, \tau; \xi, 0) f'(\xi)\, d\xi. \end{aligned}$$

Therefore,

$$\begin{aligned} |\gamma(\tau; \tilde{c}) - \gamma(\tau; c)| &\leqq \frac{|\tilde{s}(\tau) - s(\tau)|}{\tau} \int_{-b}^{b} |\xi - s^*| K(s^*, \tau; \xi, 0)\, d\xi \\ &< 2|\tilde{s}(\tau) - s(\tau)| \cdot \tau^{-1/2}. \end{aligned}$$

LEMMA 9.3. *For any* $c \in C_{[0,\sigma]}$, $\|c\| \leqq V$, $\gamma(\tau; c)$ *is a continuously differentiable function on* $(0, \sigma]$. *There exists a positive constant* $B_2 = B_2(b, V, K, L)$ *such that*

$$\sup_{0 < \tau \leqq \sigma} \tau^{1/2} |\gamma'(\tau; c)| \leqq B_2, \tag{9.4}$$

$$\lim_{\tau \downarrow 0} \tau^{1/2} \gamma'(\tau; c) = -\left(\frac{2}{\pi}\right)^{1/2} f''(b). \tag{9.5}$$

Proof. We have

$$\begin{aligned} \gamma'(\tau; c) = {} & 2\left[\frac{s(\tau)-b}{\tau} - c(\tau) - f''(b)\right] K[s(\tau), \tau; b, 0] \\ & - 2\left[\frac{s(\tau)+b}{\tau} - c(\tau) - f''(b)\right] K[s(\tau), \tau; -b, 0] \\ & + 2c(\tau) \int_{-b}^{b} K[s(\tau), \tau; \xi, 0] f''(\xi)\, d\xi \\ & + \int_{-b}^{b} K[s(\tau), \tau; \xi, 0] f'''(\xi)\, d\zeta. \end{aligned} \tag{9.6}$$

Just as in Lemma 9.1, it can be shown that there exists a positive constant such that

$$\left|\int_{-b}^{b} K[s(\tau), \tau; \xi, 0]f''(\xi)\,d\xi - \frac{f''(b)}{2}\right| + \left|\int_{-b}^{b} K[s(\tau), \tau; \xi, 0]f''''(\xi)\,d\xi - \frac{f''''(b)}{2}\right| \leqq \text{const} \cdot \tau^{1/2}.$$

Also,

$$\left|\frac{s(\tau)+b}{\tau} - c(\tau) - f''(b)\right| K[s(\tau), \tau; -b, 0] \leqq (2\pi)^{-1/2}(3b + V + K)\tau^{-3/2} \cdot \exp\left(\frac{-9b^2}{8\tau}\right) \to 0,$$

as $\tau \downarrow 0$. From (9.6), it follows that $\tau^{1/2}\gamma'(\tau; c)$ is bounded on $(0, \sigma]$, uniformly in c, which is precisely the assertion of (9.4). On the other hand, passing to the limit as $\tau \downarrow 0$ and observing that $\lim_{\tau \downarrow 0} \tau^{1/2} K[s(\tau), \tau; b, 0] = (2\pi)^{-1/2}$, one gets

$$\lim_{\tau \downarrow 0} \tau^{1/2}\gamma'(\tau, c) = -\left(\frac{2}{\pi}\right)^{1/2} f''(b).$$

LEMMA 9.4. *There exists a positive constant $B_3 = B_3(b, V, K, L)$, such that for any $c, \tilde{c} \in C_{[0,\sigma]}$; $\|c\|, \|\tilde{c}\| \leqq V$ we have*

$$\sup_{0 \leqq \tau \leqq \sigma} \tau^{1/2} |\gamma'(\tau; \tilde{c}) - \gamma'(\tau, c)| \leqq B_3 \|\tilde{c} - c\|. \tag{9.7}$$

Proof. From (9.6) one gets the decomposition, $\gamma'(\tau; \tilde{c}) - \gamma'(\tau; c) = \sum_{i=1}^{7} I_i$, where

$$I_1 = 2\left\{\frac{\tilde{s}(\tau) - s(\tau)}{\tau} - [\tilde{c}(\tau) - c(\tau)]\right\} K[\tilde{s}(\tau), \tau; b, 0],$$

$$|I_1| \leqq 4(2\pi)^{-1/2} \|\tilde{c} - c\| \tau^{-1/2},$$

$$I_2 = 2\left[\frac{s(\tau) - b}{\tau} - c(\tau) - f''(b)\right] \{K[\tilde{s}(\tau), \tau; b, 0] - K[s(\tau), \tau; b, 0]\},$$

$$|I_2| = 4(2V + K)(1 + V)\|\tilde{c} - c\| \tau^{1/2} \quad \text{by (8.3)},$$

$$-I_3 = 2\left(\frac{\tilde{s}(\tau) - s(\tau)}{\tau} - [\tilde{c}(\tau) - c(\tau)]\right) K[\tilde{s}(\tau), \tau; -b, 0],$$

$$|I_3| \leqq 2(2\pi\tau)^{-1/2} \exp\left(\frac{-b^2}{2\tau}\right) \|\tilde{c} - c\|,$$

$$-I_4 = 2\left[\frac{s(\tau) + b}{\tau} - c(\tau) - f''(b)\right] \{K[\tilde{s}(\tau), \tau; -b, 0] - K[s(\tau), \tau; -b, 0]\},$$

and by (8.4),

$$|I_4| \leqq 28[5b + 2(V + K)]b^{-5} \cdot \|\tilde{c} - c\| \tau^{1/2},$$

$$I_5 = 2[\tilde{c}(\tau) - c(\tau)] \int_{-b}^{b} K[\tilde{s}(\tau), \tau; \xi, 0]f''(\xi)\,d\xi,$$

$$|I_5| \leqq 2K\|\tilde{c} - c\|,$$

$$I_6 = 2c(\tau)\left\{\int_{-b}^{b} K[\tilde{s}(\tau), \tau; \xi, 0]f''(\xi)\,d\xi - \int_{-b}^{b} K[s(\tau), \tau; \xi, 0]f''(\xi)\,d\xi\right\},$$

$$I_7 = 2\int_{-b}^{b} K[\tilde{s}(\tau), \tau; \xi, 0]f''''(\xi)\,d\xi - \int_{-b}^{b} K[s(\tau), \tau; \xi, 0]f''''(\xi)\,d\xi.$$

The last two terms can be estimated, as in Lemma 9.2, by $|I_6+I_7| \leqq const.\ \|\tilde{c}-c\|\tau^{1/2}$. Combining all estimates together, we get (9.7).

COROLLARY 9.1. *For any $c \in C_{[0,\sigma]}$, $\gamma(\tau; c)$ belongs to the space Z_σ defined in* § 6 *[see* (6.3)*]. Furthermore, if $c,\ \tilde{c} \in C_{[0,\sigma]}$, $\|c\|;\ \|\tilde{c}\| \leqq V$, there exists a positive constant $B_4 = B_4(b, K, L, V)$ such that*

$$\|\gamma(\tilde{c})-\gamma(c)\|_{1/2} \leqq B_4\|\tilde{c}-c\|. \tag{9.8}$$

In the remaining part of this section, we establish the solvability of a certain Volterra integral equation of the first kind.

LEMMA 9.5. *For any $c \in C_{[0,\sigma]}$, $\|c\| \leqq V$ and $\lambda \in Z_\sigma$, $\|\lambda\|_{1/2} \leqq \Lambda$, the Volterra integral equation of the first kind,*

$$2\int_0^\tau G[s(\tau), \tau; s(u), u] \cdot \nu(u)\, du = \lambda(\tau), \qquad 0 \leqq \tau \leqq \sigma, \tag{9.9}$$

has a unique solution $\nu(\tau) \in C_{[0,\sigma]}$. The dependence of this solution on the pair (c, λ) is continuous, in the sense that if ν is the solution corresponding to the pair (c, λ) and $\tilde{\nu}$ the solution corresponding to $(\tilde{c}, \tilde{\lambda})$ with $\|c\|, \|\tilde{c}\| \leqq V; \|\lambda\|_{1/2}, \|\tilde{\lambda}\|_{1/2} \leqq \Lambda$, there exists a constant $B_5 = B_5(b, V, \Lambda)$ such that

$$\|\tilde{\nu}-\nu\| \leqq B_5(\|\tilde{c}-c\| + \|\tilde{\lambda}-\lambda\|_{1/2}). \tag{9.10}$$

Proof. Consider (9.9) with t replacing τ, multiply both sides by $(\tau-t)^{-1/2}$ and integrate the resulting equation from 0 to τ. We come up with the new Volterra equation of the first kind,

$$\int_0^\tau M(\tau, u; c)\nu(u)\, du = \psi(\tau; \lambda), \qquad 0 \leqq \tau \leqq \sigma, \tag{9.11}$$

where

$$\begin{aligned} M(\tau, u; c) &= \int_u^\tau \frac{\mu(\tau, u; c)}{(\tau-t)^{1/2}(t-u)^{1/2}}\, dt, \\ \mu(\tau, u; c) &= 2(\tau-u)^{1/2} G[s(\tau), \tau; s(u), u], \qquad 0 \leqq u < \tau \leqq \sigma, \end{aligned} \tag{9.12}$$

$$\psi(\tau; \lambda) = \int_0^\tau \frac{\lambda(u)\, du}{(\tau-u)^{1/2}}, \qquad 0 \leqq \tau \leqq \sigma. \tag{9.13}$$

In the equivalent form (9.11), the integral equation can be reduced to a Volterra equation of the second kind and thus be solved by iteration, as is presently shown. Since

$$-\frac{\partial}{\partial t}\left[2\tan^{-1}\left(\frac{\tau-t}{t-u}\right)^{1/2}\right] = (\tau-t)^{-1/2}(t-u)^{-1/2},$$

integration by parts in (9.12) yields

$$M(\tau, u; c) = \pi \cdot \mu(u, u; c) + 2\int_u^\tau \tan^{-1}\left(\frac{\tau-t}{t-u}\right)^{1/2} \cdot \mu_1(t, u; c)\, dt, \tag{9.14}$$

so that

$$M_\tau(\tau, u; c) = \frac{1}{\tau-u}\int_u^\tau \left(\frac{t-u}{\tau-t}\right)^{1/2} \mu_1(t, u; c)\, dt. \tag{9.15}$$

It can be shown that $M_\tau(\tau, u; c)$, $0 \leqq u < \tau \leqq \sigma$ is a bounded, continuous kernel,

continuously dependent on c. Indeed,

$$\mu_1(\tau, u; c) = \frac{s(\tau)-s(u)}{(\tau-u)^{1/2}}\left[\frac{s(\tau)-s(u)}{\tau-u} - 2c(\tau)\right] K[s(\tau), \tau; s(u), u] \tag{9.16}$$
$$-\frac{s(\tau)+s(u)}{(\tau-u)^{1/2}}\left[\frac{s(\tau)+s(u)}{\tau-u} - 2c(\tau)\right] K[s(\tau), \tau; -s(u), u],$$

and it is not hard to get the bound $|\mu_1(\tau, u; c)| \leqq (2\pi)^{-1/2}(3V^2 + 29b^{-2})$. Therefore, $M_\tau(\tau, u; c)$ is bounded:

$$|M_\tau(\tau, u; c)| \leqq \left(\frac{\pi}{8}\right)^{1/2}(3V^2 + 29b^{-2}). \tag{9.17}$$

To conclude continuity in c, we use (8.3), (8.4) to get

$$|\mu_1(\tau, u; \tilde{c}) - \mu_1(\tau, u; c)| \leqq \mu\|\tilde{c} - c\|, \qquad \mu = 3V + 4V^2(1+V) + 343b^{-3}, \tag{9.18}$$

whence

$$|M_\tau(\tau, u; \tilde{c}) - M_\tau(\tau, u; c)| \leqq \frac{\mu\|\tilde{c}-c\|}{\tau-u}\int_u^\tau \left(\frac{t-u}{\tau-t}\right)^{1/2} dt = \frac{\pi\mu}{2}\|\tilde{c}-c\|. \tag{9.19}$$

Finally, we compute $M(\tau, \tau; c) = \lim_{u\uparrow\tau} M(\tau, u; c)$.
Write $M(\tau, u; c) = (2/\pi)^{1/2}(I_1 - I_2)$, with

$$I_1 = \int_u^\tau \frac{\exp\left\{-\dfrac{[s(\tau)-s(u)]^2}{2(\tau-u)}\right\}}{(\tau-t)^{1/2}(t-u)^{1/3}}\, dt,$$

and I_2 similar to I_1, with $-s(u)$ instead of $s(u)$. Note that

$$0 \leqq I_2 \leqq \exp\left[\frac{-b^2}{2(\tau-u)}\right]\int_u^\tau \frac{dt}{(\tau-t)^{1/2}(\tau-u)^{1/2}} = \pi \cdot \exp\left[\frac{-b^2}{2(\tau-u)}\right], \tag{9.20}$$
$$\exp\left[-\frac{V^2(\tau-u)}{2}\right]\int_u^\tau \frac{dt}{(\tau-t)^{1/2}(t-u)^{1/2}} = \pi \exp\left[\frac{-V^2(\tau-u)}{2}\right] \leqq I_1 \leqq \pi.$$

Therefore $\lim I_2 = 0$ and $\lim I_1 = \pi$ as $u \uparrow \tau$, and so

$$M(\tau, \tau; c) = (2\pi)^{1/2}.$$

Now, one can formally differentiate both sides of the integral equation (9.11) with respect to τ and get

$$(2\pi)^{1/2}\nu(\tau) + \int_0^\tau M_\tau(\tau, u; c)\nu(u)\, du = \psi'(\tau, \lambda), \qquad 0 \leqq \tau \leqq \sigma, \tag{9.21}$$

where $\psi(\tau; \lambda) = \int_0^\tau \lambda(u)(\tau-u)^{-1/2}\, du = 2\int_0^\tau \lambda'(u)(\tau-u)^{1/2}\, du$ by partial integration, $\lambda \in Z_\sigma$. Therefore

$$\psi'(\tau; \lambda) = \int_0^\tau \lambda'(u)(\tau-u)^{-1/2}\, du, \qquad 0 \leqq \tau \leqq \sigma, \tag{9.22}$$

and hence

$$\|\psi'(\lambda)\| = \sup_{0\leqq\tau\leqq\sigma} |\psi'(\tau; \lambda)| \leqq \pi\|\lambda\|_{1/2} \leqq \pi\Lambda,$$
$$\|\psi'(\tilde{\lambda}) - \psi'(\lambda)\| \leqq \pi\|\tilde{\lambda} - \lambda\|_{1/2}. \tag{9.23}$$

The integral equation (9.21) is Volterra of the second kind in $\nu(\tau)$, and can be solved by iteration. Continuous dependence of the solution to (9.21) on the pair (c, λ), and hence, (9.10), is a direct consequence of (9.19), (9.23).

10. Appendix C. In this section we discuss the properties of the mapping

$$h_2^\alpha(\tau; c) = 2\int_b^\infty K[s(\tau) - \alpha\tau, \tau; \xi, 0] f'(\xi)\, d\xi - 1, \qquad 0 \leqq \tau \leqq \sigma, \tag{10.1}$$

where $s(\tau) = b + \int_0^\tau c(u)\, du$, $c \in C_{[0,\sigma]}$, $\|c\| \leqq V$ and σ satisfies (8.1).

LEMMA 10.1. *Under the above assumptions,*

$$|h_2^\alpha(\tau; c)| \leqq K[1 + (1+V)(2+3b)]\tau^{1/2}, \qquad 0 \leqq \tau \leqq \sigma. \tag{10.2}$$

Proof. Because $f'(b) = 1$, $h_2^\alpha(\tau; c)/2 = I_1 + I_2$, with $I_2 = \int_b^\infty K(b, \tau; \xi, 0) \cdot [f'(\xi) - f'(b)]\, d\xi$ and $I_1 = \int_b^\infty \{K[s(\tau) - \alpha\tau, \tau; \xi, 0] - K[b, \tau; \xi, 0]\} f'(\xi)\, d\xi$. The following estimates hold:

$$|I_2| \leqq K\int_b^\infty (\xi - b) K(b, \tau; \xi, 0)\, d\xi = K(2\pi)^{-1/2}\tau^{1/2}$$

and

$$\begin{aligned} |I_1| &\leqq |s(\tau) - b - \alpha\tau| \int_b^\infty \frac{|s^* - \xi|}{\tau} K(s^*, \tau; \xi, 0)\, K\xi\, d\xi \\ &\leqq K(1+V)\int_b^\infty |s^* - \xi| K(s^*, \tau; \xi, 0)\xi\, d\xi, \end{aligned} \tag{10.3}$$

by virtue of the mean value theorem, where s^* is a number between b and $s(\tau) - \alpha\tau$. There are two cases to be considered separately:

Case I. $s^* \leqq b$. The integral on the right-hand side of (10.3) is

$$\begin{aligned} \int_b^\infty (\xi - s^*)\xi K(s^*, \tau; \xi, 0)\, d\xi &= \int_b^\infty (\xi - s^*)^2 K(s^*, \tau; \xi, 0)\, d\xi \\ &\quad + s^* \int_b^\infty (\xi - s^*) K(s^*, \tau; \xi, 0)\, d\xi \\ &\leqq \tau + b\tau^{1/2}/(2\pi)^{1/2} \end{aligned}$$

and therefore, $|I_1| \leqq K(1+V)(1 + b/2)\tau^{1/2}$, $0 \leqq \tau \leqq \sigma$.

Case II. $s^* > b$. The interval of integration is divided into two parts as follows. $\int_b^\infty = \int_b^{s^*} + \int_{s^*}^\infty = Y_1 + Y_2$, where

$$Y_1 = \int_b^{s*} (s^* - \xi)\xi K(s^*, \tau; \xi, 0)\, d\xi \leqq \frac{3b}{2}\int_b^{s^*} (s^* - \xi) K(s^*, \tau; \xi, 0)\, d\xi \leqq \frac{3b}{2(2\pi)^{1/2}}\tau^{1/2},$$

$$Y_2 = \int_{s^*}^\infty (\xi - s^*)\xi K(s^*, \tau; \xi, 0)\, d\xi \leqq \tau + \frac{3b}{2(2\pi)^{1/2}}\tau^{1/2},$$

in analogy with the estimate of I_1. Therefore in any case $|I_1| = K(1+V)(1 + 3b/2)\tau^{1/2}$. Inequality (10.2) follows from the estimates of both I_1 and I_2.

LEMMA 10.2. Continuity of $h_2^\alpha(\tau; c)$ in (c, α). *For any* $\tilde{c}$, $c \in C_{[0,\sigma]}$; $\|c\|, \|\tilde{c}\| \leqq V$ *and* $\alpha, \tilde{\alpha} \in [0, 1]$,

$$\begin{aligned} |h_2^{\tilde{\alpha}}(\tau; \tilde{c}) - h_2^\alpha(\tau; c)| &\leqq K(2+3b)|\tilde{s}(\tau) - s(\tau)|\tau^{-1/2} \\ &\leqq K(2+3b)\|\tilde{c} - c\|\tau^{1/2}, \qquad 0 \leqq \tau \leqq \sigma, \end{aligned} \tag{10.4}$$

$$|h_2^{\tilde{\alpha}}(\tau; c) - h_2^{\alpha}(\tau; c)| \leqq K(2+3b)|\tilde{\alpha} - \alpha|\tau^{1/2}, \qquad 0 \leqq \tau \leqq \sigma. \tag{10.4}'$$

Besides, if in addition to (8.1) *the initial step* σ *satisfies the restriction*

$$\sigma \leqq \frac{2}{(1+V)^2} \ln \frac{3bK}{3bK-1}, \tag{10.5}$$

then

$$|\tilde{s}(\tau) - s(\tau)| \leqq 3\tau^{1/2}|h_2^{\alpha}(\tau; \tilde{c}) - h_2^{\alpha}(\tau; c)|, \qquad 0 \leqq \tau \leqq \sigma. \tag{10.6}$$

Proof. Consider any $c, \tilde{c}$ as above, $s(\tau) = b + \int_0^\tau c(u)\,du$, $\tilde{s}(\tau) = b + \int_0^\tau \tilde{c}(u)\,du$ and suppose $s(\tau) \leqq \tilde{s}(\tau)$. By the mean value theorem, there exists a number s^*, $s(\tau) \leqq s^* \leqq \tilde{s}(\tau)$ such that, if $\mu = s^* - \alpha\tau$,

$$\begin{aligned} h_2^{\alpha}(\tau; \tilde{c}) - h_2^{\alpha}(\tau; c) &= 2[\tilde{s}(\tau) - s(\tau)] \int_b^\infty K_x(s^* - \alpha\tau, \tau; \xi, 0) f'(\xi)\,d\xi \\ &= 2\frac{\tilde{s}(\tau) - s(\tau)}{\tau} \int_b^\infty (\xi - \mu) K(\mu, \tau; \xi, 0) f'(\xi)\,d\xi. \end{aligned} \tag{10.7}$$

Case I. $\mu \leqq b$. In this case the integral on the right-hand side of (10.7) is dominated by

$$\begin{aligned} K\int_b^\infty (\xi - \mu)\xi K(\mu, \tau; \xi, 0)\,d\xi &\leqq K\int_\mu^\infty (\xi - \mu)^2 K(\mu, \tau; \xi, 0)\,d\xi \\ &\quad + Kb\int_\mu^\infty (\xi - \mu) K(\mu, \tau; \xi, 0)\,d\xi \\ &\leqq K\left(\tau + \frac{b}{2}\tau^{1/2}\right); \end{aligned}$$

hence,

$$|h_2^{\alpha}(\tau; \tilde{c}) - h_2^{\alpha}(\tau; c)| \leqq K(2+b)[\tilde{s}(\tau) - s(\tau)]\tau^{-1/2}, \qquad 0 \leqq \tau \leqq \sigma. \tag{10.8}$$

On the other hand,

$$\begin{aligned} \int_b^\infty (\xi - \mu) K(\mu, \tau; \xi, 0) f'(\xi)\,d\xi &\geqq \int_b^\infty (\xi - \mu) K(\mu, \tau; \xi, 0)\,d\xi \\ &= \left(\frac{\tau}{2\pi}\right)^{1/2} \exp\left[-\frac{(b-\mu)^2}{2\tau}\right] \\ &\geqq e^{-1/4}\left(\frac{\tau}{2\pi}\right)^{1/2} \geqq (0.31)\tau^{1/2}, \end{aligned}$$

because

$$0 \leqq b - \mu \leqq b - s(\tau) + \alpha\tau \leqq (1+V)\tau, \quad \frac{(b-\mu)^2}{2\tau} \leqq \frac{(1+V)^2\tau^2}{2\tau} \leqq \frac{(1+V)^2\sigma}{2} \leqq \frac{1}{4},$$

by (8.1). Consequently,

$$h_2^{\alpha}(\tau, \tilde{c}) - h_2^{\alpha}(\tau; c) \geqq (0.62)\frac{\tilde{s}(\tau) - s(\tau)}{\tau^{1/2}}, \qquad 0 \leqq \tau \leqq \sigma. \tag{10.9}$$

Case II. $\mu > b$. In this case the integral in (10.7) becomes $I_2 - I_1$, where

$$I_1 = \int_b^\mu (\mu-\xi)K(\mu,\tau;\xi,0)f'(\xi)\,d\xi \leqq f'\left(\frac{3b}{2}\right)\int_b^\mu (\mu-\xi)K(\mu,\tau;\xi,0)\,d\xi$$

$$\leqq K\frac{3b}{2}\left(\frac{\tau}{2\pi}\right)^{1/2},$$

$$I_2 = \int_\mu^\infty (\xi-\mu)K(\mu,\tau;\xi,0)f'(\xi)\,d\xi \leqq K\int_\mu^\infty \xi(\xi-\mu)K(\mu,\tau;\xi,0)\,d\xi$$

$$= K\int_\mu^\infty (\xi-\mu)^2 K(\mu,\tau;\xi,0)\,d\xi$$

$$+ K\mu\int_\mu^\infty (\xi-\mu)K(\mu,\tau;\xi,0)\,d\xi,$$

and therefore,

$$I_2 \leqq K\left[1+\frac{3b}{2}(2\pi)^{-1/2}\right]\tau^{1/2}.$$

Combining the estimates for I_1, I_2, one gets in this case

$$|h_2^\alpha(\tau;\tilde{c}) - h_2^\alpha(\tau;c)| \leqq K(2+3b)[\tilde{s}(\tau)-s(\tau)]\tau^{-1/2}, \qquad 0 \leqq \tau \leqq \sigma. \tag{10.10}$$

Now (10.4) follows readily from (10.8), (10.10), while (10.5) is an immediate consequence thereof. On the other hand,

$$I_2 \geqq \int_\mu^\infty (\xi-\mu)K(\mu,\tau;\xi,0)\,d\xi = \left(\frac{\tau}{2\pi}\right)^{1/2},$$

and

$$I_1 \leqq K\mu\int_b^\mu (\mu-\xi)K(\mu,\tau;\xi,0)\,d\xi$$

$$\leqq K\frac{3b}{2}\left[1-\exp\left\{-\frac{(\mu-b)^2}{2\tau}\right\}\right]\cdot\left(\frac{\tau}{2\pi}\right)^{1/2}$$

$$\leqq K\frac{3b}{2}\left(\frac{\tau}{2\pi}\right)^{1/2}\left[1-\exp\left\{-\frac{\sigma}{2}(1+V)^2\right\}\right],$$

the last inequality being valid because $0 < \mu - b \leqq \tilde{s}(\tau) - \alpha\tau - b \leqq (1+V)\tau$, $(\mu - b)^2/2\tau \leqq (\sigma/2)(1+V)^2$. Therefore

$$h_2^\alpha(\tau;\tilde{c}) - h_2^\alpha(\tau;c) = 2\frac{\tilde{s}(\tau)-s(\tau)}{\tau}(I_2 - I_1)$$

$$\geqq 2\frac{\tilde{s}(\tau)-s(\tau)}{\tau}\left(\frac{\tau}{2\pi}\right)^{1/2}\left[1-\frac{3bK}{2}\left(1-\exp\left\{-\frac{\sigma}{2}(1+V)^2\right\}\right)\right].$$

If σ satisfies (10.5), the term in the brackets is not less than $\frac{1}{2}$. So

$$h_2^\alpha(\tau;\tilde{c}) - h_2^\alpha(\tau;c) \geqq (0.38)\frac{\tilde{s}(\tau)-s(\tau)}{\tau^{1/2}}; \qquad 0 \leqq \tau \leqq \sigma. \tag{10.11}$$

A comparison of (10.9) and (10.11) shows that the latter, and hence also (10.6), is true in any possible case.

LEMMA 10.3. *For any* $c \in C_{[0,\sigma]}$, $\|c\| \leqq V$ *and* $0 \leqq \alpha \leqq 1$, $h_2^{\alpha}(\tau; c)$ *is a continuously differentiable function on* $(0, \sigma]$. *Besides, there exists a positive constant* $B_6 = B_6(b, V, K, L)$ *such that*

$$\sup_{0<\tau\leqq\sigma} \tau^{1/2}|h_2'^{\alpha}(\tau; c)| \leqq B_6. \tag{10.12}$$

Also,

$$\lim_{\tau\downarrow 0} \tau^{1/2} h_2'^{\alpha}(\tau; c) = \frac{c(0)-\alpha+f''(b)}{(2\pi)^{1/2}}. \tag{10.13}$$

Proof.

$$\begin{aligned} h_2'^{\alpha}(\tau; c) = &\left\{2[c(\tau)-\alpha]-\left[\frac{s(\tau)-b}{\tau}-\alpha\right]+f''(b)\right\} K[s(\tau)-\alpha\tau, \tau; b, 0] \\ &+\int_b^{\infty} K[s(\tau)-\alpha\tau, \tau; \xi, 0] f'''(\xi)\, d\xi \\ &+2[c(\tau)-\alpha]\int_b^{\infty} K[s(\tau)-\alpha\tau, \tau; \xi, 0] f''(\xi)\, d\xi. \end{aligned} \tag{10.14}$$

By analogy to Lemma 10.1,

$$\left|\int_b^{\infty} K[s(\tau)-\alpha\tau, \tau; \xi, 0] f''(\xi)\, d\xi - \frac{f''(b)}{2}\right| + \left|\int_b^{\infty} K[s(\tau)-\alpha\tau, \tau; \xi, 0] f'''(\xi)\, d\xi - \frac{f'''(b)}{2}\right| \leqq \text{const. } \tau^{1/2}.$$

From (10.14) it follows that $\tau^{1/2} h_2'^{\alpha}(\tau; c)$ is bounded in $(0, \sigma]$ uniformly in c, α, which gives (10.12). Passing to the limit as $\tau\downarrow 0$ in (10.14), we get

$$\lim_{\tau\downarrow 0} \tau^{1/2} h_2'^{\alpha}(\tau; c) = \frac{c(0)-\alpha+f''(b)}{(2\pi)^{1/2}}.$$

LEMMA 10.4. *There exists a positive constant* $B_7 = B_7(b, V, K, L)$ *such that for any* $\alpha, \tilde{\alpha} \in [0, 1]$; $c, \tilde{c} \in C_{[0,\sigma]}$; $\|c\|, \|\tilde{c}\| \leqq V$,

$$\sup_{0\leqq\tau\leqq\sigma} \tau^{1/2}|h_2'^{\alpha}(\tau; \tilde{c}) - h_2'^{\alpha}(\tau; c)| \leqq B_7\|\tilde{c}-c\|, \tag{10.15}$$

$$\sup_{0\leqq\tau\leqq\sigma} \tau^{1/2}|h_2'^{\tilde{\alpha}}(\tau; c) - h_2'^{\alpha}(\tau; c)| \leqq B_7|\tilde{\alpha}-\alpha|. \tag{10.16}$$

The proof of (10.15) is similar to that of (9.8), Lemma 9.4. Once (10.15) has been established, (10.16) follows readily.

COROLLARY 10.1. *For any* $\alpha \in [0, 1]$, $c \in C_{[0,\sigma]}$, $h_2^{\alpha}(\tau; c)$ *belongs to the space* Z_{σ}. *Furthermore, if* $c, \tilde{c} \in C_{[0,\sigma]}$, $\|c\|, \|\tilde{c}\| \leqq V$ *and* $\alpha, \tilde{\alpha} \in [0, 1]$, *there exists a positive constant* $B_8 = B_8(b, K, V, L)$ *such that,*

$$\|h_2^{\alpha}(\tilde{c}) - h_2^{\alpha}(c)\|_{1/2} \leqq B_8(\|\tilde{c}-c\| + |\tilde{\alpha}-\alpha|). \tag{10.17}$$

Note added in proof. Recently, L. A. Caffarelli and A. Friedman (*A free boundary problem associated with a semilinear parabolic equation*, to appear in Comm. Partial Diff. Eqs.) gave a simple proof for the C^{∞} differentiability of the free boundary in this problem. Their method is similar to that of Schaeffer [1976].

REFERENCES

V. E. BENEŠ [1971], *Existence of optimal stochastic control laws*, this Journal, 9, pp. 446–472.

——— [1974], *Girsanov functionals and optimal bang-bang laws for final value stochastic control*, Stochastic Processes Appl., 2, pp. 127–140.

——— [1975], *Composition and invariance methods for solving some stochastic control problems*, Adv. in Appl. Probab., 7, pp. 299–329.

F. E. BROWDER [1976], *Nonlinear Operators and Nonlinear Equations of Evolution in Banach Spaces*, Proc. Symp. Pure Math., Vol. XVIII, part 2, American Mathematical Society, Providence, RI.

J. R. CANNON AND C. D. HILL [1967], *Existence, uniqueness, stability and monotone dependence in a Stefan problem for the heat equation*, J. Math. Mech., 17, pp. 1–19.

JANE CRONIN [1964], *Fixed Points and Topological Degree in Nonlinear Analysis*, Math. Surveys, Vol. 11, American Mathematical Society, Providence, R[1].

M. H. A. DAVIS AND P. P. VARAIYA [1973], *Dynamic programming conditions for partially observable stochastic systems*, this Journal, 11, pp. 226–261.

R. C. DAVIS [1968], *Stochastic final value control systems with a fuel constraint*, J. Math. Anal. Appl., 21, pp. 62–78.

W. H. FLEMING AND R. W. RISHEL [1975], *Deterministic and Stochastic Optimal Control*, Springer-Verlag, Berlin.

A. FRIEDMAN [1959], *Free boundary problems for parabolic equations, I: melting of solids*, J. Math. Mech., 8, pp. 499–517.

——— [1964], *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ.

——— [1975], *Stochastic Differential Equations and Applications, vol. I*, Academic Press, New York.

I. V. GIRSANOV [1960], *On transforming a certain class of stochastic processes by absolutely continuous substitution of measures*, Theory Prob. Appl., 5, pp. 285–301.

J. M. HYMAN [1976], *The Method of Lines Solution of Partial Differential Equations*, ERDA Research and Development Report, Courant Institute of Mathematical Sciences, New York.

N. IKEDA AND S. WATANABE [1977], *A comparison theorem for solutions of stochastic differential equations and its applications*, Osaka J. Math., 14, pp. 619–633.

K. ITO AND H. P. MCKEAN, JR. [1974], *Diffusion Processes and Their Sample Paths*, 2nd ed., Springer-Verlag, Berlin.

I. KARATZAS [1979], *A free boundary problem in stochastic optimal control*, Ph.D. thesis, Columbia University, New York.

O. A. LADYŽENSKAJA, V. A. SOLONNIKOV AND N. N. URAL'CEVA [1968], *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monographs, Vol. 23, American Mathematical Society, Providence, RI.

R. S. LIPTSER AND A. N. SHIRYAYEV [1977], *Statistics of Random Processes, vol. I* (English Translation), Springer-Verlag, Berlin.

H. P. MCKEAN, JR. [1962], *A Hölder condition for Brownian local time*, J. Math. Kyôto Univ., 1–2, pp. 195–201.

L. I. RUBINŠTEIN [1967], *The Stefan Problem*, Transl. Math. Monographs, American Mathematical Society, Providence, RI.

D. G. SCHAEFFER [1976], *A new proof of the infinite differentiability of the free boundary in the Stefan problem*, J. Diff. Eqs. 20, pp. 266–269.

A. K. ZVONKIN [1974], *A transformation of the phase space of a diffusion process that removes the drift*, Math. USSR-Sb., 22, 1, pp. 129–149.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 3, May 1981

0363-0129/81/1903-0002 $01.00/0

ON SOME IMPULSE CONTROL PROBLEMS WITH LONG RUN AVERAGE COST*

MAURICE ROBIN†

Abstract. Some particular impulse control problems with infinite horizon and long run average cost are considered for Markov processes having "nice" ergodicity properties.

The main feature of the method is to start with discounted cost and then let the discount factor go to zero. This also gives an opportunity to study the asymptotic behavior of some optimal stopping time problems when the discount factor goes to zero. Probabilistic and analytical methods are used and examples are given, especially for Markov jump process and diffusion processes with reflection.

1. Introduction. Let us consider a preliminary example. Assume that one has a machine which deteriorates, and that its state is described by a Markov process x_t. At any time one can replace the machine (immediately) by another one which has the same kind of evolution. There is an operating cost $f(x)$ and a replacement cost $c(x)$, and it is desired to minimize the overall cost or infinite horizon. Moreover, one can choose the initial state of the new machine (eventually second-hand). That kind of stochastic control problem is an impulse control problem as introduced in a general setting by Bensoussan–Lions (see [2] and its bibliography) especially for diffusion processes. Impulse control was studied in [11] for a general class of Markov processes and for discounted costs or finite horizon.

The long run average cost has been considered by Lasry [8] for particular diffusion processes and with methods based on the maximum principle for partial differential equations. For optimal stopping of diffusion processes and the corresponding variational inequalities, the same kind of problem was studied by A. Bensoussan and J. L. Lions [3b].

In this paper, we shall study the long run average cost for some particular impulse control problems allowing us to give results for a class of Markov processes having nice ergodicity properties. The approach which is adopted is to start with a discounted cost criterion and to study the limiting behavior of the problem as the discount factor goes to zero. Section 2 deals with the basic definitions and assumptions, § 3 with the main results. We study successively the asymptotic behavior of the linear equation, the optimal stopping problem and the impulse control problem.

Examples are given in § 4, and in some cases quasi-variational inequalities (in the sense of [3b]) can be obtained by analytical methods.

2. Notation, assumptions and statement of the problem. Let $\Omega = D(R^+, E)$ be the space of right continuous, left limited functions from R^+ into E, a *compact* metric space.

Let $x_t(\omega) = \omega(t)$ for any $\omega \in \Omega$, $F_t^0 = \sigma\{x_s, s \leqq t\}$, $F^0 = F_\infty^0$, F_t, F the universally completed σ-field of F_t^0, F^0 (respectively).

Let θ_t be the translation operator on Ω, and C the Banach space of continuous functions on E.

Let $X = (\Omega, F_t, \theta_t, x_t, P_x)$ be a nonterminating[1] homogeneous Markov process with semigroup $\phi(t)$.

We will assume:

(2.1) $\phi(t)$ is a Feller semigroup (see Dynkin [5]).

* Received by the editors October 30, 1979 and in revised form June 12, 1980.

† INRIA, Institut National de Recherche en Information et Automatique, Domaine de Voluceau, 78 Rocquencourt, Le Chesnay, France.

[1] i.e., $\phi(t)1 = 1\ \forall t \geqq 0$.

$$\begin{aligned} &\forall f \in C, \qquad \phi(t) f \in C, \\ &\forall f \in C, \qquad \lim_{t \to 0} \phi(t) f(x) = f(x) \quad \forall x \in E. \end{aligned} \tag{2.2}$$

Moreover, we are given

$$\begin{aligned} &f \in C, \qquad f \geqq k_1 > 0, \\ &c \in C, \qquad c \geqq k_2 > 0. \end{aligned} \tag{2.3}$$

D_A will denote the domain of the infinitesimal generator A of $\phi(t)$ in C.

Problem Π_α. Let $\alpha > 0$. An admissible control ν will be a sequence of stopping times σ^n and an element $\xi \in E$ such that

$$\text{if } \tau^n = \tau^{n-1} + \sigma^n \circ \theta_{\tau^{n-1}}, \text{ then } \tau^n \uparrow \infty.$$

Then the cost is given by

$$\begin{aligned} J_x^\alpha(\nu) = E_x\left(\int_0^{\tau^1} e^{-\alpha s} f(x_s)\, ds + e^{-\alpha \tau^1} c(x_{\tau^1})\right) \\ + E_x \sum_{n \geqq 2} e^{-\alpha \tau^{n-1}} E_\xi\left(\int_0^{\sigma^n} e^{-\alpha s} f(x_s)\, ds + e^{-\alpha \sigma^n} c(x_{\sigma^n})\right). \end{aligned}$$

The optimal cost function is defined by

$$u_\alpha(x) = \inf_\nu J_x^\alpha(\nu), \tag{2.4}$$

and the control problem is to find an admissible control ν^* achieving the minimum in (2.4).

Actually, it is possible to formulate Π_α with no restrictions on τ^n and general assumptions on the ξ^n (instead of $\xi^n = \xi$), but the results of [11, Chap. 5] show that there is no loss of generality in assuming that form which contains an optimal policy.

From [11], we have the following result:

THEOREM 2.1. *Under the assumptions* (2.1), (2.3) u_α *is the maximum element of the set of functions* w *satisfying*

$$\begin{aligned} &w \leqq Mw, \qquad Mw \equiv c(x) = \inf_\xi w(\xi), \\ &w \leqq e^{-\alpha t} \phi(t) w + \int_0^t e^{-as} \phi(s) f\, ds, \\ &w \in C. \end{aligned} \tag{2.5}$$

Moreover, the following admissible control is optimal:

$$\begin{aligned} &\hat{\sigma} = \inf\,(t \geqq 0,\ u_\alpha(x_t) = M u_\alpha(x_t)), \\ &\hat{\xi} = \arg\min u_\alpha, \\ &\tau^n = \tau^{n-1} + \hat{\sigma} \circ \theta_{\tau^{n-1}}, \qquad \tau^1 = \hat{\sigma}, \\ &\xi^n = \hat{\xi}. \end{aligned} \tag{2.6}$$

Moreover, u_α *is the unique solution of*

$$u_\alpha(x) = \inf_\tau E_x\left(\int_0^\tau e^{-\alpha s} f(x_s)\, ds + e^{-\alpha \tau} M u_\alpha(x_\tau)\right). \tag{2.7}$$

It is convenient to introduce another expression for J_x^α. Let $\tilde{\Omega} = \Omega^N$, $\tilde{F} = F^{\otimes N}$, $\tilde{F}_t = F_t^{\otimes N}$.

From $(\sigma^n, \xi)_{n \geqq 1}$, one can define the sequence (where $\tilde{w} = \{w_1, w_2, \cdots, w_n, \cdots\} \in \tilde{\Omega}$)

$$\tau^1(\tilde{w}) = \tau^1(w_1),$$

$$\tau^2(\tilde{w}) = \tau^1(w_1) + \tau^2(w_2) \circ \theta_{\tau^1(w^1)}, \quad \text{etc.} \cdots.$$

Then $y_t(\tilde{w}) = x_t(w_n)$ for $t \in [\tau^{n-1}, \tau^n[$, $(\tau^0 = 0)$.

It is shown in [11, Chap. 5], that for any $x \in E$ and $\nu = (\sigma^n, \xi)_{n \geqq 1}$ admissible, one can construct a probability measure P_x^ν on $(\tilde{\Omega}, \tilde{F})$ such that

$$P_x^\nu(y_0 = x) = 1,$$

$$P_x^\nu(y_{\tau^n} = \xi) = 1 \quad \forall n \geqq 1,$$

$$E_x^\nu[\phi(y_{\tau^{n-1}+t}) \chi_{\{\tau^n > \tau^{n-1}+t\}} | \tilde{F}_{\tau^{n-1}}] = E_\xi[\phi(x_t) \chi_{\{\sigma^n > t\}}], \quad P_x^\nu \text{ a.s.},$$

for any bounded measurable function ϕ (on E) (χ_B is the indicator function of the set B).

Intuitively, P_x^ν gives the behavior of the controlled process which is, in some sense, "Markovian" between τ^{n-1} and τ^n, with the semigroup $\phi(t)$, and which takes the value ξ at τ^n for all $n \geqq 1$. Then,

$$J_x^\alpha(\nu) = E_x^\nu\left[\int_0^\infty e^{-\alpha t} f(y_t)\, dt + \sum_{n \geqq 1} c(y_{\tau^n})\, e^{-\alpha \tau^n}\right]. \tag{2.8}$$

Now the problem without discount should be

Minimize $J_x(\nu)$, where

$$J_x(\nu) = \liminf_{T \uparrow \infty} \frac{E_x^\nu \int_0^T f(y_t)\, dt + \sum_{n \geqq 1} c(y_{\tau^n}) \chi_{\{\tau^n < T\}}}{T}. \tag{2.9}$$

When one can restrict σ^n to be such that $\sup_{x \in E} E_x \sigma^n < +\infty$, then it is enough to consider (if $\xi^n = \xi$)

$$J_x(\nu) = \liminf_{N \uparrow \infty} \frac{E_x \int_0^{\sigma^1} f(x_s) + c(x_{\sigma^1}) + \sum_{n=2}^N [E_\xi \int_0^{\sigma^n} f(x_s)\, ds + c(x_{\sigma^n})]}{E_x \sigma^1 + \sum_{n=2}^N E_\xi \sigma^n}. \tag{2.10}$$

3. Study of undiscounted problems.

3.1. Preliminary remarks. Because E is compact, $\phi(t)$ has at least one invariant probability measure; indeed, for any probability q on E,

$$Q^n(\Gamma) = \frac{1}{n}\left[\int_E q(dx) \int_0^n P_x(x_t \in \Gamma)\, dt\right] \text{ is a probability on } E,$$

Q^n converges weakly since E is compact, and any weak limit of Q^n is an invariant measure of Φ (see the Appendix).

We will study the asymptotic behavior of Π_α under the following additional assumption.[2]

There exist an invariant probability measure μ and positive constants B and γ such that

$$|P_x(x_t \in \Gamma) - \mu(\Gamma)| \leqq B e^{-\gamma t}, \tag{3.1}$$

(for any Borel set of E).

[2] We will see below some examples for which this assumption is satisfied.

This assumption implies that μ is the only invariant measure of $\Phi(t)$: indeed, if $\tilde{\mu}$ is any invariant measure,

$$\tilde{\mu}(\Gamma) = \tilde{\mu}\Phi(t)(\Gamma) \quad \text{(by definition)}$$

$$\text{implies } \tilde{\mu}\Phi(t)(\Gamma) \to \mu(\Gamma) \text{ when } t \to \infty, \text{ therefore } \tilde{\mu}(\Gamma) = \mu(\Gamma).$$

Let us state some other useful consequences of (3.1). The proofs are given in the Appendix.

LEMMA 3.1. *Under* (2.1), (2.3) *and* (3.1), *there exist positive constants* B_1, γ_1 *such that, for any bounded measurable* g *on* E,

$$\left|\phi(t)g(x) - \int_E g(x)\mu(dx)\right| \leqq B_1 e^{-\gamma_1 t}\|g\|.$$

LEMMA 3.2. *Under* (2.1), (2.3) *and* (3.1), *a necessary and sufficient condition for the equation* $-Au = f (f \in C)$ *to have a solution* (*in* D_A) *is that*

$$\int_E f(x)\mu(dx) = 0.$$

LEMMA 3.3. *Any two solutions of the equation* $-Au = f$ *differ from each other only by a constant.*

3.2. Asymptotic behavior of the linear equation. For $f \in C$, let u_α be the unique solution of

$$-Au_\alpha + \alpha u_\alpha = f, \qquad u_\alpha \in D_A. \tag{3.2}$$

It is known that

$$u_\alpha(x) = \int_0^\infty e^{-\alpha t}\phi(t)f(x)\,dt. \tag{3.3}$$

Let

$$v_\alpha(x) = u_\alpha(x) - \min u_\alpha.$$

LEMMA 3.4. *Under* (2.1), (2.3), (3.1) *we have, as* $\alpha \downarrow 0$,

$$\alpha u_\alpha \to \bar{f}, \qquad v_\alpha \to \bar{v},$$

the unique solution of

$$-A\bar{v} = f - \bar{f}, \qquad \min \bar{v} = 0, \quad \textit{where } \bar{f} = \int_E f(x)\mu(dx). \tag{3.4}$$

Proof. It is clear from (3.3) that $\|\alpha u_\alpha\|$ is bounded by $\|f\|$. Let $g_\alpha = f - \alpha u_\alpha$; then

$$(3.2) \Rightarrow -Aw = g_\alpha \text{ has a solution in } D_A.$$

Therefore by Lemma 3.2 we have $\int_E g_\alpha\,d\mu = 0$, and any solution of this equation can be written as

$$w = \int_0^\infty \phi(t)g_\alpha\,dt + c_\alpha.$$

Let $\tilde{v}_\alpha = \int_0^\infty \phi(t)g_\alpha\,dt$. Since $\|g_\alpha\| \leqq 2\|f\|$, we have $\|\tilde{v}_\alpha\| \leqq$ constant (independent of

α). In fact, by Lemma 3.1, we have

$$\|\phi(t)g_\alpha\| \leqq B_1 e^{-\gamma_1 t}\|g_\alpha\| \leqq 2B_1 e^{-\gamma_1 t}\|f\|.$$

Therefore the integral $\tilde{v}_\alpha$ is well defined.

Now v_α is also a solution of

$$-Aw = g_\alpha \tag{3.5}$$

(it is the solution of (3.5) such that $\min w = 0$); therefore, $v_\alpha = \tilde{v}_\alpha - \min \tilde{v}_\alpha$. Since $\tilde{v}_\alpha$ is uniformly bounded, so is v_α. Hence,

$$\alpha u_\alpha \to \lambda \quad \text{constant} \tag{3.6}$$

($\alpha v_\alpha \to 0$, so $\lim_{\alpha \downarrow 0} \alpha u_\alpha(x) = \lim_{\alpha \downarrow 0} \alpha \min u_\alpha$ for all $x \in E$). Moreover,

$$\int_E g_\alpha \, d\mu = 0 \Rightarrow \int \alpha u_\alpha \, d\mu = \int f \, d\mu = \bar{f};$$

therefore $\lambda = \bar{f}$.

Define now $\tilde{g}_\alpha = \bar{f} - \alpha u_\alpha$. We have $\int \tilde{g}_\alpha \, d\mu = 0$, and $\tilde{g}_\alpha \to 0$ uniformly when $\alpha \to 0$. Let

$$\tilde{v} = \int_0^\infty \phi(t)(f - \bar{f}) \, dt,$$

$$\bar{v} = \tilde{v} - \min \tilde{v}.$$

Then

$$\|v_\alpha - \bar{v}\| \leqq \int_0^\infty \|\phi(t)\tilde{g}_\alpha\| \, dt$$

$$\leqq \|\tilde{g}_\alpha\| \cdot \text{constant} \qquad \text{(via Lemma 3.1).}$$

Therefore, $v_\alpha \to \bar{v}$ in C, and the fact that $\bar{v}$ is a solution of $-Aw = f - \bar{f}$ is a consequence of Lemma 3.2. □

Remark 3.1. We also have $Av_\alpha \to A\bar{v}$ in C, and

$$\bar{f} = \lim_{T \uparrow \infty} \frac{\int_0^T \phi(t) f \, dt}{T},$$

since

$$\bar{v}(x) = \int_0^T \phi(t)(f - \bar{f}) \, dt + \phi(T)\bar{v},$$

$$\bar{f} = \frac{\int_0^T \phi(t) f \, dt}{T} + \frac{\phi(T)\bar{v} - \bar{v}}{T}.$$

3.3. Asymptotic behavior of some optimal stopping problems. Let

$$f_1 \in C, \qquad f_1 \geqq \beta > 0, \tag{3.7}$$

$$\psi_1 \in C.$$

We consider, in this section, the optimal stopping time problem

$$J_x^1(\tau) = E_x\left[\int_0^\tau f_1(x_s) \, ds + \psi_1(x_\tau)\chi_{\tau<\infty}\right], \tag{3.8}$$

$$v(x) = \inf_\tau J_x^1(\tau).$$

In fact, since $f_1 \geqq \beta > 0$, it is enough to consider the stopping times τ such that

$$E_x\tau \leqq \frac{\|\psi_1\|}{\beta}. \tag{3.9}$$

Indeed, we can restrict ourselves to τ such that

$$J_x^1(\tau) \leqq J_x^1(0) \leqq \psi_1 \leqq \|\psi_1\|,$$

and, since $f_1 \geqq \beta > 0$, we have

$$\beta E_x\tau \leqq \|\psi_1\|,$$

giving (3.9). Let us also define

$$v_\alpha(x) = \inf_\tau E_x\left[\int_0^\tau e^{-\alpha t} f_1(x_t)\, dt + e^{-\alpha\tau}\psi_1(x_\tau)\right].$$

We can now state the following:

THEOREM 3.1. *Under the assumptions* (2.1), (3.1), (3.7):

(i) *v is the maximum element of the set of functions h such that*

$$\begin{aligned} &h \in C, \\ &h \leqq \phi(t)h + \int_0^t \phi(s) f_1\, ds, \\ &h \leqq \psi_1. \end{aligned} \tag{3.10}$$

(ii) $\hat{\tau} = \inf(t, v(x_t) = \psi_1(x_t))$ *is an optimal solution of the problem* (3.8).

(iii) $v_\alpha \to v$ *uniformly.*

Proof. The proof will follow several steps. We first transform the problem to the situation where

$$f = \text{constant} > 0 \quad \text{and} \quad \psi \geqq 0, \qquad \text{(instead of } f_1, \psi_1\text{)}. \tag{3.11}$$

Indeed, let $\gamma = \inf \psi_1$, $\psi_2 = \psi_1 - \gamma \geqq 0$; we have

$$v(x) - \gamma = \inf_\tau E_x\left[\int_0^\tau f_1\, dt + \psi_2(x_\tau)\right].$$

Then let v^0 be the solution of

$$-Av^0 = f_1 - \bar{f}_1 \text{ with } \max v^0 = 0 \qquad \text{(see § 3.2):}$$

then we have

$$w(x) = v(x) - \gamma - v^0 = \inf_\tau E_x\left[\int_0^\tau \bar{f}_1\, dt + \psi(x_\tau)\right],$$

where $\psi = \psi_2 - v^0 \geqq 0$.

Since $f_1 \geqq \beta > 0$, we have $\bar{f}_1 = \int f_1\, d\mu \geqq \beta > 0$.

In the following, we will therefore study

$$\begin{aligned} &w(x) = \inf_\tau J_x(\tau) = \inf_\tau E_x\left[\int_0^\tau f\, dt + \psi(x_\tau)\right], \\ &f = \text{constant} > 0 \quad \text{and} \quad \psi \geqq 0, \quad \psi \in C, \end{aligned} \tag{3.12a}$$

and we will introduce

$$w_\alpha(x) = \inf_\tau E_x\left[\int_0^\tau f e^{-\alpha t}\,dt + e^{-\alpha\tau}\psi(x_\tau)\right]. \tag{3.12b}$$

For problems like w_α, we will need the following result proved in [11, Chap. 1]:

LEMMA 3.5. *Assume* (2.1), $f, \psi \in C$. *Then*:

(i) w_α *is the maximum element of the set of functions* h *such that*

$$h \in C,$$

$$h \leqq \psi,$$

$$h \leqq e^{-\alpha t}\phi(t)h + \int_0^t e^{-\alpha s}\phi(s) f\,ds,$$

and $\tau^\alpha = \inf(t, w^\alpha(x_t) = \psi(x_t))$ *is an optimal solution.*

(ii) *The equation*

$$w_\alpha^\varepsilon(x) = \int_0^\infty e^{-\alpha t}\phi(t)\left[f - \frac{1}{\varepsilon}(w_\alpha^\varepsilon - \psi)^+\right] dt \tag{3.13}$$

has a unique solution, $w_\alpha^\varepsilon \in C$, *which has the following interpretation*:

$$\begin{aligned} w_\alpha^\varepsilon(x) &= \inf_\nu J_x^{\varepsilon,\alpha}(\nu), \\ J_x^{\varepsilon,\alpha}(\nu) &= E_x \int_0^\infty e^{-\alpha t} \exp\left(-\frac{1}{\varepsilon}\int_0^t \nu_s\,ds\right)\left[f + \frac{1}{\varepsilon}\nu\psi\right] dt, \end{aligned} \tag{3.14}$$

where ν *is any adapted process with value in* $[0, 1]$.

(iii) $w_\alpha^\varepsilon \searrow w_\alpha$ *uniformly when* $\varepsilon \searrow 0$.

LEMMA 3.6. *Under the assumptions of Theorem* 3.1, *assume* $\psi \in D_A$. *Then*:

(i) w_α^ε *is increasing to a function* w^ε *when* $\alpha \searrow 0$.

(ii) $\|w^\varepsilon - w\| \leqq \varepsilon\|f - A\psi\|$.

Proof of Lemma 3.6. We have

$$0 \leqq w_\alpha^\varepsilon(x) \leqq J_x^{\varepsilon,\alpha}(1) \leqq \|\psi\| + \varepsilon f.$$

Clearly, under (3.11), w_α^ε is increasing when α decreases; therefore, $\lim_{\alpha\searrow 0} w_\alpha^\varepsilon(x) = w^\varepsilon(x)$ defines a lower semicontinuous function which also satisfies

$$0 \leqq w^\varepsilon(x) \leqq \|\psi\| + \varepsilon f.$$

Since $w_\alpha^\varepsilon(x) \leqq J_x^{\varepsilon,\alpha}(\nu)$ for all ν, we get

$$w^\varepsilon(x) \leqq J_x^\varepsilon(\nu), \tag{3.15}$$

since, when $\alpha \searrow 0$, for any ν we have

$$J_x^{\varepsilon,\alpha}(\nu) \nearrow J_x^\varepsilon(\nu) = E_x \int_0^\infty \exp\left(-\frac{1}{\varepsilon}\int_0^t \nu_s\,ds\right)\left[f + \frac{1}{\varepsilon}\nu\psi\right] dt$$

(which is eventually equal to $+\infty$).

Now let

$$\tau^\varepsilon = \inf(t \geqq 0, w^\varepsilon(x_t) \geqq \psi(x_t));$$

then $w_\alpha^\varepsilon \leqq w^\varepsilon$ implies $w_\alpha^\varepsilon(x_t) < \psi(x_t)$, for all $t \in [0, \tau^\varepsilon[$.

One can see from (3.13) that

$$e^{-\alpha t}w_\alpha^\varepsilon(x_t)+\int_0^t e^{-\alpha s}\left[f-\frac{1}{\varepsilon}(w_\alpha^\varepsilon-\psi)^+(x_s)\right]ds$$

is a martingale. Hence, using $w_\alpha^\varepsilon(x_t)<\psi(x_t)$ for $t\in[0,\tau^\varepsilon[$, we get

$$w_\alpha^\varepsilon(x)=E_x\left[\int_0^{t\wedge\tau^\varepsilon} e^{-\alpha s}f\,ds+e^{-\alpha t\wedge\tau^\varepsilon}w_\alpha^\varepsilon(x_{t\wedge\tau^\varepsilon})\right].$$

When $\alpha\searrow 0$, this becomes

$$w^\varepsilon(x)=E_x\left[\int_0^{t\wedge\tau^\varepsilon} f\,ds+w^\varepsilon(x_{t\wedge\tau^\varepsilon})\right]. \tag{3.16}$$

As for (3.9), one can see that

$$f\cdot E_x t\wedge\tau^\varepsilon\leqq\|\psi\|+\varepsilon f,$$

hence,

$$E_x\tau^\varepsilon\leqq\frac{\|\psi\|+\varepsilon f}{f}.$$

Now, since x_t is quasicontinuous from the left, namely

$$x_{t\wedge\tau^\varepsilon}\to x_{\tau^\varepsilon}P_x\quad\text{a.s. on }\{\tau^\varepsilon<+\infty\},\quad\text{as } t\to\infty,$$

using Fatou's lemma and the lower semicontinuity of w^ε we get

$$\begin{aligned}\liminf_{t\nearrow\infty}E_x w^\varepsilon(x_{t\wedge\tau^\varepsilon})&\geqq E_x\liminf_{t\nearrow\infty}w^\varepsilon(x_{t\wedge\tau^\varepsilon})\\&\geqq E_x w^\varepsilon(x_{\tau^\varepsilon}).\end{aligned}$$

Therefore, (3.16) becomes

$$w^\varepsilon(x)\geqq E_x\left[\int_0^{\tau^\varepsilon} f\,ds+w^\varepsilon(x_{\tau^\varepsilon})\right],$$

and since $w^\varepsilon(x_{\tau^\varepsilon})\geqq\psi(x_{\tau^\varepsilon})$,

$$w^\varepsilon(x)\geqq E_x\left(\int_0^{\tau^\varepsilon} f\,ds+\psi(x_{\tau^\varepsilon})\right)=J_x(\tau^\varepsilon),$$

and finally,

$$w^\varepsilon(x)\geqq w(x). \tag{3.17}$$

Now let τ be any stopping time (one can assume $E_x\tau\leqq\|\psi\|+f$, for instance), and let ν_τ be the process

$$\nu_\tau(t)=0\quad\text{if } t<\tau,$$

$$\nu_\tau(t)=1\quad\text{if } t\geqq\tau.$$

Then

$$J_x^\varepsilon(\nu_\tau)-J_x(\tau)=E_x\left[\int_\tau^\infty\exp\left(-\frac{1}{\varepsilon}(s-\tau)\right)\left(f+\frac{1}{\varepsilon}\psi\right)ds-\psi(x_\tau)\right]. \tag{3.18}$$

Now assume $\psi\in D_A$.

Then, using the Markov property, one can see that

$$\psi(x_{\tau+t})\exp\left(-\frac{1}{\varepsilon}t\right)-\int_0^t \exp\left(-\frac{1}{\varepsilon}s\right)\left[A\psi-\frac{1}{\alpha}\psi\right](x_{\tau+s})\,ds,$$

is a martingale w.r.t. $G_t = F_{\tau+t}$; therefore,

$$E_x\psi(x_\tau)=E_x\left[\int_0^\infty \exp\left(-\frac{1}{\varepsilon}s\right)\left(A\psi-\frac{1}{\varepsilon}\psi\right)(x_{\tau+s})\,ds\right],$$

or

$$E_x\psi(x_\tau)=E_x\left[\int_\tau^\infty \exp\left(-\frac{1}{\varepsilon}(s-\tau)\right)\left(A\psi-\frac{1}{\varepsilon}\psi\right)(x_s)\,ds\right].$$

Then (3.18) becomes

$$J_x^\varepsilon(\nu_\tau)-J_x(\tau)=E_x\left[\int_\tau^\infty \exp\left(-\frac{1}{\varepsilon}(s-\tau)\right)(f-A\psi)(x_s)\,ds\right],$$

$$J_x^\varepsilon(\nu_\tau)-J_x(\tau)\leqq\varepsilon\|f-A\psi\|. \tag{3.19}$$

Now using (3.15) and (3.17) we get

$$w\leqq w^\varepsilon\leqq w+\varepsilon\|f-A\psi\|. \quad \square$$

COROLLARY 3.6. *Under the assumptions of Theorem* 3.1, *we have* $w_\alpha(x)\nearrow w(x)$ *when* $\alpha\searrow 0$ (*and therefore* w *is lower semicontinuous*).

Proof. Clearly, $w_\alpha(x)$ is increasing when $\alpha\searrow 0$. Now, from the previous lemma, when $\psi\in D_A$,

$$\begin{aligned}|w_\alpha(x)-w(x)|&\leqq\|w_\alpha-w_\alpha^\varepsilon\|+|w_\alpha^\varepsilon(x)-w^\varepsilon(x)|+\|w^\varepsilon-w\|\\&\leqq 2\varepsilon\|f-A\psi\|+|w_\alpha^\varepsilon(x)-w^\varepsilon(x)|.\end{aligned}$$

Therefore, $\lim_{\alpha\searrow 0} w_\alpha(x)=w(x)$.

If now $\psi\in C$, since D_A is dense in C we can take $\psi^n\in D_A$, $\psi^n\to\psi$ in C. Let w^n, w_α^n be the cost functions corresponding to ψ^n; we clearly have

$$\|w_\alpha^n-w^n\|\leqq\|\psi^n-\psi\|,\qquad \|w-w^n\|\leqq\|\psi^n-\psi\|.$$

Therefore,

$$|w_\alpha(x)-w(x)|\leqq\|w_\alpha-w_\alpha^n\|+|w_\alpha^n(x)-w^n(x)|+\|w^n-w\|.$$

Taking successively $\alpha\to 0$ and $n\to\infty$, we obtain $w_\alpha(x)\nearrow w(x)$. $\square$

Remark 3.2. Although this was not used in the proof of Lemma 3.6, we can show that

$$w^\varepsilon(x)=\inf_\nu E_x\int_0^\infty \exp\left(-\frac{1}{\varepsilon}\int_0^t \nu_s\,ds\right)\left[f+\frac{1}{\varepsilon}\nu\psi\right]dt.$$

In fact, since

$$w^\varepsilon(x_t)+\int_0^t\left(f-\frac{1}{\varepsilon}(w^\varepsilon-\psi)^+\right)(x_s)\,ds$$

is a martingale,

$$\exp\left(-\frac{1}{\varepsilon}\int_0^t \nu_s\,ds\right)w^\varepsilon(x_t)+\int_0^t \exp\left(-\frac{1}{\varepsilon}\int_0^s \nu_\tau\,d\tau\right)\left[\left(f-\frac{1}{\varepsilon}(w^\varepsilon-\psi)^+ +\frac{1}{\varepsilon}\nu w^\varepsilon\right)\right]ds$$

is also a martingale.

Therefore, taking $\nu^\varepsilon(t)=0$ when $w^\varepsilon(x_t)<\psi(x_t)$, and $\nu^\varepsilon(t)=1$ when $w^\varepsilon(x_t)\geqq\psi(x_t)$, we have $(1/\varepsilon)(w^\varepsilon-\psi)^+=(1/\varepsilon)\nu^\varepsilon(w^\varepsilon-\psi)$. Hence,

$$w^\varepsilon(x)=E_x\left[\int_0^t \exp\left(-\frac{1}{\varepsilon}\int_0^s \nu_r^\varepsilon\,dr\right)\left[f+\frac{1}{\varepsilon}\nu^\varepsilon\psi\right]+\exp\left(-\frac{1}{\varepsilon}\int_0^t \nu_s^\varepsilon\,ds\right)w^\varepsilon(x_t)\right].$$

Since the first term on the right-hand side is increasing and must be bounded, we have $E_x \exp(-1/\varepsilon\int_0^t \nu_s\,ds)\to_{t\uparrow\infty}0$, and therefore, $w^\varepsilon(x)=J_x^\varepsilon(\nu^\varepsilon)$. We can conclude also that $w^\varepsilon\to w$ uniformly when $\psi\in C$ without $\psi\in D_A$, since, as in the previous corollary,

$$\|w^\varepsilon-w\|\leqq\|w^\varepsilon-w^{\varepsilon,n}\|+\|w^{\varepsilon,n}-w^n\|+\|w^n-w\|,$$
$$\|w^\varepsilon-w\|\leqq 2\|\psi^n-\psi\|+\|w^{\varepsilon,n}-w^n\|. \qquad \square$$

LEMMA 3.7. *Under the assumptions of Theorem* 3.1,

$$w\in C.$$

Proof. We will use the discrete time analogue of our stopping time problem to obtain the result.

Let $\Delta>0$, and define $y_m=x_{m\Delta}$, $G_m=F_{m\Delta}$; then (y_m, G_m, P_x) define a Markov chain with the transition probability $P(x,\Gamma)=P_x(x_\Delta\in\Gamma)$. We consider the stopping time problem

$$z(x)=\inf_{\tau\in\mathcal{T}}(g\tau+\psi(y_\tau)),$$

where $g=f\cdot\Delta$ and $\mathcal{T}$ is the set of G_m-stopping times with values in $\mathbb{N}$.

Another way to write $z(x)$ is the following: let

$$V_\Delta=\{\tau,\text{ stopping times w.r.t. the family }(F_{m\Delta}, m\in\mathbb{N}),\ \text{with values in }(k\Delta, k\in\mathbb{N})\}.$$

Then we have, clearly,

$$z(x)=\inf_{\tau\in V_\Delta}(f\tau+\psi(x_\tau)). \tag{3.20}$$

Now, from [3c] (one can also use [12]) we can see that z is a solution of the equation

$$z=\min\{\psi, g+Pz\}, \tag{3.21}$$

where $Pz(x)=\phi(\Delta)z(x)$ by the definition of $P(x,\Gamma)$. It can be shown that (3.21) has a *unique* solution obtained by the iterations

$$z^0=0,\qquad z^{k+1}=\min\{\psi, g+Pz^k\},$$

and we have, by induction,

$$0\leqq z^k\leqq z^{k+1}\leqq\psi,\qquad z^k\in C\quad\forall k.$$

Therefore $z^k(x)\nearrow z(x)$, the l.s.c., solution of (3.21) and the uniqueness is proved in [3c].

Now let us consider another iterative process. Define

$$Q\phi=\min\{\phi, g+P\phi\},$$

and

$$z^0=Q\psi=\min(\psi, g+P\psi),$$
$$z^{k+1}=Qz^k=\min(z^k, g+Pz^k).$$

We have $\psi \geqq z^0 \geqq 0$, $z^0 \in C$ and, by recurrence,

$$\psi \geqq z^k \geqq z^{k+1} \geqq 0, \qquad z^k \in C \quad \forall k.$$

Therefore, $z^k(x) \searrow z^*(x)$ which is upper semicontinuous. In order to show that $z^* = z$, we take

$$z'^0 = Qz, \qquad z'^k = Qz'^{k-1};$$

from (3.21), we have $z'^k = z$, for all k. But since $z \leqq \psi$, we have $z'^k \leqq z^k$, for all k. Therefore, we get $z \leqq z^*$. But since $z^* \leqq \psi$, we have also $z^* \leqq g + Pz^*$, which gives easily, for any stopping time $\tau \in \mathcal{T}$,

$$z^*(x) \leqq E_x[g\tau + \psi(x_\tau)],$$

which would mean $z^* \leqq z$. Therefore, $z^* = z$. We conclude that z is continuous. Now take $\Delta = 2^{-n}$, and denote by z^n the solution of (3.21) for $\Delta = 2^{-n}$. Then

$$z^n(x) = \inf_{\tau \in V_n} E_x[f\tau + \psi(x_\tau)],$$

where

$$V_n = \{\text{stopping times w.r.t. the family } (F_{k2^{-n}}, k \in \mathbb{N}), \text{ and taking values in } (m2^{-n}, m \in \mathbb{N})\}.$$

Then, since $V_{n+1} \supseteq V_n$, it is clear that z^n is a decreasing sequence and $0 \leqq z^n \leqq z^{n-1} \leqq \psi$, $z^n \in C$.

Now any stopping time in V_n (for all n), is an F_t stopping time; therefore

$$z^n \geqq w.$$

Let $z(x) = \lim_{n \nearrow \infty} z^n(x)$; then z is upper semicontinuous (3.22) and $z \geqq w$.

We will show $z = w$, which will give the lemma. We have

$$z^n \leqq fE_x\tau + E_x\psi(x_\tau) \qquad \forall \tau \in V_n$$

and, since $z^n \searrow z$,

$$z \leqq f \cdot E_x\tau + E_x\psi(x_\tau) \qquad \forall \tau \in \bigcup_n V_n$$

Let τ be any F_t stopping time with finite expectation (since it is enough to consider τ such that $E_x\tau \leqq K$ with K large enough); then τ is the decreasing limit of the sequence of F_τ stopping times

$$\tau^j = \sum_{m \geqq 1} m2^{-j} \chi_{[(m-1)2^{-j} \leqq \tau < m2^{-j}]}$$

(see, for example, [14]).

We have $\tau^j \in V_j$, and therefore

$$z \leqq fE_x\tau^j + E_x\psi(x_{\tau^j}).$$

Using the right continuity of x_t and $\tau^j \searrow \tau$, we get

$$z \leqq J_x(\tau) = fE_x\tau + E_x\psi(x_\tau).$$

Therefore $z \leqq w$, and with (3.22) we have $z = w$. Since w is l.s.c. and z is u.s.c., we have w continuous. □

LEMMA 3.8. *Under the assumptions of Theorem* 3.1,
(i) $w_\alpha \nearrow w$ *uniformly*,
(ii) $v_\alpha \to v$ *uniformly*.

Proof. Since $w_\alpha(x) \nearrow w(x)$ pointwise and w and w_α are continuous, (i) is clear. Define $v'_\alpha = v_\alpha - \gamma - v^0$. Recall that $\gamma = \inf \psi_1$ and v^0 is chosen such that

$$-Av^0 = f_1 - f, \quad (f = \bar{f}_1) \quad \text{and} \quad v^0 \leqq 0.$$

We have, for any τ,

$$E_x v^0(x_\tau)\, e^{-\alpha\tau} = v^0(x) + E_x \int_0^\tau e^{-\alpha s}[Av^0 - \alpha v^0]\, ds$$

and

$$v^0(x) = E_x \int_0^\tau e^{-\alpha s}(f_1 - f + \alpha v^0)\, ds + E_x\, e^{-\alpha\tau} v^0(x_\tau),$$

$$\gamma = E_x \int_0^\tau e^{-\alpha s}\alpha\gamma\, ds + E_s \gamma\, e^{-\alpha\tau}.$$

Therefore,

$$v'_\alpha = \inf_\tau E_x\left[\int_0^\tau e^{-\alpha s}(f - \alpha v^0 - \alpha\gamma)\, ds + e^{-\alpha\tau}\psi(x_\tau)\right].$$

Since $v^0 \leqq 0$ and since one can assume without loss of generality that $\gamma \leqq 0$, we see that it is enough to consider those τ such that

$$E_x \int_0^\tau e^{-\alpha s} f\, ds \leqq \|v'_\alpha\| \leqq \|\psi\| \quad \forall \alpha \leqq \alpha_0,$$

for α_0 small enough (and the same is true for w_α). Then

$$\|v'_\alpha - w_\alpha\| \leqq \alpha(\|v^0\| + |\gamma|) \cdot \|\psi\| \cdot \frac{1}{f}. \tag{3.23}$$

Therefore, since $v'_\alpha = v_\alpha - \gamma - v^0$ and $w = v - \gamma - v^0$,

$$\|v'_\alpha - w\| = \|v_\alpha - v\| \leqq \|v'_\alpha - w_\alpha\| + \|w_\alpha - w\|,$$

and we obtain (ii) from (i) and (3.23). □

The end of the proof of Theorem 3.1 is now strictly identical to the discounted case described in detail in [11, Chap. 1], first with w and then, by translation, with v.

Remark 3.3. Examining the proof of Theorem 3.1, one can see that instead of $f_1 \geqq \beta > 0$, one can slightly generalize in assuming only $\int f_1\, d\mu > 0$. On the other hand, if $f_1 \geqq \beta > 0$ and $\psi_1 \geqq 0$, the assumption (3.1) is not necessary to obtain the theorem. □

3.4. Asymptotic behavior of Π_α. Let us go back to the problem Π_α and define $v_\alpha = u_\alpha - \min u_\alpha$, which gives

$$v_\alpha = \inf_\tau E_x\left[\int_0^\tau e^{-\alpha t}(f(x_t) - \alpha \min u_\alpha)\, dt + e^{-\alpha\tau} c(x_\tau)\right]. \tag{3.24}$$

From (2.5), we have

$$0 \leqq v_\alpha \leqq c,$$

and therefore

$$\alpha u_\alpha \to \lambda = \text{constant independent from } \alpha. \tag{3.25}$$

One can notice that

$$0 \leqq \lambda \leqq \bar{f} \quad \text{since } u_\alpha \leqq \int_0^\infty e^{-\alpha t}\phi(t) f\, dt = u_\alpha^0.$$

In the following, we will have to consider separately the cases

$$\lambda < \bar{f} \quad \text{and} \quad \lambda = \bar{f}.$$

THEOREM 3.2. *Under the assumptions* (2.1), (2.3), (3.1) *and if* $\lambda < \bar{f}$, *then*:

(i) $v_\alpha \to \bar{v}$ *in* C, *as* $\alpha \downarrow 0$, *where*

$$\bar{v}(x) = \inf_\tau E_x\left[\int_0^\tau (f(x_s) - \lambda)\, ds + c(x_\tau)\right]. \tag{3.26}$$

(ii) $\hat{\tau} = \inf\,(t \geqq 0,\ \bar{v}(x_t) = c(x_t))$ *is an optimal solution for the problem* (3.26).

(iii) $\bar{v}$ *is the maximum element of the set of functions* h *such that*

$$\begin{aligned} &h \in C, \\ &h \leqq c, \\ &h \leqq \phi(t)h + \int_0^t \phi(s)(f - \lambda)\, ds; \end{aligned} \tag{3.27}$$

moreover,

$$\min \bar{v} = 0.$$

(iv)

$$\begin{aligned} &\lambda = \frac{E_{\bar{x}}[\int_0^{\hat{\tau}} f(x_s)\, ds + c(x_{\hat{\tau}})]}{E_x \hat{\tau}}, \quad \textit{when } \bar{x} = \arg\min \bar{v}, \\ &\lambda = \inf_{\tau, x}\left[\liminf_{t \uparrow \infty} \frac{E_x(\int_0^{t \wedge \tau} f(x_s)\, ds + c(x_{t \wedge \tau}))}{E_x(\tau \wedge t)}\right]. \end{aligned} \tag{3.28}$$

Proof. Let v^0 be the solution of $-Av^0 = f - \bar{f}$ such that $\max v^0 = 0$, and define

$$w = \bar{v} - v^0 = \inf_\tau E_x\left[\int_0^\tau (\bar{f} - \lambda)\, dt + \psi(x_\tau)\right], \tag{3.29}$$

where $\psi = c - v^0$. Since $v^0 \leqq 0$, we have $\psi \geqq 0$. Then we can use the results of Theorem 3.1 for this case; this immediately gives (ii) and (iii). Now let

$$w_\alpha = \inf_\tau E_x\left[\int_0^\tau e^{-\alpha t}(\bar{f} - \lambda)\, dt + e^{-\alpha\tau}\psi(x_\tau)\right].$$

Then from Theorem 3.1 $w_\alpha \nearrow w$ uniformly, and we define

$$v'_\alpha = v_\alpha - v^0 = \inf_\tau E_x\left[\int_0^\tau e^{-\alpha t}(\bar{f} - \alpha \min u_\alpha - \alpha v^0)\, dt + e^{-\alpha\tau}\psi(x_\tau)\right].$$

Now, since $\bar{f} - \lambda > 0$ and $\alpha \min u_\alpha \to \lambda$, for α_0 small enough, we have $\bar{f} - \alpha \min u_\alpha \geqq \delta >$

0 and $\bar{f}-\lambda \geqq \delta$ for some δ. Therefore, it is enough to consider τ such that

$$\delta E_x \int_0^\tau e^{-\alpha t}\, dt \leqq \|\psi\|.$$

Hence

$$\|v'_\alpha - w_\alpha\| \leqq (|\lambda - \alpha \min u_\alpha| + \alpha \|v^0\|) \cdot \|\psi\| \cdot \frac{1}{\delta},$$

and

$$\|v_\alpha - \bar{v}\| = \|v'_\alpha - w\| \leqq \|v'_\alpha - w_\alpha\| + \|w_\alpha - w\|,$$

which gives (i) of the theorem.

Proof of (iv). $\min v_\alpha = 0$ and $v_\alpha \to \bar{v}$ uniformly implies $\min \bar{v} = 0$; therefore there exists $\bar{x}$ with $\bar{v}(\bar{x}) = 0$. A consequence of (3.27) is that the process

$$\bar{v}(x_t) + \int_0^t (f(x_s) - \lambda)\, ds$$

is a submartingale. Therefore,

$$\bar{v}(x) \leqq E_x\left[\bar{v}(x_{\tau \wedge T}) + \int_0^{\tau \wedge T} f(x_s)\, ds\right] - \lambda E_x \tau \wedge T.$$

Since $\bar{v}(x) \leqq c(x)$ and $\bar{v} \geqq 0$, we get

$$\lambda \leqq \frac{E_x \int_0^{\tau \wedge T} f(x_t)\, dt + c(x_{\tau \wedge T})}{E_x \tau \wedge T} \quad \forall x, T, \tau.$$

Moreover, from the optimality of $\hat{\tau}$,

$$\bar{v}(x) = E_x \int_0^{\hat{\tau}} (f(x_t) - \lambda)\, dt + c(x_{\hat{\tau}}),$$

which completes the proof of (iv) when written for $x = \bar{x}$. □

Remark 3.4. One can also consider the case where the cost for changing the state from x to ξ is given by

$$c_1(x) + c_2(\xi), \qquad \text{(instead of } c(x)).$$

Assume that $c_2(x) + c_1(x) \geqq k > 0$, $c_1, c_2 \in C$, $c_2 \in D_A$. Let

$$c(x) = c_1(x) + c_2(x).$$

With that cost, the discounted cost function will be the maximum solution of

$$u_\alpha \leqq c_1(x) + \inf_\xi (c_2(\xi) + u_\alpha(\xi)),$$

$$u_\alpha \leqq e^{-\alpha t} \phi(t) u_\alpha + \int_0^t e^{-\alpha s} \phi(s) f(x)\, ds.$$

Now let $\tilde{u}_\alpha(x) = c_2(x) + u_\alpha(x)$ since $c_2 \in D_A$; we get

$$\tilde{u}_\alpha \leqq c(x) + \inf_\xi \tilde{u}_\alpha(\xi),$$

$$\tilde{u}_\alpha \leqq e^{-\alpha t} \phi(t) \tilde{u}_\alpha + \int_0^t e^{-\alpha s} \phi(s) \tilde{f}\, ds,$$

where $\tilde{f} = f - Ac_2 + \alpha c_2$. Since

$$\int_E \tilde{f}\, d\mu = \bar{f} + \alpha \bar{c}_2,$$

we get a problem similar to the one we have considered before.

Remark 3.5. (3.28) shows that λ is the minimal "cost per cycle" when the cycle is the period of time between two "replacements".

THEOREM 3.3. *Under the assumptions of Theorem* 3.2

$$\lambda = \inf_{\nu} J_x(\nu), \qquad (\textit{defined in } (2.9)),$$

and the following impulse control is optimal for the problem in (2.9):

$$\begin{aligned} &\sigma^n = \hat{\tau} \quad \forall n \geqq 1 \\ &\xi^n = \bar{x} = \arg\min \bar{v}. \end{aligned} \tag{3.30}$$

Proof. The proof is essentially identical to those of discounted impulse control problems (see [11, Chap. 5], [3a]); therefore we give only a brief outline of it.

From the optimality of $\hat{\tau}$ for (3.26), we have

$$\bar{v}(x) = E_x\left(\int_0^{\hat{\tau}} (f(x_s) - \lambda)\, ds + c(x_{\hat{\tau}})\right).$$

Since $\bar{v}(\bar{x}) = 0$, this is also

$$\bar{v}(x) = E_x\left[\int_0^{\hat{\tau}} (f(x_s) - \lambda)\, ds + c(x_{\hat{\tau}}) + \bar{v}(\bar{x})\right].$$

But

$$\bar{v}(\bar{x}) = E_{\bar{x}}\left[\int_0^{\hat{\tau}} (f(x_s) - \lambda)\, ds + c(x_{\hat{\tau}})\right].$$

Hence,

$$\lambda = \frac{E_x \int_0^{\hat{\tau}} f(x_s)\, ds + E_{\bar{x}} \int_0^{\hat{\tau}} f(x_s)\, ds + E_x c(x_{\hat{\tau}}) + E_{\bar{x}} c(x_{\hat{\tau}}) + \bar{v}(\bar{x})}{E_x \hat{\tau} + E_{\bar{x}} \hat{\tau}},$$

and more generally,

$$\lambda = \frac{E_x\left(\int_0^{\hat{\tau}} f(x_s)\, ds + c(x_{\hat{\tau}})\right) + nE_{\bar{x}}\left(\int_0^{\hat{\tau}} f(x_s)\, ds + c(x_{\hat{\tau}})\right)}{E_x \hat{\tau} + nE_{\bar{x}} \hat{\tau}},$$

when $n \to \infty$; we have (see (2.10)) $\lambda = J_x(\hat{\nu})$, where $\hat{\nu}$ is defined by (3.30).

To prove $\lambda \leqq J_x(\nu)$ for any admissible control, we use the fact that

$$\bar{v}(x) \leqq E_x \int_0^{\tau} (f(x_s) - \lambda)\, ds + c(x_\tau) \quad \forall \tau \geqq 0,$$

and since $\bar{v}(x) \geqq \min \bar{v} = 0$,

$$\bar{v}(x) \leqq E_x\left(\int_0^{\tau^1} (f(x_s) - \lambda)\, ds + c(x_{\tau^1}) + \bar{v}(\xi^1)\right),$$

for $\nu = (\tau^n, \xi^n)_{n \geqq 1}$; then we have to express $\bar{v}(\xi^1)$ with the same kind of inequality. We refer to [11, Chap. 5] for details. □

THEOREM 3.4. *Under the assumption of Theorem* 3.2, *for any* $0 \leqq \delta < \bar{f}$, *the set of functions w satisfying*

$$w \in C,$$

$$w \leqq c, \tag{3.31}$$

$$w \leqq \phi(t)w + \int_0^t \phi(s)(f-\delta)\,ds$$

has a maximum element $\bar{w}$ *given by*

$$\bar{w}(x) = \inf_{\tau} E_x\left[\int_0^{\tau} (f(x_s)-\delta)\,ds + c(x_\tau)\right]. \tag{3.32}$$

But there is only one value of δ, *namely,* $\delta = \lambda$, *such that* $\bar{w}$ *satisfies*

$$\min \bar{w} = 0.$$

Proof. By the argument already used for v_α, we can check that, since $\delta < \bar{f}$, it is enough to consider $E_x\tau \leqq K$ large enough. Then $\bar{w}$ will be the uniform limit of the discounted cost

$$w_\alpha = \inf_{\tau} E_x\left[\int_0^{\tau} e^{-\alpha s}(f(x_s)-\delta)\,ds + e^{-\alpha\tau}c(x_\tau)\right],$$

and therefore will be continuous and will be the maximum solution of (3.31). Now if $\min \bar{w} = 0$, the proof of Theorem 3.2 (iv) implies

$$\delta = \inf_{x,\tau}\left[\liminf_{T\uparrow\infty} \frac{E_x \int_0^{\tau\wedge T} f(x_s)\,ds + c(x_{\tau\wedge T})}{E_x\tau \wedge T}\right];$$

therefore, $\delta = \lambda$. □

We now investigate some conditions on the data (namely, $\phi(t)$, f, c) under which $\lambda < \bar{f}$. We will denote by $\bar{v}^0$ the unique solution of

$$-A\bar{v}^0 = f - \bar{f}, \quad \bar{f} = \int_E f\,d\mu, \quad \bar{v}^0 \in D_A, \quad \min \bar{v}^0 = 0$$

(see Lemma 3.4).

THEOREM 3.5. *Under the assumptions* (2.1), (2.3), (3.1), *and if, moreover, for any open set* $\mathcal{O} \subset E$, *the first hitting time* $\tau_{\mathcal{O}}$ *is such that*

$$\sup_{x\in E} E_x\tau_{\mathcal{O}} < +\infty, \tag{3.33}$$

then

$$\{\bar{v}^0 > c\} \neq \varnothing \Leftrightarrow \lambda < \bar{f}. \tag{3.34}$$

Proof. Sufficiency. If $\{\bar{v}^0 > c\} \neq \varnothing$, let $\bar{\tau}$ be the first hitting time of $\{\bar{v}^0 > c\}$. Using the properties of v_α, we have

$$v_\alpha(x) \leqq E_x\left(\int_0^{\bar{\tau}} e^{-\alpha s}(f(x_s) - \alpha \min u_\alpha)\,ds + e^{-\alpha\bar{\tau}}c(x_{\bar{\tau}})\right).$$

Since $v_\alpha(x) \geqq 0$, we get

$$E_x\int_0^{\bar{\tau}} e^{-\alpha s}\alpha \min u_\alpha\,ds \leqq E_x\left[\int_0^{\bar{\tau}} e^{-\alpha s}f(x_s)\,ds + e^{-\alpha\bar{\tau}}c(x_{\bar{\tau}})\right].$$

Since $E_x\bar{\tau} < +\infty$, one can go to the limit ($\alpha \to 0$) to get

$$\lambda E_x\bar{\tau} \leqq E_x\left(\int_0^{\bar{\tau}} f(x_s)\,ds + c(x_{\bar{\tau}})\right).$$

Then by the definition of $\bar{\tau}$

$$\lambda E_x\bar{\tau} < E_x\left[\int_0^{\bar{\tau}} f(x_s)\,ds + \bar{v}^0(x_{\bar{\tau}})\right].$$

Since $E_x\bar{v}^0(x_{\bar{\tau}}) = \bar{v}^0(x) - E_x\int_0^{\bar{\tau}}(f-\bar{f})\,ds$, we get

$$\lambda E_x\bar{\tau} < \bar{f}E_x\bar{\tau} + \bar{v}^0(x).$$

Since $c \geqq k > 0$, $\bar{x}$ (such that $\bar{v}(\bar{x}) = 0$ as before) is in $\{\bar{v}^0 < c\}$; therefore, $E_{\bar{x}}\bar{\tau} > 0$ and since $\bar{v}^0(\bar{x}) = 0$, we get $\lambda < \bar{f}$.

Necessity. Suppose $\lambda < \bar{f}$ and that $\bar{v}^0 \leqq c$ on E. Then $\bar{v}^0$ satisfies

$$\begin{aligned}
&\bar{v}^0 \in C,\\
&\min \bar{v}^0 = 0,\\
&\bar{v}^0 \leqq c,\\
&\bar{v}^0 = \phi(t)\bar{v}^0 + \int_0^t \phi(s)(f-\bar{f})\,ds.
\end{aligned}$$

Going back to the proof of Theorem 3.2 (iv), we conclude that

$$\bar{f} = \inf_{x,\tau}\left[\liminf_{T\uparrow\infty} \frac{E_x\int_0^{\tau\wedge T} f(x_s)\,ds + c(x_{\tau\wedge T})}{E_x\tau \wedge T}\right]$$

(with the optimum obtained for $\tau = +\infty$), which would imply $\lambda = \bar{f}$, contradicting the assumption. Therefore, $\{\bar{v}^0 > c\} \neq \varnothing$. □

We now investigate the case $\lim_{\alpha\downarrow 0} \alpha u_\alpha = \bar{f}$.

LEMMA 3.9. *Under the assumptions* (2.1), (2.3), (3.1), *then*

$$\bar{v}^0 \leqq c \text{ in } E \Rightarrow \lambda = \lim_{\alpha\to 0} \alpha u_\alpha = \bar{f}. \tag{3.35}$$

Proof. We have seen that $\lambda < \bar{f}$ implies $\{\bar{v}^0 > c\} \neq \varnothing$; therefore, the conclusion of the lemma is clear. (Notice that this does not involve the property (3.33) as the proof of necessity in Theorem 3.5.) □

The following result is also an easy consequence of the previous theorems:

LEMMA 3.10. *Under the assumption of Lemma* 3.9, *if the property* (3.33) *is satisfied, then*

$$\lambda = f \Rightarrow \bar{v}^0 \leqq c. \tag{3.36}$$

THEOREM 3.6. *Under the assumptions* (2.1), (2.3), (3.1), *and if moreover,* $\bar{v}^0 \leqq c$ *on* E, *then*

$$\bar{f} = \inf_{\nu} J_x(\nu),$$

and the policy "do nothing" is optimal.

Proof. Under the assumptions of the theorem, $\bar{v}^0$ satisfies

$$\bar{v}^0 \in C,$$

$$\min \bar{v}^0 = 0,$$

$$\bar{v}^0 = \phi(t)\bar{v}^0 + \int_0^t \phi(s)(f - \bar{f})\, ds,$$

$$\bar{v}^0 \leqq c,$$

and the proof is identical to the proof of Theorem 3.3. □

Remark. Of course a slightly different version of such a result is "if $\lambda = \bar{f}$ and (3.24) is satisfied then $\bar{f} = \inf J_x(\nu)$".

4. Examples.

4.1. Jump Markov processes. Let X be a jump Markov process defined by the rate of jump $a(x)$ and the law of jump $q(x, dy)$, such that

$$a \in C, \qquad a(x) \geqq \eta > 0 \quad \forall x,$$
$$\int_E q(x, dy) g(y) \in C \quad \forall g \text{ bounded, measurable on } E.^3 \tag{4.1}$$

The generator is given by

$$Ag(x) = a(x)\left[\int_E q(x, dy) g(y) - g(x)\right]. \tag{4.2}$$

Let us recall the following result (from Doob [4, p. 197]).

LEMMA 4.1. *Let $Q(x, \Gamma)$ be a probability transition function (Γ a Borel subset of E). Assume that there exists a measure ϕ such that $\phi(\Gamma) > 0$ for some Borel set Γ in $\bar{v}$, and that there exists $k > 0$, $\delta > 0$ such that*

$$Q^k(x, \Delta) > \delta\phi(\Delta) \quad \forall \Delta \subset \Gamma, \quad \forall x \in E. \tag{4.3}$$

Then there exists an invariant measure μ for Q such that

$$|Q^n(x, \Gamma) - \mu(\Gamma)| < (1 - \delta\phi(\Delta))^{(n/k)-1}. \tag{4.4}$$

Notice that, under the assumptions of the lemma, Q has only one invariant probability measure.

LEMMA 4.2. *Assume (4.1) and that $q(x, \cdot)$ satisfies the assumption of the previous lemma. Then there exist constants B, γ and an invariant measure μ such that, if $P(t, x, \Gamma)$ is the transition function of the jump Markov process,*

$$|P(t, x, \Gamma) - \mu(\Gamma)| \leqq B e^{-\gamma t}. \tag{4.5}$$

(As seen before, μ is the only invariant probability measure of $\phi(t)$.) Moreover, if $\tilde{\mu}$ is the only invariant probability measure of the kernel q, then

$$\mu(\Gamma) = \bar{a} \int_\Gamma \frac{1}{a(x)} \tilde{\mu}(dx). \tag{4.6}$$

Proof. First, let us prove that, under the previous hypothesis, there is a T such that

$$P(T, x, \Gamma) \geqq \delta' q^k(x, \Gamma) \quad \text{for some } \delta' > 0, \tag{4.7}$$

where k is the number involved in Lemma 4.1.

[3] It is assumed here that E is a metric compact space as before (not necessarily a finite set).

We have, for all T, Γ,

$$P(T, x, \Gamma) \geqq P_x(x\tau^k \in \Gamma, \tau^k < T, \tau^{k+1} > T),$$

and the right-hand side is equal to

$$Y = \int_0^T p_x^k(ds) \int_\Gamma q^k(x, dy) \int_{T-s}^\infty a(y)\, e^{-a(y)\sigma}\, d\sigma,$$

if $p_x^k(\delta) = P_x(\tau^k \in d\delta)$, $a(y) \leqq M$ implies

$$Y \geqq q^k(x, \Gamma) \cdot P_x(\tau^k < T) \cdot e^{-MT};$$

now

$$P_x(\tau^k < T) = 1 - P_x(\tau^k > T) \geqq 1 - \frac{E_x \tau^k}{T}.$$

But $a(y) \geqq \eta > 0$ for all y; therefore, $E_x\tau^k \leqq k/\eta$. Then, if $k/\eta T < 1$, we get

$$P_x(\tau^k < T) \geqq 1 - \frac{k}{\eta T};$$

hence

$$P(T, x, \Gamma) \geqq \left(1 - \frac{k}{\eta T}\right) \cdot e^{-MT} q^k(x, \Gamma),$$

and (4.7) is proved.

Therefore, applying Lemma 4.1 to $P(nT, x, \Gamma)$ we obtain that there exists an invariant probability μ for $P(T, x, \Gamma)$ such that

$$|P(nT, x, \Gamma) - \mu(\Gamma)| \leqq B\rho^n,$$

for some $0 \leqq \rho < 1$.

The end of the proof can be done in the same way as in Theorem 4 of Freidlin [6] using the continuity of $P(t, x, \Gamma)$ with respect to t.

Now, we know that μ invariant for ϕ is equivalent to

$$\int_E \mu(dx) Ag(x) = 0 \quad \forall g \in D_A.$$

Here, this means

$$\int_E \mu(dx) a(x) \left(\int q(x, dy) g(y) - g(x) \right) = 0,$$

i.e., $a \cdot \mu$ is invariant for q.

Since μ is the only invariant probability measure of q, and since $0 < \eta \leqq a(x) \leqq M$, we get (4.6) by normalizing $a \cdot \mu$. □

Therefore, the results of the previous sections can be applied.

4.2. Diffusion process with reflection. Let $\phi \in C_b^2(\mathbb{R}^d)$, and $\mathcal{O} = \{x \mid \phi(x) > 0\}$. It will be assumed that

(4.8) $\mathcal{O}$ is bounded and $|\nabla\phi| \geqq 1$ on $\partial\mathcal{O}$.

Let

$$a_{ij} \in C = C^0(\bar{\mathcal{O}}), \qquad i, j = 1, d,$$

such that

$$a_{ij} = a_{ji},$$

$$\sum_{i,j=1}^{d} a_{ij}(x)\xi_i\xi_j \geqq \alpha|\xi|^2 \quad \forall \xi \in \mathbb{R}^d, \quad \forall x \in \mathbb{R}^d, \tag{4.9}$$

$$b_i \in C, \qquad i = 1, d, \tag{4.10}$$

$$\begin{aligned} &d_i \text{ Lipschitz functions on } \bar{\mathcal{O}} \text{ such that} \\ &(d(x), \nabla\phi(x)) \geqq \beta > 0 \quad \forall x \in \partial\mathcal{O}. \end{aligned} \tag{4.11}$$

Denote by A the operator

$$Ag = \sum_{i,j=1}^{d} a_{ij}\frac{\partial^2 g}{\partial x_i\,\partial x_j} + \sum_{i=1}^{d} b_i\frac{\partial g}{\partial x_i}.$$

Then it is shown by Stroock–Varadhan [13] that there exist
• a unique probability measure on $(\Omega, \mathcal{F})$ (with the notation of § 2),
• a process ξ_t adapted to $\mathcal{F}_t$, continuous, (increasing only at those times t such that $x_t \in \partial\mathcal{O}$), such that, for any $g \in C^2(\bar{\mathcal{O}})$,

$$E_x g(x_t) - g(x) = E_x\left\{\int_0^t Ag(x_s)\,ds + \int_0^t (d, \nabla g)(x_s)\,d\xi_s\right\}. \tag{4.12}$$

THEOREM 4.1. *Under the assumptions* (4.8), (4.9), (4.10), (4.11):
(i) $(\Omega, \mathcal{F}_t x_t, P_x)$ *is a continuous* (*strong*) *Feller process.*
(ii) *For any* $f \in L^p(\mathcal{O})$, $p > d+1$, *the equations*

$$-Au + \alpha u = f, \qquad \sum d_i \cdot \frac{\partial u}{\partial x_i} = 0 \quad \text{on } \partial\mathcal{O}$$

have a unique solution in $W^{2,p}(\mathcal{O})$ *given by*

$$w(x) = E_x\int_0^\infty e^{-\alpha s} f(x_s)\,ds. \tag{4.13}$$

Proof. See Stroock–Varadhan [13] for (i) and [2] for (ii). (4.13) is a consequence of the generalized Ito formula as in Bensoussan–Lions [2]. See also Puterman [10]. □

LEMMA 4.3. *Under the assumptions of Theorem* 4.1, *if moreover* a_{ij} *is Hölder continuous, then there exists an invariant probability measure* μ *and constants* B *and* γ *such that*

$$|P(t, x, \Gamma) - \mu(\Gamma)| \leqq B\,e^{-\gamma t},$$

for any Borel subject of $E = \bar{\mathcal{O}}$.

Proof. See Kogan [7]. □

As before, this implies that μ is the unique invariant probability measure of $\phi(t)$. We are now going to look at a direct proof of results such as those of § 3 using partial differential inequalities.

Beginning with the problem Π_α of § 3, we have:

THEOREM 4.2. *Assume that* X *is the diffusion process described above with the hypothesis of Theorem* 4.1 *and with* $c \in W^{2,p}$, $p > d+1$. *Then* u_α *is the unique solution of the quasi-variational inequality*

$$
\begin{aligned}
&-Au_\alpha + \alpha u_\alpha \leqq f \quad \text{on } \mathcal{O},\\
&u_\alpha \leqq Mu_\alpha\\
&(-Au_\alpha + \alpha u_\alpha - f)(u_\alpha - Mu_\alpha) = 0,\\
&\left(d, \frac{\partial u_\alpha}{\partial x}\right)_{\partial\mathcal{O}} = 0, \qquad u_\alpha \in W^{2,p}(\mathcal{O}).
\end{aligned} \tag{4.14}
$$

Proof. We will only sketch the proof, since it is an easy example of methods developed in Bensoussan–Lions [2], [3a] (see also [11]).

Let us begin with the variational inequality, for $\psi \in W^{2,p}$,

$$
\begin{aligned}
&-Au_\alpha + \alpha u_\alpha \leqq f \quad \text{on } \mathcal{O}\\
&u_\alpha \leqq \psi,\\
&(-Au_\alpha + \alpha u_\alpha - f)(u_\alpha - \psi),\\
&\left(d, \frac{\partial u_\alpha}{\partial x}\right) = 0 \quad \text{on } \partial\mathcal{O}, \qquad u_\alpha \in W^{2,p}.
\end{aligned} \tag{4.15}
$$

From [3] and regularity results like Theorem 4.1, there is a unique solution to the penalized equations

$$
-Au_\alpha^\varepsilon + \alpha u_\alpha^\varepsilon + \frac{1}{\varepsilon}(u_\alpha^\varepsilon - \psi)^+ = f, \qquad \left(d, \frac{\partial u_\alpha^\varepsilon}{\partial x}\right) = 0 \quad \text{on } \partial\mathcal{O}.
$$

Then as in § 3, Lemma 3.5, we have

$$
\|u_\alpha^\varepsilon\|_C \leqq \text{constant}, \qquad \frac{1}{\varepsilon}(u_\alpha^\varepsilon - \psi)^+ \leqq \text{constant}.
$$

Then $\|Au_\alpha^\varepsilon\|_{L^p} \leqq$ constant (in fact, this is true in L_∞). Therefore, by estimates of Agmon–Douglis–Nirenberg [1], we get that $\|u_\alpha^\varepsilon\|_{W^{2,p}} \leqq$ constant, and this allows to go to the limit $\varepsilon \to 0$ in $W^{2,p}$ weakly to obtain (4.15). Uniqueness can be proved for example by the stochastic interpretation (using the generalized Ito's formula; see Bensoussan–Lions [2]).

The existence of a solution of (4.14) is shown by iteration as follows:

$$
\begin{aligned}
&-Au_\alpha^0 + \alpha u_\alpha^0 = f,\\
&-Au_\alpha^n + \alpha u_\alpha^n \leqq f,\\
&u_\alpha^n \leqq Mu_\alpha^{n-1},\\
&(-Au_\alpha^n + \alpha u_\alpha^n - f)(u_\alpha^n - Mu_\alpha^{n-1}) = 0,
\end{aligned}
$$

since $Mu_\alpha^{n-1} = c(x) + \inf_\xi u_\alpha^{n-1}(\xi) \in W^{2,p}$. And as for the variational inequality, one gets $W^{2,p}$ estimates which are uniform w.r.t. n, allowing the limit $n \to \infty$.

Again, uniqueness results from the stochastic interpretation. □

THEOREM 4.3. *Under the assumptions of Theorem* 4.2,

$$
\lim_{\alpha \to 0} \alpha u_\alpha = \lambda, \text{ constant},
$$

$$
v_\alpha = u_\alpha - \min u_\alpha \text{ converges to some } v \text{ in } W^{2,p} \text{ weakly when } \alpha \to 0,
$$

and

$$\left.\begin{aligned} &-Av+\lambda \leqq f, \\ &v \leqq c, \\ &(-Av+\lambda-f)(v-c)=0 \end{aligned}\right\} \quad \text{a.e. in } \mathcal{O}, \tag{4.16}$$

$$\left(d, \frac{\partial v}{\partial x}\right)_{\partial \mathcal{O}} = 0.$$

Proof. As in § 3, we have $0 \leqq v_\alpha \leqq c$; therefore $\alpha u_\alpha \to$ constant, denoted by λ. Moreover, the method used for Theorem 4.2 gives here

$$\|Av_\alpha^\varepsilon\|_{L^\infty} \leqq \text{constant, independent of } \varepsilon \text{ and } \alpha.$$

Therefore, v_α is bounded in $W^{2,p}$, and this allows us to go to the limit of

$$\begin{aligned} &-Av_\alpha+\alpha u_\alpha \leqq f, \\ &v_\alpha \leqq c, \\ &(-Av_\alpha+\alpha u_\alpha-f)(v_\alpha-c)=0, \\ &\left(d, \frac{\partial v_\alpha}{\partial x}\right)_{\partial \mathcal{O}} = 0, \end{aligned}$$

to obtain (4.16). □

Remark 4.1. *Extensions.* The previous results can be extended to

$$Mu(x) = C_1(x) + \inf_\xi [C_2(\xi) + u(\xi)].$$

But this can be reduced to the previous case by taking

$$\bar{u}(x) = u(x) + C_2(x), \qquad c(x) = C_1(x) + C_2(x).$$

A more general case would be

$$Mu(x) = \inf_\xi [c(x, \xi) + u(\xi)].$$

Then the previous method is not usable, but if, for example, $|c(x, \xi) - c(y, \xi)| \leqq k|x-y|$ uniformly in ξ, and if we assume that A can be put in divergence form, then we can obtain a quasi-variational inequality when $\alpha \to 0$, though in a weaker form ($v \in H^1(\mathcal{O})$). □

4.3. Spin flip process. This example is likely to be an academic one in the present context (namely, replacement-type problems). We refer to Liggett [9] for a detailed study of such kinds of processes.

Let S be a countable set and $E = \{0, 1\}^S$ with the product topology. Define for $\nu \in S$ and $\eta \in E$

$$\eta_\nu(x) = \begin{cases} \eta(x) & \text{if } x \neq \nu \\ 1-\eta(x) & \text{if } x = \nu, \quad x \in S. \end{cases}$$

For $\nu \in S$ define $\Delta_\nu : C \to C$ by

$$\Delta_\nu f(\eta) = f(\eta_\nu) - f(\eta)$$

and

$$C^1(E)=\left\{ f\in C,\ |||f|||=\sum_\nu \|\Delta_\nu f\|<\infty \right\}.$$

Given $c(x,\eta)\geqq 0$, continuous and bounded on $S\times E$ and defining A on $C^1(E)$ by

$$Af(\eta)=\sum_x c(x,\eta)\Delta_x f(\eta),$$

one can show (see [9]) that A generates a unique Markov semigroup, at least when

(4.17) $\{c(x,\eta),\ x\in S\}$ is a bounded set in $C^1(E)$.

Let

$$c(\nu)=\inf_\eta [c(\nu,\eta)+c(\nu,\eta_\nu)],$$

$$c=\inf_\nu c(\nu),$$

$$M=\sup_x \sum_{\nu\neq x} \|\Delta_\nu c(x,\eta)\|.$$

Then (see [9]):

THEOREM 4.4. *Assume that* $M\leqq \sup_x |||c(x)|||<+\infty$. *Then the closure of* A *generates a unique Markov semigroup* $\phi(t)$ *and, if* $M<c$, *then the corresponding processes are ergodic with unique invariant probability measure* μ *on* E *and*

$$\left\| \phi(t)g-\int g\,d\mu \right\| \leqq B\,e^{(M-c)t}|||g|||.$$

Therefore, the results of § 3 *can be applied.*

Appendix. Invariant measures of Feller processes. Let $\phi(t)$ be a semigroup on $C^0(E)$, corresponding to a transition probability $P(t,x,\Gamma)$, $\Gamma\in B(E)$, the Borel σ-field of E compact metric space.

LEMMA A.1. *Any weak limit of the sequence* $Q^n(\Gamma)=1/n\int_0^n P(t,x,\Gamma)\,dt$ (*which is relatively compact in the space of measures in* $(E,B(E))$ *since* E *is compact*), *is an invariant probability measure for* $\phi(t)$.

Proof. Recalling that

$$\mu\phi(t)=\mu \Leftrightarrow \int_E Af\,d\mu=0 \quad \forall f\in D_A$$

(which is a consequence of the Hille–Yoshida theorem), we have, for any $f\in D_A$,

$$\int_E Af\,dQ^n=\frac{1}{n}[\phi(n)g-g]$$

(since $\phi(n)g=g+\int_0^n \phi(t)Ag\,dt$). Therefore, when $n\to\infty$, the weak convergence of Q^n to μ gives $\int_E Af\,d\mu=0$, which gives the result. □

Remark. One can also take

$$\frac{1}{n}\int_0^n \int_E dq(x)\, P(t, x, \Gamma)\, dt$$

for any probability measure q on E. □

LEMMA A.2. *Assume that $\phi(t)$ has an invariant probability measure μ such that*

$$|P_x(x_t \in \Gamma) - \mu(\Gamma)| \leqq B\, e^{-\gamma t}, \tag{A1}$$

for some constants $B>0$, $\gamma>0$. Then there exist B_1, $\gamma_1>0$ such that, for any bounded measurable functions f on E,

$$\left|\phi(t)f - \int_E f\, d\mu\right| \leqq B_1\, e^{-\gamma_1 t}\|f\|.$$

Proof. (See, for example, Friedlin [6]). It is enough to take f such that $\|f\| = 1$; let

$$\gamma_i = \left\{y, \frac{i}{n} > f(y) \geqq \frac{i-1}{n}\right\}.$$

Then

$$\left|\phi(t)f(x) - \sum_{i=-n}^{+n} \frac{i}{n} P(t, x, \gamma_i)\right| < \frac{1}{n},$$

$$\left|\int f\, d\mu - \sum_{-n}^{+n} \frac{i}{n}\mu(\gamma_i)\right| < \frac{1}{n}$$

and

$$\begin{aligned}\left|\phi(t)f - \int f\, d\mu\right| &\leqq \left|\phi(t)f - \sum_{-n}^{+n} \frac{i}{n} P(t, x, \gamma_i)\right| \\ &\quad + \left|\sum_{-n}^{+n} \frac{i}{n} P(t, x, \gamma_i) - \sum_{-n}^{+n} \frac{i}{n}\mu(\gamma_i)\right| + \left|\sum_{-n}^{+n} \frac{i}{n}\mu(\gamma_i) - \int f\, d\mu\right|.\end{aligned}$$

Therefore, by (A1),

$$\left|\phi(t)f - \int f\, d\mu\right| \leqq \frac{2}{n} + 2nB\, e^{-\gamma t}.$$

Taking $n \sim e^{\gamma t/2}$, we get the result.

LEMMA A.3. *Assuming* (A1), *a necessary and sufficient condition for the equation*

$$-Au = f \qquad (f \in C)$$

to have a solution in D_A is that

$$\int f\, d\mu = 0.$$

Proof. Necessity (which does not use A1). If $-Au = f$ has a solution in D_A, then

for any μ invariant by $\phi(t)$

$$\int f\,d\mu = -\int Au\,d\mu = -\int \lim_{t\downarrow 0}\frac{1}{t}(\phi(t)u-u)\,d\mu$$

$$= -\lim_{t\downarrow 0}\frac{1}{t}\int [\phi(t)u-u]\,d\mu$$

$$= -\lim_{t\downarrow 0}\frac{1}{t}\left[\int \phi(t)u\,d\mu - \int u\,d\mu\right]$$

$$= 0.$$

Sufficiency. Assume $\int f\,d\mu = 0$ and (A1). Let us take

$$u(x) = \int_0^\infty \phi(t)f(x)\,dt. \tag{A2}$$

Since by (A1) $\|\phi(t)f\| \leqq B_1 e^{-\gamma_1 t}\|f\|$, (A2) is well defined and, since the integral is uniformly convergent, $u(x)$ is continuous. Then we have

$$\phi(h)u(x) = \int_0^\infty \phi(h)\phi(t)f(x)\,dt$$

$$= \int_0^\infty \phi(t+h)f(x)\,dt$$

$$= \int_h^{+\infty} \phi(t)f(x)\,dt.$$

This implies

$$\phi(h)u(x) - u(x) = -\int_0^h \phi(t)f(x)\,dt.$$

Since $f \in C$, $\phi(t)f(x)$ is continuous w.r.t. t; therefore $(1/h)\int_0^h \phi(t)f(x)\,dt \to f(x)$ when $h \to 0$. Therefore $-Au = f$. □

COROLLARY. $u(x) = \int_0^\infty \phi(t)f(x)\,dt$ *is a solution of*

$$-Au = f.$$

LEMMA A.4. *Under the assumption* (A1), *any two solutions of* $-Au = f$, *when* $\int f\,d\mu = 0$, *differ from each other only by a constant.*

Proof. It is enough to prove that if $Au = 0$, then $u =$ constant. But $Au = 0$ implies

$$\phi(t)u(x) = u(x), \quad \text{for any } t.$$

Then by (1), $\phi(t)u \to \int u\,d\mu$ when $t \to \infty$. Therefore, $u(x) = \int u\,d\mu =$ constant. □

Acknowledgment. The author would like to thank the referee for his useful comments.

REFERENCES

An extensive bibliography can be found in [2] and [3].

[1] S. AGMON, A. DOUGLIS AND L. NIRENBERG, *Estimates near the boundary for solutions of elliptic* PDE, Comm. Pure Appl. Math., 12 (1959) pp. 623–727.

[2] A. BENSOUSSAN AND J. L. LIONS, *Application des Inéquations Variationnelles au Contrôle Stochastique*, Dunod, Paris, 1978.

[3a] ———, *Contrôle Impulsionnel et Inéquations Quasi-variationnelles*, Dunod, to appear.
[3b] ———, *On the asymptotic behavior of the solution of variational inequalities*, Summer School on the Theory of Nonlinear Operator, Akademie Verlag, Berlin, 1978.
[3c] A. BENSOUSSAN, *Contrôle optimal des chaînes de Markov*, to appear.
[4] J. L. DOOB, *Stochastic Processes*, John Wiley, New York, 1953.
[5] E. B. DYNKIN, *Markov Processes*, Springer-Verlag, New York, 1965.
[6] M. J. FRIEDLIN, *Diffusion processes with reflection and problem with a directional derivative*, Theory Prob. Appl., 8 (1963), pp. 75–83.
[7] Y. A. KOGAN, *On optimal control of a nonterminating diffusion process with reflection*, Theory Prob. Appl., 14 (1969), pp. 496–502.
[8] J. M. LASRY, *Contrôle Stochastique Ergodique*, thesis, University of Paris IX, 1974.
[9] T. M. LIGGETT, *The stochastic evolution of infinite system of interacting particles*, Ecole d'été de probabilités St. Fleur, 1976, Lecture Notes in Mathematics, Vol. 598, Springer-Verlag, New York, 1977.
[10] M. PUTERMAN, *On the optimal control of diffusion processes*, TR n° 24, Stanford University, 1972.
[11] M. ROBIN, *Contrôle Impulsionnel des Processus de Markov*, Thèse d'état, Université de Paris IX, 1978.
[12] A. N. SHIRYAEV, *Statistical sequential analysis*, American Mathematical Society, Providence, RI, 1973.
[13] D. STROOCK AND S. VARADHAN, *Diffusion processes with boundary conditions*, Comm. Pure Appl. Math., 24 (1971), pp. 147–225.
[14] C. DELLACHERIE, *Capacités et processus stochastiques*, Springer-Verlag, Berlin, 1972.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 3, May 1981

0363-0129/81/1903-0003 $01.00/0

BOUNDARY CONTROLLABILITY OF HYPERBOLIC PARTIAL DIFFERENTIAL EQUATIONS*

TUNC GEVECI†

Abstract. It is shown that for a large class of first-order hyperbolic systems with constant coefficients, including strictly hyperbolic or symmetric systems, and where the boundary conditions result in well-posed mixed initial-boundary value problems, any initial state $u_0 \in H^1(\Omega)$ can be steered by boundary control to any final state $u_T \in H^1(\Omega)$ at any time $T > T_0$ (where $\Omega \subset \mathbb{R}^n$, $n \geqq 2$, is a bounded domain with smooth boundary and T_0 depends on the system and on Ω) provided that all rays escape to infinity.

1. Introduction. Russell [17] has obtained fairly comprehensive results on the boundary controllability of first-order symmetric hyperbolic systems in one space variable. A comparable theory for such systems in more than one space variable has yet to be developed. Some results are available for the wave equation [16], [17] and similar equations [5]. Clarke has obtained approximate controllability results for symmetric hyperbolic systems [2], and Littman has obtained exact controllability results for strictly hyperbolic systems with constant coefficients [9]. Our result is comparable to Littman's results, but the method is entirely different.

We use a technique due to Russell [16] and obtain controllability for a large class of homogeneous first-order hyperbolic systems with constant coefficients, including strictly hyperbolic or symmetric systems, subject to boundary conditions which result in well-posed mixed initial-boundary value problems.

It is shown that any initial state $u_0 \in H^1(\Omega)$ can be steered to any final state $u_T \in H^1(\Omega)$ at any time $T > T_0$, where $T_0 > 0$ depends on the system and on Ω, $\Omega \subset \mathbb{R}^n$, provided that all rays escape to infinity. We do not have to assume nonzero speeds of propagation, as is required in Littman's method [9].

In § 2 we state the controllability result and discuss our hypotheses. In § 3 we prove a theorem along the lines of [16]. We give a short and almost self-contained proof of the crucial lemma (Lemma 3.1) on the decay of the H^1-norm of the restriction to compact sets of the solution of the Cauchy problem with compactly supported H^1 initial data. We hope that the inclusion of the proof will be convenient for the reader, since at present such information can be extracted only from long, technical papers on the theory of hyperbolic differential equations [1], [10], [13]. In § 4 we discuss possible extensions to nonhomogeneous equations with variable coefficients.

2. Statement of the controllability theorem and the basic hypotheses. We consider a mixed initial-boundary value problem

$$\begin{aligned} &\frac{\partial u}{\partial t}(t, x) = \sum_{j=1}^{n} A_j \frac{\partial u}{\partial x^j}(t, x), \qquad (t, x) \in [0, T] \times \Omega, \\ &M(x)u(t, x) = g(t, x), \qquad (t, x) \in [0, T] \times \partial\Omega, \\ &u(0, x) = u_0(x), \qquad x \in \Omega, \end{aligned} \tag{2.1}$$

where $x = (x^1, x^2, \cdots, x^n) \in \mathbb{R}^n (n \geqq 2)$, $\Omega \subset \mathbb{R}^n$ is a bounded domain with smooth (say, C^∞) boundary $\partial\Omega$; $u(t, x)$ and $u_0(x)$ are $\mathbb{C}^k$-valued functions; $g(t, x)$ is a $\mathbb{C}^l$-valued

* Received by the editors November 8, 1979, and in revised form June 19, 1980. This work was partially supported by a grant from Control Data Corporation.

† Mathematics Division, National Research Institute for Mathematical Sciences of the CSIR, P.O. Box 395, Pretoria, South Africa.

function; each A_j, $j = 1, 2, \cdots, n$, is a $k \times k$ matrix with complex entries; and $M(x)$ is a smooth, $l \times k$ matrix-valued function defined on $\partial\Omega$.

We shall first state and discuss our hypotheses concerning (2.1).

(H.1) There exists a smooth invertible $k \times k$ matrix-valued function $\Gamma(\xi)$, $\xi = (\xi_1, \xi_2, \cdots \xi_n) \in \mathbb{R}_n \backslash 0$, homogeneous of degree 0 in ξ, such that, for each $\xi \in \mathbb{R}_n \backslash 0$,

$$\Gamma^{-1}(\xi) A(\xi) \Gamma(\xi) = \operatorname{diag}(\lambda_1(\xi), \lambda_2(\xi), \cdots, \lambda_k(\xi)), \tag{2.2}$$

where $A(\xi) = \sum_{j=1}^{n} \xi_j A_j$ and the eigenvalues $\lambda_j(\xi)$, $j = 1, 2, \cdots, k$ of $A(\xi)$ are real smooth functions of ξ and are homogeneous of degree 1.

Hypothesis (H.1) ensures that the Cauchy problem

$$\begin{aligned} &\frac{\partial u}{\partial t}(t, x) = \sum_{j=1}^{n} A_j \frac{\partial u}{\partial x^j}(t, x), \qquad (t, x) \in \mathbb{R}^{n+1}, \\ &u(0, x) = u_0(x), \qquad x \in \mathbb{R}^n \end{aligned} \tag{2.3}$$

leads to a strongly continuous group in $\operatorname{Hom}(H^s(\mathbb{R}^n))$ for any $s \in \mathbb{R}$, and is satisfied, for example, if the system is symmetric or strictly hyperbolic (i.e., if the eigenvalues $\lambda_j(\xi)$, $j = 1, 2, \cdots, k$ of $A(\xi)$ are real and distinct for $\xi \in \mathbb{R}_n \backslash 0$) [18].

(H.2) $$\sigma_{\min} := \min_{\substack{j=1,\cdots,k \\ |\xi|=1}} |\nabla \lambda_j(\xi)| > 0,$$

where

$$\nabla \lambda_j(\xi) = \left(\frac{\partial \lambda_j}{\partial \xi_1}, \frac{\partial \lambda_j}{\partial \xi_2}, \cdots, \frac{\partial \lambda_j}{\partial \xi_n} \right)$$

is the gradient of $\lambda_j(\xi)$.

(H.2) ensures that the decay property (Lemma 3.1) holds, and corresponds to the geometric condition, mentioned in the Introduction, that all rays escape to infinity. A ray corresponding to $\lambda_j(\xi)$ is given by

$$x(t; \lambda_j(\xi)) = x_0 + (t - t_0) \nabla \lambda_j(\xi), \qquad t \in \mathbb{R},$$

for some $t_0 \in \mathbb{R}$ and $x_0 \in \mathbb{R}^n$, so that the condition

$$\lim_{|t| \to \infty} |x(t; \lambda_j(\xi))| = \infty,$$

for each $\lambda_j(\xi)$, $\xi \in \mathbb{R}_n \backslash 0$, is equivalent to (H.2) since $\nabla \lambda_j(\xi)$, $j = 1, 2, \cdots, k$, are smooth functions, homogeneous of degree 0 in ξ.

If

$$G := \sum_{j=1}^{n} A_j \frac{\partial}{\partial x^j},$$

is assumed to be elliptic, i.e., each $\lambda_j(\xi) \neq 0$, $\xi \in \mathbb{R}^n \backslash 0$, $j = 1, 2, \cdots, k$, then (H.2) is automatically satisfied, since each $\lambda_j(\xi)$ is homogeneous of degree 1 in ξ and we have the Euler identity

$$\sum_{k=1}^{n} \xi_k \frac{\partial \lambda_j}{\partial \xi_k}(\xi) = \lambda_j(\xi), \qquad j = 1, \cdots, k.$$

Littman has to assume ellipticity in his paper [9]. The importance of (H.2) instead of the ellipticity of G as far as the asymptotic behavior of the solutions of (2.3) is concerned is discussed in a recent paper by Rauch [13].

The important special case of linearized shallow-water equations [11],

$$\frac{\partial}{\partial t}\begin{bmatrix} u \\ v \\ \varphi \end{bmatrix} = -\begin{bmatrix} U & 0 & 1 \\ 0 & U & 0 \\ \Phi & 0 & U \end{bmatrix}\frac{\partial}{\partial x}\begin{bmatrix} u \\ v \\ \varphi \end{bmatrix} - \begin{bmatrix} V & 0 & 0 \\ 0 & V & 1 \\ 0 & \Phi & V \end{bmatrix}\frac{\partial}{\partial y}\begin{bmatrix} u \\ v \\ \varphi \end{bmatrix},$$

where $0 < U^2 + V^2 < \Phi$, satisfies (H.1) and (H.2). If we denote the variables dual to x and y by ξ and η, respectively, the eigenvalues of

$$A(\xi, \eta) = -\xi\begin{bmatrix} U & 0 & 1 \\ 0 & U & 0 \\ \Phi & 0 & U \end{bmatrix} - \eta\begin{bmatrix} V & 0 & 0 \\ 0 & V & 1 \\ 0 & \Phi & V \end{bmatrix}$$

are

$$\lambda_1(\xi, \eta) = -(\xi U + \eta V),$$
$$\lambda_{2,3}(\xi, \eta) = -(\xi U + \eta V) \pm C\sqrt{\xi^2 + \eta^2}.$$

It is readily seen that

$$|\nabla\lambda_1(\xi, \eta)|^2 = U^2 + V^2 > 0,$$
$$|\nabla\lambda_{2,3}(\xi, \eta)|^2 \geqq (U^2 + V^2)\left(\frac{\sqrt{\Phi}}{\sqrt{U^2 + V^2}} - 1\right)^2 > 0,$$

for $\xi^2 + \eta^2 = 1$, whereas $\lambda_1(\xi, \eta)$ may be zero.

(H.3) The matrices $A_j (j = 1, 2, \cdots, n)$ and $M(x)$ are such that the mixed problem (2.1) can have at most one strong solution if $u_0 \in L^2(\Omega)$ and $g \in L^2([0, T] \times \partial\Omega)$.

For the precise definition of strong solutions and for the discussion of conditions under which (H.3) is valid, we refer the reader to [4], [6], [12], [14]. We emphasize that (H.3) is only a uniqueness hypothesis, and make the following remark concerning the way in which (H.3) will be utilized.

Remark 2.1. Assume u solves the Cauchy problem (2.3) with $u_0 \in H^1(\mathbb{R}^n)$. Since $u \in C^0(\mathbb{R}; H^1(\mathbb{R}^n))$ by (H.1), v, the restriction of u to $[0, T] \times \Omega$, is in $C^0([0, T); H^1(\Omega))$, so that we can define $g(t, \cdot)$ as the trace of $M(\cdot)v(t, \cdot)$ on $\partial\Omega$. We have that $g \in C^0([0, T]; H^{1/2}(\partial\Omega))$ since $M(\cdot)$ is smooth. Let $v_0 = u_0|_\Omega \in H^1(\Omega)$. A fortiori, v is a strong solution of the mixed problem (2.1) with L^2-data $\{v_0, g\}$, and the uniqueness hypothesis (H.3) ensures that it is the unique solution of (2.1) corresponding to the data $\{v_0, g\}$.

We can now state the controllability theorem.

THEOREM 2.1. *If* (H.1), (H.2) *and* (H.3) *are valid and* $T_0 > 0$ *is sufficiently large* ($T_0 = T_0(A_1, A_2, \cdots, A_n, \Omega)$), *then for any* $T > T_0$, $u_0 \in H^1(\Omega)$ *and* $u_T \in H^1(\Omega)$ *there exists* $g \in C^0([0, T]; H^{1/2}(\partial\Omega))$ *such that the solution of the mixed problem* (2.1) *with initial condition* u_0 *and boundary data* g *satisfies* $u(T, \cdot) = u_T(\cdot)$.

We shall prove Theorem 2.1 in the next section.

3. Proof of the controllability theorem. We first present the lemma on the decay of the H^1-norm.

LEMMA 3.1. *Under* (H.1) *and* (H.2), *and given* $\varphi \in C_0^\infty(\mathbb{R}^n)$ *there exists* $T_0' > 0$ ($T_0' = T_0'(A_1, A_2, \cdots, A_n, \varphi)$) *and, for* $|T| > T_0'$, $C_s(T) > 0$ ($C_s(T) = C_s(T; A_1, \cdots, A_n, \varphi)$, $s \in \mathbb{R}$) *such that*

$$\lim_{|T| \to \infty} C_s(T) = 0, \tag{3.1}$$

and for any $u_0 \in H^s(\mathbb{R}^n)$ *with* $\varphi(x) = 1$ *for all* x *in a neighborhood of* $\operatorname{supp}(u_0)$, *the solution* u *of the Cauchy problem,*

$$\frac{\partial u}{\partial t}(t, x) = \sum_{j=1}^{n} A_j \frac{\partial u}{\partial x^j}(t, x), \qquad (t, x) \in \mathbb{R}^{n+1}, \tag{3.2}$$

$$u(t_0, x) = u_0(x), \qquad x \in \mathbb{R}^n,$$

satisfies

$$\|\varphi(\cdot)u(T, \cdot)\|_{H^s(\mathbb{R}^n)} \leqq C_s(T - t_0)\|u_0\|_{H^s(\mathbb{R}^n)} \tag{3.3}$$

for $|T - t_0| > T_0'$.

Proof. We can assume $t_0 = 0$. We can also assume that $u_0 \in C_0^\infty(\mathbb{R}^n)$, since for general $u_0 \in H^s(\mathbb{R}^n)$ with $\operatorname{supp}(u_0) \subset \operatorname{int}\{x \in \mathbb{R}^n : \varphi(x) = 1\}$ we can find, by means of regularization, a sequence $u_0^{(j)} \in C_0^\infty(\mathbb{R}^n)$, with $\operatorname{supp}(u_0^{(j)}) \subset \operatorname{int}\{x \in \mathbb{R}^n : \varphi(x) = 1\}$, $j = 1, 2, \cdots$, and $\lim u_0^{(j)} = u_0$ in $H^s(\mathbb{R}^n)$. Since the Cauchy problem is well-posed in $H^s(\mathbb{R}^n)$, $\lim u^{(j)}(t, \cdot) = u(t, \cdot)$ in $H^s(\mathbb{R}^n)$ for each $t \in \mathbb{R}$, $u^{(j)}$ being the solution corresponding to $u_0^{(j)}$. Thus inequality (3.3) for $u_0^{(j)} \in C_0^\infty(\mathbb{R}^n)$ yields the inequality for u_0.

Denote the Fourier transform of $u(t, x)$ with respect to x by $\tilde{u}(t, \xi)$ and the Fourier transform of $u_0(x)$ by $\tilde{u}_0(\xi)$, $\xi \in \mathbb{R}_n$. From (3.2) (with $t_0 = 0$) we obtain

$$\frac{\partial \tilde{u}}{\partial t}(t, \xi) = iA(\xi)\tilde{u}(t, \xi), \qquad t \in \mathbb{R}, \quad \xi \in \mathbb{R}_n,$$

$$\tilde{u}(0, \xi) = \tilde{u}_0(\xi), \qquad \xi \in \mathbb{R}_n \tag{3.4}$$

in the notation of § 2. Let $\tilde{v}(t, \xi) = \Gamma^{-1}(\xi)\tilde{u}(t, \xi)$, so that, by (3.4),

$$\begin{aligned} \frac{\partial \tilde{v}}{\partial t}(t, \xi) &= i\Gamma^{-1}(\xi)A(\xi)u(t, \xi) \\ &= i\Gamma^{-1}(\xi)A(\xi)\Gamma(\xi)v(t, \xi) \\ &= i\Lambda(\xi)\tilde{v}(t, \xi), \end{aligned} \tag{3.5}$$

where $\Lambda(\xi) := \operatorname{diag}(\lambda_1(\xi), \lambda_2(\xi), \cdots, \lambda_k(\xi))$ by (H.1). From (3.5) we obtain

$$\tilde{v}(t, \xi) = e^{it\Lambda(\xi)}\tilde{v}(0, \xi),$$

and therefore

$$\tilde{u}(t, \xi) = \Gamma(\xi)\, e^{it\Lambda(\xi)}\Gamma^{-1}(\xi)\tilde{u}_0(\xi), \qquad \xi \in \mathbb{R}_n \backslash 0. \tag{3.6}$$

From (3.6) we obtain formally, with $\langle x, \xi \rangle$ denoting $\sum_{j=1}^n \xi_j x^j$,

$$\begin{aligned} u(t, x) &= \frac{1}{(2\pi)^n} \int e^{i\langle x, \xi \rangle}\Gamma(\xi)\, e^{it\Lambda(\xi)}\Gamma^{-1}(\xi)\tilde{u}_0(\xi)\, d\xi \\ &= \frac{1}{(2\pi)^n} \int \left(\int e^{i\langle x-y, \xi \rangle}\Gamma(\xi)\, e^{it\Lambda(\xi)}\Gamma^{-1}(\xi)u_0(y)\, dy \right) d\xi \\ &= \frac{1}{(2\pi)^n} \int \left(\int e^{i\langle x-y, \xi \rangle}\Gamma(\xi)\, e^{it\Lambda(\xi)}\Gamma^{-1}(\xi)\, d\xi \right) u_0(y)\, dy. \end{aligned} \tag{3.7}$$

This formal calculation is justified when the integrals are interpreted as oscillatory integrals, as defined by Hörmander [3]. In our case, corresponding to the equations with constant coefficients, the meaning of (3.7) in terms of oscillatory integrals coincides, of

course, with its meaning as the convolution of the distribution $\mathscr{F}_\xi^{-1}(\Gamma(\xi)\, e^{it\Lambda(\xi)}\Gamma^{-1}(\xi))$ ($\mathscr{F}_\xi^{-1}$ denoting the inverse Fourier transform) with u_0.

We have thus represented the distribution kernel $R(t; x, y)$, which is the Riemann function corresponding to the Cauchy problem (3.2) ($t_0 = 0$), by the oscillatory integral

$$R(t; x, y) = \frac{1}{(2\pi)^n} \int e^{i\langle x-y,\, \xi\rangle} \Gamma(\xi)\, e^{it\Lambda(\xi)} \Gamma^{-1}(\xi)\, d\xi. \tag{3.8}$$

From (3.7) and (3.8), since $\varphi(x) = 1$ for $x \in \operatorname{supp}(u_0)$, we obtain

$$\varphi(x) u(t, x) = \int \varphi(x)\varphi(y) R(t; x, y) u_0(y)\, dy. \tag{3.9}$$

Now, making the change of variable $\xi' = t\xi$, we have

$$R(t; x, y) = \frac{1}{(2\pi)^n |t|^n} \int e^{i\langle (x-y)/t, \xi'\rangle} \Gamma(\xi')\, e^{i\Lambda(\xi')} \Gamma^{-1}(\xi')\, d\xi', \tag{3.10}$$

since $\Lambda(\xi) = \operatorname{diag}(\lambda_1(\xi), \lambda_1(\xi), \cdots, \lambda_k(\xi))$ is homogeneous of degree 1 in ξ and $\Gamma(\xi)$ is homogeneous of degree 0 in ξ.

Letting

$$R(x) = \frac{1}{(2\pi)^n} \int e^{i\langle x, \xi\rangle} \Gamma(\xi)\, e^{i\Lambda(\xi)} \Gamma^{-1}(\xi)\, d\xi, \tag{3.11}$$

by (3.9) and (3.10) we have

$$\varphi(x) u(t, x) = \frac{1}{|t|^n} \int \varphi(x)\varphi(y) R\left(\frac{x-y}{t}\right) u_0(y)\, dy. \tag{3.12}$$

Assuming for the moment that $R(x)$ is infinitely differentiable for $|x| < \sigma_{\min}$ (see (H.2)), (3.12) yields

$$\|\varphi(\cdot) u(t, \cdot)\|_{H^s(\mathbb{R}^n)} \leqq \frac{C_s}{|t|^n} \|u_0\|_{L^2(\mathbb{R}^n)} \tag{3.13}$$

for any $s \in \mathbb{R}$, and $|t| > T_0' := \operatorname{diam}(\operatorname{supp} \varphi)/\sigma_{\min}$, since we then have $|(x-y)/t| < \sigma_{\min}$ for $x, y \in \operatorname{supp} \varphi$. (3.3) is, then, a special case of (3.13).

Since $\Lambda(\xi) = \operatorname{diag}(\lambda_1(\xi), \lambda_2(\xi), \cdots, \lambda_k(\xi))$, to show that $R(x)$ is C^∞ in $|x| < \sigma_{\min}$ we need to know that an oscillatory integral of the form

$$A(x) = \int e^{i(\langle x, \xi\rangle + \lambda(\xi))} a(\xi)\, d\xi, \tag{3.14}$$

where $a(\xi)$ is a (scalar) smooth function of ξ, homogeneous of degree 0, and $\lambda(\xi)$ is one of the eigenvalues of $A(\xi)$ (homogeneous of degree 1 in ξ), is C^∞ for

$$|x| < \sigma_{\min} \left(= \min_{\substack{j=1,\cdots,k \\ |\xi|=1}} |\nabla \lambda_j(\xi)| \right).$$

This is concluded immediately by appealing to Proposition 1.2.3 in Hörmander's paper [3], since $\phi(x, \xi) = \langle x, \xi\rangle + \lambda(\xi)$ has gradient $\nabla_\xi \phi(x, \xi) = x + \nabla\lambda(\xi)$, and $\nabla_\xi \phi(x, \xi) \neq 0$ if $|x| < \sigma_{\min}$ and $\xi \in \mathbb{R}_n \backslash 0$. □

We have thus established Lemma 3.1, and can now prove the following null-controllability result.

LEMMA 3.2. *Under* (H.1), (H.2) *and* (H.3) *we have null controllability for* $T > T_0$, $T_0 = T_0(A_1, \cdots, A_n, \varphi)$, *in the sense that*:

(A) *If* $u_0 \in H^1(\Omega)$, *there exists* $g \in C^0([0, T]; H^{1/2}(\partial\Omega))$ *such that the solution of the initial-boundary value problem* (2.1) *with data* $\{u_0, g\}$, *satisfies* $u(T, x) = 0$ *a.e. in* Ω.

(A′) *If* $u_T \in H^1(\Omega)$, *there exists* $g \in C^0([0, T]; H^{1/2}(\partial\Omega))$ *such that the solution* u *of* (2.1) *with* $u_0(x) = 0$ *a.e. in* Ω, *and boundary data* g *satisfies* $u(T, x) = u_T(x)$ *a.e. in* Ω.

We shall prove only (A). The reader will then readily appreciate that (A′) can be proved similarly, since in the statement of Lemma 3.1 $(T - t_0)$ may be positive or negative. This, in turn, is tied up with the fact that the Cauchy problem (2.3) leads not only to a semigroup but to a group.

Proof of Lemma 3.2(A). Since $\partial\Omega$ is smooth and Ω is bounded, there is a compact set K containing $\bar{\Omega}$ in its interior and a continuous linear extension operator $p : H^1(\Omega) \to H^1(\mathbb{R}^n)$ such that $(pu)(x) = u(x)$ a.e. in Ω and supp $(pu) \subset K$ for each $u \in H^1(\Omega)$ [8].

In particular, there exists a constant $C(\Omega) > 0$ such that

$$\|pu\|_{H^1(\mathbb{R}^n)} \leqq C(\Omega)\|u\|_{H^1(\Omega)}, \tag{3.15}$$

for each $u \in H^1(\Omega)$.

Let us fix two functions φ, ψ in $C_0^\infty(\mathbb{R}^n)$ such that $\varphi(x) = 1$ for each x in a neighborhood of K, $\psi(x) = 1$ for $x \in \Omega$ and supp $\psi \subset K$, and choose $T_0' > 0$ in accordance with Lemma 3.1. ($T_0' = T_0'(A_1, \cdots, A_n, \varphi)$ is therefore $T_0'(A_1, A_2, \cdots, A_n, \Omega)$.) We then define $L_T : H^1(\Omega) \to H^1(\Omega)$ for $T > T_0'$ as follows.

Given $u_0 \in H^1(\Omega)$, determine w as the solution of the Cauchy problem

$$\begin{aligned} &\frac{\partial w}{\partial t}(t, x) - Gw(t, x) = 0, \qquad (t, x) \in \mathbb{R}^{n+1}, \\ &w(0, x) = (pu_0)(x), \qquad x \in \mathbb{R}^n \end{aligned} \tag{3.16}$$

($G = \sum_{j=1}^n A_j \partial/\partial x^j$, and the equalities hold for almost all x).

We then determine z as the solution of the Cauchy problem

$$\begin{aligned} &\frac{\partial z}{\partial t}(t, x) - Gz(t, x) = 0, \qquad (t, x) \in \mathbb{R}^{n+1}, \\ &z(T, x) = \psi(x)w(T, x), \qquad x \in \mathbb{R}^n. \end{aligned} \tag{3.17}$$

Then

$$(L_T u_0)(\cdot) := z(0, \cdot)|_\Omega \in H^1(\Omega). \tag{3.18}$$

By applying Lemma 3.1 to (3.16) we obtain, using (3.15),

$$\begin{aligned} \|\varphi(\cdot)w(T, \cdot)\|_{H^1(\mathbb{R}^n)} &\leqq C(T)\|pu_0\|_{H^1(\mathbb{R}^n)} \\ &\leqq C(T)C(\Omega)\|u_0\|_{H^1(\Omega)}, \end{aligned}$$

and therefore

$$\|\psi(\cdot)w(T, \cdot)\|_{H^1(\mathbb{R}^n)} \leqq C(T)C(\Omega)C(\psi)\|u_0\|_{H^1(\Omega)}. \tag{3.19}$$

By applying Lemma 3.1 to (3.17) we obtain

$$\|\varphi(\cdot)z(0, \cdot)\|_{H^1(\mathbb{R}^n)} \leqq C(-T)\|\psi(\cdot)w(T, \cdot)\|_{H^1(\mathbb{R}^n)}. \tag{3.20}$$

Combining (3.19) and (3.20), and letting $C'(T) = C(T)C(-T)C(\Omega)C(\psi)$, we have

$$\|\varphi(\cdot)z(0, \cdot)\|_{H^1(\mathbb{R}^n)} \leqq C'(T)\|u_0\|_{H^1(\Omega)}, \tag{3.21}$$

with $\lim_{T\to\infty} C'(T) = 0$.

Since $\varphi(x)=1$ for $x\in\Omega$, by the definition of L_T (3.18) and by (3.21), if $T>T_0'$ is sufficiently large, say, $T>T_0$, we have

$$\|L_T u_0\|_{H^1(\Omega)} \leqq \alpha(T)\|u_0\|_{H^1(\Omega)}, \tag{3.22}$$

with $0<\alpha(T)<1$, for all $u_0\in H^1(\Omega)$ (notice that we indeed have $T_0=T_0(A_1,\cdots,A_n,\Omega)$).

By (3.22), for $T>T_0$, $I-L_T: H^1(\Omega)\to H^1(\Omega)$ has a bounded inverse. Given $u_0\in H^1(\Omega)$, determine $\hat{u}_0\in H^1(\Omega)$ such that $(I-L_T)\hat{u}_0=u_0$.

From the definition of L_T, replacing u_0 by $\hat{u}_0$ in (3.16), we see that $v:=w-z$ satisfies

$$\begin{aligned} &\frac{\partial v}{\partial t}(t,x)-Gv(t,x)=0, \qquad (t,x)\in\mathbb{R}^{n+1},\\ &v(T,x)=0, \qquad x\in\Omega,\\ &v(0,x)=(I-L_T)\hat{u}_0(x)=u_0(x), \qquad x\in\Omega, \end{aligned} \tag{3.23}$$

since $\varphi(x)=1$, $\psi(x)=1$ for $x\in\Omega$, and $(pu)(x)=u(x)$ a.e. in Ω, for all $u\in H^1(\Omega)$.

We now let $u=v|_{[0,T]\times\Omega}$, and let $g(t,\cdot)$ be the trace of $M(\cdot)v(t,\cdot)$ on $\partial\Omega$. By (H.3) (see Remark 2.1), u is the unique solution of the initial-boundary value problem (2.1) with initial value u_0 and boundary data g, and furthermore $u(T,x)=0$, $x\in\Omega$.

We have thus established Lemma 3.2, which immediately yields the controllability theorem, Theorem 2.1.

Proof of Theorem 2.1. By Lemma 3.2(A), if $T>T_0$, there exists $g_1\in C^0([0,T];H^{1/2}(\partial\Omega))$ such that v determined by

$$\begin{aligned} &\frac{\partial v}{\partial t}(t,x)-Gv(t,x)=0, \qquad (t,x)\in[0,T]\times\Omega,\\ &M(x)v(t,x)=g_1(t,x), \qquad (t,x)\in[0,T]\times\partial\Omega,\\ &v(0,x)=u_0(x), \quad x\in\Omega \end{aligned} \tag{3.24}$$

satisfies $v(T,x)=0$, $x\in\Omega$.

By Lemma 3.2(A'), there exists $g_2\in C^0([0,T];H^{1/2}(\partial\Omega))$ such that w determined by

$$\begin{aligned} &\frac{\partial w}{\partial t}(t,x)-Gw(t,x)=0, \qquad (t,x)\in[0,T]\times\Omega,\\ &M(x)w(t,x)=g_2(t,x), \qquad (t,x)\in[0,T]\times\partial\Omega,\\ &w(0,x)=0, \qquad x\in\Omega \end{aligned} \tag{3.25}$$

satisfies $w(T,x)=u_T(x)$, $x\in\Omega$.

Setting $u=v+w$ and $g=g_1+g_2$, we see from (3.24) and (3.25) that g is as required. □

Remark 2.2. We could have assumed the initial and final values to be in $H^s(\Omega)$ for any $s>\frac{1}{2}$, since the decay result is valid in any H^s-norm and the trace theorem can be applied for $s>\frac{1}{2}$.

4. Remarks concerning equations with variable coefficients. Let us consider the generalization of the mixed initial-boundary value problem (2.1) to nonhomogeneous

systems with variable coefficients,

$$\frac{\partial u}{\partial t}(t, x) = \sum_{j=1}^{n} A_j(x)\frac{\partial u}{\partial x^j}(t, x) + B(x)u(t, x), \tag{4.1}$$

where A_j, $j = 1, 2, \cdots, n$, and B are C^∞ matrix-valued functions with

$$A_j(x) = A_j^0 \quad \text{(a constant matrix)}, \quad j = 1, 2, \cdots, n,$$
$$B(x) = 0,$$

for $|x| > R$, for some $R > 0$.

Clearly, if the uniqueness of the solution of the mixed problem (H.3) is assumed, and if the decay result (Lemma 3.1) is valid with (2.1) replaced by (4.1), the controllability theorem (Theorem 2.1) holds in the more general case.

It is not to be expected, however, that we may be able to obtain a decay result for the perturbed system (4.1) under general hypotheses comparable to (H.1) and (H.2). Far reaching results concerning the asymptotic behaviour of the solution of the Cauchy problem for the perturbed system have been obtained by Rauch [13]. With a hyperbolicity assumption generalizing (H.1), an assumption generalizing (H.2) and guaranteeing that the singularities of the Riemann matrix are not trapped in any compact set, and the assumption that the group associated with the Cauchy problem is contractive in a space H (which is $L^2(\mathbb{R}^n)$ equipped with an inner product inducing a norm equivalent to the usual norm), Rauch proves the following result:

There exist at most finitely many $\omega_j \in \mathbb{R}$, $\omega_j \neq 0$, $j = 1, 2, \cdots, m$, for which there exists $\varphi \in H$, $\varphi \not\equiv 0$, with $G\varphi = i\omega_j\varphi$, where

$$G = \sum_{j=1}^{n} A_j \frac{\partial}{\partial x^j} + B.$$

The span H_j of such eigenfunctions is finite dimensional. Let H_0 be the closure of the kernel of G. Then, $H_j \perp H_k (j \neq k, j, k = 0, 1, \cdots, m)$. Let $H_b = H_0 \oplus H_1 \oplus \cdots \oplus H_m$ be the space of bound states, using the terminology of scattering theory, and let $H_s = H \ominus H_b$ be the space of scattering states. Denoting the propagator associated with the Cauchy problem by $P(t)$, the projection onto H_j by π_j, $j = 0, 1, \cdots, m$, and the projection onto H_s by π_s, we have

$$P(t)\varphi = P(t)(\pi_s\varphi) + \pi_0\varphi + \sum_{j=1}^{m} e^{i\omega_j t}\pi_j\varphi, \tag{4.2}$$

and

$$\|P(t)(\pi_s\varphi)\|_{H^s(B)} \leqq \frac{C}{\log t}\|\pi_s\varphi\|_H,$$

for any $\varphi \in H$ with supp $\varphi \subset B$, a ball in $\mathbb{R}^n$. H_s and H_b are invariant under $P(t)$.

This result shows clearly that the validity of a decay result is tied up with the nonexistence of eigenvalues and triviality of the kernel of G. Incidentally, Lemma 3.1 shows that, in the homogeneous constant-coefficient case that we have considered, $H_b = \{0\}$. Direct determination of H_b is difficult and general results are not available. In classical scattering theory perturbations of the Laplacian are examined in detail. The following sample result illustrates what can happen even in the case of apparently trivial perturbations:

Assume $V \in C_0^\infty(\mathbb{R}^n)$ $(n = 1, 2)$, $V \leqq 0$ and that V is not identically zero. For any $\lambda > 0$, $-\Delta + \lambda V$ has eigenvalues (even though $-\Delta$ has none) [15, p. 100].

Lax and Phillips have considered symmetric systems where G is elliptic [7]. They assume a unique continuation property which eliminates nonzero eigenvalues but does not guarantee the triviality of the kernel.

Let us finally remark that Rauch's result enables us to determine a subspace of controllable states, namely those $\varphi \in H^1(\Omega)$ for which $p\varphi \in H_s$ (p being the extension operator of § 3). Since H_s is invariant under $P(t)$, the proof of controllability holds for initial and final states having extensions in the space of scattering states.

5. Conclusion. We have extended and clarified the scope of the technique used by Russell in [16]. At the same time, the use of the concept of oscillatory integrals has simplified the implementation of the technique.

The method is based on the local decay property of the solution of the Cauchy problem and the discussion concerning the case of variable cofficients establishes the link with scattering theory. The only aspect of the mixed initial-boundary value problem that was needed was the uniqueness of the solution. In order to obtain results for general L^2-data and results concerning the case when control is exercised only on part of the boundary, the mixed problem itself has to be understood better in the case of more than one space dimension, in a way that is comparable to the case of one space dimension [17].

REFERENCES

[1] M. ATIYAH, R. BOTT AND L. GÅRDING, *Lacunas for hyperbolic differential operators with constant coefficients I*, Acta Math., 124 (1970), pp. 109–189.

[2] B. M. N. CLARKE, *Boundary controllability of linear symmetric hyperbolic systems*, J. Inst. Maths. Applics., 20 (1977), pp. 283–298.

[3] L. HÖRMANDER, *Fourier integral operators I*, Acta Math., 127 (1971), pp. 79–183.

[4] H. O. KREISS, *Initial boundary value problems for hyperbolic systems*, Comm. Pure Appl. Math., 23 (1970), pp. 277–298.

[5] J. LAGNESE, *Exact boundary controllability of a class of hyperbolic equations*, this Journal, 16 (1978), pp. 1000–1017.

[6] P. D. LAX AND R. S. PHILLIPS, *Local boundary conditions for dissipative symmetric linear differential operators*, Comm. Pure Appl. Math., 13 (1960), pp. 427–455.

[7] ———, *Scattering theory*, Rocky Mount. J. Math., 1 (1971), pp. 173–223.

[8] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications*, vol. I, Springer-Verlag, Berlin, 1972.

[9] W. LITTMAN, *Boundary control theory for hyperbolic and parabolic partial differential equations with constant coefficients*, Ann. Sc. Norm. Sup. Pisa Ser. IV, 5 (1978), pp. 567–580.

[10] D. LUDWIG, *Exact and asymptotic solutions of the Cauchy problem*, Comm. Pure Appl. Math., 13 (1960), pp. 473–508.

[11] J. OLIGER AND A. SUNDSTRÖM, *Theoretical and practical aspects of some initial boundary value problems in fluid dynamics*, SIAM J. Appl. Math., 35 (1978), pp. 419–446.

[12] J. RAUCH, *L_2 is a continuable initial condition for Kreiss' mixed problems*, Comm. Pure Appl. Math., 25 (1972), pp. 265–286.

[13] ———, *Asymptotic behavior of solutions to hyperbolic partial differential equations with zero speeds*, Comm. Pure Appl. Math., 31 (1978), pp. 431–480.

[14] J. RAUCH AND F. J. MASSEY III, *Differentiability of solutions to hyperbolic initial-boundary value problems*, Trans. Am. Math. Soc., 189 (1974), pp. 303–318.

[15] M. REED AND B. SIMON, *Methods of Modern Mathematical Physics, Vol. IV: Analysis of Operators*, Academic Press, New York, 1978.

[16] D. L. RUSSELL, *A unified boundary controllability theory for hyperbolic and parabolic partial differential equations*, Studies in Applied Math., 52 (1973), pp. 189–211.

[17] ———, *Controllability and stabilizability theory for linear partial differential equations: recent progress and open questions*, SIAM Rev., 20 (1978), pp. 639–739.

[18] F. TREVES, *Basic Linear Partial Differential Equations*, Academic Press, New York, 1964.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 3, May 1981

0363-0129/81/1903-0004 $01.00/0

GLOBAL AND ASYMPTOTIC CONVERGENCE RATE ESTIMATES FOR A CLASS OF PROJECTED GRADIENT PROCESSES*

J. C. DUNN†

Abstract. Projected gradient processes of the Goldstein–Levitin–Polyak type are considered for constrained minimization problems, $\min_\Omega F$, with Ω a convex set in a Hilbert space X and $F: X \to \mathbb{R}^1$ a differentiable functional. Global and local convergence theorems are established for a large class of these processes, including those generated with implicit step length rules proposed by Bertsekas and Goldstein. In this analysis, traditional uniform strong positivity conditions on the Hessian $\nabla^2 F$ are replaced by weaker pseudoconvexity conditions and growth conditions on F. When F has a unique minimizer in Ω, convergence rates are shown to depend on how rapidly the function $\gamma(\sigma) = \inf\{r = F(x) - F(\xi) | x \in \Omega, \|x - \xi\| \geqq \sigma\}$ grows with increasing $\sigma > 0$. If $\gamma(\sigma) \geqq B\sigma^\nu$ for some $B > 0$, the processes $\{F_n\}$ in question converge to $\inf F$ like $O(n^{-\nu/(\nu-2)})$, linearly, superlinearly, or in finitely many steps according to whether $\nu > 2$, $\nu = 2$, $2 > \nu > 1$, or $\nu = 1$. The growth properties of $\gamma(\sigma)$ are in turn dependent upon the structure of F, Ω and the norm on X. Close connections also exist here with a hierarchy of extremal types constructed in a recent study of conditional gradient algorithms, and with long-standing notions of singularity for constrained optimal control problems and unconstrained minimization problems on $\mathbb{R}^n$.

Introduction. The projected gradient methods of Goldstein [1] and Levitin and Polyak [2] are useful successive approximation techniques for certain constrained minimization problems of the general form

$$\min_{x \in \Omega} F(x), \tag{1.1}$$

with Ω a nonempty closed convex subset of a real Hilbert space X, and $F: X \to \mathbb{R}^1$ a Fréchet differentiable functional. These methods generate iterate sequences $\{x_n\}$ in Ω via the simple recursion

$$x_{n+1} = P(x_n - \alpha_n \nabla F_n), \qquad x_0 \in \Omega, \tag{1.2}$$

where ∇F_n is the gradient of F at x_n, $P(z)$ is the unique projection of $z \in X$ into Ω, i.e., the unique solution of

$$\|z - P(z)\| = \min_{y \in \Omega} \|z - y\|, \tag{1.3}$$

and where $\{\alpha_n\}$ is a sequence of nonnegative step lengths related in some suitable way to x_n and/or the iteration index n. The scheme (1.2) is embedded in a still larger class of *relaxed* projected gradient methods,

$$x_{n+1} = x_n + \omega_n(P(x_n - \alpha_n \nabla F_n) - x_n), \qquad x_0 \in \Omega, \quad \omega_n \in [0, 1], \tag{1.4}$$

treated by Demyanov and Rubinov in [3]. In a general way, (1.4) resembles the tangent manifold projection process of Rosen [4], [5] and the geodesic projection method of Luenberger [6]; all three methods reduce to classical steepest descent when $\Omega = X$; however, no one method contains any other for arbitrary $\Omega \subset X$. Only the basic recursion (1.2) is considered here.

Goldstein and Levitin and Polyak apparently were the first to formulate Hilbert space convergence theorems for (1.2) with simple open loop step length rules of the threshold type, viz.,

$$0 < \alpha \leqq \alpha_n \leqq a, \tag{1.5}$$

* Received by the editors August 9, 1979. This investigation was supported by the National Science Foundation under Research Grant ENG 78-03385.

† Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27650.

with a "sufficiently small." In particular, it is shown in [2] that $F_n - \inf_\Omega F = O(n^{-1})$, provided that Ω is closed convex and bounded, F is convex and bounded below on Ω, ∇F is Lipschitz continuous on Ω and

$$0 < \alpha \leqq \alpha_n \leqq a < \frac{2}{L}, \tag{1.6}$$

where L is a Lipschitz constant for ∇F. Moreover, if F is twice continuously differentiable with a Hessian $\nabla^2 F$ satisfying

$$\mu_{\max}\|y\|^2 \geqq \langle \nabla^2 F(x)y, y\rangle \geqq \mu_{\min}\|y\|^2 \tag{1.7}$$

for some $\mu_{\min} > 0$, all $y \in X$ and all x in Ω, then F has a unique minimizer ξ, and the iterates of (1.2) converge linearly to ξ, provided

$$\alpha_n = \alpha \tag{1.8a}$$

with

$$0 < \alpha < \frac{2}{L} \tag{1.8b}$$

and $L \geqq \mu_{\max}$. Unfortunately, in order to implement the threshold rule (1.6) or (1.8) one must first have a *value* for the Lipschitz constant L, and this may present serious difficulties in practice (see Polak's comments in [7]). Open loop step length rules of the type

$$\lim_{n\to\infty} \alpha_n = 0, \tag{1.9a}$$

$$\sum_{n=0}^{\infty} \alpha_n = \infty \tag{1.9b}$$

require no Lipschitz constants, but even when (1.7) holds, the corresponding sequences $\{x_n\}$ generated by (1.2) converge slowly [8], [9], [10]. The implicit line minimization rule treated in [3] and [11], namely

$$F(P(x_n - \alpha_n \nabla F_n)) = \min_{\alpha > 0} F(P(x_n - \alpha \nabla F_n)), \tag{1.10}$$

also requires no Lipschitz constants and in the unconstrained case, $\Omega = X$, is known to produce linearly convergent steepest descent sequences when (1.7) is satisfied [12], [13]; however, (1.10) has serious practical shortcomings of its own. Bertsekas elaborates on this last point in [14], and proposes an alternative implicit rule of a type devised originally by Armijo [15] and Goldstein [16] for unconstrained steepest descent. Goldstein investigates a closely related implicit scheme for (1.2) in [17]. When Ω is a Cartesian product of Euclidean hypercubes or balls (a common circumstance in optimal control problems [3][14]) the projection operation in (1.3) is virtually trivial and the Bertsekas–Armijo and Goldstein rules are then readily implemented for (1.2).

In Bertsekas' generalized Armijo rule, the step length α_n is determined by the condition

$$\alpha_n = (\beta_n)^{m_n} a_n, \tag{1.11a}$$

with m_n the least nonnegative integer m satisfying

$$F_n - F(P(x_n - (\beta_n)^m a_n \nabla F_n)) \geqq \delta_n \langle \nabla F_n, x_n - (\beta_n)^m a_n \nabla F_n \rangle, \tag{1.11b}$$

where a_n, β_n, and δ_n are specified numbers in $(0, \infty)$, $(0, 1)$ and $(0, 1)$, respectively. Actually, a_n, β_n and δ_n are set equal to constants for all n in the original version of (1.11) [14]; however, for reasons which will become apparent later on, certain practical advantages may accrue from allowing these parameters to vary during the course of the iteration. In the analogous Goldstein version of (1.11), the given numbers a_n and δ_n fall in $(0, \infty)$ and $(0, \frac{1}{2}]$ respectively. One puts $\alpha_n = a_n$ if

$$F_n - F(P(x_n - a_n \nabla F_n)) \geqq \delta_n \langle \nabla F_n, x_n - P(x_n - a_n \nabla F_n) \rangle, \tag{1.12a}$$

or else chooses any $\alpha_n \in (0, a_n)$ satisfying

$$\begin{aligned} (1 - \delta_n) \langle \nabla F_n, x_n - P(x_n - \alpha_n \nabla F_n) \rangle &\geqq F_n - F(P(x_n - \alpha_n \nabla F_n)) \\ &\geqq \delta_n \langle \nabla F_n, x_n - P(x_n - \alpha_n \nabla F_n) \rangle. \end{aligned} \tag{1.12b}$$

Both rules are feasible (i.e., (1.11) and (1.12) actually have solutions $\alpha_n \in (0, a_n])$ at a general vector x_n in a general convex set Ω, without continuity restrictions on ∇F. These points are developed at greater length in § 2.

By construction, the projected gradient sequences $\{x_n\}$ obtained from Bertsekas' rule (1.11) or Goldstein's rule (1.12) satisfy the condition

$$F_n - F_{n+1} \geqq \delta_n \langle \nabla F_n, x_n - x_{n+1} \rangle \tag{1.13}$$

with $\delta_n > 0$ for all $n \geqq 0$, and this is also true of the threshold rule (1.6) when ∇F is globally Lipschitz continuous on Ω. In all cases where (1.2) and (1.3) hold, there are two simple prerequisites for "fast" convergence of the sequence $\{F_n\}$, namely, that

$$\delta_n \geqq \delta \tag{1.14a}$$

and

$$\alpha_n \geqq \alpha \tag{1.14b}$$

for some positive numbers δ and α, and all $n \geqq 0$. It is shown in § 3 that increasing values of the product $\delta\alpha$ are associated with more rapidly convergent *upper bounds* on $F_n - \inf_\Omega F$; even more specifically, if one puts

$$\delta_\infty = \lim_{n \to \infty} \inf \delta_n \tag{1.15a}$$

and

$$\alpha_\infty = \lim_{n \to \infty} \inf \alpha_n, \tag{1.15b}$$

then larger values of the product $(\delta\alpha)_\infty$ are associated with better asymptotic convergence rates for certain *upper bounds* on $F_n - \inf_\Omega F$. Condition (1.14b) is explicit in the threshold rule (1.6), but (1.14a) must be *deduced* from (1.14b) and (1.6); this is indeed possible (see § 2). On the other hand, for (1.11) and (1.12) one *chooses* the parameters δ_n in accordance with (1.14a) and then tries to deduce (1.14b). When ∇F is Lipschitz continuous on Ω, and when $a_n = a > 0$, $\beta_n = \beta \in (0, 1)$ and $\delta_n = \delta \in (0, 1)$ for all n, Bertsekas establishes (1.14b) with

$$\alpha = \min \left\{ a, \frac{2\beta(1 - \delta)}{L} \right\},$$

where L is any positive Lipschitz constant for ∇F. His argument works equally well for variable parameters δ_n and a_n, and in this case one readily obtains (1.14b) with

$$\alpha = \min\left\{a, \frac{2\beta(1-d)}{L}\right\}, \tag{1.16}$$

where a is now a positive lower bound on the upper thresholds a_n, β is a positive lower bound on β_n and $d \in [\delta, 1)$ is an upper bound on δ_n. Similarly, for the Goldstein rule it is shown in § 2 that (1.14b) holds with

$$\alpha \geqq \min\left\{a, \frac{2\delta}{L}\right\}. \tag{1.17}$$

From (1.16) and (1.17) it is just a short step to analogous lower bounds on $\lim_{n\to\infty} \inf \alpha_n$ in (1.15b), except that now a, β, d and δ are replaced by $\lim_{n\to\infty} \inf a_n$, $\lim_{n\to\infty} \inf \beta_n$, $\lim_{n\to\infty} \sup \delta_n$ and $\lim_{n\to\infty} \inf \delta_n$ respectively, and the global Lipschitz constant L is replaced by a limit of local Lipschitz norms.

According to what has just been said, there are (at least) three different step length rules capable of generating projected gradient sequences satisfying (1.13) and (1.14) under reasonable continuity conditions on ∇F. In all such cases, it turns out that the convergence properties of $\{x_n\}$ and $\{F_n\}$ depend mainly on the quantities α, δ, α_∞ and δ_∞ and the local structure of Ω and F near minimizers of F in Ω. This justifies the broader viewpoint adopted in § 4, whose principal convergence theorems are formulated for the general class of projected gradient sequences $\{x_n\}$ satisfying (1.2), (1.3) and (1.14), with no explicit reference to specifics of the method used to generate the step lengths α_n, or to global Lipschitz continuity conditions on ∇F. When F is bounded below and continuously differentiable on the closed convex set Ω, it is shown in Theorem 3.1 that all such sequences $\{x_n\}$ have the descent property relative to F, and every subsequential limit point ξ of $\{x_n\}$ is an extremal of F in Ω; i.e.,

$$\langle \nabla F(\xi), x - \xi \rangle \geqq 0 \tag{1.18}$$

for all $x \in \Omega$. This extends Bertsekas' result in [14] for constant parameter sequences $a_n = a$, $\beta_n = \beta$ and $\delta_n = \delta$ (in particular, Theorem 4.1 applies to the case, $a_n \to \infty$, which is not susceptible to Bertsekas' method of proof). At this level of generality, F_n may converge to some limit $f > \inf F$ and $\{x_n\}$ need not converge at all. However, it is possible to say much more if F satisfies certain conditions of the pseudoconvexity type, e.g.,

$$F(x) - F(y) > 0 \Rightarrow \langle \nabla F(x), x - y \rangle \geqq \kappa(F(x) - F(y))$$

for some $\kappa \in (0, 1]$ and all $x, y \in \Omega$, or the still weaker requirement,

$$\inf_{\substack{x\in\Omega \\ \sigma \geqq F(x) > \inf_\Omega F}} \inf_{\substack{y\in\Omega \\ F(x) > F(y)}} \frac{\langle \nabla F(x), x - y \rangle}{F(x) - F(y)} > 0 \tag{1.19}$$

for all σ in $F(\Omega)$ with $\sigma > \inf_\Omega F$. When (1.19) holds, F is pseudoconvex and therefore strictly quasiconvex on Ω [18]; hence the level sets of F in Ω are convex and every extremal of F in Ω is automatically a global minimizer of F in Ω. Moreover, it is shown in Theorem 4.2 that $\{F_n\}$ always converges to $\inf_\Omega F$, that $F_n - \inf_\Omega F = O(n^{-1})$ if $\{x_n\}$ is bounded, and that $F_n - \inf_\Omega F = o(n^{-1})$ if $\{x_n\}$ converges. Still more can be said when F has a unique minimizer ξ; here everything depends on how fast $F(x)$ grows as x moves

away from ξ within Ω. If the uniform growth condition

$$0<\gamma(\sigma)=\inf_{\substack{x\in\Omega\\ \|x-\xi\|\geq\sigma}} F(x)-F(\xi) \tag{1.20}$$

holds for $\sigma>0$, then $\{x_n\}$ converges to ξ and consequently $F_n-\inf_\Omega F=o(n^{-1})$. More specifically, if $\gamma(\cdot)$ satisfies

$$\lim_{s\to 0^+}\left(\inf_{s\geq\sigma>0}\frac{\gamma(\sigma)}{\sigma^\nu}\right)>0 \tag{1.21}$$

for some $\nu>0$, then: (i) $F_n-\inf_\Omega F=O(n^{-\nu/(\nu-2)})$, or (ii) $F_n-\inf_\Omega F=O(\lambda^n)$ for some $\lambda\in[0,1)$, or (iii) $F_n=\inf_\Omega F$ for n sufficiently large, according to whether (i) $\nu>2$, or (ii) $2\geq\nu>1$, or (iii) $\nu=1$. Furthermore, if ∇F is locally Lipschitz continuous near ξ, then $\{x_n\}$ converges *superlinearly* to ξ for $1<\nu<2$. These results are established in Theorems 4.3, 4.4 and 4.5.

Close connections exist between the growth laws (1.20)–(1.21) and a hierarchy of extremal types constructed in recent investigations of the conditional gradient method [19], [20]. Convergence rates for conditional gradient sequences depend on how fast the *local linear approximation* $\langle\nabla F(\xi),\cdot\rangle$ to F grows near ξ; however, the counterparts of (1.20)–(1.21) for $\langle\nabla F(\xi),\cdot\rangle$ are purely manifestations of "curvature" in the boundary of Ω at ξ. On the other hand, for projected gradient sequences it is the growth rate of $F(\cdot)$ itself that matters, and (1.20)–(1.21) is jointly dependent on the local structure of $\partial\Omega$ *and* F near ξ. This difference between the two methods has important implications when the minimizers of F are singular. For projected gradient sequences, $\{F_n\}$ may converge *linearly* to $F(\xi)$ even though ξ is singular to first order (i.e., $\langle F(\xi),\cdot\rangle$ has multiple minimizers [9]); however, under the same circumstances an example of Cannon and Cullum [21] shows that one may expect no better than $O(n^{-1})$ convergence for conditional gradient methods. Furthermore, even if ξ is singular to second order (i.e., the local *quadratic* approximation to F at ξ has multiple minimizers [22]), the growth law (1.20)–(1.21) can still hold for some $\nu>2$. In such cases the corresponding $O(n^{-\nu/(\nu-2)})$ estimate fills a conspicuous gap in existing convergence theories for gradient processes, and suggests that the projected gradient algorithms treated here are superior to analogous conditional gradient methods for constrained minimization problems with singular solutions.

Taken as a whole, the theory developed in this article significantly extends the convergence analyses for projected gradient methods in [1], [2], [14] and earlier investigations of classical steepest descent processes in [23]–[27]. The present analysis reveals a *continuum* of increasingly singular extremal types for constrained and unconstrained minimization problems and shows clearly how the degree of singularity, as measured by growth rate of f within Ω near a minimizer ξ, affects the convergence behavior of gradient processes. All the convergence rate estimates presented here, including the standard linear convergence result, have been derived for a large class of nonconvex functionals and are therefore more broadly applicable than the classical results for convex functionals. Moreover, the $O(n^{-\nu/(\nu-2)})$ estimate for singular extremals with $\infty>\nu>2$ is completely new (even for unconstrained steepest descent), as are the superlinear convergence and finite termination theorems for constrained problems with $2>\delta\geq 1$. The present theory also has certain interesting implications for bounded optimal control problems, since a close connection exists between the growth rate of the objective functional F near an optimal control $\xi(\cdot)$, and the structure of the zero-crossing sets for $\xi(\cdot)$'s switching functions; this point is discussed briefly in § 4.

2. Preliminaries. This section records a number of basic facts which have a bearing on the analysis of (1.2) in § 4.

Let Ω be a nonempty closed convex subset of a real Hilbert space X with inner product $\langle \cdot, \cdot \rangle$ and associated norm $\| \cdot \|$. Suppose that $F: X \to \mathbb{R}^1$ is Fréchet differentiable on Ω, with gradient ∇F. Then ξ is a minimizer of F in Ω only if ξ is an extremal of F in Ω, i.e., ξ satisfies (1.18); moreover, if F is convex and ξ is an extremal, then ξ is a minimizer [3].

For $x \in X$ let $P(x)$ denote the projection of x into Ω. Since $\| \cdot \|^2$ is convex and Fréchet differentiable with gradient $\nabla(\|u\|^2) = 2u$, condition (1.3) holds if and only if for all $z \in \Omega$,

$$\langle x - P(x), z - P(x) \rangle \leqq 0. \tag{2.1}$$

It follows immediately from the Schwarz inequality and (2.1) that $P: X \to \Omega$ is non-expansive; thus,

$$\begin{aligned} \|x - y\| \|P(x) - P(y)\| &\geqq \langle x - y, P(x) - P(y) \rangle \\ &\geqq \|P(x) - P(y)\|^2, \end{aligned}$$

and consequently

$$\|P(x) - P(y)\| \leqq \|x - y\| \tag{2.2}$$

for all $x, y \in X$. In geometric terms, (2.1) states that $x - P(x)$ falls in the cone of normals to Ω at $P(x)$; i.e.,

$$x - P(x) \in K_\Omega(P(x)), \tag{2.3}$$

where

$$K_\Omega(u) = \{w \in X | \langle w, v - u \rangle \leqq 0, \forall v \in \Omega\}. \tag{2.4}$$

For $\alpha > 0$ and $x \in \Omega$, put

$$x(\alpha) = P(x - \alpha \nabla F(x)).$$

As an immediate corollary of (2.1), one finds that

$$\alpha \langle \nabla F(x), z - x(\alpha) \rangle \geqq \langle x(\alpha) - x, x(\alpha) - z \rangle \tag{2.5}$$

for all $z \in \Omega$; in particular, for $z = x$ this yields

$$\langle \nabla F(x), x - x(\alpha) \rangle \geqq \frac{1}{\alpha} \cdot \|x(\alpha) - x\|^2 \geqq 0. \tag{2.6}$$

It follows from (2.5) and (2.6) that x is an extremal of F in Ω if and only if $x = x(\alpha)$; i.e.,

$$x = P(x - \alpha \nabla F(x)) \tag{2.7}$$

for $\alpha > 0$. Moreover, if x is *not* an extremal, the directional derivative $\langle \nabla F(x), x(\alpha) - x \rangle$ is *strictly negative* at x for all $\alpha > 0$; this suggests, but does not immediately prove, the important descent property,

$$F(x(\alpha)) - F(x) < 0, \tag{2.8}$$

for small $\alpha > 0$ and x not an extremal. When $x \in \operatorname{Int} \Omega$ and α is small, one has $x(\alpha) = P(x - \alpha \nabla F(x)) = x - \alpha \nabla F(x)$, and therefore

$$F(x) - F(x(\alpha)) = \langle \nabla F(x), x - x(\alpha) \rangle + o(\|x - x(\alpha)\|) = \alpha \|\nabla F(x)\|^2 + o(\alpha).$$

Under these circumstances,

$$\lim_{\alpha \to 0^+} \frac{F(x)-F(x(\alpha))}{\langle \nabla F(x), x-x(\alpha)\rangle} = 1, \tag{2.9}$$

provided $\nabla F(x) \neq 0$, in which case (2.8) certainly holds for sufficiently small positive α. On the other hand, when x is on the boundary of Ω the issue is less easily resolved. If x is not an extremal, one still has

$$\begin{aligned} \left| \frac{F(x)-F(x(\alpha))}{\langle \nabla F(x), x-x(\alpha)\rangle} - 1 \right| &\leqq \frac{o(\|x-x(\alpha)\|)}{\langle \nabla F(x), x-x(\alpha)\rangle} \\ &\leqq \frac{\alpha \cdot o(\|x-x(\alpha)\|)}{\|x-x(\alpha)\|^2}, \end{aligned} \tag{2.10}$$

in view of (2.6). Moreover, for $0 \leqq \alpha_1 \leqq \alpha_2$ it follows from (2.2) that

$$\|x(\alpha_2)-x(\alpha_1)\| \leqq \|\nabla F(x)\| \cdot |\alpha_2-\alpha_1|, \tag{2.11}$$

and therefore

$$\lim_{\alpha \to 0^+} x(\alpha) = x(0) = x. \tag{2.12}$$

Consequently (2.9) follows from (2.10) if

$$\frac{\alpha}{\|x-x(\alpha)\|} = O(1),$$

a condition which is automatically fulfilled when Ω is a "simple" convex set [17] and x is not an extremal. Actually, (2.9) holds on *any* convex set Ω when x is not an extremal. This can be seen by using (2.10) and (2.11) to obtain

$$\left| \frac{F(x)-F(x(\alpha))}{\langle \nabla F(x), x-x(\alpha)\rangle} - 1 \right| \leqq \frac{\alpha}{\langle \nabla F(x), x-x(\alpha)\rangle} \cdot \frac{o(\alpha)}{\alpha}.$$

Evidently, (2.9) will follow at once if the positive quantity $\langle \nabla F(x), x-x(\alpha)\rangle/\alpha$ is bounded away from zero as $\alpha \to 0^+$. Therefore, suppose that this last condition does *not* hold; i.e., suppose that

$$\lim_{k \to \infty} \frac{\langle \nabla F(x), x-x(\alpha_k)\rangle}{\alpha_k} = 0$$

for some sequence of positive α_k's converging to zero. With (2.5), (2.6) and (2.11) one then obtains

$$\begin{aligned} \langle \nabla F(x), z-x\rangle &\geqq -\langle \nabla F(x), x-x(\alpha_k)\rangle - \frac{\|x-x(\alpha_k)\|}{\alpha_k} \cdot \|z-x(\alpha_k)\| \\ &\geqq -\alpha_k \|\nabla F(x)\|^2 - \|z-x(\alpha_k)\| \cdot \left(\frac{\langle \nabla F(x), x-x(\alpha_k)\rangle}{\alpha_k} \right)^{1/2} \end{aligned}$$

for all $z \in \Omega$ and all k, and thus

$$\langle \nabla F(x), z-x\rangle \geqq 0$$

for all $z \in \Omega$, which means that x is an extremal. This contradiction establishes (2.9) (and as a consequence, (2.8)) at nonextremal points x in arbitrary convex sets Ω.

As yet, no continuity conditions have been imposed on ∇F. When ∇F is locally Lipschitz continuous, it is possible to improve the estimate (2.10) by observing that

$$
\begin{aligned}
|F(y)-F(x)-\langle \nabla F(x), y-x\rangle| &= \left|\int_0^1 \frac{d}{d\sigma} F(x+\sigma(y-x))\, d\sigma - \langle \nabla F(x), y-x\rangle\right| \\
&\leqq \left|\int_0^1 \langle \nabla F(x+\sigma(y-x))-\nabla F(x), y-x\rangle\, d\sigma\right| \\
&\leqq \int_0^1 L\|y-x\|^2 \sigma\, d\sigma \\
&= \frac{1}{2} L\|y-x\|^2,
\end{aligned}
\tag{2.13}
$$

for some $L>0$ and all y sufficiently near x. Together, (2.6), (2.12) and (2.13) produce

$$
\left|\frac{F(x)-F(x(\alpha))}{\langle \nabla F(x), x-x(\alpha)\rangle}-1\right| \leqq \frac{1}{2}\alpha L \tag{2.14}
$$

for sufficiently small α.

When (2.9) is satisfied, the Goldstein rule (1.12) is always feasible. To see this, observe that F is differentiable and therefore continuous, and that $x(\alpha)$ is continuous in α with x fixed (see (2.11)). Hence, the quotient

$$
\frac{F_n - F(P(x_n-\alpha \nabla F_n))}{\langle \nabla F_n, x_n - P(x_n-\alpha \nabla F_n)\rangle} \tag{2.15}
$$

is defined and continuous for $\alpha>0$, provided x_n is not an extremal. If the latter provision is met and (1.12a) is violated, it then follows from (2.9) and the intermediate value theorem that (1.12b) holds for *at least one* α_n in $(0, a_n)$. On the other hand, if x_n is an extremal, then (1.12a) is always true since both sides of this inequality vanish. As there are no other possibilities to consider, the Goldstein rule is feasible whenever (2.9) is satisfied [17]. The feasibility of the Bertsekas–Armijo rule (1.11) is also an immediate consequence of (2.9).

Suppose that the sequences $\{x_n\}\subset \Omega$, $\{\alpha_n\}\subset (0,\infty)$, $\{a_n\}\subset (0,\infty)$ and $\{\delta_n\}\subset (0, \frac{1}{2}]$ obey (1.2) and the Goldstein rule (1.12). It then follows from (2.6) that $F_n \geqq F_{n+1}$ for all $n\geqq 0$, or equivalently, that x_{n+1} belongs to the level set $S_n = \{x\in\Omega | F(x)\leqq F(x_n)\}$ for all $n\geqq 0$. Suppose also that ∇F is Lipschitz continuous on the convex hull of S_n, with Lipschitz norm

$$
L_n = \sup_{\substack{x,y\in \mathrm{Co}(S_n)\\ x\neq y}} \frac{\|\nabla F(x)-\nabla F(y)\|}{\|x-y\|} < \infty. \tag{2.16}
$$

Then (2.13) is satisfied for x, y in S_n with L replaced by L_n; in particular, if x_n is not an extremal and (1.12b) holds, one has

$$
-\frac{1}{2}\alpha_n L_n \leqq \frac{F_n - F(P(x_n-\alpha_n \nabla F_n))}{\langle \nabla F_n, x_n - P(x_n-\alpha_n \nabla F_n)\rangle} - 1 \leqq -\delta_n,
$$

and therefore $L_n\alpha_n \geqq 2\delta_n$.

In all cases, the parameters α_n, a_n, and δ_n satisfy

$$
\begin{aligned}
\alpha_n &\geqq \left[\max\left\{\frac{1}{a_n}, \frac{L_n}{2\delta_n}\right\}\right]^{-1} \geqq \left[\max\left\{\frac{1}{a}, \frac{L_0}{2\delta}\right\}\right]^{-1} \\
&= \alpha > 0,
\end{aligned}
\tag{2.17}
$$

where a and δ are positive lower bounds for $\{a_n\}$ and $\{\delta_n\}$ respectively. Moreover, if one puts

$$L_\infty = \lim_{n\to\infty} L_n, \quad a_\infty = \lim_{n\to\infty} \inf a_n, \quad \delta_\infty = \lim_{n\to\infty} \inf \delta_n, \tag{2.18}$$

then (2.17) yields the asymptotic estimate

$$\liminf \alpha_n = \alpha_\infty, \tag{2.19a}$$

with

$$\alpha_\infty \geqq \begin{cases} a & \text{if } L_\infty = 0, \\ \min\left\{a_\infty, \dfrac{2\delta_\infty}{L_\infty}\right\} & \text{if } L_\infty > 0. \end{cases} \tag{2.19b}$$

Finally, suppose that $\{x_n\}$ converges to ξ, and that ∇F is locally Lipschitz continuous near ξ, with Lipschitz norms

$$L(\sigma) = \sup_{\substack{x,y\in B(\xi,\sigma)\\ x\neq y}} \frac{\|\nabla F(x) - \nabla F(y)\|}{\|x-y\|} < \infty \tag{2.20a}$$

for sufficiently small $\sigma > 0$, where

$$B(\xi,\sigma) = \{z \in \Omega \,|\, \|z-\xi\| \leqq \sigma\}. \tag{2.20b}$$

Then for each small positive σ there is a corresponding $N(\sigma)$ such that $\|x_n - \xi\| \leqq \sigma$ and

$$\alpha_n \geqq \left[\max\left\{\frac{1}{a_n}, \frac{L(\sigma)}{2\delta_n}\right\}\right]^{-1} \tag{2.21}$$

for all $n \geqq N(\sigma)$. Since σ can be arbitrarily small here, one now obtains the asymptotic estimate (2.19) with

$$L_\infty = \lim_{\sigma\to 0^+} L(\sigma). \tag{2.22}$$

By a straightforward adaptation of Bertsekas' arguments in [14], it is possible to obtain estimates similar to (2.17), (2.19) and (2.21) for the Bertsekas–Armijo rule (1.11), viz.,

$$\begin{aligned} \alpha_n &\geqq \left[\max\left\{\frac{1}{a_n}, \frac{L_n}{2\beta_n(1-\delta_n)}\right\}\right]^{-1} \geqq \left[\max\left\{\frac{1}{a}, \frac{L_0}{2\beta(1-d)}\right\}\right]^{-1} \\ &= \alpha > 0 \end{aligned} \tag{2.23}$$

and

$$\alpha_n \geqq \left[\max\left\{\frac{1}{a_n}, \frac{L(\sigma)}{2\beta_n(1-\delta_n)}\right\}\right]^{-1}, \tag{2.24}$$

and finally,

$$\liminf \alpha_n = \alpha_\infty, \tag{2.25a}$$

with

$$\alpha_\infty \geqq \begin{cases} a_\infty & \text{if } L_\infty = 0, \\ \min\left\{a_\infty, \dfrac{2\beta_\infty}{L_\infty}(1 - \lim\limits_{n\to\infty} \sup \delta_n)\right\} & \text{if } L_\infty > 0, \end{cases} \tag{2.25b}$$

where a and β are positive lower bounds for $\{a_n\}$ and $\{\beta_n\}$, $d \in (0, 1)$ is an upper bound for $\{\delta_n\}$, $\beta_\infty = \lim_{n\to\infty} \inf \beta_n$, and L_∞ is specified by (2.18) or (2.22) depending on whether ∇F is Lipschitz continuous on $\mathrm{Co}(S_n)$, or locally Lipschitz continuous near $\xi = \lim_{n\to\infty} x_n$.

It has just been shown that the Bertsekas–Armijo and Goldstein step length rules generate sequences $\{\alpha_n\}$ which satisfy (1.14b) under conditions of the Lipschitz continuity type on $\nabla F(x)$; for both of these rules, the constraint (1.14a) is satisfied a priori. However, in the case of the simple threshold rule (1.6) the situation is reversed; here (1.14b) is satisfied by construction, and it is necessary to prove that (1.14a) follows from the rightmost inequalities in (1.6), viz.,

$$\alpha_n \leqq a < \frac{2}{L}.$$

Again this follows easily from (2.14) under Lipschitz continuity conditions on ∇F. If x_n is not an extremal and if L is a Lipschitz constant for ∇F on the convex set Ω, then (1.2), (1.6) and (2.14) give

$$\frac{F_n - F_{n+1}}{\langle \nabla F_n, x_n - x_{n+1}\rangle} \geqq 1 - \frac{1}{2}\alpha_n L \geqq 1 - \frac{1}{2} aL.$$

Thus in all cases (1.14a) is satisfied for all $n \geqq 0$, with

$$\delta = 1 - \tfrac{1}{2}aL > 0.$$

3. The modulus of pseudoconvexity. For gradient-like methods, theorems establishing the global minimizing property, $F_n \to \inf_\Omega F$, are traditionally formulated for convex functionals F; however, it appears that many of these results remain valid for a substantially larger class of functionals which are *pseudoconvex* in Mangasarian's sense [18]; in any case, this is true for the main theorems in § 4.

A differentiable function $F : X \to \mathbb{R}^1$ is pseudoconvex on $\Omega \subset X$ if and only if

$$F(x) > F(y) \Rightarrow \langle \nabla F(x), x - y\rangle > 0 \tag{3.1}$$

for all $x, y \in \Omega$. If F is pseudoconvex, then the level sets of F in Ω are convex (i.e., F is quasiconvex) and every extremal of F in Ω is a global minimizer of F in Ω. Every convex functional F is pseudoconvex, since the inequality

$$F(y) - F(x) \geqq \langle \nabla F(x), y - x\rangle \tag{3.2}$$

holds for all x, y when F is convex. In a certain *quantitative* sense to be made precise below, convexity is an extreme case of pseudoconvexity.

LEMMA 3.1. *Let $F : X \to \mathbb{R}^1$ be pseudoconvex on the nonempty convex set $\Omega \subset X$. Suppose that*

$$F(x) > \inf_\Omega F \tag{3.3}$$

at some fixed x in Ω, and put

$$\hat{\kappa}(x) = \inf_{\substack{y\in\Omega \\ F(x)>F(y)}} \frac{\langle \nabla F(x), x - y\rangle}{F(x) - F(y)}. \tag{3.4}$$

Then

$$0 \leqq \hat{\kappa}(x) \leqq 1. \tag{3.5}$$

Furthermore, if F is convex on Ω then

$$\hat{\kappa}(x) = 1. \tag{3.6}$$

Proof. In view of (3.1) and (3.3), the infimum in (3.4) is finite and nonnegative. If $x, y \in \Omega$ and $F(x) > F(y)$, then

$$\langle \nabla F(x), x - y \rangle > 0$$

and consequently

$$F(x) > F(x + \alpha(y - x)),$$

for α sufficiently small and positive. Since Ω is convex, it follows that $x + \alpha(y - x) \in \Omega$ for $\alpha \in [0, 1]$, consequently (3.1) and (3.4) give

$$\begin{aligned} \langle \nabla F(x), \alpha(x - y) \rangle &\geqq \hat{\kappa}(x)[F(x) - F(x + \alpha(y - x))] \\ &\geqq \hat{\kappa}(x)[\langle \nabla F(x), \alpha(x - y) \rangle + o(\alpha)] \end{aligned}$$

for small $\alpha > 0$. In the limit as $\alpha \to 0^+$ this yields

$$\langle \nabla F(x), x - y \rangle \geqq \hat{\kappa}(x) \langle \nabla F(x), x - y \rangle,$$

and therefore $1 \geqq \hat{\kappa}(x)$. If F is convex, then (3.2) and (3.4) also imply $\hat{\kappa}(x) \geqq 1$. □

DEFINITION 3.1. Let $F : X \to \mathbb{R}^1$ be pseudoconvex on the nonempty convex set $\Omega \subset X$. For $\sigma > \inf_\Omega F$, put

$$\kappa(\sigma) = \inf_{\substack{x \in \Omega \\ \sigma \geqq F(x) > \inf_\Omega F}} \hat{\kappa}(x) \geqq 0, \tag{3.7}$$

where $\hat{\kappa}(x)$ is specified by (3.4). The function $\kappa(\cdot) : (\inf_\Omega F, \infty) \to [0, \infty]$ will be called the *modulus of pseudoconvexity* for F on Ω.

The parameter $\kappa(\sigma)$ arises in a natural way in the global and asymptotic analyses of § 4; its principal characteristics are summarized in the following lemma.

LEMMA 3.2. *Let $F : X \to \mathbb{R}^1$ be pseudoconvex on the nonempty convex set $\Omega \subset X$, with associated modulus of pseudoconvexity $\kappa(\cdot)$ defined by* (3.7). *If F is constant on Ω, then $\kappa(\sigma) = +\infty$ for all $\sigma > \inf_\Omega F$. Otherwise, if F is continuous but not constant on Ω, then $\kappa(\cdot)$ is nonincreasing, with*

$$0 \leqq \kappa(\sigma) \leqq 1 \tag{3.8}$$

for all $\sigma > \inf_\Omega F$, and consequently

$$\kappa(\sigma) \leqq \lim_{\sigma \to (\inf_\Omega F)^+} \kappa(\sigma) \in [0, 1] \tag{3.9}$$

for all $\sigma > \inf_\Omega F$. Finally, if F is convex, and continuous but not constant, then

$$\kappa(\sigma) = 1 \tag{3.10}$$

for all $\sigma > \inf_\Omega F$.

Proof. For $\sigma > \inf_\Omega F$, put

$$S(\sigma) = \{x \in \Omega \mid \sigma \geqq F(x) > \inf_\Omega F\}.$$

If F is constant on Ω, then $S(\sigma)$ is empty, and consequently $\kappa(\sigma) = +\infty$. If F is continuous but not constant on the convex set Ω, it follows from the intermediate value theorem that $S(\sigma)$ is not empty for all $\sigma < \inf_\Omega F$, in which case (3.8) follows at once

from (3.7) and Lemma 3.1. Furthermore, $\sigma_2 \geqq \sigma_1 > \inf_\Omega F \Rightarrow S(\sigma_2) \supset S(\sigma_1) \Rightarrow \kappa(\sigma_2) \leqq \kappa(\sigma_1)$; therefore $\kappa(\cdot)$ is nonincreasing and converges upward to some limit in $[0, 1]$ as σ converges to $\inf_\Omega F$ from above. If F is actually convex, then (3.10) is immediate from (3.7) and Lemma 3.1. □

Theorems 4.2–4.4 in § 4 are formulated for differentiable pseudoconvex functionals with strictly positive moduli $\kappa(\sigma)$ on the set $F(\Omega) - \{\inf_\Omega F\}$; these theorems and the following lemma demonstrate that the class of such functionals is a nontrivial extension of the class of all differentiable convex functionals.

LEMMA 3.3. *Suppose that, for all x in the convex set $\Omega \subset X$,*

$$F(x) = \Phi(G(x)),$$

where $G : X \to \mathbb{R}^1$ is bounded below, convex and continuously differentiable in the Fréchet sense on Ω, and $\Phi : \mathbb{R}^1 \to \mathbb{R}^1$ is continuously differentiable, with

$$\frac{d\Phi}{dt}(t) > 0 \tag{3.11}$$

for all t in the set $G(\Omega)_+ = G(\Omega) \cup \{\inf_\Omega G\}$. The composite function $F : X \to \mathbb{R}^1$ is then bounded below, continuously differentiable and pseudoconvex on Ω. If G is not constant on Ω then F is not constant on Ω, and

$$0 < \kappa(\sigma) \leqq 1 \tag{3.12}$$

for all σ in the nonempty interval $F(\Omega)_- = F(\Omega) - \{\inf_\Omega F\}$, where $\kappa(\cdot)$ is the modulus of pseudoconvexity in (3.7); *moreover,*

$$\lim_{\sigma \to (\inf_\Omega F)^+} \kappa(\sigma) = 1. \tag{3.13}$$

Proof. G is continuous on Ω, and Ω is convex and therefore connected. Thus, by the intermediate value theorem, $G(\Omega)_+$ is an interval containing its left endpoint $\inf_\Omega G > -\infty$. According to (3.11), Φ is continuous and strictly increasing on $G(\Omega)_+$; therefore, the set $F(\Omega)_+ = F(\Omega) \cup \{\inf_\Omega F\} = \Phi(G(\Omega)_+)$ is an interval containing its left endpoint $\inf_\Omega F$. Furthermore, Φ has a continuous single-valued inverse Φ^{-1} of $F(\Omega)_+$, and

$$F(x) > F(y) \Leftrightarrow G(x) > G(y) \tag{3.14}$$

for all $x, y \in \Omega$.

Suppose that $F(x) > F(y)$ for some $x, y \in \Omega$. Since G is convex, the mean value theorem, (3.11) and (3.14) give

$$\begin{aligned} 0 < F(x) - F(y) &= \frac{d\Phi}{dt}(\tau)[G(x) - G(y)] \\ &\leqq \frac{d\Phi}{dt}(\tau)\langle \nabla G(x), x - y \rangle \end{aligned} \tag{3.15}$$

for some τ in the interval

$$[G(y), G(x)] \subset G(\Omega)_+. \tag{3.16}$$

Moreover, by the chain rule, F is continuously differentiable on Ω with $\nabla F(x) = \frac{d\Phi}{dt}(G(x)) \nabla G(x)$, and therefore

$$\langle \nabla F(x), x - y \rangle = \frac{d\Phi}{dt}(G(x))\langle \nabla G(x), x - y \rangle \tag{3.17}$$

for all x, y in Ω. The inequalities (3.11), (3.15) and (3.17) now yield

$$F(x)>F(y)\Rightarrow\frac{\langle\nabla F(x),\,x-y\rangle}{F(x)-F(y)}\geqq\frac{\dfrac{d\Phi}{dt}(G(x))}{\dfrac{d\Phi}{dt}(\tau)}>0 \tag{3.18}$$

for some τ in the interval (3.17) and x, y in Ω. Thus F is pseudoconvex on Ω.

If G is not constant on Ω, the set $F(\Omega)_-$ is an interval with a nonempty interior and left endpoint $\inf_\Omega F>-\infty$. For fixed $\sigma\in F(\Omega)_-$, the continuous function $\dfrac{d\Phi}{dt}(\cdot)$ attains its minimum and maximum values on the closed bounded interval $[\inf_\Omega G,\,\Phi^{-1}(\sigma)]\subset G(\Omega)_+$. Furthermore, according to (3.11),

$$0<m(\sigma)=\min_{\Phi^{-1}(\sigma)\geqq t\geqq\inf_\Omega G}\frac{d\Phi}{dt}(t) \tag{3.19a}$$

and

$$0<M(\sigma)=\max_{\Phi^{-1}(\sigma)\geqq t\geqq\inf_\Omega G}\frac{d\Phi}{dt}(t). \tag{3.19b}$$

Consequently, (3.7), (3.19), (3.20) and Lemma 3.2 give

$$0<\frac{m(\sigma)}{M(\sigma)}\leqq\kappa(\sigma)\leqq 1 \tag{3.20}$$

for all $\sigma\in F(\Omega)_-$. Finally, since Φ^{-1} is continuous and strictly increasing on $F(\Omega)_+$, one has

$$\lim_{\sigma\to(\inf_\Omega F)^+}\Phi^{-1}(\sigma)=\Phi^{-1}(\inf_\Omega F)=\inf_\Omega G,$$

and therefore

$$\lim_{\sigma\to(\inf_\Omega F)^+}\frac{m(\sigma)}{M(\sigma)}=1. \tag{3.21}$$

Condition (3.13) is now immediate from (3.20) and (3.21). □

Example 3.1. Let $G:X\to\mathbb{R}^1$ satisfy the hypotheses of Lemma 3.3, and in addition, let $G(x)$ have nonnegative values as x ranges over a given nonempty convex set $\Omega\subset X$. For $t\in[0,\infty)$, put

$$\Phi(t)=\frac{t}{1+t}.$$

Then

$$\frac{d\Phi}{dt}(t)=\frac{1}{(1+t)^2}>0$$

on $[0,\infty)$. Consequently, the composite function

$$F(x)=\frac{G(x)}{1+G(x)}$$

is pseudoconvex on Ω. Notice that if G is a *linear* functional then the pseudoconvex functional F is actually *concave*.

Note 3.1. Reference [19] formulates a convergence theorem for certain conditional gradient processes and composite pseudoconvex functionals of the type considered in Lemma 3.3.

4. Convergence theorems. The results in this section are formulated for projected gradient sequences $\{x_n\}$ satisfying (1.2), (1.13) and (1.14). Lipschitz continuity of the gradient of F is not invoked explicitly in Theorems 4.1–4.3; however, conditions of this sort play an important (and perhaps essential) part in establishing (1.13) and (1.14) for Goldstein's rule (1.12) and for Bertsekas' generalized Armijo rule (1.11) (see § 2 and [14]). Bertsekas obtains the counterpart of Theorem 4.1 below, for projected gradient sequences satisfying (1.13) with $\delta_n = \delta = a$ constant, and *both* threshold conditions in (1.5). His proof is modified here in order to circumvent the upper threshold condition and thereby permit an analysis of (1.11) and (1.12) with variable and unbounded upper thresholds a_n; for such schemes, the associated asymptotic convergence rate bounds in Theorems 4.2 and 4.3 are "optimized" in a certain sense when $\delta_n \to \frac{1}{2}$ and $a_n \to \infty$ as $n \to \infty$ (see Notes 4.4 and 4.8).

THEOREM 4.1. *Let Ω be a nonempty closed convex subset of a real Hilbert space X, and let $F : X \to \mathbb{R}^1$ be bounded below and continuously differentiable in the Fréchet sense on Ω. Furthermore, let $\{x_n\} \subset \Omega$ be a projected gradient sequence satisfying* (1.2), (1.13) *and* (1.14) *with $\{\delta_n\} \subset (0, \infty)$ and $\{\alpha_n\} \subset (0, \infty)$; i.e.,*

$$x_{n+1} = P(x_n - \alpha_n \nabla F_n), \qquad x_0 \in \Omega,$$
$$F_n - F_{n+1} \geqq \delta_n \langle \nabla F_n, x_n - x_{n+1} \rangle,$$
$$\delta_n \geqq \delta, \qquad \alpha_n \geqq \alpha$$

for some fixed positive numbers δ and α and all integers $n \geqq 0$. Then $\{F_n\}$ is nonincreasing and converges to some limit $f \geqq \inf_\Omega F > -\infty$; moreover,

$$\lim_{n \to \infty} \langle \nabla F_n, x_n - x_{n+1} \rangle = 0, \tag{4.1}$$

and every subsequential limit point ξ of $\{x_n\}$ is an extremal of F in Ω; i.e., ξ satisfies (1.18). *In particular, if x_N is an extremal for some $N \geqq 0$, then $x_n = x_N$ for all $n \geqq N$. Finally, if $\alpha_n \leqq b$ for some $b \geqq \alpha$ and all $n \geqq 0$, then*

$$\lim_{n \to \infty} \|x_{n+1} - x_n\| = 0. \tag{4.2}$$

Proof. Conditions (1.2), (1.13) and (2.6) give

$$\begin{aligned} F_n - F_{n+1} &\geqq \delta_n \langle \nabla F_n, x_n - x_{n+1} \rangle \\ &\geqq \frac{\delta_n}{\alpha_n} \|x_n - x_{n+1}\|^2 \geqq 0 \end{aligned} \tag{4.3}$$

for $n \geqq 0$. Thus $\{F_n\}$ is nonincreasing and bounded below, and must therefore converge to a limit $f \geqq \inf_\Omega F > -\infty$. It follows that

$$\lim_{n \to \infty} (F_n - F_{n+1}) = 0.$$

Equation (4.1) is now an immediate consequence of (1.14a) and (4.3).

Suppose that $\lim_{k\to\infty} x_{n_k} = \xi \in \Omega$. Consider that for fixed $z \in \Omega$, (1.2) and (2.5) give

$$\begin{aligned}\langle \nabla F_{n_k}, z - x_{n_k}\rangle &= \langle \nabla F_{n_k}, z - x_{n_k+1}\rangle + \langle \nabla F_{n_k}, x_{n_k+1} - x_{n_k}\rangle \\ &\geqq \frac{1}{\alpha_{n_k}}\langle x_{n_k+1} - x_{n_k}, x_{n_k+1} - z\rangle + \langle \nabla F_{n_k}, x_{n_k+1} - x_{n_k}\rangle.\end{aligned} \tag{4.4}$$

Suppose that $\{x_{n_k+1}\}$ is bounded. Then (4.4), (2.6), (1.14b) and the Schwarz inequality yield

$$\langle \nabla F_{n_k}, z - x_{n_k}\rangle \geqq -\frac{C}{\alpha^{1/2}}\langle \nabla F_{n_k}, x_{n_k} - x_{n_k+1}\rangle^{1/2} + \langle \nabla F_{n_k}, x_{n_k+1} - x_{n_k}\rangle,$$

where C is any upper bound on $\{\|x_{n_k+1} - z\|\}$. In the limit as $k \to \infty$, it now follows from (4.1) and the continuity of ∇F that

$$\langle \nabla F(\xi), z - \xi\rangle \geqq 0. \tag{4.5}$$

On the other hand, suppose that $\{x_{n_k+1}\}$ is *not* bounded. Then there is a subsequence $\{x_{n_{k_j}}\}$ for which

$$\lim_{j\to\infty} x_{n_{k_j}} = \xi, \tag{4.6a}$$

while

$$\lim_{j\to\infty} \|x_{n_{k_j}+1}\| = \infty. \tag{4.6b}$$

To simplify the notation, write $m_j = n_{k_j}$ and consider that

$$\begin{aligned}2\langle x_{m_j+1} - x_{m_j}, x_{m_j+1}\rangle &= \|x_{m_j+1} - x_{m_j}\|^2 + \|x_{m_j+1}\|^2 - \|x_{m_j}\|^2 \\ &\geqq \|x_{m_j+1}\|^2 - \|x_{m_j}\|^2.\end{aligned}$$

Because of (4.6), the right side of this inequality is positive for large j; hence (4.4), (2.6), (1.14b), and the Schwarz inequality give

$$\langle \nabla F_{m_j}, z - x_{m_j}\rangle \geqq -\frac{\|z\|}{\alpha^{1/2}}\langle \nabla F_{m_j}, x_{m_j} - x_{m_j+1}\rangle^{1/2} + \langle \nabla F_{m_j}, x_{m_j+1} - x_{m_j}\rangle$$

for j sufficiently large. In view of (4.1) this yields (4.5) once again, in the limit as $j \to \infty$. Thus (4.5) holds in all cases, and since z can be any element of Ω, this means that ξ is an extremal.

If x_N is an extremal it follows at once from (1.2), (1.14b) and (2.7) that $x_n = x_N$ for all $n \geqq N$.

Finally, if $\alpha_n \leqq b$ for all $n \geqq 0$, then (2.6) gives

$$\langle \nabla F_n, x_n - x_{n+1}\rangle \geqq \frac{1}{b}\|x_{n+1} - x_n\|^2 \geqq 0,$$

and (4.2) is now an immediate consequence of (4.1). □

Note 4.1. If the level set $S_0 = \{x \in \Omega \mid F(x) \leqq F(x_0)\}$ is compact and if all subsequential limit points of $\{x_n\} \subset S_0$ are extremals of F in Ω, then $\{x_n\}$ converges to the set, E_0, of extremals in S_0; i.e.,

$$\lim_{n\to\infty} \left(\inf_{x\in E_0} \|x_n - x\|\right) = 0.$$

Moreover, if E_0 is finite and if $\lim_{n\to\infty}\|x_{n+1}-x_n\|=0$, then $\{x_n\}$ converges to some *specific* extremal in E_0 [16], [20].

Theorem 4.1 can establish the fact of convergence for certain projected gradient processes, but it says nothing about *rates* of convergence. For pseudoconvex functionals, it turns out that much of the convergence rate analysis for (1.2) ultimately reduces to an investigation of the recursive inequalities $r_n - q_n r_n^k \geqq r_{n+1} \geqq 0$, with $n \geqq 0$, q_n positive and k in the range $2 \geqq k \geqq 1$. When $k>1$, a consideration of the differential inequality

$$\frac{dr}{dt} \leqq -q(t)r^k$$

suggests the transformation $s_n = r_n^{1-k}$, and indeed this transformation greatly simplifies the proof of the following basic lemma.

LEMMA 4.1. *Suppose that* $\{r_n\}\subset[0,\infty)$ *and* $\{q_n\}\subset[0,\infty)$ *satisfy*

$$r_n - q_n r_n^k \geqq r_{n+1} \tag{4.7}$$

for $n \geqq 0$, *with* k *a fixed exponent in the range* $(1,\infty)$. *If*

$$\liminf_{n\to\infty} q_n \geqq q_\infty > 0, \tag{4.8}$$

then

$$\limsup_{n\to\infty} r_n n^{1/(k-1)} \leqq [(k-1)q_\infty]^{-1/(k-1)}. \tag{4.9}$$

Moreover, if $\lim_{n\to\infty} q_n = \infty$ *then* $r_n = o(n^{-1/(k-1)})$; *i.e.*, $\lim_{n\to\infty} r_n n^{1/(k-1)} = 0$. *Finally, if*

$$q_n \geqq q > 0 \tag{4.10}$$

for $0 \leqq n < N$, *then*

$$r_n \leqq r_0 \cdot [1+(k-1)r_0^{k-1}qn]^{-1/(k-1)} \tag{4.11}$$

for $0 \leqq n \leqq N$.

Proof. If $r_m = 0$ for some $m \geqq 0$, then $r_n = 0$ for all $n \geqq m$. If $r_n > 0$ for $0 \leqq n < m$, put $s_n = r_n^{1-k}$ for n in this range, and observe that the mean value theorem gives

$$s_{n+1} - s_n = \frac{r_n^{k-1} - r_{n+1}^{k-1}}{(r_n \cdot r_{n+1})^{k-1}} = \frac{(k-1)\zeta_n^{k-2}(r_n - r_{n+1})}{(r_n \cdot r_{n+1})^{k-1}},$$

for $0 \leqq n < m-1$ and some ζ_n in the interval

$$r_n \geqq \zeta_n \geqq r_{n+1} > 0.$$

In view of (4.7), one then has

$$s_{n+1} - s_n \geqq (k-1)q_n\left(\frac{r_n}{\zeta_n}\right)\left(\frac{\zeta_n}{r_{n+1}}\right)^{k-1} \geqq (k-1)q_n,$$

for k fixed in $(1,\infty)$ and $0 \leqq n < m-1$. Consequently,

$$s_n - s_0 = \sum_{i=0}^{n-1}(s_{i+1}-s_i) \geqq (k-1)\sum_{i=0}^{n-1} q_i$$

for $0 \leqq n < m$. In all cases, therefore, one has

$$0 \leqq r_n \leqq r_0 \cdot \left[1+(k-1)r_0^{(k-1)} \cdot \sum_{i=0}^{n-1} q_i\right]^{1/(1-k)} \tag{4.12}$$

for $n \geqq 0$, and thus

$$0 \leqq r_n n^{1/(k-1)} \leqq r_0 \left[n^{-1} + (k-1) r_0^{k-1} n^{-1} \cdot \sum_{i=0}^{n-1} q_i \right]^{1/(1-k)} \tag{4.13}$$

for all $n \geqq 1$. Since

$$\lim_{n\to\infty} \inf q_n \geqq q_\infty \Rightarrow \lim_{n\to\infty} \inf \left(n^{-1} \cdot \sum_{i=1}^{n-1} q_i \right) \geqq q_\infty,$$

the estimate (4.9) follows at once from (4.8) and (4.13). If $q_n \to \infty$ then (4.8), and consequently (4.9), holds for all $q_\infty > 0$ no matter how large, and this means that $r_n n^{1/(k-1)} \to 0$. Finally, (4.11) is immediate from (4.10) and (4.12). □

Note 4.2. Demyanov and Rubinov [3] use a different method of proof to obtain the estimate $r_n = O(n^{-1})$ when (4.7) holds with $k = 2$; the same estimate is derived in [19] by a special version of the proof of Lemma 4.1. The principal content of Lemma 4.1 is asserted without proof in [3, remark, p. 130] for general $k > 1$, but the result is actually applied there only for the special case $k = 2$.

THEOREM 4.2. *Let Ω be a nonempty closed convex subset of a real Hilbert space X, and let $F : X \to \mathbb{R}^1$ be pseudoconvex, bounded below, and continuously differentiable in the Fréchet sense. Furthermore, let $\{x_n\} \subset \Omega$ be a projected gradient sequence satisfying* (1.2), (1.13) *and* (1.14) *with $\{\delta_n\} \subset (0, \infty)$ and $\{\alpha_n\} \subset (0, \infty)$; i.e.,*

$$x_{n+1} = P(x_n - \alpha_n \nabla F_n), \quad x_0 \in \Omega,$$

$$F_n - F_{n+1} \geqq \delta_n \langle \nabla F_n, x_n - x_{n+1} \rangle,$$

$$\delta_n \geqq \delta, \qquad \alpha_n \geqq \alpha$$

for some fixed positive numbers δ and α and all $n \geqq 0$. Put

$$r_n = F_n - \inf_{\Omega} F,$$

and suppose that $r_0 > 0$ implies $\kappa_0 > 0$, where $\kappa_0 = \kappa(F_0)$ and $\kappa(\cdot)$ is the pseudoconvexity modulus in (3.7); *i.e.,*

$$\kappa(\sigma) = \inf_{\substack{x \in \Omega \\ \sigma \geqq F(x) > \inf_\Omega F}} \left(\inf_{\substack{y \in \Omega \\ F(x) > F(y)}} \frac{\langle \nabla F(x), x - y \rangle}{F(x) - F(y)} \right)$$

for $\sigma > \inf_\Omega F$. Then

$$\lim_{n\to\infty} r_n = 0, \tag{4.14}$$

and every subsequential limit point of $\{x_n\}$ is a minimizer of F in Ω. In particular, if $r_N = 0$ for some $N \geqq 0$, then $r_n = 0$ and $x_n = x_N$ for all $n \geqq N$.

Suppose also that $\{x_n\}$ has a bounded range with diameter $D > 0$. Then, for all $n \geqq 0$,

$$r_n \leqq r_0 \cdot [1 + r_0 q n]^{-1} \tag{4.15a}$$

with

$$q = r \kappa_0^2 \delta \alpha \cdot [D + (D^2 + 4\kappa_0 r_0 \alpha)^{1/2}]^{-2}. \tag{4.15b}$$

Furthermore, if

$$0 < \delta_\infty = \lim_{n\to\infty} \inf \delta_n < \infty, \tag{4.16}$$

$$0<\alpha_\infty=\lim_{n\to\infty}\inf\alpha_n<\infty, \tag{4.17}$$

$$0<D_\infty=\lim_{n\to\infty}\sup\,(\lim_{m\to\infty}\sup\|x_n-x_m\|), \tag{4.18}$$

and

$$0<\kappa_\infty=\lim_{\sigma\to(\inf_\Omega F)^+}\kappa(\sigma)\leqq 1, \tag{4.19}$$

then

$$0\leqq\lim_{n\to\infty}\sup\,(nr_n)\leqq\frac{1}{q_\infty}, \tag{4.20a}$$

where

$$q_\infty=\left(\frac{\kappa^2\delta\alpha}{D^2}\right)_\infty. \tag{4.20b}$$

Finally, if $\{x_n\}$ *converges, or if* $\lim_{n\to\infty}\alpha_n=\infty$, *then* $r_n=o(n^{-1})$, *i.e.*,

$$\lim_{n\to\infty}nr_n=0. \tag{4.21}$$

Proof. According to Theorem 4.1, F_n converges monotonically downward to some limit f, with

$$F_0\geqq F_n\geqq f\geqq\inf_\Omega F$$

for all $n\geqq 0$; consequently, $r_0=0\Rightarrow F_0=f=\inf_\Omega F$. On the other hand, suppose that $r_0>0$ and $f>\inf_\Omega F$. Choose a $z\in\Omega$ for which

$$f>F(z)\geqq\inf_\Omega F. \tag{4.22}$$

Then $F_n-F(z)\geqq f-F(z)>0$, and therefore (1.2), (2.5) and (3.7) give

$$\begin{aligned}\kappa_0(F_n-F(z))&\leqq\langle\nabla F_n,\,x_n-z\rangle\\&=\langle\nabla F_n,\,x_n-x_{n+1}\rangle+\langle\nabla F_n,\,x_{n+1}-z\rangle\\&\leqq\langle\nabla F_n,\,x_n-x_{n+1}\rangle-\frac{1}{\alpha_n}\langle x_{n+1}-x_n,\,x_{n+1}-z\rangle\end{aligned} \tag{4.23}$$

for all $n\geqq 0$, with $\kappa_0>0$. Suppose that $\{x_n\}$ is bounded. Then by virtue of (1.14b), (2.6), and the Schwarz inequality, one can carry (4.23) further to

$$\kappa_0(F_n-F(z))\leqq\langle\nabla F_n,\,x_n-x_{n+1}\rangle+\frac{C}{(\alpha)^{1/2}}\langle\nabla F_n,\,x_n-x_{n+1}\rangle^{1/2},$$

where C is any upper bound on $\{\|x_{n+1}-z\|\}$. By Theorem 4.1, both terms on the right side of this inequality converge to zero as $n\to\infty$, while F_n converges to f. Consequently,

$$\kappa_0(f-F(z))\leqq 0,$$

which contradicts (4.22). On the other hand, suppose that $\{x_n\}$ is *not* bounded. Then there is a subsequence $\{n_k\}$ such that

$$\|x_{n_k+1}\|>\|x_{n_k}\|$$

for all $k \geqq 0$ (otherwise $\{\|x_n\|\}$ is eventually nonincreasing and therefore bounded). It follows that

$$2\langle x_{n_k+1} - x_{n_k}, x_{n_k+1}\rangle = \|x_{n_k+1} - x_{n_k}\|^2 + \|x_{n_k+1}\|^2 - \|x_{n_k}\|^2 > 0,$$

in which case (1.2), (1.14b), (2.6) and (4.23) give

$$\kappa_0(F_{n_k} - F(z)) \leqq \langle \nabla F_{n_k}, x_{n_k} - x_{n_k+1}\rangle + \frac{\|z\|}{\alpha^{1/2}}\langle \nabla F_{n_k}, x_{n_k} - x_{n_k+1}\rangle^{1/2},$$

for all $k \geqq 0$. Again, this produces a contradiction of (4.22) in the limit as $k \to \infty$. Since every alternative has now been considered, it follows that (4.22) is impossible; i.e., (4.14) holds in all cases (and because F is continuous, this means that every subsequential limit point of $\{x_n\}$ is a minimizer). In particular, if $r_N = 0$ for some $N \geqq 0$, then x_N is a minimizer (and therefore an extremal) of F in Ω, and consequently $x_n = x_N$ for all $n \geqq N$, by Theorem 4.1.

Fix n and suppose that $r_n > 0$. In view of (4.14), one then has

$$F_n > F_m \geqq \inf_{\Omega} F$$

for all large m, in which case (1.2), (2.5), and (3.7) give

$$\begin{aligned} \kappa_n(F_n - F_m) &\leqq \langle \nabla F_n, x_n - x_m\rangle \\ &= \langle \nabla F_n, x_n - x_{n+1}\rangle + \langle \nabla F_n, x_{n+1} - x_m\rangle \\ &\leqq \langle \nabla F_n, x_n - x_{n+1}\rangle + \frac{1}{\alpha_n}\|x_n - x_{n+1}\|\,\|x_{n+1} - x_m\| \end{aligned} \tag{4.24}$$

for all large m, where

$$\kappa_n = \kappa(F_n) \geqq \kappa(F_0) = \kappa_0 > 0,$$

according to Lemma 3.3 and Theorem 4.1. In the limit as $m \to \infty$, (1.13), (2.6), and (4.24) give

$$\begin{aligned} 0 < \kappa_n r_n &\leqq \langle \nabla F_n, x_n - x_{n+1}\rangle + \frac{D_{n+1}}{\alpha_n}\|x_{n+1} - x_n\| \\ &\leqq \langle \nabla F_n, x_n - x_{n+1}\rangle + \frac{D_{n+1}}{\alpha_n^{1/2}}\langle \nabla F_n, x_n - x_{n+1}\rangle^{1/2} \\ &\leqq \theta_n^2 + \frac{D_{n+1}}{\alpha_n^{1/2}}\theta_n, \end{aligned}$$

where

$$D_{n+1} = \lim_{m\to\infty} \sup \|x_{n+1} - x_m\| \tag{4.25}$$

and

$$\theta_n = \left(\frac{r_n - r_{n+1}}{\delta_n}\right)^{1/2}.$$

Complete the square to obtain

$$0 < \kappa_n r_n + \frac{D_{n+1}^2}{4\alpha_n} \leqq \left(\theta_n + \frac{D_{n+1}}{2\alpha_n^{1/2}}\right)^2.$$

Since $\theta_n \geqq 0$, this requires that

$$\theta_n \geqq -\frac{D_{n+1}}{2\alpha_n^{1/2}} + \left(\frac{D_{n+1}^2}{4\alpha_n} + \kappa_n r_n\right)^{1/2}$$

$$= \kappa_n r_n \cdot \left[\frac{D_{n+1}}{2\alpha_n^{1/2}} + \left(\frac{D_{n+1}^2}{4\alpha_n} + \kappa_n r_n\right)^{1/2}\right]^{-1}.$$

Equivalently,

$$r_n - r_{n+1} \geqq q_n r_n^2 > 0, \tag{4.26a}$$

where

$$q_n = \kappa_n^2 \delta_n \cdot \left[\frac{D_{n+1}}{2\alpha_n^{1/2}} + \left(\frac{D_{n+1}^2}{4\alpha_n} + \kappa_n r_n\right)^{1/2}\right]^{-2} \geqq q \tag{4.26b}$$

and q is given by (4.15b). There are now just two cases to consider: either there is an $N \geqq 0$ such that $r_n = 0$ for $n \geqq N$ and $r_n > 0$ for $0 \leqq n < N$, or else $r_n > 0$ for all n. In either case, (4.26) and Lemma 4.1 produce the estimate (4.15). In the first case (4.21) (and a fortiori, (4.20)) is immediate. In the second case, the numbers $\kappa_n = \kappa(F_n)$ exist for all n and converge monotonically upward to κ_∞ in (4.19), hence it follows from (4.16), (4.17) and (4.18) that (4.26) holds for all $n \geqq 0$, with

$$\liminf_{n\to\infty} q_n \geqq q_\infty \tag{4.26c}$$

and q_∞ given by (4.20b). Finally, if $\{x_n\}$ converges, then $D_\infty = 0$ and (4.26c) holds for arbitrarily large q_∞, which means that $\lim_{n\to\infty} q_n = \infty$; this is also true if $\lim_{n\to\infty} \alpha_n = \infty$. The estimates (4.18) and (4.21) now follow once again from Lemma 4.1. □

Note 4.3. Goldstein's article on unconstrained steepest descent processes [28] contains a model for the proof of (4.14) in the special case $\Omega = X$ and F convex. A somewhat different version of the same basic argument was utilized earlier in the proof of Theorem 4.1.

When the hypotheses of Theorem 4.2 are satisfied, and when F has a unique minimizer ξ in Ω, the sequence $\{x_n\}$ will converge to ξ if Ω is compact, or if F is uniformly quasiconvex in the sense of [3], or if F is convex and ξ is strongly nonsingular in the sense of [19]. In any of these cases, and others besides, $F_n - \inf_\Omega F$ will converge to zero more rapidly than n^{-1}. The question is, how *much* more rapidly? An inspection of (4.25) and (4.26) in the proof of Theorem 4.2 suggests that the convergence rate for r_n will depend upon how quickly $\{\|x_n - \xi\|\}$ converges to zero. Some further consideration reveals that the convergence rate for $\{\|x_n - \xi\|\}$ is constrained by how rapidly F grows in Ω near ξ. These two simple observations are brought together in the following theorem.

THEOREM 4.3. *Let X, Ω, F, $\kappa(\cdot)$, $\{x_n\}$, $\{\delta_n\}$ and $\{\alpha_n\}$ satisfy the hypotheses of Theorem* 4.2. *In addition, suppose that F has a unique minimizer ξ in Ω, and satisfies the uniform growth condition* (1.20); *i.e.*,

$$0 < \gamma(\sigma) = \inf_{\substack{x\in\Omega \\ \|x-\xi\|\geqq\sigma}} r(x)$$

for $\sigma > 0$, where $r(x) = F(x) - F(\xi)$. Then $\{x_n\}$ converges to ξ and $r_n = o(n^{-1})$. Suppose

that (1.21) *also holds, with*

$$0<B_\infty=\lim_{s\to 0^+}\left(\inf_{s\geqq\sigma>0}\frac{\gamma(\sigma)}{\sigma^\nu}\right)<\infty \tag{4.27}$$

and ν *fixed in* $[1,\infty)$. *Then there is a* $B>0$ *such that*

$$r_n\geqq B\|x_n-\xi\|^\nu \tag{4.28}$$

for all $n\geqq 0$. *If* $\nu>2$, *it follows from* (4.28) *that*

$$r_n\leqq r_0\cdot[1+\tau r_0^\tau qn]^{-1/\tau} \tag{4.29a}$$

for all $n\geqq 0$, *with*

$$\tau=(\nu-2)/\nu, \tag{4.29b}$$

$$q=\kappa_0 C\delta\cdot[1+(1+Cr_0^\tau)^{1/2}]^{-2} \tag{4.29c}$$

and

$$C=4\kappa_0 B^{2/\nu}\alpha. \tag{4.29d}$$

Furthermore, if δ_∞, α_∞, *and* κ_∞ *satisfy* (4.16), (4.17) *and* (4.19), *then*

$$\limsup_{n\to\infty} r_n n^{1/\tau}\leqq[\tau q_\infty]^{-1/\tau}, \tag{4.30a}$$

with

$$q_\infty=(\kappa^2 B^{2/\nu}\delta\alpha)_\infty. \tag{4.30b}$$

On the other hand, if $\lim_{n\to\infty}\alpha_n=\infty$, *then* $r_n=o(n^{-1/\tau})$; *i.e.*,

$$\lim_{n\to\infty} r_n n^{1/\tau}=0. \tag{4.31}$$

For ν *in the interval* $[1,2]$, (4.28) *gives*

$$r_n\leqq r_0\lambda^n \tag{4.32a}$$

for all $n\geqq 0$, *where*

$$0\leqq\lambda=\max\{0,1-r_0^\tau q\}<1, \tag{4.32b}$$

with q *specified by* (4.30b). *If* $\nu=2$, *if* $r_n>0$ *for all* $n\geqq 0$ *and if* δ_∞, α_∞ *and* κ_∞ *satisfy* (4.16), (4.17) *and* (4.19), *then*

$$\limsup_{n\to\infty}\frac{r_{n+1}}{r_n}\leqq\lambda_\infty, \tag{4.33a}$$

where

$$\lambda_\infty=1-q'_\infty \tag{4.33b}$$

and

$$0<q'_\infty=4(\kappa^2 B\delta\alpha)_\infty(1+[1+4(\kappa B\alpha)_\infty]^{1/2})^{-2}<1. \tag{4.33c}$$

Finally, if $\nu\in[1,2)$, *or if* $\nu=2$ *and* $\lim_{n\to\infty}\alpha_n=\infty$, *then* (4.33a) *and* (4.33b) *hold with*

$$0<q'_\infty=(\kappa\delta)_\infty<1. \tag{4.34}$$

Proof. If x_n does not converge to ξ, there is an $\varepsilon > 0$ and a subsequence $\{n_k\}$ such that $\|x_{n_k} - \xi\| \geqq \varepsilon$ for all $k \geqq 0$. Since $\gamma(\sigma)$ is nondecreasing and positive for $\sigma > 0$, it follows that $r_{n_k} \geqq \gamma(\|x_{n_k} - \xi\|) \geqq \gamma(\varepsilon) > 0$ for all $k \geqq 0$; on the other hand, $r_{n_k} \to 0$ according to Theorem 4.2. This contradiction proves that $x_n \to \xi$, and therefore that $r_n = o(n^{-1})$, by Theorem 4.2.

Fix B' in $(0, B_\infty)$. Since $\|x_n - \xi\| \to 0$, it follows from (4.27) that, for some sufficiently large $N \geqq 0$,

$$n \geqq N \Rightarrow \|x_n - \xi\| = 0 \quad \text{or} \quad r_n \geqq \gamma(\|x_n - \xi\|) \geqq B' \|x_n - \xi\|^\nu.$$

Consequently, $\sup\{b \in [0, \infty) | r_n \geqq b\|x_n - \xi\|^\nu, \forall n \geqq 0\}$ is strictly positive. This establishes (4.28).

Since $\|x_n - \xi\| \to 0$, it follows that $D_{n+1} = \|x_{n+1} - \xi\|$ in (4.25) and (4.26). If $\nu > 2$ and $r_{n+1} > 0$, put

$$\phi_n = \frac{\|x_{n+1} - \xi\|}{2\alpha_n^{1/2} r_{n+1}^{1/\nu}} \leqq \frac{1}{2\alpha^{1/2} B^{1/\nu}}. \tag{4.35}$$

By Theorem 4.1 one has $r_n \geqq r_{n+1}$; therefore r_n is positive and (4.26) yields

$$r_n - r_{n+1} \geqq q_n r_n^{2-2/\nu}, \tag{4.36a}$$

where

$$q_n = \kappa_n^2 \delta_n [\phi_n + (\phi_n^2 + \kappa_n r_n^{1-2/\nu})^{1/2}]^{-2} \geqq q, \tag{4.36b}$$

with q specified by (4.29c). According to Theorem 4.2, there are now just two cases to consider: either there is an $N \geqq 0$ such that $r_n = 0$ for $n \geqq N$ and $r_n = 0$ for $0 \leqq n < N$, or else $r_n > 0$ for all $n \geqq 0$. In either case, (4.35) and Lemma 4.1 produce the estimate (4.29). In the first case (4.31), and a fortiori, (4.30), is immediate. In the second case, the numbers κ_n converge upward to κ_∞, therefore it follows from (1.20), (4.16), (4.17), (4.19), (4.27) and (4.35) that (4.36) holds for all $n \geqq 0$, with

$$\liminf_{n \to \infty} q_n \geqq q_\infty \tag{4.37}$$

and q_∞ given by (4.30b). Moreover, if $\alpha_n \to \infty$, then $\phi_n \to 0$ and therefore $q_n \to \infty$. The estimates (4.30) and (4.31) now follow from Lemma 4.1.

Finally, if $\nu \in [1, 2]$ and $r_{n+1} > 0$, then $r_n > 0$ and (4.26) yields

$$r_n - r_{n+1} \geqq q_n r_n > 0, \tag{4.38a}$$

where

$$q_n = \kappa_n^2 \delta_n \cdot [\phi_n r_n^{(2-\nu)/2\nu} + (\phi_n^2 r_n^{(2-\nu)/\nu} + \kappa_n)^{1/2}]^{-2} \geqq r_0^\tau q, \tag{4.38b}$$

and τ and q are specified in (4.29). In all cases, the estimate (4.32) follows from (4.38). If $r_n > 0$ for all $n \geqq 0$, then (4.38) gives

$$0 < \frac{r_{n+1}}{r_n} \leqq (1 - q_n) r_n, \tag{4.39}$$

for all $n \geqq 0$, and this leads directly to the estimate (4.33)–(4.34). □

The convergence theory developed in [19], [20] for conditional gradient algorithms suggests that the linear convergence rate estimate (4.33)–(4.34) is probably conservative for $1 \leqq \nu < 2$. This is indeed the case. If ∇F is continuous and condition (4.27) holds at ξ with $\nu = 1$, then the sequence $\{x_n\}$ actually *terminates* at ξ for some value of n; moreover, if ∇F is locally Lipschitz continuous at ξ and if (4.27) holds with

$\nu \in (1, 2)$, then $\{x_n\}$ converges *superlinearly* to ξ. These results are established in the following theorem.

THEOREM 4.4. *Let X, Ω, F, $\kappa(\cdot)$, $\{x_n\}$, $\{\delta_n\}$, and $\{\alpha_n\}$ satisfy the hypotheses of Theorem* 4.2. *In addition, suppose that F has a unique minimizer ξ in Ω, and satisfies the uniform growth condition* (1.20)–(4.27) *with $\nu = 1$; i.e.,*

$$0 < \gamma(\sigma) = \inf_{\substack{x \in \Omega \\ \|x-\xi\| \geqq \sigma}} r(x)$$

for $\sigma > 0$, and

$$0 < B_\infty = \lim_{s \to 0^+} \left(\inf_{s \geqq \sigma > 0} \frac{\gamma(\sigma)}{\sigma} \right) < \infty, \tag{4.40}$$

where $r(x) = F(x) - F(\xi)$. Then there is a $B \in (0, B_\infty]$ such that

$$r_n \geqq B\|x_n - \xi\| \tag{4.41}$$

for all $n \geqq 0$, and an $N_* \geqq 0$ such that

$$n > N_* \Rightarrow x_n = \xi. \tag{4.42}$$

Moreover, suppose that ∇F is Lipschitz continuous on the level set $S_0 = \{x \in \Omega | F(x) \leqq F(x_0)\}$, and put

$$q = 4r_0\kappa_0^2 B^2 \delta\alpha \cdot [r_0^{1/2} + (r_0 + 4\kappa_0 B^2 \alpha)^{1/2}]^{-2}. \tag{4.43}$$

If $q < r_0$, then (4.42) *holds for*

$$N_* \geqq 1 + \frac{\ln [r_0(1 + \alpha L_0)/\alpha B^2]}{\ln (1/\lambda)}, \tag{4.44a}$$

where

$$1 > \lambda = 1 - q/r_0 > 0 \tag{4.44b}$$

and L_0 is a Lipschitz constant for ∇F on S_0. On the other hand, if $q \geqq r_0$, then (4.42) *holds with $N_* = 0$.*

Finally, suppose that ∇F is just locally Lipschitz continuous at ξ, and that (4.27) *is satisfied with ν fixed in* $(1, 2)$. *Then either* (4.42) *holds for some $N_* \geqq 0$, or else $\{x_n\}$ converges to ξ superlinearly, with $\|x_n - \xi\| > 0$ for all $n \geqq 0$ and*

$$\lim_{n \to \infty} \sup \frac{\|x_{n+1} - \xi\|}{\|x_n - \xi\|^{1/(\nu-1)}} \leqq C_\infty^{1/(\nu-1)}, \tag{4.45a}$$

where

$$0 \leqq L_\infty = \lim_{\sigma \to 0^+} \sup_{\substack{x \in \Omega \\ \sigma \geqq \|x-\xi\| > 0}} \frac{\|\nabla F(x) - \nabla F(\xi)\|}{\|x - \xi\|}, \tag{4.45b}$$

$$C_\infty = \begin{cases} (L/B)_\infty & \text{if } \alpha_\infty = \infty, \\ \left(\dfrac{1 + \alpha L}{\alpha B}\right)_\infty & \text{if } \alpha_\infty < \infty \end{cases} \tag{4.45c}$$

and

$$\alpha_\infty = \lim_{n \to \infty} \inf \alpha_n. \tag{4.45d}$$

Proof. It was established in the proof of Theorem 4.3 that (4.40) $\Rightarrow$ (4.41) for some $B>0$.

Since $\xi\in\Omega$, it follows from (1.2) and (2.5) that

$$\|x_{n+1}-\xi\|^2 \leqq -\alpha_n\langle\nabla F(\xi), x_{n+1}-\xi\rangle - \alpha_n\langle\nabla F_n-\nabla F(\xi), x_{n+1}-\xi\rangle + \langle x_n-\xi, x_{n+1}-\xi\rangle$$

$$\leqq -\alpha_n\langle\nabla F(\xi), x_{n+1}-\xi\rangle + (\alpha_n\|\nabla F_n-\nabla F(\xi)\|+\|x_n-\xi\|)\cdot\|x_{n+1}-\xi\|. \tag{4.46}$$

Furthermore,

$$r(x)=\langle\nabla F(\xi), x-\xi\rangle + o(\|x-\xi\|)$$

in the limit as $x\to\xi$; consequently, (4.40) gives

$$\lim_{s\to 0^+}\Big(\inf_{\substack{x\in\Omega\\ s\geqq\|x-\xi\|>0}}\frac{\langle\nabla F(\xi), x-\xi\rangle}{\|x-\xi\|}\Big)=\lim_{s\to 0^+}\Big(\inf_{\substack{x\in\Omega\\ s\geqq\|x-\xi\|>0}}\frac{r(x)}{\|x-\xi\|}\Big)\geqq B_\infty>0,$$

and therefore

$$\langle\nabla F(\xi), x-\xi\rangle\geqq B_\infty\|x-\xi\|$$

for *all* x in the convex set Ω. According to (4.46) one then has

$$\|x_{n+1}-\xi\|^2\leqq[\alpha_n(-B_\infty+\|\nabla F_n-\nabla F(\xi)\|)+\|x_n-\xi\|]\cdot\|x_{n+1}-\xi\|. \tag{4.47}$$

By Theorem 4.3, $x_n\to\xi$ and therefore $\nabla F_n\to\nabla F(\xi)$; for n sufficiently large, this means that

$$-B_\infty+\|\nabla F_n-\nabla F(\xi)\|<0$$

and

$$\alpha_n(-B_\infty+\|\nabla F_n-\nabla F(\xi)\|)+\|x_n-\xi\|\leqq\alpha(-B_\infty+\|\nabla F_n-\nabla F(\xi)\|)+\|x_n-\xi\|<0. \tag{4.48}$$

Condition (4.42) is now immediate from (4.47) and (4.48). Furthermore, if L_0 is a Lipschitz constant for ∇F on S_0, and if $q<r_0$, it follows from Theorem 4.3 and from (4.41) and (4.48) that

$$\alpha_n(-B_\infty+\|\nabla F_n-\nabla F(\xi)\|)+\|x_n-\xi\|\leqq-\alpha_nB_\infty+(1+\alpha_nL_0)\|x_n-\xi\|$$

$$\leqq-\alpha_nB+(1+\alpha_nL_0)\Big(\frac{r_n}{B}\Big)$$

$$\leqq-\alpha_nB+(1+\alpha_nL_0)\Big(\frac{r_0}{B}\Big)\lambda^n,$$

with λ specified by (4.44b). In view of (4.47) one therefore has $x_{n+1}=\xi$ if

$$\lambda^n<\frac{\alpha_nB^2}{r_0(1+\alpha_nL_0)}\leqq\frac{\alpha B^2}{r_0(1+\alpha L_0)}.$$

This establishes (4.44) for $q<r_0$. On the other hand, if $q\geqq r_0$, it follows from (4.34) that $x_n=\xi$ for all $n\geqq 1$.

Finally, if $\|x_n-\xi\|>0$ for all $n\geqq 0$, then (4.46) yields

$$\|x_{n+1}-\xi\|\leqq-\alpha_n\frac{\langle\nabla F(\xi), x_{n+1}-\xi\rangle}{\|x_{n+1}-\xi\|^\nu}\cdot\|x_{n+1}-\xi\|^{\nu-1}$$

$$+\Big(1+\alpha_n\frac{\|\nabla F_n-\nabla F(\xi)\|}{\|x_n-\xi\|}\Big)\|x_n-\xi\|,$$

or equivalently,

$$(4.49)\quad \frac{\|x_{n+1}-\xi\|}{\|x_n-\xi\|^{1/(\nu-1)}} \leqq \left[\frac{1+\alpha_n(\|\nabla F_n-\nabla F(\xi)\|/\|x_n-\xi\|)}{\|x_{n+1}-\xi\|^{2-\nu}+\alpha_n(\langle \nabla F(\xi), x_{n+1}-\xi\rangle/\|x_{n+1}-\xi\|^{\nu})}\right]^{1/(\nu-1)},$$

for $\nu \in (1, 2)$ and all $n \geqq 0$. If ∇F is locally Lipschitz continuous at ξ, one has

$$r_{n+1} = \langle \nabla F(\xi), x_{n+1}-\xi\rangle + O(\|x_{n+1}-\xi\|^2)$$

in the limit as $n \to \infty$ (see (2.13)); consequently, for $\nu \in (1, 2)$,

$$(4.50)\quad \liminf_{n\to\infty} \frac{\langle \nabla F(\xi), x_{n+1}-\xi\rangle}{\|x_{n+1}-\xi\|^{\nu}} = \liminf_{n\to\infty} \frac{r_n}{\|x_{n+1}-\xi\|^{\nu}} \geqq B_\infty,$$

in view of (4.27). The estimate (4.45) now follows at once from (4.49) and (4.50). □

Note 4.4. Suppose that $\{\alpha_n\}$ is generated by Goldstein's rule (1.12) with $\delta_n \to \delta_\infty \in (0, \frac{1}{2}]$. When ∇F is locally Lipschitz continuous, formulas (2.19) and (2.22) provide a lower bound on the quantity $\alpha_\infty = \lim_{n\to\infty} \inf \alpha_n$ appearing in the error estimates of Theorems 4.2, 4.3 and 4.4. In particular, when the upper thresholds a_n diverge to $+\infty$, one obtains

$$(4.51a)\quad \alpha_\infty \geqq \underline{\alpha}_\infty,$$

with

$$(4.51b)\quad \underline{\alpha}_\infty = \begin{cases} \infty & \text{if } L_\infty = 0, \\ \dfrac{2\delta_\infty}{L_\infty} & \text{if } L_\infty > 0 \end{cases}$$

and L_∞ specified by (2.20) and (2.22). Observe now that the right sides of the error estimates (4.20), (4.30), (4.33), and (4.45) are decreasing functions of α_∞; consequently these inequalities remain valid if α_∞ is replaced by $\underline{\alpha}_\infty$. Furthermore, the new coarser error bounds obtained in this way are optimized by taking $\delta_\infty = \frac{1}{2}$. Similar considerations yield "optimized" error bounds at some $\delta_\infty \in [\frac{1}{2}, 1)$ for the Bertsekas–Armijo rule (1.11). However, it is essential to understand that a conservative upper bound on the convergence rate is being optimized in this scheme, and not the convergence rate itself. One should also bear in mind that the computational costs entailed in implementing (1.11) and (1.12) tend to increase without limit as $a_n \to \infty$ or $\beta_n \to 1$ in the first case, and as $a_n \to \infty$ or $\delta_n \to \frac{1}{2}$ in the second case.

Note 4.5. Let F be quasiconvex on the convex set Ω, and let ξ be a minimizer of F in Ω. Then the level sets of F in Ω are convex, and the infimum on the right side of (1.20) can therefore be restricted to the sphere $S(\xi, \sigma) = \{x \in \Omega \mid \|x-\xi\| = \sigma\}$ in Ω; i.e.,

$$(4.52)\quad \gamma(\sigma) = \inf_{x\in S(\xi,\sigma)} F(x) - F(\xi).$$

If Ω is compact, or if Ω is closed and X is finite dimensional, the sphere $S(\xi, \sigma)$ is compact and the infimum in (4.52) is actually attained in $S(\xi, \sigma)$, provided F is continuous; under these circumstances, the uniform growth condition (1.20) holds if ξ is a *proper* minimizer, i.e., if $x \in \Omega$ and $x \neq \xi \Rightarrow F(x) > F(\xi)$. Condition (1.20) also holds if F is uniformly quasiconvex, i.e., if

$$(4.53)\quad F\left(\frac{x+y}{2}\right) \leqq \max\{F(x), F(y)\} - \delta(\|x-y\|)$$

for some nondecreasing function $\delta(\cdot): (0, \infty) \to (0, \infty)$, and all $x, y \in \Omega$. If ξ is a mini-

mizer of F in Ω, (4.53) yields

$$F(\xi) \leqq F\left(\frac{x+\xi}{2}\right) \leqq F(x) - \delta(\|x-\xi\|), \tag{4.54}$$

and therefore

$$F(x) - F(\xi) \geqq \delta(\|x-\xi\|) > 0 \tag{4.55}$$

for all $x \in \Omega - \{\xi\}$. It then follows that

$$\gamma(\sigma) = \inf_{\substack{x \in \Omega \\ \|x-\xi\|=\sigma}} F(x) - F(\xi) \geqq \delta(\sigma) > 0$$

for $\sigma > 0$. Evidently, (4.55) implies that ξ is a proper minimizer of F; however, no restrictions are imposed here on the compactness of Ω or the dimensionality of X.

As an illustration, consider the quadratic convex functional F defined for x in l^2 by

$$F(x) = \sum_{i=1}^{\infty} c_i x_i^2, \tag{4.56}$$

with uniformly bounded positive coefficients c_i. F has a unique minimizer in l^2 at $\xi = 0$, where $F(\xi) = 0$. Moreover,

$$(\inf_{1 \leqq i < \infty} c_i) \cdot \sigma^2 \leqq \gamma(\sigma) = \inf_{\substack{x \in l^2 \\ \|x\|=\sigma}} F(x) \leqq (\liminf_{i \to \infty} c_i) \cdot \sigma^2.$$

The left inequality in this expression is immediate from (4.56); the right inequality can be established by observing that

$$F(\sigma e^{(i)}) = \sigma^2 c_i,$$

where

$$e_j^{(i)} = \begin{cases} 0, & i \neq j, \\ 1, & i = j \end{cases}$$

and $\|e^{(i)}\| = 1$. If $\inf_{1 \leqq i < \infty} c_i = c > 0$, then (1.20) is satisfied with $\gamma(\sigma) \geqq c\sigma^2$ relative to the l^2-norm; in fact, one has

$$F\left(\frac{x+y}{2}\right) \leqq \frac{1}{2}(F(x) + F(y)) - \frac{1}{4} c\|x-y\|^2$$

for all $x, y \in \Omega$ with $x \neq y$. Hence F is uniformly convex (and a fortiori uniformly quasiconvex) on Ω. On the other hand, if $\lim_{i \to \infty} \inf c_i = 0$, then $\gamma(\sigma)$ vanishes identically *relative to the* l^2*-norm.* However, notice that for *arbitrary* bounded positive c_i's, F is itself the square of a weighted inner product-induced norm, and therefore $\gamma(\sigma) = \sigma^2$ with respect to that weighted norm. This underscores a fundamental point: in infinite dimensional spaces, the growth condition (1.20) is *norm dependent.*

Note 4.6. Let F be convex on Ω and let ξ be a minimizer of F in Ω. Suppose that Ω is *uniformly convex* in the sense that for some increasing function $\delta(\cdot) : (0, \infty) \to (0, \infty)$,

$$x, y \in \Omega \text{ and } \|z\| \leqq \delta(\|x-y\|) \Rightarrow \frac{x+y}{2} + z \in \Omega. \tag{4.57}$$

It then follows from [19, Theorem 3.4] that

$$F(x) - F(\xi) \geqq \langle \nabla F(\xi), x - \xi \rangle \geqq 2\|\nabla F(\xi)\| \cdot \delta(\sigma),$$

for $x \in \Omega$ and $\|x-\xi\| \geqq \sigma > 0$. Consequently, if $\nabla F(\xi) \neq 0$ the growth condition (1.20) holds with

$$\gamma(\sigma) \geqq 2\|\nabla F(\xi)\| \cdot \delta(\sigma) > 0.$$

In any Hilbert space, a ball with radius $R > 0$ satisfies the uniform convexity condition (4.57) with

$$\delta(\sigma) = (1/8R) \cdot \sigma^2.$$

(See [19, Note 3.4].)

More generally, suppose that F is convex on the convex set Ω, that ξ is a minimizer of F in Ω, and that ξ is *strongly nonsingular* in the sense of [19], i.e.,

$$0 < a(\sigma) = \inf_{\substack{x \in \Omega \\ \|x-\xi\| \geqq \sigma}} \langle \nabla F(\xi), x-\xi \rangle \tag{4.58}$$

for $\sigma > 0$. Then

$$F(x) - F(\xi) \geqq \langle \nabla F(\xi), x-\xi \rangle \geqq a(\sigma)$$

for all $x \in \Omega$ with $\|x-\xi\| \geqq \sigma > 0$, in which case the uniform growth condition (1.20) holds with

$$\gamma(\sigma) \geqq a(\sigma) > 0$$

for $\sigma > 0$. In [19], a strongly nonsingular extremal is said to be *regular* if $a(\sigma) \geqq A\sigma^2$ for some $A > 0$, and *strongly regular* if $a(\sigma) \geqq A\sigma$ for some $A > 0$. Roughly speaking, ξ will be regular if $\nabla F(\xi) \neq 0$, if ξ is on the boundary $\partial\Omega$ of Ω, and if the boundary has "positive curvature" at ξ; on the other hand, ξ will be strongly regular if ξ is a "vertex" in $\partial\Omega$ and $-\nabla F(\xi)$ lies in the *interior* of the cone of normals to Ω at ξ.

As a simple illustration, consider the following convex set in $\mathbb{R}^2$:

$$\Omega = \{(x_1, x_2) \in \mathbb{R}^2 | x_2 \geqq |x_1|^l\}$$

with fixed $l \geqq 1$. Suppose that a differentiable function $F: \mathbb{R}^2 \to \mathbb{R}^1$ has a minimizer at $\xi = (0, 0)$. If $l > 1$, condition (1.18) requires that

$$\frac{\partial F}{\partial x_1}(0) = 0 \quad \text{and} \quad \frac{\partial F}{\partial x_2}(0) \geqq 0,$$

in which case

$$\langle \nabla F(\xi), x-\xi \rangle = \frac{\partial F}{\partial x_2}(0) \cdot x_2 \geqq \frac{\partial F}{\partial x_2}(0) \cdot |x_1|^l$$

for all $x \in \Omega$. It then follows that, for $\sigma > 0$,

$$\begin{aligned}
\inf_{\substack{x \in \Omega \\ \|x-\xi\| \geqq \sigma}} \langle \nabla F(\xi), x-\xi \rangle &= \inf_{\substack{x \in \Omega \\ \|x-\xi\| = \sigma}} \langle \nabla F(\xi), x-\xi \rangle \\
&= \frac{\partial F}{\partial x_2}(0) \cdot x_2 \Big|_{x_1^2 + x_2^2 = \sigma^2, x_2 = |x_1|^l} \\
&= \frac{\partial F}{\partial x_2}(0) \cdot |x_1|^l \Big|_{x_1^2 + |x_1|^{2l} = \sigma^2} \\
&\geqq \frac{\partial F}{\partial x_2}(0) \cdot \frac{\sigma^l}{[1+\sigma^{2(l-1)}]^{l/2}}.
\end{aligned}$$

Therefore, ξ is strongly nonsingular for $l>1$ and $\dfrac{\partial F}{\partial x_2}(0)>0$. Moreover, if F is convex on Ω then $F(x)-F(\xi)\geqq\langle\nabla F(\xi), x-\xi\rangle$ for all $x\in\Omega$, and consequently F must satisfy the growth condition (1.20)–(4.27) with $\nu=l$ and some $B_\infty\geqq(\partial F/\partial x_2)(0)>0$. Notice that at $\xi=0$ the curvature of $\partial\Omega$ is zero for $l>2$, positive for $l=2$, and infinite for $2>l>1$. When $l=1$, the vector $\xi=0$ is a vertex of Ω, and the cone of normals to Ω at ξ is specified by

$$K_\Omega(0)=\{(y_1, y_2)\in\mathbb{R}^2 | y_2\leqq -|y_1|\}.$$

If F has an extremal at $\xi=0$ and if $-\nabla F(\xi)$ falls in the interior of $K_\Omega(0)$, i.e., if

$$\frac{\partial F}{\partial x_2}(0)\geqq\left|\frac{\partial F}{\partial x_1}(0)\right|+\varepsilon$$

for some $\varepsilon>0$, then

$$\begin{aligned}\inf_{\substack{x\in\Omega\\ \|x-\xi\|\geqq\sigma}}\langle\nabla F(\xi), x-\xi\rangle &= \inf_{\substack{x\in\Omega\\ \|x-\xi\|=\sigma}}\langle\nabla F(\xi), x-\xi\rangle\\ &= -\left|\frac{\partial F}{\partial x_1}(0)\right|\frac{\sqrt{2}}{2}\sigma+\frac{\partial F}{\partial x_2}(0)\frac{\sqrt{2}}{2}\sigma\\ &\geqq\frac{\varepsilon}{\sqrt{2}}\sigma,\end{aligned}$$

for all $\sigma>0$. Thus, $\xi=0$ is strongly regular for $l=1$ and $-\nabla F(\xi)\in\operatorname{int}K_\Omega(\xi)$.

Reference [19] contains a lengthy discussion of the foregoing extremal classification scheme and its connections with established notions of singularity and nonsingularity for nonlinear optimal control problems with bounded control vector components entering linearly into the equations of state. For such problems, the growth rate of the linear functional $\langle\nabla F(\xi), x-\xi\rangle$ near an extremal control $\xi(\cdot)$ is determined by the structure of the zero-crossing set θ for the switching function associated with $\xi(\cdot)$. In particular, if θ has zero measure, then $\xi(\cdot)$ satisfies the strong nonsingularity condition (4.58) in the Hilbert space $X=\mathscr{L}^2$. It is also shown in [19] that $\xi(\cdot)$ satisfies analogues of the foregoing regularity or strong regularity conditions in the Banach space $\mathscr{L}^1$, according to whether θ consists of finitely many simple zeros (bang–bang control) or is empty (constant control); while this $\mathscr{L}^1$-theory cannot be applied directly in the Hilbert space setting of the present analysis, it does give some insight into why projected gradient methods behave differently for optimal control problems which have bang–bang solutions and those which do not.

Note 4.7. If F has a continuous Hessian ∇^2F, then $\nabla^2F(x)$ is self-adjoint at each $x\in\Omega$, with

$$\begin{aligned}\|\nabla^2F(x)\| &= \sup_{\|h\|=1}\|\nabla^2F(x)h\|\\ &= \max\{|\mu_{\min}(x)|, |\mu_{\max}(x)|\}<\infty,\end{aligned}\tag{4.59}$$

where

$$\begin{aligned}\mu_{\min}(x) &= \inf_{\|h\|=1}\langle h, \nabla^2F(\xi)h\rangle\\ &\leqq \sup_{\|h\|=1}\langle h, \nabla^2F(\xi)h\rangle=\mu_{\max}(x).\end{aligned}\tag{4.60}$$

Moreover, according to the mean value theorem, ∇F is locally Lipschitz continuous on the convex set Ω. Therefore, the Lipschitz norms $L(\sigma)$ in (2.20) are finite for sufficiently small $\sigma > 0$, with

$$\lim_{\sigma \to 0^+} L(\sigma) \leqq \|\nabla^2 F(\xi)\|. \tag{4.61}$$

Under these circumstances, it is possible to replace L_∞ by $\|\nabla^2 F(\xi)\|$ in the α-threshold formulas (4.51) and (4.52) for the Goldstein rule (1.12) and analogous formulas for the Bertsekas–Armijo rule (1.11). In particular, if $\mu_{\min}(\xi) > 0$, then $\{\alpha_n\}$ sequences generated by (1.11) or (1.12) with $a_n \to \infty$, $\beta_n \to \beta_\infty$, and $\delta_n \to \delta_\infty$, satisfy the condition (4.17), with

$$\alpha_\infty \geqq \underline{\alpha}_\infty = \frac{2\beta_\infty(1-\delta_\infty)}{\mu_{\max}(\xi)} \tag{4.62}$$

or

$$\alpha_\infty \geqq \underline{\alpha}_\infty = \frac{2\delta_\infty}{\mu_{\max}(\xi)}, \tag{4.63}$$

respectively.

Let $Q(\xi, \cdot)$ denote the local quadratic approximation to $F(\cdot)$ at ξ; i.e.,

$$Q(\xi, x) = \langle \nabla F(\xi), x - \xi \rangle + \tfrac{1}{2}\langle x - \xi, \nabla^2 F(\xi)(x - \xi) \rangle. \tag{4.64}$$

The minimizer ξ is said to be *regular to second* order (cf. [22]) if and only if

$$Q(\xi, x) \geqq C\|x - \xi\|^2 \tag{4.65}$$

for some $C > 0$ and all $x \in \Omega$. If F is quasiconvex and ξ is regular to second order, then, according to Note 4.5 and (4.65), one has

$$\begin{aligned} \gamma(\sigma) &= \inf_{\substack{x \in \Omega \\ \|x-\xi\| \geqq \sigma}} F(x) - F(\xi) \\ &= \inf_{\substack{x \in \Omega \\ \|x-\xi\| = \sigma}} F(x) - F(\xi) \\ &\geqq \left(C + \frac{o(\sigma^2)}{\sigma^2} \right) \sigma^2. \end{aligned} \tag{4.66}$$

The growth condition (1.20)–(4.27) now follows with $\nu = 2$ and $B_\infty = C$. For ξ to be regular to second order it is sufficient (but not necessary) that $\mu_{\min}(\xi) > 0$, or that ξ is regular (Note 4.6) and $\mu_{\min}(\xi) \geqq 0$. In the former case, (4.65) is satisfied with $C = \frac{1}{2}\mu_{\min}(\xi)$ since the first term on the right in (4.64) is always nonnegative at the extremal ξ. If $\{\alpha_n\}$ is generated by the Bertsekas–Armijo rule (1.11) or the Goldstein rule (1.12), it then follows from (4.62), (4.63) and (4.66) that the corresponding asymptotic upper bound on r_{n+1}/r_n derived from (4.33) is a decreasing function of the ratio, $\rho(\xi) = \mu_{\min}(\xi)/\mu_{\max}(\xi)$. This suggests that larger values of $\rho(\xi)$ favor more rapid convergence of projected gradient sequences generated by the Bertsekas–Armijo or Goldstein rules when $\mu_{\min}(\xi) > 0$.

Convergence rate estimates derived from (4.33) for the Bertsekas–Armijo and Goldstein rules are actually quite conservative when $\Omega = X$, $F \in C^2(X, \mathbb{R}^1)$ and $\mu_{\min}(\xi) > 0$; sharper estimates are established for this special case in the next and final theorem.

THEOREM 4.5. *Let X be a real Hilbert space and let $F: X \to \mathbb{R}^1$ have a continuous Hessian, $\nabla^2 F$, with associated spectral limits*

$$\begin{aligned} \mu_{\min}(x) &= \inf_{\|h\|=1} \langle h, \nabla^2 F(x) h \rangle \\ &\leqq \sup_{\|h\|=1} \langle h, \nabla^2 F(x) h \rangle = \mu_{\max}(x). \end{aligned} \tag{4.67}$$

Suppose that ξ is an extremal of F in X, with

$$\mu_{\min}(\xi) > 0. \tag{4.68}$$

Furthermore, let $\{x_n\} \subset X$ be a gradient sequence satisfying (1.2), (1.13), *and* (1.14) *with $\{\delta_n\} \subset (0, \infty)$ and $\{\alpha_n\} \subset (0, \infty)$; i.e.,*

$$x_{n+1} = x_n - \alpha_n \nabla F_n,$$

$$F_n - F_{n+1} \geqq \delta_n \langle \nabla F_n, x_n - x_{n+1} \rangle,$$

$$\delta_n \geqq \delta, \qquad \alpha_n \geqq \alpha.$$

Finally, put $r(x) = F(x) - F(\xi)$, and suppose that

$$\lim_{n \to \infty} x_n = \xi. \tag{4.69}$$

Then either $x_n = \xi$ for all sufficiently large n, or else $r_n > 0$ for all n, with

$$0 \leqq \limsup_{n \to \infty} \frac{r_{n+1}}{r_n} \leqq 1 - 2(\delta\alpha)_\infty \mu_{\min}(\xi), \tag{4.70a}$$

where

$$\delta_\infty = \liminf_{n \to \infty} \delta_n \tag{4.70b}$$

and

$$\alpha_\infty = \liminf_{n \to \infty} \alpha_n. \tag{4.70c}$$

Proof. Suppose that $\{x_n\}$ and $\{\alpha_n\}$ satisfy (1.2) and that $x_N = \xi$. Since $\nabla F(\xi) = 0$ at the extremal ξ, one then has $x_{N+1} = x_N - \alpha_N \nabla F(x_N) = \xi$, and therefore $x_n = \xi$ for all $n \geqq N$, by induction. On the other hand, suppose that $\|x_n - \xi\| > 0$ for all n. In view of (4.68) the extremal ξ is a proper local minimizer of F, hence it follows from (4.69) that $r_n > 0$ for sufficiently large n. Moreover, (1.2) and (1.13) give

$$r_n - r_{n+1} \geqq \delta_n \langle \nabla F_n, x_n - x_{n+1} \rangle \geqq \delta_n \alpha_n \|\nabla F_n\|^2$$

for all n, and therefore

$$0 < \frac{r_{n+1}}{r_n} \leqq 1 - \frac{\delta_n \alpha_n \|\nabla F_n\|^2}{r_n} \tag{4.71}$$

for all sufficiently large n.

Since $\nabla^2 F$ is continuous, the quantity $\langle \nabla F(\xi + t(x - \xi)), u \rangle$, is continuously differentiable in t, with x and u fixed in X. Consequently,

$$\langle \nabla F(x), u \rangle = \int_0^1 \frac{d}{dt} \langle \nabla F(\xi + t(x - \xi)), u \rangle \, dt = \int_0^1 \langle u, H(t)(x - \xi) \rangle \, dt, \tag{4.72a}$$

where

$$H(t) = \nabla^2 F(\xi + t(x - \xi)) \tag{4.72b}$$

for $0 \leqq t \leqq 1$. Put $u = \nabla F(x)$ in (4.72) to obtain

$$\|\nabla F(x)\|^2 = \int_0^1 \langle \nabla F(x), H(t)(x - \xi) \rangle \, dt.$$

Put $u = H(t)(x - \xi)$ and apply (4.72) once again to get

$$\|\nabla F(x)\|^2 = \int_0^1 \int_0^1 \langle H(t)(x - \xi), H(\tau)(x - \xi) \rangle \, d\tau \, dt. \tag{4.73}$$

Observe now that Taylor's formula gives

$$F(x) - F(\xi) = \tfrac{1}{2}\langle x - \xi, H(t^*)(x - \xi) \rangle \geqq \tfrac{1}{2}\mu_{\min}(x^*)\|x - \xi\|^2 \tag{4.74}$$

for some $t^* \in [0, 1]$, with $x^* = \xi + t^*(x - \xi)$. For $0 \leqq t \leqq 1$, put

$$\bar{\psi}(t) = H(t) - H(t^*) \tag{4.75}$$

in (4.73) to obtain

$$\|\nabla F(x)\|^2 \geqq \|H(t^*)(x - \xi)\|^2 - \psi(x)\|x - \xi\|^2, \tag{4.76a}$$

where

$$\psi(x) = 2\|H(t^*)\| \cdot \max_{0 \leqq t \leqq 1} \|\bar{\psi}(t)\| + (\max_{0 \leqq t \leqq 1} \|\bar{\psi}(t)\|)^2. \tag{4.76b}$$

Since $\nabla^2 F$ is continuous, the spectral limit $\mu_{\min}(x)$ is also continuous and therefore $\mu_{\min}(z) > 0$ for all z sufficiently near ξ. Consequently, (4.74), (4.76) and the Schwarz inequality yield

$$\begin{aligned} \|\nabla F(x)\|^2 &\geqq \frac{\langle x - \xi, H(t^*)(x - \xi) \rangle^2}{\|x - \xi\|^2} - \psi(x)\|x - \xi\|^2 \\ &\geqq 2\left(\mu_{\min}(x^*) - \frac{\psi(x)}{\mu_{\min}(x^*)}\right) r(x) \end{aligned} \tag{4.77}$$

for x sufficiently near ξ. In view of (4.71) one therefore has

$$0 < \frac{r_{n+1}}{r_n} \leqq 1 - 2\delta_n \alpha_n \left(\mu_{\min}(x_n^*) - \frac{\psi(x_n)}{\mu_{\min}(x_n^*)}\right) \tag{4.78}$$

for n sufficiently large. Finally, since $\nabla^2 F$ is continuous it follows that $\psi(x_n) \to 0$ and $\mu_{\min}(x_n^*) \to \mu_{\min}(\xi) > 0$ as $n \to \infty$. The estimate (4.70) is now immediate from (4.78). □

Note 4.8. If the hypotheses of Theorem 4.5 hold, and if the gradient sequence $\{x_n\} \subset X$ is generated by the Bertsekas–Armijo rule (1.11) with $\lim_{n \to \infty} a_n = \infty$, $\lim_{n \to \infty} \beta_n = \beta_\infty \in (0, 1]$, and $\lim_{n \to \infty} \delta_n = \delta_\infty \in (0, 1)$, then it follows from (4.62) and (4.70) that

$$0 \leqq \limsup \frac{r_{n+1}}{r_n} \leqq 1 - C_\infty \cdot \frac{\mu_{\min}(\xi)}{\mu_{\max}(\xi)}, \tag{4.79a}$$

with

$$C_\infty = [4\beta\delta(1 - \delta)]_\infty. \tag{4.79b}$$

Similarly, for the Goldstein rule (1.12), the inequalities (4.63) and (4.70) yield (4.79a) once again, with

$$C_\infty = 4\delta_\infty^2, \tag{4.79c}$$

provided $\lim_{n\to\infty} a_n = \infty$ and $\lim_{n\to\infty} \delta_n = \delta_\infty \in (0, \frac{1}{2}]$. In both cases, the best bound is obtained when $\delta_\infty = \frac{1}{2}$. For the Goldstein rule, one then has

$$0 \leqq \lim_{n\to\infty} \sup \frac{r_{n+1}}{r_n} \leqq 1 - \frac{\mu_{\min}(\xi)}{\mu_{\max}(\xi)}, \tag{4.80}$$

and this estimate also applies to the Bertsekas–Armijo rule if $\beta_\infty = 1$. Notice that r_n converges *superlinearly* to 0 according to (4.80), if $\mu_{\max}(\xi) = \mu_{\min}(\xi) > 0$.

The inequality

$$\frac{r_{n+1}}{r_n} \leqq 1 - \frac{\mu_{\min}(\xi)}{\mu_{\max}(\xi)}$$

is a classic result for quadratic F and gradient sequences generated by the line minimization step length rule (cf. [28]); that the right side of this expression should also appear in the asymptotic bound (4.80) is not surprising if one considers that (1.11) and (1.12) *approximate* the line minimization rule more and more accurately as $n \to \infty$ when $a_n \to \infty$, $\delta_n \to \frac{1}{2}$, $\beta_n \to 1$ and $x_n \to \xi$.

REFERENCES

[1] A. A. GOLDSTEIN, *Convex programming in Hilbert space*, Bull. Amer. Math. Soc., 70 (1964), pp. 709–710.

[2] E. S. LEVITIN AND B. T. POLYAK, *Constrained minimization problems*, USSR J. Comput. Math. Phys., 6 (1966), pp. 1–50.

[3] V. F. DEMYANOV AND A. M. RUBINOV, *Approximate Methods in Optimization Problems*, American Elsevier, New York, 1971.

[4] J. B. ROSEN, *The gradient projection method for nonlinear programming: Part I, linear constraints*, SIAM J. Appl. Math., 8 (1960), pp. 181–217.

[5] ———, *The gradient projection method for nonlinear programming: Part II, nonlinear constraints*, SIAM J. Appl. Math., 9 (1961), pp. 514–532.

[6] D. G. LUENBERGER, *The gradient projection method along geodesics*, Management Sci., 18 (1972), pp. 620–631.

[7] E. POLAK, *An historical survey of computational methods in optimal control*, SIAM Rev., 15 (1973), pp. 553–584.

[8] R. E. BRUCK, JR., *The iterative solution of the equation $y \in x + Tx$ for a monotone operator T in Hilbert space*, Bull. Amer. Math. Soc., 79 (1973), pp. 1258–1261.

[9] B. E. RHOADES, *Fixed point iterations using infinite matrices, III*, in Proceedings of the Conference on Computing Fixed Points with Applications, Clemson University, Academic Press, New York, 1974.

[10] J. C. DUNN, *Iterative construction of fixed points for multi-valued operators of the monotone type*, J. Functional Anal., 27 (1978), pp. 38–50.

[11] G. P. MCCORMICK AND R. A. TAPIA, *The gradient projection method under mild differentiability conditions*, SIAM J. Control, 10 (1972), pp. 93–98.

[12] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1969.

[13] A. A. GOLDSTEIN, *Constructive Real Analysis*, Harper and Row, New York, 1967.

[14] D. P. BERTSEKAS, *On the Goldstein–Levitin–Polyak gradient projection method*, IEEE Trans. Autom. Contr., AC-21 (1976), pp. 174–184.

[15] L. ARMIJO, *Minimization of functionals having continuous partial derivatives*, Pacific J. Math., 16 (1966), pp. 1–3.

[16] A. A. GOLDSTEIN, *On steepest descent*, SIAM J. Control, Ser. A, 3 (1965), pp. 147–151.

[17] ———, *On gradient projection*, Proc. 12th Allerton Conf. Circuit and System Theory, University of Illinois, 1974, pp. 38–40.

[18] O. L. MANGASARIAN, *Pseudo-convex functions*, SIAM J. Control, Ser. A, 3 (1965), pp. 281–290.
[19] J. C. DUNN, *Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals*, this Journal, 17 (1979), pp. 187–211.
[20] ———, *Convergence rates for conditional gradient sequences generated by implicit step length rules*, this Journal, 18 (1980), pp. 473–487.
[21] M. D. CANNON AND C. D. CULLUM, *A tight upper bound on the rate of convergence of the Frank–Wolfe algorithm*, SIAM J. Control, Ser. A, 6 (1968), pp. 509–516.
[22] J. C. DUNN, *Newton's method and the Goldstein step length rule for constrained minimization problems*, this Journal, 18 (1980), pp. 659–674.
[23] L. V. KANTOROVICH, *Functional analysis and applied mathematics*, Uspehi Mat. Nauk 3 (1948), pp. 89–185.
[24] M. M. VAINBERG, *On the convergence of the method of steepest descents for nonlinear equations*, Soviet Math. Dokl., 1 (1960), pp. 1–4.
[25] B. T. POLYAK, *Gradient methods for the minimization of functionals*, USSR J. Comput. Math. Phys., 3 (1963), pp. 643–653.
[26] M. Z. NASHED, *Steepest descent for singular linear operator equations*, SIAM J. Numer. Anal., 7 (1970), pp. 358–362.
[27] A. A. GOLDSTEIN, *Minimizing functionals on Hilbert space*, in Computer Methods in Optimization Problems, Academic Press, New York, 1964.
[28] D. G. LUENBERGER, *Vector Space Methods in Optimization*, John Wiley, New York, 1969.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 3, May 1981

0363-0129/81/1903-0005 $01.00/0

STOCHASTIC CONTROL ON HILBERT SPACE FOR LINEAR EVOLUTION EQUATIONS WITH RANDOM OPERATOR-VALUED COEFFICIENTS*

N. U. AHMED†

Abstract. We consider a problem of optimal control of the stochastic evolution equation $d\xi = (A(t)\xi + B(t)u)\,dt + \sigma(t)\,dw$, on a separable Hilbert space, where $\{A(t), B(t), \sigma(t), t \geqq 0\}$ are progressively measurable operator-valued random processes with A generally unbounded. We prove the existence and uniqueness of (weak) solutions of the evolution equation. Then we present the existence of optimal controls and necessary conditions of optimality for a quadratic (random) cost function. For optimal feedback controls we solve a random operator Riccati equation and a backward stochastic evolution equation. The backward equation is solved by transposing a random isomorphism generated from a forward evolution equation. The optimal feedback control is given by a random affine transformation of the state. Some examples are presented to indicate usefulness of the results. This work is a partial extension of the results of Bismut [SIAM J. Control Optim., 14 (1976), pp. 419–444; 15 (1977), pp. 1–4] and Bensoussan and Voit [SIAM J. Control Optim., 13 (1975), pp. 904–926].

1. Introduction. We consider the problem of optimal control for a class of stochastic linear evolution equations on a Hilbert space with random operator-valued coefficients. The cost function is assumed to be quadratic with random operator-valued weighting functions. In § 2 we prove the existence and uniqueness of solutions of the system equations. In § 3 existence of optimal controls and necessary conditions of optimality are presented. In § 4 we prove the existence and uniqueness of solutions of a class of backward stochastic evolution equations and use this result to rewrite the necessary (sufficient) conditions of § 3 in terms of the familiar adjoint state. In § 5 we consider the question of existence and uniqueness of solutions of a stochastic operator-Riccati integral (or differential) equation arising from the decoupled feedback control system. The solution of the integral (or differential) equation defines the optimal control as a random affine transformation of the state. In § 6 we present a few interesting special cases as examples.

Our results extend certain recent results of Bismut [2], [3] and Bensoussan and Voit [1]. In the former the system is governed by a finite dimensional stochastic linear differential equation with random matrix-valued coefficients, and in the latter the system is a stochastic evolution equation with deterministic operator-valued coefficients. Our system model covers that of Bensoussan and Voit but not that of Bismut since we do not consider state dependent noise. However our results can be readily extended to this case at the cost of cumbersome notation. Our Example (i) includes the results of Sworder [7], [8], and Example (iii) covers Example 4 of Bismut as a special case.

2. Notation, formulation of the problem and solutions of stochastic evolution equations. Let H be a separable Hilbert space and V a linear subspace of H having the structure of a reflexive Banach space and dense in H. Identify H with its adjoint H'; then $V \subset H \subset V'$, where V' is the dual of V. We use the notation $\langle \cdot, \cdot \rangle$ for $V' - V$ duality pairing and $(\cdot, \cdot)_F$ for scalar product in any Hilbert space F with the associated norm $|\cdot|_F$. In general, for any Banach space K with the dual K', we write the corresponding duality pairing as $(\cdot, \cdot)_{K'-K}$ and the norm as $|\cdot|_K$. Let (Ω, β, μ) be a

* Received by the editors October 8, 1979, and in revised form July 31, 1980. A preliminary version of this paper was presented at the Ninth Conference on Stochastic Processes and Their Applications, August 6–10, 1979, Northwestern University, Evanston, Illinois 60201.

† Department of Electrical Engineering, University of Ottawa, Ottawa, Ontario K1N 6N5, Canada.

complete probability space, K a real Banach space, $p \in [1, \infty]$ and $L_p(\Omega, K)$ the equivalence classes of strongly measurable functions on Ω with values in K so that $|\cdot|_K^p$ is μ-integrable on Ω. The norm in $L_p(\Omega, K)$ is denoted by

$$\|f\|_{L_p(\Omega,K)} = \left(\int_\Omega |f|_K^p \, d\mu\right)^{1/p} \equiv (E|f|_K^p)^{1/p},$$

where $E(\cdot)$ denotes the mathematical expectation of its argument. When K is the real line and $p = 2$, $L_2(\Omega, R)$ is the space of square integrable real-valued random variables. It is clear that the Banach spaces $L_2(\Omega, V)$, $L_2(\Omega, H)$ and $L_2(\Omega, V')$ are well defined, and they are Hilbert spaces whenever V is a Hilbert space. It is known that if K is a separable Banach space then strong and weak measurability are equivalent; in that case we will omit the qualifying statement.

Let β_t, $t \in I$ be an increasing family of complete subsigma algebras of the sigma algebra β satisfying the right continuity property (Meyer [9, D30, p. 65]). Let $\mathcal{F}$ denote the σ-field of the progressively measurable subsets (Meyer [9, D45, p. 68; 50, p. 71]) of the set $I \times \Omega$, and $\mathcal{F}^*$ its completion with respect to the measure $dt \times d\mu$. For K a Banach space and $p \in [1, \infty]$, we denote by $L_p^\omega(I, K)$ the space of equivalence classes of strongly $\mathcal{F}^*$-measurable random processes on I with values in K so that for any $f \in L_p^\omega(I, K)$, $E\int_I |f(t)|_K^p \, dt < \infty$. The space $L_p^\omega(I, K)$ equipped with the norm $\|f\|_{L_p^\omega(I,K)} \equiv (E\int_I |f(t)|_K^p \, dt)^{1/p}$ becomes a Banach space. For K Hilbert and $p = 2$, this is a Hilbert space. For example, if V is a Hilbert space, then $L_2^\omega(I, V)$, $L_2^\omega(I, H)$ and $L_2^\omega(I, V')$ are all Hilbert spaces of $\mathcal{F}^*$-measurable abstract random processes. By $L_p(\beta_t, K)$ we mean the equivalence classes of K-valued strongly β_t-measurable random variables with $\int_\Omega |f|_K^p \, d\mu_t < \infty$, where μ_t is the probability measure restricted to the σ-algebra β_t. It is clear that for $t_1 < t_2$, $L_p(\beta_{t_1}, K) \subset L_p(\beta_{t_2}, K)$ and that they are closed subspaces of the space $L_p(\beta, K)$. In the sequel we will make extensive use of these spaces.

For any pair of normed linear spaces X and Y we use $\mathcal{L}(X, Y)$ to denote the normed linear space of bounded linear operators from X into Y. Note that if Y is a Banach space then so also is $\mathcal{L}(X, Y)$. Let $L_\infty(\mathcal{F}^*, \mathcal{L}(X, Y))$ denote the class of strongly $\mathcal{F}^*$-measurable operator-valued functions on $I \times \Omega$ with values in $\mathcal{L}(X, Y)$ so that for any $T \in L_\infty(\mathcal{F}^*, \mathcal{L}(X, Y))$, $\|T(t, \omega)\|_{\mathcal{L}(X,Y)}$ is $dt \times d\mu$ essentially bounded on $I \times \Omega$ and $(t, \omega) \to T(t, \omega)x$ is $\mathcal{F}^*$-measurable for each $x \in X$.

We consider the optimal control problem for the system governed by the following stochastic evolution equation:

S
$$d\xi = [A(t)\xi + B(t)u]\, dt + \sigma(t)\, dW(t), \qquad t \in (0, T),$$
$$\xi(0) = \xi_0 \in L_2(\beta_0, H),$$

where A, B, σ are operator-valued stochastic processes on I and ξ and u are the states and controls respectively. The problem is to find a control u from an admissible class (to be defined shortly) that minimizes the cost function

$$J(u) = E\int_0^T \{|C\xi - Z_d|_{\mathcal{H}}^2 + (Nu, u)\}\, dt, \tag{2.1}$$

where C and N are also operator-valued stochastic processes on I and Z_d is a given $\mathcal{F}^*$-measurable stochastic process defined on I with values in a real separable Hilbert space $\mathcal{H}$.

It will be assumed throughout the paper that all the random processes A, B, σ, C, N, W, Z_d are $\mathcal{F}^*$-measurable; that is, for each $t \in I$, $A(t)$, $B(t)$, $\sigma(t)$, $C(t)$, $N(t)$, $W(t)$ and $Z_d(t)$ are all β_t-measurable (or equivalently adapted to β_t). The σ-algebras β_t are

usually generated by the inverse images of topological Borel-fields of the topological spaces in which the random variables take their values at time t.

For the solution of the control problem we will introduce the following basic assumptions:

(Ai) $A(t), t \in [0, T] \equiv I, 0 < T < \infty$, is a family of random evolution operators from V to V' such that, for each $x, y \in V$, $(t, \omega) \to \langle A(t, \omega)x, y\rangle$ is $\mathcal{F}^*$-measurable on $[0, T] \times \Omega$ and $dt \times d\mu$ essentially bounded; i.e., $A \in L_\infty(\mathcal{F}^*, \mathcal{L}(V, V'))$. In the rest of the paper we will suppress the variable ω.

(Aii) There exist $\lambda \geqq 0$ and $\alpha > 0$ such that, for each $x \in L_2(\beta_t, V)$,

$$-E\langle A(t)x, x\rangle + \lambda E|x|_H^2 \geqq \alpha E|x|_V^2 \quad \text{uniformly in } t \text{ on } I.$$

(Aiii) $\{W(t), t \in I\}$ is an H-valued Wiener process with the usual properties:

(a) $W(0) = 0 \quad \mu - \text{a.e.}$

(b) $E\{(W(t) - W(s), h)|\beta_s\} = 0$ for all $h \in H$ and $t \geqq s$.

(c) There exists a positive self-adjoint operator $Q \in L_\infty(I, \mathcal{L}(H, H))$ such that for all $h, f \in H$

$$E\{(W(t) - W(s), h) \cdot (W(t) - W(s), f)|\beta_s\} = \int_s^t (Q(\theta)h, f)\, d\theta,$$

where $E\{\cdot|\beta_s\}$ denotes the conditional expectation with respect to the σ-algebra β_s.

(Aiv) $\{\sigma(t), t \in I\}$ is a family of β_t-measurable bounded linear operators from H to V', or equivalently σ is an $\mathcal{F}^*$-measurable $\mathcal{L}(H, V')$-valued random variable so that, for any arbitrary sequence of basis vectors $\{v_i\} \subset V$ with $|v_i|_H = 1$, $i = 1, 2, \cdots$,

$$\sum_{i=1}^{\infty} \int_I E\langle \sigma(\theta)Q(\theta)\sigma^*(\theta)v_i, v_i\rangle\, d\theta < \infty,$$

where σ^* is the adjoint of the operator σ. Clearly by definition $\sigma(t)$ is independent of $\sigma\{W(\theta) - W(s), \theta \geqq s \geqq t\}$ for all $t \in I$.

(Av) F is a real separable Hilbert space and $\{B(t), t \in I\}$ is a family of β_t-measurable bounded linear operators with values in $\mathcal{L}(F, V')$, or equivalently B is a $\mathcal{F}^*$-measurable $\mathcal{L}(F, V')$-valued function on $I \times \Omega$ with

$$\text{ess sup}\{\|B(t)\|_{\mathcal{L}(F,V')}, (t, \omega) \in I \times \Omega\} \leqq b < \infty,$$

that is,

$$B \in L_\infty(\mathcal{F}^*, \mathcal{L}(F, V')).$$

(Avi) With F as in (Av) let $L_2^\omega(I, F)$ denote the equivalence classes of (norm)-square integrable $\mathcal{F}^*$-measurable random variables with values in the Hilbert space F and equipped with the norm

$$\|u\|_{L_2^\omega(I,F)} = \left(E\int_I |u(t)|_F^2\, dt\right)^{1/2},$$

which, by Fubini's theorem, is equal to $(\int_I E|u(t)|_F^2\, dt)^{1/2}$. For the class of admissible controls we take any closed convex subset $\mathcal{U}_a$ of $\mathcal{U} \equiv L_2^\omega(I, F)$.

It can be shown, as in the stochastic calculus for finite dimensional spaces, that for the Wiener process W satisfying (Aiii) and the operator σ satisfying (Aiv), the stochastic integral $\int_I \sigma(t)\, dW(t)$ is a well-defined random variable with values in V' and that $\nu(t) \equiv \int_0^t \sigma(\theta)\, dW(\theta)$ is a V'-valued β_t martingale. Let $M = M(\mathcal{F}^*, \mathcal{L}(H, V'))$ denote the class of progressively measurable random processes on $I \times \Omega$ with values in

$\mathscr{L}(H, V')$ and suppose M is given the topology $\mathscr{T}_M$ induced by the system of neighborhoods

$$N_\varepsilon(\sigma_0) \equiv \left\{ \sigma \in M \,\middle|\, \int_I E \operatorname{Tr} (\sigma - \sigma_0) Q (\sigma - \sigma_0)\, dt < \varepsilon \right\}, \qquad \sigma_0 \in M, \quad \varepsilon > 0.$$

With this preparation we can now present a result on the existence of a solution of the stochastic evolution equation S. Let X be a separable Banach space, and let $C^\omega_{1,T}(I, X)$ denote the class of once continuously differentiable progressively measurable random processes on $I \times \Omega$ with values in X vanishing at $t = T$ with probability one.

By a solution of the evolution equation S we mean any function $\xi \in L^\omega_2(I, V) \cap L^\omega_\infty(I, H)$ such that

$$\begin{aligned} -\int_0^T \langle \xi(t), \dot{\phi}(t) \rangle\, dt = (\xi_0, \phi(0)) + \int_0^T \{ \langle A^* \phi, \xi \rangle + \langle Bu, \phi \rangle \}\, dt \\ + \int_0^T \langle \phi, \sigma\, dW \rangle \quad \mu\text{-a.e.} \end{aligned} \tag{2.2}$$

for every $\phi \in C^\omega_{1,T}(I, V)$.

We shall refer to this solution as the weak solution.

THEOREM 2.1. *Consider the system* S *and suppose the operator* A *satisfies* (Ai) *and* (Aii), *the injection map* $V \hookrightarrow H$ *is compact, the operator* B *satisfies* (Av), *the operator* σ *satisfies* (Aiv) *and belongs to* M, *and the Wiener process* W *satisfies* (Aiii). *Then for each* $\xi_0 \in L_2(\beta_0, H)$ *and* $u \in \mathscr{U}_a \subset L^\omega_2(I, F)$ *the system* S *has a unique (weak) solution* $\xi \in L^\omega_2(I, V) \cap L^\omega_\infty(I, H)$, *and that* $u \to \xi$ *is an affine continuous map from* $L^\omega_2(I, F)$ *into* $L^\omega_2(I, V)$. *More generally,* $(\xi_0, u, \sigma) \to \xi$ *is a continuous map from* $L_2(\beta_0, H) \times L^\omega_2(I, F) \times M$ *into* $L^\omega_2(I, V)$.

Proof. Defining $\xi(t) = \eta(t)\, e^{-\lambda t}$ one can verify that the system equation S reduces to

$$d\eta(t) = (A(t) - \lambda I)\eta\, dt + e^{-\lambda t} B(t) u(t)\, dt + e^{-\lambda t} \sigma(t)\, dW(t).$$

Thus in (Aii) λ can be assumed to be zero without loss of generality, and from now on we will do so.

For existence, since the injection map of V into H is compact one can construct a common basis for V, H and V'. Let Λ be the canonical isomorphism of V onto V' so that $\langle \Lambda x, y \rangle_{V'-V} = (x, y)_V$ for all $x, y \in V$. Then Λ^{-1} is linear continuous from V' into V and consequently from H into V, and therefore a linear self-adjoint compact operator in H. As a consequence there exists a sequence of orthonormal vectors $\{v_i\}$ in H forming the eigenvectors of the operator Λ^{-1} such that

$$\begin{aligned} &\Lambda v_i = \rho_i^2 v_i, \\ &(v_i, v_j)_H = \delta_{ij}, \qquad \rho_i^2 > 0, \quad i, j = 1, 2, \cdots. \end{aligned} \tag{2.3}$$

The sequence $\{v_i\}$ also forms an orthogonal basis for V and V' so that

$$\begin{aligned} &(v_i, v_j)_V = \langle \Lambda v_i, v_j \rangle = \rho_i^2 \delta_{ij}, \\ &(v_i, v_j)_{V'} = (\Lambda^{-1} v_i, v_j) = (1/\rho_i^2) \delta_{ij}. \end{aligned} \tag{2.4}$$

Since V is also dense in H and $\xi_0 \in L_2(\beta_0, H)$ we can choose a sequence of square integrable real random variables $\{y_i\}$ so that

$$\xi_0^n = \sum_{i=1}^n y_i v_i \in L_2(\beta_0, V) \subset L_2(\beta_0, H)$$

and

$$\xi_0^n \to \xi_0 \text{ strongly in } L_2(\beta_0, H).$$

Clearly $\sum_{i=1}^{\infty} E(y_i)^2 < \infty$. Define an approximate solution of the problem S by

$$\xi^n(t) \equiv \sum_{i=1}^{n} x_i^n(t) v_i,$$

in the sense that

S_n

$$\langle d\xi^n(t), v_j\rangle = \{\langle A(t)\xi^n(t), v_j\rangle + \langle B(t)u(t), v_j\rangle\}\, dt + (\sigma^*(t)v_j, dW(t)),$$

$$\xi^n(0) = \sum_{i=1}^{n} x_i^n(0) v_i = \sum_{i=1}^{n} y_i v_i, \qquad 1 \leqq j \leqq n.$$

Equivalently the system S_n can be written in the form of a system of linear stochastic differential equations of the form

$\tilde{\mathrm{S}}_n$

$$dx^n(t) = [\mathscr{A}^n(t)x^n(t) + f^n(t)]\, dt + dm^n(t),$$

$$x^n(0) = y^n,$$

where $x^n(t) = (x_1^n(t), x_2^n(t), \cdots, x_n^n(t))'$, $\mathscr{A}^n$ is a $(n \times n)$ matrix-valued function with elements $\mathscr{A}_{ij}(t) = \langle A(t)v_i, v_j\rangle$, $1 \leqq i,\ j \leqq n$, $f^n(t) = (f_1(t), \cdots, f_n(t))'$, with $f_i(t) \equiv \langle B(t)u(t), v_i\rangle$, $m^n(t) = (m_1(t), \cdots, m_n(t))'$ with $m_i(t) \equiv \int_0^t (\sigma^*(\theta)v_i, dW(\theta))_H$ and $y^n = (y_1, \cdots, y_n)'$. Here "$'$" denotes transposition. By a result due to Bismut we will be able to justify the existence of a solution to the problem $\tilde{\mathrm{S}}_n$. For this we note the properties of the coefficients of the systems $\tilde{\mathrm{S}}_n$.

(i) Since $A \in L_\infty(\mathscr{F}^*, \mathscr{L}(V, V'))$, the elements of the matrix-valued function $\mathscr{A}^n$ are $dt \times d\mu$ essentially bounded.

(ii) Since $B \in L_\infty(\mathscr{F}^*, \mathscr{L}(F, V'))$, $u \in L_2(I, F)$ there exists a constant k dependent on the bound b of the norm of the operator B and the Lebesgue measure $l(I)$ of the interval I so that

$$E\left(\int_I |f^n(t)|\, dt\right)^2 \leqq k\|u\|^2_{L_2^\omega(I,F)}. \tag{2.5}$$

(iii) By hypothesis, the operator $\sigma \in L_\infty(\mathscr{F}^*, \mathscr{L}(H, V'))$, $Q \in L_\infty(I, \mathscr{L}(H, H))$ and $\sum_{i=1}^{\infty} \int_I E\langle \sigma(t)Q(t)\sigma^*(t)v_i, v_i\rangle\, dt < \infty$. Therefore

$$E|m^n(t)|^2 \equiv \sum_{i=1}^{n} E(m_i^n(t))^2 = \sum_{i=1}^{n} \int_0^t E\langle \sigma Q \sigma^* v_i, v_i\rangle\, d\theta < \infty. \tag{2.6}$$

Since W is a Wiener process and $\sigma \in L_\infty(\mathscr{F}^*, \mathscr{L}(H, V'))$, that is, for each $t \in I$, $\sigma(t)$ is β_t-measurable with values in $\mathscr{L}(H, V')$, the process $\nu(t) \equiv \int_0^t \sigma(\theta)\, dW(\theta)$ is a V'-valued β_t martingale. Consequently $m_i(t) \equiv \int_0^t (\sigma^*(\theta)v_i,\ dW(\theta))$, $i = 1, 2, \cdots$, is a scalar-valued β_t martingale for each $v_i \in V$. Therefore it follows from (2.6) that $m^n(t)$ is an R^n-valued square integrable β_t martingale with

$$E|m^n(t)|^2 \leqq \sum_{i=1}^{n} \int_0^t E(Q\sigma^* v_i, \sigma^* v_i)\, d\theta < \infty. \tag{2.7}$$

(iv) Clearly $E|y^n|^2 \equiv \sum_{i=1}^{n} E(y_i)^2 < \infty$. Let $L_{2,1}$ denote the space of n-vector-valued $\mathscr{F}^*$-measurable stochastic processes $\{f\}$ defined on I so that $E(\int_I |f(t)|\, dt)^2 < \infty$. This is a Banach space and was introduced by Bismut [2, p. 421]. Under conditions (i)–(iv), it follows from a result due to Bismut [2, Theorem A, p. 441] that the system $\tilde{\mathrm{S}}_n$

has a unique solution x^n which is continuous with probability one and has bounded second moment (considering $B=0$, $v=0$ in Bismut [2, A1, p. 441]. The continuity of the solution x^n follows from the fact that the martingale m^n is itself continuous with probability one. Thus $\xi^n \in L_2^\omega(I, V) \cap C^\omega(I, H)$, where $C^\omega(I, H)$ is the Banach space of continuous second order random processes defined on I, adapted to $\{\beta_t, t \geqq 0\}$ and assuming values in H.

By application of Ito's lemma to the function $q(x^n(t)) \equiv |x^n(t)|^2$, where x^n is the solution of the finite dimensional system equation $\tilde{S}_n$, it is easily verified that

$$d|x^n(t)|^2 = 2\Big[(\mathscr{A}^n(t)x^n(t), x^n(t)) + (x^n(t), f^n(t)) + \frac{1}{2}\sum_{i=1}^{n}\langle\sigma(t)Q\sigma^*(t)v_i, v_i\rangle\Big]dt + 2(x^n(t), dm^n(t)). \tag{2.8}$$

Equation (2.8) is equivalent to

$$d|\xi^n(t)|_H^2 = 2\Big[\langle A\xi^n, \xi^n\rangle + \langle Bu, \xi^n\rangle + \frac{1}{2}\sum_{i=1}^{n}\langle\sigma Q\sigma^* v_i, v_i\rangle\Big]dt + 2(\sigma^*\xi^n, dW)_H. \tag{2.9}$$

Since $Q \in L_\infty(I, \mathscr{L}(H, H))$ and $\sigma \in L_\infty(\mathscr{F}^*, \mathscr{L}(H, V'))$ with $\sigma^* \in L_\infty(\mathscr{F}^*, \mathscr{L}(V, H))$, it is clear that

$$\text{ess sup}\{\|\sigma(t)Q(t)\sigma^*(t)\|_{\mathscr{L}(V,V')}, (t, \omega) \in I \times \Omega\} < \infty,$$

and therefore there exists a nonnegative finite number δ so that

$$E\Big(\int_I (\sigma^*\xi^n, dW)\Big)^2 = E\int_I \langle\sigma Q\sigma^*\xi^n, \xi^n\rangle\, dt \leqq \delta E\int_I |\xi^n(t)|_V^2\, dt.$$

Thus, if we recall the martingale property of the Wiener process W it follows that $E\int_I(\sigma^*\xi^n, dW) = 0$ and $E\int_0^t(\sigma^*\xi^n, dW) = 0$ for every $t \in I$. Using this fact in (2.9), we obtain

$$E|\xi^n(t)|_H^2 - 2E\int_0^t \langle A(\theta)\xi^n(\theta), \xi^n(\theta)\rangle\, d\theta = E|\xi_0^n|_H^2 + 2E\int_0^t \langle B(\theta)u(\theta), \xi^n(\theta)\rangle\, d\theta + \sum_{i=1}^{n} E\int_0^t \langle\sigma Q\sigma^* v_i, v_i\rangle\, d\theta. \tag{2.10}$$

Utilizing (Aii), Fubini's theorem, and the Schwarz inequality we can reduce (2.10) to

$$E|\xi^n(t)|_H^2 + 2\alpha\int_0^t E|\xi^n(\theta)|_V^2\, d\theta \leqq E|\xi_0^n|_H^2 + 2\Big(\int_0^t E|Bu|_{V'}^2\, d\theta\Big)^{1/2}\Big(\int_0^t E|\xi^n|_V^2\, d\theta\Big)^{1/2} + \sum_{i=1}^{n}\int_0^t E\langle\sigma Q\sigma^* v_i, v_i\rangle\, d\theta. \tag{2.11}$$

Further, using the elementary inequality

$$ab \leqq \frac{1}{2\varepsilon}a^2+\frac{\varepsilon}{2}b^2, \qquad a, b \in R,$$

which is valid for every $\varepsilon > 0$, we obtain from (2.11)

$$E|\xi^n(t)|_H^2+2\left(\alpha-\frac{\varepsilon}{2}\right)\int_0^t E|\xi^n(\theta)|_V^2\,d\theta$$

$$\leqq E|\xi_0^n|_H^2+\frac{1}{\varepsilon}\int_0^t E|(Bu)(\theta)|_{V'}^2\,d\theta+\sum_{i=1}^{n}\int_0^t E\langle\sigma Q\sigma^* v_i, v_i\rangle\,d\theta, \tag{2.12}$$

since ε is an arbitrary positive number, for $\varepsilon = \alpha$ we have

$$E|\xi^n(t)|_H^2+\alpha\int_0^t E|\xi^n(\theta)|_V^2\,d\theta$$

$$\leqq E|\xi_0^n|_H^2+\frac{1}{\alpha}\int_0^t E|(Bu)(\theta)|_{V'}^2\,d\theta+\sum_{i=1}^{n}\int_0^t E\langle\sigma Q\sigma^* v_i, v_i\rangle\,d\theta. \tag{2.13}$$

Since $B \in L_\infty(\mathscr{T}^*, \mathscr{L}(F, V'))$, $u \in L_2^\omega(I, F)$ and by assumption

$$E\int_I \operatorname{Tr}(\sigma Q\sigma^*)\,d\theta = \sum_{i=1}^{\infty} E\int_I \langle\sigma Q\sigma^* v_i, v_i\rangle\,d\theta < \infty,$$

by Bessel's inequality

$$E|\xi_0^n|_H^2 \leqq E|\xi_0|_H^2,$$

it follows from (2.13) that $\{\xi^n\}$ is contained in a bounded subset of $L_2^\omega(I, V) \cap L_\infty^\omega(I, H)$. V being a Hilbert space and (Ω, β, μ) a complete probability space, $L_2^\omega(I, V)$ is a reflexive Banach space, in fact a Hilbert space. Therefore there exists a subsequence of the sequence $\{\xi^n\}$ again denoted by $\{\xi^n\}$ (for convenience of notation) and an element η so that $\xi^n \to \eta$ weakly in $L_2^\omega(I, V)$. A reflexive Banach space being weakly complete, the weak limit η belongs to $L_2^\omega(I, V)$. It is also clear that $\xi^n \to \eta$ in $L_\infty^\omega(I, H)$ in its ω^*-topology. We show that η solves the problem S. Let $\phi \in C_{1,T}^\omega(I, R)$; multiply on both sides of the first equation of S_n by $\phi(t)$ and integrate by parts to obtain

$$-\int_0^T \langle\xi^n(t), \dot{\phi}^j(t)\rangle\,dt$$

$$= (\xi^n(0), \phi^j(0))+\int_0^T \{\langle A^*\phi^j, \xi^n\rangle+\langle Bu, \phi^j\rangle\}\,dt+\int_0^T \langle\phi^j, \sigma\,dW\rangle \quad \mu\text{-a.e.}, \tag{2.14}$$

where $\dot{\phi}^j(t) = ((d/dt)\phi(t))v_j$. Since $\phi \in C_{1,T}^\omega(I, R)$ $\phi^j, \dot{\phi}^j \in L_2^\omega(I, V) \subset L_2^\omega(I, H)$. Let G denote the completion of the σ-algebra generated by the class of sets $\{J \times D : J \in B_I, D \in \beta\}$ where B_I is the σ-algebra of Lebesgue measurable subsets of the interval I, and β is

the σ-algebra associated with the probability space (Ω, β, μ). Note that the σ-algebra (of progressively measurable subsets of the set $I \times \Omega$) $\mathcal{F}^* \subset G$, and $L_2^\omega(I, V)$ $(\equiv L_2(\mathcal{F}^*, V))$ is a closed subspace of the space $L_2(G, V)$. Thus any sequence $\{f_n\} \subset L_2^\omega(I, V)$ that converges weakly in $L_2^\omega(I, V)$ also converges weakly in the larger space $L_2(G, V)$, and the two limits are one and the same element. Since $\xi^n \to \eta$ weakly in $L_2^\omega(I, V)$ it is clear that $\xi^n \to \eta$ weakly also in $L_2(G, V)$. If we recall that $\dot{\phi}^j + A^*\phi^j \in L_2^\omega(I, V')$, it follows from the above facts that, for any bounded random variable $z (z \in L_\infty(\beta, R), \beta = \beta_T)$,

$$\begin{aligned} &\int_{I\times\Omega} z \cdot \langle \xi^n, \dot{\phi}^j + A^*\phi^j \rangle \, dt \, d\mu \\ &\qquad \to \int_{I\times\Omega} z \cdot \langle \eta, \dot{\phi}^j + A^*\phi^j \rangle \, dt \, d\mu. \end{aligned} \tag{2.15}$$

Similarly, by virtue of the facts that $\xi^n(0) \in L_2(\beta_0, H) \subset L_2(\beta, H)$ and $\xi^n(0) \to \xi_0$ strongly in $L_2(\beta_0, H)$, it follows that for any $z \in L_\infty(\beta, R)$,

$$\int_\Omega z \cdot (\xi^n(0), \phi^j(0)) \, d\mu \to \int_\Omega z \cdot (\xi_0, \phi^j(0)) \, d\mu. \tag{2.16}$$

Since the integrals $\int_0^T \langle \phi^j, Bu \rangle \, dt$ and $\int_0^T \langle \phi^j, \sigma \, dW \rangle$ belong to $L_1(\beta, R)$, multiplying (2.14) by $z \in L_\infty(\beta, R)$ and letting $n \to \infty$ gives from (2.15) and (2.16) that

$$\begin{aligned} &-E\left\{ z \cdot \int_0^T \langle \eta, \dot{\phi}^j \rangle \, dt \right\} \\ &\quad = E\{z \cdot (\xi_0, \phi^j(0))\} + E\left\{ z \cdot \int_0^T \langle \eta, A^*\phi^j \rangle \, dt \right\} \\ &\qquad + E\left\{ z \cdot \left(\int_0^T \langle \phi^j, Bu \rangle \, dt + \int_0^T \langle \phi^j, \sigma \, dW \rangle \right) \right\}. \end{aligned} \tag{2.17}$$

For any $D \in \beta$ let us define $E^D\{\cdot\} = \int_D \{\cdot\} \, d\mu$. Since (2.17) is valid for arbitrary $z \in L_\infty(\beta, R)$, we can choose for z the indicator function χ_D of an arbitrary set $D \in \beta$. This reduces (2.17) to the equivalent form

$$\begin{aligned} -E^D \int_0^T \langle \eta, \dot{\phi}^j \rangle \, dt &= E^D(\xi_0, \phi^j(0)) + E^D \int_0^T \{\langle A^*\phi^j, \eta \rangle + \langle Bu, \phi^j \rangle\} \, dt \\ &\quad + E^D \int_0^T \langle \phi^j, \sigma \, dW \rangle. \end{aligned} \tag{2.18}$$

Since (2.18) is valid for arbitrary $D \in \beta = \beta_T$, we have

$$\begin{aligned} -\int_0^T \langle \eta, \dot{\phi}^j \rangle \, dt &= (\xi_0, \phi^j(0)) + \int_0^T \{\langle A^*\phi^j, \eta \rangle + \langle Bu, \phi^j \rangle\} \, dt \\ &\quad + \int_0^T \langle \phi^j, \sigma \, dW \rangle \quad \mu - \text{a.e.} \end{aligned} \tag{2.19}$$

for all $\phi \in C_{1,T}^\omega(I, R)$. Let $C_\infty^\omega(I, R)$ denote the class of $\mathcal{F}^*$-measurable real-valued C^∞-functions with compact support in I. For $\phi \in C_\infty^\omega(I, R)$, the equation (2.19) reduces

to

$$(2.19)' \quad -\int_0^T \langle \eta, \dot{\phi}^j \rangle \, dt = \int_0^T \{\langle A^*\phi^j, \eta\rangle + \langle Bu, \phi^j\rangle\} \, dt + \int_0^T \langle \phi^j, \sigma \, dW\rangle \quad \mu\text{-a.e.}$$

This is true for all $\{v_j\}$ and all $\phi \in C_\infty^\omega(I, R)$. Hence η has the Ito differential given by $d\eta = (A\eta + Bu)\, dt + \sigma\, dw$. This statement will be made more precise in Theorem 4.1 while we solve a backward stochastic evolution equation. Evaluating the Ito-differential of the scalar $\langle \eta, \phi^j \rangle$ for $\phi^j = \phi v_j$ with arbitrary $\phi \in C_{1,T}^\omega(I, R)$, we obtain

$$\begin{aligned} -\int_0^T \langle \eta, \dot{\phi}^j \rangle \, dt = (\eta(0), \phi^j(0)) + \int_0^T \{\langle A^*\phi^j, \eta\rangle + \langle Bu, \phi^j\rangle\} \, dt \\ + \int_0^T \langle \phi^j, \sigma \, dW\rangle \quad \mu\text{-a.e.} \end{aligned} \tag{2.20}$$

Since v_j can be chosen from any dense subset of V and V is assumed to be dense in H, we conclude upon comparing (2.19) with (2.20) that $\eta(0) = \xi_0$ μ-a.e. and $\eta(0) \in L_2(\beta_0, H)$. Therefore, for all $\psi \in C_{1,T}^\omega(I, V)$,

$$\begin{aligned} -\int_0^T \langle \eta, \dot{\psi} \rangle \, dt = (\eta(0), \psi(0)) + \int_0^T \{\langle A^*\psi, \eta\rangle + \langle Bu, \psi\rangle\} \, dt \\ + \int_0^T \langle \psi, \sigma \, dW\rangle \quad \mu\text{-a.e.} \end{aligned} \tag{2.21}$$

Hence η is a weak solution of the evolution equation S in the sense of definition (2.2). For uniqueness we note that the difference of any two solutions, $\tilde{\eta} = \eta_1 - \eta_2$, satisfies the equality

$$\int_0^T \langle \tilde{\eta}, \dot{\psi} + A\psi \rangle \, dt = 0 \quad \text{for all } \psi \in C_{1,T}^\omega(I, V). \tag{2.22}$$

Thus $\tilde{\eta}$ must necessarily be zero, proving uniqueness. We may then replace the statement $\xi^n \to \eta$ weakly by $\xi^n \to \xi$ weakly in $L_2^\omega(I, V)$. This completes the proof of existence and uniqueness. That the mapping $u \to \xi$ from $L_2^\omega(I, F)$ into $L_2^\omega(I, V)$ is affine is obvious. For continuity we note that the weak limit of any sequence in a normed space is bounded in norm by the same bound as the sequence itself. Therefore it follows from (2.13) and the inequalities following this expression that

$$\begin{aligned} E|\xi(t)|_H^2 + \alpha \int_0^t E|\xi(\theta)|_V^2 \, d\theta \\ \leqq E|\xi_0|_H^2 + \frac{1}{\alpha}\int_0^t E|Bu|_{V'}^2 \, d\theta + \int_0^t E \operatorname{Tr}(\sigma Q \sigma^*) \, d\theta \end{aligned} \tag{2.23}$$

for $t \in I$. From this inequality we can prove the continuity of the map $(\xi_0, u, \sigma) \to \xi$. Indeed, it is clear from the inequality (2.23) that if (s = strongly)

$$\xi_{0,n} \xrightarrow{\text{s}} 0 \quad \text{in } L_2(\beta_0, H),$$

$$u_n \xrightarrow{\text{s}} 0 \quad \text{in } L_2^\omega(I, F),$$

$$\sigma_n \xrightarrow{\mathscr{T}_M} 0 \quad \text{in } M(\mathscr{F}^*, \mathscr{L}(H, V')),$$

then

$$\xi(\xi_{0n}, u_n, \sigma_n) \equiv \xi^n \xrightarrow{s} 0 \quad \text{in } L_2^\omega(I, V) \cap L_\infty^\omega(I, H).$$

This completes the proof of the theorem.

For stochastic control problems, in which controls are usually exercised on the basis of past information, it is natural to consider, instead of S, the family of Cauchy problems

$$\begin{aligned} &d\xi = (A\xi + Bu)\,dt + \sigma\,dW, \qquad t \in (s, T), \\ &\xi(s) = h, \end{aligned} \tag{2.24}$$

for $s \in (0, T)$. As an immediate consequence of Theorem 2.1 we have the following result.

COROLLARY 2.1. *Suppose the hypotheses of Theorem* 2.1 *are satisfied. Then for each* $h \in L_2(\beta_s, H)$, $s \in (0, T)$, *and* $u \in \mathcal{U}^s \subset L_2(s, T; F)$ *the problem* (2.24) *has a unique (weak) solution* $\xi_{s,h} \in L_2^\omega(s, T; V) \cap L_\infty^\omega(s, T; H)$ *such that* $u \to \xi_{s,h}$ *is an affine continuous map from* $L_2^\omega(s, T; F)$ *into* $L_2^\omega(s, T; V)$ *and* $h \to \xi_{s,h}$ *is a continuous map from* $L_2(\beta_s, H)$ *into* $L_2^\omega(s, T; V)$.

3. Existence of optimal controls and necessary conditions of optimality. For convenience of notation we use $\xi(u)$ to denote the response of the system S to the control action $u \in \mathcal{U}_a \subset L_2^\omega(I, F)$. Suppose $\mathcal{H}$ is another separable Hilbert space, and define the Hilbert space $L_2^\omega(I, \mathcal{H})$ as before, with the scalar product

$$(f, g)_{L_2^\omega(I,\mathcal{H})} = E\int_0^T (f(t), g(t))_{\mathcal{H}}\,dt. \tag{3.1}$$

Let $C \in \mathcal{L}(L_2^\omega(I, V), L_2^\omega(I, \mathcal{H}))$ denote the output or observation operator, and let $Z \equiv C\xi$ be the output. Let $N \in \mathcal{L}(L_2^\omega(I, F), L_2^\omega(I, F))$ and $Z_d \in L_2^\omega(I, \mathcal{H})$ be given. The problem of optimal control is to find a control u from the class $\mathcal{U}_a$ so that the cost function $J(u)$, given by

$$J(u) \equiv \|C\xi(u) - Z_d\|^2_{L_2^\omega(I,\mathcal{H})} + (Nu, u)_{L_2^\omega(I,F)}, \tag{3.2}$$

is minimum. The solution of this problem is given by the following theorem.

THEOREM 3.1. *Suppose the hypotheses of Theorem* 2.1 *hold. Let* $\mathcal{U}_a$ *be a closed convex subset of* $\mathcal{U} \equiv L_2^\omega(I, F)$ *containing the zero element,* $C \in \mathcal{L}(L_2^\omega(I, V), L_2^\omega(I, \mathcal{H}))$, $Z_d \in L_2^\omega(I, \mathcal{H})$ *and* $N \in \mathcal{L}(L_2^\omega(I, F), L_2^\omega(I, F))$ *a self-adjoint positive operator with the property* $(Nu, u)_{L_2^\omega(I,F)} \geq \nu\|u\|^2_{L_2^\omega(I,F)}$, $\nu > 0$. *Then the optimal control problem* (3.2) *subject to the dynamic constraint S has a unique solution.*

Proof. The proof is classical.

For the necessary conditions of optimality it is convenient to rewrite the cost function J in the form

$$J(u) = E\int_I |C(t)\xi(u)(t) - Z_d(t)|^2_{\mathcal{H}}\,dt + E\int_I (N(t)u(t), u(t))_F\,dt \tag{3.3}$$

while considering $C \in L_\infty(\mathcal{F}^*, \mathcal{L}(V, \mathcal{H}))$ and $N \in L_\infty(\mathcal{F}^*, \mathcal{L}(F, F))$. The following lemmas are classical.

LEMMA 3.1. *The functional* J *has a Gateaux differential at each point* $u \in \mathcal{U}_a$ *given by*

$$J'_u(v - u) = 2E\int_I (C(t)\xi(u) - Z_d, C(t)[\xi(v) - \xi(u)])_{\mathcal{H}}\,dt + 2E\int_I (Nu, v - u)_F\,dt$$

in the direction $(v - u)$ *for each* $v \in L_2^\omega(I, F)$.

LEMMA 3.2. *Let J be a Gateaux differentiable functional on a Hilbert space $\mathcal{U}$, and $\mathcal{U}_a$ a closed convex subset of $\mathcal{U}$. Then in order that $u \in \mathcal{U}_a$ be a minimizing element it is necessary that*

$$J'_u(v-u) \geqq 0 \quad \text{for all } v \in \mathcal{U}_a. \tag{3.4}$$

Further, if $v \to J(v)$ is strictly convex, then (3.4) *is also the sufficient condition.*

4. Backward stochastic evolution equation and the necessary conditions of optimality. For the solution of the control problem in terms of the adjoint state or in the form of the state feedback, it is required to solve certain stochastic evolution equations backward in time. Using semimartingale theory with Meyer's formula for change of variables, Bismut [2] has given a meaning to the solution of the backward equation in the finite dimensional case. Since this formula is not applicable to the infinite dimensional case we use a different technique known as the "principle of transposition". Essentially, in this approach one constructs a suitable isomorphism which is then transposed to solve the original problem. We introduce here an isomorphism suitable for the purpose.

Consider the evolution equation

$$\begin{aligned} &d\phi = (A-\Gamma)\phi\, dt + g\, dt, \qquad t \in I = (0, T), \\ &\phi(0) = 0, \end{aligned} \tag{4.1}$$

and suppose A satisfies (Ai), (Aii); $\Gamma \in L_\infty(\mathcal{F}^*, \mathcal{L}(H, H))$ with $\Gamma \geqq 0\ dt \times d\mu$ a.e. and $g \in L_2^\omega(I, V')$. Then it follows from Theorem 2.1 ($\sigma = 0$) that for every $g \in L_2^\omega(I, V')$, (4.1) has a unique solution $\phi = \phi(g) \in L_2^\omega(I, V)$, and that in this particular case $d\phi/dt$ exists in the sense of generalized random processes or equivalently in the sense of distribution and belongs to $L_2^\omega(I, V')$. Define

$$\tau \equiv \frac{d}{dt} - (A-\Gamma) \tag{4.2}$$

and the set

$$X_0 \equiv \{\phi : \phi \in L_2^\omega(I, V),\ \phi(0) = 0\ \mu\text{-a.e.},\ \tau\phi \in L_2^\omega(I, V')\}. \tag{4.3}$$

The set X_0 provided with the norm

$$\|\phi\|_{X_0} \equiv (\|\phi\|^2_{L_2^\omega(I,V)} + \|\tau\phi\|^2_{L_2^\omega(I,V')})^{1/2}$$

is a Hilbert space, and τ is an isomorphism of X_0 onto $L_2^\omega(I, V')$. By transposition of this isomorphism we can solve the backward evolution equation

$$\begin{aligned} &dy + \{(A^* - \Gamma)y + f\}\, dt + \tilde{\sigma}\, dW = 0, \qquad t \in (0, T), \\ &y(T) = \eta \end{aligned} \tag{4.4}$$

in the weak sense. More precisely, (4.4) is said to have a weak solution if there exists a $y \in L_2^\omega(I, V)$ so that

$$\int_0^T \langle y, \tau\phi \rangle\, d\theta = l(\phi) \quad \mu\text{-a.e.} \tag{4.5}$$

for all $\phi \in X_0$, where

$$l(\phi) \equiv (\eta, \phi(T)) + \int_0^T \langle f, \phi \rangle\, d\theta + \int_0^T (\tilde{\sigma}^*\phi, dW). \tag{4.6}$$

THEOREM 4.1. *Consider the system* (4.4). *Suppose the operator* A *satisfies* (Ai), (Aii); *the operator* $\Gamma \in L_\infty(\mathcal{F}^*, \mathcal{L}(H, H))$ *is self-adjoint and positive*; $f \in L_2^\omega(I, V')$; $\tilde{\sigma}$ *satisfies* (Aiv); W *is the Wiener process satisfying* (Aiii) *and* $\eta \in L_2(\beta, H)$. *Then* (4.4) *has a unique weak solution* $y \in L_2^\omega(I, V)$ *and, for each* $t \in I$, *the solution* y *satisfies the equality*

$$\langle y(t), \phi(t)\rangle + E^{\beta_t}\int_t^T \langle y, \tau\phi\rangle\, d\theta = E^{\beta_t}\langle \eta, \phi(T)\rangle + E^{\beta_t}\int_t^T \langle f, \phi\rangle\, d\theta \tag{4.7}$$

(β_t, μ)-*a.e. for all* $\phi \in X_0$. *Further,* $y \in C_0^\omega(I, V')$.

Proof. First we show that there exists a unique $y \in L_2^\omega(I, V)$ so that

$$\int_0^T \langle y, \tau\phi\rangle\, d\theta = l(\phi) \quad \mu\text{-a.e.}$$

for all $\phi \in X_0$, and then we prove that this y solves (4.4) in the sense of distribution. Finally we prove (4.7) using these results.

Let $C_0^\omega(I, H)$ denote the vector space of H-valued progressively measurable continuous random processes on I with the topology given by the norm

$$\|x\|_{C_0^\omega(I,H)} = \left(E\left(\sup_{t\in I}|x(t)|_H^2\right)\right)^{1/2}.$$

Equipped with this topology $C_0^\omega(I, H)$ is a Banach space. For each $\phi \in X_0$, $(A-\Gamma)\phi \in L_2^\omega(I, V')$ and consequently $\dot{\phi} \in L_2^\omega(I, V')$. Further, $V \subset H \subset V'$ both algebraically and topologically and V is dense in H. Therefore with a modification over a set of dt-measure zero, $\phi \in C_0^\omega(I, H)$ whenever $\phi \in X_0$. Loosely speaking, $X_0 \subset C_0^\omega(I, H)$. The proof of this is similar to that of Carroll [11, Lemma 3.7, p. 176]. Therefore, whenever $\phi \in X_0$, $\phi(T)$ is a well-defined β_T-measurable H-valued random variable and $\phi(T) \in L_2(\beta_T, H)$. Consequently, for a given $\eta \in L_2(\beta, H)$, $\beta = \beta_T$, $(\eta, \phi(T)) \in L_1(\beta, R)$. Since, by hypothesis, $f \in L_2^\omega(I, V')$, $\tilde{\sigma} \in L_\infty(\mathcal{F}^*, \mathcal{L}(H, V'))$ and $Q \in L_\infty(I, \mathcal{L}(H, H))$, it is easily verified that

$$\int_0^T \langle f, \phi\rangle\, dt \in L_1(\beta, R),$$

$$\int_0^T (\tilde{\sigma}^*\phi, dW) \in L_1(\beta, R).$$

Thus, for each $\phi \in X_0$, $l(\phi) \in L_1(\beta, R)$ and consequently, for any $z \in L_\infty(\beta, R)$,

$$F_z(l(\phi)) \equiv E\{z \cdot l(\phi)\} \tag{4.9}$$

defines a bounded linear functional on X_0. In fact $(z, \phi) \to F_z(l(\phi))$ is a bilinear functional on $L_\infty(\beta, R) \times X_0$. Thus for a fixed but arbitrary $z \in L_\infty(\beta, R)$, $\phi \to F_z(l(\phi))$ is a continuous linear functional on X_0.

Further, since τ is an isomorphism of X_0 onto $L_2^\omega(I, V')$, τ^{-1} is a continuous linear operator from $L_2^\omega(I, V')$ into X_0. Thus, for each $z \in L_\infty(\beta, R)$, $\phi \to F_z(l\tau^{-1}(\phi))$ is a continuous linear functional on $L_2^\omega(I, V')$. In other words $(l\tau^{-1})$ is a continuous linear operator from $L_2^\omega(I, V')$ into $L_1(\beta, R)$. Thus there exists a unique $y \in (L_2^\omega(I, V'))'$ such

that

$$F_z(l\tau^{-1}(\psi)) = F_z\Big(\int_0^T \langle y, \psi\rangle\, dt\Big) = E\Big\{z \cdot \int_0^T \langle y, \psi\rangle\, dt\Big\} \tag{4.10}$$

for all $\psi \in L_2^{\omega}(I, V')$ and all $z \in L_\infty(\beta, R)$. τ being an isomorphism of X_0 onto $L_2^{\omega}(I, V')$, this is equivalent to

$$F_z(l(\phi)) = E\Big\{z \cdot \int_0^T \langle y, \tau\phi\rangle\, dt\Big\}, \tag{4.11}$$

for all $\phi \in X_0$ and $z \in L_\infty(\beta, R)$. Since V is reflexive, $y \in L_2^{\omega}(I, V)$. Since z is any arbitrary element of $L_\infty(\beta, R)$, taking $z = \chi_D$ for arbitrary $D \in \beta$ we get from (4.9) and (4.11) that

$$E^D \int_0^T \langle y, \tau\phi\rangle\, dt = E^D(l(\phi)) \equiv l_D(\phi) \tag{4.12}$$

for arbitrary $D \in \beta$ $(= \beta_T)$ and all $\phi \in X_0$. Since $D \in \beta$ is arbitrary, it follows from (4.12) that there exists a unique $y \in L_2^{\omega}(I, V)$ such that

$$\int_0^T \langle y, \tau\phi\rangle\, dt = l(\phi) \quad \mu\text{-a.e.},$$

for all $\phi \in X_0$.

For convenience of notation we write $E^{\beta_t}\{x\}$ for $E\{x|\beta_t\}$, where $E\{x|\beta_t\}$ denotes the conditional expectation of the random variable x given the σ-algebra β_t. Returning to the proof we note that the equality (4.12) remains valid for arbitrary $D \in \beta_t \subset \beta_T$ for each $t \in (0, T)$. Thus, for arbitrary $t \in (0, T)$ and arbitrary $D \in \beta_t$, we have

$$\begin{aligned} E^D\Big(\int_0^T \langle y, \tau\phi\rangle\, dt\Big) &= E^D\Big(E^{\beta_t}\Big\{\int_0^T \langle y, \tau\phi\rangle\, dt\Big\}\Big) \\ &= E^D\Big(E^{\beta_t}\Big\{\int_0^t \langle y, \tau\phi\rangle\, dt\Big\} + E^{\beta_t}\Big\{\int_t^T \langle y, \tau\phi\rangle\, dt\Big\}\Big) \\ &= E^D\Big(\int_0^t \langle y, \tau\phi\rangle\, dt + E^{\beta_t}\Big\{\int_t^T \langle y, \tau\phi\rangle\, dt\Big\}\Big). \end{aligned} \tag{4.13}$$

The last equality follows from the fact that y and $\tau\phi$ are progressively measurable and consequently the integral $\int_0^t \langle y, \tau\phi\rangle\, dt$ is β_t-measurable. Similarly, for $D \in \beta_t$,

$$\begin{aligned} l_D(\phi) &= E^D[l(\phi)] \\ &= E^D[E^{\beta_t}\{l(\phi)\}] \\ &= E^D\Big[E^{\beta_t}\Big\{(\eta, \phi(T)) + \int_t^T \langle f, \phi\rangle\, dt + \int_t^T (\tilde{\sigma}^*\phi, dW)\Big\} \\ &\qquad + \int_0^t \langle f, \phi\rangle\, dt + \int_0^t (\tilde{\sigma}^*\phi, dW)\Big]. \end{aligned} \tag{4.14}$$

Since for each $\theta \in (0, T)$ $(\tilde{\sigma}^*\phi)(\theta)$ is β_θ-measurable, $E^{\beta_t}\{\int_t^T (\tilde{\sigma}^*\phi, dW) = 0$. Thus for

$D \in \beta_t$

$$l_D(\phi) = E^D\Big[E^{\beta_t}\Big\{(\eta, \phi(T)) + \int_t^T \langle f, \phi\rangle\, dt\Big\} + \int_0^t \langle f, \phi\rangle\, dt + \int_0^t (\tilde{\sigma}^*\phi, dW)\Big]. \tag{4.15}$$

Therefore, for each $t \in (0, T)$ and arbitrary $D \in \beta_t$, it follows from (4.12), (4.13) and (4.15) that

$$E^D\Big[\int_0^t \langle y, \tau\phi\rangle\, dt - \int_0^t \langle f, \phi\rangle\, dt - \int_0^t (\tilde{\sigma}^*\phi, dW)\Big] = E^D\Big[E^{\beta_t}\Big\{(\eta, \phi(T)) + \int_t^T \langle f, \phi\rangle\, dt - \int_t^T \langle y, \tau\phi\rangle\, dt\Big\}\Big], \tag{4.16}$$

for all $\phi \in X_0$.

Since D is an arbitrary element of β_t, (4.16) implies that for each $t \in (0, T)$

$$\int_0^t \langle y, \tau\phi\rangle\, d\theta - \int_0^t \langle f, \phi\rangle\, d\theta - \int_0^t (\tilde{\sigma}^*\phi, dW) = E^{\beta_t}\Big\{(\eta, \phi(T)) + \int_t^T \langle f, \phi\rangle\, d\theta - \int_t^T \langle y, \tau\phi\rangle\, d\theta\Big\} \tag{4.17}$$

(β_t, μ)-a.e. for all $\phi \in X_0$. Thus for arbitrary $\phi \in \mathscr{D}(0, t; V) = (C_0^\infty(0, t; V)) \subset \mathscr{D}(0, T; V) \subset X_0$ we have, due to (4.17),

$$\int_0^t \langle y, \tau\phi\rangle\, d\theta = \int_0^t \langle f, \phi\rangle\, d\theta + \int_0^t (\tilde{\sigma}^*\phi, dW) \tag{4.17'}$$

(β_t, μ)-a.e. for all $\phi \in \mathscr{D}(0, t; V)$. This equality implies that the stochastic differential of y exists in the sense of generalized random process or stochastic vector valued measures on Borel subsets of $(0, t)$. We write this symbolically as

$$dy + ((A^* - \Gamma)y + f)\, dt + \tilde{\sigma}\, dW = 0. \tag{4.18}$$

For the proof of this we introduce the linear form

$$L(\phi) \equiv -\int_0^t \langle y, \tau\phi\rangle\, d\theta + \int_0^t \langle f, \phi\rangle\, d\theta + \int_0^t (\tilde{\sigma}^*\phi, dW),$$

and taking $\phi = \psi \cdot v$ with v an arbitrary element of V and $\psi \in \mathscr{D}(0, t)$ we define

$$\begin{aligned} L_v(\psi) &\equiv L(\psi \cdot v) \\ &= -\int_0^t \dot{\psi}(\theta)\langle y(\theta), v\rangle\, d\theta + \int_0^t \psi(\theta)\langle[(A^* - \Gamma)y + f]\, d\theta + \tilde{\sigma}\, dW, v\rangle \\ &= \int_0^t \psi(\theta)\langle dy + [(A^* - \Gamma)y + f]\, d\theta + \tilde{\sigma}\, dW, v\rangle \equiv \int_0^t \psi \nu_v(d\theta). \end{aligned} \tag{4.19}$$

It is clear that $\psi \to L_v(\psi)$ is a well-defined β_t-measurable continuous linear form on $\mathscr{D}(0, t)$, since this is so for the map $\phi \to L(\phi)$ on $\mathscr{D}(0, t; V)$. Even more is true; $\psi \to L_v(\psi)$ is a β_t-measurable continuous linear form on $B_0'(0, t)$ where $B_0'(0, t)$ denotes the vector space of once continuously differentiable functions on the open interval $(0, t)$ whose

members along with the first derivatives vanish on the boundary $\{0, t\}$. $B_0'(0, t)$ is given the locally convex topology generated by the family of seminorms $q_s(\phi) \equiv \max_{\theta\in(0,t)} |\partial^s\phi(\theta)|$, $s = 0, 1$; $\phi \in B_0'(0, t)$. Every continuous linear form η on $B_0'(0, t)$ has the representation (Horvath [10, Prop. 3, p. 346])

$$\eta(\psi) = \nu^0(\psi) - \nu^1(\dot{\psi}), \qquad \psi \in B_0'(0, t), \qquad \text{“}\cdot\text{”} = \frac{d}{dt}, \tag{4.19}'$$

where $\nu^\alpha(\xi) \equiv \int_0^t \xi(\theta)\nu^\alpha(d\theta)$, $\alpha = 0, 1$, are integrable measures. We note that (4.19) has precisely the same form as given by (4.19)′, where we can identify the measures ν^0 and ν^1 as, respectively,

$$\nu^0(d\theta) = \langle((A^* - \Gamma)y + f, v\rangle\, d\theta + \langle\tilde{\sigma}\, dW(\theta), v\rangle,$$

and

$$\nu^1(d\theta) = \langle y(\theta), v\rangle\, d\theta.$$

Since the injection $\mathcal{D}(0, t) \hookrightarrow B_0'(0, t)$ is continuous and $\mathcal{D}$ is dense in B_0', the dual of B_0' is a subspace of the space of distributions $\mathcal{D}'$ and consequently the measures ν is differentiable in the sense of distribution and we can write $\nu_v = \nu^0 + \partial\nu^1$, or equivalently $\nu_v(d\theta) = \langle((A^* - \Gamma)y + f)(\theta), v\rangle\, d\theta + \langle\tilde{\sigma}(\theta)\, dW(\theta), v\rangle + \langle dy, v\rangle$. Due to (4.17)′ and the definition of L_v we have $L_v(\psi) = \int_0^t \psi\nu_v(d\theta) = 0$ for all $\psi \in \mathcal{D}(0, t)$. Therefore $\nu_v(d\theta) = 0$ in the distribution sense and consequently $\langle dy(\theta), v\rangle + \langle((A^* - \Gamma)y + f)(\theta), v\rangle\, d\theta + \langle\tilde{\sigma}(\theta)\, dW(\theta), v\rangle = 0$. Since the measure $\nu^0(d\theta)$ is integrable and $\langle dy(\theta), v\rangle = -\nu^0(d\theta)$, we conclude that the measure $\partial\nu'(d\theta) \equiv \langle dy(\theta), v\rangle$ is also integrable. Thus, for s, $s + h \in (0, t)$ and $v \in V$,

$$\langle y(s+h) - y(s), v\rangle + \int_s^{s+h} \langle(A^* - \Gamma)y + f, v\rangle\, d\theta + \int_s^{s+h} (\tilde{\sigma}^*(\theta)v, dW(\theta)) = 0.$$

Since $t \in (0, T)$ is arbitrary and $\eta(t) \equiv \int_0^t \tilde{\sigma}\, dW$, $t \in I$, is a V'-valued β_t martingale this shows that $y \in C_0^\omega(0, T; V')$ and therefore $y(t)$ is defined for each $t \in (0, T)$. Further, since $v \in V$ is arbitrary we have $dy + ((A^* - \Gamma)y + f)\, d\theta + \tilde{\sigma}\, dW(\theta) = 0$. For $\phi \in X_0$ we have $\dot{\phi} \in L_2^\omega(0, T; V')$ and consequently for each $t \in (0, T)$

$$\int_0^t \langle\phi, dy\rangle = \langle\phi(t), y(t)\rangle - \int_0^t \langle\dot{\phi}, y\rangle\, d\theta. \tag{4.20}$$

Due to (4.18) we have

$$\int_0^t \langle\phi, dy\rangle = -\int_0^t \langle\phi, (A^* - \Gamma)y\rangle\, d\theta - \int_0^t \langle\phi, f\rangle\, d\theta - \int_0^t (\tilde{\sigma}^*\phi, dW), \tag{4.21}$$

which is well defined for all $\phi \in X_0$. Thus for arbitrary $\phi \in X_0$ it follows from (4.20) and (4.21) that for each $t \in (0, T)$

$$\langle\phi(t), y(t)\rangle = \int_0^t \langle y, \tau\phi\rangle\, d\theta - \int_0^t \langle f, \phi\rangle\, d\theta - \int_0^t (\tilde{\sigma}^*\phi, dW) \quad \mu\text{-a.e.} \tag{4.22}$$

Combining (4.17) and (4.22) we obtain (4.7).

Remark 4.1. Given $\tilde{\sigma} = 0$, any function $y \in L_2^\omega(I, V)$ that satisfies (4.7) is a solution of the problem (4.4) in the weak sense. In this case $y \in C_0^\omega(I, H)$ also.

Remark 4.2. For $\tilde{\sigma} \not\equiv 0$ we observe from the proof of the previous theorem that $y \in C_0^\omega(I, V')$. We are, however, unable to prove whether or not y belongs to $C_0^\omega(I, H)$.

With the help of the above results we can solve the control problem. Let $\Lambda_{\mathcal{H}}$ (Λ_F) denote the canonical isomorphisms of $\mathcal{H}$ onto $\mathcal{H}'$ (F onto F').

THEOREM 4.2. *Consider the system* S *and the cost functional* $J(u)$ *of* (2.1), *and suppose the hypotheses of Theorem* 2.1 *and Theorem* 3.1 *hold. Then in order that* $u \in \mathcal{U}_a$ *be an optimal control it is necessary and sufficient that there exist a* $p(u) \in L_2^{\omega}(I, V)$ *so that*:

$$\text{(i)} \qquad E\int_0^T (\Lambda_F^{-1}B^*p(u)+Nu, v-u)_F\,dt \geqq 0 \quad \textit{for all } v \in \mathcal{U}_a,$$

$$\text{(ii)} \quad \text{(AS)} \qquad -dp(u) = \{A^*p(u)+C^*\Lambda_{\mathcal{H}}(C\xi(u)-Z_d)\}\,dt,$$
$$p(u)(T) = 0 \quad \mu\text{-a.e.},$$

$$\text{(iii)} \quad \text{(S)} \qquad d\xi(u) = (A\xi(u)+Bu)\,dt+\sigma\,dW,$$
$$\xi(u)(0) = \xi_0.$$

Proof. Since $L_2^{\omega}(I, F)$ is a Hilbert space, $\mathcal{U}_a$ is a closed convex subset of $L_2^{\omega}(I, F)$ and J is Gateaux differentiable (Lemma 3.1) and strictly convex, being quadratic, Lemma 3.2 applies. Therefore it follows from this lemma that a necessary and sufficient condition for $u \in \mathcal{U}_a$ to be optimal is that

$$E\int_0^T (C\xi(u)-Z_d, C(\xi(v)-\xi(u)))_{\mathcal{H}}\,dt + E\int_0^T (Nu, v-u)_F\,dt \geqq 0 \tag{4.23}$$

for all $v \in \mathcal{U}_a$. Using the canonical map $\Lambda_{\mathcal{H}}$ in (4.23) and noting the fact that $C^* \in \mathcal{L}(L_2^{\omega}(I, \mathcal{H}'), L_2^{\omega}(I, V'))$ we have

$$E\int_0^T \langle C^*\Lambda_{\mathcal{H}}(C\xi(u)-Z_d), \xi(v)-\xi(u)\rangle\,dt + E\int_0^T (Nu, v-u)_F\,dt \geqq 0 \tag{4.24}$$

for all $u \in \mathcal{U}_a$. In (4.24) $\langle\cdot\,,\cdot\rangle$ denotes $V'-V$ duality pairing. Since $C^*\Lambda_{\mathcal{H}}(C\xi(u)-Z_d) \in L_2^{\omega}(I, V')$ and $p(u)(T) = 0 \in L_2(\beta, H)$, it follows from Theorem 4.1 with $\Gamma \equiv 0$, $\tilde{\sigma} \equiv 0$ and $f = C^*\Lambda_{\mathcal{H}}(C\xi(u)-Z_d)$ that the system (AS) has a unique weak solution $p(u) \in L_2^{\omega}(I, V)$ corresponding to the control $u \in \mathcal{U}_a$. Therefore, according to (4.5) of Theorem 4.1,

$$\int_0^T \langle p(u), d\phi\rangle - \int_0^T \langle p(u), A\phi\rangle\,dt = \int_0^T \langle C^*\Lambda_{\mathcal{H}}(C\xi(u)-Z_d), \phi\rangle\,dt \tag{4.25}$$

μ-a.e. for all $\phi \in X_0$. Let $\xi(v)$ denote the solution of the problem S corresponding to the control $v \in \mathcal{U}_a$. Since $(\xi(v)-\xi(u))(0) = 0$ and $\xi(v)-\xi(u)$ solves the problem $d\eta = (A\eta+B(v-u))\,dt$, $\eta(0) = 0$ and $Bv-Bu \in L_2^{\omega}(I, V')$, it is clear that $\xi(v)-\xi(u) \in X_0$.

Thus, taking $\xi(v)-\xi(u)$ for ϕ in (4.25), we have

$$\int_0^T \langle p(u), B(v-u)\rangle\,dt = \int_0^T \langle C^*\Lambda_{\mathcal{H}}(C\xi(u)-Z_d), \xi(v)-\xi(u)\rangle\,dt$$

μ-a.e., and consequently

$$E\int_0^T \langle p(u), B(v-u)\rangle\,dt = E\int_0^T \langle C^*\Lambda_{\mathcal{H}}(C\xi(u)-Z_d), \xi(v)-\xi(u)\rangle\,dt. \tag{4.26}$$

Therefore it follows from (4.24) and (4.26) that

$$E\int_0^T \langle p(u), B(v-u)\rangle\,dt + E\int_0^T (Nu, v-u)_F\,dt \geqq 0 \tag{4.27}$$

for all $v \in \mathcal{U}_a$. Since $B \in \mathcal{L}(L_2^{\omega}(I, F), L_2^{\omega}(I, V'))$ we have $B^* \in \mathcal{L}(L_2^{\omega}(I, V), L_2^{\omega}(I, F'))$;

therefore, using the canonical map Λ_F in (4.27), we obtain

$$E\int_0^T (\Lambda_F^{-1}B^*p(u)+Nu, v-u)_F\,dt \geqq 0 \quad \text{for all } v \in \mathcal{U}_a. \tag{4.28}$$

This completes the proof of the theorem.

Remark. Note that the inequality (4.28) characterizing the optimal control does not change if we add in the adjoint equation (AS) an additional noise term $\tilde{\sigma}\,dW$, with $\tilde{\sigma} \in L_\infty(\mathcal{F}^*, \mathcal{L}(H, V'))$.

5. Feedback control and operator Riccati equation. By removing constraints on the control, that is, taking $\mathcal{U}_a = L_2^\omega(I, F)$, we get from (4.28) that the optimal control is of the form

$$u = -N^{-1}\Lambda_F^{-1}B^*p. \tag{5.1}$$

Substituting the expression for u from (5.1) into the equations (S) and (AS) we obtain a coupled system of random evolution equations:

$$\text{(CS)} \qquad \begin{aligned} &d\xi = (A\xi - Kp)\,dt + \sigma\,dW, && \xi(0) = \xi_0, \\ &-dp = (A^*p + L\xi)\,dt + g\,dt, && p(T) = 0, \end{aligned} \qquad t \in I = (0, T),$$

where $K \equiv BN^{-1}\Lambda_F^{-1}B^*$, $L \equiv C^*\Lambda_{\mathcal{H}}C$ and $g = -C^*\Lambda_{\mathcal{H}}Z_d$. The optimal control is obtained from the solution of (CS) and the expression (5.1). Clearly, due to our hypotheses on B, N, C and Z_d, K, $L \in \mathcal{L}(L_2^\omega(I, V), L_2^\omega(I, V'))$ or more precisely $L_\infty(\mathcal{F}^*, \mathcal{L}(V, V'))$, and $g \in L_2^\omega(I, V')$. Since we are interested in the feedback control and u is given in terms of the adjoint state p, our aim is to find an expression for p in terms of the state ξ. Loosely speaking, we wish to show that there exists a $P \in L_\infty(\mathcal{F}^*, \mathcal{L}(H, H))$ and a $\gamma \in L_2^\omega(I, V) \cap L_\infty^\omega(I, H)$ so that

$$p(t) = P(t)\xi(t) + \gamma(t) \quad \text{for } t \in I = (0, T). \tag{5.2}$$

We take an approach similar to that of Lions [5, Chapter 3, § 4, p. 132].

LEMMA 5.1. *For each $h \in L_2(\beta_s, H)$ the system*

$$\text{(OS)} \qquad \begin{aligned} &d\phi = (A\phi - K\psi)\,dt + \sigma\,dW, && \phi(s) = h, \\ &-d\psi = (A^*\psi + L\phi)\,dt + g\,dt, && \psi(T) = 0, \end{aligned} \qquad t \in (s, T)$$

has a unique solution ϕ, $\psi \in L_2^\omega(s, T; V) \cap L_\infty^\omega(s, T; H)$. Furthermore, $\psi \in C_0^\omega(s, T; H)$ also.

Proof. The optimality system (OS) arises from the following control problem:

$$\begin{aligned} &d\phi(v) = (A\phi(v) + Bv)\,dt + \sigma\,dW, \qquad t \in (s, T), \\ &\phi(v)(s) = h, \\ &J(s, h, v) = E_{s,h}\left\{\int_s^T |C\phi(v) - Z_d|_{\mathcal{H}}^2\,dt + \int_s^T (Nv, v)_F\,dt\right\} = \min, \end{aligned} \tag{5.3}$$

where $E_{s,h}\{\cdot\}$ denotes the conditional expectation given that $\phi(v)(s) = h$ and the min is taken over $\mathcal{U}^s = L_2^\omega(s, T; F)$. It follows from Corollary 2.1 and Theorem 3.1 that this problem has a unique solution $u \in \mathcal{U}^s$ with $\phi(u) \in L_2^\omega(s, T; V)$. By Theorem 4.2 the control is given by $u = -N^{-1}\Lambda_F^{-1}B^*\psi$, where ψ satisfies the equation

$$\begin{aligned} &-d\psi = (A^*\psi + L\phi(u))\,dt + g\,dt, \qquad t \in (s, T), \\ &\psi(T) = 0, \end{aligned} \tag{5.4}$$

in the sense of Theorem 4.1. Since $L\phi(u)+g \in L_2^\omega(s, T; V')$ (5.4) has a unique solution $\psi \in L_2^\omega(s, T; V)$ in the sense of Theorem 4.1. Further, from (5.4) it follows that $d\psi/dt$ exists in the distribution sense and belongs to $L_2^\omega(s, T; V')$ and consequently $\psi \in C_0^\omega(s, T; H)$. This completes the proof of the lemma.

LEMMA 5.2. *Consider the optimality system* (OS) *for* $s \in (0, T)$ *arbitrary. Then* $h \to \{\phi, \psi\}$ *is a continuous mapping from the strong topology of* $L_2(\beta_s, H)$ *into the weak topology of* $L_2(s, T; V)$.

Proof. Let $\{h_n\} \in L_2(\beta_s, H)$, and suppose $h_n \to h$ strongly in $L_2(\beta_s, H)$. Let v be an arbitrary element of $\mathcal{U}^s$ and $\phi_n(v)$ the solution of the problem

$$\begin{aligned} d\phi_n(v) &= (A\phi_n(v)+Bv)\,dt+\sigma\,dW, \\ \phi_n(v)(s) &= h_n. \end{aligned} \tag{5.5}$$

Since $\{h_n\}$ is bounded in $L_2(\beta_s, H)$, it follows that $\{\phi_n(v)\}$ is contained in a bounded subset of $L_2^\omega(s, T; V) \cap L_\infty^\omega(s, T; H)$. $L_2^\omega(s, T; V)$ being reflexive, there exists a subsequence again denoted by $\{\phi_n\}$ and an element $\phi \in L_2^\omega(s, T; V)$ so that $\phi_n(v) \to \phi$ weakly. Following the same procedure as in the proof of Theorem 2.1, we conclude that $\phi = \phi(v)$ is the solution of the problem (5.5) with h_n replaced by h and that $\phi(v) \in L_\infty^\omega(s, T; H)$ also. In fact for a fixed $v \in \mathcal{U}^s$ as $h_n \to h$ strongly in $L_2(\beta_s, H)$ it follows from the inequality

$$E|\phi_n(v)(t)-\phi(v)(t)|_H^2+2\alpha E\int_s^t |\phi_n(v)(\theta)-\phi(v)(\theta)|_V^2\,d\theta \leqq E|h_n-h|_H^2$$

that $\phi_n(v) \to \phi(v)$ strongly in $L_2^\omega(s, T; V) \cap L_\infty^\omega(s, T; H)$. Thus, C being a bounded linear operator from $L_2^\omega(I, V)$ to $L_2^\omega(I, \mathcal{H})$, for a fixed $v \in \mathcal{U}^s$,

$$\lim_n J(s, h_n, v) = J(s, h, v) \quad \beta_s\text{-a.s.} \tag{5.6}$$

whenever $h_n \to h$ strongly in $L_2(\beta_s, H)$. Let $\{u_n\} \in \mathcal{U}^s$ be the sequence of controls (see Theorem 3.1) so that

$$J(s, h_n, u_n) = \inf\{J(s, h_n, v),\ v \in \mathcal{U}^s\}. \tag{5.7}$$

Let u be the optimal control corresponding to problem (5.3). Then clearly

$$J(s, h_n, u_n) = \inf\{J(s, h_n, v),\ v \in \mathcal{U}^s\} \leqq J(s, h_n, u). \tag{5.8}$$

Since the sequence $\{\phi_n(u)\}$ (the solutions of (5.5) with $v = u$) is bounded in $L_2^\omega(s, T; V) \cap L_\infty^\omega(s, T; H)$, $\{J(s, h_n, u)\}$ is a bounded sequence of numbers. Thus by (5.8) $J(s, h_n, u_n) < \infty$ independently of n; and due to the fact that

$$J(s, h_n, u_n) \geqq E\int_0^T (Nu_n, u_n)_F\,dt \geqq \nu\int_s^T E|u_n|_F^2\,dt \quad \beta_s\text{-a.s.}, \tag{5.9}$$

we have a sequence of controls $\{u_n\}$ contained in a bounded subset of $\mathcal{U}^s$. Since $L_2^\omega(I, F)$ is a Hilbert space, there exists a subsequence of the sequence $\{u_n\}$ again denoted by $\{u_n\}$ and an element $\tilde{u} \in \mathcal{U}^s$ so that $u_n \to \tilde{u}$ weakly. Consequently the sequence of solutions $\{\phi_n(u_n)\}$ of the problem (5.5) with $v = u_n$ is contained in a bounded subset of $L_2^\omega(s, T; V) \cap L_\infty^\omega(s, T; H)$. Therefore there exists a subsequence of the sequence $\{\phi_n(u_n)\}$ again denoted by $\{\phi_n\}$ and an element $\phi \in L_2^\omega(s, T; V) \cap L_\infty^\omega(s, T; H)$ so that $\phi_n \to \phi$ weakly in $L_2^\omega(s, T; V)$. In fact, $\phi_n \to \phi$ in the weak star topology of $L_\infty^\omega(s, T; H)$ also. Again by the same technique as in the proof of Theorem 2.1 (see (2.14)–(2.19)′, with $\xi^n(s) = h_n$, $u = u_n$, and $\xi^n = \phi_n$), we obtain $\phi = \phi(\tilde{u})$ where $\phi(\tilde{u})$ is the solution of the problem (5.5) with v replaced by $\tilde{u}$ and h_n by h. Since J is quadratic in ϕ and u it is

weakly lower semicontinuous on $L_2^w(I, V) \times L_2(I, F)$, and consequently as $u_n \to \tilde{u}$ weakly in $L_2^w(s, T; F)$ and $\phi_n \to \phi(\tilde{u})$ weakly in $L_2^w(s, T; V)$ we have

$$J(s, h, \tilde{u}) \leqq \underline{\lim}_n J(s, h_n, u_n). \tag{5.10}$$

Therefore it follows from (5.6), (5.8) and (5.10) that

$$\begin{aligned} J(s, h, \tilde{u}) \leqq \underline{\lim}_n J(s, h_n, u_n) &\leqq \overline{\lim}^n J(s, h_n, u_n) \\ &\leqq \overline{\lim}^n J(s, h_n, u) = J(s, h, u). \end{aligned} \tag{5.11}$$

Since u is the optimal control, $\tilde{u} = u$ and consequently the original sequence u_n itself converges weakly to u. As a consequence, $\phi(u)$ is the weak limit of $\phi_n(u_n)$. Using this fact for the problem (5.4) it is easy to verify that $\psi_n(u_n) \to \psi(u)$ weakly in $L_2^w(s, T; V)$. This completes the proof of the lemma.

LEMMA 5.3. *Let $\{\phi, \psi\}$ be the solution of the optimality problem* (OS) *corresponding to $h \in L_2(\beta_s, H)$ with $s \in (0, T)$ arbitrary. Then $h \to \psi(s)$ is a continuous affine mapping of $L_2(\beta_s, H)$ into $L_2(\beta_s, H)$, and there exists $P(s) \in \mathscr{L}(L_2(\beta_s, H), L_2(\beta_s, H))$ and $\gamma(s) \in L_2(\beta_s, H)$ so that $\psi(s) = P(s)h + \gamma(s)$.*

Proof. By Lemma 5.2, the mapping $h \to \{\phi, \psi\}$ is continuous from $L_2(\beta_s, H)$ into $L_2^w(s, T; V) \times L_2^w(s, T; V)$. Obviously the mapping $\{\phi, \psi\} \to \psi$ is continuous linear from $L_2^w(s, T; V) \times L_2^w(s, T; V)$ into $L_2^w(s, T; V)$. By Lemma 5.1 the adjoint problem

$$\begin{aligned} -d\psi &= (A^*\psi + L\phi(u) + g)\, dt, \\ \psi(T) &= 0 \end{aligned}$$

has a unique solution in the class $C_0^w(s, T; H)$. More specifically, ψ belongs to the class $\mathscr{X} = \{\eta \mid \eta \in L_2^w(I, V), d\eta/dt \in L_2^w(I, V')\}$, which endowed with the norm $\|\eta\|_{\mathscr{X}} = (\|\eta\|^2_{L_2^w(I,V)} + \|d\eta/dt\|^2_{L_2^w(I,V')})^{1/2}$ becomes a Hilbert space, and $\mathscr{X} \subset C_0(I, H)$. Therefore $\psi \to \psi(s)$ is a continuous linear mapping from $\mathscr{X}$ into $L_2(\beta_s, H)$. As a consequence of the above facts, $h \to \psi(s)$ is a continuous affine mapping from $L_2(\beta_s, H)$ into itself. This implies that there exists an operator $P(s) \in \mathscr{L}(L_2(\beta_s, H), L_2(\beta_s, H))$ and a vector $\gamma(s) \in L_2(\beta_s, H)$ so that $\psi(s) = P(s)h + \gamma(s)$.

Remarks. Since $s \in (0, T)$ is arbitrary it follows from the above result and the existence of unique solution $\{\xi, p\}$ of the problem (CS) (a consequence of Lemma 5.1) that

$$p(t) = P(t)\xi(t) + \gamma(t) \tag{5.12}$$

for each $t \in (0, T)$ with P and γ defined by $P(t)\xi(t) = \tilde{p}(t)$ and $\gamma(t) = \tilde{\tilde{p}}(t)$ where $\tilde{p}$ and $\tilde{\tilde{p}}$ are the solutions of the problems

$$\begin{aligned} d\tilde{\xi}(\theta) &= (A\tilde{\xi} - K\tilde{p})\, d\theta, \qquad \tilde{\xi}(t) = \xi(t), \\ -d\tilde{p}(\theta) &= (A^*\tilde{p} + L\tilde{\xi})\, d\theta, \qquad \tilde{p}(T) = 0, \end{aligned} \quad \text{for } \theta \in (t, T) \tag{5.13}$$

and

$$\begin{aligned} d\tilde{\tilde{\xi}}(\theta) &= (A\tilde{\tilde{\xi}} - K\tilde{\tilde{p}})\, d\theta + \sigma\, dW(\theta), \qquad \tilde{\tilde{\xi}}(t) = 0, \\ -d\tilde{\tilde{p}} &= (A^*\tilde{\tilde{p}} + L\tilde{\tilde{\xi}})\, d\theta + g\, d\theta, \qquad \tilde{\tilde{p}}(T) = 0, \end{aligned} \quad \text{for } \theta \in (t, T) \tag{5.14}$$

respectively.

With slight modification of the procedures of Lions [5, Lemmas 4.4, 4.5, 4.6, p. 136] we can verify that for each $t \in (0, T)$, $P(t)$ is a positive, self-adjoint, bounded linear operator in $L_2(\beta_t, H)$. Indeed, the optimality system (5.13) arises from the optimal

control problem

$$
\begin{aligned}
&d\tilde{\xi}(\theta) = (A\tilde{\xi} + Bu)\, d\theta, \qquad \theta \in (t, T),\\
&\tilde{\xi}(t) = h \in L_2(\beta_t, H),\\
&J(t, h, v) \equiv E\left\{\int_t^T [\langle L\tilde{\xi}, \tilde{\xi}\rangle + (Nv, v)_F]\, d\theta | \beta_t\right\} = \min,
\end{aligned} \tag{5.15}
$$

where the min is taken over the class $L_2^{\omega}(t, T; F)$. We verify that for $u = -N^{-1}\Lambda_F^{-1}B^*\tilde{p}$, $J(t, h, u) = (P(t)h, h)$. By Theorem 4.1 the solution of the adjoint system of (5.13) ($\Gamma = 0$, $\tau = (d/dt - A)$, $f = L\tilde{\xi}$, $\tilde{\sigma} = 0$, $\eta = 0$) satisfies the equality

$$
\langle \tilde{p}(t), \phi(t)\rangle + E^{\beta_t}\int_t^T \langle \tilde{p}(\theta), \tau\phi\rangle\, d\theta = E^{\beta_t}\int_t^T \langle L\tilde{\xi}, \phi\rangle\, d\theta
$$

for $t \in (0, T)$ and for all $\phi \in X_0$. Since $\tilde{p} \in L_2^{\omega}(I, V)$, $K\tilde{p} \in L_2^{\omega}(I, V')$, and consequently by Corollary 2.1 the problem

$$
d\tilde{\xi}(\theta) = (A\tilde{\xi} - K\tilde{p})\, d\theta, \qquad \tilde{\xi}(t) = \xi(t), \quad \theta > t
$$

has a unique solution, where $\xi(\theta)$, $\theta \in (0, T)$, is any solution of the problem

$$
d\xi(\theta) = (A\xi + Bv)\, d\theta, \quad \xi(0) = 0, \quad v \in L_2^{\omega}(I, F).
$$

Define

$$
\phi(\theta) = \begin{cases} \xi(\theta), & 0 < \theta \leqq t,\\ \tilde{\xi}(\theta), & \theta \geqq t. \end{cases}
$$

Clearly $\phi \in X_0$, and for this choice of ϕ we have

$$
\begin{aligned}
\langle \tilde{p}(t), \tilde{\xi}(t)\rangle &= -E^{\beta_t}\int_t^T \langle \tilde{p}(\theta), (\tau\tilde{\xi})(\theta)\rangle\, d\theta + E^{\beta_t}\int_t^T \langle L\tilde{\xi}, \tilde{\xi}\rangle\, d\theta\\
&= E^{\beta_t}\left\{\int_t^T \langle \tilde{p}(\theta), (K\tilde{p})(\theta)\rangle\, d\theta + \int_t^T \langle L\tilde{\xi}, \tilde{\xi}\rangle\, d\theta\right\}.
\end{aligned}
$$

Substituting $u = -N^{-1}\Lambda_F^{-1}B^*\tilde{p}$ for v in the integrand of (5.15) one can easily verify that the expression in the right-hand side of the above equality equals $J(t, \xi(t), u)$. Therefore for $\xi(t) = h$ we have $J(t, h, u) = \langle \tilde{p}(t), h\rangle = \langle P(t)h, h\rangle = (P(t)h, h)$.

Since the random evolution equation $dy = Ay\, d\theta$ with initial condition $y(t) = y_0$, $\theta \in (t, T)$, has a unique solution $y \in L_2^{\omega}(t, T; V)$ for every $y_0 \in L_2(\beta_t, H)$ (see Corollary 2.1) it is clear that A admits an essentially (norm)-bounded random evolution (or transition) operator in H. Further, since $C^*\Lambda_{\mathscr{K}}C = L \in L_{\infty}(\mathscr{F}^*, \mathscr{L}(V, V'))$, there exists a constant $k > 0$ so that $J(t, h, 0) \leqq k|h|_H^2$ μ-a.e. for each $h \in L_2(\beta_t, H)$. Thus $(P(t)h, h) = J(t, h, u) \leqq J(t, h, 0) \leqq k|h|_H^2$ μ-a.e. for each $h \in L_2(\beta_t, H)$. That $P(t)$ is positive is obvious, and that it is self-adjoint follows from the fact that the functional

$$
\nu_t(h_1, h_2) \equiv E\left\{\int_t^T [\langle L\tilde{\xi}_1, \tilde{\xi}_2\rangle + (Nu_1, u_2)]\, d\theta | \beta_t\right\}
$$

is symmetric in h_1 and h_2 on $L_2(\beta_t, H) \times L_2(\beta_t, H)$, where u_1 and u_2 are the optimal controls corresponding to the initial states h_1, h_2 of the problem (5.15); $\tilde{\xi}_1$, $\tilde{\xi}_2$ are the corresponding trajectories.

The above results only indicate that for each $t \in [0, T]$ there exists a positive self-adjoint operator $P(t) \in \mathscr{L}(L_2(\beta_t, H), L_2(\beta_t, H))$ and a vector $\gamma(t) \in L_2(\beta_t, H)$ so that

$p(t) = P(t)\xi(t) + \gamma(t)$. This, however, does not tell us anything about the regularity of the maps $t \to P(t)$, $t \to \gamma(t)$. Below we give a formal derivation of differential and integral equations that govern the evolution of P and γ as abstract random processes on $[0, T]$. We also give existence theorems for solutions of these equations.

PROPOSITION 5.1. *Suppose the operators A, B, C, N are all independent of the Wiener process W in addition to satisfying the previous assumptions. Then the evolution of the operator-valued process P and the vector-valued process γ is formally governed by the stochastic differential equations*

$$\begin{aligned} &\langle dP \cdot h, \eta\rangle + \{\langle Ah, P\eta\rangle + \langle Ph, A\eta\rangle + \langle (L - PKP)h, \eta\rangle\}\, dt = 0, \\ &P(T) = 0, \qquad h, \eta \in V \end{aligned} \tag{5.16}$$

and

$$\begin{aligned} &d\gamma + [(A^* - PK)\gamma + g]\, dt + P\sigma\, dW = 0, \\ &\gamma(T) = 0. \end{aligned} \tag{5.17}$$

Proof. The proof is formally obtained by direct substitution of the expression for P given by (5.12) into the system of equations (CS). In that the second equation is used in the sense of Theorem 4.1, specifically the equality (4.11).

The question of existence of solutions of the operator Riccati equation in infinite dimensional spaces has received considerable attention in recent years (Lions [5], Temam [9], Curtain and Pritchard [41]). Recently Bismut [2], [3] has considered the same question for a class of general linear quadratic stochastic control problems in finite dimension. Infinite dimensional problem with random operator-valued coefficients does not appear to have been considered in the literature. The operator Riccati equation in the differential form (5.16) is difficult to solve. We give a formal derivation of an integral equation or better a functional equation satisfied by the operator P which is later justified by an existence theorem following a procedure very similar to that used by Curtain and Pritchard [4]. With this end in view let us consider the Cauchy problem

$$\begin{aligned} &\tau\psi = g, \qquad t \in (0, T) = I, \\ &\psi(0) = 0 \end{aligned} \tag{5.18}$$

with $\tau \equiv (d/dt - A)$, $g \in L_2^{\omega}(I, V')$, and redefine $X_0 \equiv \{\psi\colon \psi \in L_2^{\omega}(I, V),\ \psi(0) = 0,\ \tau\psi \in L_2^{\omega}(I, V')\}$. Let, for each $t \in (0, T)$, $\Omega(t) \equiv \{\eta\colon \eta = \psi(t)$ for some $\psi \in X_0\}$ denote the attainable set of the system (5.18). As usual we call the system (5.18) controllable if for each $t \in (0, T)$ the set $\Omega(t)$ is dense in $L_2(\beta_t, H)$.

PROPOSITION 5.2. *Suppose the operators A, B, C, N satisfy the basic assumptions, are independent of the Wiener process W, and the system* (5.18) *is controllable. Further let $B \in L_{\infty}(\mathcal{F}^*, \mathcal{L}(F, H))$ and $C \in L_{\infty}(\mathcal{F}^*, \mathcal{L}(H, \mathcal{H}))$. Then the operator-valued process P satisfies the equality*

$$(P(t)h, \eta) = E^{\beta_t} \int_t^T d\theta (\Phi^*(\theta, t)[L(\theta) + (PKP)(\theta)]\Phi(\theta, t)h, \eta) \quad (\beta_t, \mu)\text{-a.e.} \tag{5.19}$$

for every $h, \eta \in L_2(\beta_t, H)$ and $t \in (0, T)$, where Φ is the evolution operator corresponding to the generator $(A - KP)$.

Proof. Under the present assumptions both K $(\equiv BN^{-1}\Lambda_F^{-1}B^*)$ and L $(\equiv C^*\Lambda_{\mathcal{H}}C)$ belong to $L_{\infty}(\mathcal{F}^*, \mathcal{L}(H, H))$ and K, $L \geqq 0$ $(dt \times d\mu)$-a.e. Further, from the remarks following Lemma 5.3, for each $t \in (0, T)$, $P(t)$ is a μ-a.e. positive, self-adjoint bounded linear operator in the Hilbert space $L_2(\beta_t, H)$.

In fact, we shall see later that $P \in L_\infty(\mathscr{F}^*, \mathscr{L}(H, H))$, and hence $KP \in L_\infty(\mathscr{F}^*, \mathscr{L}(H, H))$ also. Therefore the operator $(A - KP)$ satisfies the property (Aii), with possibly a different λ. Consequently, by the existence and uniqueness of the solution of the random evolution equation $d\xi = (A - KP)\xi\, dt$ and its continuous dependence with respect to Cauchy data (Corollary 2.1), we conclude that there exists a random transition or evolution operator Φ with generator $A - KP$ so that for each $h \in L_2(\beta_t, H)$, $t \in (0, T)$, the equation has a unique (mild) solution given by $\xi(\theta) = \Phi(\theta, t)h$, $\theta \geqq t$. This fact can also be proved by converting the evolution equation into an integral equation with the (Volterra) kernel given by the transition operator corresponding to the generator A. The properties (Ai) and (Aii) ensure the existence of such a transition operator. Since the operator P is determined by the system (5.13) and the operators A, B, C, N are all independent of the Wiener process W, we can write

$$\tilde{\xi}(\theta) = \Phi(\theta, t)h, \qquad \theta \geqq t, \quad \text{with } \tilde{\xi}(t) = \xi(t) = h, \quad h \in L_2(\beta_t, H). \tag{5.20}$$

The solution of the adjoint system in (5.13) is defined in the sense of (4.7) of Theorem 4.1 (see Remark 4.1). In this case $\tau = (d/dt - A)$, $\Gamma \equiv 0$, $f = L\tilde{\xi}$ and $\tilde{\sigma} \equiv 0$. Therefore the solution of (5.13) is given by the solution of the functional equation

$$E^{\beta_t} \int_t^T \langle \tilde{p}, \tau\phi \rangle\, d\theta + (\tilde{p}(t), \phi(t)) = E^{\beta_t} \int_t^T \langle L\tilde{\xi}, \phi \rangle\, d\theta \quad (\beta_t, \mu)\text{-a.e.} \tag{5.21}$$

for all $t \in (0, T)$ and $\phi \in X_0$.

From the representation $\tilde{p} = P\tilde{\xi}$ and the first equation of (5.13) we have $\tilde{\xi}(\theta, h) = \Phi(\theta, t)h$, $\theta \geqq t > 0$, $h \in L_2(\beta_t, H)$. Similarly, let $\tilde{\phi}(\theta, \eta) = \Phi(\theta, t)\eta$, $\theta \geqq t > 0$ denote the solution of the same problem with $\eta \in \Omega(t) \subset L_2(\beta_t, H)$. Let ν be any arbitrary element of X_0 with $\nu(t) = \eta$ and define

$$\phi(\theta) = \begin{cases} \nu(\theta), & 0 \leqq \theta \leqq t, \\ \tilde{\phi}(\theta, \eta), & t \leqq \theta \leqq T. \end{cases} \tag{5.22}$$

Since $-KP\phi \in L_2^\omega(I, H) \subset L_2^\omega(I, V')$ and the Cauchy problem (5.18) has a unique solution for each $g \in L_2^\omega(I, V')$, it is clear that $\phi \in X_0$ and $\tau\phi = -KP\phi$ for $\theta \geqq t$. Substituting these in (5.21) we obtain

$$\begin{aligned} -E^{\beta_t} \int_t^T \langle P(\theta)\Phi(\theta, t)h, K(\theta)P(\theta)\Phi(\theta, t)\eta \rangle\, d\theta + (P(t)h, \eta) \\ = E^{\beta_t} \int_t^T \langle L(\theta)\Phi(\theta, t)h, \Phi(\theta, t)\eta \rangle\, d\theta \quad (\beta_t, \mu)\text{-a.e.} \end{aligned} \tag{5.23}$$

for each $t \in (0, T)$.

Since both K and $L \in L_\infty(\mathscr{F}^*, \mathscr{L}(H, H))$ and $\tau\phi = -KP\phi \in L_2^\omega(t, T; H)$, the $V' - V$ pairing in (5.23) reduces to a scalar product in H. Thus (5.23) is equivalent to

$$(P(t)h, \eta) = E^{\beta_t} \int_t^T (\Phi^*(\theta, t)[L(\theta) + (PKP)(\theta)]\Phi(\theta, t)h, \eta)\, d\theta \quad (\beta_t, \mu)\text{-a.e.} \tag{5.24}$$

which holds for each $t \in (0, T)$ and for all $h \in L_2(\beta_t, H)$ and $\eta \in \Omega(t) \subset L_2(\beta_t, H)$. Since the system (5.18) is controllable, or equivalently $\Omega(t)$ is dense in $L_2(\beta_t, H)$, (5.24) holds for all h, $\eta \in L_2(\beta_t, H)$. This completes the proof of the proposition.

Remarks. Since L, P, K are all positive and belong to $L_\infty(\mathscr{F}^*, \mathscr{L}(H, H))$, it follows from (5.19) that the process (ν_t, β_t, μ) is a supermartingale where

$$\nu_t \equiv E^{\beta_t} \int_t^T (\Phi^*[L + PKP]\Phi h, h)\, d\theta.$$

Since $\eta \in L_2(\beta_t, H)$ is arbitrary, as an immediate consequence of the above proposition we have

COROLLARY 5.1. *Suppose the assumptions of Proposition* 5.2 *are satisfied. Then the operator-valued process P formally satisfies the random functional equation*

$$P(t)h = E^{\beta_t}\int_t^T \Phi^*(\theta, t)[L(\theta)+(PKP)(\theta)]\Phi(\theta, t)h\, d\theta \quad (\beta_t, \mu)\text{-a.e.} \tag{5.25}$$

for $t \in (0, T)$ *and for each* $h \in L_2(\beta_t, H)$.

Curtain and Pritchard [4] have developed a general perturbation technique for solving the operator Riccati equation as an integral equation on Hilbert space. In their approach a sequence of admissible controls is generated in the feedback form, thereby generating a sequence of operators $\{P_n\}$ which is then shown to converge in the strong operator topology to an operator P that satisfies the integral equation. This is a constructive approach, and in principle can be used to generate suboptimal controls for practical application. Even though the technique due to Curtain and Pritchard was developed for the deterministic operator Riccati equation, it can be suitably modified to handle stochastic problems. Following their procedure we can prove the existence of solution of the random functional equation (5.25). With this end in view let us call an operator Φ an almost sure mild evolution operator in H if

(i) $\Phi(\gamma, s)\Phi(s, t) = \Phi(\gamma, t)$ μ-a.e. for $0 \leqq t \leqq s \leqq \gamma \leqq T$,

(ii) for each $h \in L_2(\beta_t, H)$, $\Phi(\gamma, t)h \in L_2(\beta_\gamma, H)$ for $0 \leqq t \leqq \gamma \leqq T$, and

(iii) $\Phi(\gamma, t)$ is almost surely weakly continuous in t in $[0, \gamma]$ and in γ on $[t, T]$.

THEOREM 5.1. *Suppose the hypotheses of Proposition* 5.2 *hold; further, that A admits an almost sure mild evolution operator* Φ_1 *in H satisfying* $\operatorname{ess\,sup}_{0\leqq\theta\leqq\gamma\leqq T}\|\Phi_1(\gamma, \theta)\|_{\mathscr{L}(H,H)} \leqq c_1 < \infty$ μ*-almost surely and that there exists a* $\nu > 0$ *so that* $(N(t)u, u)_F \geqq \nu|u|_F^2$ μ*-almost surely for all* $t \in I = [0, T]$. *Then there exists a unique* $P \in L_\infty(\mathscr{F}^*, \mathscr{L}(H, H))$ *that satisfies the integral equation* (5.25).

Proof. For each $t \in I$, the operator $P(t)$ is entirely determined by the optimality system (5.13) which is equivalent to the optimization problem (5.15). Therefore it suffices to construct a sequence of feedback controls $\{u_n\}$, for the problem 5.15, in terms of certain positive self-adjoint operators $\{P_n\} \in L_\infty(\mathscr{F}^*, \mathscr{L}(H, H))$, so that while u_n converges to the optimal control for the problem (5.15) P_n converges to a self-adjoint positive operator $P \in L_\infty(\mathscr{F}^*, \mathscr{L}(H, H))$ that satisfies the functional equation (5.25). Define, for positive integers n,

$$\begin{aligned}
&F_n \equiv -N^{-1}\Lambda_F^{-1}B^*P_{n-1}, \qquad P_0 = 0,\\
&u_n \equiv F_n\xi, \qquad (\xi = \text{state corresponding to control } u_n, (5.15)),\\
&K_n \equiv BF_n = -KP_{n-1},\\
&M_n \equiv L + P_{n-1}KP_{n-1},\\
&P_n(t)h \equiv E^{\beta_t}\int_t^T \Phi_n^*(\theta, t)M_n(\theta)\Phi_n(\theta, t)h\, d\theta, \qquad h \in L_2(\beta_t, H),
\end{aligned} \tag{5.26}$$

where Φ_n is the evolution operator corresponding to the perturbed generator $(A + BF_n) = (A - KP_{n-1})$ of the problem (5.13). By the result of Curtain and Pritchard [4, Theorem 1.1, p. 953] the generators of mild evolution operators are closed under perturbation by operators from the class $B_\infty(\mathscr{F}^*, \mathscr{L}(H, H))$ denoted in this paper by $L_\infty(\mathscr{F}^*, \mathscr{L}(H, H))$. Thus for each n, Φ_n is an almost sure mild evolution operator in H since $BF_n \in L_\infty(\mathscr{F}^*, \mathscr{L}(H, H))$.

From (5.26) and (5.15) it is easily verified that

$$(P_n(t)h, h) = E^{\beta_t}\int_t^T \{\langle L\xi_n, \xi_n\rangle + (Nu_n, u_n)\}\, d\theta = J(t, h, u_n),$$

which is the cost of control u_n for the problem (5.15). Following closely the same technique as given by Curtain and Pritchard [4, Lemma 2.1, 2.2, p. 963], we obtain

$$(P_{n+1}(t)h, h) \leqq (P_n(t)h, h) \tag{5.27}$$

μ-almost surely for all $t \in I$ and $n \geqq 1$. Since $M_1 = L \in L_\infty(\mathscr{F}^*, \mathscr{L}(H, H))$ and by assumption $\operatorname{ess\,sup}_{0\leqq t\leqq\gamma\leqq T}\|\Phi_1(\gamma, t)\|_{\mathscr{L}(H,H)} \leqq c_1$, μ-almost surely, it follows from the defining equation for P_n in (5.26) that there exists a $c_2 > 0$ so that $\operatorname{ess\,sup}_{0\leqq t\leqq T}\|P_1(t)\|_{\mathscr{L}(H,H)} \leqq c_2$ μ-almost surely. Therefore due to (5.27) we have

$$\operatorname*{ess\,sup}_{0\leqq t\leqq T}\|P_n(t)\|_{\mathscr{L}(H,H)} \leqq c_2 \quad \mu\text{-a.s.} \quad \text{for all } n \geqq 1. \tag{5.28}$$

Since L is positive, the sequence $\{P_n\}$ is positive. Therefore it follows from (5.27) and (5.28) that for each $t \in I$, $P_n(t)$ converges strongly (strong operator topology) (β_t, μ)-almost surely to a positive self-adjoint operator $P(t)$, and $\operatorname{ess\,sup}_{0\leqq t\leqq T}\|P(t)\|_{\mathscr{L}(H,H)} \leqq c_2$ μ-almost surely. Being the almost sure strong limit uniformly in $t \in I$, $P \in L_\infty(\mathscr{F}^*, \mathscr{L}(H, H))$. Therefore, for each t, $F_n(t)$ converges strongly μ-almost surely to an operator $F(t) \in \mathscr{L}(H, F)$ and $M_n(t)$ converges weakly (weak operator topology) μ-almost surely to an operator $M(t) \in \mathscr{L}(H, H)$, and both F and M are μ-almost surely uniformly bounded in norm on $I = [0, T]$; that is, $F \in L_\infty(\mathscr{F}^*, \mathscr{L}(H, F))$ and $M \in L_\infty(\mathscr{F}^*, \mathscr{L}(H, H))$ and $M = L + PKP$. By definition, the evolution operator corresponding to the perturbed generator $(A + BF_n) = (A - KP_{n-1})$ is Φ_n and is given by the solution of the random integral equation

$$\Phi_n(\gamma, t)h = \Phi_1(\gamma, t)h + \int_t^\gamma [\Phi_1(\gamma, \theta)B(\theta)F_n(\theta)]\Phi_n(\theta, t)h\, d\theta \tag{5.29}$$

for $h \in L_2(\beta_t, H)$, $n \geqq 1$. From (5.29) and the estimates given above we have

$$\operatorname*{ess\,sup}_{0\leqq t\leqq\gamma\leqq T}\|\Phi_n(\gamma, t)\|_{\mathscr{L}(H,H)} \leqq c_1 \exp\{c_1c_2c_3b^2T/\nu\} \quad \mu\text{-a.s.}, \tag{5.30}$$

where $c_3 = \|\Lambda_F^{-1}\|$ and $b = \|B\|_{L_\infty(\mathscr{F}^*;\mathscr{L}(F,H))} = \|B^*\|$.

Let Φ be the solution of the integral equation

$$\Phi(\gamma, t)h = \Phi_1(\gamma, t)h + \int_t^\gamma [\Phi_1(\gamma, \theta)B(\theta)F(\theta)]\Phi(\theta, t)h\, d\theta \tag{5.31}$$

corresponding to the generator $(A + BF) = (A - KP)$. Then Φ also satisfies the estimate (5.30). Since F_n converges strongly μ-almost surely uniformly on I, there exists a $d > 0$ so that $\|F_n\|_{\mathscr{L}(H,F)} \leqq d$, μ-almost surely uniformly on I. Therefore by the Lebesgue dominated convergence theorem it follows from the equality

$$(\Phi_n h - \Phi h) = \int_t^\gamma \Phi_1 B(F_n - F)\Phi h\, d\theta + \int_t^\gamma \Phi_1 BF_n(\Phi_n h - \Phi h)\, d\theta$$

that Φ_n converges to Φ strongly (strong operator topology) μ-almost surely uniformly on $0 \leqq t \leqq \gamma \leqq T$. Since $P_n \to P$ strongly μ-almost surely uniformly on I, $\Phi_n \to \Phi$ strongly μ-almost surely uniformly on $0 \leqq t \leqq \gamma \leqq T$, $M_n \to M = L + PKP$ weakly μ-almost surely uniformly on I, and the operator $E^{\beta_t}(\cdot)$ is a positive idempotent linear contraction on

$L_1(\Omega, \beta, \mu)$, again by the Lebesgue dominated convergence theorem we have

$$E^{\beta_t}\int_t^\gamma (M_n(\theta)\Phi_n(\theta, t)h, \Phi_n(\theta, t)\eta)\, d\theta \to E^{\beta_t}\int_t^\gamma (M(\theta)\Phi(\theta, t)h, \Phi(\theta, t)\eta)\, d\theta$$

for arbitrary $\eta \in L_2(\beta_t, H)$. Therefore taking limits in (5.26) we obtain

$$(P(t)h, \eta) = E^{\beta_t}\int_t^T (M(\theta)\Phi(\theta, t)h, \Phi(\theta, t)\eta)\, d\theta$$

for every $\eta \in L_2(\beta_t, H)$. Thus P satisfies the integral equation

$$P(t)h = E^{\beta_t}\int_t^T \Phi^*(\theta, t)[L(\theta) + (PKP)(\theta)]\Phi(\theta, t)h\, d\theta$$

for each $t \in I = [0, T]$, and $P \in L_\infty(\mathcal{F}^*, \mathcal{L}(H, H))$.

Remarks. That the feedback control

$$u_0 = -N^{-1}\Lambda_F^{-1}B^*P\xi$$

is optimal for the problem (5.15) follows from precisely the same arguments as in Curtain and Pritchard, [4, Theorem 2.2, p. 965]. Similarly the uniqueness of the solution P follows from identical arguments as in Curtain and Pritchard [4, Theorem 2.3, p. 966].

To obtain a differential version of the operator Riccati equation, recently Curtain and Pritchard introduced the concept of a quasi-evolution operator [4, p. 957]. This concept can be obviously generalized to cover random evolution operators. Following them we call an evolution operator $\Phi(\gamma, s): \{0 \leqq s \leqq \gamma \leqq T\} \to \mathcal{L}(H, H)$ an almost sure quasi-evolution operator if (i) it is an almost sure mild evolution operator, and (ii) there exists a nonzero $h \in L_2(\beta_s, H)$ and an operator $A \in L_\infty(\mathcal{F}^*, \mathcal{L}'(H, H))$ with closed values $A(t, \omega)$ so that

$$(f, \Phi(\gamma, s)h - h) = \int_s^\gamma (f, \Phi(\gamma, \theta)A(\theta)h)\, d\theta \tag{5.32}$$

μ-almost surely for every $f \in L_2(\beta_\gamma, H)$. Here $\mathcal{L}'(H, H)$ denotes the class of linear not necessarily bounded operators in H.

As an immediate consequence of the definition (5.32) we have

$$\frac{\partial}{\partial s}(f, \Phi(\gamma, s)h) = -(f, \Phi(\gamma, s)A(s)h) \quad (dt \times d\mu)\text{-a.e.} \tag{5.33}$$

for $h \in D(A(s)) \subset L_2(\beta_s, H)$ and for all $f \in L_2(\beta_\gamma, H)$.

THEOREM 5.2. *Suppose the assumptions of Theorem* 5.1 *hold, and let* A *be the generator of an almost sure quasi-evolution operator* Φ_1 *in* H *with* $D_A \subset L_2^\omega(I, V)$ *so that for each* $x \in D_A$, $Ax \in L_2^\omega(I, H)$ *where* $I = (0, T)$. *Then* P *satisfies the operator Riccati equation in the differential form given by*

$$\begin{aligned} &E^{\beta_t}\left(\frac{dP}{dt}h, \eta\right) + (Ah, P\eta) + (Ph, A\eta) - (PKPh, \eta) + (Lh, \eta) = 0, \\ &\qquad\qquad (\beta_t, \mu)\text{-a.s. for almost all } t \in I, \\ &P(T) = 0, \end{aligned} \tag{5.34}$$

for all h, $\eta \in D_{A(t)}$. *Further,* $t \to P(t)$ *is right continuous in the weak operator topology* μ-*a.s.*

Proof. By Theorem 5.1 there exists a $P \in L_\infty(\mathscr{F}^*, \mathscr{L}(H, H))$ such that

$$(P(t)h, \eta) = E^{\beta_t} \int_t^T ([L(\theta) + (PKP)(\theta)]\Phi(\theta, t)h, \Phi(\theta, t)\eta)\, d\theta \tag{5.35}$$

for all $h, \eta \in L_2(\beta_t, H)$, where Φ is the almost sure mild evolution operator corresponding to the generator $(A - KP)$. Since, by assumption, A is the generator of an almost sure quasi-evolution operator and further $KP \in L_\infty(\mathscr{F}^*, \mathscr{L}(H, H))$, it can be easily verified that Φ is an almost sure quasi-evolution operator with $(A - KP)$ as its generator. Define

$$\begin{aligned} &\Lambda(\theta, t) \equiv ([L(\theta) + (PKP)(\theta)]\Phi(\theta, t)h, \Phi(\theta, t)\eta), \qquad 0 \leqq t \leqq \theta \leqq T, \\ &\nu_t \equiv (P(t)h, \eta), \qquad t \in (0, T) \quad \text{for } h, \eta \in D_{A(t)}. \end{aligned} \tag{5.36}$$

For $\Delta t > 0$ we can write

$$\begin{aligned} \nu_{t+\Delta t} = {} & E^{\beta_{t+\Delta t}} \int_{t+\Delta t}^T [\Lambda(\theta, t+\Delta t) - \Lambda(\theta, t)]\, d\theta \\ & + E^{\beta_{t+\Delta t}} \int_t^T \Lambda(\theta, t)\, d\theta - E^{\beta_{t+\Delta t}} \int_t^{t+\Delta t} \Lambda(\theta, t)\, d\theta, \end{aligned}$$

and since $\beta_t \subset \beta_{t+\Delta t}$ and ν_t is β_t-measurable we have

$$\begin{aligned} E^{\beta_t}\nu_{t+\Delta t} = {} & E^{\beta_t} \int_{t+\Delta t}^T [\Lambda(\theta, t+\Delta t) - \Lambda(\theta, t)]\, d\theta + E^{\beta_t} \int_t^T \Lambda(\theta, t)\, d\theta \\ & - E^{\beta_t} \int_t^{t+\Delta t} \Lambda(\theta, t)\, d\theta. \end{aligned} \tag{5.37}$$

Consequently, for $\Delta t > 0$ we have

$$\begin{aligned} E^{\beta_t}\left(\frac{\nu_{t+\Delta t} - \nu_t}{\Delta t}\right) = {} & E^{\beta_t} \int_{t+\Delta t}^T \left\{\frac{\Lambda(\theta, t+\Delta t) - \Lambda(\theta, t)}{\Delta t}\right\} d\theta \\ & - E^{\beta_t} \frac{1}{\Delta t} \int_t^{t+\Delta t} \Lambda(\theta, t)\, d\theta. \end{aligned} \tag{5.38}$$

Since Φ is an almost sure quasi-evolution operator and $Ax \in L_2^\omega(I, H)$ for any $x \in D_A$, the expressions in the right-hand side of (5.38) are well defined for all $\Delta t > 0$. On letting $\Delta t \downarrow 0$, recalling the definition of ν_t and noting that $\Lambda(t, t)$ is β_t-measurable, we have

$$E^{\beta_t}\left(\frac{dP}{dt}h, \eta\right) = E^{\beta_t} \int_t^T \frac{\partial \Lambda}{\partial t}(\theta, t)\, d\theta - \Lambda(t, t). \tag{5.39}$$

Using (5.33) and recalling that $(A - KP)$ is the generator of Φ we have

$$\begin{aligned} \frac{\partial \Lambda}{\partial t}(\theta, t) = {} & -\{([L(\theta) + (PKP)(\theta)]\Phi(\theta, t)(A(t) - (KP)(t))h, \Phi(\theta, t)\eta) \\ & + ([L(\theta) + (PKP)(\theta)]\Phi(\theta, t)h, \Phi(\theta, t)(A(t) - (KP)(t))\eta)\}, \end{aligned}$$

and since $(L + PKP)$ is self-adjoint we can write the above expression as

$$\frac{\partial \Lambda}{\partial t}(\theta, t) = -\{(\Phi^*(L + PKP)\Phi\eta, (A - KP)h) + (\Phi^*(L + PKP)\Phi h, (A - KP)\eta)\}.$$

Therefore we have, due to (5.19),

$$E^{\beta_t}\int_t^T \frac{\partial}{\partial t}\Lambda(\theta, t)\,d\theta = -(A(t)h, P(t)\eta) - (A(t)\eta, P(t)h) + 2(K(t)P(t)h, P(t)\eta), \tag{5.40}$$

$$\Lambda(t, t) = ([L(t) + (PKP)(t)]h, \eta).$$

Substituting (5.40) into (5.39) we obtain (5.34). The continuity assertion follows from (5.37). This completes the proof of the theorem.

For the backward stochastic evolution equation (5.17) we have the following result.

THEOREM 5.3. *Suppose the assumptions of Theorem* 5.1 *hold and the operator* σ *belongs to* $L_\infty(\mathcal{F}^*, \mathcal{L}(H, H))$ *in addition to satisfying the basic assumption* (Aiv). *Then the system* (5.17) *has a unique weak solution* $\gamma \in L_2^\omega(I, V)$ *in the sense that*

$$\int_0^T \langle \gamma, \tau\psi\rangle\,d\theta = l(\psi) \quad \mu\text{-a.s.} \quad \text{for all } \psi \in X_0$$

where

$$\tau \equiv \frac{d}{dt} - (A - KP)$$

and

$$l(\psi) \equiv \int_0^T \langle g, \psi\rangle\,d\theta + \int_0^T (\sigma^* P\psi, dW).$$

Further, $t \to \gamma(t)$ *is weakly right continuous* μ*-a.s.*

Proof. The proof follows from Theorem 4.1 with $\Gamma = KP$, which is self-adjoint and, due to Theorem 5.1, belongs to $L_\infty(\mathcal{F}^*, \mathcal{L}(H, H))$ as required, and $\eta = 0$.

THEOREM 5.4. *Under the assumptions of Theorem* 5.3 *the feedback control* $u = -N^{-1}\Lambda_F^{-1}B^*(P\xi + \gamma)$ *obtained from* (5.1) *and* (5.12) *is optimal.*

Proof. Let P and γ denote the solutions of (5.16) and (5.17). Substitute the expression for u into the equation $d\xi = (A\xi + Bu)\,dt + \sigma \cdot dW$. Let $s \in (0, T)$ and consider the problem

$$\text{S}' \qquad \begin{aligned} &d\xi' = (A - KP)\xi'\,dt - K\gamma\,dt + \sigma\,dW, \qquad t \in (s, T), \\ &\xi'(s) = h \in L_2(\beta_s, H). \end{aligned}$$

Since $K \in L_\infty(\mathcal{F}^*, \mathcal{L}(H, H))$ and $\gamma \in L_2^\omega(I, V)$ we have $K\gamma \in L_2^\omega(I, H)$; similarly since $P \in L_\infty(\mathcal{F}^*, \mathcal{L}(H, H))$ (Theorem 5.1) we have $KP \in L_\infty(\mathcal{F}^*, \mathcal{L}(H, H))$. Thus by Theorem 2.1 the problem S′ has a unique solution $\xi' \in L_2^\omega(s, T; V) \cap L_\infty^\omega(s, T; H)$. Define $p' = (P\xi' + \gamma)$. Then using (5.16) and (5.17) in the equality $\langle dp', \eta\rangle = \langle dP \cdot \xi', \eta\rangle + \langle P \cdot d\xi', \eta\rangle + \langle d\gamma, \eta\rangle$ one obtains

$$\text{S}'' \qquad \begin{aligned} &-dp' = (A^*p' + L\xi')\,dt + g\,dt, \qquad t \in (s, T), \\ &p'(T) = 0. \end{aligned}$$

The two systems in S′ and S″ are identical to those in the optimality system (OS) of Lemma 5.1. Since these problems have unique solutions (Lemma 5.1) we have $\{\xi', p'\} = \{\phi, \psi\}$. Therefore by Theorem 4.2 and Lemma 5.1 the control

$$u = -N^{-1}\Lambda_F^{-1}B^*\psi = -N^{-1}\Lambda_F^{-1}B^*p' = -N^{-1}\Lambda_F^{-1}B^*(P\xi' + \gamma)$$

is optimal (see the proof of Lemma 5.1). Since $s \in (0, T)$ and $h \in L_2(\beta_s, H)$ are arbitrary, this proves the theorem.

6. Examples. In this section we present a few special but interesting examples.

Example i. Let $\{(A_i, B_i), i = 1, 2, \cdots, N\}$ be a set of fixed elements of $\mathscr{L}(V, V') \times \mathscr{L}(F, H)$ with the property that for all $x, y \in V$ there exist finite positive numbers δ, α such that

$$\langle A_i x, y\rangle \leqq \delta |x|_V |y|_V,$$
$$\langle A_i x, x\rangle \geqq \alpha |x|_V^2,$$
$$\|B_i\|_{\mathscr{L}(F,H)} \leqq \delta,$$

for all $i = 1, 2, \cdots, N$. Let us denote this set by $\Sigma \subset \mathscr{L}(V, V') \times L(F, H)$, and suppose the process $\{(A(t), B(t)), t \in I\}$ is a separable homogeneous Markov chain with state space Σ and parameters a_i, $i, j = 1, 2, \cdots, N$ so that $a_{ij} \geqq 0$ for $i \neq j$ and $a_{ii} = -\sum_{j \neq i} a_{ij}$.

Let N, C be deterministic observable processes and belong to $L_\infty(I, \mathscr{L}(F, F))$ and $L_\infty(I, \mathscr{L}(V, \mathscr{H}))$ respectively; and suppose $Z_d \equiv 0$, $\sigma \equiv 0$. In this case the optimal feedback control at time t, given that $A(t) = A_i$, $B(t) = B_i$ and $\xi(t) = x$, is

$$u^0(t, x, i) = -N^{-1}(t)\Lambda_F^{-1} B_i^* (P_i(t)x), \tag{6.1}$$

where $\{P_i, i = 1, 2, \cdots, N\}$ satisfy the system of deterministic (operator) Riccati equations

$$\frac{d}{dt}(P_i h, \eta) + (A_i h, P_i \eta) + (P_i h, A_i \eta) + \sum_{j=1}^{N} a_{ij}(P_j h, \eta) - (P_i K_i P_i h, \eta) + (Lh, \eta) = 0, \tag{6.2}$$
$$P_i(T) = 0, \qquad K_i(t) \equiv B_i N^{-1}(t)\Lambda_F^{-1} B_i^*, \qquad i = 1, 2, \cdots, N, \quad t \in (0, T), \quad h, \eta \in V.$$

This equation follows from (5.34) of Theorem 5.2. Further, since $Z_d \equiv 0$ and $\sigma \equiv 0$, we have $\gamma \equiv 0$. For practical realization of the control given by (6.1) it is required to follow the Markov chain $\{A(t), B(t)), t \in I\}$ to determine its switching time and the state in addition to observing the state $\xi(t)$. This example extends the result of Sworder [6] from finite to infinite dimensional space.

Example ii. Suppose that all the assumptions of Example i hold except that $Z_d \not\equiv 0$. In this case the optimal feedback control is given by

$$u(t, x, i) = -N^{-1}(t)\Lambda_F^{-1} B_i^* (P_i(t)x + \gamma_i(t)), \tag{6.3}$$

where $\{P_i, i = 1, 2, \cdots, n\}$ satisfy the system (6.2) and $\{\gamma_i, i = 1, 2, \cdots, n\}$ satisfy the system of equations

$$\frac{d\gamma_i}{dt} + (A_i^* - P_i K_i)\gamma_i + \sum_{j=1}^{N} a_{ij}\gamma_j + g = 0, \qquad t \in (0, T),$$
$$\gamma_i(T) = 0, \qquad g = -C^* \Lambda_{\mathscr{H}} Z_d, \quad i = 1, 2, \cdots, N \tag{6.4}$$

in the weak sense; that is, for all $v \in V$

$$\frac{d}{dt}\langle \gamma_i, v\rangle + \langle \gamma_i, (A_i - K_i P_i)v\rangle + \sum_{j=1}^{N} a_{ij}\langle \gamma_j, v\rangle + \langle g, v\rangle = 0,$$
$$\gamma_i(T) = 0, \qquad t \in (0, T), \quad i = 1, 2, \cdots, N. \tag{6.4'}$$

Example iii. Let (A_1, B_1, C_1, N_1) and (A_2, B_2, C_2, N_2) be two families of constant operators, with $A_i \in \mathscr{L}(V, V')$, $B_i \in \mathscr{L}(F, H)$, $C_i \in \mathscr{L}(V, \mathscr{H})$ and $N_i \in \mathscr{L}(F, F)$ having the properties given in § 2. Let τ be a positive random variable with density $\mu(\tau \in dt) = \lambda e^{-\lambda t}\, dt$, $\lambda > 0$, and $\{\eta_t, t \geqq 0\}$ the increasing family of σ-algebras generated by the indicator function of the set $\{t < \tau\}$. Clearly $\eta_t = \sigma(1_{t \wedge \tau})$ and it is a right continuous family of σ-algebras with no time of discontinuity, and τ is a totally inaccessible stopping time (Meyer [9, VII, 54b]). Define $\tau_0 = T \wedge \tau$ and consider the system

$$\begin{aligned} d\xi &= 1_{t<\tau_0}(A_1\xi + B_1 u)\, dt + 1_{t \geqq \tau_0}(A_2\xi + B_2 u)\, dt, \\ \xi(0) &= \xi_0. \end{aligned} \tag{6.5}$$

It is desired to minimize the cost function

$$J(u) = E\int_0^T \{1_{t<\tau_0}(|C_1\xi|_{\mathscr{H}}^2 + (N_1 u, u)_F) + 1_{t \geqq \tau_0}(|C_2\xi|_{\mathscr{H}}^2 + (N_2 u, u)_F)\}\, dt. \tag{6.6}$$

From the results of the previous section it follows that the optimal feedback control, given the present state $\xi(t) = x$, and the structural state $\mathbf{i} \equiv (A_i, B_i, C_i, N_i)$, is

$$u(t, x, i) = -N_i^{-1} \Lambda_F^{-1} B_i^* P_i(t) x, \tag{6.7}$$

where P_i $(i = 1, 2)$ are the solutions of the operator Riccati equations

$$\begin{aligned} &\frac{d}{dt}(P_1 h, \eta) + (A_1 h, P_1 \eta) + (P_1 h, A_1 \eta) + \lambda((P_2 - P_1)h, \eta) + (P_1 K_1 P_1 h, \eta) + (L_1 h, \eta) = 0, \\ &\frac{d}{dt}(P_2 h, \eta) + (A_2 h, P_2 \eta) + (P_2 h, A_2 \eta) - (P_2 K_2 P_2 h, \eta) + (L_2 h, \eta) = 0, \\ &P_1(T) = P_2(T) = 0, \qquad t \in (0, T) \quad \text{for all } h, \eta \in V. \end{aligned} \tag{6.8}$$

where $K_i \equiv B_i N_i^{-1} \Lambda_F^{-1} B_i^*$ and $L_i \equiv C_i^* \Lambda_{\mathscr{H}} C_i$, $i = 1, 2$.

The equations (6.8) follow from (6.2) with the parameters a_{ij}, $i, j = 1, 2$, having values $a_{11} = -\lambda$, $a_{12} = \lambda$; $a_{21} = a_{22} = 0$. Since Z_d and hence $g \equiv 0$ it follows that $\gamma \equiv 0$. The operator P is given by

$$P(t) = 1_{t<\tau_0} P_1(t) + 1_{t \geqq \tau_0} P_2(t). \tag{6.9}$$

This example is similar to the finite dimensional example given by Bismut [2, ex. 4, p. 439]. It is clear that many other variants of this example are possible and covered by our Example i: for instance, one in which there are three structural states $\mathbf{i}$, $i = 1, 2, 3$ and transitions are possible only from **1** to **2**, **2** to **3** and from **2** to **1**, with **3** being the absorbing state. In this case

$$\begin{gathered} a_{11} = -\lambda_1, \qquad a_{12} = \lambda_1, \qquad a_{13} = 0, \\ a_{21} = \lambda_3, \qquad a_{22} = -(\lambda_2 + \lambda_3), \qquad a_{23} = \lambda_2, \\ a_{31} = a_{32} = a_{33} = 0. \end{gathered}$$

Using (6.2) one can immediately write the three operator equations for P_1, P_2 and P_3.

Acknowledgment. The author would like to thank Professor P. P. Varaiya of the University of California, Berkeley, Professor S. K. Mitter of MIT and Professor D. A. Dawson of Carleton University for many valuable discussions during the preparation of this paper.

REFERENCES

[1] A. BENSOUSSAN AND B. VIOT, *Optimal control of stochastic linear distributed parameter systems*, this Journal, 13 (1975), pp. 904–926.

[2] J. M. BISMUT, *Linear quadratic optimal stochastic control with random coefficients*, this Journal, 14 (1976), pp. 419–444.

[3] ———, *On optimal control of linear stochastic equations with a linear-quadratic criterion*, this Journal, 15 (1977), pp. 1–4.

[4] R. CURTAIN AND A. J. PRITCHARD, *The infinite dimensional Riccati equation for systems defined by evolution operators*, this Journal, 14 (1976), pp. 951–983.

[5] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer, New York, 1971.

[6] D. D. SWORDER, *Feedback control of a class of linear systems with jump parameters*, IEEE Trans. Automat. Control, 14 (1969), pp. 9–14.

[7] ———, *On the stochastic maximum principle*, J. Math. Anal. Appl., 24 (1968), pp. 627–640.

[8] R. TEMAM, *Sur l'équation de Riccati associée a des opérateurs nonbornés en dimension infinie*, J. Funct. Anal., 7 (1971), pp. 85–115.

[9] P. A. MEYER, *Probabilités et potentiel*, (Eng. trans.) Blaisdell, Waltham, MA, 1966.

[10] J. HORVATH, *Topological Vector Spaces and Distributions*, Vol. 1, Addison-Wesley, Reading, MA, 1966.

[11] R. W. CARROLL, *Abstract Methods in Partial Differential Equations*, Harper and Row, New York, 1969.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 3, May 1981

0363-0129/81/1903-0006 $01.00/0

ERRATUM: SUFFICIENT CONDITIONS FOR KUHN-TUCKER VECTORS IN CONVEX PROGRAMMING*

P. LEVINE† AND J. CH. POMEROL†

The proof of Theorem 5.1 (p. 694) is not correct because the set $\bar{V}$ depends on u. This theorem should read as follows.

THEOREM 5.1. *The following assertions are equivalent:*

(i) $\varphi(0)$ *is finite, and there exist* U_0 *and* V_0, *two* 0*-neighborhoods in* U *and* V *respectively, such that*

$$\forall (u, v) \in U_0 \times V_0, \qquad \varphi_v(u) \leqq a < +\infty.$$

(ii) (C_3) *is satisfied.*

The inequality (1) in the proof results from $\gamma_u(v) = \operatorname{lsc} \varphi_v(u)$ which entails $-g^*_{V_0}(u) \leqq a$ whenever $u \in \operatorname{int}(U_0)$. The remainder of the proof is unchanged.

As a consequence Theorem 5.1′ must be changed. For any 0-neighborhood W in V we introduce the functional $\delta_W(u) = \sup_{v \in W} \varphi_v(u) = \sup_{v \in W} \inf_{x \in X} (F(x, u) - \langle x, v \rangle)$. Then Theorem 5.1′ becomes

THEOREM 5.1′. *The following assertions are equivalent:*

(i) *One has* $-\infty < \operatorname{lsc} \varphi(0)$ *and there exists a* 0*-neighborhood* W *in* V *such that* $0 \in \operatorname{core} \operatorname{dom} \delta_W$.

(ii) *There exist a* 0*-neighborhood* W *in* V *and a real* β *satisfying* $\beta < \gamma(0)$ *such that*

$$\{y \mid \exists v \in W \text{ satisfying } G(y, v) \geqq \beta\} \text{ is bounded.}$$

Also, we have wrongly asserted in the proofs of Theorems 4.1 and 4.2 (p. 692) that a weak*-convergent generalized sequence in the dual of a Banach space is strongly bounded. It follows that the condition (C_4) is not sufficient when X is a Banach space, whereas Theorem 4.2 remains true from the Krein–Smulian theorem or C-closed mappings (see [11]). The same erroneous property was used in Theorem 5.2; hence, when X is a Banach space, we only have (i)$\Rightarrow$(ii) in Theorem 5.2 and Corollaries 5.1 and 5.1′.

Finally, when these two errors are accounted for, Theorem 5.2 should read:

THEOREM 5.2. *If* V *is normed in a topology compatible with the pairing, then the following assertions are equivalent*:

(i) $\varphi(0)$ *is finite and there exist a* 0*-neighborhood* U_0 *in* U *and a ball* V_k *of radius* $k > 0$ *centered at the origin in* V *such that*

$$\forall (u, v) \in U_0 \times V_k, \qquad \varphi_v(u) \leqq a < +\infty.$$

(ii) *There exist* $k > 0$ *and* $\beta < \gamma(0)$ *such that the set* $\{y \mid \exists v$ *satisfying* $\|v\| \leqq k$ *and* $G(y, v) \geqq \beta\}$ *is equicontinuous.*

Moreover (i) *implies* (ii) *when* X *is a Banach space.*

An analogous modification has to be made with Corollaries 5.1 and 5.1′.

In conclusion, when V is normed in a compatible topology, the tables (p. 698) remain unchanged due to the following proposition.

* This Journal, 17 (1979), pp. 689–699. Received by the editors November 19, 1979.
† Laboratoire d'Econométrie, Université P. et M. Curie, 75230 Paris CEDEX 5, France.

PROPOSITION. *Assume that either X is a Banach space or V is normed in a topology compatible with the pairing. Then the following assertions are equivalent*:

(i) $0 \in \text{core dom } \varphi_0$.

(ii) *There exists a ball W centered at the origin in V such that*

$$0 \in \text{core dom } \delta_W.$$

Furthermore, when X is a Banach space (C_4) and (CS_4) must be suppressed in the tables, and we have $(C_2) \Rightarrow (C_8)$ instead of the equivalence.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 4, July 1981

0363-0129/81/1904-0001 $01.00/0

LOWER SEMICONTINUITY, OPTIMIZATION AND REGULARIZING EXTENSIONS OF INTEGRAL FUNCTIONALS*

FRANCESCO FERRO†

Abstract. We prove the lower semicontinuity, relative to the local convergence topology induced in $W^{1,1}(\Omega)$ by $L^1_{\text{loc}}(\Omega)$, of a large class of integral functionals, under two different sets of hypotheses. We obtain also an optimization theorem and apply these results to the study of some extensions of integral functionals.

1. Introduction. In this work we study the lower semicontinuity of integral functionals of the following type:

$$I(u)=\int_\Omega L(x, u(x), \nabla u(x))\,dx, \qquad u\in W^{1,1}(\Omega),$$

where L satisfies very mild regularity conditions (see § 2). In § 2 we prove two theorems which assure the lower semicontinuity of I relative to the strong convergence topology induced in $W^{1,1}(\Omega)$ by $L^1(\Omega)$. The first theorem is proved under a mild growth condition about L and a supplementary condition about the subdifferential of L (we remark that these conditions are satisfied, for example, by the well-known integrand $(1+|v|^2)^{1/2}$), which does not allow for L the value $+\infty$. On the contrary in the second theorem the value $+\infty$ is allowed for L, but we require a stronger growth condition on L. For integrands which are greater than a summable function (e.g., nonnegative integrands) we obtain also lower semicontinuity results relative to the local convergence topology induced in $W^{1,1}(\Omega)$ by $L^1_{\text{loc}}(\Omega)$ (Corollaries 2.1 and 2.2). In § 3 we consider some extensions of I to a larger class of functions and, by the lower semicontinuity theorems, we obtain new results about the behavior of these extensions on "regular" functions (i.e., on the elements of $W^{1,1}(\Omega)$).

We prove also a result about the existence of a minimum (Theorem 3.3) which improves results of [3] and extends to the n-dimensional case the arguments in [12]. Our results differ from the classical ones in [9], [14] since, among other things, we do not assume any continuity hypothesis on L.

2. Lower semicontinuity. Throughout this paper Ω will be an open and bounded set in $\mathbb{R}^n$ whose boundary $\partial\Omega$ satisfies the local Lipschitz condition. $W^{1,1}(\Omega)$ is the space of all functions in $L^1(\Omega)$ whose first partial derivatives (in the distribution sense) are summable in Ω.

Let $L:\Omega\times\mathbb{R}\times\mathbb{R}^n\to\mathbb{R}\cup\{+\infty\}$ be a proper normal integrand (see [13]); that is $L(x,\cdot,\cdot)$ is lower semicontinuous and not identically $+\infty$ for every $x\in\Omega$ and L is $\mathcal{L}\times\mathcal{B}$-measurable (we mean that L is measurable relative to the σ-algebra generated by all products of Lebesgue-measurable sets in Ω, and Borel sets in $\mathbb{R}\times\mathbb{R}^n$). We define

$$I(u)=\int_\Omega L(x, u(x), \nabla u(x))\,dx, \qquad u\in W^{1,1}(\Omega). \tag{2.1}$$

We remark that our assumptions on L assure that $L(x, u(x), v(x))$ is measurable (i.e., Lebesgue-measurable) if u and v are; then $I(u)$ is well defined by (2.1) if $L(x, u(x), \nabla u(x))$ is summable; otherwise we put $I(u)=-\infty$ if $L(x, u(x), \nabla u(x))$ is

* Received by the editors June 11, 1979.

† Istituto di Matematica, via L. B. Alberti 4, 16100 Genoa, Italy. This work was partially supported by Laboratorio per la Matematica Applicata del C.N.R., Genoa.

majorized by a summable function and $I(u) = +\infty$ in every other case. Throughout this paper we suppose that there exists $u \in W^{1,1}(\Omega)$ such that $I(u) \in \mathbb{R}$.

If $L(x, u, \cdot)$ is convex we denote by $\partial_v L(x, u, v_0)$ the subdifferential of $L(x, u, \cdot)$ at the point v_0; i.e.,

$$\partial_v L(x, u, v.) = \{p \in \mathbb{R}^n : L(x, u, w) \geqq L(x, u, v_0) + p(w - v_0), \text{ for every } w \in \mathbb{R}^n\}.$$

We define

$$H(x, u, p) = \sup\{pv - L(x, u, v) : v \in \mathbb{R}^n\}.$$

$H(x, u, \cdot)$ is convex, and we denote by $\partial_p H(x, u, p_0)$ the subdifferential of $H(x, u, \cdot)$ at the point p_0.

If $L(x, u, \cdot)$ is convex, by standard results about convex analysis (see [10]), we have

$$L(x, u, v) = \sup\{pv - H(x, u, p) : p \in \mathbb{R}^n\}.$$

An easy calculation proves the following.

LEMMA 2.1. *Let $L(x, u, \cdot)$ be convex for every $(x, u) \in \Omega \times \mathbb{R}$; let $h_0 \in L^1(\Omega)$, $M_0 > 0$, $K_0 > 0$.*
Then we have

$$L(x, u, v) \geqq K_0|v| - M_0|u| - h_0(x) \quad \text{for every } (x, u, v) \in \Omega \times \mathbb{R} \times \mathbb{R}^n, \tag{2.2}$$

if and only if,

$$H(x, u, p) \leqq h_0(x) + M_0|u| \quad \text{for every } (x, u) \in \Omega \times \mathbb{R} \quad \text{and} \quad |p| \leqq K_0. \tag{2.3}$$

The following result is proved by a proper use of the techniques given in [12].

LEMMA 2.2. *Let $L(x, u, \cdot)$ be convex for every $(x, u) \in \Omega \times \mathbb{R}$ and let (2.2) hold. Then $H(x, \cdot, p)$ is upper semicontinuous for every $x \in \Omega$ and $|p| < K_0$.*

Proof. Let $\bar{p} \in \mathbb{R}^n$ such that $|\bar{p}| < K_0$. We set

$$H_1(x, u, \bar{p}, v) = L(x, u, v) - \bar{p}v.$$

We have

$$-H(x, u, \bar{p}) = \inf\{H_1(x, u, \bar{p}, v) : v \in \mathbb{R}^n\}.$$

By [12, equivalence theorem] $H(x, \cdot, \bar{p})$ is upper semicontinuous if the set $M(x, r, \alpha) = \{v$: there exists $u \in \mathbb{R}$ such that $|u| \leqq r$, $H_1(x, u, \bar{p}, v) \leqq \alpha\}$ is bounded. This is true since

$$\begin{aligned} \alpha \geqq H_1(x, u, \bar{p}, v) = L(x, u, v) - \bar{p}v &\geqq K_0|v| - M_0|u| - h_0(x) - |\bar{p}||v| \\ &= (K_0 - |\bar{p}|)|v| - M_0 r - h_0(x). \end{aligned}$$

□

Now we may prove our first lower semicontinuity result.

THEOREM 2.1. *Let $L(x, u, \cdot)$ be convex for every $(x, u) \in \Omega \times \mathbb{R}$ and let (2.2) hold; moreover, suppose that for every $u \in W^{1,1}(\Omega)$ there exists $p_u \in (L^\infty(\Omega))^n$ such that $\|p_u\|_{(L^\infty(\Omega))^n} \leqq K_0$ and*

$$\nabla u(x) \in \partial_p H(x, u(x), p_u(x)) \quad \text{a.e. in } \Omega. \tag{2.4}$$

Then I is lower semicontinuous in $W^{1,1}(\Omega)$ relative to strong convergence in $L^1(\Omega)$.

Proof. Let $u \in W^{1,1}(\Omega)$; by (2.4) and [10, Thm. 23.5] we have

$$p_u(x)\nabla u(x) - H(x, u(x), p_u(x)) = L(x, u(x), \nabla u(x)).$$

(We remark that $H(x, u(x), p)$ is a normal integrand on $\Omega \times \mathbb{R}^n$ by [13, Prop. 2S].)

By (2.3) and [13, Thm. 3C] we have

$$\int_\Omega L(x, u(x), \nabla u(x))\, dx = \sup\left\{\int_\Omega p(x)\nabla u(x)\, dx - \int_\Omega H(x, u(x), p(x))\, dx : p \in (L^\infty(\Omega))^n\right\},$$

and so

$$\begin{aligned}
&\int_\Omega p_u(x)\nabla u(x)\, dx - \int_\Omega H(x, u(x), p_u(x))\, dx \\
&\quad = \int_\Omega L(x, u(x), \nabla u(x))\, dx \\
&\quad = \sup\left\{\int_\Omega p(x)\nabla u(x)\, dx - \int_\Omega H(x, u(x), p(x))\, dx : p \in (L^\infty(\Omega))^n\right\} \\
&\quad \geqq \int_\Omega p_u(x)\nabla u(x)\, dx - \int_\Omega H(x, u(x), p_u(x))\, dx.
\end{aligned}$$

Hence we may write

$$\begin{aligned}
&\int_\Omega L\,(x, u(x), \nabla u(x))\, dx \\
&\quad = \int_\Omega p_u(x)\nabla u(x)\, dx - \int_\Omega H(x, u(x), p_u(x))\, dx \\
&\quad = \sup\left\{\int_\Omega p(x)\nabla u(x)\, dx - \int_\Omega H(x, u(x), p(x))\, dx : \|p\|_{(L^\infty(\Omega))^n} \leqq K_0\right\}.
\end{aligned}$$

Now let $p \in (L^\infty(\Omega))^n$ and $p_h(x) = (1-h)p(x)$, $h > 0$; then

$$\|p_h\|_{(L^\infty(\Omega))^n} < \|p\|_{(L^\infty(\Omega))^n} \quad \text{and} \quad \lim_{h\to 0} p_h(x) = p(x) \quad \text{a.e. in } \Omega.$$

Therefore if $\|p\|_{(L^\infty(\Omega))^n} \leqq K_0$ we have

$$H(x, u(x), p_h(x)) \leqq h_0(x) + M_0.|u(x)|$$

and, by Fatou's lemma,

$$\begin{aligned}
&\liminf_{h\to 0}\left(\int_\Omega p_h(x)\nabla u(x)\, dx - \int_\Omega H(x, u(x), p_h(x))\, dx\right) \\
&\quad = \int_\Omega p(x)\nabla u(x)\, dx - \limsup_{h\to 0}\int_\Omega H(x, u(x), p_h(x))\, dx \\
&\quad \geqq \int_\Omega p(x)\nabla u(x)\, dx - \int_\Omega \lim_{h\to 0} H(x, u(x), p_h(x))\, dx \\
&\quad = \int_\Omega p(x)\nabla u(x)\, dx - \int_\Omega H(x, u(x), p(x))\, dx,
\end{aligned}$$

where the last equality holds since $H(x, u(x), \cdot)$, which is convex and lower semicontinuous by definition, is finite (by (2.3)), and so continuous, in the interior of the ball of

radius K_0. If $|p(x)| = K_0$, we have $\lim_{h\to 0} H(x, u(x), p_h(x)) = H(x, u(x), p(x))$, since $p_h(x)$ approaches $p(x)$ along a line segment from the interior of the ball of radius K_0 (see [10, Corollary 7.5.1]). Then we have

$$\int_\Omega L(x, u(x), \nabla u(x))\, dx$$
$$= \sup\left\{\int_\Omega p(x)\nabla u(x)\, dx - \int_\Omega H(x, u(x), p(x))\, dx : \|p\|_{(L^\infty(\Omega))^n} < K_0\right\}.$$

Finally if $p \in (L^\infty(\Omega))^n$ and $\|p\|_{(L^\infty(\Omega))^n} < K_0$, there exists $p_h \in (C_0^\infty(\Omega))^n$ such that $\|p_h\|_{(L^\infty(\Omega))^n} \leqq \|p\|_{(L^\infty(\Omega))^n} < K_0$ and $\lim_{h\to 0} p_h(x) = p(x)$ a.e. in Ω.

As before, by the continuity of $H(x, u(x), \cdot)$ in the interior of the ball of radius K_0, we obtain

$$\liminf_{h\to 0}\left(\int_\Omega p_h(x)\nabla u(x)\, dx - \int_\Omega H(x, u(x), p_h(x))\, dx\right)$$
$$\geqq \int_\Omega p(x)\nabla u(x)\, dx - \int_\Omega H(x, u(x), p(x))\, dx;$$

so we have

$$\int_\Omega L(x, u(x), \nabla u(x))\, dx$$
$$= \sup\left\{\int_\Omega p(x)\nabla u(x)\, dx - \int_\Omega H(x, u(x), p(x))\, dx : \|p\|_{(L^\infty(\Omega))^n} < K_0,\ p \in (C_0^\infty(\Omega))^n\right\}. \tag{2.5}$$

Now let $\{u_m\} \subset W^{1,1}(\Omega)$ and $u_0 \in W^{1,1}(\Omega)$ be such that $u_m \to u_0$ in $L^1(\Omega)$; it follows that $\nabla u_m \to \nabla u_0$ in the distribution sense. Moreover if $p \in (C_0^\infty(\Omega))^n$ and $\|p\|_{(L^\infty(\Omega))^n} < K_0$, we have by (2.3) that $H(x, u_m(x), p(x)) - M_0|u_m(x)| \leqq h_0(x)$, so we may use Fatou's lemma, and for a suitable subsequence $\{u_r\}$ of $\{u_m\}$ such that $u_r \to u_0$ a.e. in Ω and $\limsup_{m\to+\infty} \int_\Omega H(x, u_m(x), p(x))\, dx = \lim_{r\to+\infty} \int_\Omega H(x, u_r(x), p(x))\, dx$, we have

$$\liminf_{m\to+\infty}\left(\int_\Omega p(x)\nabla u_m(x)\, dx - \int_\Omega H(x, u_m(x), p(x))\, dx\right)$$
$$= \int_\Omega p(x)\nabla u_0(x)\, dx - \limsup_{m\to+\infty} \int_\Omega H(x, u_m(x), p(x))\, dx$$
$$= \int_\Omega p(x)\nabla u_0(x)\, dx - \lim_{r\to+\infty}\left(\int_\Omega H(x, u_r(x), p(x))\, dx - M_0\int_\Omega |u_r(x)|\, dx\right)$$
$$- \lim_{r\to+\infty} M_0 \int_\Omega |u_r(x)|\, dx$$
$$\geqq \int_\Omega p(x)\nabla u_0(x)\, dx - \int_\Omega \limsup_{r\to+\infty} (H(x, u_r(x), p(x)) - M_0|u_r(x)|)\, dx$$
$$- M_0 \int_\Omega |u_0(x)|\, dx$$
$$\geqq \int_\Omega p(x)\nabla u_0(x)\, dx - \int_\Omega H(x, u_0(x), p(x))\ dx \text{ by Lemma 2.2.}$$

Therefore the functional

$$u \to \int_\Omega p(x)\nabla u(x)\,dx - \int_\Omega H(x, u(x), p(x))\,dx$$

is lower semicontinuous in $W^{1,1}(\Omega)$ relative to strong convergence in $L^1(\Omega)$ for every $p \in (C_0^\infty(\Omega))^n$ such that $\|p\|_{(L^\infty(\Omega))^n} < K_0$. By (2.5) the same is true for the functional I. □

In what follows we use the following condition:

$$H(x, u, p) \leqq h(x, p) + M(p)|u|, \quad \text{for every } (x, u, p) \in \Omega \times \mathbb{R} \times \mathbb{R}^n, \tag{2.6}$$

where h is a real-valued positive function such that $h(\cdot, p) \in L^1(\Omega)$ for every $p \in \mathbb{R}^n$ and $h(x, \cdot)$ is convex (and so continuous) for every $x \in \Omega$, and M is a real-valued positive continuous function.

LEMMA 2.3. *Let* (2.6) *hold. Then* $-H$ *is a normal integrand.*

Proof. The result may be proved as in [12, Prop. 4]. □

LEMMA 2.4. *Let* (2.6) *hold. Then for every* $s \geqq 0$ and $p \in (L^\infty(\Omega))^n$ *such that* $\|p\|_{(L^\infty(\Omega))^n} \leqq s$ *there exist* $h_s \in L^1(\Omega)$, $h_s > 0$, *and* $M_s > 0$, *with* M_s *continuous relative to* s, *such that*

$$H(x, u, p(x)) \leqq h_s(x) + M_s|u|, \quad \textit{for every } (x, u) \in \Omega \times \mathbb{R}. \tag{2.7}$$

Proof. We argue as in [3, Prop. 2.2]. Let $h_s(x) = \max\{h(x, p): |p| \leqq s\}$. By the continuity of $h(x, \cdot)$ there exists p_s such that $|p_s| \leqq s$ and $h_s(x) = h(x, p_s)$. Then there exist $p_1^{(s)}, \cdots, p_\nu^{(s)}$ such that

$$p_s \in \{p: |p| \leqq s\} \subset \operatorname{co}\{p_1^{(s)}, \cdots, p_\nu^{(s)}\};$$

so we obtain, for suitable $\lambda_i^{(s)} \in \mathbb{R}$, where $0 \leqq \lambda_i^{(s)} \leqq 1$ and $\sum_{i=1}^{\nu} \lambda_i^{(s)} = 1$,

$$h_s(x) = h\Big(x, \sum_{i=1}^{\nu} \lambda_i^{(s)} p_i^{(s)}\Big) \leqq \sum_{i=1}^{\nu} \lambda_i^{(s)} h(x, p_i^{(s)}),$$

and h_s is summable by (2.6). Finally we define $M_s = \max\{M(p): |p| \leqq s\}$. □

We state our second semicontinuity result, which improves [3, Thm. 3.1]:

THEOREM 2.2. *Let* $L(x, u, \cdot)$ *be convex for every* $(x, u) \in \Omega \times \mathbb{R}$ *and let* (2.6) *hold. Then* I *is lower semicontinuous in* $W^{1,1}(\Omega)$ *relative to strong convergence in* $L^1(\Omega)$.

Proof. As in the proof of Theorem 2.1 we have by (2.6) and [13, Thm. 3C]

$$I(u) = \sup\left\{\int_\Omega p(x)\nabla u(x)\,dx - \int_\Omega H(x, u(x), p(x))\,dx : p \in (L^\infty(\Omega))^n\right\}.$$

Now let $p \in (L^\infty(\Omega))^n$ and $p_h \in (C_0^\infty(\Omega))^n$ be such that $\|p_h\|_{(L^\infty(\Omega))^n} \leqq \|p\|_{(L^\infty(\Omega))^n}$ and $\lim_{h\to 0} p_h(x) = p(x)$

a.e. in Ω. We have, by Lemma 2.4,

$$H(x, u(x), p_h(x)) \leqq h_s(x) + M_s|u(x)|,$$

where $s = \|p\|_{(L^\infty(\Omega))^n}$.

Hence by Fatou's lemma and the continuity of $H(x, u(x), \cdot)$ (see Lemma 2.3) we have

$$\limsup_{h\to 0} \int_\Omega H(x, u(x), p_h(x))\,dx \leqq \int_\Omega H(x, u(x), p(x))\,dx,$$

and so

$$I(u)=\sup\left\{\int_\Omega p(x)\nabla u(x)\,dx-\int_\Omega H(x,u(x),p(x));\ p\in(C_0^\infty(\Omega))^n\right\}. \tag{2.8}$$

Now, as in the proof of Theorem 2.1, it may be proved that the functional

$$u\to\int_\Omega p(x)\nabla u(x)\,dx-\int_\Omega H(x,u(x),p(x))\,dx,$$

is lower semicontinuous in $W^{1,1}(\Omega)$ relative to strong convergence in $L^1(\Omega)$ for every $p\in(C_0^\infty(\Omega))^n$, (in this case there is no restriction on p); the proof is complete by (2.8). □

Finally we obtain the following statements.

COROLLARY 2.1. *Let the hypotheses of Theorem* 2.1 *hold and suppose there exists* $g\in L^1(\Omega)$ *such that* $L(x,u,v)\geqq -g(x)$ *for every* $(x,u,v)\in\Omega\times\mathbb{R}\times\mathbb{R}^n$; *then* I *is lower semicontinuous in* $W^{1,1}(\Omega)$ *relative to local convergence in* $L^1(\Omega)$.

Proof. We have $L+g\geqq 0$; then

$$\int_\Omega L(x,u(x),\nabla u(x))\,dx+\int_\Omega g(x)\,dx=\sup\left\{\int_{\Omega_h}(L(x,u(x),\nabla u(x))+g(x))\,dx:\ h>0\right\},$$

where $\Omega_h=\{x\in\Omega:\ d(x,\partial\Omega)>h\}$. For every h the functional

$$u\to\int_{\Omega_h}L(x,u(x),\nabla u(x))\,dx$$

is lower semicontinuous relative to local convergence in $L^1(\Omega)$ (since by Theorem 2.1 it is lower semicontinuous relative to strong convergence in $L^1(\Omega_h)$) and so I is lower semicontinuous relative to local convergence in $L^1(\Omega)$. □

COROLLARY 2.2. *Let the hypotheses of Theorem* 2.2 *hold and suppose there exists* $g\in L^1(\Omega)$ *such that* $L(x,u,v)\geqq -g(x)$ *for every* $(x,u,v)\in\Omega\times\mathbb{R}\times\mathbb{R}^n$; *then* I *is lower semicontinuous in* $W^{1,1}(\Omega)$ *relative to local convergence in* $L^1(\Omega)$.

Proof. We obtain the result as in the proof of Corollary 2.1 (obviously, we use Theorem 2.2 instead of Theorem 2.1). □

3. Optimization and regularizing extensions of integral functionals. In this section we use the results of § 2 to derive some properties of the regularizing extensions (in the sense of [5]) of I, defined in [1].

We recall some definitions. Let $C_0(\Omega)$ be the space of all continuous functions whose support is compact in Ω, endowed with the uniform convergence topology, and let $\bar{C}_0(\Omega)$ be its closure; then $M_n(\Omega)=((\bar{C}_0(\Omega))^*)^n$ is the Banach space of all Radon n-dimensional measures in Ω whose total variation is finite in Ω. We consider the space

$$BV(\Omega)=\{u\in L^1(\Omega);\ \nabla u\in M_n(\Omega)\},$$

which is a Banach space if we put

$$\|u\|_{BV(\Omega)}=\|u\|_{L^1(\Omega)}+\|\nabla u\|_{M_n(\Omega)}. \tag{3.1}$$

We recall some results given in [6], [7], [8]. Let ν be the unit outer normal to $\partial\Omega$, and H_{n-1} be the $(n-1)$-dimensional Hausdorff measure in $\mathbb{R}^n$; if $u\in BV(\Omega')$, where $\bar{\Omega}\subset\Omega'$, there exist $\gamma^-(u)$ and $\gamma^+(u)\in L^1(\partial\Omega)$ such that

$$\int_{\bar{\Omega}}G\nabla u+\int_\Omega u\ \mathrm{div}\ G=\int_{\partial\Omega}\gamma^+(u)G\nu\,dH_{n-1}\quad\text{for every } G\in(C_0^1(\Omega'))^n,$$

and

$$\int_\Omega G\nabla u + \int_\Omega u \operatorname{div} G = \int_{\partial\Omega} \gamma^-(u) G\nu \, dH_{n-1} \quad \text{for every } G \in (C_0^1(\Omega'))^n.$$

$\gamma^+(u)$ and $\gamma^-(u)$ are called, respectively, the outer trace and the inner trace of u on $\partial\Omega$. If given merely $u \in BV(\Omega)$ we may deal on $\partial\Omega$ only with $\gamma^-(u)$; if $u \in W^{1,1}(\Omega)$ we have $\gamma^-(u) = \gamma(u)$, where $\gamma(u)$ is the trace of u on $\partial\Omega$ in the sense of Sobolev spaces. We remark that

$$\left|\int_\Omega u\right| + \|\nabla u\|_{M_n(\Omega)}$$

and

$$\|\gamma^-(u)\|_{L^1(\partial\Omega)} + \|\nabla u\|_{M_n(\Omega)}$$

are norms in $BV(\Omega)$ equivalent to (3.1) (e.g., see [4]). We defined in [1] a topology on $BV(\Omega)$ which we called the w_q^* topology. We proved the following statements about the w_q^* topology:

(A) The w_q^* topology is metrizable on the balls of $BV(\Omega)$ (see [1]).

(B) The balls of $BV(\Omega)$ are w_q^*-compact (see [1]).

(C) A sequence $\{u_m\}$, w_q^*-converges to u if and only if $\int_\Omega u_m \to \int_\Omega u$ and $\int_\Omega G\,\nabla u_m \to \int_\Omega G\,\nabla u$ for every $G \in (\bar{C}_0(\Omega))^n$ (see [2, Thm. 1.3]).

(D) If a sequence $\{u_m\}$ w_q^*-converges to u, then $\lim_{m\to+\infty} u_m = u$ in $L^1(\Omega)$ (see [2, Thm. 1.2]).

(E) $W^{1,1}(\Omega)$ is w_q^*-dense in $BV(\Omega)$ (see [1, Prop. 1.2].)

A more detailed approach to the w_q^* topology may be found in [1] and [2]. By (E) we may define (see [1, (3.3)])

$$J_L(u) = \min\{\liminf_\alpha I(u_\alpha) : \{u_\alpha\} \text{ is a net, } u_\alpha \xrightarrow{w_q^*} u\}, \qquad u \in BV(\Omega). \tag{3.2}$$

Now we consider $L^1(\partial\Omega)$ as a space of n-valued measures on $\partial\Omega$ in the following way:

$$\langle f, G\rangle = \int_{\partial\Omega} fG\nu \, dH_{n-1} \quad \text{for every } f \in L^1(\partial\Omega) \text{ and } G \in (C(\partial\Omega))^n.$$

So, $L^1(\partial\Omega)$ may be endowed with the induced weak topology of dual space which we shall call w_2^* topology (see [1, Appendix]). We remark that a net $\{f_\alpha\} \subset L^1(\partial\Omega)$ w_2^*-converges to f if and only if

$$\int_{\partial\Omega} f_\alpha G\nu \, dH_{n-1} \to \int_{\partial\Omega} fG\nu \, dH_{n-1} \quad \text{for every } G \in (C(\partial\Omega))^n.$$

The balls of $((C(\partial\Omega))^*)^n$ are w_2^*-compact and their induced w_2^* topology is metrizable.

We consider on $BV(\Omega)\oplus L^1(\partial\Omega)$ the $w_q^* \times w_2^*$ topology. If $u \in W^{1,1}(\Omega)$ we have $(u, \gamma(u)) \in BV(\Omega)\oplus L^1(\partial\Omega)$; in this sense we have $W^{1,1}(\Omega) \subset BV(\Omega)\oplus L^1(\partial\Omega)$. We proved:

(F) $W^{1,1}(\Omega)$ is $w_q^* \times w_2^*$-dense in $BV(\Omega)\oplus L^1(\partial\Omega)$ (see [1, Lemma 1.3]).

By (F), we may define (see [1, Appendix])

$$J_2(u, f) = \min\{\liminf_\alpha I(u_\alpha) : \{u_\alpha\} \text{ is a net,}$$
$$(u_\alpha, \gamma(u_\alpha)) \xrightarrow{w_q^* \times w_2^*} (u, f)\}, \qquad (u, f) \in BV(\Omega)\oplus L^1(\partial\Omega). \tag{3.3}$$

The following inequalities follow from (3.2), (3.3):

$$J_L(u) \leqq J_2(u, \gamma^-(u)) \leqq I(u) \quad \text{for every } u \in W^{1,1}(\Omega), \tag{3.4}$$

$$J_L(u) \leqq J_2(u, f) \quad \text{for every } (u, f) \in BV(\Omega) \oplus L^1(\partial\Omega). \tag{3.5}$$

Remark 3.1. By general topological arguments we have

(i) $J_L(u) = I(u)$ for every $u \in W^{1,1}(\Omega)$ if and only if I is w_q^*-lower semicontinuous;

(ii) $J_2(u, \gamma(u)) = I(u)$ for every $u \in W^{1,1}(\Omega)$ if and only if I is $w_q^* \times w_2^*$-lower semicontinuous.

LEMMA 3.1. *Let* $\{u_\alpha\} \subset BV(\Omega)$ *be a net and* $u \in BV(\Omega)$ *such that* $u_\alpha \overset{w_q^*}{\to} u$; *if there exists* $c > 0$ *such that* $\|\nabla u_\alpha\|_{M_n(\Omega)} \leqq c$, *then there exists a subnet* $\{u_\gamma\}$ *of* $\{u_\alpha\}$ *such that* $\lim_\gamma u_\gamma = u$ *in* $L^1(\Omega)$.

Proof. By [1, Remark 1.1] there exists a subnet $\{u_\gamma\}$ of $\{u_\alpha\}$ such that $\int_\Omega u_\gamma \to \int_\Omega u$ and $\int_\Omega G\nabla u_\gamma \to \int_\Omega G\nabla u$ for every $G \in (\bar{C}_0(\Omega))^n$.

Let $u'_\gamma = u_\gamma - (\text{mes } \Omega)^{-1} \int_\Omega u_\gamma$ and $u' = u - (\text{mes } \Omega)^{-1} \int_\Omega u$. We have $\int_\Omega u'_\gamma = \int_\Omega u' = 0$ and so, by previous remarks on equivalent norms in $BV(\Omega)$ and the hypothesis, we have $\|u'_\gamma\|_{BV(\Omega)} \leqq c'$ for a suitable constant $c' > 0$. Then $\{u'_\gamma\}$ is relatively compact in $L^1(\Omega)$; therefore if $\{u'_{\gamma_1}\}$ is a subnet of $\{u'_\gamma\}$ there exist a subnet $\{u'_{\gamma_2}\}$ of $\{u'_{\gamma_1}\}$ and $u'_2 \in L^1(\Omega)$ such that $\lim_{\gamma_2} u'_{\gamma_2} = u'_2$ in $L^1(\Omega)$. We have also

$$\int_\Omega G\nabla u'_{\gamma_2} \to \langle G, \nabla u'_2 \rangle \quad \text{for every } G \in (C_0^\infty(\Omega))^n,$$

where $\langle \cdot, \cdot \rangle$ denotes the duality between $(C_0^\infty(\Omega))^n$ and the space of n-valued distributions. Since $\int_\Omega G\nabla u'_{\gamma_2} = \int_\Omega G\nabla u_{\gamma_2} \to \int_\Omega G\nabla u = \int_\Omega G\nabla u'$, we obtain $u'_2 = u'$ and the whole net $\{u'_\gamma\}$ converges to u' in $L^1(\Omega)$. Now we may write

$$\|u_\gamma - u\|_{L^1(\Omega)} \leqq \|u'_\gamma - u'\|_{L^1(\Omega)} + (\text{mes } \Omega)^{-1} \left| \int_\Omega u_\gamma - \int_\Omega u \right|,$$

and this completes the proof. □

The following is easily proved.

LEMMA 3.2. *Let* $L(x, u, \cdot)$ *be convex for every* $(x, u) \in \Omega \times \mathbb{R}$; *let* $K > 0$ *and* $\theta \in L^1(\Omega)$.

Then we have,

$$L(x, u, v) \geqq K|v| - \theta(x) \quad \textit{for every } (x, u, v) \in \Omega \times \mathbb{R} \times \mathbb{R}^n, \tag{3.6}$$

if and only if

$$H(x, u, p) \leqq \theta(x), \quad \textit{for every } (x, u) \in \Omega \times \mathbb{R} \textit{ and } |p| \leqq K. \tag{3.7}$$

THEOREM 3.1. *Let* (3.6) *hold. Then*

$$J_L(u) = \min \{ \lim_{m \to +\infty} \inf I(u_m) : \{u_m\} \textit{ is a sequence, } u_m \xrightarrow{w_q^*} u \}, \qquad u \in BV(\Omega), \tag{3.8}$$

$$J_2(u, f) = \min \{ \lim_{m \to +\infty} \inf I(u_m) : \{u_m\} \textit{ is a sequence, } (u_m, \gamma(u_m)) \xrightarrow{w_q^* \times w_2^*} (u, f) \}, \quad (u, f) \in BV(\Omega) \oplus L^1(\partial\Omega), \tag{3.9}$$

$$J_2(u, f) \geqq J_L(u) > -\infty, \quad \textit{for every } (u, f) \in BV(\Omega) \oplus L^1(\partial\Omega). \tag{3.10}$$

Proof. If $J_L(u) = +\infty$, then (3.8) holds; otherwise let $\{u_\alpha\}$ be such that $u_\alpha \xrightarrow{w_q^*} u$ and $J_L(u) = \lim_\alpha I(u_\alpha)$. Then there exists $z \in \mathbb{R}$ such that $I_L(u_\alpha) \leqq z$, for $\alpha > \alpha_z$, where α_z is a suitable index. By (3.6) we have $\|\nabla u_\alpha\| \leqq c_z$ for a suitable constant c_z. Then by Lemma 3.1 there exists a subnet $\{u_\beta\}$ of $\{u_\alpha\}$ such that $\lim_\beta u_\beta = u$ in $L^1(\Omega)$ and

$$\int_\Omega |u_\beta| \leqq \int_\Omega |u_\beta - u| + \int_\Omega |u| \leqq 1 + \int_\Omega |u| \quad \text{for } \beta > \beta_0,$$

where β_0 is a suitable index. Therefore, for $\beta > \alpha_z$ and $\beta > \beta_0$ we have

$$\|u_\beta\|_{W^{1,1}(\Omega)} = \|u_\beta\|_{BV(\Omega)} \leqq \text{constant}.$$

So, the value $J_L(u)$ may be calculated by considering only the elements of a ball in $W^{1,1}(\Omega)$; we have (3.8), since the w_q^* topology is metrizable on the balls.

We may prove (3.9) in an analogous way (it suffices to remember that $\|\gamma(u)\|_{L^1(\partial\Omega)} \leqq c\|u\|_{W^{1,1}(\Omega)}$ for a suitable constant c independent of u). Now we prove (3.10): we have

$$-\infty < K\int_\Omega |\nabla u| - \int_\Omega \theta \leqq I(u), \quad \text{for every } u \in W^{1,1}(\Omega), \tag{3.11}$$

and (3.10) follows by (3.8), (3.9), (3.11) since the functional

$$u \to \int_\Omega |\nabla u|$$

is w_q^*-lower semicontinuous in $BV(\Omega)$. □

THEOREM 3.2. *Let the hypotheses of Theorem* 2.1 *hold; then I is sequentially w_q^*-lower semicontinuous (and also sequentially $w_q^* \times w_2^*$-lower semicontinuous). If, also,* (3.6) *holds, then I is w_q^*-lower semicontinuous (and also $w_q^* \times w_2^*$-lower semicontinuous) and*

$$J_L(u) = J_2(u, \gamma^-(u)) = I(u), \qquad u \in W^{1,1}(\Omega). \tag{3.12}$$

Proof. The first part of the theorem follows by Theorem 2.1 and the property (D) of the w_q^* topology (i.e., that the w_q^* topology is sequentially stronger than the norm topology in $L^1(\Omega)$ and the $w_q^* \times w_2^*$ topology is obviously stronger than the w_q^* topology). Afterwards, if (3.6) holds, by Theorem 3.1 and the sequential w_q^*-lower semicontinuity of I, we obtain $I(u) \leqq J_L(u)$ for every $u \in W^{1,1}(\Omega)$ and we obtain (3.12) by (3.4). The lower semicontinuity result about I follows by Remark 3.1. □

The following theorem improves [3, Thms. 3.2, 3.4].

THEOREM 3.3. *Let the hypotheses of Theorem* 2.2 *hold. Then I is sequentially w_q^*-lower semicontinuous (and also sequentially $w_q^* \times w_2^*$-lower semicontinuous); moreover, if $M = M(p)$ is a positive constant, the sets*

$$A_{z,s}(I) = \{u : I(u) \leqq z\} \cap \{u : \|u\|_{L^1(\Omega)} \leqq s\}$$

are compact relative to the weak topology of $W^{1,1}(\Omega)$ and I has an absolute minimum on every weakly closed subset of $W^{1,1}(\Omega)$ which is bounded relative to the norm of $L^1(\Omega)$.

Proof. The semicontinuity result follows by Theorem 2.2 and the property (D) of the w_q^* topology. Now, we prove the compactness of $A_{z,s}(I)$. If $u \in A_{z,s}(I)$ we have, by [13, Thm. 3C] applied to the normal integrands $h(x,p) + M|u(x)|$ and $h(x,p) +$

$M\ (\text{mes}\ \Omega)^{-1}s$,

$$
\begin{aligned}
z \geqq I(u) &\geqq \int_\Omega \sup\{p\nabla u(x) - h(x,p) - M|u(x)|\colon p \in \mathbb{R}^n\}\,dx \\
&= \sup\left\{\int_\Omega (p(x)\nabla u(x) - h(x,p(x)) - M|u(x)|)\,dx\colon p \in (L^\infty(\Omega))^n\right\} \\
&\geqq \sup\left\{\int_\Omega (p(x)\nabla u(x) - h(x,p(x)) - M(\text{mes}\ \Omega)^{-1}s)\,dx\colon p \in (L^\infty(\Omega))^n\right\} \\
&= \int_\Omega \sup\{p\nabla u(x) - h(x,p) - M(\text{mes}\ \Omega)^{-1}s\colon p \in \mathbb{R}^n\}\,dx \\
&= \int_\Omega \phi(x, \nabla u(x))\,dx,
\end{aligned}
$$

where $\phi(x,v) = \sup\{pv - h(x,p) - M(\text{mes}\ \Omega)^{-1}s\colon p \in \mathbb{R}^n\}$. Hence, the set

$$
B_{z,s}(\phi) = \{u \in W^{1,1}(\Omega)\colon \int_\Omega \phi(x, \nabla u(x))\,dx \leqq z, \|u\|_{L^1(\Omega)} \leqq s\},
$$

which is weakly compact in $W^{1,1}(\Omega)$ by [11, Corollary 2B], contains $A_{z,s}(I)$, which is weakly sequentially closed by the first part of the theorem (and also closed since the weak topology is metrizable on compact sets). So $A_{z,s}(I)$ is weakly compact.

The last part of the theorem follows easily by the lower semicontinuity of I and the compactness of $A_{z,s}(I)$. □

The following theorem gives new information about the extended functionals J_L and J_2.

THEOREM 3.4. *Let the hypotheses of Theorem* 2.2 *and* (3.6) *hold. Then I is w_q^*-lower semicontinuous (and also $w_q^* \times w_2^*$-lower semicontinuous) and*

$$
J_L(u) = \begin{cases} I(u) & \text{if } u \in W^{1,1}(\Omega), \\ +\infty & \textit{otherwise}, \end{cases} \tag{3.13}
$$

$$
J_2(u,f) = \begin{cases} I(u) & \text{if } u \in W^{1,1}(\Omega) \textit{ and } f = \gamma^-(u), \\ +\infty & \text{otherwise.} \end{cases} \tag{3.14}
$$

Proof. By Theorem 3.3 and Theorem 3.1 we obtain the w_q^*-lower semicontinuity of I as in the proof of Theorem 3.2. Now we must prove (3.13); let

$$
\tilde{J}_L(u) = \begin{cases} I(u) & \text{if } u \in W^{1,1}(\Omega), \\ +\infty & \text{otherwise.} \end{cases}
$$

$\tilde{J}_L$ is w_q^*-lower semicontinuous if and only if (3.13) holds. Then let $u_\alpha \xrightarrow{w_q^*} u$; if $\lim_\alpha \inf \tilde{J}_L(u_\alpha) = +\infty$ we have, obviously, $\tilde{J}_L(u) \leqq \lim_\alpha \inf \tilde{J}_L(u_\alpha)$; otherwise, if $\lim_\alpha \inf \tilde{J}_L(u_\alpha) < +\infty$, let $\{u_\gamma\} \subset \{u_\alpha\}$ be a subnet such that $\lim_\alpha \inf \tilde{J}_L(u_\alpha) = \lim_\gamma \tilde{J}_L(u_\gamma) < +\infty$. Then there exist suitable $z \in \mathbb{R}$ and $\gamma(z)$ such that $\tilde{J}_L(u_\gamma) < z$ if $\gamma > \gamma(z)$; by the definition of $\tilde{J}_L$, we have $\tilde{J}_L(u_\gamma) = I(u_\gamma) < z$ and $u_\gamma \in W^{1,1}(\Omega)$, if $\gamma > \gamma(z)$. By (3.6), we have $\|\nabla u_\gamma\|_{M_n(\Omega)} \leqq$ constant and, as in the proof of Theorem 3.1, $\|u_\gamma\|_{W^{1,1}(\Omega)} \leqq$ constant. Hence by Theorem 3.3 $\{u_\gamma\}$ is contained in a weakly compact subset of $W^{1,1}(\Omega)$ and so there exists a subnet $\{u_\delta\} \subset \{u_\gamma\}$, and $u_0 \in W^{1,1}(\Omega)$ such that $\lim_\delta u_\delta = u_0$, relative to the weak topology of $W^{1,1}(\Omega)$; we have also $u = u_0$. Finally we

have

$$\tilde{J}_L(u) = I(u) \leqq \liminf_{\delta} I(u_\delta) = \lim_{\delta} \tilde{J}_L(u_\delta) = \liminf_{\alpha} \tilde{J}_L(u_\alpha),$$

since we proved that I is w_q^*-lower semicontinuous and the w_q^* topology on the balls is weaker then the weak topology of $W^{1,1}(\Omega)$; so we obtain (3.13). By (3.4) we have also $J_2(u, \gamma(u)) = I(u)$ if $u \in W^{1,1}(\Omega)$ (and so we have a part of (3.14)). We obtain completely (3.14) if we prove that the functional

$$\tilde{J}_2(u, f) = \begin{cases} I(u) & \text{if } u \in W^{1,1}(\Omega) \quad \text{and} \quad f = \gamma(u), \\ +\infty & \text{otherwise,} \end{cases}$$

is $w_q^* \times w_2^*$-lower semicontinuous.

Let $(u_\alpha, f_\alpha) \xrightarrow{w_q^* \times w_2^*} (u, f)$; as in the proof of (3.13), if $\lim_\alpha \inf \tilde{J}_2(u_\alpha, f_\alpha) < +\infty$ there exists a subnet $\{(u_\delta, f_\delta)\} \subset \{(u_\alpha, f_\alpha)\}$ such that

$$\lim_{\delta} \tilde{J}_2(u_\delta, f_\delta) = \liminf_{\alpha} \tilde{J}_2(u_\alpha, f_\alpha), \qquad u_\delta \in W^{1,1}(\Omega), \qquad f_\delta = \gamma(u_\delta)$$

and $\lim_\delta u_\delta = u$ relative to the weak topology of $W^{1,1}(\Omega)$ and (by Lemma 3.1) relative to the norm topology of $L^1(\Omega)$. Moreover we have for every $G \in (C^1(\mathbb{R}^n))^n$,

$$\int_{\partial\Omega} f_\delta G\nu \, dH_{n-1} \to \int_{\partial\Omega} fG\nu \, dH_{n-1},$$

and

$$\int_{\partial\Omega} f_\delta G\nu \, dH_{n-1} = \int_{\partial\Omega} \gamma(u_\delta) G\nu \, dH_{n-1} = \int_\Omega G\nabla u_\delta + \int_\Omega u_\delta \operatorname{div} G \to \int_\Omega G\nabla u + \int_\Omega u \operatorname{div} G$$

$$= \int_{\partial\Omega} \gamma(u) G\nu \, dH_{n-1};$$

hence, we have $f = \gamma(u)$ and

$$\tilde{J}_2(u, f) = \tilde{J}_2(u, \gamma(u)) = I(u) \leqq \liminf_{\delta} I(u_\delta) = \lim_{\delta} \tilde{J}_2(u_\delta, \gamma(u_\delta)) = \liminf_{\alpha} \tilde{J}_2(u_\alpha, f_\alpha). \qquad \square$$

We observe by an example the difference between Theorem 3.2 and Theorem 3.4 (i.e., between (3.12) and, (3.13), (3.14)): the integrand $L(x, u, v) = (1 + |v|^2)^{1/2}$ satisfies the hypotheses of Theorem 3.2, then (3.12) holds; moreover we proved in [1] that in this case $J_2(u, f) = \int_\Omega (1 + |\nabla u|^2)^{1/2} + \int_{\partial\Omega} |f - \gamma^-(u)| \, dH_{n-1}$ and so (3.13) and (3.14) do not hold.

REFERENCES

[1] F. FERRO, *Variational functionals defined on spaces of BV functions and their dependence on boundary data*, Ann. Mat. Pura. Appl., 122 (1979), pp. 269–287.

[2] ———, *Integral characterization of functionals defined on spaces of BV functions*, Rend. Sem. Mat. Univ. Padova, 61 (1979), pp. 177–201.

[3] ———, *Optimization theorems in the n-dimensional calculus of variations*, Boll. Un. Mat. Ital., 16-B (1979), pp. 255–265.

[4] ———, *A minimal surface theorem with generalized boundary data*, Boll. Un. Mat. Ital., 16-B (1979), pp. 962–971.

[5] A. D. IOFFE AND V. W. TIHOMIROV, *Extension of variational problems*, Trans. Moscow Math. Soc., 18 (1968), pp. 207–273.
[6] M. MIRANDA, *Distribuzioni aventi derivate misure. Insiemi di perimetro localmente finito*, Ann. Scuola Norm. Sup. Pisa Sci. Fis. Mat., serie III, 19 (1964), pp. 27–56.
[7] ———, *Un teorema di esistenza e unicità per il problema dell'area minima in n variabili*, Ann. Scuola Norm. Sup. Pisa Sci. Fis. Mat., serie III, 19 (1965), pp. 233–249.
[8] ———, *Comportamento delle successioni convergenti di frontiere minimali*, Rend. Sem. Mat. Univ. Padova, 38 (1967), pp. 238–257.
[9] C. B. MORREY, *Multiple Integrals in the Calculus of Variations*, Springer-Verlag, Berlin, Heidelberg, New York, 1966.
[10] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton Univ. Press, Princeton, NJ, 1970.
[11] ———, *Integrals which are convex functionals II*, Pacific J. Math., 39 (1971), pp. 439–469.
[12] ———, *Existence theorems for general control problems of Bolza and Lagrange*, Advances Math., 15 (1975), pp. 312–333.
[13] ———, *Integral functionals, normal integrands and measurable selections*, in Nonlinear Operators and the Calculus of Variations, L. Waelbroeck, ed., Lectures Notes in Math. 543, Springer-Verlag, Berlin, Heidelberg, New York, 1976.
[14] J. SERRIN, *On the definition and properties of certain variational integrals*, Trans. Amer. Math. Soc., 101 (1961), pp. 139–167.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 4, July 1981

0363-0129/81/1904-0002 $01.00/0

CAUSAL FACTORIZATION AND LINEAR FEEDBACK*

JACOB HAMMER† AND MICHAEL HEYMANN†

Abstract. An algebraic framework for the investigation of linear dynamic output feedback is introduced. Pivotal in the present theory is the problem of causal factorization, i.e. the problem of factoring two systems over each other through a causal factor. The basic issues are resolved with the aid of the new concept of latency kernels.

1. Introduction. In recent years the system theory literature has seen a rapidly growing interest in questions associated with linear feedback. In the early 1960's, linear control theory centered chiefly around quadratic (Gaussian) optimal problems and the resulting feedback designs. Later, interest in feedback shifted to a variety of so-called "synthesis" problems. These included the well-known problem of observer design (see Luenberger [1966]), the pole shifting theorem and related issues (Wonham [1967], Simon and Mitter [1968], Brash and Pearson [1970], Heymann [1968]) as well as the decoupling problem (Falb and Wolovich [1967], Gilbert [1969], Wonham and Morse [1970], Morse and Wonham [1970]). All of these feedback synthesis problems, as well as many others, were formulated and resolved within the framework of state space representations. While most of the work was done with the use of conventional state equations, the work of Wonham and Morse was distinguished by its "coordinate free" setting and initiated what later developed into the celebrated "geometric theory" of linear control (see, e.g., Wonham [1979]).

The current growing interest in linear feedback differs significantly from that of the past both in character and in its source of motivation. While previously the study of feedback was largely oriented at problem solving, the current interest is motivated by a desire of gaining insight into the general nature of linear feedback—chiefly from an algebraic point of view. Much of the motivation for the present trend can be traced back to the work of Rosenbrock [1970], in which polynomial matrix techniques were used for the study of a variety of (linear) control theoretic questions. Particularly useful turned out to be techniques based on polynomial fraction representations of transfer functions (see, e.g., Heymann [1972], Wolovich [1974], Forney [1975], Fuhrmann [1976]). In this setting of fraction representations, feedback was first studied in Heymann [1972] (see especially Chapter 6 therein), and in a polynomial module framework the study of feedback was initiated by Eckberg [1974]. State feedback also received attention in an algebraic framework by Morse [1975]. A different approach to the study of linear feedback was taken in Hautus and Heymann [1978], where the fundamental underlying object was taken to be the input–output map of the system. There, static linear state feedback was investigated in an algebraic framework consistent with the setting of the (classical) module theory of linear realization as introduced by Kalman (see, e.g., Kalman et al. [1969, Chapter 10]). More recently, state feedback was also examined in Fuhrmann [1979] using what he termed "polynomial models", and in Münzner and Prätzel-Wolters (1979a], [1979b], [1979c] in a module and category theoretic framework.

While these various approaches to the study of feedback differ from each other substantially both in the underlying concept and in philosophy, they commonly converge on essentially the same (standard) issues that characterize state feedback. It is

* Received by the editors January 2, 1980.
† Department of Electrical Engineering, Technion–Israel Institute of Technology, Haifa, Israel.

significant, however, that no success (and, in fact, very little effort, if any) has been reported in respect to *output*, as opposed to *state* feedback. When various fundamental questions in regard to output feedback are examined, it becomes immediately clear that difficulties arise that are completely absent in the state-feedback setting. In fact, one discovers immediately that crucial insight is missing. It turns out that the chief reason for this state of affairs is the fact that all of the presently existing algebraic theory of linear systems, and especially that of feedback, rests in one way or another on the theory of modules over the ring $K[z]$ of polynomials and on polynomial matrices. This algebraic machinery is completely satisfactory to develop a fairly comprehensive framework for state feedback. It is not adequate, though, to deal with output-feedback where issues associated with causality become significantly more intricate.

The present paper deals in a comprehensive way with the problem of causal output feedback. A related question which receives a great deal of attention in the paper and on which much of the theory hinges is the so-called causal factorization problem. This is the problem of when a given linear input-output map can be factored over another one by a causal linear map. Through the resolution of this issue, questions associated with dynamic causal output feedback are then also resolved. Attention is also given to the static factorization problem as well as the problem of static feedback where special emphasis is placed on the state-feedback case.

A crucial role in the present theory is played by the newly introduced concept of *latency*. In the discrete time setting, latency expresses "degree of causality" and (intuitively) refers to the intrinsic delay which inputs encounter before output responses are produced. Latency is algebraically expressed by modules over the ring $K[[z^{-1}]]$ of power series (in z^{-1} over a field K). These modules arise in a natural way when the concept of causality is studied algebraically and in fact are readily seen to be the natural algebraic device for the study of feedback.

The paper is organized as follows. In § 2 the basic concepts of ΛK-linear maps, causality, linear i/o maps as well as linear i/s maps, which have been investigated in detail in Hautus and Heymann [1978], are reviewed. The conceptual viewpoint, on which the present investigation of feedback rests, is discussed in § 3. An important technical concept that arises in the algebraic study of linear systems both in connection with the $K[z]$-module theory and the $K[[z^{-1}]]$-module theory is that of "proper bases" and "proper independence". This is the topic of § 4. Section 5 is devoted to the investigation of causal factorization, the main result being Theorem 5.2 and its corollaries. Results are also obtained on static feedback (Theorems 5.10 and 5.14). In § 6 the problem of invariants is investigated in detail and explicit characterizations are derived and exhibited. The role of the latency kernels and latency indices is also discussed. The paper is concluded in § 7 with an investigation of the interesting question of feedback (design) limitations. It is shown that the essential limitation to the possibility of causal feedback implementation of precompensators is the system's latency. In particular, precompensators can be implemented as causal feedback devices modulo a "precompensator remainder" whose dynamic order need not exceed the sum of the system's latency indices.

2. ΛK-linear maps, causality and input-output behavior. We shall adopt a terminology and setup consistent with that of Hautus and Heymann [1978].

Let K be a field and let S be a K-linear space. The class of all truncated S-valued Laurent series of the form

$$s = \sum_{t=t_0}^{\infty} s_t z^{-t} \tag{2.1}$$

is denoted by $S((z^{-1}))$ or alternatively by ΛS. The polynomial subset of S, i.e., the set of all elements of ΛS of the form $\sum_{t\leqq 0} s_t z^{-t}$, is denoted $\Omega^+ S$. The power series subset of ΛS, i.e., the set of all elements of the form $\sum_{t\geqq 0} s_t z^{-t}$, is denoted $\Omega^- S$. The set $\Lambda K = K((z^{-1}))$ of K-valued Laurent series is endowed with a field structure under the operation of convolution as multiplication and coefficientwise addition. In particular, for $\alpha = \sum_{t=t_0}^{\infty} \alpha_t z^{-t}$ and $\alpha' = \sum_{t=t_0'}^{\infty} \alpha_t' z^{-t}$ in ΛK, the product $\alpha\alpha'$ is given by

$$\alpha\alpha' = \sum_{t=t_0+t_0'}^{\infty} \left[\sum_{j=t_0}^{t-t_0'} \alpha_t \alpha'_{t-j} \right] z^{-t}$$

and the sum $\alpha + \alpha'$ is given by

$$\alpha + \alpha' = \sum_{t=\min(t_0,\, t_0')}^{\infty} (\alpha_t + \alpha_t') z^{-t}.$$

With ΛK as the underlying field it then follows that, with convolution as the scalar multiplication and with the usual coefficientwise addition, the set ΛS becomes a ΛK-linear space. When S is a finite dimensional K-linear space, say of dimension n, then so is ΛS as a ΛK-linear space. It is readily observed that, under the same operations of convolution as multiplication and coefficientwise addition, the field ΛK contains (as subobjects) also (i) the ring $K[z]$, or in our notation $\Omega^+ K$, of polynomials in z; (ii) the ring $K[[z^{-1}]]$, or in our notation $\Omega^- K$, of formal power series in z^{-1}; and finally, (iii) the field K itself. It, thus, follows immediately that the set ΛS is not only a ΛK-linear space, but is simultaneously also an $\Omega^+ K$-module, an $\Omega^- K$-module and a K-linear space. As we shall see, these facts turn out to be of central importance in the theory.

Now, we let $\mathbb{Z}$ denote the integers and for an element $s \in \Lambda S$, given by (2.1), we define the *order* of s by

$$\text{ord } s := \begin{cases} \min\{t \in \mathbb{Z} \mid s_t \neq 0\} & \text{if } s \neq 0, \\ \infty & \text{if } s = 0. \end{cases} \tag{2.2}$$

If $s \neq 0$ and $t_0 = \text{ord } s$, we call the coefficient s_{t_0} the *leading coefficient* of s.

Let U and Y be K-linear spaces. We shall call U the *input value space* and Y the *output value space* of an underlying linear system Σ. The ΛK-linear spaces ΛU and ΛY are then called the *extended input space* and *extended output space*, respectively. Elements $u = \Sigma u_t z^{-t} \in \Lambda U$ and $y = \Sigma y_t z^{-t} \in \Lambda Y$, called, respectively, (extended) inputs and (extended) outputs, are identified with time sequences $\{u_t\}$ and $\{y_t\}$ (with t being identified as time marker).

Let $\bar{f}: \Lambda U \to \Lambda Y$ be a K-linear map. We say that $\bar{f}$ is *time invariant* if

$$\bar{f}(z \cdot u) = z \cdot \bar{f}(u)$$

for all $u \in \Lambda U$, so that $\bar{f}$ is time invariant whenever it is a ΛK-linear map (Wyman [1972]). Next, for a ΛK-linear map $\bar{f}: \Lambda U \to \Lambda Y$ we define the *order* of $\bar{f}$ by

$$\text{ord } \bar{f} := \inf\{\text{ord } \bar{f}(u) - \text{ord } u \mid 0 \neq u \in \Lambda U\}. \tag{2.3}$$

If the map $\bar{f}$ is the zero map then ord $\bar{f} := \infty$; otherwise ord $\bar{f} < \infty$. While it is possible that ord $\bar{f} = -\infty$ we shall not concern ourselves here with this case and confine our attention to maps of finite order. This is clearly always the case when U (and hence also ΛU) is finite dimensional.

A ΛK-linear map $\bar{f}: \Lambda U \to \Lambda Y$ is called *causal* if ord $\bar{f} \geqq 0$ and *strictly causal* if ord $\bar{f} > 0$. The map $\bar{f}$ is called *order consistent* if for each $0 \neq u \in \Lambda U$

$$\text{ord } \bar{f}(u) - \text{ord } u = \text{ord } \bar{f}.$$

Clearly, an invertible ΛK-linear map $\bar{l}: \Lambda S \to \Lambda S$ is order consistent if and only if ord $\bar{l}^{-1} = -\text{ord } \bar{l}$. A ΛK-linear map $\bar{f}$ is said to be *order preserving* (or *instantaneous*) if it is order consistent and ord $\bar{f} = 0$. An invertible order preserving (and hence causal) ΛK-linear map $\bar{l}: \Lambda S \to \Lambda S$ is called *a bicausal isomorphism* (or simply *bicausal*) since its inverse is then also causal. Finally, we call $\bar{f}$ *nonlatent* if it is order consistent and ord $\bar{f} = 1$.

We now introduce the following (see also Hautus and Heymann [1978]).

DEFINITION 2.4. A map $\bar{f}: \Lambda U \to \Lambda Y$ is called an *extended linear input-output map* (or *extended linear* i/o *map*) if it is strictly causal (i.e., ord $\bar{f} > 0$) and ΛK-linear.

Let L denote the K-linear space of K-linear maps $U \to Y$ and let ΛL denote the ΛK-linear space of all L-Laurent series. We identify this space with the space of ΛK-linear maps $\Lambda U \to \Lambda Y$ of finite order as follows. We define the K-linear maps

$$\begin{aligned} &\bar{\imath}_u : U \to \Lambda U : u \mapsto u \quad \text{(canonical injection)}, \\ &\bar{p}_k : \Lambda Y \to Y : \Sigma y_t z^{-t} \mapsto y_k. \end{aligned} \tag{2.5}$$

and with every ΛK-linear map $\bar{f}: \Lambda U \to \Lambda Y$ we associate the Laurent series

$$Z_{\bar{f}}(z^{-1}) := \Sigma A_t z^{-t}, \tag{2.6}$$

where, for each $k \in \mathbb{Z}$,

$$A_k := A_k(\bar{f}) := \bar{p}_k \cdot \bar{f} \cdot \bar{\imath}_u. \tag{2.7}$$

The Laurent series (2.6) is called the *impulse response* or the *transfer function* of $\bar{f}$. If $u = \Sigma u_t z^{-t} \in \Lambda U$ is any element, then the action of $\bar{f}$ on u is given by

$$\bar{f} \cdot u = (\Sigma A_t(\bar{f}) z^{-t}) \cdot (\Sigma u_t z^{-t}) = \sum_t \sum_k (A_k(\bar{f}) u_{t-k}) z^{-t}. \tag{2.8}$$

It is thus immediately seen that

$$\text{ord } \bar{f} = \min \{k \mid A_k(\bar{f}) \neq 0\}, \tag{2.9}$$

whence we have the following characterization of causality in terms of the transfer function: *The map $\bar{f}$ is causal if and only if $A_k(\bar{f}) = 0$ for $k < 0$ and strictly causal if and only if $A_k(\bar{f}) = 0$ for $k \leqq 0$.* We also have the following easily verified proposition.

PROPOSITION 2.10. *Let $\bar{f}: \Lambda U \to \Lambda Y$ be a ΛK-linear map of order k_0 ($< \infty$) and transfer function $Z_{\bar{f}}(z^{-1}) = \sum_{k=k_0}^{\infty} A_k z^{-k}$. Then $\bar{f}$ is order consistent if and only if A_{k_0} is injective (i.e.,* $\ker A_{k_0} = 0$).

The following is an immediate corollary to Proposition 2.10.

COROLLARY 2.11. *Let $\bar{l}: \Lambda S \to \Lambda S$ be a causal ΛK-linear map with transfer function $\sum_{k=0}^{\infty} A_k(\bar{l}) z^{-k}$. Then $\bar{l}$ is a bicausal isomorphism if and only if $A_0(\bar{l})$ is invertible, in which case $A_0(\bar{l}^{-1}) = (A_0(\bar{l}))^{-1}$.*

We associate with an extended linear i/o map $\bar{f}$ a *restricted linear i/o* map $\tilde{f}$ which is obtained as follows (see also Hautus and Heymann [1978]). Inputs are restricted to the subset $\Omega^+ U \subset \Lambda U$, called the *restricted input space*, and consist of all inputs that terminate at $t = 0$, i.e., elements of the form $\sum_{t \leqq 0} u_t z^{-t}$. Outputs are observed only for $t \geqq 1$, that is, in the subset $z^{-1}\Omega^- Y$ which is, of course, in bijective correspondence with the $\Omega^+ K$-quotient module $\Gamma^+ Y := \Lambda Y / \Omega^+ Y$ which we call the *restricted output space*. The *restricted linear* i/o map $\tilde{f}: \Omega^+ U \to \Gamma^+ Y$ associated with $\bar{f}$ is then defined by

$$\tilde{f} = \pi^+ \cdot \bar{f} \cdot j^+,$$

where $j^+: \Omega^+ U \to \Lambda U$ is the canonical injection and $\pi^+: \Lambda Y \to \Gamma^+ Y$ is the canonical projection. Clearly, since π^+ and j^+ are $\Omega^+ K$-module homomorphisms, so is also $\tilde{f}$ and

we have the following:

DEFINITION 2.12. A map $\tilde{f}: \Omega^+U \to \Gamma^+Y$ is called *a restricted linear* i/o *map* if it is an Ω^+K-module homomorphism.

Next, we define the *linear output response* (or *output value*) *map* $f: \Omega^+U \to Y$ associated with a given linear i/o map $\bar{f}$ (or $\tilde{f}$) as follows:

$$f: \Omega^+U \to Y: u \mapsto f(u) = \bar{p}_1 \cdot \bar{f}(u) = p_1 \cdot \tilde{f}(u), \tag{2.13}$$

where (identifying Γ^+Y with $z^{-1}\Omega^-Y$)

$$p_1: \Gamma^+Y \to Y: \sum_{t=1}^{\infty} y_t z^{-t} \mapsto y_1. \tag{2.14}$$

A linear i/o map $\bar{f}$ (or $\tilde{f}$) is called *reachable* if the associated output value map f is surjective.

If $f: \Omega^+U \to Y$ is any K-linear map, it can be regarded as an output value map of a linear system. In particular, the restricted and extended linear i/o maps associated with f are then given by

$$\tilde{f}(u) = \sum_{t \geqq 0} f(z^t u) z^{-t-1}, \qquad u \in \Omega^+U, \tag{2.15}$$

and

$$\bar{f}(u) = \sum_{t \in \mathbb{Z}} f(\mathscr{S}^+(z^t u)) z^{-t-1}, \qquad u \in \Lambda U, \tag{2.16}$$

where $\mathscr{S}^+: \Lambda U \to \Omega^+U: \Sigma\, u_t z^{-t} \mapsto \sum_{t \geqq 0} u_t z^{-t}$ is the *truncation operator*.

The relation between the maps $\bar{f}$, $\tilde{f}$ and f is summarized by the commutative diagram, Fig. 2.1, in which i denotes the identity map.

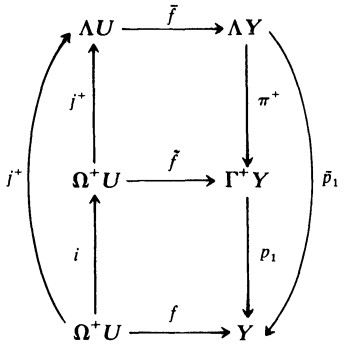


FIG. 2.1

The output value map f, which gives for each (restricted) input the value of the output at time $t = 1$, is clearly a K-linear map. In some special cases, there exists an Ω^+K-module structure on Y, compatible with its K-vector space structure, such that the output value map f is not just K-linear but is also an Ω^+K-module homomorphism. When this is the case, then for each $u \in \Omega^+U$ and for each positive integer k, $f(z^k u) = z^k f(u)$, whence, by (2.15), knowledge of the output value at time $t = 1$ implies knowledge of the whole ensuing output sequence. This is therefore precisely the case when the system's output "qualifies" as state, a fact which motivates the following definition (for greater detail the reader is referred to Hautus and Heymann [1978]):

DEFINITION 2.17. An extended linear i/o map $\bar{f}: \Lambda U \to \Lambda Y$ is called *an extended linear input-state* (or i/s) *map* if there exists an Ω^+K-module structure on Y, compatible

with its K-linear structure, such that the output value map $f = \bar{p}_1 \cdot \bar{f} \cdot j^+$ is an Ω^+K-homomorphism. The associated restricted map $\tilde{f}$ is called *a restricted linear* i/s *map.*

If Y and W are K-linear spaces and $H\colon Y \to W$ is a K-linear map, then it induces in a natural way a ΛK-linear map which we call *static* as follows:

$$H : \Lambda Y \to \Lambda W : \Sigma y_t z^{-t} \mapsto \Sigma (Hy_t) z^{-t}. \tag{2.18}$$

In a similar way H induces also static Ω^+K and Ω^-K-homomorphisms.

We shall need the following characterizations of linear i/s maps, from Hautus and Heymann [1978].

THEOREM 2.19. *If $\bar{f} : \Lambda U \to \Lambda Y$ is an extended linear* i/s *map then*

$$\ker f = \ker \tilde{f}. \tag{2.20}$$

THEOREM 2.21. *Let $\tilde{f} : \Lambda U \to \Lambda Y$ be a reachable extended linear* i/o *map. Then the following are equivalent:*

(i) *$\bar{f}$ is an extended reachable linear* i/s *map.*

(ii) *Condition* (2.20) *holds.*

(iii) *For every extended linear* i/o *map $\tilde{g}\colon \Lambda U \to \Lambda W$ satisfying $\ker \tilde{f} \subset \ker \tilde{g}$ (where $\tilde{f}$ and $\tilde{g}$ are the corresponding restricted* i/o *maps and where W is a K-linear space) there exists a unique static map $H : \Lambda Y \to \Lambda W$ such that $\bar{g} = H \cdot \bar{f}$.*

3. Feedback and causal factorization—general considerations. We shall be concerned with the setup described by the block diagram in Fig. 3.1.

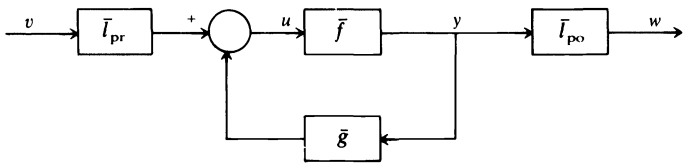


FIG. 3.1

Here $\bar{f}\colon \Lambda U \to \Lambda Y$ is an extended linear i/o map, called the *open loop system,* $\bar{g} : \Lambda Y \to \Lambda U$ is a causal ΛK-linear map called the (*output*) *feedback compensator,* $\bar{l}_{pr} : \Lambda U \to \Lambda U$ is a ΛK-linear bicausal isomorphism called (*bicausal*) *precompensator* and $\bar{l}_{po} : \Lambda Y \to \Lambda Y$ is a ΛK-linear bicausal isomorphism called (*bicausal*) *postcompensator.* In case any of the maps $\bar{g}$, $\bar{l}_{pr}$ or $\bar{l}_{po}$ is static we shall call it, respectively a *static* feedback, pre or post compensator.

Now, since the map $\bar{g}$ is causal and $\bar{f}$ is strictly causal, it readily follows that the composite maps $\bar{f} \cdot \bar{g} : \Lambda Y \to \Lambda Y$ and $\bar{g} \cdot \bar{f} : \Lambda U \to \Lambda U$ are both strictly causal. Letting I denote both of the corresponding identity maps, we see that both of the maps $(I + \bar{g}\bar{f}) : \Lambda U \to \Lambda U$ and $(I + \bar{f}\bar{g}) : \Lambda Y \to \Lambda Y$ are bicausal isomorphisms. It follows that the setup of Fig. 3.1 is "well-posed" in the sense that there is a strictly causal ΛK-linear map $\Lambda U \to \Lambda Y : v \mapsto w$ given by either of the following composite maps:

$$v \mapsto w = [\bar{l}_{po} \cdot \bar{f} \cdot (I + \bar{g}\bar{f})^{-1} \cdot \bar{l}_{pr}](v), \tag{3.1}$$

$$v \mapsto w = [\bar{l}_{po} \cdot (I + \bar{f}\bar{g})^{-1} \cdot \bar{f} \cdot \bar{l}_{pr}](v). \tag{3.2}$$

Using again block diagrams, (3.1) and (3.2) can be described, respectively, as in Fig. 3.2a and 3.2b.

In both descriptions, the dashed blocks represent bicausal mappings, so that the compensator configuration of Fig. 3.1 can always be represented equivalently

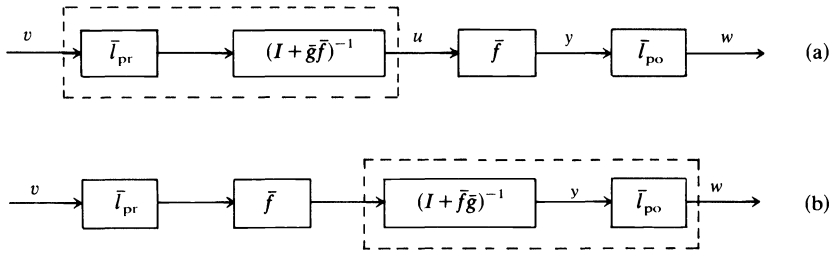


FIG. 3.2

by the original system preceded and followed by bicausal compensators, *with the feedback compensator represented, as one chooses, either as a precompensator or a postcompensator.*

Because of the obvious duality between the precompensator situation and the postcompensator situation, there is no need to discuss both of them in detail. Since practical interest in postcompensators is at best limited, we shall henceforth confine our attention to precompensation, and discuss postcompensators only in connection with certain mathematical questions.

For various reasons, not to be elaborated on here, feedback compensation is preferred over external compensation whenever possible. Thus, one is interested in the following problem.

Causal feedback problem 3.3. Let $\bar{f}: \Lambda U \to \Lambda Y$ be an extended linear i/o map.

(a) Under what conditions can a given bicausal ΛK-linear isomorphism $\bar{l}: \Lambda U \to \Lambda U$ be represented as feedback, i.e. under what conditions do there exist a static map $L: \Lambda U \to \Lambda U$ and a causal ΛK-linear map $\bar{g}: \Lambda Y \to \Lambda Y$, such that $\bar{l}^{-1} = L + \bar{g}\bar{f}$?

(b) Under what conditions (on $\bar{f}$) can every bicausal $\bar{l}$ be represented as feedback?

Let $\bar{l}: \Lambda U \to \Lambda U$ be a bicausal ΛK-linear map, and let

$$Z_{\bar{l}^{-1}}(z^{-1}) = \sum_{t=0}^{\infty} L_t z^{-t}$$

denote the transfer function of $\bar{l}^{-1}$. We can then write

$$Z_{\bar{l}^{-1}}(z^{-1}) = L_0 + \sum_{t=1}^{\infty} L_t z^{-t} = L_0 + Z_{\bar{h}}(z^{-1}),$$

where L_0 is a static ΛK-linear map and $Z_{\bar{h}}(z^{-1})$ is the transfer function of a strictly causal map $\bar{h}: \Lambda U \to \Lambda U$ representing the strictly causal part of $\bar{l}^{-1}$. Hence we can always decompose the map $\bar{l}^{-1}$ as

$$\bar{l}^{-1} = L + \bar{h},$$

with L static and $\bar{h}$ strictly causal. The causal feedback problem 3.3 is therefore essentially equivalent to the following.

Causal factorization problem 3.4. Let $\bar{f}: \Lambda U \to \Lambda Y$ be a given strictly causal ΛK-linear map.

(a) Under what conditions can a strictly causal ΛK-linear map $\bar{h}: \Lambda U \to \Lambda U$ be factored causally over $\bar{f}$, i.e., when does there exist a causal map $\bar{g}: \Lambda Y \to \Lambda U$ such that $\bar{h} = \bar{g} \cdot \bar{f}$?

(b) Under what conditions can every strictly causal ΛK-linear map $\bar{h}: \Lambda U \to \Lambda U$ be factored causally over $\bar{f}$?

It is readily noted that the strict causality of the maps $\bar{f}$ and $\bar{h}$ is inessential to the causal factorization problem, and arises in problem 3.4 only because of the specific requirements of the feedback problem. Indeed, if $\bar{h}$ factors causally over $\bar{f}$, i.e., if there exists a causal $\bar{g}$ such that $\bar{h} = \bar{g} \cdot \bar{f}$, then for each integer k we also have $z^k\bar{h} = z^k\bar{g}\bar{f} = \bar{g} \cdot (z^k\bar{f})$ so that $z^h\bar{h}$ factors causally over $z^k\bar{f}$, and for sufficiently large positive k (unless $\bar{h}$ or $\bar{f}$ are zero) the maps $z^k\bar{h}$ and $z^k\bar{f}$ are not causal. Thus, the causal factorization problem can be stated in the following less restrictive way:

Given two ΛK-linear maps $\bar{f}: \Lambda S \to \Lambda Y$ and $\bar{h}: \Lambda S \to \Lambda W$ (where S, Y and W are K-linear spaces), when does there exist a causal ΛK-linear map $\bar{g}: \Lambda Y \to \Lambda W$ such that the following diagram in Fig. 3.3 commutes

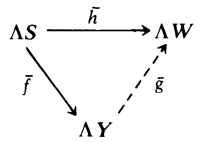


FIG. 3.3

If the causality requirement of $\bar{g}$ is dropped, the factorization problem is standard (see, e.g., Greub [1967]) and $\bar{h}$ factors over $\bar{f}$ if and only if $\ker \bar{f} \subset \ker \bar{h}$. Yet this condition does not say anything about the causality of $\bar{g}$. To deal efficiently with the causality issue, we reintroduce the concept of causality using an approach which is algebraically more tractable.

Let $\bar{f}: \Lambda U \to \Lambda Y$ be a ΛK-linear map. We can characterize causality of $\bar{f}$ as follows (compare with our definitions of causality in § 2):

(3.5) The map $\bar{f}$ is *causal* if and only if $u \in \Omega^- U$ implies $\bar{f}(u) \in \Omega^- Y$.

Similarly, we have:

(3.6) The map $\bar{f}$ is *strictly causal* if and only if $u \in z\Omega^- U$ implies $\bar{f}(u) \in \Omega^- Y$.

Let us denote the $\Omega^- K$-quotient module $\Lambda Y / \Omega^- Y$ by $\Gamma^- Y$, and let $\pi^-: \Lambda Y \to \Gamma^- Y$ denote the canonical projection. The following can then be easily verified by the reader.

PROPOSITION 3.7. *Let $\bar{f}: \Lambda U \to \Lambda Y$ be a ΛK-linear map.*

(a) *The map $\bar{f}$ is causal if and only if $\Omega^- U \subset \ker \pi^- \bar{f}$.*

(b) *The map $\bar{f}$ is strictly causal if and only if $z\Omega^- U \subset \ker \pi^- \bar{f}$.*

(c) *The map $\bar{f}$ is order consistent if and only if, for some integer k, $z^k \Omega^- U = \ker \pi^- \bar{f}$.*

(d) *The map $\bar{f}$ is instantaneous if and only if $\Omega^- U = \ker \pi^- \bar{f}$.*

(e) *The map $\bar{f}$ is nonlatent if and only if $z\Omega^- U = \ker \pi^- \bar{f}$.*

We shall use the characterizations of the above proposition extensively in the following sections.

4. Proper independence and proper bases. Let K be a field and let $S := K^m$. For an element $0 \neq s \in \Lambda S$, denote by $\hat{s}$ the leading coefficient of s. If $s = 0$ we shall say that $\hat{s} = 0$.

DEFINITION 4.1. A set of vectors $s_1, \cdots, s_k \in \Lambda S$ is called *properly independent* if their leading coefficients $\hat{s}_1, \cdots, \hat{s}_k \in S$ are K-linearly independent.

Below we derive a variety of properties of properly independent sets, of proper bases and of proper direct sum decompositions. Our objective is to develop this theory here only to the extent required in the sequel. Many further results have been omitted, and the reader can, for example, easily verify that the converses of a number of our results are also valid. A more extensive exposition of this and related topics will be published elsewhere.

LEMMA 4.2. *If $s_1, \cdots, s_k \in \Lambda S$ is a properly independent set of vectors, then* (i) *it is ΛK-linearly independent, and* (ii) *for every set of scalars $\alpha_1, \cdots, \alpha_k \in \Lambda K$ the following holds*:

$$\operatorname{ord} \sum_{i=1}^{k} \alpha_i s_i = \min \{\operatorname{ord} \alpha_i s_i | i = 1, \cdots, k\}.$$

Proof. We shall prove the lemma by showing that if either (i) or (ii) fails to hold then the set $s_1, \cdots, s_k$ is not properly independent. If $\alpha_1, \cdots, \alpha_k \in \Lambda K$ is any set of scalars then, by definition, $\operatorname{ord} \sum_{i=1}^{k} \alpha_i s_i \geqq r := \min \{\operatorname{ord} \alpha_i s_i | i = 1, \cdots, k\}$. If either (i) or (ii) fails to hold, there exist $\alpha_1, \cdots, \alpha_k \in \Lambda K$, not all zero, such that either $\sum_{i=1}^{k} \alpha_i s_i = 0$ or $\operatorname{ord} \sum_{i=1}^{k} \alpha_i s_i > r$. For each $i = 1, \cdots, k$ define

$$\varepsilon_i := \begin{cases} 1 & \text{if } \operatorname{ord} \alpha_i s_i = r, \\ 0 & \text{if } \operatorname{ord} \alpha_i s_i > r, \end{cases}$$

and consider the terms of order r in $\Sigma \alpha_i s_i$. This yields $\sum_{i=1}^{k} \varepsilon_i \hat{\alpha}_i \hat{s}_i = 0$, implying that $\hat{s}_1, \cdots, \hat{s}_k$ are K-linearly dependent since not all the $\varepsilon_i \hat{\alpha}_i$ are zero. Hence $s_1 \cdots, s_k$ are not properly independent, completing the proof. □

The condition of Lemma 4.2(ii) has been called the "predictable degree property" in Forney [1975], in the (analogous) setting of "minimal polynomial bases" for rational vector spaces. We shall adopt this terminology and call the property of Lemma 4.2(ii) the *predictable order property*.

DEFINITION 4.3. Let $\mathcal{R} \subset \Lambda S$ be a ΛK-linear subspace. A basis $\{s_1, \cdots, s_k\}$ of $\mathcal{R}$ is called *proper* if the vectors $s_1, \cdots, s_k$ are properly independent. The basis is called *normalized* if for each $i = 1, \cdots, k$, $\operatorname{ord} s_i = 0$.

To avoid possible confusion in the ensuing discussion where we shall deal with both K-linear and ΛK-linear spaces, we shall use subscripts to emphasize the field. Thus, for example, $\operatorname{span}_{\Lambda K} \{s_1, \cdots, s_k\}$ denotes the ΛK-linear subspace spanned by $s_1, \cdots, s_k \in \Lambda S$, whereas $\operatorname{span}_K \{\hat{s}_1, \cdots, \hat{s}_k\}$ denotes the K-linear subspace spanned by $\hat{s}_1, \cdots, \hat{s}_k \in S$. Similarly, $\dim_{\Lambda K} \mathcal{R}$ denotes the dimension of a subspace $\mathcal{R} \subset \Lambda S$ as a ΛK-linear space (to distinguish from K-linear). We next have the following theorem.

THEOREM 4.4. *Every nonzero ΛK-linear subspace $\mathcal{R} \subset \Lambda S$ has a proper basis. Moreover, every properly independent subset of $\mathcal{R}$ can be extended to a proper basis.*

Proof. Let $0 \neq s_1 \in \mathcal{R}$ be any vector. Then s_1 is properly independent. We shall complete the proof by showing that if $s_1, \cdots, s_k \in \mathcal{R}$ are a properly independent set and if $\mathcal{R}_k := \operatorname{span}_{\Lambda K} \{s_1, \cdots, s_k\}$ is a proper subspace of $\mathcal{R}$, we can find a vector $s_{k+1} \in \mathcal{R}$ such that the set $\{s_1, \cdots, s_k, s_{k+1}\}$ is also properly independent. The proof is by contradiction. Assume that $\mathcal{R}_k \subset \mathcal{R}$ is a proper subspace, let $s_{k+1}^{\circ} \in \mathcal{R}$ be such that the set $\{s_1, \cdots, s_k, s_{k+1}^{\circ}\}$ is ΛK-linearly independent and, without loss of generality, assume that this set is also normalized. Let $\mathcal{R}_{k+1} := \operatorname{span}_{\Lambda K} \{s_1, \cdots, s_k, s_{k+1}^{\circ}\}$ and suppose that there is no vector $s \in \mathcal{R}_{k+1}$ such that the set $\{s_1, \cdots, s_k, s\}$ is properly independent. This means that for each $s \in \mathcal{R}_{k+1}$, $\hat{s} \in \hat{\mathcal{R}}_k := \operatorname{span}_K \{\hat{s}_1, \cdots, \hat{s}_k\}$, contradicting, as we shall see, the ΛK-linear independence of $s_1, \cdots, s_k, s_{k+1}^{\circ}$. Indeed, we observe that there are scalars $\alpha_1^{\circ}, \cdots, \alpha_k^{\circ} \in K$ such that $\hat{s}_{k+1}^{\circ} = \sum_{i=1}^{k} \alpha_i^{\circ} \hat{s}_i$. Let $n_0 := 0$ and set $s_{k+1}^{1} := s_{k+1}^{\circ} - \sum_{i=1}^{k} \alpha_i^{\circ} z^{-n_0} s_i$, so that $\operatorname{ord} s_{k+1}^{1} > \operatorname{ord} s_{k+1}^{\circ}$. We now form a sequence of vectors $\{s_{k+1}^{t}\}$, $t = 0, 1, 2, \cdots$, with $s_{k+1}^{t} \in \mathcal{R}_{k+1}$, such that $\operatorname{ord} s_{k+1}^{t+1} > \operatorname{ord} s_{k+1}^{t}$ for all $t = 0, 1, 2, \cdots$ as follows: For each t, set $n_t = \operatorname{ord} s_{k+1}^{t}$ and let $s_{k+1}^{t+1} := s_{k+1}^{t} - \sum_{i=1}^{k} \alpha_i^{t} z^{-n_t} s_i$, where the scalars $\alpha_1^{t}, \cdots, \alpha_k^{t} \in K$ satisfy the condition that $\hat{s}_{k+1}^{t} = \sum_{i=1}^{k} \alpha_i^{t} \hat{s}_i$. Upon defining $\alpha_i := \sum_{t=0}^{\infty} \alpha_i^{t} z^{-n_t} \in \Lambda K$, $i = 1, \cdots, k$, it is readily verified that $s_{k+1}^{\circ} - \sum_{i=1}^{k} \alpha_i s_i = 0$, whence $s_{k+1}^{\circ} \in \mathcal{R}_k$, a contradiction. □

COROLLARY 4.5. *Let $\mathcal{R} \subset \Lambda S$ be a ΛK-linear subspace. Then* $\dim_{\Lambda K} \mathcal{R} = \dim_K \hat{\mathcal{R}}$, *where* $\hat{\mathcal{R}} := \operatorname{span}_K \{\hat{s} | s \in \mathcal{R}\}$.

Let $\mathcal{R} \subset \Lambda S$ be a ΛK-linear subspace. If $\mathcal{R} = \mathcal{R}_1 \oplus \mathcal{R}_2$ is a direct sum decomposition of $\mathcal{R}$ into ΛK-linear subspaces $\mathcal{R}_1$ and $\mathcal{R}_2$, then, in general, $\hat{\mathcal{R}}_1 \cap \hat{\mathcal{R}}_2 \neq 0$ so that $\hat{\mathcal{R}} \neq \hat{\mathcal{R}}_1 + \hat{\mathcal{R}}_2$. This leads us to the following

DEFINITION 4.6. A direct sum decomposition $\mathcal{R} = \mathcal{R}_1 \oplus \mathcal{R}_2$ of a ΛK-linear subspace $\mathcal{R} \subset \Lambda S$ into ΛK-linear subspaces $\mathcal{R}_1$ and $\mathcal{R}_2$ is called *proper* if $\hat{\mathcal{R}}_1 \cap \hat{\mathcal{R}}_2 = 0$. The subspace $\mathcal{R}_2$ is then called a *proper direct summand* of $\mathcal{R}_1$.

With the aid of Corollary 4.5 it is readily seen that a direct sum decomposition is proper if and only if $\hat{\mathcal{R}} = \hat{\mathcal{R}}_1 + \hat{\mathcal{R}}_2$. Thus, $\mathcal{R} = \mathcal{R}_1 \oplus \mathcal{R}_2$ is a proper decomposition if and only if there are proper bases $s_{11}, \cdots, s_{1k_1}$ of $\mathcal{R}_1$ and $s_{21}, \cdots s_{2k_2}$ of $\mathcal{R}_2$ such that the set $s_{11}, \cdots, s_{1k_1}, s_{21}, \cdots, s_{2k_2}$ is a proper basis of $\mathcal{R}$. We then have the following further corollary to Theorem 4.4.

COROLLARY 4.7. *Let $\mathcal{R} \subset \Lambda S$ be a ΛK-linear subspace. Then every ΛK-linear subspace $\mathcal{R}_1 \subset \mathcal{R}$ has a proper direct summand in $\mathcal{R}$.*

Finally, we also have the following variant of the predictable order property.

COROLLARY 4.8. *Let $\mathcal{R} = \mathcal{R}_1 \oplus \mathcal{R}_2$ be a proper direct sum decomposition of a ΛK-linear subspace $\mathcal{R} \subset \Lambda S$. Let $s = s_1 + s_2$ be the representation of any vector $s \in \mathcal{R}$, with $s_i \in \mathcal{R}_i$, $i = 1, 2$. Then* $\operatorname{ord} s = \min \{\operatorname{ord} s_1, \operatorname{ord} s_2\}$.

Proof. By definition, $\operatorname{ord} s \geqq \min \{\operatorname{ord} s_1, \operatorname{ord} s_2\}$. If the above inequality is strict, there exist scalars $\alpha_1, \alpha_2 \in K$, not both zero, such that $\alpha_1 \hat{s}_1 + \alpha_2 \hat{s}_2 = 0$ contradicting the fact that $\hat{\mathcal{R}}_1 \cap \hat{\mathcal{R}}_2 = 0$. □

5. Causal factorization. We turn now to the causal factorization problem (3.4). As we mentioned earlier, there is no essential need, in characterizing causal factorizability, to assume strict causality, or even causality, of the maps under consideration. We shall therefore begin with the general case and turn to specific consideration of i/o maps later on. We shall assume that the spaces U and Y are finite dimensional, in particular that $U = K^m$ and $Y = K^p$. For convenience of notation, we shall temporarily use the notation ΛU and ΛY also in connection with ΛK-linear maps $\bar{f}: \Lambda U \to \Lambda Y$ that are *not* necessarily i/o maps (i.e., are not necessarily strictly causal).

Let $\bar{f}: \Lambda U \to \Lambda Y$ be a ΛK-linear map and let $\pi^-: \Lambda Y \to \Gamma^- Y := \Lambda Y / \Omega^- Y$ be the canonical projection. Since $\Omega^- Y$ is an $\Omega^- K$-module, so is the quotient $\Lambda Y / \Omega^- Y$. Thus the map π^- is an $\Omega^- K$-homomorphism and so is also the composite $\pi^- \bar{f}$. We have

LEMMA 5.1. *Let $\bar{f}: \Lambda U \to \Lambda Y$ be a ΛK-linear map and let $\pi^-: \Lambda Y \to \Gamma^- Y$ be the canonical projection. If $\mathcal{R} \subset \ker \pi^- \bar{f}$ is a ΛK-linear subspace, then $\mathcal{R} \subset \ker \bar{f}$.*

Proof. Assume $u \in \mathcal{R} \subset \ker \pi^- \bar{f}$, where $\mathcal{R}$ is a ΛK-linear subspace. Then $\alpha u \in \ker \pi^- \bar{f}$ for all $\alpha \in \Lambda K$. Thus $\bar{f}(\alpha u) = \alpha \bar{f}(u) \in \Omega^- Y$ for all $\alpha \in \Lambda K$, whence $\bar{f}(u) = 0$ and $u \in \ker \bar{f}$ as claimed. □

Next we have the following central theorem.

THEOREM 5.2. *Let $\bar{f}: \Lambda U \to \Lambda Y$ and $\bar{h}: \Lambda U \to \Lambda W$ be ΛK-linear maps, where U, Y and W are finite dimensional K-linear spaces. There exists a causal ΛK-linear map $\bar{g}: \Lambda Y \to \Lambda W$ such that $\bar{h} = \bar{g} \cdot \bar{f}$ if and only if* $\ker \pi^- \bar{f} \subset \ker \pi^- \bar{h}$.

Proof. Suppose $\bar{h} = \bar{g} \cdot \bar{f}$ with $\bar{g}$ causal. Let $u \in \ker \pi^- \bar{f}$. Then $\bar{f}(u) \in \Omega^- Y$, and by causality of $\bar{g}$ (see Proposition 3.7(a)) $\Omega^- Y \subset \ker \pi^- \bar{g}$. It follows that $\bar{f}(u) \in \ker \pi^- \bar{g}$ whence $u \in \ker \pi^- \bar{g} \cdot \bar{f} = \ker \pi^- \bar{h}$. Conversely, assume that $\ker \pi^- \bar{f} \subset \ker \pi^- \bar{h}$. By Lemma 5.1 this implies that $\ker \bar{f} \subset \ker \bar{h}$ whence by a standard theorem of linear algebra (see, e.g., Greub [1967]) a ΛK-linear map $\bar{g}: \Lambda Y \to \Lambda W$ such that $\bar{h} = \bar{g} \cdot \bar{f}$ exists. It remains to be shown that the map $\bar{g}$ can be selected to be causal. To this end write $\Lambda Y = \operatorname{Im} \bar{f} \oplus \mathcal{R}$, where $\operatorname{Im} \bar{f}$ is the image of $\bar{f}$ and $\mathcal{R}$ is any proper direct summand

(see Corollary 4.7). Let $\bar{g}_0 : \Lambda Y \to \Lambda W$ be any ΛK-linear map that satisfies the condition that $\bar{h} = \bar{g}_0 \cdot \bar{f}$ and let $\bar{g}_1 : \operatorname{Im} \bar{f} \to \Lambda W$ be the restriction of $\bar{g}_0$ to the image of $\bar{f}$. Let $p : \Lambda Y \to \operatorname{Im} \bar{f}$ denote the projection onto $\operatorname{Im} \bar{f}$ along $\mathscr{R}$; that is, if $y = y_1 + y_2 \in \Lambda Y$ is the decomposition of y into its components $y_1 \in \operatorname{Im} \bar{f}$ and $y_2 \in \mathscr{R}$, then $py = y_1$. Clearly, p is ΛK-linear, and we shall see that the map $\bar{g} = \bar{g}_1 \cdot p$ satisfies the conditions of the theorem. First observe that for $u \in \Lambda U$,

$$\bar{g} \cdot \bar{f}(u) = \bar{g}_1 \cdot p\bar{f}(u) = \bar{g}_0 \bar{f}(u) = \bar{h}(u),$$

so that $\bar{g} \cdot \bar{f} = \bar{h}$. To see that $\bar{g}$ is causal, let $y = y_1 + y_2 \in \Omega^- Y$, where $y_1 \in \operatorname{Im} \bar{f}$ and $y_2 \in \mathscr{R}$. By Proposition 3.7(a), the proof will be complete if we show that $y \in \ker \pi^- \bar{g}$. Indeed, Corollary 4.8 implies that both y_1 and y_2 are in $\Omega^- Y$ so that $\bar{g} \cdot y = \bar{g}_1 \cdot py = \bar{g}_1 \cdot y_1 = \bar{g}_0 \cdot \bar{f}(u)$ for some $u \in \ker \pi^- \bar{f}$. But by hypothesis $\ker \pi^- \bar{f} \subset \ker \pi^- \bar{h}$, whence $\bar{g} \cdot y = \bar{g}_0 \cdot \bar{f}(u) = \bar{h}(u) \in \Omega^- W$ so that $y \in \ker \pi^- \bar{g}$ as claimed. □

Theorem 5.2 clarifies the significance of the $\Omega^- K$-module $\ker \pi^- \bar{f}$ in connection with the causal factorization problem (and consequently also with feedback). We call this module the *latency module* or *latency kernel* of $\bar{f}$.

COROLLARY 5.3. *Let $\bar{f} : \Lambda U \to \Lambda Y$ be a ΛK-linear map of finite order. Then $\bar{f}$ is order consistent if and only if for every ΛK-linear map $\bar{h} : \Lambda U \to \Lambda W$ which satisfies* $\operatorname{ord} \bar{h} \geqq \operatorname{ord} \bar{f}$ *there exists a causal ΛK-linear map $\bar{g} : \Lambda Y \to \Lambda W$ such that $\bar{h} = \bar{g} \cdot \bar{f}$.*

Proof. Recall that a map $\bar{f}$ is order consistent if $\operatorname{ord} \bar{f}(u) - \operatorname{ord} u = \operatorname{ord} \bar{f}$ for each $0 \neq u \in \Lambda U$. Suppose $\bar{f}$ is order consistent and $\operatorname{ord} \bar{h} \geqq \operatorname{ord} \bar{f}$. Let $0 \neq u \in \ker \pi^- \bar{f}$. Then $\bar{f}(u) \in \Omega^- Y$ and $\operatorname{ord} \bar{f}(u) \geqq 0$. Now $\operatorname{ord} \bar{h}(u) - \operatorname{ord} u \geqq \operatorname{ord} \bar{h} \geqq \operatorname{ord} \bar{f} = \operatorname{ord} \bar{f}(u) - \operatorname{ord} u$, whence $\operatorname{ord} \bar{h}(u) \geqq \operatorname{ord} \bar{f}(u) \geqq 0$, so that $u \in \ker \pi^- \bar{h}$, implying that $\ker \pi^- \bar{f} \subset \ker \pi^- \bar{h}$. By Theorem 5.2 the existence of a causal $\bar{g}$ such that $\bar{h} = \bar{g} \cdot \bar{f}$ is thus assured. Conversely, suppose $\bar{f}$ is not order consistent and that $\bar{h}$ is an order consistent map satisfying $\operatorname{ord} \bar{h} = \operatorname{ord} \bar{f}$. Then there exists $0 \neq u \in \Lambda U$ such that $\operatorname{ord} \bar{f}(u) > \operatorname{ord} \bar{f} + \operatorname{ord} u = \operatorname{ord} \bar{h} + \operatorname{ord} u = \operatorname{ord} \bar{h}(u)$. If $k := \operatorname{ord} \bar{f}(u)$, then $0 = \operatorname{ord} \bar{f}(z^k u) > \operatorname{ord} \bar{h}(z^k u)$ so that $z^k u \in \ker \pi^- \bar{f}$ but $z^k u \notin \ker \pi^- \bar{h}$. Hence $\ker \pi^- \bar{f} \not\subset \ker \pi^- \bar{h}$ and by Theorem 5.2 there does not exist a causal $\bar{g}$ such that $\bar{h} = \bar{g} \cdot \bar{f}$, completing the proof. □

The following corollary which is an immediate consequence of Corollary 5.3 is of central interest in our study of causal factorization since it deals with linear i/o maps and gives us an important characterization of nonlatency.

COROLLARY 5.4. *Let $\bar{f} : \Lambda U \to \Lambda Y$ be an extended linear* i/o *map. Then $\bar{f}$ is nonlatent if and only if for every strictly causal ΛK-linear map $\bar{h} : \Lambda U \to \Lambda W$ there exists a causal ΛK-linear map $\bar{g} : \Lambda Y \to \Lambda W$ such that $\bar{h} = \bar{g} \cdot \bar{f}$.*

Let $\bar{f} : \Lambda U \to \Lambda Y$ be an extended linear i/o map and let $\bar{l} : \Lambda U \to \Lambda U$ be a bicausal isomorphism, i.e., a bicausal precompensator for $\bar{f}$. Let $\bar{h}$ be the strictly causal part of $\bar{l}^{-1}$, i.e., $\bar{l}^{-1} = L + \bar{h}$ where L is static. As we have seen in § 3, $\bar{l}$ can be realized as feedback around $\bar{f}$ if $\bar{h}$ factors causally over $\bar{f}$. Theorem 5.2 tells us essentially that the only barrier to realizing a bicausal precompensator as feedback is the relative latency of $\bar{f}$ and $\bar{h}$. Corollary 5.4 characterizes the class of i/o maps over which every bicausal precompensator can be realized as feedback. These i/o maps are, as we have seen, the nonlatent maps (a fact which motivated our choice of terminology). Now, a very special and important class of nonlatent maps is that of injective i/s maps. This fact is proved in the following theorem.

THEOREM 5.5. *Let $\bar{f} : \Lambda U \to \Lambda Y$ be an injective linear* i/s *map. Then $\bar{f}$ is nonlatent.*

Proof. By strict causality of $\bar{f}$ we have that $z\Omega^- U \subset \ker \pi^- \bar{f}$, so that to prove nonlatency we need only to show that $\ker \pi^- \bar{f} \subset z\Omega^- U$. Let $u \in \ker \pi^- \bar{f}$ so that $\bar{f}(u) \in \Omega^- Y$. Write $u = u^+ + u^-$, where $u^+ \in z^2 \Omega^+ U$ and $u^- \in z\Omega^- U$. The proof will be completed by showing that $u^+ = 0$ so that $u \in z\Omega^- U$ as claimed. Note that $\bar{f}(u^-) \in \Omega^- Y$

by the strict causality of $\bar{f}$ so that, in view of the fact that $\bar{f}(u) = \bar{f}(u^+) + \bar{f}(u^-)$, it follows that $\bar{f}(u^+) \in \Omega^- Y$. By (2.16) we have

$$\bar{f}(u^+) = \sum_{t \in \mathbb{Z}} f(\mathcal{S}^+(z^t u^+)) z^{-t-1} \in \Omega^- Y,$$

so that, in particular, $f(\mathcal{S}^+(z^{-2}u^+)) = 0$. But $z^{-2}u^+ \in \Omega^+ U$, whence $f(\mathcal{S}^+(z^{-2}u^+)) = f(z^{-2}u^+) = 0$ implying that $z^{-2}u^+ \in \ker f = \ker \tilde{f}$ (the equality being a consequence of the i/s property (2.20)). It follows that $\bar{f}(z^{-2}u^+) \in \Omega^+ Y$, or alternatively, that $\bar{f}(u^+) \in z^2 \Omega^+ Y$. Since $z^2 \Omega^+ Y \cap \Omega^- Y = 0$, we conclude that $\bar{f}(u^+) = 0$ or that $u^+ = 0$ by the injectivity of $\bar{f}$. □

While Theorem 5.5 deals only with *injective* i/s maps, it is important to observe that this is not a serious restriction. Indeed, it is shown in Proposition 5.6 below that in the special case of i/s maps (in contrast to i/o maps in general), the kernel is "static"; i.e., if $\bar{f}$ is a noninjective i/s map, then $\ker \bar{f} = \Lambda U^0$ where $U^0 \subset U$ is a subspace. This means that the whole degeneracy lies in the input value space U which has been chosen too large, and by restricting the input value space to a proper summand of U^0 in U, the injectivity is restored.

PROPOSITION 5.6. *Let $\bar{f} : \Lambda U \to \Lambda Y$ be an extended linear* i/s *map. Then there exists a subspace $U^0 \subset U$ such that* $\ker \bar{f} = \Lambda U^0$.

Proof. Let $i_u : U \to \Omega^+ U : u \mapsto u$ be the canonical injection and define the subspace $U^0 \subset U$ as $U^0 := \ker f \cdot i_u$, where f is the output value map associated with $\bar{f}$. Since $\bar{f}$ is an i/s map we have $\ker f \cdot i_u = \ker \tilde{f} \cdot i_u = \ker \bar{f} \cdot \bar{i}_u$ with the last equality holding by the strict causality of $\bar{f}$. Thus $\bar{i}_u(U^0) \subset \ker \bar{f}$, and since $\ker \bar{f}$ is a ΛK-linear space we conclude that $\Lambda U^0 \subset \ker \bar{f}$. To prove that $\ker \bar{f} \subset \Lambda U^0$, it suffices to prove that if $0 \neq u = \sum_{t=t_0}^{\infty} u_t z^{-t} \in \ker \bar{f}$ then $u_{t_0} \in U^0$. By recursive application of the same argument this will then imply that $u_t \in U^0$ for all $t \geqq t_0$. Now by formula (2.16) we have $f(\mathcal{S}^+(z^k u)) = 0$ for all $k \in \mathbb{Z}$, and since $\mathcal{S}^+(z^{t_0} u) = u_{t_0}$ the results follow. □

The importance of Theorem 5.5 lies in the fact that it tells us that bicausal precompensation is equivalent, in the sense of solvability, to dynamic state feedback. Let $\bar{f} : \Lambda U \to \Lambda Y$ be an extended linear i/o map. We write (see Hautus and Heymann [1978]) $\bar{f} = H \cdot \bar{f}_s$, where H is a static output map and $\bar{f}_s$ is a reachable i/s map. If $\bar{f}_s$ is injective (which is always the case when $\ker \bar{f}$ does not contain a subspace of the form ΛS, $0 \neq S \subset U$), then every bicausal precompensator can be realized as feedback around $\bar{f}_s$. That is, we can write every bicausal $\bar{l} : \Lambda U \to \Lambda U$ as $\bar{l}^{-1} = L + \bar{g}\bar{f}_s$, where $\bar{g} : \Lambda Y \to \Lambda U$ is a causal ΛK-linear map and L is static.

Before we proceed with our general investigation, it is worthwhile to record one more consequence of Theorem 5.2.

COROLLARY 5.7. *Let $\bar{f}_1, \bar{f}_2 : \Lambda U \to \Lambda Y$ be two extended linear* i/o *maps with U and Y finite dimensional K-linear spaces. There exists a bicausal ΛK-linear map $\bar{l} : \Lambda Y \to \Lambda Y$ such that $\bar{f}_2 = \bar{l} \cdot \bar{f}_1$ if and only if* $\ker \pi^- \bar{f}_1 = \ker \pi^- \bar{f}_2$.

Proof. First, observe that if a bicausal $\bar{l}$ exists then, by Theorem 5.2, it follows immediately that $\ker \pi^- \bar{f}_1 = \ker \pi^- \bar{f}_2$. Conversely, assume that $\ker \pi^- \bar{f}_1 = \ker \pi^- \bar{f}_2$ and write $\Lambda Y = \operatorname{Im} \bar{f}_1 \oplus \mathcal{R}_1 = \operatorname{Im} \bar{f}_2 \oplus \mathcal{R}_2$ where $\mathcal{R}_1$ and $\mathcal{R}_2$ are proper direct summands. By Theorem 5.2 there exist causal maps $\bar{l}^1, \bar{l}^2 : \Lambda Y \to \Lambda Y$ such that $\bar{l}^1 \bar{f}_1 = \bar{f}_2$ and $\bar{l}^2 \bar{f}_2 = \bar{f}_1$. Hence $\bar{l}^2 \cdot \bar{l}^1 \bar{f}_1 = \bar{f}_1$, and letting $\bar{l}_1 : \operatorname{Im} \bar{f}_1 \to \Lambda Y$ denote the restriction of $\bar{l}^1$ to the image of $\bar{f}_1$, it is readily verified that $\bar{l}_1$ is order preserving. Now, $\ker \pi^- \bar{f}_1 = \ker \pi^- \bar{f}_2$ implies that $\ker \bar{f}_1 = \ker \bar{f}_2$, whence $\dim \operatorname{Im} \bar{f}_1 = \dim \operatorname{Im} \bar{f}_2$ and $\dim \mathcal{R}_1 = \dim \mathcal{R}_2$. Let $\bar{l}_2 : \mathcal{R}_1 \to \Lambda Y$ be an order preserving map satisfying $\operatorname{Im} \bar{l}_2 = \mathcal{R}_2$ and let $p : \Lambda Y \to \operatorname{Im} \bar{f}_1$ denote the projection along $\mathcal{R}_1$. We claim that the map $\bar{l} : \Lambda Y \to \Lambda Y$:

$y \mapsto \bar{l}_1 py + \bar{l}_2(I-p)y$ is a bicausal isomorphism and that $\bar{l} \cdot \bar{f}_1 = \bar{f}_2$. Indeed, to see the latter property, note that for any $u \in \Lambda U$ we have

$$\bar{l}\bar{f}_1(u) = \bar{l}_1 p \bar{f}_1(u) + \bar{l}_2(I-p)\bar{f}_1(u) = \bar{l}_1 \cdot \bar{f}_1(u) = \bar{l}^1 \bar{f}_1(u) = \bar{f}_2(u).$$

To see the bicausality of $\bar{l}$ it suffices to show that it is order preserving. Indeed, let $y = y_1 + y_2 \in \Lambda Y$ be any element with $y_1 \in \operatorname{Im} \bar{f}_1$ and $y_2 \in \mathscr{R}_1$. Then $\bar{l}y = \bar{l}_1 y_1 + \bar{l}_2 y_2$ and, using Corollary 4.8 together with the fact that $\operatorname{Im} \bar{l}_1$ and $\operatorname{Im} \bar{l}_2$ form a proper direct sum, we have that $\operatorname{ord} \bar{l}y = \min\{\operatorname{ord} \bar{l}_1 y_1, \operatorname{ord} \bar{l}_2 y_2\} = \min\{\operatorname{ord} y_1, \operatorname{ord} y_2\}$, where the last equality follows from the order preserving property of $\bar{l}_1$ and $\bar{l}_2$. Using Corollary 4.8 again, together with the fact that $\operatorname{Im} \bar{f}_1$ and $\mathscr{R}_1$ form a proper direct sum, gives that $\min\{\operatorname{ord} y_1, \operatorname{ord} y_2\} = \operatorname{ord} y$ whence $\operatorname{ord} \bar{l}y = \operatorname{ord} y$ as claimed and the proof is complete. □

Clearly, the bicausal ΛK-linear map $\bar{l}$ of Corollary 5.7 can be regarded as a bicausal postcompensator for $\bar{f}_1$, and there is a kind of duality between feedback and compensation which deserves some further comments.

Let $\bar{f}: \Lambda U \to \Lambda Y$ be an extended linear i/o map and let $\bar{l}_{\mathrm{pr}}: \Lambda U \to \Lambda U$ be a bicausal precompensator for $\bar{f}$. If $\bar{w}: \Lambda U \to \Lambda U$ is the strictly causal part of $\bar{l}_{\mathrm{pr}}$, then the causal feedback problem is that of existence of a causal ΛK-linear map $\bar{g}: \Lambda Y \to \Lambda U$ such that $\bar{w} = \bar{g} \cdot \bar{f}$. The map $\bar{g}$ can be regarded essentially as a causal (but not necessarily bicausal) postcompensator for $\bar{f}$. Conversely, if $\bar{l}_{\mathrm{po}}: \Lambda Y \to \Lambda Y$ is a bicausal postcompensator and if $\bar{w}: \Lambda Y \to \Lambda Y$ the strictly causal part of $\bar{l}_{\mathrm{po}}$, the dual of the above causal factorization problem is that of the existence of a causal ΛK-linear map $\bar{g}: \Lambda Y \to \Lambda U$ such that $\bar{w} = \bar{f} \cdot \bar{g}$. Here $\bar{g}$ can be viewed as a causal, but again not necessarily bicausal, precompensator for $\bar{f}$. Thus the pre- and postcompensator problems become interrelated through feedback. We can also write down the dual of Corollary 5.7 regarding the problem of bicausal precompensation.

COROLLARY 5.8. *Let $\bar{f}_1, \bar{f}_2: \Lambda U \to \Lambda Y$ be two extended linear* i/o *maps with U and Y finite dimensional K-linear spaces. There exists a bicausal ΛK-linear map $\bar{l}: \Lambda U \to \Lambda U$ such that $\bar{f}_2 = \bar{f}_1 \cdot \bar{l}$ if and only if $\ker \pi^- \bar{f}_1^* = \ker \pi^- \bar{f}_2^*$, where $\bar{f}_1^*$ and $\bar{f}_2^*$ denote the dual maps of $\bar{f}_1$ and $\bar{f}_2$ respectively.*

In Corollary 5.8 the dual maps $\bar{f}_1^*$ and $\bar{f}_2^*$ can of course be identified with the transposes of the corresponding maps (or transfer functions) in view of the finite dimensionality of the underlying spaces.

In Hautus and Heymann [1978], the static state feedback problem was investigated. This is the following problem: Given an extended linear i/s map $\bar{f}: \Lambda U \to \Lambda Y$, under what conditions can a bicausal precompensator $\bar{l}: \Lambda U \to \Lambda U$ be written as $\bar{l}^{-1} = L + G\bar{f}$, where L and G are static maps. It was shown there that a necessary and sufficient condition for the static state feedback problem to have a solution is that

$$\bar{l}^{-1}(\ker \tilde{f}) \subset \Omega^+ U, \tag{5.9}$$

where $\tilde{f}: \Omega^+ U \to \Gamma^+ Y$ is the restricted i/s map associated with $\bar{f}$. We now turn to the more general question of static output (rather than state) feedback. As we have been doing throughout this paper, we focus our attention on the *static* factorization problem which is characterized in the following

THEOREM 5.10. *Let $\bar{f}: \Lambda U \to \Lambda Y$ and $\bar{h}: \Lambda U \to \Lambda W$ be ΛK-linear maps. There exists a static ΛK-linear map $G: \Lambda Y \to \Lambda W$ such that $\bar{h} = G \cdot \bar{f}$ if and only if* $\ker \bar{p}_1 \cdot \bar{f} \subset \ker \bar{p}_1 \cdot \bar{h}$.

Proof. Assume first that G exists so that $\bar{h} = G \cdot \bar{f}$. Then $u \in \ker \bar{p}_1 \cdot \bar{f}$ implies that $\bar{p}_1 \cdot \bar{f}(u) = 0$, whence $\bar{p}_1 \cdot \bar{h}(u) = \bar{p}_1 \cdot G \cdot \bar{f}(u) = G \cdot \bar{p}_1 \cdot \bar{f}(u) = 0$, so that $u \in \ker \bar{p}_1 \cdot \bar{h}$.

Conversely, assume that $\ker \bar{p}_1 \cdot \bar{f} \subset \ker \bar{p}_1 \cdot \bar{h}$. This implies the existence of a K-linear map $G: Y \to W$ such that $\bar{p}_1 \cdot \bar{h} = G \cdot \bar{p}_1 \cdot \bar{f}$. By definition of static maps (see (2.18)), we have that $G \cdot \bar{p}_1 = \bar{p}_1 \cdot G$ so that $\bar{p}_1(\bar{h} - G \cdot \bar{f}) = 0$. That this implies $\bar{h} = G\bar{f} = 0$ is seen as follows. Suppose to the contrary that $(\bar{h} - G \cdot \bar{f})(u) = \sum_{t \in \mathbb{Z}} y_t z^{-t} \neq 0$ for some $u \in \Lambda U$. Then there exists $k \in \mathbb{Z}$ such that $y_k \neq 0$. Let $\hat{u} = z^{k-1}u$ and note that $p_1(\bar{h} - G\bar{f})(\hat{u}) = p_1 \sum_{t \in \mathbb{Z}} y_t z^{-t+k-1} = y_k \neq 0$, a contradiction. □

We shall conclude the present discussion by specializing our static factorization results to the case of linear i/s maps. We need the following lemma.

LEMMA 5.11. *Let $\bar{f}: \Lambda U \to \Lambda Y$ be an injective extended linear* i/s *map. Then* $\ker \pi^+ \bar{f} \subset \Omega^+ U$.

Proof. Let $u \in \ker \pi^+ \bar{f}$ be any element. Then $\bar{f}(u) \in \Omega^+ Y$ so that $\bar{p}_1 \cdot \bar{f}(u) = 0$. Write $u = u^+ + u^-$, where $u^+ \in \Omega^+ U$ and $u^- \in z^{-1}\Omega^- U$. Then by the strict causality of $\bar{f}$ it follows that $\bar{f}(u^-) \in z^{-2}\Omega^- Y$ and $\bar{p}_1 \cdot \bar{f}(u^-) = 0$. Hence $\bar{p}_1 \cdot \bar{f}(u^+) = \bar{p}_1 \cdot \bar{f}(u) - \bar{p}_1 \cdot \bar{f}(u^-) = 0$ and $u^+ \in \ker \bar{p}_1 \cdot \bar{f} \cdot j^+ = \ker f = \ker \tilde{f}$, the last equality following from the i/s property of $\bar{f}$. We conclude that $\bar{f}(u^+) \in \Omega^+ Y$ so that also $\bar{f}(u^-) = \bar{f}(u) - \bar{f}(u^+) \in \Omega^+ Y$. Hence $\bar{f}(u^-) \in \Omega^+ Y \cap z^{-2}\Omega^- Y = 0$ and, by the injectivity of $\bar{f}$, $u^- = 0$ concluding the proof. □

COROLLARY 5.12. *Let $\bar{f}: \Lambda U \to \Lambda Y$ be an injective extended linear* i/s *map and let $\bar{h}: \Lambda U \to \Lambda W$ be a strictly causal ΛK-linear map. Then there exists a static map G: $\Lambda Y \to \Lambda W$ such that $\bar{h} = G \cdot \bar{f}$ if and only if* $\ker \pi^+ \bar{f} \subset \ker \pi^+ \bar{h}$.

Proof. If G exists such that $\bar{h} = G \cdot \bar{f}$, then $u \in \ker \pi^+ \bar{f}$ implies that $\bar{f}(u) \in \Omega^+ Y$, so that $\bar{h}(u) = G \cdot \bar{f}(u) \in \Omega^+ W$ and $u \in \ker \pi^+ \bar{h}$. Conversely, suppose $\ker \pi^+ \bar{f} \subset \ker \pi^+ \bar{h}$. We will show that this implies that $\ker \bar{p}_1 \cdot \bar{f} \subset \ker \bar{p}_1 \cdot \bar{h}$, from which the existence of G is insured by Theorem 5.10. Let $u \in \ker \bar{p}_1 \cdot \bar{f}$ be any element and write $u = u^+ + u^-$, where $u^+ \in \Omega^+ U$ and $u^- \in z^{-1}\Omega^- U$. Then, by strict causality of both $\bar{f}$ and $\bar{h}$ it follows that $\bar{f}(u^-) \in z^{-2}\Omega^- Y$ and $\bar{h}(u^-) \in z^{-2}\Omega^- W$ yielding $\bar{p}_1 \bar{f}(u^-) = 0$ and $\bar{p}_1 \bar{h}(u^-) = 0$. Hence, $u^+ = u - u^- \in \ker \bar{p}_1 \bar{f}$ so that $u^+ \in \ker f = \ker \tilde{f}$, the last equality following from the i/s property of $\bar{f}$. Consequently $u^+ \in \ker \tilde{f} \subset \ker \pi^+ \bar{f} \subset \ker \pi^+ \bar{h} \subset \ker \bar{p}_1 \bar{h}$, the last inclusion holding by definition. Thus $u = u^+ + u^- \in \ker \bar{p}_1 \bar{h}$, and the proof is complete. □

Let $\bar{f}: \Lambda U \to \Lambda Y$ be a reachable linear i/s map. Let $\bar{l}: \Lambda U \to \Lambda U$ be a bicausal isomorphism and write $\bar{l}^{-1} = L + \bar{h}$, where L is static and $\bar{h}$ is strictly causal. Corollary 5.12 can then be interpreted as a solvability condition of the static state feedback problem. Clearly, the condition of the corollary must be equivalent with condition (5.9) which was obtained in Hautus and Heymann [1978]. We shall see next (Theorem 5.14 below) that this is indeed the case. We require the following lemma.

LEMMA 5.13. *Let $\bar{f}: \Lambda U \to \Lambda Y$ be an extended linear* i/s *map and let $\bar{h}: \Lambda U \to \Lambda W$ be a strictly causal ΛK-linear map. Then* $\ker \tilde{f} \subset \ker \tilde{h}$ *only if* $\ker \bar{f} \subset \ker \bar{h}$.

Proof. Assume that $\ker \bar{f} \not\subset \ker \bar{h}$ and let $u \in \ker \bar{f}$ satisfy $\bar{h}(u) \neq 0$. Then there exists $k \in \mathbb{Z}$ such that $\pi^+ \bar{h}(z^k u) \neq 0$ so that by the strict causality of $\bar{h}$ we have that $0 \neq \mathscr{S}^+(z^k u) \in \Omega^+ U$ and $\pi^+ \bar{h}(\mathscr{S}^+(z^k u)) = \tilde{h}(\mathscr{S}^+(z^k u)) \neq 0$. However, $\bar{f}(z^k u) = 0$ and upon application of Proposition 5.6 we also have that $\bar{f}(\mathscr{S}^+(z^k u)) = 0$, whence $\mathscr{S}^+(z^k u) \in \ker \tilde{f}$. Thus $\ker \tilde{f} \not\subset \ker \tilde{h}$ and the proof is complete. □

THEOREM 5.14. *Let $\bar{f}: \Lambda U \to \Lambda Y$ be a reachable extended linear* i/s *map. Let $\bar{l}: \Lambda U \to \Lambda U$ be a bicausal ΛK-linear map and write $\bar{l}^{-1} = L + \bar{h}$ where L is static and $\bar{h}$ is strictly causal. Then* $\ker \pi^+ \bar{f} \subset \ker \pi^+ \bar{h}$ *if and only if* $\bar{l}^{-1}(\ker \tilde{f}) \subset \Omega^+ U$.

Proof. Suppose $\ker \pi^+ \bar{f} \subset \ker \pi^+ \bar{h}$. Let $u \in \ker \tilde{f}$ be any element. Then $u \in \ker \pi^+ \bar{h}$, and since $u \in \Omega^+ U$ we also have that $u \in \ker \pi^+ L$. Hence $u \in (\ker \pi^+ \bar{h}) \cap (\ker \pi^+ L) \subset \pi^+(\bar{h} + L) = \ker \pi^+ \bar{l}^{-1}$ so that $\bar{l}^{-1}(u) \in \Omega^+ U$. Conversely,

assume that $\bar{l}^{-1}(\ker \tilde{f}) \subset \Omega^+ U$. This immediately implies that $\ker \tilde{f} \subset \ker \tilde{h}$ whence, by Lemma 5.13, $\ker \bar{f} \subset \ker \bar{h}$. Now let $u \in \ker \pi^+ \bar{f}$ and write $u = u^+ + u^-$ with $u^+ \in \Omega^+ U$ and $u^- \in z^{-1}\Omega^- U$. Then $\bar{f}(u^-) \in z^{-2}\Omega^- Y$, and since $\bar{f}(u) \in \Omega^+ Y$ we conclude that $\bar{p}_1 \cdot \bar{f}(u^+) = 0$. This implies that $u^+ \in \ker f = \ker \tilde{f}$ (with the equality holding since $\bar{f}$ is an i/s map) so that $u^+ \in \ker \tilde{h} \subset \ker \pi^+ \bar{h}$. Finally, $u^+ \in \ker \tilde{f}$ implies that $\bar{f}(u^+) \in \Omega^+ U$ whence $\bar{f}(u^-) = \bar{f}(u) - \bar{f}(u^+) \in \Omega^+ Y$. But then $\bar{f}(u^-) \in \Omega^+ Y \cap z^{-2}\Omega^- Y = 0$, so that $u^- \in \ker \bar{f} \subset \ker \bar{h}$, and hence $u^- \in \ker \pi^+ \bar{h}$. This implies that $u = u^+ + u^- \in \ker \pi^+ \bar{h}$, concluding the proof. □

6. Factorization invariants—explicit calculation. Throughout this section we shall assume that $U = K^m$ and $Y = K^p$, and we shall study properties of ΛU as an $\Omega^- K$-module as well as properties of submodules thereof.

The ring $\Omega^- K$ is of course a principal ideal domain, and clearly also a Euclidean domain. The units of $\Omega^- K$ are precisely those elements whose order is zero and each element $0 \neq \alpha \in \Omega^- K$ can be expressed as

$$\alpha = z^{-\operatorname{ord}\alpha}\alpha_0,$$

where $\alpha_0 \in \Omega^- K$ is a unit. It is clear, therefore, that all the ideals of $\Omega^- K$ are of the form (z^{-k}), forming a chain with (z^{-1}) being the unique maximal ideal and the only prime. Thus, the ring $\Omega^- K$ is also a local ring and $\Omega^- K/(z^{-1})$ is a field, isomorphic to the field $\mathcal{K}_0$ which consists of the units of $\Omega^- K$ augmented by zero. We shall make use of the special properties of the ring $\Omega^- K$ in the ensuing discussion.

For a fixed integer k, consider the subset $z^{-k}\Omega^- U \subset \Lambda U$. Clearly, this subset is an $\Omega^- K$ submodule of ΛU. Moreover, while ΛU itself is not a finitely generated $\Omega^- K$-module, the submodule $z^{-k}\Omega^- U$ is (and hence is a free module). In fact, it is readily noted that $\operatorname{rank}_{\Omega^- K} z^{-k}\Omega^- U = \dim_{\Lambda K} \Lambda U = \dim_K U$. Indeed, if $\{e_1, \cdots, e_m\}$ is a basis for U (as well as for ΛU), then $\{z^{-k}e_1, \cdots, z^{-k}e_m\}$ is a basis (i.e., a free generator) for $z^{-k}\Omega^- U$.

Let $0 \neq \Delta \subset \Lambda U$ be an $\Omega^- K$-submodule. We say that Δ is *of finite order* if there exists a finite integer k such that $\Delta \subset z^{-k}\Omega^- U$. The maximal integer k for which the above holds, and which is the least order of elements in Δ, is denoted k_Δ and is called the *order* of Δ. We define the order of the zero module as infinity. We have the following:

PROPOSITION 6.1. *Let $0 \neq \Delta \subset \Lambda U$ be an $\Omega^- K$-submodule. Then Δ is finitely generated if and only if it has finite order.*

Proof. If Δ has finite order there exists a finite integer k such that Δ is a submodule of $z^{-k}\Omega^- U$ which is, of course, finitely generated. Since $\Omega^- K$ is a principal ideal domain, Δ is then also finitely generated. Conversely, if Δ is finitely generated, say by elements $d_1, \cdots, d_m \in \Delta$, then clearly $\Delta \subset z^{-k_\Delta}\Omega^- U$, where $k_\Delta := \min\{\operatorname{ord} d_i, i = 1, \cdots, m\}$. □

Let $\Delta \subset \Lambda U$ be a finitely generated $\Omega^- K$-submodule. Then, by Proposition 6.1, it is of finite order and hence $\operatorname{rank} \Delta \leqq \dim U\ (= m)$. Let Δ be of rank n and let $d_1, \cdots, d_n$ be a basis for Δ. Define the $\Omega^- K$-homomorphism D: $\Omega^- K^n \to \Delta$ by $De_i = d_i$, $i = 1, \cdots, n$, where $e_1, \cdots, e_n$ denotes the natural basis for K^n (as well as for $\Omega^- K^n$). We can view D also as a matrix with entries in ΛK by regarding $d_i \in \Lambda K^m (= \Lambda U)$ as the ith column of D. Conversely, if D is an $m \times n$ matrix with entries in ΛK, we can regard D as an $\Omega^- K$-homomorphism $\Omega^- K^n \to \Lambda U$: $e_i \mapsto d_i$, $i = 1, \cdots, n$, where $d_i \in \Lambda U$ is the ith column of D. The image $\Delta = D\Omega^- K^n := \{Dw \mid w \in \Omega^- K^n\}$ is an $\Omega^- K$-submodule of ΛU. Clearly, $\operatorname{rank} \Delta = \operatorname{rank} D$, where $\operatorname{rank} D$ is the matrix rank of D over the ring $\Omega^- K$ (or over ΛK).

Consider now the special case when $n = m$ (that is, $K^m = U$) and let D be a nonsingular $m \times m$ matrix with entries in ΛK. Then D defines, as above, an $\Omega^- K$-homomorphism $\Omega^- U \to \Lambda U$ and also (when simply regarded as a transfer function) a ΔK-linear map $\Lambda U \to \Lambda U$. Denoting both maps by the same symbol D, it is readily verified that the diagram in Fig. 6.1 is commutative,

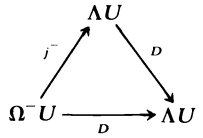


FIG. 6.1.

where j^- denotes the canonical injection. Since the matrix D is nonsingular, the ΛK-linear map D is invertible. We shall say that the matrix D is *bicausal* if the associated ΛK-linear map is bicausal, i.e., if the entries of D are in $\Omega^- K$ and its determinant is a unit in this ring (that is, has order zero). In analogy we shall say that a matrix D is *strictly causal* or *causal* if so is the associated ΛK-linear map. Finally, an $\Omega^- K$-submodule $\Delta = D\Omega^- U \subset \Lambda U$ is called a *full* submodule if rank $\Delta = m$, i.e., if the matrix D is nonsingular.

THEOREM 6.2. *Let $\Delta_1, \Delta_2 \subset \Lambda U$ be finitely generated $\Omega^- K$-submodules given by $\Delta_1 = D_1\Omega^- U$ and $\Delta_2 = D_2\Omega^- U$. Then $\Delta_2 \subset \Delta_1$ if and only if there exists a causal matrix R (i.e., with entries in $\Omega^- K$) such that $D_2 = D_1 R$.*

The proof of Theorem 6.2 is elementary and will be omitted. The following corollary will be useful in the sequel.

COROLLARY 6.3. *Let $\Delta_1, \Delta_2 \subset \Lambda U$ be finitely generated $\Omega^- K$-submodules given by $\Delta_1 = D_1\Omega^- U$ and $\Delta_2 = D_2\Omega^- U$. Assume that Δ_1 is full and define $R := D_1^{-1}D_2$. Then $\Delta_2 \subseteq \Delta_1$ if and only if R is causal with equality if and only if R is bicausal.*

Let $\Delta \subset \Lambda U$ be a finitely generated $\Omega^- K$-submodule of rank n and order k_Δ. Then for all integers $j \leqq k_\Delta$, $\Delta \subset z^{-j}\Omega^- U$ and for each integer $j \geqq k_\Delta$ we define the submodule $\Delta_j \subset \Delta$ by

$$\Delta_j := \Delta \cap z^{-j}\Omega^- U. \tag{6.4}$$

Clearly $z^{-j}\Omega^- U \subset z^{-k}\Omega^- U$ for all $j \geqq k$, and it follows that

$$\Delta = \Delta_{k_\Delta} \supset \Delta_{k_\Delta+1} \supset \cdots \supset \Delta_j \supset \Delta_{j+1} \cdots. \tag{6.5}$$

As an immediate consequence of the fact that if $u \in \Delta_j$ then $z^{-1}u \in \Delta_{j+1}$, it is clear that rank Δ = rank Δ_j for all j and the quotient modules

$$\mathscr{D}_i := \Delta_j / \Delta_{j+1} \tag{6.6}$$

are all torsion modules with z^{-1} as annihilators, that is, for each j and for each $[u] \in \mathscr{D}_j$, $z^{-1}[u] = 0$. Next we shall show that the sequence of quotient modules $\{\mathscr{D}_j\}$ is isomorphic to a chain $\{S_j\}$ of (finite dimensional) K-linear subspaces of U, that is, each $\mathscr{D}_j$ is isomorphic to a subspace $S_j \subset U$ and

$$0 = S_{k_\Delta - 1} \subset S_{k_\Delta} \subset S_{k_\Delta+1} \subset \cdots \subset S_j \subset \cdots \subset U. \tag{6.7}$$

Indeed, each element in $\mathscr{D}_j$ is an equivalence class $[u]$ of elements in Δ_j. A representative $u \in [u]$ can be expressed as $u = \sum_{k=j}^{\infty} u_k z^{-k}$. If $u' = \sum_{k=j}^{\infty} u'_k z^{-k}$ and $u'' = \sum_{k=j}^{\infty} u''_k z^{-k}$ are any two elements in the same equivalence class $[u]$ then, since $u' - u'' \in \Delta_{j+1}$, it follows that $u'_j = u''_j$. Thus, with each equivalence class $[u]$ is associated a unique leading coefficient u_j (of z^{-j}). We can now define the map $\gamma_j : \mathscr{D}_j \to U : [u] \mapsto u_j$. Naturally the

map γ_j is K-linear since $\gamma_j([u]+[u'])=\gamma_j([u+u'])=u_j+u_j'$ and $\gamma_j(\alpha[u])=\gamma_j([\alpha u])=\alpha u_j$. It is also clear that γ_j is injective, since $\ker \gamma_j=\Delta_{j+1}=[0]$. Now, for each integer j, we define $S_j := \operatorname{Im}(\gamma_j)$. Clearly S_j is then K-linearly isomorphic to $\mathscr{D}_j$ and $S_j \subset S_{j+1}$ with $S_{k_\Delta - j-1}=0$ for all $j \geqq 0$. Also, by the finite dimensionality of U, there exists an integer $k^\Delta (\geqq k_\Delta)$ such that $S_{k^\Delta-1} \neq S_{k^\Delta}$ and $S_{k^\Delta+j}=S_{k^\Delta}$ for all $j \geqq 0$. We call the chain $\{S_j\}$ the *order chain* of Δ, and the sequence of integers $\{\mu_j\}$, $\mu_j := \dim S_j$, we call the *order list* of Δ. In the special case when $\Delta=\ker \pi^- \bar{f}$ where $\bar{f}$ is a linear i/o map, we refer to the order chain and the order list of Δ, respectively, also as the *latency chain* and *latency list* of $\bar{f}$.

It is interesting to observe that the integer k^Δ is also the least integer satisfying the condition that $z^{-1}\Delta_j=\Delta_{j+1}$ for all $j \geqq k^\Delta$. Indeed, we have seen that $z^{-1}\Delta_j \subset \Delta_{j+1}$ for all j. To see that $z^{-1}\Delta_j \supset \Delta_{j+1}$ if and only if $j \geqq k^\Delta$, let $u=\sum_{k=j+1}^{\infty} u_k z^{-k} \in \Delta_{j+1}$ be any element. Then we can write $u=z^{-1}u'$ where $u'=\sum_{k=j}^{\infty} u_{k+1} z^{-k} \in z^{-j}\Omega^- U$, and clearly $u \in z^{-1}\Delta_j$ if and only if $u' \in \Delta_j$. This can hold for every $u \in \Delta_{j+1}$ only if $S_{j+1}=S_j$, whence the necessity that $j \geqq k^\Delta$. The sufficiency of the condition is an immediate consequence of Theorem 6.11 below.

Next we have the following useful result.

LEMMA 6.8. *Let $\Delta \subset \Lambda U$ be a finitely generated $\Omega^- K$-submodule with order chain $\{S_j\}$ and order list $\{\mu_i\}$. Then* $\dim S_{k^\Delta}=\operatorname{rank}\Delta$.

Proof. Let $\operatorname{rank}\Delta=\mu$, let $d_1, \cdots, d_\mu$ be a basis of Δ and define $\mathscr{R} := \operatorname{span}_{\Lambda K}\{d_1, \cdots, d_\mu\}$. It is easily seen that $\mathscr{R}$ is the smallest ΛK-linear space containing Δ and $\dim_{\Lambda K}\mathscr{R}=\operatorname{rank}\Delta$. The ΛK-linear space $\mathscr{R}$ has a proper basis and (by Corollary 4.5) $\dim_{\Lambda K}\mathscr{R}=\dim_K \hat{\mathscr{R}}$. But clearly $\hat{\mathscr{R}}=S_{k^\Delta}$ and the proof is complete. □

Let $\{S_j\}$ and $\{S_j'\}$ be the order chains and $\{\mu_j\}$ and $\{\mu_j'\}$ the order lists, respectively, of submodules Δ and Δ' of ΛU. We shall say that $\{S_j'\}$ is a *subchain* of $\{S_j\}$, denoted $\{S_j'\} \subset \{S_j\}$ if, for all j, $S_j' \subset S_j$. Similarly we say that the list $\{\mu_j'\}$ is *smaller* than the list (μ_j), denoted $\{\mu_j'\} \leqq \{\mu_j\}$ if $\mu_j' \leqq \mu_j$ for all integers j. As an immediate consequence of the definition we have the following,

PROPOSITION 6.9. *Let $\Delta,\Delta' \subset \Lambda U$ be $\Omega^- K$-submodules with order chains $\{S_j\}$ and $\{S_j'\}$ and order lists $\{\mu_j\}$ and $\{\mu_j'\}$, respectively. If $\Delta' \subset \Delta$ then $\{S_j'\} \subset \{S_j\}$ and $\{\mu_j'\} \leqq \{\mu_j\}$.*

Let $\Delta \subset \Lambda U$ be a finitely generated $\Omega^- K$-submodule. A set of elements $d_1, \cdots, d_k \in \Delta$ is called *properly free* if the elements are properly independent as elements of ΛU (regarded as a ΛK-linear space), that is, if the leading coefficients $\hat{d}_1, \cdots, \hat{d}_k$ are K-linearly independent. It is then clear that if $d_1, \cdots, d_k$ are properly free they are also *free* (i.e. independent over the ring $\Omega^- K$).

DEFINITION 6.10. Let $\Delta \subset \Lambda U$ be a finitely generated $\Omega^- K$-submodule. A basis $d_1, \cdots, d_\mu$ of Δ is called *proper* if $d_1, \cdots, d_\mu$ are properly free. The basis will be called *ordered* if $\operatorname{ord} d_{i+1} \geqq \operatorname{ord} d_i$ for all $i=1, \cdots, \mu-1$.

THEOREM 6.11. *Let $\Delta \subset \Lambda U$ be an $\Omega^- K$-submodule or rank μ and of order k_Δ, with order chain $\{S_j\}$ and order list $\{\mu_j\}$. Then (i) there exists an ordered proper basis for Δ. (ii) If $d_1, \cdots, d_\mu$ is any ordered proper basis for Δ, then the following conditions are satisfied:*

$$\operatorname{ord} d_j = i \quad \text{for } \mu_{i-1} < j \leqq \mu_i \text{ and } i = k_\Delta, k_{\Delta+1}, \cdots. \tag{6.12}$$

For each $j=1, \cdots, \mu$, the set $\hat{d}_1, \cdots, \hat{d}_j \in S_i$, where i is the least integer such that $j \leqq \mu_i$. (6.13)

Proof. (i) We shall construct an ordered proper basis for Δ which, in particular, satisfies (6.12) and (6.13). Consider the sequence $\{\mathscr{D}_j\}$ of quotient modules $\mathscr{D}_j$ defined by (6.6), of which $\mathscr{D}_{k_\Delta}$ is the first nonzero one. Choose any equivalence class $0 \neq [d_1] \in \mathscr{D}_{k_\Delta}$ and let $d_1 \in \Delta$ be any representative of $[d_1]$. Then $\operatorname{ord} d_1 = k_\Delta$ and d_1 is clearly properly free. We proceed stepwise and assume that for $j>0$, $d_1, \cdots, d_j$ are properly

free elements of Δ satisfying (6.12) and (6.13). If $j<\mu$, let k denote the least integer such that $j<\mu_k$. Then $\hat{d}_1, \cdots, \hat{d}_j \in S_k$ are K-linearly independent, but they do not span S_k, since $\dim S_k = \mu_k$. Thus, there exists an element $[d_{j+1}] \in \mathscr{D}_k$ such that for any representative $d_{j+1} \in [d_{j+1}]$, the set $\hat{d}_1, \cdots, \hat{d}_j, \hat{d}_{j+1} \in S_k$ are K-linearly independent and hence the set $d_1, \cdots, d_{j+1}$ is properly free. Clearly (6.13) is satisfied, and since $\operatorname{ord} d_{j+1} = k$ so is also (6.12). By Lemma 6.8, $\dim S_{k^\Delta} = \operatorname{rank} \Delta = \mu$, so that we finally obtain an ordered, properly free set of elements $d_1, \cdots, d_\mu \in \Delta$ satisfying (6.12) and (6.13). Let Δ' denote the $\Omega^- K$-submodule of ΛU generated by $d_1, \cdots, d_\mu$. It remains to be shown that $\Delta' = \Delta$. Obviously $\Delta' \subset \Delta$ and since $\operatorname{ord} d_i \leqq k^\Delta$ for all $i = 1, \cdots, \mu$ and since $\operatorname{span}_K\{\hat{d}_1, \cdots, \hat{d}_\mu\} = S_{k^\Delta}$, it follows also that $\Delta_{k^\Delta} \subset \Delta'$. Let $u \in \Delta$ be any element and let $\operatorname{ord} u = j$. Then $\hat{u} \in S_j$ whence there are elements $\alpha_1, \cdots, \alpha_{\mu_j} \in \Omega^- K$ such that $\sum_{k=1}^{\mu_j} \hat{\alpha}_k \hat{d}_k = \hat{u}$ and $\operatorname{ord}(u - \sum_{k=1}^{\mu_j} \alpha_k d_k) > j$. Proceeding stepwise the same way, we conclude that there are elements $\alpha_1, \cdots, \alpha_\mu \in \Omega^- K$ such that $u = \sum_{i=1}^{\mu} \alpha_i d_i + u'$, with $\operatorname{ord} u' \geqq k^\Delta$. Clearly, $\sum_{i=1}^{\mu} \alpha_i d_i \in \Delta'$, and since $u' \in \Delta_{k^\Delta} \subset \Delta'$, it follows also that $u \in \Delta'$ and the proof of (i) is complete. To see that (ii) holds, it suffices to observe that for each integer j, every ordered proper basis $d_1, \cdots, d_\mu$ of Δ has precisely μ_j elements whose order is less than or equal to j and $\operatorname{span}_K \{\hat{d}_1, \cdots, \hat{d}_{\mu_j}\} = S_j$. □

The following immediate corollary to Theorem 6.11 gives a sharp insight to the relation between ordered proper bases of $\Omega^- K$-modules and their order chain.

COROLLARY 6.14. *Let $\Delta \subset \Lambda U$ be an $\Omega^- K$-submodule of rank μ with order chain $\{S_j\}$ and order list $\{\mu_j\}$. Then $d_1, \cdots, d_\mu$ is an ordered proper basis of Δ if and only if for each j, $\hat{d}_1, \cdots, \hat{d}_{\mu_j}$ is a basis for S_j.*

We now return to questions connected with our primary objective of studying causal factorization and feedback. First we have some preliminary facts.

LEMMA 6.15. *Let U be an m-dimensional K-linear space and let $\bar{f}: \Lambda U \to \Lambda Y$ be a ΛK-linear map. For each integer j let $\Delta_j(\bar{f})$ be the $\Omega^- K$-subodule of ΛU defined by $\Delta_j(\bar{f}) := \ker \pi^- \bar{f} \cap z^{-j} \Omega^- U$. Then $\operatorname{rank} \Delta_j(\bar{f}) = m$.*

Proof. First note that since $\Delta_j(\bar{f}) \subset z^{-j}\Omega^- U$, $\operatorname{rank} \Delta_j(\bar{f}) \leqq m$, with equality obviously holding when $\bar{f} = 0$, since then $\ker \pi^- \bar{f} = \Lambda U$. Assume now that $\bar{f} \neq 0$, define $t := \max\{j\text{-ord } \bar{f}, \ -\text{ord } \bar{f}\}$ and let $u \in z^{-t}\Omega^- U$ be any element. Then $\operatorname{ord} \bar{f}u \geqq \operatorname{ord} \bar{f} + \operatorname{ord} u \geqq \operatorname{ord} \bar{f} + t \geqq \max\{j, 0\}$ and $u \in \Delta_j(\bar{f})$. Hence $z^{-t}\Omega^- U \subset \Delta_j(\bar{f})$ so that $\operatorname{rank} \Delta_j(\bar{f}) \geqq m$ and the proof is complete. □

PROPOSITION 6.16. *Let U be an m-dimensional K-linear space and let $\bar{f}: \Lambda U \to \Lambda Y$ be a ΛK-linear map. Then the following are equivalent*

(i) *$\bar{f}$ is injective.*

(ii) *$\ker \pi^- \bar{f}$ is finitely generated.*

(iii) *$\operatorname{rank} \ker \pi^- \bar{f} = m$.*

Proof. That (ii) and (iii) are equivalent follows immediately from Lemma 6.15 and the fact that if $\ker \pi^- \bar{f}$ is finitely generated it is of finite order, say t, so that $\ker \pi^- \bar{f} = \Delta_t(\bar{f})$. To see that (ii) implies (i), recall that $\ker \bar{f} \subset \ker \pi^- \bar{f}$ so that if $\ker \bar{f} \neq 0$ then $\ker \pi^- \bar{f}$ is not of finite order and hence is not finitely generated. It remains to be shown that (i) implies (ii). Assume that (i) holds, let $y_1, \cdots, y_m$ be a normalized proper basis for $\operatorname{Im} \bar{f} \subset \Lambda Y$ and let $u_1, \cdots, u_m$ be the (unique) elements of ΛU satisfying $\bar{f}(u_i) = y_i$, $i = 1, \cdots, m$. The proof will be complete upon showing that $\ker \pi^- \bar{f}$ is of finite order and, in fact, we claim that $\ker \pi^- \bar{f} \subset z^{-t}\Omega^- U$ where $t := \min\{\operatorname{ord} u_i | i = 1, \cdots, m\}$. Indeed, if $u \in \ker \pi^- \bar{f}$, then $\bar{f}(u) \in \Omega^- Y$ and there are elements $\alpha_1, \cdots, \alpha_m \in \Omega^- K$ such that $\bar{f}(u) = \sum_{i=1}^m \alpha_i y_i = \sum_{i=1}^m \alpha_i \bar{f}(u_i) = \bar{f}(\sum_{i=1}^m \alpha_i u_i)$, whence $u = \sum_{i=1}^m \alpha_i u_i$ so that $\operatorname{ord} u \geqq t$. □

In view of Proposition 6.16, it follows that the latency kernel of a given linear i/o map $\bar{f}$ is finitely generated if and only if $\bar{f}$ is injective, the case which receives, of course,

most of our attention. Before proceeding further, a remark on the noninjective case is in order.

Remark 6.17. It is readily noted that if $\bar{f}: \Lambda U \to \Lambda Y$ is a ΛK-linear map, then $\ker \pi^- \bar{f}$ can (always) be written as

$$\ker \pi^- \bar{f} = \ker \bar{f} + \mathcal{R},$$

where $\mathcal{R}$ is a finitely generated full $\Omega^- K$-submodule of ΛU. However, in the above representation, $\mathcal{R}$ is nonunique except in the special case when $\bar{f}$ is injective and $\ker \bar{f} = 0$. If $\bar{f}_1$ and $\bar{f}_2$ are two ΛK-linear maps then $\ker \pi^- \bar{f}_1 \subset \ker \pi^- \bar{f}_2$ if and only if $\ker \bar{f}_1 + \mathcal{R}_1 \subset \ker \bar{f}_2 + \mathcal{R}_2$. While this condition necessarily implies $\ker \bar{f}_1 \subset \ker \bar{f}_2$, it cannot be claimed, except in the injective case, that $\mathcal{R}_1 \subset \mathcal{R}_2$. Hence, for computational purposes it is convenient in the noninjective case to resort to the fact that $\ker \pi^- \bar{f}_1 \subset \ker \pi^- \bar{f}_2$ if and only if $\Delta_j(\bar{f}_1) \subset \Delta_j(\bar{f}_2)$ for all j, where $\Delta_j(\bar{f}_i)$ is as defined in Lemma 6.15. However, $\Delta_j(\bar{f}_1) \subset \Delta_j(\bar{f}_2)$ for all j if and only if $\Delta_j(\bar{f}_1) \subset \Delta_j(\bar{f}_2)$ for any $j \leqq \min\{\operatorname{ord} \mathcal{R}_1, \operatorname{ord} \mathcal{R}_2\}$ where $\mathcal{R}_i$, $i = 1, 2$, are any submodules in the corresponding representations of $\ker \pi^- \bar{f}_i$. By Lemma 6.15 both $\Delta_j(\bar{f}_1)$ and $\Delta_j(\bar{f}_2)$ are full finitely generated $\Omega^- K$-submodules of ΛU so that the situation is thus similar to that in the injective case. □

Let $\bar{f}: \Lambda U \to \Lambda Y$ be an injective extended linear i/o map and let $\Delta = \ker \pi^- \bar{f}$. Then $\Delta = D\Omega^- U$ is a full, finitely generated $\Omega^- K$-submodule of ΛU and the columns $d_1, \cdots, d_m$ of the generating matrix D form a basis of Δ. We shall next establish certain properties of possible selections of the matrix D.

PROPOSITION 6.18. *Let $\bar{f}: \Lambda U \to \Lambda Y$ be an injective extended linear i/o map. Write $\ker \pi^- \bar{f} = D\Omega^- U$. Then D^{-1} exists and is strictly causal; i.e., the elements of D^{-1} are in $z^{-1}\Omega^- K$.*

Proof. The existence of D^{-1} follows immediately from Proposition 6.16. From the strict causality of $\bar{f}$ it follows that $z\Omega^- U \subset \ker \pi^- \bar{f}$, whence by Theorem 6.2 there exists a causal matrix R such that $zI = DR$. Thus $D^{-1} = z^{-1}R$ and $z^{-1}R$ is clearly strictly causal. □

Let $\Delta \subset \Lambda U$ be a full finitely generated $\Omega^- K$-submodule and write $\Delta = D\Omega^- U$. We call the columns $d_1, \cdots, d_m$ of D a *polynomial basis* of Δ if the matrix D is a polynomial matrix, i.e., with elements in $\Omega^+ K$. We call the basis a *strictly polynomial basis* if its elements are strict polynomials, i.e., with elements in $z\Omega^+ K$. If in addition D is a proper basis we call it a *proper polynomial basis*, respectively, *proper strictly polynomial basis* for Δ.

THEOREM 6.19. *Let $\bar{f}: \Lambda U \to \Lambda Y$ be an injective extended linear* i/o *map. Then $\ker \pi^- \bar{f}$ has a proper strictly polynomial basis.*

Proof. Let $\tilde{d}_1, \cdots, \tilde{d}_m$ be a proper basis for $\ker \pi^- \bar{f}$ and for each i write $\tilde{d}_i = \Sigma d_{ij} \cdot z^{-j} = d_i + d_i^-$, where $d_i = \sum_{i<0} d_{ij} z^{-j} \in z\Omega^+ U$ and $d_i^- = \sum_{j \geqq 0} d_{ij} z^{-j} \in \Omega^- U$. Then $zd_i^- \in z\Omega^- U \subset \ker \pi^- \bar{f}$, the inclusion following from the strict causality of $\bar{f}$. Thus, there are elements $\alpha_{ij} \in \Omega^- K$, $j = 1, \cdots, m$, so that $zd_i^- = \sum_{j=1}^m \alpha_{ij} \tilde{d}_j$. Defining the matrices $D := [d_1, \cdots, d_m]$, $\tilde{D} := [\tilde{d}_1, \cdots, \tilde{d}_m]$ and $A := [\alpha_{ji}]$ we can thus write $\tilde{D} = D + z^{-1}\tilde{D}A$, or alternatively, $D = \tilde{D}(I - z^{-1}A)$. Since A is causal by definition of the α_{ij}, it follows that $(I - z^{-1}A)$ is a bicausal matrix. Consequently, by Corollary 6.3, we have $\ker \pi^- \bar{f} = \tilde{D}\Omega^- U = D\Omega^- U$ so that the columns $d_1, \cdots, d_m$ of D also form a proper basis for $\ker \pi^- \bar{f}$. That this basis is strictly polynomial follows directly from the definition of the d_i. □

For an injective extended linear i/o map $\bar{f}$ it is convenient to define a set of nonnegative integers, called *latency indices*, which are associated in one–one correspondence with the latency list of $\bar{f}$. We proceed as follows. Let $d_1, \cdots, d_m$ be an

ordered proper basis for $\ker \pi^- \bar{f}$. Then, as we have seen, for each $i = 1, \cdots, m$, ord $d_i \leqq -1$. We define the *latency indices* $\{\nu_1, \cdots, \nu_m\}$ of $\bar{f}$ by $\nu_i := -\text{ord } d_i - 1$. The relation of the latency indicates with the latency list is clearly established by Corollary 6.14, and if $\{\mu_i\}$ is the latency list of $\bar{f}$ then we have

$$\nu_i = -j - 1 \quad \text{for } \mu_{j-1} < i \leqq \mu_j, \quad j = k_\Delta, k_\Delta + 1, \cdots, \tag{6.20}$$

where $k_\Delta = \text{ord ker } \pi^- \bar{f}$. Clearly $\nu_i \geqq 0$ for all $i = 1, \cdots, m$, and $\bar{f}$ is nonlatent if and only if all its latency indices are zero.

We conclude this section with the discussion of certain invariance properties of the latency indices. We have seen previously that if $\bar{f}_1 : \Lambda U \to \Lambda Y$ and $\bar{f}_2 : \Lambda U \to \Lambda Y$ are two extended linear i/o maps and if $\bar{l}_{po} : \Lambda Y \to \Lambda Y$ is a ΛK-linear bicausal isomorphism such that $\bar{f}_2 = \bar{l}_{po} \cdot \bar{f}_1$, then $\bar{f}_1$ and $\bar{f}_2$ have the same latency kernels; i.e., $\ker \pi^- \bar{f}_1 = \ker \pi^- \bar{f}_2$. If there exist both a bicausal postcompensator as above and a ΛK-linear bicausal precompensator $\bar{l}_{pr} : \Lambda U \to \Lambda U$ such that $\bar{f}_2 = \bar{l}_{po} \cdot \bar{f}_1 \cdot \bar{l}_{pr}$, then $\ker \pi^- \bar{f}_2 = \ker \pi^- \bar{f}_1 \cdot \bar{l}_{pr}$, and since $u \in \ker \pi^- \bar{f}_1 \cdot \bar{l}_{pr}$ if and only if $\bar{l}_{pr} u \in \ker \pi^- \bar{f}_1$, it follows that $\bar{l}_{pr} \ker \pi^- \bar{f}_2 = \ker \pi^- \bar{f}_1$. Since the map $\bar{l}_{pr}$ is, in particular, also an $\Omega^- K$-homorphism (which we denote l_{pr}) we interpret it as an order preserving $\Omega^- K$-isomorphism $l_{pr} : \ker \pi^- \bar{f}_2 \to \ker \pi^- \bar{f}_1$. Suppose, conversely, that there exists an order preserving $\Omega^- K$-isomorphism l_{pr} as above. Fix an integer j and define (as in Lemma 6.15) $\Delta_j(\bar{f}_2) \subset \ker \pi^- \bar{f}_2$. Then, by the same lemma, $\Delta_j(\bar{f}_2)$ is a full finitely generated $\Omega^- K$-submodule of ΛU, and if $d_1, \cdots, d_m$ is a proper basis for $\Delta_j(\bar{f}_2)$, it is clearly also a basis for ΛU. Let $\bar{l}_{pr} : \Lambda U \to \Lambda U$ be the (unique) ΛK-linear map whose action on the d_i's is that of l_{pr}. Then, $\bar{l}_{pr}$ is order preserving and thus a bicausal isomorphism $\Lambda U \to \Lambda U$. Moreover, since $\bar{l}_{pr} u = l_{pr} u$ for all elements $u \in \ker \pi^- \bar{f}_2$, it follows that $\bar{l}_{pr} \ker \pi^- \bar{f}_2 = \ker \pi^- \bar{f}_1$ whence $\ker \pi^- \bar{f}_2 = \ker \pi^- \bar{f}_1 \cdot \bar{l}_{pr}$. Applying now Corollary 5.7 to the above kernel equality, we conclude that there exists a bicausal ΛK-linear postcompensator $\bar{l}_{po} : \Lambda Y \to \Lambda Y$ such that $\bar{f}_2 = \bar{l}_{po} \bar{f}_1 \bar{l}_{pr}$. We have just proved the following.

THEOREM 6.21. *Let $\bar{f}_1, \bar{f}_2 : \Lambda U \to \Lambda Y$ be two extended linear* i/o *maps with U and Y finite dimensional K-linear spaces. There exist bicausal ΛK-linear compensators $\bar{l}_{pr} : \Lambda U \to \Lambda U$ and $\bar{l}_{po} : \Lambda Y \to \Lambda Y$ such that $\bar{f}_2 = \bar{l}_{po} \cdot \bar{f}_1 \cdot \bar{l}_{pr}$ if and only if there exists an order preserving $\Omega^- K$-isomorphism $l_{pr} : \ker \pi^- \bar{f}_2 \to \ker \pi^- \bar{f}_1$.*

We now restrict Theorem 6.21 to the injective case to obtain the following invariance characterization of the latency indices.

COROLLARY 6.22. *Let $\bar{f}_1, \bar{f}_2 : \Lambda U \to \Lambda Y$ be two injective extended linear* i/o maps *with U and Y finite dimensional K-linear spaces. There exist bicausal ΛK-linear compensators $\bar{l}_{pr} : \Lambda U \to \Lambda U$ and $\bar{l}_{po} : \Lambda Y \to \Lambda Y$ such that $\bar{f}_2 = \bar{l}_{po} \cdot \bar{f}_1 \cdot \bar{l}_{pr}$ if and only if $\bar{f}_1$ and $\bar{f}_2$ have the same latency indices.*

Proof. By the injectivity of $\bar{f}_1$ and $\bar{f}_2$, both $\Delta_1 = \ker \pi^- \bar{f}_1$ and $\Delta_2 = \ker \pi^- \bar{f}_2$ are of rank m, where $m = \dim U$, and in view of Theorem 6.21 it needs only to be shown that Δ_1 and Δ_2 have the same latency indices (or latency lists) if and only if there exists an order preserving $\Omega^- K$-isomorphism $l_{pr} : \Delta_2 \to \Delta_1$. Let $d_{11}, \cdots, d_{1m}$ and $d_{21}, \cdots, d_{2m}$ be ordered proper bases for Δ_1 and Δ_2, respectively, and let D_1 and D_2 be the corresponding matrices. Then an order preserving isomorphism $l_{pr} : \Delta_2 \to \Delta_1$ exists if and only if the matrix $D_1 D_2^{-1}$ is bicausal which is easily seen to be the case if and only if ord $d_{1j} =$ ord d_{2j} for all $j = 1, \cdots, m$. Employing Corollary 6.14 completes the proof. □

Theorem 6.21 and Corollary 6.22 could, of course, have been stated for any ΛK-linear maps and not only strictly causal ones. The proofs did in no way depend on the causality properties of the maps involved. Also, Corollary 6.22 could have been obtained as an application of the existence of, so called, Smith canonical forms for matrices over Euclidean rings (see, e.g., MacDuffee [1934]).

7. Precompensation and feedback. Let $\bar{f}: \Lambda U \to \Lambda Y$ be an extended linear i/o map and let $\bar{l}: \Lambda U \to \Lambda U$ be a ΛK-linear bicausal precompensator. Write $\bar{l}^{-1} = L + h$ where $L: \Lambda U \to \Lambda U$ is static and $\bar{h}: \Lambda U \to \Lambda U$ is strictly causal. We have seen in § 5 that $\bar{l}$ can be realized by a static precompensator (i.e., coordinate change in the input value space) and output feedback around $\bar{f}$ (i.e., $\bar{h} = \bar{g} \cdot \bar{f}$ for causal ΛK-linear map $\bar{g}: \Lambda Y \to \Lambda U$) if and only if $\ker \pi^- \bar{f} \subset \ker \pi^- \bar{h}$ (see Theorem 5.2). When $\bar{f}$ is a nonlatent map, feedback realization as above is thus possible for every bicausal map $\bar{l}$. In general, however, feedback realization is not possible for every precompensator $\bar{l}$. We shall say that $\bar{l}$ has *a* $(\bar{v}, \bar{g})$ *representation* if it can be expressed as $\bar{l} = \bar{l}_{(\bar{v},\bar{g})} = (I + \bar{g}\bar{f})^{-1}\bar{v}$ where $\bar{v}: \Lambda U \to \Lambda U$ is a bicausal isomorphism and $\bar{g}: \Lambda Y \to \Lambda U$ is a causal ΛK-linear map. We call the map $\bar{v}$ in the above representation the *precompensator remainder* of the representation. The precompensator $\bar{l}$ can thus be realized as feedback whenever $\bar{l}$ has a $(\bar{v}, \bar{g})$ representation with $\bar{v} = V$, a static map.

In general, the precompensator remainder $\bar{v}$ is dynamic and can be represented as $\bar{v} = V + \bar{v}_c$ where V is the static part of $\bar{v}$ and $\bar{v}_c: \Lambda U \to \Lambda U$ is strictly causal, i.e., an extended linear i/o map. We recall (see, in particular, Hautus and Heymann [1978]) that the dynamic characteristics of $\bar{v}_c$ are determined by $\ker \pi^+ \bar{v}_c \cdot j^+$ which is an Ω^+K-submodule of $\Omega^+ U$ and can be represented by

$$\ker \pi^+ \bar{v}_c \cdot j^+ = \ker \pi^+ \bar{v} \cdot j^+ = D\Omega^+ U, \tag{7.1}$$

where D is a polynomial matrix whose columns form a basis for $\ker \pi^+ \bar{v} \cdot j^+$. The degree n of the determinant of D (when D is nonsingular) is the dimension of the minimal state space realizing $\bar{v}_c$. More specifically, if D in (7.1) is selected to be proper, i.e., the columns of D are properly free (in the sense that the leading coefficient vectors are K-linearly independent just as in § 4 above), then the column degrees σ_i, $i = 1, \cdots, m$ are the reachability indices of $\bar{v}_c$ and their sum is $\sum_{i=1}^{m} \sigma_i = n = \deg \cdot \det D$.

It is of interest in selecting a $(\bar{v}, \bar{g})$ pair representing a given precompensator $\bar{l}$ to choose the representation in such a way that the precompensator remainder $\bar{v}$ has least dynamic order, i.e., is realizable by a state space of least possible dimension. In this way the precompensator is realized "as much as possible" by feedback. The following theorem provides a bound on the dynamic order of the precompensator remainder $\bar{v}$ which need not be exceeded in the realization of any bicausal precompensator $\bar{l}$, and which is dependent only on the dynamic properties (latency) of the i/o map $\bar{f}$ under consideration.

THEOREM 7.2. *Let $\bar{f}: \Lambda U \to \Lambda Y$ be an injective extended linear* i/o *map with latency indices $\nu_1 \geqq \cdots \geqq \nu_m$. Let $\bar{l}: \Lambda U \to \Lambda U$ be a bicausal ΛK-linear map. There exists a $(\bar{v}, \bar{g})$ representation for $\bar{l}$ such that the precompensator remainder $\bar{v}$ has (ordered) reachability indices $\sigma_1 \geqq \cdots \geqq \sigma_m$ satisfying $\sigma_i \leqq \nu_i$, $i = 1, \cdots, m$.*

Remark. 7.3. It is interesting to observe that Theorem 7.2 explicitly implies what we have seen previously, namely, that if $\bar{f}$ is a nonlatent i/o map, then every bicausal $\bar{l}$ can be realized as output feedback. Indeed, if $\bar{f}$ is nonlatent, its latency indices ν_i are all zero, whence by Theorem 7.2 there exists a pair $(\bar{v}, \bar{g})$ with $\bar{v}$ having reachability indices all zero, that is, with $\bar{v}$ static. □

To prove Theorem 7.2 we shall need the following lemmas.

LEMMA 7.4. *Let U be a finite dimensional K-linear space and let $\bar{v}: \Lambda U \to \Lambda U$ be a bicausal ΛK-linear isomorphism. Then* $\ker \pi^+ \bar{v} \cdot j^+$ *and* $\ker \pi^+ \bar{v}^{-1} \cdot j^+$ *have the same lists of reachability indices.*

Proof. By Hautus and Heymann [1978, Theorem 6.11] the lemma will be proved upon showing that there exists an order-preserving $\Omega^+ K$-isomorphism $\ker \pi^+ \bar{v} \cdot j^+ \to \ker \pi^+ \bar{v}^{-1} \cdot j^+$. We shall see that the map $\bar{v}$ itself, which is in particular also an order

preserving Ω^+K-isomorphism, satisfies the required properties. Indeed, let $\xi \in \text{ket } \pi^+\bar{v} \cdot j^+$ be any element. Then $\bar{v} \cdot j^+\xi = \bar{v}\xi \in \Omega^+U$ and since also $\xi \in \Omega^+U$ we have $\xi = \bar{v}^{-1}(\bar{v}\xi) = \bar{v}^{-1}j^+(\bar{v}\xi) \in \Omega^+U$, whence $\bar{v}\xi \in \ker \pi^+\bar{v}^{-1} \cdot j^+$, completing the proof. □

Let $\bar{f}: \Lambda U \to \Lambda Y$ be an injective extended linear i/o map and let $d_1, \cdots, d_m$ be a proper strictly polynomial basis for $\ker \pi^-\bar{f}$ (see Theorem 6.19), and write $\ker \pi^-\bar{f} = D\Omega^-U$ where $D = [d_1, \cdots, d_m]$. Then $z^{-1}D$ is also polynomial and the column degrees of $z^{-1}D$ are (by definition) the latency indices of $\bar{f}$. Below we shall not distinguish sharply between maps and their transfer functions. Let $\mathscr{S}^-: \Lambda U \to \Omega^-U: \Sigma u_t z^{-t} \mapsto \sum_{t \geqq 0} u_t z^{-t}$ denote the *causal truncation*. Let $N: \Lambda U \to \Omega^-U$ be defined as the (unique) ΛK-linear map whose transfer function is given by

$$N := \mathscr{S}^-(\bar{l}^{-1}D), \tag{7.5}$$

and define the ΛK-linear maps

$$\bar{\phi}: \Lambda U \to \Lambda U: u \mapsto ND^{-1}u, \tag{7.6}$$

$$\bar{v}^{-1} := \bar{l}^{-1} - \bar{\phi}. \tag{7.7}$$

LEMMA 7.8. *With* $\bar{\phi}$ *and* $\bar{v}^{-1}$ *as defined in* (7.6) *and* (7.7) *the following hold true*:
(i) $\ker \pi^-\bar{f} \subset \ker \pi^-\bar{\phi}$.
(ii) $z^{-1}D\Omega^+U \subset \ker \pi^+ \cdot \bar{v}^{-1} \cdot j^+$.

Proof. (i) Let $u \in \ker \pi^-\bar{f}$. Then $u = Dw$ for some $w \in \Omega^-U$ and we have $\bar{\phi}u = ND^{-1}u = ND^{-1}Dw = Nw \in \Omega^-U$ since N is a causal map, and hence $\pi^-\bar{\phi}u = 0$ so that $u \in \ker \pi^-\bar{\phi}$. (ii) If $u \in z^{-1}D\Omega^+U$ then $u = z^{-1}Dw$ for some $w \in \Omega^+U$, and we have, using the definitions of $\bar{v}^{-1}$ and of $\bar{\phi}$, $\bar{v}^{-1}j^+u = \bar{v}^{-1}z^{-1}Dw = (\bar{l}^{-1} - \bar{\phi})z^{-1}Dw = z^{-1}(\bar{l}^{-1}D - N)w$. Now, in view of (7.5) the map $(\bar{l}^{-1}D - N)$ has a strictly polynomial transfer function so that $z^{-1}(\bar{l}^{-1}D - N)$ is polynomial. Since also w is polynomial it follows that $z^{-1}(\bar{l}^{-1}D - N)w \in \Omega^+U$, whence $u \in \ker \pi^+\bar{v}^{-1}j^+$ as claimed. □

Proof of Theorem 7.2. If $\bar{l}$ is a bicausal precompensator for $\bar{f}$ and $(\bar{v}, \bar{g})$ is a representation of $\bar{l}$, then $\bar{l} = (I + \bar{g} \cdot \bar{f})^{-1}\bar{v}$, whence $\bar{l}^{-1} = \bar{v}^{-1} + \bar{v}^{-1} \cdot \bar{g} \cdot \bar{f} = \bar{v}^{-1} + \bar{\rho} \cdot \bar{f}$, where the map $\bar{\rho} = \bar{v}^{-1}\bar{g}$ is clearly also causal. By Lemma 7.4, $\bar{v}$ and $\bar{v}^{-1}$ have the same reachability indices. Hence the theorem will be proved if we can show that $\bar{l}^{-1}$ can be represented as

$$\bar{l}^{-1} = \bar{v}^{-1} + \bar{\phi}$$

satisfying the following requirements: (a) $\bar{v}^{-1}: \Lambda U \to \Lambda U$ is a bicausal ΛK-linear map such that its reachability indices σ_i satisfy $\sigma_i \leqq \nu_i$, $i = 1, \cdots, m$. (b) The ΛK-linear map $\bar{\phi}: \Lambda U \to \Lambda U$ is strictly causal and can be represented as $\bar{\phi} = \bar{\rho} \cdot \bar{f}$ for some causal ΛK-linear map $\bar{\rho}: \Lambda Y \to \Lambda U$. As we see below, the maps $\bar{\phi}$ and $\bar{v}^{-1}$ as defined in (7.6) and (7.7) satisfy the required conditions. Indeed, Lemma 7.8(i) combined with Theorem 5.2 implies that $\bar{\phi} = \bar{p} \cdot \bar{f}$ for some causal $\bar{p}$. Since $\bar{f}$ is strictly causal by definition, it follows that so also is $\bar{\phi}$. Hence condition (b) above holds. To see that (a) is also satisfied note first that the difference between a bicausal ΛK-linear map and a strictly causal one is bicausal (see e.g. Corollary 2.11). Hence the map $\bar{v}^{-1}$ is bicausal. Now Lemma 7.8(ii) implies the requirement on the reachability indices since, in particular, it implies that $\bar{v}^{-1}$ can be realized with state space $\Omega^+U/z^{-1}D\Omega^+U$ whose reachability indices are the column degrees of $z^{-1}D$. (The reader is referred to Hautus and Heymann [1978] for relevant details on the problem of realization.) □

While Theorem 7.2 gives an upper bound on the required dynamic order of precompensator remainders, it has been, so far, seen only in the nonlatent case that this bound is tight. It is clear that in general, except in the case of nonlatent i/o maps, the

maximal required order of precompensator remainders depends not only on the i/o map $\bar{f}$ but also on the specific precompensator $\bar{l}$ under consideration. It turns out that the bound of Theorem 7.2 is tight, however, in the following sense: There always exist bicausal isomorphisms $\bar{l}$ for which all precompensator remainders satisfy the condition that $n = \sum_{i=1}^{m} \sigma_i \geqq \sum_{i=1}^{m} \nu_i$, where n is the minimal state space dimension and the σ_i are reachability indices of the precompensator remainder, and the ν_i are the latency indices of the i/o, map $\bar{f}$.

THEOREM 7.9. *Let $\bar{f}: \Lambda U \to \Lambda Y$ be an injective linear* i/o *map with latency indices $\nu_1, \cdots, \nu_m$. There exists a ΛK-linear bicausal isomorphism $\bar{l}: \Lambda U \to \Lambda U$ such that the following holds: If $(\bar{v}, \bar{g})$ is any representation of $\bar{l}$ and if $\sigma_1, \cdots, \sigma_m$ are the reachability indices of the precompensator remainder $\bar{v}$, then $\sum_{i=1}^{m} \sigma_i \geqq \sum_{i=1}^{m} \nu_i$.*

Proof. Let $d_1, \cdots, d_m$ be a proper strictly polynomial basis for $\ker_{\cdot} \pi^{-} \bar{f}$ and write $\ker \pi^{-} \bar{f} = D\Omega^{-} U$ where $D = [d_1, \cdots, d_m]$. Then the matrix $D_1 := z^{-1} D$ is also polynomial and D_1^{-1} is causal (see Proposition 6.18). Below we shall use the same notation interchangeably for matrices and their associated ΛK-linear maps. Let $L: \Lambda U \to \Lambda U$ be any static ΛK-linear map such that $L + D_1^{-1}$ is bicausal. Consider the bicausal precompensator $\bar{l} := (L + D_1^{-1})^{-1}$. If $\bar{v}$ is any precompensator remainder for $\bar{l}$, then $\bar{v}^{-1} = \bar{l}^{-1} - \bar{\rho}\bar{f} = L + D_1^{-1} - \bar{\rho}\bar{f}$ for some causal map $\bar{\rho}$. By Lemma 7.4, $\bar{v}$ has the same reachability indices as $\bar{v}^{-1}$ and the latter has the same reachability indices as $D_1^{-1} - \bar{\rho}\bar{f}$. Now, we have

$$D_1^{-1} - \bar{\rho} \cdot \bar{f} = (I - \bar{\rho} \cdot \bar{f} \cdot D_1) D_1^{-1} = \bar{l}^{*} \cdot D_1^{-1}$$

where $\bar{l}^{*} = I - \bar{\rho} \cdot \bar{f} \cdot D_1$ is bicausal because the composite $\bar{f} \cdot D_1$ is strictly causal, the latter following since $\ker \pi^{-} \bar{f} \cdot D_1 = D_1^{-1} \ker \pi^{-} \bar{f} = D_1^{-1}(zD_1)\Omega^{-} U = z\Omega^{-} U$. Let $\bar{l}^{*} D_1^{-1} = P \cdot Q^{-1}$ be a coprime fraction representation of $\bar{l}^{*} \cdot D_1^{-1}$ (see, e.g., Heymann [1972] or Hautus and Heymann [1978]). Then clearly P is nonsingular, and computing determinantal degrees gives us (because $\bar{l}^{*}$ is bicausal) that

$$n := \deg \det Q = \deg \det P + \deg \det D_1 \geqq \deg \det D_1.$$

Since n equals the sum of the reachability indices of the i/o map $P \cdot Q^{-1}$ the proof is complete. □

Note added in proof. The reader is also referred to Emre and Hautus [1980], where certain solvability conditions for rational matrix equations are given that are related to the causal factorization problem.

REFERENCES

F. M. BRASH AND J. B. PEARSON [1970], *Pole placement using dynamic compensators*, IEEE Trans. Automat. Control, AC-15, pp. 34–43.

A. E. ECKBERG, JR. [1974], *A characterization of linear systems via polynomial matrices and module theory*, MIT Electronic Systems Laboratory Rep. ESL-R-528, Mass. Inst. of Tech., Cambridge, MA.

E. EMRE AND M. L. J. HAUTUS [1980], *A polynomial characterization of $(t, \mathcal{B})$-invariant and reachability subspaces*, this Journal, 18, pp. 420–436.

P. L. FALB AND W. A. WOLOVICH [1967], *Decoupling in the design and synthesis of multivariable control systems*, IEEE Trans. Automat. Control, AC-12, pp. 651–659.

G. D. FORNEY, JR. [1975], *Minimal bases of rational vector spaces, with applications to multivariable linear systems*, SIAM J. Control, 13, pp. 493–520.

P. A. FUHRMANN [1976], *Algebraic system theory: an analyst's point of view*, J. Franklin Inst., 301, pp. 521–540.

—— [1979], *Linear feedback via polynomial models*, Int. J. Control, to appear.

E. G. GILBERT [1969], *The decoupling of multivariable systems by state feedback*, SIAM J. Control, 7, pp. 50–64.

W. H. GREUB [1967], *Linear Algebra*, 3rd edition, Springer Verlag, Berlin.

M. L. J. HAUTUS AND M. HEYMANN [1978], *Linear feedback—an algebraic approach*, this Journal, 16, pp. 83–105.

M. HEYMANN [1968], *Comments on pole assignment in multi-input controllable linear systems*, IEEE Trans. Automat. Control, AC-13, pp. 748–749.

M. HEYMANN [1972], *Structure and realization problems in the theory of dynamical systems*, Lecture Notes, International Center for Mechanical Sciences, Udine, Italy; also Springer-Verlag, New York, 1975.

R. E. KALMAN, P. L. FALB AND M. A. ARBIB [1969], *Topics in mathematical system theory*, McGraw Hill, New York.

D. G. LUENBERGER [1966], *Observers for multivariable systems*, IEEE Trans. Automat. Control, Ac-11, pp. 190–197.

C. C. MACDUFFEE [1934], *The Theory of Matrices*, Chelsea, New York.

A. S. MORSE [1975], *System invariants under feedback and cascade control*, Proceedings of the conference on mathematical systems theory, Udine, Italy, pp. 61–74; Lecture Notes in Economics and Mathematical Systems 131, Springer Verlag, Berlin.

A. S. MORSE AND W. M. WONHAM [1970], *Decoupling and pole assignment by dynamic compensation*, SIAM J. Control, 8, pp. 317–337.

H. F. MÜNZER AND D. PRÄTZEL-WOLTERS [1979a], *Minimal bases of polynomial modules, structural indices and Brunovsky-transformations*, Int. J. Control, 30, pp. 291–318.

——— [1979b], *Geometric and moduletheoretic approach to linear systems, Part 1: basic categories and functors*, Proceedings of the Delft Conference on Systems and Networks, July.

——— [1979c] *Geometric and moduletheoretic approach to linear systems, Part 2: moduletheoretic characterization of reachability subspaces*, Internal report, Universität Bremen, Bremen, Germany.

H. H. ROSENBROCK [1970], *State space and multivariable theory*, Nelson, London.

J. D. SIMON AND S. K. MITTER [1968], *A theory of modal control*, Information and Control, 13, pp. 316–353.

W. A. WOLOVICH [1974], *Linear multivariable systems*, Applied Mathematical Sciences Series, 11, Springer-Verlag, New York.

W. M. WONHAM [1967], *On pole assignment in multi-input controllable linear systems*, IEEE Trans. Automat. Control. AC-12, pp. 660–665.

——— [1979] *Linear Multivariable Control: A Geometric Approach*, 2nd ed., Springer-Verlag, New York.

W. M. WONHAM AND A. S. MORSE [1970], *Decoupling and pole assignment in linear multivariable systems: A geometric approach*, SIAM J. Control, 8, pp. 1–18.

B. F. WYMAN [1972], *Linear systems over commutative rings*, Lecture notes, Stanford Univ., Stanford, CA.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 4, July 1981

0363-0129/81/1904-0003 $01.00/0

REFLECTED BROWNIAN MOTION IN THE "BANG-BANG" CONTROL OF BROWNIAN DRIFT*

STEVEN E. SHREVE†

Abstract. Let $w(t)$ be a standard one-dimensional Brownian motion, and define the Brownian motion $z_x(t) = \int_0^t -\operatorname{sgn}(w(s)+x)\,dw(s)$. It is shown that a reflecting Brownian motion related to $\{z_x(s), s \geqq 0\}$ coincides with $|w(t)+x|$. A related computation yields the joint distribution of $w(t)$ and its local time. These results are applied to the control problem of minimization of $Ey^2(T)$ subject to

$$y(0) = x, \qquad dy(t) = u\,dt + dw(t).$$

The optimal control law is known to be $u = -\operatorname{sgn} y(t)$. We compute the optimal transition density and value function.

1. Introduction Let $\{w(t), t \geqq 0\}$ be a standard, one-dimensional Brownian motion defined on a probability space $(\Omega, \mathcal{F}, P)$, and let $\mathcal{F}(t)$ be the σ-field generated by $\{w(s): 0 \leqq s \leqq t\}$. By a result due to Skorohod [12] and reported by McKean [11, § 3.9], for each $x \geqq 0$, there is a unique pair of nonanticipating processes $(\xi(t), \zeta(t))$ satisfying

$$\xi(t) = x + \int_0^t \operatorname{sgn}(w(s)+x)\,dw(s) + \zeta(t) \tag{1}$$

such that $\{\xi(t), t \geqq 0\}$ is nonnegative and $\{\zeta(t), t \geqq 0\}$ is nondecreasing with $\zeta(0) = 0$. Moreover, $\zeta(t)$ is the local time at zero of $\xi(t)$.

It is easy to see (§ 2) that Tanaka's formula gives the solution to (1). This solution shows that $\{\xi(t), t \geqq 0\}$ is a reflecting Brownian motion beginning at x, and $\zeta(t)$ is the local time of $\{w(s), s \geqq 0\}$ at $-x$. We use these facts in § 3 to identify two apparently very different reflecting Brownian motions. In § 4 we compute the joint distribution of $w(t)$ and its local time at x. The distribution was previously known for the case $x = 0$ [9, p. 45, Problem 2.3.3]. In § 5 we apply the results of §§ 1–4 to the computation of the transition density for the system

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

$$dy(t) = -\operatorname{sgn} y(t)\,dt + dw(t). \tag{3}$$

The origin of this system as a solution to an optimal control problem is also discussed in § 5.

2. Tanaka's formula. For the moment, we allow x to be any real number. We single out the negative of the Ito integral

$$z_x(t) = -\int_0^t \operatorname{sgn}(w(s)+x)\,dw(s) \tag{4}$$

appearing in (1) for further attention. It is well known that $\{z_x(t), t \geqq 0\}$ is a standard Brownian motion adapted to $\{\mathcal{F}(t), t \geqq 0\}$. A simple proof of this can be given using time substitution as described in [11, § 2.5].

It is perhaps worth noting that for fixed $t > 0$, $w(t)$ and $z_0(t)$ are uncorrelated normal random variables which are not independent. The correlation coefficient can be computed by applying Ito's lemma to $w(t)z_0(t)$. The lack of independence follows from (6). In fact, the joint density can be obtained by setting $x = 0$ in (20)–(22).

* Received by the editors February 1, 1980, and in revised form July 7, 1980. This work was supported by the University of Delaware Research Foundation.

† Department of Mathematics, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213.

Corresponding to the Brownian motion $\{w(t), t \geqq 0\}$ there exists a function $l: [0, \infty) \times R \times \Omega \to [0, \infty)$ called *local time*, characterized by the following properties:

(L1) For each $\omega \in \Omega$, $l(t, \xi, \omega)$ is jointly continuous in (t, ξ).

(L2) For each $\omega \in \Omega$ and $\xi \in R$, $l(t, \xi, \omega)$ is nondecreasing in t.

(L3) There is a subset Ω^* of Ω of full measure such that for each $\omega \in \Omega^*$,

$$\int_0^t 1_{[a,b]}(w(s, \omega))\, ds = \int_a^b l(t, \xi, \omega)\, d\xi, \qquad t \geqq 0,\ a, b \in R,\ a < b. \tag{5}$$

The existence of local time for Brownian motion was first proved by Trotter [13]. (See also Ito and McKean [9, § 2.8]). Local time has the following interpretation. For each $\omega \in \Omega$ and $t > 0$, there is a measure $\mu(t, \omega)$ on the Borel subsets of R given by

$$\mu(t, \omega)(B) = \text{Lebesgue measure } \{s: 0 \leqq s \leqq t,\ w(s, \omega) \in B\}.$$

Condition (L3) states that $l(t, \cdot, \omega)$ is the density for this measure. Hereafter we suppress the sample point ω.

The Brownian motions $\{z_x(t), t \geqq 0\}$ and $\{w(t), t \geqq 0\}$ are related by the formula

$$z_x(t) = |x| - |w(t) + x| + l(t, -x), \qquad x \in R, \quad t \geqq 0. \tag{6}$$

Equation (6), which is easily seen to be equivalent to Tanaka's formula [11, pp. 68–70], holds except on a null subset of Ω independent of t. A heuristic proof of (6) (and Tanaka's formula) can be given by differentiating (5) with respect to b and evaluating at $b = -x$ to obtain

$$\int_0^t \delta(w(s) + x)\, ds = l(t, -x),$$

where δ is the Dirac delta. Now apply Ito's lemma to $|w(t) + x|$ to obtain

$$|w(t) + x| = |x| + \int_0^t \operatorname{sgn}(w(s) + x)\, dw(s) + \int_0^t \delta(w(s) + x)\, ds,$$

from which (6) follows. The argument can be made rigorous by taking smooth approximations to the functions involved.

3. Reflecting Brownian motion. In this section we assume the number x in (4) and (5) is nonnegative. Due to the symmetry with respect to the origin of $\{w(t), t \geqq 0\}$, this represents no loss of generality, and the corresponding results for $x \leqq 0$ can be written down with only a moment's reflection.

Corresponding to any standard Brownian motion $\{w(t), t \geqq 0\}$ and $x \geqq 0$, there are two "reflecting Brownian motions." The first is

$$w^+(t) = |w(t) + x|.$$

To obtain the second, define

$$\begin{aligned} \tau &= \min \{s \geqq 0: w(s) = -x\}, \\ w_{\max}(t) &= \max \{w(s): \tau \leqq s \leqq t\}. \end{aligned} \tag{7}$$

The second reflecting Brownian motion is

$$w^-(t) = \begin{cases} w(t) + x, & 0 \leqq t \leqq \tau, \\ w_{\max}(t) - w(t), & t > \tau. \end{cases}$$

After time τ, the trajectories of $w^+(t)$ and $w^-(t)$ are in general different for the same ω, but $\{w^+(t), t \geqq 0\}$ and $\{w^-(t), t \leqq 0\}$ are identical in law [9, § 2.1]. The question arises

whether we can find a standard Brownian motion $\{b(t), t \geqq 0\}$ such that $\{w^+(t), t \geqq 0\}$ and $\{b^-(t), t \geqq 0\}$ are the same process. The answer of the next theorem is that such a $\{b(t), t \geqq 0\}$ exists and is intimately connected with $\{z_x(t), t \geqq 0\}$.

THEOREM 1. *Let $\{w(t), t \geqq 0\}$ be a standard Brownian motion, let $x \geqq 0$ be given, and let $\{z_x(t), t \geqq 0\}$ be defined by* (4). *Define τ by* (7) *and*

$$b(t) = \begin{cases} w(t), & 0 \leqq t \leqq \tau, \\ z_x(t) - 2x, & t > \tau. \end{cases}$$

Then $\{b(t), t \geqq 0\}$ is a standard Brownian motion, and for ω in a set of full measure,

$$w^+(t) = b^-(t) \quad \forall t \geqq 0. \tag{8}$$

Proof. Note first of all that $b(\tau) = -x$, while $z_x(\tau) = x$, so $b(\tau) = z_x(\tau) - 2x$. For $t > \tau$, we have

$$\begin{aligned} b(t) &= [z_x(t) - z_x(\tau)] + b(\tau) \\ &= \int_\tau^t -\operatorname{sgn}(w(s) + x)\, dw(s) + b(\tau), \end{aligned}$$

while for $t \leqq \tau$

$$b(t) = \int_0^t dw(s).$$

We summarize this as

$$b(t) = \int_0^t [1_{[0,\tau]}(s) - 1_{(\tau,\infty)}(s) \operatorname{sgn}(w(s) + x)]\, dw(s),$$

and from this representation we can use the time substitution argument [11, § 2.5] to show $\{b(t), t \geqq 0\}$ is a standard Brownian motion.

It is clear from the definitions that for $t \leqq \tau$, $w^+(t) = b^-(t)$. For $t \geqq \tau$, we have

$$b_{\max}(t) = \max\{b(s): \tau \leqq s \leqq t\},$$

and from (6),

$$\begin{aligned} b(t) &= z_x(t) - 2x \\ &= -|w(t) + x| - x + l(t, -x) \\ &\leqq -x + l(t, -x), \end{aligned}$$

so

$$b_{\max}(t) \leqq -x + l(t, -x). \tag{9}$$

Furthermore,

$$b^-(t) = b_{\max}(t) - b(t) \leqq w^+(t). \tag{10}$$

Since both $\{b^-(t), t \geqq 0\}$ and $\{w^+(t), t \geqq 0\}$ are reflecting Brownian motions, and for both of them τ is the time of first passage to the origin, they must have the same distribution on $\{\tau \leqq t\}$. Thus (10) implies that for almost all ω satisfying $\tau(\omega) \leqq t$, we have $b^-(t) = w^+(t)$. Since these processes are continuous, there is a subset Ω^* of Ω of full measure such that $\omega \in \Omega^*$ implies $b^-(t) = w^+(t)$ for every $t \geqq \tau(\omega)$. □

COROLLARY 2. *Under the conditions of Theorem* 1, *for* ω *in a set of full measure,*

$$l(t, -x) = x + b_{\max}(t) \quad \forall t \geqq \tau. \tag{11}$$

Proof. Equality in (10) implies equality in (9). □

Let us define for $t \geqq 0$

$$z_{\max}(t) = \max \{z_x(s): 0 \leqq s \leqq t\}.$$

Note that the maximum is taken over $s \in [0, t]$, rather than $s \in [\tau, t]$.

COROLLARY 3. *Assume the hypotheses of Theorem* 1. *Then, for all* ω *in a set of full measure,*

$$\tau = \min \{s \geqq 0: z_x(s) = x\}, \tag{12}$$

$$z_{\max}(t) > x \Leftrightarrow l(t, -x) > 0 \Leftrightarrow \tau < t, \tag{13}$$

$$z_{\max}(t) = x \Leftrightarrow \tau = t, \tag{14}$$

$$w(t) = -z_x(t) \qquad \forall t \in [0, \tau], \tag{15}$$

$$|w(t) + x| = z_{\max}(t) - z_x(t) \qquad \forall t \in [\tau, \infty), \tag{16}$$

$$z_{\max}(t) = l(t, -x) + x \quad \forall t \in [\tau, \infty). \tag{17}$$

Proof. Equation (15) is an immediate consequence of (4) and (7), and (7) and (15) imply (12). Since $z_x(t) < x$ for $\tau > t$ and $z_x(\tau) = x$, we have

$$b_{\max}(t) = z_{\max}(t) - 2x, \tag{18}$$

and (17) follows from (11). Suppose ω is in the set of full measure for which (12), (15) and (17) hold, and for which the law of the iterated logarithm holds for the Brownian motion $\{z_x(\tau + t), t \geqq 0\}$. Then $z_{\max}(t) > x$ for $\tau < t$, so $z_{\max}(t) = x$ is equivalent to $\tau = t$, and $z_{\max}(t) > x$ is equivalent to $\tau < t$. Furthermore, $z_{\max}(t) > x$ and (17) imply $l(t, -x) > 0$. On the other hand, (17) implies $l(\tau, -x) = 0$, and since $l(t, -x)$ is nondecreasing in t, $l(t, -x) > 0$ implies $\tau < t$. This establishes the equivalences of (13). Equation (16) follows from (8), (18) and the definitions. □

4. The joint distribution of Brownian motion and its local time. To simplify notation, we fix $x \geqq 0$ and $t > 0$ and set

$$W = w(t), \qquad M = z_{\max}(t),$$

$$Y = w(t) + x, \qquad K = l(t, -x).$$

$$Z = z_x(t),$$

The joint density of (Z, M) is [9, § 2.1]

$$f_{Z,M}(z, m) = \frac{2(2m - z)}{t\sqrt{2\pi t}} \exp\left[-\frac{(2m - z)^2}{2t}\right], \qquad z \leqq m, \quad m \geqq 0. \tag{19}$$

LEMMA 4. *On the set* $\{M \geqq x\}$, *the distribution of* Y *conditioned on* Z *is symmetric with respect to the origin.*

Proof. Let $y(s) = w(s) + x$. From Corollary 3 we know that $\{M \geqq x\}$ is almost surely equal to $\{\tau \leqq t\}$, and on this set

$$Z = x + \int_\tau^t -\operatorname{sgn}(y(s))\, dy(s)$$

$$= x + \int_\tau^t -\operatorname{sgn}(-y(s))\, d(-y(s)).$$

Since the processes $\{y(s)1_{\{M \geqq x\}}, s \geqq 0\}$ and $\{-y(s)1_{\{M \geqq x\}}, s \geqq 0\}$ are identical in law, they have the same joint distribution with Z on $\{M \geqq x\}$. In particular, (Z, Y) and $(Z, -Y)$ are identical in law on this set. The lemma follows. □

We now compute the joint distribution of (Y, Z). For this purpose, we introduce three subsets of the (y, z)-plane. Define

$$R_1 = \{(y, z): z + y = x, y > 0\},$$
$$R_2 = \{(y, z): z + y > x, y > 0\},$$
$$R_3 = \{(y, z): z - y > x, y < 0\}.$$

The joint distribution of (Y, Z) can be characterized by a nonnormalized density on $R_2 \cup R_3$ plus a singular distribution on R_1. The singular part arises from (15). To see this, observe first that $P\{M = x\} = 0$, so we need only consider the cases $\{M < x\}$ and $\{M > x\}$. When $M < x$, we have $\tau > t$ and $(Y, Z) \in R_1$ (see (13)–(15)). If $\{M > x\}$, $\tau < t$ and $|Y| = M - Z$ (see (13)–(16)). Depending on whether $Y > 0$ or $Y < 0$ ($P\{Y = 0\} = 0$), we have $(Y, Z) \in R_2$ or $(Y, Z) \in R_3$, and for fixed Z, Lemma 4 states that each of these two cases is equally likely. Therefore, for any $z \leqq x$,

$$\begin{aligned} P\{(Y, Z) \in R_1, Z \leqq z\} &= \int_{-\infty}^{z} \int_{\zeta}^{x} f_{Z,M}(\zeta, m)\, dm\, d\zeta \\ &= \frac{1}{\sqrt{2\pi t}} \int_{z-2x}^{z} \exp\left(-\frac{\zeta^2}{2t}\right) d\zeta, \end{aligned} \tag{20}$$

while for any Borel set B contained in $R_2 \cup R_3$,

$$P\{(Y, Z) \in B\} = \int_B f_{Y,Z}(y, z)\, dy\, dz, \tag{21}$$

where

$$f_{Y,Z}(y, z) = \begin{cases} \frac{1}{2} f_{Z,M}(z, z + y) & \text{if } (y, z) \in R_2, \\ \frac{1}{2} f_{Z,M}(z, z - y) & \text{if } (y, z) \in R_3. \end{cases} \tag{22}$$

We define three subsets of the (w, k)-plane:

$$S_1 = \{(w, k): w + x > 0, k = 0\},$$
$$S_2 = \{(w, k): w + x > 0, k > 0\},$$
$$S_3 = \{(w, k): w + x < 0, k > 0\}.$$

It is a consequence of Corollary 3 that there is a subset of Ω of full measure on which $(W, K) \in S_i$ if and only if $(Y, Z) \in R_i$, $i = 1, 2, 3$. We have from (20),

$$\begin{aligned} P\{(W, K) \in S_1, W > w\} &= P\{(Y, Z) \in R_1, Z < -w\} \\ &= \frac{1}{\sqrt{2\pi t}} \int_{-w-2x}^{-w} \exp\left(-\frac{\zeta^2}{2t}\right) d\zeta, \qquad w \geqq -x, \end{aligned} \tag{23}$$

and

$$P\{(W, K) \in S_1\} = \frac{1}{\sqrt{2\pi t}} \int_{-x}^{x} \exp\left(-\frac{\zeta^2}{2t}\right) d\zeta. \tag{24}$$

This describes the singular part of the distribution of (W, K), the case where $K = 0$. For any Borel set B contained in $S_2 \cup S_3$,

$$P\{(W, K) \in B\} = \int_B f_{W,K}(w, k)\, dw\, dk,$$

where, from (16), (17), (19) and (23),

$$(25)\quad f_{W,K}(w, k) = \begin{cases} \dfrac{k + 2x + w}{t\sqrt{2\pi t}} \exp\left[-\dfrac{(k + 2x + w)^2}{2t}\right], & w + x > 0, \quad k > 0, \quad x \geqq 0, \\ \dfrac{k - w}{t\sqrt{2\pi t}} \exp\left[-\dfrac{(k - w)^2}{2t}\right], & w + x < 0, \quad k > 0, \quad x \geqq 0. \end{cases}$$

The distribution of (W, K) for $x \leqq 0$ can be obtained from (23)–(25) and symmetry considerations. We summarize these results as a theorem, in which we give the joint distribution of $(w(t), l(t, x))$ rather than $(w(t), l(t, -x))$.

THEOREM 5. *Let $\{w(t), t \geqq 0\}$ be a standard Brownian motion and let $l(t, x)$, $t \geqq 0$, $x \in R$, be its local time. Fix $t > 0$, $x \in R$ and set $W = w(t)$, $L = l(t, x)$. Then*

$$P\{L = 0\} = \frac{2}{\sqrt{2\pi t}} \int_0^{|x|} \exp\left(-\frac{\zeta^2}{2t}\right) d\zeta,$$

and

$$P\{W < w, L = 0\} = \begin{cases} \dfrac{1}{\sqrt{2\pi t}} \displaystyle\int_{w-2x}^{w} \exp\left(-\frac{\zeta^2}{2t}\right) d\zeta, & w \leqq x, \quad x \geqq 0, \\ P\{L = 0\} - \dfrac{1}{\sqrt{2\pi t}} \displaystyle\int_{-w+2x}^{-w} \exp\left(-\frac{\zeta^2}{2t}\right) d\zeta, & w \geqq x, \quad x \leqq 0. \end{cases}$$

For any Borel set B contained in the half-plane $\{(w, l)\colon l > 0\}$,

$$P\{(W, L) \in B\} = \int_B f_{W,L}(w, l)\, dw\, dl,$$

where

$$f_{W,L}(w, l) = \begin{cases} \varphi(l + 2x - w), & w \leqq x, \quad x \geqq 0, \quad l > 0, \\ \varphi(l + w), & w \geqq x, \quad x \geqq 0, \quad l > 0, \\ \varphi(l - w), & w \leqq x, \quad x \leqq 0, \quad l > 0, \\ \varphi(l - 2x + w), & w \geqq x, \quad x \leqq 0, \quad l > 0, \end{cases}$$

and

$$\varphi(u) = \frac{u}{t\sqrt{2\pi t}} \exp\left(-\frac{u^2}{2t}\right).$$

5. The "bang-bang" control problem. The control problem which motivated this inquiry is the following.

Minimize

$$(26)\qquad Ey^2(T),$$

subject to

$$dy(t) = u(t, y(t))\,dt + dw(t), \tag{27}$$

$$y(0) = x, \tag{28}$$

$$|u(t, y)| \leqq 1, \qquad 0 \leqq t \leqq T, \quad y \in R. \tag{29}$$

We assume also that $u: [0, T] \times R \to [-1, 1]$ is jointly Lebesgue measurable. This model is a special case of the one studied by Beneš [2], and the result of [2] specialized to our case says that the "bang-bang" control

$$u(t, y(t)) = -\operatorname{sgn}(y(t)) \tag{30}$$

is optimal. Beneš [3] shows that for almost every sample path of the solution of (27), where (28) and (30) hold, the set of t for which $y(t) = 0$ has Lebesgue measure zero. The definition of sgn (0) is thus inconsequential. We take sgn (0) = 0. The optimal control law for this problem has also been derived by Davis and Clark [5] using martingale methods, Ikeda and Watanabe [8] using a stochastic ordering principle, and Haussmann [7] using a stochastic maximum principle.

The purpose of this section is to "solve" the differential equation (27), where (28) and (30) hold, and evaluate $Ey^2(T)$. A "solution" consists of using the Girsanov transformation to compute the transition density for $y(t)$, and this involves the process $\{z_x(t), t \geqq 0\}$. We then obtain an explicit formula for $Ey^2(T)$ (Corollary 7). Beneš, Shepp and Witsenhausen [4] have recently used the forward and backward equations of the controlled system to determine the Laplace transform of the transition density. Balakrishnan [1] has found the transition density, albeit in a less explicit form than given here.

We wish to solve the stochastic differential equation

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

$$dy(t) = -\operatorname{sgn} y(t)\,dt + dw(t). \tag{3}$$

Although the drift coefficient is discontinuous, there is a unique continuous stochastic process $\{y(t), t \geqq 0\}$ adapted to $\{\mathcal{F}(t), t \geqq 0\}$ satisfying (2) and (3) [14, Theorem 3]. This is the strong solution of (2), (3). The uniqueness of this solution implies that its transition density will agree with the transition density of any weak solution. We compute the transition density of the weak solution given by the Girsanov transformation.

To use the Girsanov transformation, define processes

$$\begin{aligned} z_x(t) &= \int_0^t -\operatorname{sgn}(w(s) + x)\,dw(s), \\ \varphi_x(t) &= \exp\left[z_x(t) - \frac{t}{2}\right]. \end{aligned} \tag{4}$$

Since sgn $(w(s) + x)$ is bounded and

$$\int_0^t |\operatorname{sgn} w(s) + x|^2\,ds = \frac{t}{2} \quad \text{a.s.},$$

$\varphi_x(t)$ is of the form

$$\exp\left[\int_0^t \beta(s)\,dw(s) - \frac{1}{2}\int_0^t |\beta(s)|^2\,ds\right]$$

and $\{(\varphi_x(t), \mathcal{F}(t)), t \geq 0\}$ is a martingale [10, eg. 1, p. 220]. We define a new probability measure $\tilde{P}_x$ on $\mathcal{F}(T)$ by

$$\tilde{P}_x(A) = E[1_A \varphi_x(T)], \qquad A \in \mathcal{F}(T), \tag{31}$$

where 1_A is the indicator of A. If $A \in \mathcal{F}(T)$ and $T' > T$, then

$$\begin{aligned}\tilde{P}_x(A) &= E[1_A \varphi_x(T)] \\ &= E\{1_A E[\varphi_x(T') | \mathcal{F}(T)]\} \\ &= E[1_A \varphi_x(T')],\end{aligned}$$

so the definition of $\tilde{P}_x$ on $\mathcal{F}(t)$ is independent of T, provided only that $T \geq t$. We denote by $\tilde{E}_x$ the expectation corresponding to $\tilde{P}_x$.

We now define $y_x(t) = w(t) + x$ and

$$\tilde{w}_x(t) = w(t) + \int_0^t \operatorname{sgn}(y_x(s))\, ds. \tag{32}$$

Girsanov's theorem [10, Thm. 6.3, p. 232] states that $\{\tilde{w}_x(t), 0 \leq t \leq T\}$ is a standard Brownian motion on $(\Omega, \mathcal{F}, \tilde{P}_x)$. From (32) we have

$$dy_x(t) = -\operatorname{sgn}(y_x(t))\, dt + d\tilde{w}_x(t),$$

and so $\{y_x(t), 0 \leq t \leq T\}$ is a solution of (3), except, of course, $w(t)$ in (3) must be replaced by $\tilde{w}_x(t)$. Furthermore, $\tilde{P}_x\{y_x(0) = x\} = E_x\{1_{\{w(0)=0\}} \varphi_x(T)\} = 1$, so $y_x(t)$ satisfies (2) as well. Our goal is to calculate the transition density $p(t, x, y) = (\partial/\partial y)\tilde{P}_x\{y_x(t) \leq y\}$ and to compute the optimal cost $J(T, x) = \tilde{E}_x y_x^2(T)$.

Recalling the notation introduced at the beginning of § 4 and equations (20)–(22), we write

$$\begin{aligned}p(t, x, y) &= \frac{\partial}{\partial y}\tilde{P}\{Y \leq y\} \\ &= \frac{\partial}{\partial y} E\left\{1_{\{Y \leq y\}} \exp\left(Z - \frac{t}{2}\right)\right\}.\end{aligned}$$

For $y > 0$,

$$\begin{aligned}p(t, x, y) = \frac{\partial}{\partial y}\int_0^y \int_{x-\eta}^{\infty} \frac{1}{2} \exp\left(z - \frac{t}{2}\right) f_{Z,M}(z, z+\eta)\, dz\, d\eta \\ + \frac{\partial}{\partial y}\int_{x-y}^{x} \int_z^x \exp\left(z - \frac{t}{2}\right) f_{Z,M}(z, m)\, dm\, dz,\end{aligned}$$

where the second term on the right side comes from the singular part of the distribution of (Y, Z) as described by (20). Taking first the partial derivatives and then performing the integrations, we obtain

$$p(t, x, y) = \frac{1}{\sqrt{2\pi t}}\left[\exp\left(-\frac{(x-y-t)^2}{2t}\right) + e^{-2y}\int_{x+y}^{\infty} \exp\left(-\frac{(v-t)^2}{2t}\right) dv\right], \tag{33}$$

$$x \geq 0, \quad y \geq 0.$$

For $y < 0$, we have

$$\begin{aligned}p(t, x, y) &= \frac{\partial}{\partial y}\int_{-\infty}^{y} \int_{x+\eta}^{\infty} \frac{1}{2} \exp\left(z - \frac{t}{2}\right) f_{Z,M}(z, z-\eta)\, dz\, d\eta \\ &= \int_{x+y}^{\infty} \frac{1}{2} \exp\left(z - \frac{t}{2}\right) f_{Z,M}(z, z-y)\, dz.\end{aligned}$$

Computing the integral leads to

$$p(t,x,y)=\frac{1}{\sqrt{2\pi t}}\left[\exp\left(2x-\frac{(x-y+t)^2}{2t}\right)+e^{2y}\int_{x-y}^{\infty}\exp\left(-\frac{(v-t)^2}{2t}\right)dv\right], \tag{34}$$

$$x\geqq 0,\quad y\leqq 0.$$

Since (33) and (34) coincide when $y=0$, we can take either as the definition of $p(t,x,0)$. Symmetry allows us to relate $p(t,x,y)$ for $x\leqq 0$ to the formulas just computed. Indeed,

$$p(t,x,y)=p(t,-x,-y)\quad \forall x,y\in R. \tag{35}$$

We summarize with a theorem.

THEOREM 6. *Let* $\{w(t),\ t\geqq 0\}$ *be a standard one-dimensional Brownian motion, fix* $x\in R$, *and let* $\tilde{w}_x(t)$ *be given by* (32). *Define* $y_x(t)=w(t)+x$. *Then under the probability measure* $\tilde{P}_x$ *defined by* (31), $\{\tilde{w}_x(t),\ 0\leqq t\leqq T\}$ *is a standard one-dimensional Brownian motion,*

$$y_x(t)=x-\int_0^t \operatorname{sgn}(y_x(s))\,ds+\tilde{w}_x(t), \tag{36}$$

and

$$\tilde{P}_x\{y_x(t)\leqq y\}=\int_{-\infty}^{y}p(t,x,\eta)\,d\eta,$$

where $p(t,x,y)$ *satisfies* (33)–(35).

As one would expect, $p(t,x,y)$ is a fundamental solution of the backward equation

$$\tfrac{1}{2}p_{xx}-\operatorname{sgn}(x)p_x-p_t=0$$

and the forward equation

$$\tfrac{1}{2}p_{yy}+\operatorname{sgn}(y)p_y-p_t=0$$

associated with (36). These facts can be verified directly from (33)–(35).

COROLLARY 7. *Under the hypotheses of Theorem* 6, *let* $J(t,x)=\tilde{E}_x y_x^2(t)$. *Then for* $t>0$ *and* $x\in R$,

$$\begin{aligned}J(t,x)=&\frac{1}{2}+\sqrt{\frac{t}{2\pi}}(|x|-t-1)\exp\left[-\frac{(|x|-t)^2}{2t}\right]\\&+\frac{1}{\sqrt{2\pi}}\left[(|x|-t)^2+t-\frac{1}{2}\right]\int_{-\infty}^{(|x|-t)/(t)^{1/2}}\exp\left(-\frac{u^2}{2}\right)du\\&+\frac{1}{\sqrt{2\pi}}\left(|x|+t-\frac{1}{2}\right)e^{2|x|}\int_{(|x|+t)/(t)^{1/2}}^{\infty}\exp\left(-\frac{u^2}{2}\right)du.\end{aligned} \tag{37}$$

In particular,

$$\lim_{t\downarrow 0}J(t,x)=x^2,\qquad \lim_{t\to\infty}J(t,x)=\tfrac{1}{2}.$$

Proof. The proof of (37) for $x\geqq 0$ is by substitution of (33) and (34) into the definition of $J(t,x)$. The integration is lengthy but straightforward. For $x\leqq 0$, (37) can be obtained from (35) and the case $x\geqq 0$. □

We can now apply the Hamilton–Jacobi–Bellman equation [6, Thm. 4.1, p. 159] to $J(t,x)$ to give still another proof that the control law (30) is optimal. In this case the sufficient condition for optimality reduces to

$$\operatorname{sgn}x=\operatorname{sgn}J_x(t,x)\qquad x\in R,\quad t\geqq 0. \tag{38}$$

Since J is an even function of x, it suffices to show that $J_x(t, x) > 0$ for $x > 0$, $t > 0$. (Since $J(0, x) = x^2$, (38) is trivial for $t = 0$. It is also easily checked that $J_x(t, 0) = 0$.) Computation reveals

$$J_x(t, x) = \frac{2}{\sqrt{2\pi}}\left[(x-t)\int_{(t-x)/(t)^{1/2}}^{\infty} e^{-u^2/2}\,du + (x+t)\,e^{2x}\int_{(t+x)/(t)^{1/2}}^{\infty} e^{-u^2/2}\,du\right], \qquad x > 0.$$

If $0 < t \leqq x$, (38) is clear. If $0 < x < t$, we make the change of variables $v = (u^2 - 4x)^{1/2}$ in the second integral and bound the resulting function $v/(v^2+4x)^{1/2}$ by its value at the lower limit of integration to obtain the desired result.

6. Acknowledgment. The author wishes to acknowledge the aid of V. E. Beneš, who found an error in some preliminary work on this subject and suggested the applicability of Tanaka's formula. He also wishes to thank the referees for pointing out the uniqueness of the transition density corresponding to the weak solution in § 5. The proof of (38) is due to K. Spear.

REFERENCES

[1] A. V. Balakrishnan, *On stochastic bang-bang control*, Appl. Math. Optim., 6 (1980), pp. 91–96.
[2] V. E. Beneš, *Girsanov functionals and optimal bang-bang laws for final value stochastic control*, Stochastic Process. Appl., 2 (1974), pp. 127–140.
[3] ———, *Full "bang" to reduce predicted miss is optimal*, this Journal, 14 (1976), pp. 62–84.
[4] V. E. Beneš, L. A. Shepp and H. S. Witsenhausen, *Some solvable stochastic control problems*, Stochastics, 4 (1980), pp. 39–83.
[5] M. H. A. Davis and J. M. C. Clark, *On "predicted miss" in stochastic control*, Stochastics, 2 (1979), pp. 197–209.
[6] W. H. Fleming and R. W. Rishel, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.
[7] U. G. Haussmann, *Some examples of optimal stochastic controls*, preprint.
[8] N. Ikeda and S. Watanabe, *A comparison theorem for solutions of stochastic differential equations and its applications*, Osaka J. Math., 14 (1977), pp. 617–633.
[9] K. Ito and H. P. McKean, *Diffusion Processes and their Sample Paths*, Academic Press, New York, 1965.
[10] R. S. Liptser and A. N. Shiryayev, *Statistics of Random Processes I: General Theory*, Springer-Verlag, New York, 1977.
[11] H. P. McKean, *Stochastic Integrals*, Academic Press, New York, 1969.
[12] A. V. Skorokhod, *Stochastic equations for diffusion processes in a bounded region* 1, 2, Theory Prob. Appl., 6 (1961), pp. 264–274; 7 (1962), pp. 3–23.
[13] H. F. Trotter, *A property of Brownian motion paths*, Illinois J. Math., 2 (1958), pp. 425–433.
[14] A. K. Zvonkin, *A transformation of the phase space of a diffusion process that removes the drift*, Mat. Sb., 93 (1974), pp. 129–149 (= Math. USSR-Sb., 22 (1974), pp. 129–149).

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 4, July 1981

0363-0129/81/1904-0004 $01.00/0

ON APPLICATIONS OF CONTROL THEORY TO INTEGRAL INEQUALITIES: II*

M. B. SUBRAHMANYAM†

Abstract. In this paper a theorem is given for the attainment of the best possible constants in integral type inequalities. Using a control theoretic formulation for such inequalities, necessary conditions for an optimal control are derived in both finite and infinite interval cases. The utilization of these results is demonstrated by means of an example.

1. An existence theorem. In [8] and [9] we treat some problems related to the application of control theory to integral inequalities in the finite interval case. Here we consider a general case in which the interval of interest need not be finite. We also derive the necessary conditions for an optimal control in a number of problems.

To be more specific, consider the n-dimensional system

$$\dot{x} = \underset{n\times n}{A}(t)x + \underset{n\times r}{B}(t)u, \qquad x(t_0) = 0, \tag{1.1}$$

where $t \in [t_0, t_1]$, $t_1 \leqq \infty$. We impose a finite number of constraints on the trajectory x and the control u, such as, $\lim_{t\to t_1} x(t) = 0$, $\int_{t_0}^{t_1} u^p\, dt = 0$, $p > 0$, and so on. We lay some restrictions on these constraints later.

The functional to be minimized is

$$F(x, u) = \frac{\int_{t_0}^{t_1} \phi^1(u, t)\, dt}{[\int_{t_0}^{t_1} \phi^2(x, t) f(t)\, dt]^\alpha}, \tag{1.2}$$

where $\alpha > 0$, $f(t) \geqq 0$ is measurable and u is a measurable control. We make the following assumptions:

(a) $A(t)$ and $B(t)$ are continuous $n \times n$ and $n \times r$ matrix functions respectively.

(b) For $i = 1, 2$, ϕ^i is continuous in x, u and t. Also, for each t, ϕ^1 is convex in u.

(c) Admissible controls are measurable functions on $[t_0, t_1]$ such that $\int_{t_0}^{t_1} \phi^1\, dt < \infty$.

(d) $\phi^1(u, t) \geqq a\ |u|^p$, $a > 0$, $p > 1$, and $\phi^2(x(t), t) \geqq 0$ along any $x(t)$ which is the response to some admissible $u(t)$.

(e) For each $K < \infty$, there is an integrable $g_K(t)$ such that, if $\|u\|_p \leqq K$, then

$$|\phi^2(x(t), t)\, f(t)| \leqq g_K(t) \tag{1.3}$$

a.e. on $[t_0, t_1]$ for any admissible u whose trajectory obeys any constraints imposed.

(f) There exists $k > 0$ such that, for every $c \geqq 0$,

$$\begin{aligned} \phi^1(cu, t) &= c^k \phi^1(u, t), \\ \phi^2(cx, t) &= c^{k/\alpha} \phi^2(x, t). \end{aligned} \tag{1.4}$$

By (1.3), this assumption implies that for every $c > 0$, $F(cx, cu) = F(x, u)$.

(g) There exists an admissible control the trajectory of which satisfies the imposed constraints and is such that

$$\infty > \int_{t_0}^{t_1} \phi^2(x, t) f(t)\, dt > 0.$$

* Received by the editors May 30, 1980.

† Department of Mathematics, University of Wisconsin, Madison, Wisconsin 53706; currently at Department of Electrical Engineering, Texas A & I University, Kingsville, Texas 78363.

We call a constraint *regular* if the following two conditions hold:

(1) (x, u) satisfies the constraint $\Rightarrow (cx, cu)$ satisfies the constraint for every $c > 0$.

(2) Let (x^1, u^1), $(x^2, u^2), \cdots$ be admissible pairs such that $u^i \to u^0$ weakly in $L_p(t_0, t_1) = \{u = (u_1, \cdots, u_r) : [t_0, t_1] \to R^r \mid \|u\|^p = \int_{t_0}^{t_1} |u|^p \, dt < \infty\}$. Suppose (x^n, u^n) satisfies the constraint for each $n \geqq 1$. Then (x^0, u^0) obeys the constraint. (It is shown in the proof of Theorem 1.1 that u^0 is necessarily admissible.)

PROPOSITION 1.1. *Consider all pairs* (x, u) *that obey* (1.1) *and the constraints. Assume that all the constraints are regular, and let*

$$\lambda = \inf_{(x,u)} F(x, u) = \inf_{(x,u)} \frac{\int_{t_0}^{t_1} \phi^1(u, t) \, dt}{[\int_{t_0}^{t_1} \phi^2(x, t) f(t) \, dt]^\alpha}. \tag{1.5}$$

(λ *is well defined by assumptions* (c) *and* (g).) *Also, let*

$$\inf_u \int_{t_0}^{t_1} \phi^1(u, t) \, dt = J \quad \textit{subject to} \left[\int_{t_0}^{t_1} \phi^2(x, t) f(t) \, dt\right]^\alpha = M > 0. \tag{1.6}$$

Then $\lambda = J/M$.

Proof. Clearly $J/M \geqq \lambda$. To reverse the inequality, let $\tilde{u}$ be such that $F(\tilde{x}, \tilde{u}) \leqq \lambda + \varepsilon$ for some $\varepsilon \geqq 0$. Let $[\int_{t_0}^{t_1} \phi^2(\tilde{x}, t) f(t) \, dt]^\alpha = \tilde{M}$ ($< \infty$ by assumptions (c), (d) and (e)), and $\mu = (M/\tilde{M})^{1/k}$. Then $(\mu\tilde{x}, \mu\tilde{u})$ obeys all the constraints by the regularity of the constraints, and by assumption (f), $[\int_{t_0}^{t_1} \phi^2(\mu\tilde{x}, t) f(t) \, dt]^\alpha = M$ and $F(\mu\tilde{x}, \mu\tilde{u}) \leqq \lambda + \varepsilon$. By (1.6), $J/M \leqq \lambda + \varepsilon$. Since ε is arbitrary, the conclusion of the proposition follows. □

THEOREM 1.1. *Consider the system* (1.1) *and* (1.2) *along with assumptions* (a)–(g). *Also assume that the constraints on* x *and* u *are regular. Then there exists a control among all admissible controls that minimizes* (1.2).

Proof. By Proposition 1.1, it is sufficient to exhibit a minimizing control among all admissible controls for which $[\int_{t_0}^{t_1} \phi^2(x, t) f(t) \, dt]^\alpha = M > 0$ and the trajectories of which satisfy (1.1) and all the constraints. Let $J = \inf_u \int_{t_0}^{t_1} \phi^1(u, t) \, dt$ subject to $[\int_{t_0}^{t_1} \phi^2(x, t) f(t) \, dt]^\alpha = M$. Choose $\{(x^i, u^i)\}$ such that $\lim_{i\to\infty} \int_{t_0}^{t_1} \phi^1(u^i, t) \, dt = J$ with $[\int_{t_0}^{t_1} \phi^2(x^i, t) f(t) \, dt]^\alpha = M$ for each i. By assumption (d), $\{u^i\}$ form a bounded sequence in $L_p(t_0, t_1)$, and hence a subsequence, still denoted by $\{u^i\}$, converges weakly to some u^0 in $L_p(t_0, t_1)$. Let x^0 be the response of (1.1) to u^0. By assumption (a) and by the weak convergence, $x^i(t) \to x^0(t)$ for all $t \in [t_0, t_1)$. By the regularity of the constraints, $x^0(t)$ obeys all the constraints. Assumption (b) $\Rightarrow \phi^2(x^i(t), t)$ converges to $\phi^2(x^0(t), t)$ for all t in $[t_0, t_1)$. Since $\|u^i\|_p \leqq K$ for some $K < \infty$, by assumption (e) and by the Lebesgue dominated convergence theorem,

$$\left[\int_{t_0}^{t_1} \phi^2(x^0(t), t) f(t) \, dt\right]^\alpha = \lim_{i\to\infty} \left[\int_{t_0}^{t_1} \phi^2(x^i(t), t) f(t) \, dt\right]^\alpha = M.$$

If $t_1 < \infty$, then we have by assumption (b) (see [6, p. 209]),

$$\int_{t_0}^{t_1} \phi^1(u^0, t) \, dt \leqq \liminf_{i\to\infty} \int_{t_0}^{t_1} \phi^1(u^i, t) \, dt = J.$$

If $t_1 = \infty$, $[t_0, t_1) = \cup_{j=0}^{\infty} [t_0 + j, t_0 + j + 1]$. We have on each subinterval,

$$a_j = \int_{t_0+j}^{t_0+j+1} \phi^1(u^0, t) \, dt \leqq \liminf_{i\to\infty} \underbrace{\int_{t_0+j}^{t_0+j+1} \phi^1(u^i, t) \, dt}_{a_{ji}}$$

$$\Rightarrow \sum_{j=0}^{\infty} a_j \leqq \sum_{j=0}^{\infty} \liminf_{i\to\infty} a_{ji} \leqq \liminf_{i\to\infty} \sum_{j=0}^{\infty} a_{ji} = J,$$

by Fatou's lemma. Thus the proof is complete. □

2. Finite interval case with integrable $f(t)$. In the next few sections, we will develop the necessary conditions for optimal controls in various situations. Let $I = [t_0, t_1]$, $t_1 < \infty$, and $C(I)$ be the space of all continuous $x(t) = (x_1(t), \cdots, x_n(t))$: $I \to R^n$ such that $\|x\| = \max\{\|x_1\|, \cdots, \|x_n\|\}$, where $\|x_i\| = \sup_{[t_0,t_1]} |x_i(t)|$. Once again, consider the n-dimensional system

$$\frac{dx}{dt} = \underset{n\times n}{A}(t)x + \underset{n\times r}{B}(t)u, \tag{2.1}$$

with

$$x(t_0) = c, \qquad x(t_1) = d, \qquad t_1 < \infty, \tag{2.2}$$

$$F(x, u) = \frac{\int_{t_0}^{t_1} \phi^1(u, t)\, dt}{[\int_{t_0}^{t_1} \phi^2(x, t) f(t)\, dt]^{\alpha}}, \tag{2.3}$$

where $f(t)$ is integrable and $\alpha \in R$. We make the following assumptions.

(a) $A(t)$ and $B(t)$ are continuous $n \times n$ and $n \times r$ matrix functions respectively.

(b) Admissible controls are measurable functions such that $\int_{t_0}^{t_1} \phi^1(u, t)\, dt < \infty$.

(c) $\phi^1(u, t) \geqq a|u|^p$, $a > 0$, $p > 1$, and $\phi^2(x(t), t) f(t) \geqq 0$ a.e. on $[t_0, t_1]$ for any trajectory $x(t)$ which is the response to some $u \in L_p(t_0, t_1)$.

(d) Let ϕ^1 be continuously differentiable in u and ϕ^2 be continuously differentiable in x, and both be measurable in t. Moreover, let $\phi^1_u(u(t), t) \in L_q(t_0, t_1)$ for all $u \in L_p(t_0, t_1)$, $(1/p) + (1/q) = 1$. Also, let ϕ^2_x be bounded for bounded x, the bound being uniform for almost all t.

(e) Let ϕ^1_u and ϕ^2_x be locally Lipschitzian in u and x respectively, i.e., there exist $\delta > 0$, K_1, $K_2 > 0$ depending on (x, u) such that for all $\bar{x} \in C(I)$ and $\bar{u} \in L_p(t_0, t_1)$ with $\|\bar{x}\| \leqq \delta$, $\|\bar{u}\|_p \leqq \delta$, we have

$$|\phi^2_x(x + \bar{x}, t) - \phi^2_x(x, t)| \leqq K_1 \|\bar{x}\|$$

a.e., and

$$\|\phi^1_u(u + \bar{u}, t) - \phi^1_u(u, t)\|_q \leqq K_2 \|\bar{u}\|_p.$$

(f) $(x^0(t), u^0(t))$ minimizes (2.3) subject to (2.1) and (2.2)

$$\Rightarrow \infty > \int_{t_0}^{t_1} \phi^1(u^0(t), t)\, dt > 0, \qquad \infty > \int_{t_0}^{t_1} \phi^2(x^0(t), t) f(t)\, dt > 0.$$

(g) The pair $(A(t), B(t))$ is completely controllable (see [5] for the definition of complete controllability and related discussion).

By assumption (f), we can consider the alternative cost functional

$$G(x, u) = \ln \int_{t_0}^{t_1} \phi^1\, dt - \alpha \ln \int_{t_0}^{t_1} \phi^2 f(t)\, dt \tag{2.4}$$

in place of (2.3). In order to establish our necessary conditions, we will make use of the Dubovitskii–Milyutin theorem [3, Thm. 6.1]. For the definitions of various terms, we refer the reader to [3]. If K is a cone in a Banach space E, we mean by the dual cone K^* the set $\{g \in E^* | \ g(x) \geqq 0 \text{ for all } x \in K\}$. Now we state the necessary conditions for an optimal control.

THEOREM 2.1. *Consider the system* (2.1)–(2.3) *along with assumptions* (a)–(g). *Suppose that* $(x^0(t), u^0(t))$ *minimizes* (2.3). *Then there exists* $\psi(t) \in C(I)$ *such that*

$$\frac{d\psi}{dt} = -A^T(t)\psi - \lambda \phi^2_x(x^0, t) f(t), \tag{2.5}$$

where

$$\lambda = \frac{\int_{t_0}^{t_1} \phi^1(u^0, t)\, dt}{\int_{t_0}^{t_1} \phi^2(x^0, t) f(t)\, dt} \tag{2.6}$$

and

$$\phi_u^1(u^0(t), t) - \alpha B^T \psi(t) = 0 \quad a.e.\ on \quad [t_0, t_1]. \tag{2.7}$$

Proof. Let $E = C(I) \times L_p(I)$. Our admissible controls form a subset of $L_p(I)$. But by assumption (f), the optimal cost is finite and hence we can regard u^0 to be optimal with respect to those controls in $L_p(I)$ whose trajectories obey (2.1) and (2.2). Thus, we take our space of controls to be $L_p(I)$.

(a) *Cone of directions of decrease.* By assumptions (d) and (e), the Fréchet derivative of the functional G in (2.4) is given by

$$G'(x^0, u^0)(x, u) = \frac{\int_{t_0}^{t_1} (\phi_u^1(u^0, t), u)\, dt}{\int_{t_0}^{t_1} \phi^1(u^0, t)\, dt} - \alpha \frac{\int_{t_0}^{t_1} (\phi_x^2(x^0, t), x) f(t)\, dt}{\int_{t_0}^{t_1} \phi^2(x^0, t) f(t)\, dt}. \tag{2.8}$$

By [3, Thm. 7.5], $(x(t), u(t))$ lies in the cone K_0 of directions of decrease in E if and only if $G'(x^0, u^0)(x, u) < 0$. By assumption (f), $(x, u) \in K_0$ if and only if

$$\int_{t_0}^{t_1} (\phi_u^1(u^0, t), u)\, dt - \alpha\lambda \int_{t_0}^{t_1} (\phi_x^2(x^0, t), x) f(t)\, dt < 0, \tag{2.9}$$

where λ is defined by (2.6). If $K_0 \neq \phi$, then by [3, Thm. 10.2], for any $g_0 \in K_0^*$,

$$g_0(x, u) = -\lambda_0 \left\{ \int_{t_0}^{t_1} (\phi_u^1(u^0, t), u)\, dt - \alpha\lambda \int_{t_0}^{t_1} (\phi_x^2(x^0, t), x) f(t)\, dt \right\}, \qquad \lambda_0 \geqq 0. \tag{2.10}$$

(b) *Cone of tangent directions.* To find the tangent directions in E at (x^0, u^0), we will apply the results of [3, Lecture 9]. Let

$$Q = \left\{ (x, u) \in E \,|\, x(t) = \Phi(t)c + \Phi(t) \int_{t_0}^{t} \Phi^{-1}(s) B(s) u(s)\, ds, \right. \\ \left. t_0 \leqq t \leqq t_1,\ x(t_1) = d \right\}, \tag{2.11}$$

where $\Phi(t)$ is a fundamental matrix of $\dot{y} = A(t)y$ with $\Phi(t_0) = I$, and let

$$P(x, u) = \left(x(t) - \Phi(t)c - \Phi(t) \int_{t_0}^{t} \Phi^{-1}(s) B(s) u(s)\, ds,\ x(t_1) \right), \tag{2.12}$$

which maps E into $C(I) \times R^n$. Also

$$P'(x^0, u^0)(x, u) = \left(x(t) - \Phi(t) \int_{t_0}^{t} \Phi^{-1}(s) B(s) u(s)\, ds,\ x(t_1) \right), \tag{2.13}$$

where $P'(x^0, u^0) : E \to C(I) \times R^n$. We wish to show that $P'(x^0, u^0)$ is onto.

Let $(a(t), b) \in C(I) \times R^n$. Since $(A(t), B(t))$ is completely controllable, select $\tilde{u} \in L_p(I)$ such that

$$\Phi(t) \int_{t_0}^{t} \Phi^{-1}(s) B(s) \tilde{u}(s)\, ds = b - a(t_1).$$

Set $\tilde{x}(t) = \Phi(t) \int_{t_0}^{t} \Phi^{-1}(s) B(s) \tilde{u}(s)\, ds + a(t)$. Then $P'(x^0, u^0)(\tilde{x}, \tilde{u}) = (a(t), b)$. By [3, Thm. 9.1] the set K_1 of tangent directions at (x^0, u^0) is given by $\{(x, u) \in$

$E \mid P'(x^0, u^0)(x, u) = 0\}$. Thus K_1 consists of all (x, u) satisfying

$$\frac{dx}{dt} = A(t)x + B(t)u, \qquad x(t_0) = 0, \tag{2.14}$$

$$x(t_1) = 0. \tag{2.15}$$

Let $L_1 \subseteq E$ denote the pairs satisfying (2.14) and $L_2 \subseteq E$ the set of (x, u) satisfying (2.15). It follows that (see [3, Lecture 12]) $K_1^* = L_1^* + L_2^*$, and if $g_2 \in L_2^*$, then $g_2(x, u) = a^T x(t_1)$ for some $a \in R^n$. If $g_1 \in L_1^*$, then $g_1(x, u) = 0$ for all $(x, u) \in L_1$, since L_1 is a subspace.

(c) *Application of Dubovitskii–Milyutin theorem.* The above theorem [3, Thm. 6.1] states that there exist $g_0 \in K_0^*$, $g_1 \in L_1^*$ and $g_2 \in L_2^*$, not all zero, such that for all $(x, u) \in E$,

$$g_0(x, u) + g_1(x, u) + g_2(x, u) = 0. \tag{2.16}$$

Let u be arbitrary and x be a solution of (2.14) for this u. Then $g_1(x, u) = 0$, and hence

$$-\lambda_0 \left\{ \int_{t_0}^{t_1} (\phi_u^1, u)\, dt - \alpha\lambda \int_{t_0}^{t_1} (\phi_x^2, x) f(t)\, dt \right\} + a^T x(t_1) = 0, \qquad \lambda_0 \geqq 0. \tag{2.17}$$

λ_0 has to be positive, because if $\lambda_0 = 0$, then from (2.17), $a^T x(t_1) = 0$. If $a = 0$, we would have $g_1 = g_2 = g_3 = 0$ by (2.16), which is not possible. If $a \neq 0$, by the complete controllability of the pair $(A(t), B(t))$ in (2.14), we can select some $x(t)$ for which $x(t_1) = a$, which gives the contradiction that $a^T a = 0$. Hence $\lambda_0 > 0$. (2.17) becomes

$$\int_{t_0}^{t_1} (\phi_u^1, u)\, dt - \alpha\lambda \int_{t_0}^{t_1} (\phi_x^2, x) f(t)\, dt - \frac{1}{\lambda_0} a^T x(t_1) = 0. \tag{2.18}$$

Define ψ by

$$\frac{d\psi}{dt} = -A^T\psi - \lambda\phi_x^2(x^0, t) f(t), \qquad \psi(t_1) = \frac{a}{\alpha\lambda_0}. \tag{2.19}$$

Then

$$\begin{aligned} \lambda \int_{t_0}^{t_1} (\phi_x^2 f(t), x)\, dt &= -\int_{t_0}^{t_1} \left(\frac{d\psi}{dt} + A^T\psi, x \right) dt \\ &= -\frac{a^T}{\alpha\lambda_0} x(t_1) + \int_{t_0}^{t_1} (\psi, B(t)u)\, dt. \end{aligned} \tag{2.20}$$

Thus (2.18) becomes

$$\int_{t_0}^{t_1} (\phi_u^1 - \alpha B^T\psi, u)\, dt = 0 \tag{2.21}$$

for arbitrary u. Hence

$$\phi_u^1 - \alpha B^T\psi = 0, \quad \text{a.e. on } [t_0, t_1]. \tag{2.22}$$

(d) *Case when $K_0 = \phi$.* If $K_0 = \phi$, then

$$\int_{t_0}^{t_1} (\phi_u^1, u)\, dt - \alpha\lambda \int_{t_0}^{t_1} (\phi_x^2, x) f(t)\, dt = 0 \tag{2.23}$$

for all $(x, u) \in E$, and we can proceed on as above, letting $\psi(t_1) = 0$. □

It is possible to extend our necessary conditions to more general functionals than in (2.3); for example, to functionals of the form

$$\frac{\int_{t_0}^{t_1} \phi^1(x, u, t)\, dt}{[\int_{t_0}^{t_1} \phi^2(x, u, t) f(t)\, dt]^{\alpha}}.$$

However, additional assumptions have to be made on ϕ^1 and ϕ^2.

3. Finite interval case with nonintegrable $f(t)$. We consider the scalar system defined by

$$\dot{x} = u, \qquad x(t_0) = 0, \qquad x(t_1) = 0 \text{ or free}, \tag{3.1}$$

with

$$F(x, u) = \frac{\int_{t_0}^{t_1} \phi^1(u, t)\, dt}{[\int_{t_0}^{t_1} \phi^2(x, t) f(t)\, dt]^{\alpha}}, \qquad \alpha \in R, \qquad f(t) \geqq 0 \quad \text{a.e.} \tag{3.2}$$

Let $I = [t_0, t_1]$, $L_p(I) = \{u \mid \int_{t_0}^{t_1} |u|^p\, dt < \infty\}$, $\tilde{L}_p(I) = \{u \in L_p(I) \mid \int_{t_0}^{t_1} u\, dt = 0\}$ and $L_k(I, \mu) = \{x \mid \int_{t_0}^{t_1} |x|^k\, d\mu < \infty,\ d\mu = f(t)\, dt\}$.

We make the following assumptions:

(a) Admissible controls are measurable functions such that

$$\int_{t_0}^{t_1} \phi^1(u, t)\, dt < \infty.$$

(b) $\phi^1(u, t) \geqq a|u|^p$, $a > 0$, $p > 1$ for almost all $t \in I$.

(c) $(x^0(t), u^0(t))$ minimizes (3.2) subject to (3.1) $\Rightarrow$

$$\infty > \int_{t_0}^{t_1} \phi^1(u^0, t)\, dt > 0, \qquad \infty > \int_{t_0}^{t_1} \phi^2(x^0, t) f(t)\, dt > 0.$$

(d) Let ϕ^1 be continuously differentiable in u, ϕ^2 be continuously differentiable in x, and both be measurable in t.

(e) $\phi^2(x, t) \geqq b|x|^k$, $b > 0$, $k > 1$ for almost all $t \in I$.

(f) $\phi^1_u \in L_q(I)$ for all $u \in L_p(I)$, $(1/p) + (1/q) = 1$, and $\phi^2_x \in L_m(I, \mu)$, $(1/k) + (1/m) = 1$, whenever $x \in L_k(I, \mu)$.

(g) There exist δ, K_1, $K_2 > 0$ depending on (x, u) such that for all $\|h\|_k \leqq \delta$ and $\|\bar{u}\|_p \leqq \delta$ we have

$$\|\phi^1_u(u + \bar{u}, t) - \phi^1_u(u, t)\|_q \leqq K_1 \|\bar{u}\|_p,$$

$$\|\phi^2_x(x + h, t) - \phi^2_x(x, t)\|_m \leqq K_2 \|h\|_k.$$

(h) (i) If the boundary condition $x(t_1) = 0$, then $x \in L_k(I, \mu)$ for all $u \in \tilde{L}_p(I)$.

(ii) If $x(t_1)$ is free, then $x \in L_k(I, \mu)$ for all $u \in L_p(I)$.

THEOREM 3.1. *Consider* (3.1) *and* (3.2) *along with assumptions* (a)–(h). *Suppose* $(x^0(t), u^0(t))$ *is optimal. Then there exists* $\psi(t)$ *such that*

$$\frac{d\psi}{dt} = -\lambda \phi^2_x(x^0, t) f(t), \qquad \lambda = \frac{\int_{t_0}^{t_1} \phi^1(u^0, t)\, dt}{\int_{t_0}^{t_1} \phi^2(x^0, t) f(t)\, dt} \tag{3.3}$$

and

$$\phi^1_u(u^0, t) - \alpha\psi(t) = \text{constant a.e. on } I. \tag{3.4}$$

Moreover, if $x(t_1)$ *is free, the constant in* (3.4) *can be taken to be zero.*

Proof. Case 1. $x(t_1) = 0$.

Let $E = L_k(I, \mu) \times \tilde{L}_p(I)$. By assumption (c), we can consider the alternative cost functional

$$G(x, u) = \ln \int_{t_0}^{t_1} \phi^1(u, t)\, dt - \alpha \ln \int_{t_0}^{t_1} \phi^2(x, t) f(t)\, dt. \tag{3.5}$$

By assumptions (c), (d), (e), (f) and (g), the Fréchet derivative of G at (x^0, u^0) is defined by

$$G'(x^0, u^0)(x, u) = \frac{\int_{t_0}^{t_1} \phi_u^1(u^0, t) u\, dt}{\int_{t_0}^{t_1} \phi^1(u^0, t)\, dt} - \alpha \frac{\int_{t_0}^{t_1} \phi^2(x^0, t) x f(t)\, dt}{\int_{t_0}^{t_1} \phi^2(x^0, t) f(t)\, dt}. \tag{3.6}$$

As in the proof of Theorem 2.1, if $g_0 \in K_0^*$ where K_0 is the cone of directions of decrease, and if $K_0 \neq \phi$, then

$$g_0(x, u) = -\lambda_0 \left\{ \int_{t_0}^{t_1} \phi_u^1 u\, dt - \alpha\lambda \int_{t_0}^{t_1} \phi_x^2 x f(t)\, dt \right\}, \qquad \lambda_0 \geqq 0. \tag{3.7}$$

To get the tangent directions, the relevant equations are (using the same notation as in the proof of Theorem 2.1):

$$Q = \left\{ (x, u) \in E \,\middle|\, x(t) = \int_{t_0}^{t} u(s)\, ds,\ x(t_1) = 0 \right\}, \tag{3.8}$$

$$P(x, u) = x(t) - \int_{t_0}^{t} u(s)\, ds, \tag{3.9}$$

and

$$P'(x^0, u^0)(x, u) = x(t) - \int_{t_0}^{t} u(s)\, ds. \tag{3.10}$$

Note that by assumption (h), $\int_{t_0}^{t} u(s)\, ds$ is in $L_k(I, \mu)$ and hence P maps E into $L_k(I, \mu)$. Letting $u = 0$ in (3.10), we see that $P'(x^0, u^0)$ is onto $L_k(I, \mu)$. Hence the tangent directions are given by $\{(x, u) \in E \,|\, dx/dt = u,\ x(t_0) = 0\}$, which automatically implies that $x(t_1) = 0$ since $u \in \tilde{L}_p(I)$. Proceeding as in the proof of Theorem 2.1, we get (after an application of the Dubovitskii–Milyutin theorem)

$$\int_{t_0}^{t_1} \phi_u^1 u\, dt - \alpha\lambda \int_{t_0}^{t_1} \phi_x^2 x f(t)\, dt = 0. \tag{3.11}$$

Define ψ by

$$\frac{d\psi}{dt} = -\lambda \phi_x^2(x^0, t) f(t). \tag{3.12}$$

ψ makes sense by assumption (f). By (3.11), (3.12) and the boundary conditions, it follows that

$$\int_{t_0}^{t_1} (\phi_u^1 - \alpha\psi) u\, dt = 0 \tag{3.13}$$

for any arbitrary $u \in \tilde{L}_p(I)$. (3.13) implies that $\phi_u^1(u^0, t) - \alpha\psi(t) = \text{constant}$ a.e. on I. The case where $K_0 = \phi$ can be handled easily.

Case 2. $x(t_1)$ is free.

Take $E = L_k(I, \mu) \times L_p(I)$ and proceed as in Case 1. In this case define ψ by

$$\frac{d\psi}{dt} = -\lambda \phi_x^2(x^0, t) f(t), \qquad \psi(t_1) = 0. \tag{3.14}$$

We get in place of (3.13) for any $u \in L_p(I)$

$$\int_{t_0}^{t_1} (\phi_u^1 - \alpha\psi)u \, dt = 0, \tag{3.15}$$

which implies that $\phi_u^1 - \alpha\psi = 0$ a.e. on I. □

4. Infinite interval case. Consider the scalar system

$$\dot{x} = u, \qquad x(t_0) = 0, \tag{4.1}$$

with

$$F(x, u) = \frac{\int_{t_0}^{\infty} \phi^1(u, t) \, dt}{[\int_{t_0}^{\infty} \phi^2(x, t) f(t) \, dt]^{\alpha}}, \qquad \alpha \in \boldsymbol{R}. \tag{4.2}$$

Let $I = [t_0, \infty)$. We make the following assumptions:

(a)–(g) are identical to those in § 3.

(h) Assume that $x \in L_k(I, \mu)$ for all $u \in L_p(I)$ and $\phi_x^2(x(t), t) f(t)$ is integrable for each $u \in L_p(I)$.

(i) $\lim_{t\to\infty} x(t) \int_{\infty}^{t} \phi_x^2(x(s), s) f(s) \, ds = 0$ for every $u \in L_p(I)$.

THEOREM 4.1. *Consider* (4.1) *and* (4.2) *along with assumptions* (a)–(i). *Suppose* $(x^0(t), u^0(t))$ *minimizes* (4.2). *Then there exists* $\psi(t)$ *such that*

$$\frac{d\psi}{dt} = -\lambda\phi_x^2(x^0, t) f(t), \qquad \lambda = \frac{\int_{t_0}^{\infty} \phi^1(u^0, t) \, dt}{\int_{t_0}^{\infty} \phi^2(x^0, t) f(t) \, dt}, \tag{4.3}$$

and

$$\phi_u^1(u^0, t) - \alpha\psi(t) = 0 \quad \textit{a.e. on } I. \tag{4.4}$$

Proof. Let $E = L_k(I, \mu) \times L_p(I)$. By assumption (c), we can consider the alternative cost functional

$$G(x, u) = \ln \int_{t_0}^{\infty} \phi^1(u, t) \, dt - \alpha \ln \int_{t_0}^{\infty} \phi^2(x, t) f(t) \, dt. \tag{4.5}$$

By assumptions (c), (d), (e), (f) and (g), the Fréchet derivative of G at (x^0, u^0) is defined by

$$G'(x^0, u^0)(x, u) = \frac{\int_{t_0}^{\infty} \phi_u^1(u^0, t) u \, dt}{\int_{t_0}^{\infty} \phi^1(u^0, t) \, dt} - \alpha \frac{\int_{t_0}^{\infty} \phi_x^2(x^0, t) x f(t) \, dt}{\int_{t_0}^{\infty} \phi^2(x^0, t) f(t) \, dt}. \tag{4.6}$$

If $g_0 \in K_0^*$, where K_0 is the cone of directions of decrease, and if $K_0 \neq \phi$, then

$$g_0(x, u) = -\lambda_0 \Big\{ \int_{t_0}^{\infty} \phi_u^1 u \, dt - \alpha\lambda \int_{t_0}^{\infty} \phi_x^2 x f(t) \, dt \Big\}, \qquad \lambda_0 \geqq 0. \tag{4.7}$$

To get the tangent directions, we have

$$Q = \Big\{ (x, u) \in E \Big| \, x(t) = \int_{t_0}^{t} u(s) \, ds \Big\}, \tag{4.8}$$

$$P(x, u) = x(t) - \int_{t_0}^{t} u(s) \, ds = P'(x^0, u^0)(x, u). \tag{4.9}$$

We now mimic the steps in the proofs of earlier theorems.

Applying the Dubovitskii–Milyutin theorem, we arrive at

$$\int_{t_0}^{\infty} \phi_u^1 u \, dt - \alpha\lambda \int_{t_0}^{\infty} \phi_x^2 x f(t) \, dt = 0. \tag{4.10}$$

Define ψ by

$$\frac{d\psi}{dt} = -\lambda \phi_x^2(x^0, t) f(t), \qquad \psi(\infty) = 0, \tag{4.11}$$

which makes sense by assumption (h). Now

$$\begin{aligned} -\lambda \int_{t_0}^{\infty} \phi_x^2 x f(t) \, dt &= \int_{t_0}^{\infty} \frac{d\psi}{dt} x \, dt \\ &= x\psi \Big|_{t_0}^{\infty} - \int_{t_0}^{\infty} u\psi \, dt. \end{aligned} \tag{4.12}$$

By assumption (i), $\lim_{t\to\infty} x\psi(t) = 0$. Thus (4.10) implies that, for any $u \in L_p(I)$,

$$\int_{t_0}^{\infty} (\phi_u^1 - \alpha\psi) u \, dt = 0. \tag{4.13}$$

By assumption (f), $\phi_u^1 \in L_q(I)$ and (4.12) implies that $\psi \in L_q(I)$ (since $\int_{t_0}^{\infty} u\psi \, dt < \infty$ for every $u \in L_p(I)$). Hence $\phi_u^1 - \alpha\psi = 0$ a.e. on I. The case where $K_0 = \phi$ can be easily handled. □

Several examples in the finite interval case are given in [8] and [9]. We now consider an example in the infinite interval case.

Example. If $\dot{x} \in L_2(1, \infty)$ and $x(1) = 0$, then for $\nu > 2$

$$\int_1^{\infty} \dot{x}^2 \, dt \geqq \lambda \int_1^{\infty} \frac{x^2}{t^\nu} \, dt, \tag{4.14}$$

where λ is the least positive number such that

$$\ddot{x} + \frac{\lambda x}{t^\nu} = 0, \qquad x(1) = 0 \tag{4.15}$$

has a nonzero solution on $[1, \infty)$ such that $\dot{x} \in L_2(1, \infty)$. (That a least positive number exists such that the above conditions are satisfied follows from our theory, as we will see.) Moreover, equality in (4.14) holds if and only if x is a solution of (4.15).

Proof. Let $\dot{x} = u$ and for $x \not\equiv 0$, consider the functional

$$F(x, u) = \frac{\int_1^{\infty} u^2 \, dt}{\int_1^{\infty} x^2/t^\nu \, dt}. \tag{4.16}$$

To verify assumption (e) of Theorem 1.1, observe that

$$\begin{aligned} |x(t)| &= \left| \int_1^t u(s) \, ds \right| \leqq t^{1/2} \|u\|_2, \\ \left| \frac{x^2}{t^\nu} \right| &\leqq \frac{1}{t^{\nu-1}} \|u\|_2^2, \qquad \nu > 2. \end{aligned} \tag{4.17}$$

Thus there exists an optimal u for (4.16). Also we have

$$|\phi_x^2(x(s), s)f(s)| = \left|\frac{2x(s)}{s^\nu}\right| \leqq Cs^{-\nu+(1/2)}, \tag{4.18}$$

$$\left|\int_\infty^t \phi_x^2 f(s)\, ds\right| \leqq Ct^{-\nu+(3/2)}, \tag{4.19}$$

and

$$\left|x(t)\int_\infty^t \phi_x^2 f(s)\, ds\right| \leqq \tilde{C}t^{-\nu+2}. \tag{4.20}$$

Equations (4.17), (4.18) verify assumptions (f) and (h) of Theorem 4.1, and (4.20) verifies assumption (i). Thus, applying Theorem 4.1, we get $u = \psi/2$, $d\psi/dt = -2\lambda\, x/t^\nu$, where u is optimal and λ is the value of (4.16) for this u. Hence,

$$\ddot{x} + \lambda\frac{x}{t^\nu} = 0, \qquad x(1) = 0. \tag{4.21}$$

Thus, if equality holds in (4.14), then x satisfies (4.21). Now suppose $x \not\equiv 0$ satisfies (4.21) for some $\lambda = \bar{\lambda} > 0$, $\dot{x} \in L_2(I)$. Then $\int_1^\infty \dot{x}^2\, dt = x\dot{x}\,|_1^\infty - \int_1^\infty \ddot{x}x\, dt$, where $(x\dot{x})(\infty) = 0$ by assumption (i) of Theorem 4.1. Hence

$$\frac{\int_1^\infty \dot{x}\, dt}{\int_1^\infty x^2/t^\nu\, dt} = \bar{\lambda}. \tag{4.22}$$

So λ is the least positive value such that (4.21) has a nonzero solution with $x(1) = 0$ and $\dot{x} \in L_2(1, \infty)$. We also deduce that if x is a solution of (4.21) for the optimal λ, then equality holds in (4.14). This concludes the proof.

For the special case of $\nu = 3$, we get

$$\ddot{x} + \frac{\lambda x}{t^3} = 0, \qquad x(1) = 0, \qquad \dot{x} \in L_2(1, \infty), \tag{4.23}$$

and we are seeking the least positive λ such that (4.23) has a nonzero solution x. Letting $\tau = t^{-1}$, the conditions become

$$\frac{d^2x}{d\tau^2} + \frac{2}{\tau}\frac{dx}{d\tau} + \frac{\lambda}{\tau}x = 0, \qquad x(1) = 0, \qquad \int_0^1 \left(\frac{dx}{d\tau}\right)^2 \tau^2\, d\tau < \infty. \tag{4.24}$$

This is a special case of equation (3) on p. 97 of [10], and the general solution is given by

$$x(\tau) = A\tau^{-1/2}J_1(\sqrt{4\lambda\tau}) + B\tau^{-1/2}Y_1(\sqrt{4\lambda\tau}), \tag{4.25}$$

where J_1 and Y_1 are Bessel functions of the first and second kinds respectively. The condition that $\int_0^1 (dx/d\tau)^2\, \tau^2\, d\tau$ is finite eliminates the solution $\tau^{-1/2}Y_1(\sqrt{4\lambda\tau})$. Since $x(1) = 0$, we have $J_1(\sqrt{4\lambda}) = 0$. Since we are looking for the least positive λ, $\sqrt{4\lambda}$ has to be the first positive zero of J_1, which is approximately 3.8317059702. Thus, λ is approximately 3.67049266.

We finally remark that almost the same proof as that of Theorem 4.1 leads to the necessary conditions when $x(t_0) = c \neq 0$.

REFERENCES

[1] A. YA. DUBOVITSKII AND A. A. MILYUTIN, *Extremum problems in the presence of restrictions*, U.S.S.R. Comput. Math. Math. Phys., 5 (1965), pp. 1–80.
[2] N. DUNFORD AND J. SCHWARTZ, *Linear Operators, Part* 1, Interscience, New York, 1958.
[3] I. V. GIRSANOV, *Lecture Notes in Economics and Mathematical Systems* 67, Springer-Verlag, New York, 1972.
[4] G. H. HARDY, J. E. LITTLEWOOD AND G. PÓLYA, *Inequalities*, Cambridge University Press, New York, 1973.
[5] R. E. KALMAN, P. L. FALB AND M. A. ARBIB, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1967.
[6] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
[7] M. B. SUBRAHMANYAM, *Necessary conditions for minimum in problems with nonstandard cost functionals*, J. Math. Anal. Appl., 60, (1977), pp. 601–616.
[8] ———, *On applications of control theory to integral inequalities*, J. Math. Anal. Appl., 77 (1980), pp. 47–59.
[9] ———, *A control problem with application to integral inequalities*, J. Math. Anal. Appl., to appear.
[10] G. N. WATSON, *A Treatise on the Theory of Bessel Functions*, Macmillan, New York, 1944.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 4, July 1981

0363-0129/81/1904-0005 $01.00/0

DISTURBANCE DECOUPLING BY MEASUREMENT FEEDBACK WITH STABILITY OR POLE PLACEMENT*

JAN C. WILLEMS† AND CHRISTIAN COMMAULT‡

Abstract. In this paper we solve the disturbance decoupling problem by measurement feedback and requiring stability or pole placement on the closed loop system. The problem is attacked using the geometric approach through the concepts of $A(\text{mod } \mathcal{B})$-invariant and controllability subspaces and their duals, $A|\mathcal{K}$-invariant and complementary observability subspaces. The solution of this problem has an interesting structure consisting of a feedback processor which decomposes into (i) a disturbance decoupling loop; (ii) a disturbance input stabilization or pole placement loop, and (iii) a controlled output stabilization or pole placement loop.

1. Introduction. Consider the dynamical system with signal flow graph depicted in Fig. 1.

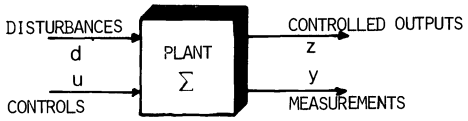


FIG. 1

If this system is controlled by means of the feedback processor shown in Fig. 2,

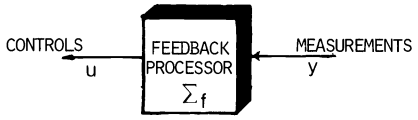


FIG. 2

then one obtains the closed loop system shown in Fig. 3.

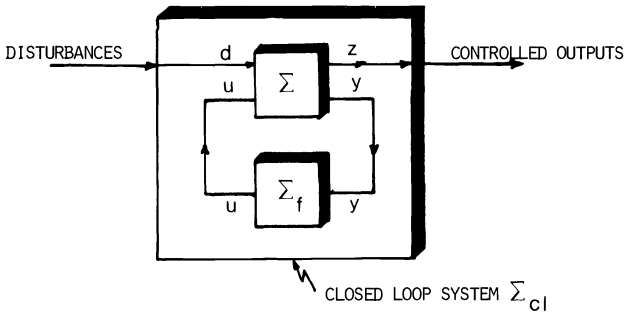


FIG. 3

One of the most easily motivated control synthesis questions is the problem of designing a feedback processor such that in the closed loop system,

* Received by the editors December 26, 1979, and in revised form June 2, 1980.

† Mathematics Institute, P.O. Box 800, 9700 AV. Groningen, the Netherlands.

‡ Ecole Nationale Supérieure d'Ingénieurs Electriciens de Grenoble, Laboratoire d'Automatique, Domaine Universitaire, BP 46, 38402 St. Martin d'Hères, France.

(i) The disturbances are completely decoupled from the controlled outputs; and
(ii) the closed loop system is internally stable or
(ii)′ (in the linear case) the closed loop poles may be arbitrarily assigned.

We will call these problems, following the acronym perversion propagated in [1]: DDPM (for (i)), the disturbance decoupling problem with measurement feedback; DDPMS (for (i) and (ii)), the disturbance decoupling problem with measurement feedback and stability; and DDPMPP (for (i) and (ii))′, the disturbance decoupling problem with measurement feedback and pole placement.

The disturbance decoupling problem in all its variations has been studied extensively before, and has motivated much of the development of the geometric approach in linear (and recently also in nonlinear) system theory. However, the early papers in this area have primarily been about disturbance decoupling using state feedback with or without stability or pole placement requirements [1, §§ 4.3, 5.6]. These results are based on the concepts of $A(\text{mod } \mathscr{B})$-invariant and controllability subspaces. There have also been a number of papers on DDEP, the disturbance decoupled estimation problem, or, what amounts to the same thing, the unknown input observer design problem (see [2], [3], and for earlier references, [4], [5], [6]). This problem will be treated in § 4. The crucial concepts in this context are those of $A|\mathscr{K}$-invariant and complementary observability subspaces. These are the duals of $A(\text{mod } \mathscr{B})$-invariant and controllability subspaces. They may be introduced by formal dualization (see [1, Ex. 5.17], or [2], where DDEP is solved this way) but they can also be defined directly, in a more intrinsic way in connection with observer synthesis questions [3], [5], [6].

In most industrial applications it will not be possible to assume that all the state variables are measured. Consequently, there is a direct practical motivation for studying the disturbance decoupling problem in the context of measurement feedback. Recently, in fact, DDPM has been solved in [7] and in [3]. Actually DDPM had already been formulated by Basile and Marro who, for this purpose, introduced the notions of *controlled* and *conditioned* invariant subspaces (we will call these $A(\text{mod } \mathscr{B})$- and $A|\mathscr{K}$-invariant subspaces) and they actually obtained as necessary conditions the conditions which, as shown in [3], [7], are in fact sufficient and hence lead to a synthesis for DDPM.

In all of the above references, the stability or pole placement question was not considered. It goes without saying that in applications one will need to consider also the stability aspects. In the present paper we will solve this problem (see (ii) and (iii) of our theorem).

It is quite surprising that DDPMS and DDPMPP have not been solved before even though their solution has been very much in reach, through the combined results in the work of Wonham [1], Basile and Marro [6] and the compensator design by output feedback of Brasch and Pearson [8] (see also [1, § 2.8]). The solution which we have obtained is in a sense what could have been conjectured from [3] or [7]. However, the resulting synthesis is a rather intricate and complex one.

We have attempted to make the paper reasonably self contained. Given the potential practical interest in this problem, one could hope that this true culmination of the disturbance decoupling circle of ideas ought to serve as the theoretical basis for some convincing specific applications.

We would like to emphasize that the disturbances could be also state or parameter dependent. The theorem which will be obtained also gives disturbance decoupling when the disturbance is of the form $d((\mathscr{F}x(\cdot))(t), \alpha, t)$, with $\mathscr{F}$ an (unknown) dynamic function of the state and α an unknown parameter.

2. Mathematical problem formulation. Consider the plant equations given by

$$\Sigma: \qquad \dot{x} = Ax + Bu + Gd, \qquad y = Cx, \qquad z = Hx,$$

with $x \in \mathbb{R}^n =: \mathscr{X}$, the state, $u \in \mathbb{R}^m =: \mathscr{U}$, the control, $d \in \mathbb{R}^q =: \mathscr{D}$, the disturbance, $y \in \mathbb{R}^p =: \mathscr{Y}$, the measurement and $z \in \mathbb{R}^l =: \mathscr{Z}$, the controlled output.

The DDPM problem is to find (real) feedback matrices $\{F, E, M, N\}$ defining the feedback processor

$$\Sigma_f: \qquad \dot{w} = Fw + Ey, \qquad u = Mw + Ny,$$

with $w \in \mathbb{R}^k =: \mathscr{W}$, the state of the feedback processor, such that the closed loop system $\Sigma_{cl} := \Sigma \times \Sigma_f$|feedback:

$$\Sigma_{c1}: \qquad \left[\begin{array}{c} \dot{x} \\ \hline \dot{w} \end{array}\right] = \left[\begin{array}{c|c} A + BNC & BM \\ \hline EC & F \end{array}\right] \left[\begin{array}{c} x \\ \hline w \end{array}\right] + \left[\begin{array}{c} G \\ \hline 0 \end{array}\right] d, \qquad z = [H \mid 0] \left[\begin{array}{c} x \\ \hline w \end{array}\right],$$

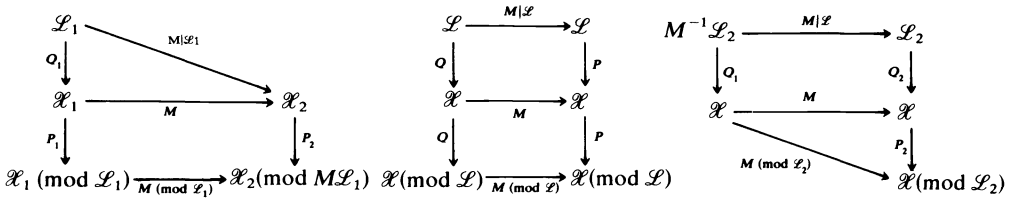
which may be written compactly as $\dot{x}^e = A^e x^e + G^e d$, $z = H^e x^e$, has zero transfer function, $H^e(Is - A^e)^{-1} G^e = 0$; i.e., the controlled output z is influenced only by the initial conditions and not by the disturbances d.

DDPMS requires in addition some conditions on the spectrum of A^e, $\sigma(A^e)$. This stability requirement is modelled, as usual, by requiring $\sigma(A^e) \subset \mathbb{C}_g$ with $\mathbb{C}_g$ a given subset of the complex plane $\mathbb{C}$ which is symmetric ($\{\lambda \in \mathbb{C}_g\} \Leftrightarrow \{\bar{\lambda} \in \mathbb{C}_g\}$; $^-$ denotes the complex conjugate), and which contains at least one point of the real axis. Simple asymptotic stability is thus obtained by taking $\mathbb{C}_g = \{\lambda \in \mathbb{C} | \operatorname{Re} \lambda < 0\}$.

DDPMPP requires pole placement in the sense that for any $\mathbb{C}_g$ which is symmetric and contains at least one point of the real axis it should be possible to achieve $\sigma(A^e) \subset \mathbb{C}_g$. (The results essentially imply that the closed loop characteristic polynomial can be chosen arbitrarily, provided that this characteristic polynomial have a sufficiently high degree and can be factored into two real factors of the right degree. These details we leave to the reader to fill in.

Some notation.

1. We will throughout use lower case letters for vectors, capitals for matrices and linear operators, and script for linear subspaces and vector spaces. If $M: \mathscr{X}_1 \to \mathscr{X}_2$ and $\mathscr{L}_1 \subset \mathscr{X}_1$, then $M|\mathscr{L}_1: \mathscr{L}_1 \to \mathscr{X}_2$, denotes $l_1 \mapsto Ml_1$, while $M(\operatorname{mod} \mathscr{L}_1): \mathscr{X}_1(\operatorname{mod} \mathscr{L}_1) \to \mathscr{X}_2(\operatorname{mod} M\mathscr{L}_1)$ denotes $x_1(\operatorname{mod} \mathscr{L}_1) \mapsto (Mx_1)(\operatorname{mod} M\mathscr{L}_1)$. If $\mathscr{L}_2 \subset \mathscr{X}_2$, then $M|\mathscr{L}_2: M^{-1}\mathscr{L}_2 \to \mathscr{X}_2$ denotes $l_2 \mapsto Ml_2$, while $M(\operatorname{mod} \mathscr{L}_2): \mathscr{X}_1 \to \mathscr{X}_2(\operatorname{mod} \mathscr{L}_2)$ denotes $x_1 \mapsto (Mx_1)(\operatorname{mod} \mathscr{L}_2)$. If $M: \mathscr{X} \to \mathscr{X}$ and $\mathscr{L} \subset \mathscr{X}$ is M-invariant then $M|\mathscr{L}: \mathscr{L} \to \mathscr{L}$ denotes $l \mapsto Ml$, while $M(\operatorname{mod} \mathscr{L})$ denotes $x(\operatorname{mod} \mathscr{L}) \mapsto (Mx)(\operatorname{mod} \mathscr{L})$. With Q's representing canonical injections ($Q: x \mapsto x$) and P's canonical projections ($P: x \mapsto x(\operatorname{mod} \mathscr{L})$) these definitions may be visualized in the commutative diagrams



If $M: \mathscr{X} \to \mathscr{X}$ and $\mathscr{L}_1, \mathscr{L}_2 \subset \mathscr{X}$, then $\mathscr{L}_1$ is said to be $M(\text{mod } \mathscr{L}_2)$-*invariant* if, for all $l_1 \in \mathscr{L}_1$, $Ml_1 \in \mathscr{L}_1(\text{mod } \mathscr{L}_2)$. It is said to be $M|\mathscr{L}_2$-*invariant* if, for all $l \in \mathscr{L}_1 \cap \mathscr{L}_2$, $Ml \in \mathscr{L}_2$. Thus $\mathscr{L}_1$ is $M(\text{mod } \mathscr{L}_2)$-invariant iff $M\mathscr{L}_1 \subset \mathscr{L}_1 + \mathscr{L}_2$ and $M|\mathscr{L}_2$-invariant iff $M(\mathscr{L}_1 \cap \mathscr{L}_2) \subset \mathscr{L}_2$. These concepts, which are very natural in the context of a linear algebra, also turn out to have very natural system theoretical interpretations!

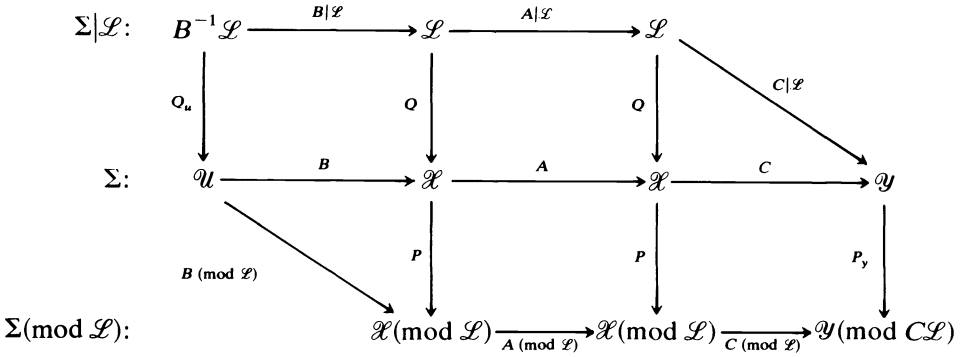
The spectrum of $M: \mathscr{X} \to \mathscr{X}$ is denoted by $\sigma(M)$. It is a set with multiplicity. The characteristic polynomial of M will be denoted by χ_M.

2. Consider the system Σ: $\dot{x} = Ax + Bu$, $y = Cx$, which we will sometimes denote by $\Sigma(A, B, C)$. Let $\mathscr{B} := \text{Im } B$ and $\mathscr{K} := \text{Ker } C$. Then $\mathscr{R} = \langle A|\mathscr{B}\rangle := \sum_{i=0}^{n-1} A^i\mathscr{B}$ denotes the reachable subspace, while $\mathscr{N} = \langle \mathscr{K}|A\rangle := \bigcap_{i=0}^{n-1} A^{-i}\mathscr{K}$ denotes the unobservable subspace. Both $\mathscr{R}$ and $\mathscr{N}$ are A-invariant subspaces. In fact, they are respectively the infimal A-invariant subspace containing $\mathscr{B}$ and the supremal A-invariant subspace contained in $\mathscr{K}$. The system is reachable iff $\mathscr{R} = \mathscr{X}$, and observable iff $\mathscr{N} = \{0\}$. If both conditions hold, then we will call the system *minimal*. If $A(\text{mod } \mathscr{R})$ and $A|\mathscr{N}$ are stable (relative to some $\mathbb{C}_g$), then we will call Σ *stabilizable and detectable.*

The *reachability index* of Σ, κ_Σ, is defined as the smallest integer l such that $\sum_{i=0}^{l-1} A^i\mathscr{B} = \mathscr{R}$, while the *observability index* of Σ, ν_Σ, is defined as the smallest integer l such that $\bigcap_{i=0}^{l-1} A^{-i}\mathscr{K} = \mathscr{N}$.

It is well known that $\{\Sigma \text{ stabilizable and detectable (relative to } \mathbb{C}_g)\} \Leftrightarrow \{\text{there exists a feedback compensator such that the closed loop system is stable (relative to } \mathbb{C}_g)\}$ and that $\{\Sigma \text{ minimal}\} \Leftrightarrow \{\text{for any } \mathbb{C}_g \text{ there exists a feedback compensator such that the closed loop poles are contained in } \mathbb{C}_g\}$. The required dimension of the feedback compensator achieving these properties is bounded above by $(\min(\kappa_\Sigma, \nu_\Sigma) - 1)$.

Let $\mathscr{L}$ be A-invariant. Then one may define the system $\Sigma|\mathscr{L}$, by $\Sigma|\mathscr{L} := \{A', B', C'\}$ with $A' := A|\mathscr{L}$, $B' := B|\mathscr{L}$ and $C' := C|\mathscr{L}$. Similarly the system $\Sigma(\text{mod } \mathscr{L})$ is defined by $\Sigma(\text{mod } \mathscr{L}) := \{A'', B'', C''\}$ with $A'' := A(\text{mod } \mathscr{L})$, $B'' := B(\text{mod } \mathscr{L})$ and $C'' := C(\text{mod } \mathscr{L})$. These are illustrated in the commutative diagram.



3. We will use, as standard notation, A_F for $A + BF$ and A^H for $A + HC$.

3. DDP. Consider the linear system Σ: $\dot{x} = Ax + Bu$, $y = Cx$. Let Σ_x denote all state trajectories of this system. Formally, $\Sigma_x := \{x: \mathbb{R} \to \mathscr{X} | x \text{ abs. cont. and } \dot{x}(t) - Ax(t) \in \mathscr{B} := \text{Im } B \text{ a.e.}\}$. A subspace $\mathscr{V}$ is said to be a *controlled invariant* subspace if for all $x_0 \in \mathscr{V}$ there exists $x \in \Sigma_x$ such that $x(0) = x_o$ and $x(t) \in \mathscr{V}$ for all t. A subspace $\mathscr{R}$ is said to be a *controllability subspace* if for all $x_0, x_1 \in \mathscr{R}$ there exists $T > 0$ and $x \in \Sigma_x$ such that $x(0) = x_0$, $x(T) = x_1$, and $x(t) \in \mathscr{R}$ for all t. We will denote the set of all controlled invariant subspaces by $\underline{\mathscr{V}}$ and the set of all controllability subspaces by $\underline{\mathscr{R}}$.

It is well known [1, Chaps. 4, 5] that $\{\mathscr{V} \text{ is controlled invariant}\} \Leftrightarrow \{\mathscr{V} \text{ is } A(\text{mod } \mathscr{B})\text{-invariant}\} \Leftrightarrow \{A\mathscr{V} \subset \mathscr{V} + \mathscr{B}\} \Leftrightarrow \{\text{there exists } F \text{ such that } \mathscr{V} \text{ is } A_F\text{-invariant}\}$. The family of

all such F's will be denoted by $\underline{F}(\mathcal{V})$. Furthermore: $\{\mathcal{R}$ is a controllability subspace$\}\Leftrightarrow$ {there exists F and $\mathcal{B}_1 \subset \mathcal{B}$ such that $\langle A_F|\mathcal{B}_1\rangle = \mathcal{R}\}\Leftrightarrow\{\mathcal{R} \in \underline{\mathcal{V}}$, and for any real polynomial p of degree $=\dim \mathcal{R}$, there exists F such that $\chi_{A_F|\mathcal{R}} = p\}\Leftrightarrow\{\mathcal{R} \in \underline{\mathcal{V}}$ and $\Sigma(A_F, B, -)|\mathcal{R}$ is controllable for all $\underline{F}(\mathcal{R})\}$.

Finally, if $\mathcal{V} \in \underline{\mathcal{V}}$ is such that there is an $F \in \underline{F}(\mathcal{V})$ such that $\sigma(A_F|\mathcal{V}) \subset \mathbb{C}_g$, then we call $\mathcal{V}$ a *stabilizable* controlled invariant subspace (relative to $\mathbb{C}_g$). The family of all stabilizable subspaces is denoted by $\underline{\mathcal{V}}_g$.

It is well known and easy to prove that $\underline{\mathcal{V}}$, $\underline{\mathcal{R}}$, and $\underline{\mathcal{V}}_g$ are closed under subspace addition, and thus there exists a supremal element of all elements of $\underline{\mathcal{V}}$, $\underline{\mathcal{R}}$, and $\underline{\mathcal{V}}_g$ contained in any given subspace $\mathcal{L}$ of $\mathcal{X}$. These subspaces will be denoted by $\mathcal{V}^*_{\mathcal{L}}$, $\mathcal{R}^*_{\mathcal{L}}$, and $\mathcal{V}^*_{g,\mathcal{L}}$, respectively. We recall the following algorithms for computing $\mathcal{V}^*_{\mathcal{L}}$ and $\mathcal{R}^*_{\mathcal{L}}$.

Algorithm (ISA) (the invariant subspace algorithm; see [1, p. 91]):

$$\mathcal{V}^{\mu+1}_{\mathcal{L}} := \mathcal{L} \cap A^{-1}(\mathcal{V}^{\mu}_{\mathcal{L}} + \mathcal{B}); \qquad \mathcal{V}^0 = \mathcal{X}.$$

Algorithm (ACSA) (the almost controllability subspace algorithm; see [1, p. 106] and [9], [10]:

$$\mathcal{R}^{\mu+1}_{\mathcal{L}} := \mathcal{L} \cap (A\mathcal{R}^{\mu}_{\mathcal{L}} + \mathcal{B}); \qquad \mathcal{R}^0 = \{0\}.$$

The sequence $\mathcal{V}^{\mu}_{\mathcal{L}}$ reaches, strictly decreasingly, its limit $\mathcal{V}^{\infty}_{\mathcal{L}} = \mathcal{V}^{\dim \mathcal{L}+1}_{\mathcal{L}}$ and $\mathcal{R}^{\mu}_{\mathcal{L}}$ reaches, strictly increasingly, its limit $\mathcal{R}^{\infty}_{\mathcal{L}} = \mathcal{R}^{\dim \mathcal{L}}_{\mathcal{L}}$. Furthermore

$$\mathcal{V}^*_{\mathcal{L}} = \mathcal{V}^{\infty}_{\mathcal{L}} \quad \text{and} \quad \mathcal{R}^*_{\mathcal{L}} = \mathcal{V}^{\infty}_{\mathcal{L}} \cap \mathcal{R}^{\infty}_{\mathcal{L}} = \mathcal{R}^{\infty}_{\mathcal{V}^{\infty}_{\mathcal{L}}}.$$

Computing a corresponding feedback matrix F such that $A_F\mathcal{V}^*_{\mathcal{L}} \subset \mathcal{V}^*_{\mathcal{L}}$ requires solving a set of linear equations. Finding an F such that

$$A_F\mathcal{R}^*_{\mathcal{L}} \subset \mathcal{R}^*_{\mathcal{L}} \quad \text{and} \quad \chi_{A_F|\mathcal{R}^*_{\mathcal{L}}} = p,$$

requires a standard pole placement computation.

One of the main applications of the above concepts is the disturbance decoupling problem. The main results are summarized in the following proposition.

PROPOSITION 1. (See [1, §'s 4.3 and 5.6]). *Consider* $\dot{x} = Ax + Bu + Gd$, $z = Hx$ *and the control law* $u = Fx$. *Then*:

(i) DDP. *There exists* F *such that* $H(Is - A_F)^{-1}G = 0$ *iff* $\operatorname{Im} G \subset \mathcal{V}^*_{\operatorname{Ker} H}$.

(ii) DDPS. *There exists* F *such that* $H(Is - A_F)^{-1}G = 0$ *and* $\sigma(A_F) \subset \mathbb{C}_g$ *iff* (A, B) *is stabilizable (relative to* $\mathbb{C}_g$*) and* $\operatorname{Im} G \subset \mathcal{V}^*_{g,\operatorname{Ker} H}$.

(iii) DDPPP. *For any* $\mathbb{C}_g$ *there exists* F *such that* $H(Is - A_F)^{-1}G = 0$ *and* $\sigma(A_F) \subset \mathbb{C}_g$, *iff* (A, B) *is reachable and* $\operatorname{Im} G \subset R^*_{\operatorname{Ker} H}$.

An important refinement of the above proposition occurs when one allows a feedforward term in the control.

PROPOSITION 2. (See [1, Ex. 4.10, 5.12]). *Consider* $\dot{x} = Ax + Bu + Gd$, $z = Hx$ *and the control law* $u = Fx + Rd$. *Then*:

(i) DDP′. *There exist* F, R *such that* $H(Is - A_F)^{-1}(G + BR) = 0$ *iff* $\operatorname{Im} G \subset \mathcal{V}^*_{\operatorname{Ker} H} + \mathcal{B}$.

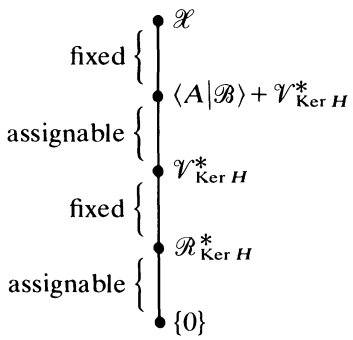
(ii) DDPS′. *There exist* F, R *such that* $H(Is - A_F)^{-1}(G + BR) = 0$, *and* $\sigma(A_F) \subset \mathbb{C}_g$, *iff* (A, B) *is stabilizable (relative to* $\mathbb{C}_g$*) and* $\operatorname{Im} G \subset \mathcal{V}^*_{g,\operatorname{Ker} H} + \mathcal{B}$.

(iii) DDPPP′. *For any* $\mathbb{C}_g$ *there exist* F, R *such that* $H(Is - A_F)^{-1}(G + BR) = 0$ *and* $\sigma(A_F) \subset \mathbb{C}_g$, iff (A, B) *is reachable and* $\operatorname{Im} G \subset \mathcal{R}^*_{\operatorname{Ker} H} + \mathcal{B}$.

Proof of Propositions 1 and 2. Because of the references given it suffices to prove (iii) which, however, follows directly from the fact that $\mathcal{V}^*_{g,\operatorname{Ker} H} = \mathcal{R}^*_{\operatorname{Ker} H}$ whenever $\mathbb{C}_g \cap \sigma((A_F|\mathcal{V}^*_{\operatorname{Ker} H}) \bmod \mathcal{R}^*_{\operatorname{Ker} H}) = \varnothing$, for any $F \in \underline{F}(\mathcal{V}^*_{\operatorname{Ker} H})$.

To contrast with what is to come we summarize some of the main features of the above results.

1. The situation with the spectrum may be illustrated (see [11]) as follows:



2. We use the following notion of genericity. Consider all (A, B, C, G, H) belonging to a given algebraic variety Z. Let $Z = \bigcup_{i=1}^{N} Z_i$, be a decomposition of Z into its irreducible components. Let $\bar{\mathscr{P}}$ denote all (A, B, C, G, H) for which a given problem (e.g., DDP) is not solvable. Then we will say that the problem is *generically solvable* iff $Z_i \cap \bar{\mathscr{P}}$ is a proper subvariety of Z_i for all i.

If we consider all elements of (A, B, C, G, H) to be free, then DDP is never generically solvable; DDP′ is generically solvable iff

$$\#\ \text{controls} \geqq \#\ \text{controlled outputs}.$$

This condition also holds for the generic solvability of DDP, if we consider the subclass of systems with $HG = 0$. For DDPPP′ the condition becomes

$$\#\ \text{controls} > \#\ \text{controlled outputs},$$

while DDPPP needs again the added a priori assumption $HG = 0$.

4. DDEP. The dual notion of controlled invariance is that of conditioned invariance which has been introduced in [6] and further studied in [3], [12] (see also [13] and [1, Ex. 5.17]). We prefer the following definition.

DEFINITION. Consider the system $\dot{x} = Ax$, $y = Cx$. A subspace $\mathscr{S} \subset X$ is said to be *conditionally invariant* if there exist matrices F, E such that $z := x(\mathrm{mod}\ \mathscr{S})$ satisfies $\dot{z} = Fz + Ey$.

This definition may seem a bit "ad hoc". In fact, its discrete time analogue may be introduced in a more intrinsic way by defining $\mathscr{S}$ to be *conditionally invariant* for $x(t+1) = Ax(t)$, $y(t) = Cx(t)$, if there exists f such that $x(t+1)(\mathrm{mod}\ \mathscr{S}) = f(x(t)(\mathrm{mod}\ \mathscr{S}), y(t))$.

The following conditions are equivalent: $\{\mathscr{S}$ is a conditioned invariant subspace$\} \Leftrightarrow \{\mathscr{S}$ is $A|\mathrm{Ker}\ C$ invariant$\} : \Leftrightarrow \{A(\mathscr{S} \cap \mathrm{Ker}\ C) \subset \mathscr{S}\} \Leftrightarrow \{L$ exists such that $A^L\mathscr{S} \subset \mathscr{S}\}$ (L is related to F, E in the above definition by $F = A^L(\mathrm{mod}\ \mathscr{S})$, and $E = -L(\mathrm{mod}\ \mathscr{S})$). Indeed, assume that $\mathscr{S}$ is a conditioned invariant subspace. Then if $x \in \mathrm{Ker}\ C$, it follows that $(\dot{x})(\mathrm{mod}\ \mathscr{S}) = (d/dt)(x(\mathrm{mod}\ \mathscr{S})) = Fx(\mathrm{mod}\ \mathscr{S}) = (Ax)(\mathrm{mod}\ \mathscr{S})$, which shows that $A(\mathscr{S} \cap \mathrm{Ker}\ C) \subset \mathscr{S}$. A simple linear algebra calculation shows that this implies the existence of L such that $A^L\mathscr{S} \subset \mathscr{S}$. For such an L there holds (for $\dot{x} = Ax$, $y = Cx$)

$$\begin{aligned}\frac{d}{dt}(x(\mathrm{mod}\ \mathscr{S})) &= (\dot{x})(\mathrm{mod}\ \mathscr{S}) = (Ax)(\mathrm{mod}\ \mathscr{S}) \\ &= (A^L x)(\mathrm{mod}\ \mathscr{S}) - L(\mathrm{mod}\ \mathscr{S})y \\ &= A^L(\mathrm{mod}\ \mathscr{S})x(\mathrm{mod}\ \mathscr{S}) - L(\mathrm{mod}\ \mathscr{S})y,\end{aligned}$$

which shows the equivalence of the above statements.

The class of all conditionally invariant subspaces will be denoted by $\underline{\mathcal{S}}$ and for $\mathcal{S} \in \underline{\mathcal{S}}$, $\underline{\mathcal{L}}(\mathcal{S}) := \{L | A^L \mathcal{S} \subset \mathcal{S}\}$.

It follows from the definitions that $A|\operatorname{Ker} C$-invariant subspaces are immediately related to the construction of observers. For the stability properties of conditionally invariant subspaces it is not $\sigma(A^L|\mathcal{S})$ but $\sigma(A^L(\operatorname{mod} \mathcal{S}))$ which is relevant. Indeed, consider the data processor

$$\dot{z} = A^L(\operatorname{mod} \mathcal{S})z - L(\operatorname{mod} \mathcal{S})y,$$

as estimator for $x(\operatorname{mod} \mathcal{S})$ in $\dot{x} = Ax$, $y = Cx$. Define $e := z - x(\operatorname{mod} \mathcal{S})$ and note that in this case we need not have $z(t) = x(t)(\operatorname{mod} \mathcal{S})$, since it is not assumed that $z(0) = x(0)(\operatorname{mod} \mathcal{S})$. Then e is governed by $\dot{e} = A^L(\operatorname{mod} \mathcal{S})e$. Consequently, the error dynamics are governed by $\sigma(A^L(\operatorname{mod} \mathcal{S}))$, which leads naturally to the following definition.

DEFINITION. A conditionally invariant subspace $\mathcal{S}$ is said to be a *complementary observability* subspace if for any given real polynomial p of degree $n - \dim \mathcal{S}$, there exists $L \in \underline{\mathcal{L}}(\mathcal{S})$ such that

$$\chi_{A^L(\operatorname{mod} \mathcal{S})} = p.$$

It is said to be a *complementary detectability* subspace (relative $\mathbb{C}_g$) if there exists $L \in \underline{\mathcal{L}}(\mathcal{S})$, such that $\sigma(A^L(\operatorname{mod} \mathcal{S})) \subset \mathbb{C}_g$.

There holds: $\{\mathcal{S}$ is a complementary observability subspace$\} \Leftrightarrow \{\mathcal{S} \in \underline{\mathcal{S}}$ and $\exists L$, $\mathcal{K}_1 \supset \mathcal{K} := \operatorname{Ker} C$ such that $\langle \mathcal{K}_1 | A^L \rangle = \mathcal{S}\} \Leftrightarrow \{\mathcal{S} \in \underline{\mathcal{S}}$ and $\Sigma(A^L, -, C)(\operatorname{mod} \mathcal{S})$ is observable for any $L \in \underline{\mathcal{L}}(\mathcal{S})\}$.

These statements follow immediately from duality. Indeed, let $\underline{\mathcal{N}}$ and $\underline{\mathcal{S}}_g$ denote all complementary observability and detectability subspaces associated with a given pair (A, C). It is easily seen that $A|\operatorname{Ker} C$-subspaces behave dually to $A^T(\operatorname{mod} \operatorname{Im} \cdot C^T)$-subspaces: $\mathcal{S} \in \underline{\mathcal{S}}$ (resp. $\underline{\mathcal{N}}$, $\underline{\mathcal{S}}_g$) relative to (A, C), iff $\mathcal{S}^\perp \in \underline{\mathcal{V}}$ (resp. $\underline{\mathcal{R}}$, $\underline{\mathcal{V}}_g$) relative to (A^T, C^T). In particular $\underline{\mathcal{S}}$, $\underline{\mathcal{S}}_g$ and $\underline{\mathcal{N}}$ are closed under subspace intersection and thus there exist infimal elements of all elements of $\underline{\mathcal{S}}$, $\underline{\mathcal{S}}_g$, and $\underline{\mathcal{N}}$ containing a given subspace $\mathcal{K}$ of $\mathcal{X}$. These subspaces will be denoted by $\mathcal{S}^*_{\mathcal{K}}$, $\mathcal{S}^*_{g,\mathcal{K}}$, and $\mathcal{N}^*_{\mathcal{K}}$ respectively.

In order to compute $\mathcal{S}^*_{\mathcal{L}}$ and $\mathcal{N}^*_{\mathcal{L}}$, it suffices to dualize the algorithms given before. Let $\mathcal{K} := \operatorname{Ker} C$, and consider the following algorithms.

Algorithm (ISA)′:

$$\mathcal{S}^{\mu+1}_{\mathcal{L}} := \mathcal{L} + A(\mathcal{S}^{\mu}_{\mathcal{L}} \cap \mathcal{K}); \qquad \mathcal{S}^0 = \{0\}.$$

Algorithm (ACSA)′:

$$\mathcal{N}^{\mu+1}_{\mathcal{L}} := \mathcal{L} + (A^{-1}\mathcal{N}^{\mu}_{\mathcal{L}}) \cap \mathcal{K}; \qquad \mathcal{N}^0 = \mathcal{X}$$

The sequence $\mathcal{S}^{\mu}_{\mathcal{L}}$ reaches (strictly increasingly) its limit $\mathcal{S}^{\infty}_{\mathcal{L}} = \mathcal{S}^{n-\dim \mathcal{L}+1}_{\mathcal{L}}$ and $\mathcal{N}^{\mu}_{\mathcal{L}}$ reaches (strictly decreasingly) its limit $\mathcal{N}^{\infty}_{\mathcal{L}} = \mathcal{N}^{n-\dim \mathcal{L}}_{\mathcal{L}}$. Furthermore,

$$\mathcal{S}^*_{\mathcal{L}} = \mathcal{S}^{\infty}_{\mathcal{L}} \quad \text{and} \quad \mathcal{N}^*_{\mathcal{L}} = \mathcal{S}^{\infty}_{\mathcal{L}} + \mathcal{N}^{\infty}_{\mathcal{L}} = \mathcal{N}^{\infty}_{\mathcal{S}^{\infty}_{\mathcal{L}}}.$$

Computing a corresponding output injection matrix L such that

$$A^L \mathcal{S}^*_{\mathcal{L}} \subset \mathcal{S}^*_{\mathcal{L}}$$

requires solving a set of linear equations. Finding an L such that

$$A^L \mathcal{N}^*_{\mathcal{L}} \subset \mathcal{N}^*_{\mathcal{L}} \quad \text{and} \quad \chi_{A^L(\operatorname{mod} \mathcal{N}^*_{\mathcal{L}})} = p,$$

requires a standard pole placement computation.

Before introducing the disturbance decoupled estimation problem, we give a simple but very useful result concerning the role of dynamic extensions of linear systems.

Let Σ: $\dot{x} = Ax + Bu$, $y = Cx$ be given. We will call the system Σ^e: $\dot{x} = Ax + Bu$, $\dot{w} = v$, considered as a system with input (u, v) and output (y, w), denoted as Σ^e: $\dot{x}^e = A^e x^e + B^e u^e$, $y^e = C^e x^e$, with $\mathscr{X}^e = \mathscr{X} \oplus \mathscr{W}$, etc., an *extension* of Σ. The dimension of $\mathscr{W}$ is called the *dimension* of this extension. We will denote by P the canonical projection $x^e = (x, w) \overset{P}{\mapsto} x$. It is important to note that static feedback around Σ^e corresponds to dynamic feedback around Σ, and that any (finite dimensional) feedback processor around Σ may be visualized in this way. We have the following simple relations between invariant subspaces of Σ and Σ^e.

PROPOSITION 3. *Let Σ^e be an extension of Σ. Then,*

$$\{\mathscr{V}^e \in \underline{\mathscr{V}}^e\} \Leftrightarrow \{P\mathscr{V}^e \in \underline{\mathscr{V}}\}, \qquad \{\mathscr{S}^e \in \underline{\mathscr{S}}^e\} \Leftrightarrow \{\mathscr{S}^e \cap \mathscr{X} \in \underline{\mathscr{S}}\},$$

$$\{\mathscr{V}_g^e \in \underline{\mathscr{V}}_g^e\} \Leftrightarrow \{P\mathscr{V}_g^e \in \underline{\mathscr{V}}_g\} \text{ and } \{\mathscr{S}_g^e \in \underline{\mathscr{S}}_g^e\} \Leftrightarrow \{\mathscr{S}_g^e \cap \mathscr{X} \in \underline{\mathscr{S}}_g\},$$

$$\{\mathscr{R}^e \in \underline{\mathscr{R}}^e\} \Leftrightarrow \{P\mathscr{R}^e \in \underline{\mathscr{R}}\}, \qquad \{\mathscr{N}^e \in \underline{\mathscr{N}}^e\} \Leftrightarrow \{\underline{\mathscr{N}}^e \cap \mathscr{X} \in \underline{\mathscr{N}}\}.$$

Consider now the plant Σ: $\dot{x} = Ax + Bu + Gd$, with observation (y, u) where $y = Cx$, and the output to be estimated $z = Hx$. The disturbance decoupled estimation problem DDEP is the problem of constructing a data processor, (an *observer*) Σ_p: $\dot{w} = Fw + Ey + Ku$; $\hat{z} = Mw + Ny + Su$, such that the resulting estimation error $e := z - \hat{z}$ depends only on the initial conditions and not on the disturbance d or on the input u. The resulting signal flow graph is then as shown in Fig. 4.

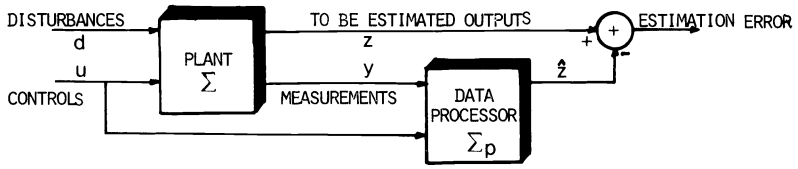


FIG. 4

We emphasize again that, as in the disturbance decoupling problem, the disturbance may also depend on the state through an unknown function or on unknown parameters.

Since in a disturbance decoupled observer the transfer function d, $u \mapsto e$ is zero, all signals $e(\cdot)$, obtainable by varying the initial conditions $x(0)$, $w(0)$, are exactly those obtainable by varying the initial conditions $v(0)$ as the output of a system of the form $\dot{v} = Pv$, $z - \hat{z} = Qv$, for some P, Q. If (P, Q) is observable (which we may always assume to be the case) then we will call $\sigma(P)$ the spectrum (or poles) of the error dynamics of the observer. Note that in an input decoupled observer $x(0) = 0$ and $w(0) = 0$ together imply $e(t) = 0$ for all t (i.e., we have the possibility of perfect tracking of the to be estimated signal by means of the observed signal).

The following proposition treats the disturbance decoupled estimation problem DDEP. This refers to the possibility of finding a disturbance decoupled observer. The problems DDEPS and DDEPPP add the stability or pole placement requirement to the estimation error dynamics.

PROPOSITION 4. *Consider the system $\dot{x} = Ax + Bu + Gd$, with observation (y, u) where $y = Cx$, and the to be estimated output $z = Hx$. Consider an observer of the form $\dot{w} = Fw + Ey + Ru$, $\hat{z} = Mw + Ny + Su$. Then:*

(i) DDEP. *There exists a disturbance decoupled observer, iff* $\mathcal{S}^*_{\mathrm{Im}\,G} \cap \mathrm{Ker}\, C \subset \mathrm{Ker}\, H$.

(ii) DDEPS. *There exists a disturbance decoupled observer with error spectrum* $\subset \mathbb{C}_g$, *iff* $\mathcal{S}^*_{g,\mathrm{Im}\,G} \cap \mathrm{Ker}\, C \subset \mathrm{Ker}\, H$.

(iii) DDEPPP. *For any* $\mathbb{C}_g$ *there exists a disturbance decoupled observer with error spectrum contained in* $\mathbb{C}_g$, *iff* $\mathcal{N}^*_{\mathrm{Im}\,G} \cap \mathrm{Ker}\, C \subset \mathrm{Ker}\, H$.

Proof. Claims (i) and (ii) are essentially proven, by duality arguments, in [2] (see also the references of this paper). For completeness, we include a short proof.

I *Necessity.* Let $e := z - \hat{z}$, and consider the dynamics of $x^e = (x, w) \in \mathcal{X} \oplus \mathcal{W}$, written as $\dot{x}^e = A^e x^e + G^e(u, d)$, $e = H^e x^e + D^e(u, d)$. This must have zero transfer function $(u, d) \mapsto e$. Equivalently, $D^e = 0$ and $\langle A^e | \mathrm{Im}\, G^e \rangle \subset \langle \mathrm{Ker}\, H^e | A^e \rangle =: \mathcal{S}^e$. Obviously, $\mathcal{S}^e$ is A^e-invariant and $\mathrm{Im}\, B^e \subset \mathcal{S}^e \subset \mathrm{Ker}\, C^e$. Now Σ^e: $\dot{x} = Ax + Bu + Gd$; $\dot{w} = Fw + Ey + Ku$ is clearly obtainable from an extension of Σ by extended output injection. Hence any A^e-invariant subspace belongs to $\underline{\mathcal{S}}^e$: $\mathcal{S}^e \in \underline{\mathcal{S}}^e$ and, from Proposition 3, $\mathcal{S}^e \cap \mathcal{X} =: \mathcal{S} \in \underline{\mathcal{S}}$, which implies $\mathrm{Im}\, G = \mathcal{X} \cap \mathrm{Im}\, G^e \subset \mathcal{S} \subset \mathrm{Ker}\, H^e = \mathrm{Ker}\,(H - NC)$, which yields $\mathcal{S} \cap \mathrm{Ker}\, C \subset \mathrm{Ker}\, H$. This proves (i). To prove (ii) it suffices to note that the spectrum of the dynamics e equals the spectrum of $A^e \pmod{\mathcal{S}^e}$. Hence $\mathcal{S}^e \in \underline{\mathcal{S}}^e_g$, and thus $\mathcal{S} \in \underline{\mathcal{S}}_g$ (see Proposition 3), if DDEPS is solvable, whereas the condition for DDEPPP follows directly from the fact that $\mathcal{S}^*_{g,\mathrm{Im}\,G} = \mathcal{N}^*_{\mathrm{Im}\,G}$, whenever $\mathbb{C}_g \cap \sigma(A^L \pmod{\mathcal{S}^*_{\mathrm{Im}\,G}} | \mathcal{N}^*_{\mathrm{Im}\,G} \pmod{\mathcal{S}^*_{\mathrm{Im}\,G}}) = \varnothing$, for any $L \in \underline{L}(\mathcal{S}^*_{\mathrm{Im}\,G})$.

II *Sufficiency.* Assume $\mathcal{S} \in \underline{\mathcal{S}}$, $\mathrm{Im}\, G \subset \mathcal{S}$, and $\mathcal{S} \cap \mathrm{Ker}\, C \subset \mathrm{Ker}\, H$. By this last inclusion there exists M, N such that $Hx = Mx \pmod{\mathcal{S}} + NC$. Let $L \in \underline{L}(\mathcal{S})$ and consider the observer $\dot{w} = A^L \pmod{\mathcal{S}}\ w - L \pmod{\mathcal{S}} y + B \pmod{\mathcal{L}}\ u$, $\hat{z} = Mw + Ny$, with $\mathcal{W} \cong \mathcal{X} \pmod{\mathcal{S}}$, and M, N such that $Hx = Mx \pmod{\mathcal{S}} + NCx$. A simple calculation then shows that the following equation holds,

$$\frac{d}{dt} x \pmod{\mathcal{S}} = A^L \pmod{\mathcal{S}} x \pmod{\mathcal{S}} - L \pmod{\mathcal{S}} y + B \pmod{\mathcal{S}} u.$$

Thus $e := z - \hat{z}$, is governed by $\dot{r} = A^L \pmod{\mathcal{S}} r$, $e = Nr$, with $r := x \pmod{\mathcal{S}} - w$. Hence the transfer function $(u, d) \mapsto e$ is zero, which yields (i). If $\mathcal{S} \in \underline{\mathcal{S}}_{g,\mathrm{Im}\,G}$, or $\mathcal{S} \in \underline{\mathcal{N}}_{\mathrm{Im}\,G}$, then this reasoning yields (ii) and (iii). □

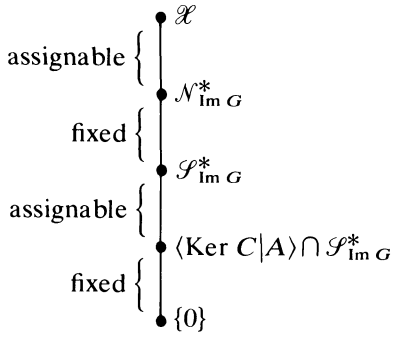
Remarks.

1. In some applications it may be desired that the observations should in any case be "filtered" before being used in $\hat{z}$. This requirement is translated into the constraints $N = 0$, $S = 0$. The results of Proposition 4 then need to be modified, respectively, to:

(i) DDEP′. $\mathcal{S}^*_{\mathrm{Im}\,G} \subset \mathrm{Ker}\, H$,
(ii) DDEPS′. $\mathcal{S}^*_{g,\mathrm{Im}\,G} \subset \mathrm{Ker}\, H$,
(iii) DDEPPP′. $\mathcal{N}^*_{\mathrm{Im}\,G} \subset \mathrm{Ker}\, H$.

2. The estimate given on the order of the observer given in the above proposition is, in general, conservative. In fact, the minimal order estimator design is the dual of the minimal dynamic cover problem, and is not solved at this point. However, it is easily seen from the above proposition that if the to be estimated output is the state, then the dimension of the required observer is at least n-Rank C. The proposition hence also shows in what sense the "Luenberger observer" is minimal. In fact, the order of the observer which achieves pole placement needs, assuming (A, C) to be observable, only be n-Rank C, whereas the above proposition would predict n. This is due to the result described in [1, Lemma 3.5, Th. 3.3]. The procedure described there may actually be generalized to the situation at hand, but we will not go into that here.

3. The situation with the spectrum of conditioned invariant subspaces may be illustrated [11] as follows:



4. If we assume all elements of (A, B, C, G, H) to be arbitrary, then DDEP is generically solvable iff

$$\#\ \text{measurements} \geqq \#\ \text{disturbances}.$$

This condition also holds for DDEP′ provided we add the a priori requirement $HG = 0$. DDEPP instead requires

$$\#\ \text{measurements} > \#\ \text{disturbances}$$

while DDEPP′ again needs the a priori assumption $HG = 0$.

5. DDPM. In this section we will give the main result of this paper: the disturbance decoupling problem with measurements and stability or pole placement requirements.

DEFINITION. Consider the system $\dot{x} = Ax + Bu$, $y = Cx$. The subspace $\mathscr{L} \subset \mathscr{X}$ is said to be an (A, B, C)-*invariant* subspace if there exists K such that $(A + BKC)\,\mathscr{L} \subset \mathscr{L}$.

We will denote all (A, B, C)-invariant subspaces by $\underline{\mathscr{L}}$. The following proposition is easily seen.

PROPOSITION 5. $\mathscr{L} = \underline{\mathscr{V}} \cap \underline{\mathscr{S}}$.

In fact, if $\mathscr{L} \in \underline{\mathscr{L}}$, then it is a matter of solving a set of linear equations to compute a suitable $\mathscr{K}$ for it.

The following elegant result of [3] shows how one can produce (A, B, C)-invariant subspaces by extension.

PROPOSITION 6. *Let $\mathscr{V} \in \underline{\mathscr{V}}$ and $\mathscr{S} \in \underline{\mathscr{S}}$, with $\mathscr{S} \subset \mathscr{V}$. Then there exist an extension of Σ of dimension $\leqq \dim \mathscr{V} - \dim \mathscr{S}$ and an $\mathscr{L}^e \in \underline{\mathscr{L}}^e$, such that $\mathscr{V} = P\mathscr{L}^e$, and $\mathscr{S} = \mathscr{L}^e \cap \mathscr{X}$.*

Proof. The idea behind this proof is shown in Fig. 5. Take $\mathscr{W} \cong \mathscr{V} (\mathrm{mod}\ \mathscr{S})$, i.e., $\dim \mathscr{W} = \dim \mathscr{V} - \dim \mathscr{S}$, and $\mathscr{X}^e := \mathscr{X} \oplus \mathscr{W}$. Let $\mathscr{V}'$ be such that $\mathscr{V} = \mathscr{S} \oplus \mathscr{V}'$, and $\mathscr{Z} \subset \mathscr{V}' \oplus \mathscr{W}$, such that $\mathscr{Z} \cap \mathscr{V}' = \mathscr{Z} \cap \mathscr{W} = \{0\}$ and $\dim \mathscr{Z} = \dim \mathscr{V}' = \dim \mathscr{W}$. Now $\mathscr{L}^e := \mathscr{Z} \oplus \mathscr{S}$, will have the required properties (see Fig. 5). □

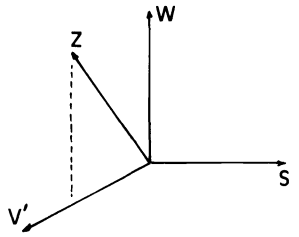


FIG. 5

Remark. The above proposition in effect shows how one can produce an $(A, B, C)^e$-invariant subspace from a pair $\mathscr{S} \subset \mathscr{V}$. Actually this result also solves the following problem. Consider $\dot{x} = Ax + Bu$, $y = Cx$. Let $\mathscr{V} \subset \mathscr{K}$ and suppose that we would like to make $\mathscr{V}$ invariant by feedback from y. Clearly for this $\mathscr{V}$ needs to be $A(\mathrm{mod}\ \mathscr{B})$-invariant. A systematic procedure for achieving such a feedback law is given in Proposition 6. First choose an $A|\mathrm{Ker}\ C$-invariant subspace $\mathscr{S} \subset \mathscr{V}$, taking $\mathscr{S} = \{0\}$ shows that it is always possible to achieve this.

Now let $\mathscr{L}^e$ be such that $P\mathscr{L}^e = \mathscr{V}$, and $\mathscr{L}^e \cap \mathscr{K} = \mathscr{S}$. Then $\mathscr{L}^e \in \underline{\mathscr{L}}^e$, and, hence, there exists K^e (defining a dynamic feedback law) such that $(A^e + B^e K^e C^e)\,\mathscr{L}^e \subset \mathscr{L}^e$. The ensuing closed loop system will have the property that if $x^e(0) \in \mathscr{L}^e$, then $x^e(t) \in \mathscr{L}^e$, i.e., $Px^e(t) \in \mathscr{V}$ for all t, as desired. Actually this procedure may be viewed in terms of separation, with an observer used to estimate the feedback law $u = Fx$, with $F \in \underline{F}(\mathscr{V})$.

We continue with a lemma which is an interesting generalization of well-known results about stabilizability and pole placement by output feedback.

LEMMA. *Let Σ: $\dot{x} = Ax + Bu$, $y = Cx$ be given, and let $\mathscr{L}$ be an A-invariant subspace of $\mathscr{K}$.*

I. *Consider the system $\Sigma|\mathscr{L}$ and assume that it is stabilizable and detectable. Then there exists an extension Σ_1^e of Σ and a static feedback law K_1^e around Σ_1^e such that, with $A_{1,cl} := A^e + B^e K_1^e C^e$,*

(i) $\mathscr{L} \oplus \mathscr{W}_1$ is $A_{1,cl}$-invariant,
(ii) $\sigma(A_{1,cl}|(\mathscr{L} \oplus \mathscr{W}_1)) \subset \mathbb{C}_g$,
(iii) $\sigma(A_{1,cl} = \sigma(A_{1,cl})|(\mathscr{L} \oplus \mathscr{W}_1)) \cup \sigma(A(\mathrm{mod}\ \mathscr{L}))$.

This can always be achieved with an extension of dimension of at most $\gamma_1 := \min(\kappa_{\Sigma|L}, \nu_{\Sigma|L}) - 1 \leqq \min(\kappa_\Sigma, \nu_\Sigma) - 1$. Moreover, if $\Sigma|\mathscr{L}$ is minimal, then given any real polynomial p_1 of degree $\geqq \dim \mathscr{L} + \gamma_1$ one can in fact achieve this with the characteristic polynomial of $A_{1,cl}|(\mathscr{L} \oplus \mathscr{W}_1)$ equal to p_1.

II. *Consider the system $\Sigma(\mathrm{mod}\ \mathscr{L})$, and assume that it is stabilizable and detectable. Then there exists an extension Σ_2^e of Σ and a static feedback law K_2^e around Σ_2^e such that, with $A_{2,cl} := A^e + B^e K_2^e C^e$:*

(i) *$\mathscr{L}$ is $A_{2,cl}$-invariant,*
(ii) $\sigma(A_{2,cl}(\mathrm{mod}\ \mathscr{L})) \subset \mathbb{C}_g$,
(iii) $\sigma(A_{2,cl}) = \sigma(A|\mathscr{L}) \dot{\cup} \sigma(A_{2,cl}(\mathrm{mod}\ \mathscr{L}))$.

This can always be achieved with an extension of dimension of at most $\gamma_2 := \min(\kappa_{\Sigma(\mathrm{mod}\,\mathscr{L})}, \nu_{\Sigma(\mathrm{mod}\,\mathscr{L})}) - 1 \leqq \min(\kappa_\Sigma, \nu_\Sigma) - 1$. Moreover, if $\Sigma(\mathrm{mod}\ \mathscr{L})$ is minimal, then given any real polynomial p_2 of degree $\geqq n\text{-}\dim \mathscr{L} + \gamma_2$, one can in fact achieve this with the characteristic polynomial of $A_{2,cl}(\mathrm{mod}\ \mathscr{L})$ equal to p_2.

Proof. In an suitable basis with $\mathscr{X} \cong \mathscr{L} \oplus \mathscr{X}(\mathrm{mod}\ \mathscr{L})$, $\mathscr{U} \cong \mathscr{B}^{-1}\mathscr{L} \oplus \mathscr{U}(\mathrm{mod}\ \mathscr{B}^{-1}\mathscr{L})$, $\mathscr{Y} \cong C\mathscr{L} \oplus \mathscr{Y}(\mathrm{mod}\ C\mathscr{L})$, Σ may be written as

$$\begin{pmatrix} \dot{x}_1 \\ \hline \dot{x}_2 \end{pmatrix} = \left(\begin{array}{c|c} A_{11} & A_{12} \\ \hline 0 & A_{22} \end{array}\right)\begin{pmatrix} x_1 \\ \hline x_2 \end{pmatrix} + \left(\begin{array}{c|c} B_{11} & B_{12} \\ \hline 0 & B_{22} \end{array}\right)\begin{pmatrix} u_1 \\ \hline u_2 \end{pmatrix}, \qquad \begin{pmatrix} y_1 \\ \hline y_2 \end{pmatrix} = \left(\begin{array}{c|c} C_{11} & C_{12} \\ \hline 0 & C_{22} \end{array}\right)\begin{pmatrix} x_1 \\ \hline x_2 \end{pmatrix},$$

In this representation, $\Sigma|\mathscr{L} \cong \{A_{11}, B_{11}, C_{11}\}$ and $\Sigma(\mathrm{mod}\ \mathscr{L}) \cong \{A_{22}, B_{22}, C_{22}\}$. In order to prove the lemma it suffices to synthesize a Brasch–Pearson stabilization or a placement compensator (see [1, § 3.8]) from y_1 to u_1 for (i), or y_2 to u_2 for (ii).

It remains to be shown that $\min(\kappa_{\Sigma|\mathscr{L}}, \nu_{\Sigma|\mathscr{L}})$, $\min(\kappa_{\Sigma(\mathrm{mod}\,\mathscr{L})}, \nu_{\Sigma(\mathrm{mod}\,\mathscr{L})}) \leqq \min(\kappa_\Sigma, \nu_\Sigma)$. This however is due to the fact that $\kappa_{\Sigma|\mathscr{L}} \leqq \kappa_\Sigma$ (this follows from the results in [14]) and that $\nu_{\Sigma|\mathscr{L}} \leqq \nu_\Sigma$, which is easily derived from first principles. Dually, $\kappa_{\Sigma(\mathrm{mod}\,\mathscr{L})} \leqq \kappa_\Sigma$ and $\nu_{\Sigma|\mathscr{L}} \leqq \nu_\Sigma$. □

The above lemma shows under what conditions stabilization or pole placement by feedback from y to u can be done in a decentralized fashion by feedback from y_1 to u_1 and from u_2 to u_2 without destroying the special subsystem structure induced by the A-invariant subspace $\mathscr{L}$

We are now in a position to state and prove our main result.

THEOREM. *Consider the system* $\Sigma: \dot{x} = Ax + Bu + Gd$, $y = Cx$, $z = Hx$, *and the feedback processor* Σ_f: $\dot{w} = Fw + Ey$, $u = Mw + Ny$. *Let* Σ_{cl} *be the resulting closed loop system with*

$$A_{cl} := \left[\begin{array}{c|c} A + BNC & BM \\ \hline EC & F \end{array}\right].$$

Then

(i) DDPM ([2], [3]). *There exists* Σ_f *such that the transfer function* $d \mapsto z$ *in* Σ_{cl} *is zero iff* $\mathscr{S}^*_{\mathrm{Im}\, G} \subset \mathscr{V}^*_{\mathrm{Ker}\, H}$. *Moreover the required* $\dim \mathscr{W} \leqq \dim \mathscr{V}^*_{\mathrm{Ker}\, H} - \dim \mathscr{S}^*_{\mathrm{Im}\, G}$.

(ii) DDPMS. *There exists* Σ_f *such that the transfer function* $d \mapsto z$ *in* Σ_{cl} *is zero and* $\sigma(A_{cl}) \subset \mathbb{C}_g$ *iff* Σ *is stabilizable and detectable and* $\mathscr{S}^*_{g,\mathrm{Im}\, G} \subset \mathscr{V}^*_{g,\mathrm{Ker}\, H}$. *Moreover the required* $\dim \mathscr{W} \leqq \dim \mathscr{V}^*_{g,\mathrm{Ker}\, H} - \dim \mathscr{S}^*_{g,\mathrm{Im}\, G} - 2(\min(\kappa_\Sigma, \nu_\Sigma) - 1)$.

(iii) DDPMPP. *For any* $\mathbb{C}_g$ *there exists* Σ_f *such that the transfer function* $d \mapsto z$ *in* Σ_{cl} *is zero and* $\sigma(A_{cl}) \subset \mathbb{C}_g$ *iff* Σ *is minimal and* $\mathscr{N}^*_{\mathrm{Im}\, G} \subset R^*_{\mathrm{Ker}\, H}$. *Moreover, the required* $\dim \mathscr{W} \leqq \dim \mathscr{R}^*_{\mathrm{Ker}\, H} - \dim \mathscr{N}^*_{\mathrm{Im}\, G} - 2(\min(\kappa_\Sigma, \nu_\Sigma) - 1)$.

Proof.

I. *Necessity.* Assume that a required Σ_f exists. Consider the extension Σ^e on which static feedback results in a Σ_{cl} with zero transfer function $d \mapsto z$, and write it as Σ_{cl}: $\dot{x}^e = A_{cl}x^e + G^e d$, $z = H^e x^e$. Hence, $\langle A_{cl} | \mathrm{Im}\, G^e \rangle \subset \langle \mathrm{Ker}\, H^e | A_{cl} \rangle =: \mathscr{L}^e$. Obviously $\mathscr{L}^e$ is A_{cl}-invariant and hence, as shown in the proof of Proposition 4, $\mathscr{L}^e \in \underline{\mathscr{L}}^e = \underline{\mathscr{S}}^e \cap \underline{\mathscr{V}}^e$. Furthermore, $\mathrm{Im}\, G^e \subset \mathscr{L}^e \subset \mathrm{Ker}\, H^e$. Hence, $\mathrm{Im}\, G \subset \mathscr{K} \cap \mathrm{Im}\, G^e \subset \mathscr{L}^e \cap \mathscr{K} =: \mathscr{S} \subset \mathscr{V} := P\mathscr{L}^e \subset P\, \mathrm{Ker}\, H^e = \mathrm{Ker}\, H$. From Proposition 3, it follows that $\mathscr{V} \in \underline{\mathscr{V}}$ and $\mathscr{S} \in \underline{\mathscr{S}}$, as desired.

The above reasoning also shows the solvability of DDPMS. Indeed, when A_{cl} is stable then $A_{cl}|\mathscr{L}^e$ is stable as well; thus $\mathscr{L}^e \in \underline{\mathscr{V}}^e_g \cap \underline{\mathscr{S}}^e_g$, which by the above reasoning shows that there exist $\mathscr{V}_g \in \underline{\mathscr{V}}_g$ and $\mathscr{S}_g \in \underline{\mathscr{S}}_g$ such that $\mathrm{Im}\, G \subset \mathscr{S}_g \subset \mathscr{V}_g \subset \mathrm{Ker}\, H$. That stabilizability and detectability of Σ is also a necessary condition follows from general principles.

Consider now DDPMPP. If $\sigma((A_F|\mathscr{V}^*_{\mathrm{Ker}\, H})(\mathrm{mod}\, \mathscr{R}^*_{\mathrm{Ker}\, H})) \cap \mathbb{C}_g = \varnothing$ for $F \in \underline{F}(\mathscr{V}^*_{\mathrm{Ker}\, H})$, then $\mathscr{V}^*_{g,\mathrm{Ker}\, H} = \mathscr{R}^*_{\mathrm{Ker}\, H}$ and, dually, if $\sigma(A^L(\mathrm{mod}\, \mathscr{S}^*_{\mathrm{Im}\, G})|\mathscr{N}^*_{\mathrm{Im}\, G}(\mathrm{mod}\, \mathscr{S}^*_{\mathrm{Im}\, G})) \cap \mathbb{C}_g = \varnothing$, for $L \in \underline{L}(\mathscr{S}^*_{\mathrm{Im}\, G})$, then $\mathscr{S}^*_{g,\mathrm{Im}\, G} = \mathscr{N}^*_{\mathrm{Im}\, G}$. Hence, there exist plenty of $\mathbb{C}_g$'s such that $\mathscr{V}^*_{g,\mathrm{Ker}\, H} = \mathscr{R}^*_{\mathrm{Ker}\, H}$ and $\mathscr{S}^*_{g,\mathrm{Im}\, G} = \mathscr{N}^*_{\mathrm{Im}\, G}$, which yields $\mathscr{N}^*_{\mathrm{Im}\, G} \subset \mathscr{R}^*_{\mathrm{Ker}\, H}$, since if DDPMPP is solvable, then DDPMS is solvable for those $\mathbb{C}_g$'s, as required. That minimality of Σ is also a necessary condition follows again from general principles.

II. *Sufficiency.* This part of the proof is constructive and the procedure may be divided into three parts.

Step 1 (*Disturbance decoupling*). Since $\mathscr{S}^*_{\mathrm{Im}\, G} \subset \mathscr{V}^*_{\mathrm{Ker}\, H}$ there exists, by Proposition 7, a first extension of Σ, Σ^e_1 of dimension $\leqq \dim V^*_{\mathrm{Ker}\, H} - \dim S^*_{\mathrm{Im}\, G}$ and an $(A, B, C)^e$-invariant subspace L^e_1 such that $\mathscr{S}^*_{\mathrm{Im}\, G} = \mathscr{L}^e_1 \cap X \subset P\mathscr{L}^e_1 = \mathscr{V}^*_{\mathrm{Ker}\, H}$. Write Σ^e_1 as $\dot{x}^e_1 = A^e_1 x^e_1 + B^e_1 u^e_1 + G^e d$, $y^e_1 = C^e_1 x_1$, $z = H^e_1 x^e_1$. Hence $\mathrm{Im}\, G^e_1 \subset \mathscr{L}^e_1 \subset \mathrm{Ker}\, H^e_1$. Thus there exists K^e_1 such that $\mathscr{L}^e_1$ is $A_{1,cl}(:= A^e_1 + B^e_1 K^e_1 C^e_1)$-invariant. Since $\langle A_{1,cl} | \mathrm{Im}\, G^e_1 \rangle \subset \mathscr{L}^e_1 \subset \langle \mathrm{Ker}\, H^e_1 | A_{1,cl} \rangle$ this yields the solution to DDPM. The resulting closed system is

$$\Sigma^e_{1,cl}: \dot{x}^e_1 = A_{1,cl}x^e_1 + B^e_1 u^e_1 + G^e_1 d,\ y^e_1 = C^e_1 x^e_1,\ z = H^e_1 x^e_1.$$

It is disturbance decoupled but enjoys no further stability properties as yet.

We now consider DDPMS. If Σ is stabilizable and detectable, so are Σ_1^e and $\Sigma_{1,cl}^e$. Let $\mathscr{L}_1^e$ be constructed as in the previous paragraph starting from $\mathscr{S}^*_{g,\operatorname{Im} G} \subset \mathscr{V}^*_{g,\operatorname{Ker} H}$. Hence $\mathscr{L}_1^e$ is $A_{1,cl}$-invariant, $\operatorname{Im} G_1^e \subset \mathscr{L}_1^e \subset \operatorname{Ker} H_1^e$, and $\mathscr{L}_1^e \in \underline{\mathscr{S}}_{1,g}^e \cap \underline{\mathscr{V}}_{1,g}^e$. Consider $\Sigma_{1,cl}^e|\mathscr{L}_1^e$ and $\Sigma_{1,cl}^e \pmod{\mathscr{L}_1^e}$. Now, $\Sigma_{1,cl}^e|\mathscr{L}_1^e$ is stabilizable and detectable; stabilizable because $\mathscr{L}_1^e \in \underline{\mathscr{V}}_{1,g}^e$, and detectable because Σ_1^e is. Dually, $\Sigma_{1,cl}^e \pmod{\mathscr{L}_1^e}$ is stabilizable and detectable; stabilizable because $\Sigma_{1,cl}^e$ is and detectable because $\mathscr{L}_1^e \in \underline{\mathscr{S}}_{1,g}^e$.

Using these properties of $\Sigma_{1,cl}^e|\mathscr{L}_1^e$ and $\Sigma_{1,cl}^e \pmod{\mathscr{L}_1^e}$ it is now possible to carry out the stabilization steps by a decentralized procedure, by first putting feedback around $\Sigma_{1,cl}^e|\Sigma_1^e$ (we will call this the *disturbance loop stabilization*) and then putting feedback around $\Sigma_{1,cl}^e \pmod{\mathscr{L}_1^e}$ (we will call this the *controlled output loop stabilization*).

Step 2 (*Disturbance loop stabilization*). Let us now use procedure (i) of the lemma on $\Sigma_{1,cl}^e|\mathscr{L}^e$. This yields a new extension Σ_2^e and a feedback such that $\mathscr{L}_2^e := \mathscr{L}_1^e \oplus \mathscr{W}_2$ is $A_{2,cl}$-invariant, $\sigma(A_{2,cl}|\mathscr{L}_2^e) \subset \mathbb{C}_g$, and $\sigma(A_{2,cl} \pmod{\mathscr{L}_2^e}) = \sigma(A_{1,cl} \pmod{\mathscr{L}_2^e}) \subset \mathbb{C}_g$. *Furthermore*, $\operatorname{Im} G_2^e \subset \mathscr{L}_2^e \subset \operatorname{Ker} H_2^e$ remains satisfied, which still yields a disturbance decoupled system.

Step 3 (*Controlled output stabilization*). Let us now use procedure (ii) of the lemma on $\Sigma_{2,cl}^e \pmod{\mathscr{L}_2^e}$. (Note that $\Sigma_{2,cl}^e \pmod{\mathscr{L}_2^e} = \Sigma_{1,cl}^e \pmod{\mathscr{L}_1^e}$.) This yields an extension Σ_3^e and a feedback such that $\mathscr{L}_3^e = \mathscr{L}_2^e$ remains $A_{3,cl}$-invariant, which implies $\operatorname{Im} G_3^e \subset \mathscr{L}_3^e \subset \operatorname{Ker} H_3^e$ and hence, the DDPM conditions will remain satisfied. Furthermore, $\sigma(A_{3,cl} \pmod{\mathscr{L}_3^e}) \subset \mathbb{C}_g$ and $\sigma((A_{3,cl})|\mathscr{L}_3^e) = \sigma((A_{2,cl})|\mathscr{L}_2^e) \subset \mathbb{C}_g$, which yields DDPMS.

We still need to show the estimate on the required order of the extension. This follows from the estimates

$$\kappa_{\Sigma_{1,cl}^e \pmod{\mathscr{L}_1^e}} \leqq \kappa_{\Sigma_{1,cl}^e} = \kappa_{\Sigma_1^e} = \kappa_\Sigma,$$

and

$$\kappa_{\Sigma_{2,cl}^e|\mathscr{L}_2^e} = \kappa_{\Sigma_{1,cl}^e|\mathscr{L}_1^e} \leqq \kappa_{\Sigma_{1,cl}^e} = \kappa_{\Sigma_1^e} = \kappa_\Sigma,$$

with similar estimates for the observability indices.

Turning now to DDPMPP we see that the procedure sketched above will also work with an arbitrary $\mathbb{C}_g$ provided $\mathscr{N}^*_{\operatorname{Im} G} \subset \mathscr{R}^*_{\operatorname{Ker} H}$, since $\Sigma_{1,cl}^e$, as constructed in Step 1, will then be such that $\Sigma_{1,cl}^e \pmod{\mathscr{L}_1^e}$ and $\Sigma_{1,cl}^e|\mathscr{L}_1^e$ are both minimal, and hence Steps 2 and 3 can be done with pole placement.

This ends the sketchy proof of the theorem. □

Remarks. 1. DDPM is solvable only if $HG = 0$. However, in this subclass we have generic solvability if and only if

$$\#\text{ controls} \geqq \#\text{ controlled outputs},$$

$$\#\text{ observations} \geqq \#\text{ disturbances}.$$

For DDPMPP this condition becomes:

$$\#\text{ controls} > \#\text{ controlled outputs},$$

$$\#\text{ observations} > \#\text{ disturbances},$$

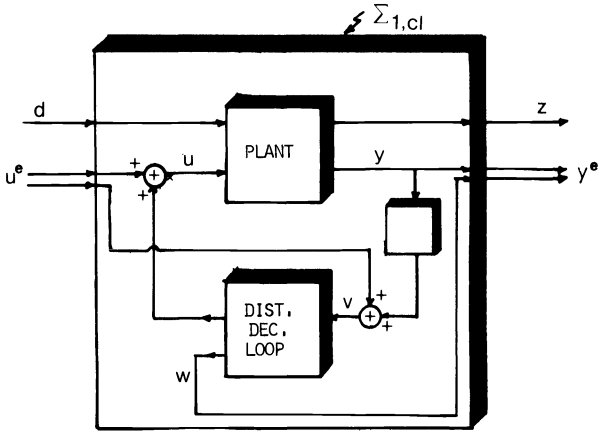
Finally, if feedforward is allowed (i.e., d is measured and available in the feedback processor) then we have solvability of DDPM, (resp. DDPMPP) iff we have it for DDP, (resp. DDPPP).

2. Note that the estimates for the dimension of the feedback processors as given in the theorem and the lemma are conservative, and may in any specific situation be improvable by analyzing the controllability and observability indices of $\Sigma|\mathscr{L}$ and $\Sigma \pmod{\mathscr{L}}$. Of course, the *minimal* order required, or generically required, is not known and will be a complex combination of the minimal cover and the minimal order stabilizing compensator design (research) problems.

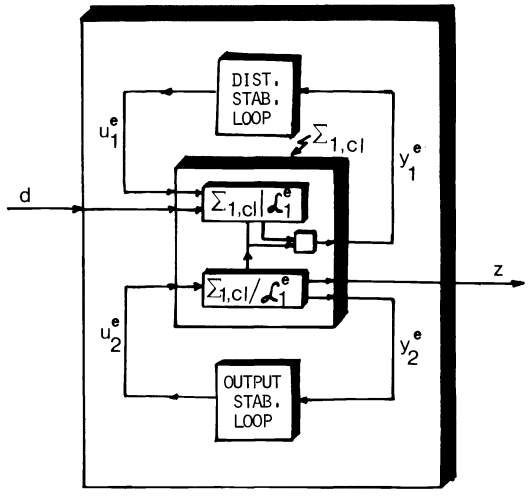
3. We have been referring to Step 2 in the proof of our theorem as *disturbance loop stabilization* because it stabilizes the loop which is influenced by external disturbances (even though it does not influence the to be controlled outputs). Step 3 is called *controlled output stabilization* because it stabilizes the loop which influences the to be controlled outputs (even though it is not influenced by the disturbances).

The total design procedure with the disturbance decoupling loop and the two stabilization loops has an appealing hierarchical structure. This structure may be made more elegant yet by viewing all three control loops in terms of a separation philosophy, with the observer elements driven by the estimation errors and having their own internal control feedback. It seems appropriate to mention at this point that, as shown in [3], in a closed loop configuration it is in general not possible to distinguish *observer error dynamic modes* and *state feedback controlled modes.*

The signal flow graph of the controller may be visualized as shown in Fig. 6.



DISTURBANCE DECOUPLING



STABILIZATION

FIG. 6

Altogether this results in a complex, but nevertheless logically structured and, from a cybernetic point of view appealing, synthesis. Even though the order of the feedback control compensator may be up to three times the dynamic order of the plant,

the resulting feedback controller could be implemented on a microprocessor for moderately complicated plants.

4. The synthesis procedure explained in the proof of the theorem can obviously be made into a computer-aided design algorithm. In the case of DDPMPP one would proceed as follows:

Data. A, B, C, G, H, (which must satisfy $HG = 0$), and the desired $\mathbb{C}_g$ (or the desired symmetric set of poles $\lambda_1, \lambda_2, \cdots, \lambda_N$, or the desired characteristic polynomial p of degree N. The synthesis will work as long as N is large enough and as long as a factorizability condition on p, which comes out of the structure of the controller, is satisfied).

Verify whether $m > l$ and $p > q$. If so, proceed with confidence (see Remark 1). If not, count on luck due to special structure of the system matrices.

Step 1. Compute $\mathcal{N}^*_{\text{Im } G}$ and $\mathcal{R}^*_{\text{Ker } H}$ using, e.g., the linear algorithms given above. If $\mathcal{N}^*_{\text{Im } G} \subset \mathcal{R}^*_{\text{Ker } H}$, proceed. Otherwise, look for some other control system design approach, e.g., an LQG approach.

Step 2. Solve DDPM by computing $\mathcal{L}^e_1$ and K^e_1, using the ideas in the proofs of Proposition 6 and Step 1 of the theorem.

Step 3. Design Brasch–Pearson compensators for $\Sigma^e_1 (\text{mod } \mathcal{L}^e_1)$ and $\Sigma^e_1 | \mathcal{L}^e_1$.

Obviously, in order to implement such procedures into good working high level computer-aided design packages, a lot of numerical work remains to be done [15]. However, it seems very important that such packages be developed, and the failure of control theorists to give adequate attention to such efforts undoubtedly contributes to the widely advertised gap between control theory and practice.

Acknowledgment. I would like to thank Dr. J. M. Schumacher for some useful discussions.

REFERENCES

[1] W. M. WONHAM, *Linear Multivariable Control, a Geometric Approach*, 2nd ed., Springer-Verlag, New York, 1979.

[2] S. P. BHATTACHARYYA, Observer design for linear systems with unknown inputs, IEEE Trans. Automatic Control, AC-23 (1978), pp. 483–484.

[3] J. M. SCHUMACHER, *Compensator synthesis using* (C, A, B)*-pairs*, IEEE Trans. on Automatic Control, AC-25 (1980).

[4] G. BASILE and G. MARRO, *L'invarianza rispetto ai disturbi studiata nello spazio degli stati*, Rendiconti della LXX Riunione Annuale AEI, 1969, paper 1.4.01.

[5] R. LASCHI AND G. MARRO, *Alcune considerazioni sull'osservabilità dei sistemi dinamici con ingressi inaccessibili*, Ibid., 1969, paper 1.1.06.

[6] G. BASILE AND G. MARRO, *Controlled and conditioned invariant subspaces in linear system theory*, J. Optim. Theory Appl., 3 (1969), pp. 306–315.

[7] H. AKASHI AND H. IMAI, *Disturbance localization and output deadbeat through an observer in discrete-time linear multivariable systems*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 621–627.

[8] F. M. BRASCH AND J. B. PEARSON, *Pole placement using dynamic compensators*, IEEE Trans. Automat. Control, AC-15 (1970), pp. 34–43.

[9] J. C. WILLEMS, *Almost* A(mod $\mathcal{B}$)*-invariant subspaces*, Astérisque, 75–76 (1980).

[10] ———, *Almost invariant subspaces: an approach to high gain feedback design—Part* 1: *Almost controlled invariant subspaces*, IEEE Trans. Automat. Control, AC-26 (1981).

[11] J. M. SCHUMACHER, *A complement on pole placement*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 281–282.

[12] G. MARRO, *Fondamenti di Teoria dei Sistemi*, Pàtron Editore, Bologna, 1975.

[13] A. S. MORSE, *Structural invariants of linear multivariable systems*, this Journal, 11 (1973), pp. 446–465.

[14] M. HEYMANN, *Controllability subspaces and feedback simulation*, this Journal, 14 (1976), pp. 769–789.

[15] B. C. MOORE AND A. J. LAUB, *Computation of supremal* (A, B)*-invariant and controllability subspaces*, IEEE Trans. Automat. Control, AC-23 (1978), pp. 783–792.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 4, July 1981

0363-0129/81/1904-0006 $01.00/0

SELECTING SUBSETS FROM THE SET OF NONDOMINATED VECTORS IN MULTIPLE OBJECTIVE LINEAR PROGRAMMING*

J. G. ECKER† AND NANCY E. SHOEMAKER‡

Abstract. In this paper, we develop methods for selecting certain subsets from the set N of nondominated points for multiple-objective linear programming problems. One such subset is the set of a points x in N for which the maximum deviation of the objective function values Cx from some ideal vector M is as small as possible. This subset can be obtained as the set of nondominated points for a multiple-objective problem that is considerably smaller than the original problem and the proposed method does not require that the set N be calculated explicitly. The method is extended to obtain another subset of N called the trade-off compromise set that has some interesting properties and that gives valuable information about possible trade-offs amongst the objectives.

1. Introduction. In a multiple objective linear programming problem, a convex polyhedron $X \subseteq R^n$ is given over which several linear objectives are to be maximized. These objectives are given as the components of a column Cx where C is a $k \times n$ matrix with k denoting the number of objectives. A point $x^0 \in X$ is called nondominated if there is no $x \in X$ with $Cx \geqq Cx^0$ and $Cx \neq Cx^0$. Nondominated points are sometimes referred to as *Pareto optimal* or *efficient* points.

Methods for generating the set of all nondominated extreme points have been developed by several authors; see for example Philip [13], Evans and Steuer [19], Yu and Zeleny [17] and Ecker and Kouada [20]. Methods have also been developed for describing the entire nondominated set N as a union of maximal nondominated faces as in Yu and Zeleny [17], Gal [14], Isermann [15] and Ecker, Hegner and Kouada [16].

Once the set N has been generated, the problem of using this (often large) set in the decision process remains. In this paper, we develop methods for selecting subsets of the nondominated set. One important class of subsets is determined by considering those points $x \in N$ for which Cx is as close as possible to some vector of *ideal* values for the k objectives. In [8], Zeleny defines an *ideal vector* M by letting

$$M_i = \max_{x \in X} C_i x, \qquad i = 1, 2, \cdots, k,$$

where C_i denotes the ith row of C. We assume throughout that X is bounded. Given these maxima M_i of the individual objectives over X, a point $x \in X$ is called a *compromise solution* (as in [5], [7] and [8]) with respect to the l_∞-norm if it is optimal for the program

$$Q_\infty(X): \min_{x \in X} \max_i (M_i - C_i x).$$

Compromise solutions in multiple-objective programming have been investigated by several authors; see, for example, references [1]–[10]. In [5], a general framework for studying the relation of compromise solutions and the set of nondominated solutions is presented. In this paper, we will be concerned with the compromise solutions defined by the l_∞-norm and we will restrict our discussion to the linear multiple-objective problem.

* Received by the editors December 8, 1978, and in revised form June 30, 1980. This research was supported in part by the National Science Foundation under grant MCS 75-09443 A02.

† Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, New York 12181.

‡ Department of Mathematics, Oakland University, Rochester, Michigan 48063.

2. Finding all nondominated compromise solutions. The set N_∞ of all nondominated compromise solutions, which we seek, is the set of alternate optima for the program

$$Q_\infty(N): \min_{x \in N} \max (M_i - C_i x).$$

If x is optimal for $Q_\infty(X)$, then x is not necessarily nondominated and the two programs $Q_\infty(N)$ and $Q_\infty(X)$ are not equivalent. However, as established by Dinkelbach and Durr in [2], there is at least one $x \in X$ that is optimal for both $Q_\infty(X)$ and $Q_\infty(N)$ and, consequently, these two programs have the same optimal value. In particular, if $Q_\infty(X)$ has a unique optimal point x^*, then it follows that x^* is the unique optimal vector for $Q_\infty(N)$ and the set of nondominated compromise solutions consists of the single x^*. In the remainder of this section, we propose a method for finding all nondominated compromise solutions when $Q_\infty(X)$ has alternate optima.

It is well known that $Q_\infty(X)$ can be reformulated as an equivalent linear program (see [11], for example). However, because N is not necessarily a convex polyhedron, the program $Q_\infty(N)$ has no such reformulation. In [8], Zeleny proposes a method for solving $Q_\infty(N)$ by using concepts from two-person zero-sum games. In some cases, however, the method in [8] may yield solutions that are dominated (see [12] for an example).

Our approach for solving $Q_\infty(N)$ first requires solving $Q_\infty(X)$ as a linear program; namely the well-known program Q below.

$$\begin{aligned} Q: \quad & \min w \\ & \text{subject to } M - Cx \leqq we \text{ and } x \in X, \end{aligned}$$

where $w \in R^1$, and e is a column vector with each component equal to one. By using the optimal value for program Q, the following theorem (which is the main result in this section) indicates how the set N_∞ of nondominated compromise solutions can easily be determined.

THEOREM 1. *Let $\bar{w} > 0$ be the optimal value for program Q. The set N_∞ of nondominated compromise solutions is equal to the set of nondominated points for the multiple objective program*

$$\begin{aligned} P: \quad & \max Cx \\ & \textit{subject to } M - Cx \leqq \bar{w}e \textit{ and } x \in X. \end{aligned}$$

Proof. Let N^* be the set of nondominated points for P. Suppose x is a nondominated compromise solution. If y is feasible for P with $Cy \geqq Cx$, then $Cy = Cx$, since $y \in X$ and $x \in N$. This shows that $N_\infty \subseteq N^*$. Now suppose that $x \in N^*$. To show that $x \in N_\infty$, we first note that the optimal set for program Q is given by $\{x \in X | M - Cx \leqq \bar{w}e\}$, which is the feasible set for P. Thus, $x \in N^*$ implies that x is a compromise solution. To see that $x \in N$, suppose that there exists $y \in X$ with $Cy \geqq Cx$ (so that $M - Cy \leqq M - Cx$). Thus, $M - Cy \leqq \bar{w}e$ and so y is feasible for program P. But then $x \in N^*$ implies that $Cy = Cx$. Therefore $x \in N$ and x is a nondominated compromise solution which shows that $N^* \subseteq N_\infty$, and the proof is complete.

In view of the above result, solving $Q_\infty(N)$ can be accomplished without explicitly finding the nondominated set N. One simply solves the linear program Q to obtain $\bar{w}$ and then the nondominated set for the multiple objective program P is generated using one of the existing methods referenced earlier.

The optimal tableau for the linear program Q can be used to obtain the initial tableau required by methods, as in [16], to generate the nondominated set for P. For details on computational considerations in solving P, see [12], where the fact that the feasible set for P is simply the set of alternate optima for Q is exploited. We will also see in the next section that in the special case where the original multiple objective problem has three objectives ($k = 3$) then the multiple objective program P reduces to an ordinary linear program. This occurs because at least two of the three objectives will be constant over X_∞.

In the next section, we consider an extension of our method for finding a subset of N_∞ which provides valuable information concerning possible objective function trade-offs amongst the set of nondominated points.

3. The trade-off compromise set. Given the ideal vector M, let $d(x)$ denote the deviation vector defined by

$$d_i(x) = M_i - C_i(x), \qquad i = 1, 2, \cdots, k.$$

A point $\bar{x} \in X$ is called a *trade-off compromise point* if $C_j x > C_j \bar{x}$ for some $x \in X$ implies that there is an index i such that $C_i x < C_i \bar{x}$ and $d_i(\bar{x}) \geqq d_j(\bar{x})$. Given a trade-off compromise point $\bar{x}$, if we wish to improve the objective function value $C_j \bar{x}$ of the jth objective, then we must be willing to accept a *decrease* in another objective function that is no better off relative to its ideal value than the jth objective. If such a decrease is not acceptable then no increase in $C_j x$ is possible.

It is clear from its definition that a trade-off compromise point is nondominated. In fact, if T denotes the *trade-off compromise set*, then we will show below that $T \subseteq N_\infty$, the set of nondominated compromise points. The name "trade-off compromise" was chosen because of the information that such points provide about possible trade-offs and because such points are nondominated compromise points.

The set T of trade-off compromise points depends on the ideal vector M. After presenting a method for determining the set T, in the next section we will show how M can be altered to provide further information about possible trade-offs amongst the objectives.

To help motivate the algorithm for generating the trade-off compromise set T, consider the following process. Suppose we solve program Q in § 2 and discover that its optimal set X_∞ does not consist of just a single point. As established below, there will, in general, be at least two objectives satisfying

$$M_i - C_i x = \bar{w}$$

for each $x \in X_\infty$. Suppose we then define a new ideal vector M' by

$$M'_i = M_i - \bar{w}, \quad \text{if } x \in X_\infty \text{ implies } C_i x = M_i - \bar{w},$$

$$M'_i = M_i, \qquad \text{if } \exists x \in X_\infty \text{ with } C_i x > M_i - \bar{w}.$$

That is, the ideal value is changed to the only attainable value for the objectives that are worst off. For the objectives that can get closer (over X_∞) to their ideal values, we could then try to find an $x \in X_\infty$ that minimizes (over this smaller set of objectives) the maximum deviation. This suggests the iterative process described in the algorithm below. Each iteration of this algorithm involves solving only a single linear program.

4. An algorithm for generating the trade-off compromise set T. In this section, we present an algorithm that consists of a finite sequence of no more than k linear programs for generating the trade-off compromise set T, where k is the number of objectives.

After presenting the algorithm we will develop the results necessary to show that the algorithm converges to T. Some additional characteristics of the set T will also be discussed.

ALGORITHM A.

Step 0. Let $J_0 = \{1, 2, \cdots, k\}$, $X^0 = X$, $n = 0$.

Step 1. Solve the linear program

$$Q_n: \quad \min w$$
$$\text{subject to } M_i - C_i x \leqq w, \qquad i \in J_n,\ x \in X^n \text{ and } w \geqq 0.$$

Let w^n be the minimal value.

Step 2. Let $X^{n+1} = \{x \in X^n \mid C_i x \geqq M_i - w^n,\ i \in J_n\}$.

Let $J_{n+1} = \{j \in J_n \mid \exists x \in X^{n+1} \text{ with } C_j x > M_j - w^n\}$.

If $J_{n+1} = \phi$, stop.

Otherwise, let $n = n + 1$ and go to Step 1.

When $n = 0$, notice that program Q_0 is simply program Q. Also notice that, since M_i is the maximum of $C_i x$ over X, it follows that if $w^n = 0$ then $J_{n+1} = \phi$. The set J_{n+1} identifies those objectives that can get closer over X^{n+1} to their ideal values. In particular, when $n = 0$, the set J_1 identities those objectives that can be improved over the set X_∞ of optimal solutions to program Q.

A proof of convergence of the algorithm depends on the following lemma.

LEMMA 1. *There exists $j \in J_n$ such that $C_j x = M_j - w^n$ for all $x \in X^{n+1}$, so $J_{n+1} \neq J_n$ for $n \geqq 0$.*

Proof. Suppose for each $j \in J_n$, $\exists x^j \in X^{n+1}$ such that $C_j x^j > M_j - w^n$. Consider the point

$$\bar{x} = \left(\frac{1}{m}\right) \sum_{j \in J_n} x^j,$$

where m is the number of indices in J_n. Notice that $\bar{x} \in X^{n+1}$, since X^{n+1} is convex, and so $\bar{x} \in X^n$, since $X^{n+1} \subseteq X^n$. We will now show that $\bar{w}$ exists such that $(\bar{x}, \bar{w})$ is feasible for Q_n with $\bar{w} < w^n$, contradicting the assumption that w^n be the minimal value for Q_n. To this end, observe that for each $i \in J_n$,

$$\begin{aligned} z^i &\equiv M_i - C_i \frac{1}{m} \sum_{j \in J_n} x^j \\ &= \frac{1}{m} \sum_{j \in J_n} (M_i - C_i x^j) \\ &= \frac{1}{m} \sum_{j \in J_n - \{i\}} (M - C_i x^j) + \frac{1}{m}(M_i - C_i x^i) \\ &\leqq \frac{1}{m}(m-1) w^n + \frac{1}{m}(M_i - C_i x^i). \end{aligned}$$

The last inequality holds since $x^j \in X^{n+1}$ for each j. Thus,

$$z^i < w^n \quad \text{for each } i \in J_n,$$

since $M_i - C_i x^i < w^n$ by choice of x^i. Letting

$$\bar{w} = \max_{i \in J_n} z^i,$$

we see that $\bar{w} < w^n$. This contradiction completes the proof.

Lemma 1 implies that eventually the algorithm will reach a stage with $J_N = \phi$ for some N. In the development below, we show that the corresponding set X^N is, in fact, the trade-off compromise set T.

THEOREM 2. *Algorithm A terminates with $J_N = \phi$ for some $N \leqq k$ and the set X^N generated is such that*:

(a) $x, y \in X^N$, *implies* $Cx = Cy$;

(b) $X^N \subseteq E$.

Proof. The existence of N is immediate from the finiteness of J_0 and Lemma 1. Let $j \in \{1, 2, \cdots, k\}$. Then since $J_N = \phi$, $j \in J_0$, and $J_{n+1} \subset J_n$ for all n, there is an n such that $j \in J_n - J_{n+1}$. From the definition of X^{n+1} and J_{n+1}, it follows that $C_j x = M_j - w^n$ for all $x \in X^{n+1}$. But $X^N \subseteq X^{n+1}$ so $C_j x = M_j - w^n$ for all $x \in X^N$. Thus each $C_j x$ is constant over X^N which proves (a). To prove (b), choose $x \in X^N$. Choose $y \in X$ with $Cy \geqq Cx$. Then we have $X = X^0$ so $y \in X^0$. For $0 \leqq n < N$, consider the *induction hypothesis* $y \in X^n$. Then $X^N \subseteq X^{n+1}$ implies $C_j x \geqq M_j - w^n$ for each $j \in J_n$. But $Cy \geqq Cx$ then implies $C_j y \geqq M_j - w^n$ for each $j \in J_n$. By the induction hypothesis, we then conclude that $y \in X^{n+1}$. Therefore by induction, $y \in X^n$ for $0 \leqq n \leqq N$. From (a), $y \in X^N$ implies $Cy = Cx$ and therefore $x \in E$. Thus, $X^N \subseteq E$ and the proof is complete.

For the next theorem we need some notation and preliminary results. Let the vector z denote the objective function values given by any point in X^N; that is, let $z = Cx$ for some $x \in X^N$. Without loss of generality assume that the sets J_n are constructed so that

$$\begin{aligned} J_0 &= \{1, 2, \cdots, k_0\}, & k_0 &= k, \\ J_1 &= \{1, 2, \cdots, k_1\}, & k_1 &< k_0, \\ J_2 &= \{1, 2, \cdots, k_2\}, & k_2 &< k_1, \\ &\vdots \\ J_N &= \phi. \end{aligned}$$

Note that if $k_{n+1} < i \leqq k_n$ (that is, $i \in J_n - J_{n+1}$), then $z_i = M_i - w^n$. From Algorithm A we know that $w^n > w^{n+1}$, $n = 0, 1, \cdots, N-1$. Furthermore, from our definition of the sets X^n, we observe that

$$\begin{aligned} X^n = \{x \in X \mid & C_i x = z_i \text{ for } i > k_{n-1} \\ & \text{and } C_i x \geqq M_i - w^{n-1} \text{ for } i \leqq K_{n-1}\}. \end{aligned}$$

THEOREM 3. *Let $z = Cx$ for some $x \in X^N$. For any $x \in X$, if $C_j x > z_j$ for some j such that $k_n < j \leqq k_{n-1}$, then there exists $i > k_n$ such that $C_i x < z_i$.*

Proof. Assume there exists a point $\bar{x} \in X$ such that $C_j \bar{x} > z_j$ and $C_i \bar{x} \geqq z_i$ for all $i > k_n$, where $k_n < j \leqq k_{n-1}$. By changing the choice of j, if necessary, we can assume that either $C_i \bar{x} = z_i$ for each $i > k_{n-1}$, or that $n = 1$, in which case $C_i \bar{x} = z_i$ for $i > k_{n-1}$, holds vacuously. We will show that the existence of such a point $\bar{x}$ is inconsistent with a property of the set $J_{n-1} - J_n$, namely that $C_i x = z_i$ for all $x \in X^n$ and all $i \in J_{n-1} - J_n = \{k_n + 1, k_n + 2, \cdots, k_{n-1}\}$.

Pick $\hat{x} \in X^N$ and let $x(\alpha) = \alpha \bar{x} + (1 - \alpha)\hat{x}$, $0 < \alpha < 1$. We will show that:

(1) $C_j x(\alpha) > z_j$ for $0 < \alpha < 1$, and

(2) $x(\alpha) \in X^n$ for α sufficiently small.

Recall that $C_j \bar{x} > z_j$ and $C_j \hat{x} = z_j$, so (1) follows immediately. For $i > k_{n-1}$, $C_i \bar{x} = z_i$ and for all i we have by definition that $C_i \hat{x} = z_i$. Thus, $C_i x(\alpha) = z_i$ for all $i > k_{n-1}$. For $k_n < i \leqq k_{n-1}$, $C_i \hat{x} = z_i = M_i - w^{n-1}$, and $C_i \bar{x} \geqq z_i = M_i - w^{n-1}$. Thus, for $k_n < i \leqq k_{n-1}$, $C_i x(\alpha) \geqq M_i - w^{n-1}$. For all i such that $k_m < i \leqq k_{m-1} \leqq k_n$, $C_i \hat{x} = z_i = M_i - w^{m-1} >$

$M_i - w^{n-1}$, so for α sufficiently small $C_i x(\alpha) > M_i - w^{n-1}$. Thus, $x(\alpha) \in X^n$ for α sufficiently small which proves (2) and completes the proof.

Before giving the characterization of the set X^N as the trade-off compromise set T, we state two corollaries of Theorem 3 related to the number of iterations needed for convergence of Algorithm A for small values of k.

COROLLARY 1. *If for any $j \in J_0 - J_1$ there is an $\bar{x} \in X$ such that $C_j\bar{x} > z_j = M_j - w^0$, then $J_0 - J_1$ contains at least two indices.*

Proof. Letting $n = 1$ in Theorem 3 provides the proof.

The next corollary shows that when $k = 3$, program P of § 2 reduces to an ordinary linear program.

COROLLARY 2. *If $k = 3$ and $\bar{w} > 0$ in program Q, then program P reduces to a linear program.*

Proof. For $j \in J_0 - J_1$, let $\bar{x}$ be optimal for max $C_j x$ subject to $x \in X$. Then $C_j\bar{x} = M_j > M_j - \bar{w}$, so by Corollary 1 at least two objectives are constant over X_∞, the feasible region for P.

The following corollary is simply a restatement of Theorem 3, in terms of the definition of a trade-off compromise point.

COROLLARY 3. *Suppose Algorithm* A *terminates with X^N. If $\bar{x} \in X^N$, then $\bar{x}$ is a trade-off compromise point.*

The following completes the characterization of X^N.

THEOREM 4. *If x is a trade-off compromise point, then $x \in X^N$.*

Proof. Given $x \in X$ with the trade-off compromise property, let $\bar{x}$ be a point in X^N. We will show that $Cx = C\bar{x}$ which implies that $x \in X^N$.

Assume that $C\bar{x} \neq Cx$. Since both $\bar{x}$ and x are efficient, we can choose an index i_1 such that $C_{i_1}(\bar{x}) > C_{i_1}(x)$ which implies

$$d_{i_1}(\bar{x}) < d_{i_1}(x). \tag{1}$$

Since $x \in T$, there is an index i_2 with

$$C_{i_2}(\bar{x}) < C_{i_2}(x) \quad \text{and} \quad d_{i_2}(x) \geqq d_{i_1}(x), \tag{2}$$

and from the strict inequality in (2) we also have

$$d_{i_2}(\bar{x}) > d_{i_2}(x). \tag{3}$$

From (1)–(3), we have

$$d_{i_1}(\bar{x}) < d_{i_1}(x) \leqq d_{i_2}(x) < d_{i_2}(\bar{x}). \tag{4}$$

Similarly, from

$$C_{i_2}(x) > C_{i_2}(\bar{x}),$$

and the fact that $\bar{x} \in T$ by Corollary 3, an analogous argument implies the existence of an index i_3 such that

$$d_{i_2}(x) < d_{i_2}\bar{x}) \leqq d_{i_3}(\bar{x}) < d_{i_3}(x). \tag{5}$$

Thus we have,

$$d_{i_1}(\bar{x}) < d_{i_1}(x) \leqq d_{i_2}(x) < d_{i_2}(\bar{x}) \leqq d_{i_3}(\bar{x}) < d_{i_3}(x). \tag{6}$$

But

$$d_{i_3}(\bar{x}) < d_{i_3}(x) \quad \text{implies } C_{i_3}(\bar{x}) > C_{i_3}(x).$$

Using the fact that $x \in T$ gives the existence of an index i_4 such that

$$d_{i_3}(\bar{x}) < d_{i_3}(x) \leqq d_{i_4}(x) < d_{i_4}(\bar{x}).$$

Notice that the indices i_1, i_2, i_3 and i_4 are distinct. Continuing, we could obtain an infinite sequence $\{i_n\}$ of distinct indices. However, these indices must be chosen from $\{1, 2, \cdots, k\}$. This contradiction implies that the assumption $Cx \neq C\bar{x}$ must be false; and the proof is complete.

Thus from Corollary 3 and Theorem 4 we conclude that the set X^N generated by Algorithm A is precisely the set T of trade-off compromise points. In the next section we present a numerical example to illustrate Algorithm A and to provide further insight into the trade-off compromise set T.

5. A numerical example. The multiple objective program used in this example was given by Yu and Zeleny in [17]. In [18], Isermann considers this same problem and lists all 29 nondominated extreme points. The problem has five objectives and is given as

$$\max Cx = \begin{bmatrix} 3 & -7 & 4 & 1 & 0 & -1 & -1 & 8 \\ 2 & 5 & 1 & -1 & 6 & 8 & 3 & -2 \\ 5 & -2 & 5 & 0 & 6 & 7 & 2 & 6 \\ 0 & 4 & -1 & -1 & -3 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_8 \end{bmatrix}$$

$$\text{subject to} \begin{bmatrix} 1 & 3 & -4 & 1 & -1 & 1 & 2 & 4 \\ 5 & 2 & 4 & -1 & 3 & 7 & 2 & 7 \\ 0 & 4 & -1 & -1 & -3 & 0 & 0 & 1 \\ -3 & -4 & 8 & 2 & 3 & -4 & 5 & -1 \\ 12 & 8 & -1 & 4 & 0 & 1 & 1 & 0 \\ -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 8 & -12 & -3 & 4 & -1 & 0 & 0 & 0 \\ 15 & -6 & 13 & 1 & 0 & 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_8 \end{bmatrix} \leqq \begin{bmatrix} 40 \\ 84 \\ 18 \\ 100 \\ 40 \\ -12 \\ 30 \\ 100 \end{bmatrix}$$

with $x_1, \cdots, x_8 \geqq 0$. The ideal vector M of individual maxima over the feasible set X is given by

$$M = \begin{bmatrix} 117.25 \\ 176.83 \\ 179.06 \\ 18.00 \\ 39.35 \end{bmatrix}$$

In the first iteration of Algorithm A for $n = 0$, we solve

$$Q_0: \quad \begin{aligned} &\min w \\ &\text{subject to } M - Cx \leqq we \text{ and } x \in X \end{aligned}$$

and obtain an optimal point (w^0, x^0) with

$$w^0 = 76.03,$$

$$x^0 = (0, 0, 0, 9.58, 17.54, 1.69, 0, 4.17)^T.$$

From the optimal tableau for Q_0 we note that there are multiple optima. Here

$$Cx^0 = \begin{bmatrix} 41.22 \\ 100.80 \\ 142.03 \\ -58.03 \\ 32.97 \end{bmatrix}$$

and we observe that

$$M_i - C_i x^0 = w^0, \quad \text{for } i = 1, 2 \text{ and } 4.$$

Thus objective functions C_3x and C_5x are the only ones that may get closer over X_∞ to their ideal point.

On the next iteration, $n = 1$, we therefore have $J_1 = \{3, 5\}$ and consider the linear program

$$\begin{aligned} Q_1: \quad & \min w \\ & \text{subject to } M_3 - C_3x \leqq w \\ & \qquad\qquad M_5 - C_5x \leqq w \\ & \qquad\qquad M - Cx \leqq w^0 e \text{ and } x \in X. \end{aligned}$$

Here we obtain an optimal point (w^1, x^1) with

$$w^1 = 37.03,$$

$$x^1 = (0, 0, 6.12, 10.80, 14.70, 2.91, 0, 1.11)^T,$$

with

$$Cx^1 = \begin{bmatrix} 41.22 \\ 100.80 \\ 142.03 \\ -58.03 \\ 35.01 \end{bmatrix}$$

In this case, there is only one objective with

$$M_i - C_i x^1 = w^1,$$

namely, for $i = 3$. This illustrates the one case where an iteration will not yield at least two objectives that are constant over the optimal set for Q_1. This happens because the third objective is, in fact, constant over X_∞. This can be determined by considering the optimal tableau for Q_0. Notice on this iteration that C_5x is closer to its ideal value and since the slack variable associated with $M_5 - C_5x \leqq w$ in Q_1 is positive at optimality, we see that $J_2 = \{5\}$ for the next iteration.

When $n = 2$, we therefore consider

$$\begin{aligned} Q_2: \quad & \min w \\ & \text{subject to } M_5 - C_5x \leqq w \\ & \qquad\qquad M_3 - C_3x \leqq w^1 \\ & \qquad\qquad M_5 - C_5x \leqq w^1 \\ & \qquad\qquad M - Cx \leqq w^0 e \text{ and } x \in X. \end{aligned}$$

On this iteration, we obtain

$$w^2 = 3.18,$$

$$x^2 = (0, 0, 1.40, 9.20, 17.10, 0, 4.59, 3.87)^T,$$

with

$$Cx^2 = \begin{bmatrix} 41.22 \\ 100.80 \\ 142.03 \\ -58.03 \\ 36.16 \end{bmatrix}$$

In this case, we have $J_3 = \phi$ and so the algorithm terminates. From the optimal tableau for Q_2 it is an easy matter to describe the set X^3, and from examining that tableau in this case we observe that the set X^3 consists of more than a single point; that is, program Q_2 has multiple optima, but the vector of objective function values Cx, is identical to Cx^2 over the set of multiple optima.

Of course, as discussed above, *this set X^3 is precisely equal to the trade-off compromise set T.* Suppose we re-order the objectives according to their distance (in the l_∞-norm) from the ideal point M. In this case we obtain the following schematic:

$$Z = \begin{array}{c|cc} & \text{objective} & \text{deviation} \\ \hline 36.16 & 5 & 3.18 \\ \hline 142.03 & 3 & 37.03 \\ \hline 41.22 & 1 & \\ 100.80 & 2 & 76.03 \\ -58.03 & 4 & \end{array}$$

Suppose for example, we wish to find a point $x \in X$ where the third objective is better off (has a deviation smaller than 37.03). We must in this case be willing to accept a decrease in the first, second or fourth objective. It is not possible, in view of the above development, to find a point in X where the third objective is increased *only* at the expense of the fifth objective.

Of course, the vector Z of objective function values for each point in the trade-off compromise set, as obtained above, is *dependent* upon the ideal vector M chosen. A different M may yield a different ordering of objectives but *once the trade-off compromise set for this ideal vector is determined*, then it too would give valuable information about other possible trade-offs.

Our intent is to develop an interactive procedure based on Algorithm A. To illustrate one such approach, we continue with the above. After obtaining the trade-off compromise set $T(M)$ and the corresponding set of objective function values $Z(M)$ given above, suppose we wish to find a nondominated point where objectives number 1 and 2 are increased from their current values of 41.22 and 100.80 respectively. Of course, since these objectives are in the "poorest class" we must be willing to accept a decrease in the fourth objective which is also in that class and we assume here that such a decrease is acceptable. One way to proceed would be as follows. Instead of using the original ideal vector M, use

$$M^1 = M + 10e_1 + 10e_2,$$

where e_i is the identity vector with all components equalling zero excepting the ith component, which is equal to one. Using this ideal vector M^1, the application of

Algorithm A yields a trade-off compromise set $T(M^1)$ given in the schematic below:

$$Z(M^1)=\begin{array}{|c|c|c|}\hline 36.56 & 5 & 2.79 \\ \hline 148.69 & 3 & 30.37 \\ \hline 44.56 & 1 & \\ 104.14 & 2 & 82.69 \\ -64.69 & 4 & \\ \hline\end{array}\quad \begin{array}{c}\text{objective}\\ \\ \\ \\ \\ \\ \end{array}\quad \begin{array}{c}\text{deviation}\\ \\ \\ \\ \\ \\ \end{array}.$$

Notice that objective number four has decreased as we know it must, and in this case, all the other objectives actually increased. Suppose now that objectives 1, 2 and 4 are at their minimum acceptable levels. Our theory tells us that it is impossible to find a nondominated point where objective number three is greater than 148.69. In addition, if objective number 5 is to be increased above 36.56 then we must be willing to accept a decrease in the third objective from 148.69 if no decrease in the poorest class is acceptable.

By appropriately changing the ideal vector M we can alter the "class structure" of the vector $Z(M)$; that is, some objectives that are, for example, in the poorest class relative to one ideal vector may not be poor relative to another ideal vector. To illustrate this point, consider the above example where the ideal vector is $M^2 = M + 60e_3$. This choice will increase the actual value $Z_3(M)$ but at the same time may put the third objective in the poorest class because its deviation from $M_3 + 60$ will be large. If we apply Algorithm A using this ideal vector M^2 then we obtain the following:

$$Z(M^2)=\begin{array}{|c|c|c|}\hline 36.78 & 5 & 2.57 \\ \hline 106.06 & 2 & 70.76 \\ 46.48 & 1 & \\ \hline 152.53 & 3 & 86.53 \\ -68.53 & 4 & \\ \hline\end{array}\quad \begin{array}{c}\text{objective}\\ \\ \\ \\ \\ \\ \end{array}\quad \begin{array}{c}\text{deviation}\\ \\ \\ \\ \\ \\ \end{array}.$$

Here the "class structure" has changed considerably and the information on possible trade-offs that is provided by this class structure is also different. Objectives 1 and 2, for example, are no longer in the poorest class. If we are unwilling to accept a decrease in objectives 3 or 4 from their current levels then it is impossible to find a feasible point where objectives 1 and 2 are increased from their current levels.

We are currently investigating an iteractive procedure for systematically using the theory developed in this paper to analyze trade-offs. For example, if a decision maker decides that an objective does not need to be increased above its current level then that objective could be eliminated by adding new constraints. The smaller multiple objective problem could then be analyzed and possible trade-offs could be explored.

Acknowledgment. The authors would like to thank Mr. Michael Kupferschmid for his valuable computer programming assistance.

REFERENCES

[1] W. DINKELBACH, *Uber einen Losungsatz zum Vektormaximumproblem*, Unternehmensforschung-Heute, M. Beckmann and H. P. Kunzi, eds., Springer-Verlag, Berlin, 1971, pp. 1–13.

[2] W. DINKELBACH AND W. DURR, *Effizienzaussagen bei Ersatzprogrammen zum Vektormaximum-problem*, Multiple Criteria Decision Making, Operations Research Verfabren, H. P. Kunzi and H. Shubert, eds., Verlag Anton Hain, Meisenheim, Germany, 1972, pp. 117–123.

[3] W. DINKELBACH AND H. ISERMANN, *On decision making under multiple criteria*, Multiple Criteria Decision Making, J. L. Cochrane and M. Zeleny, eds., University of South Carolina Press, Columbia, SC, 1973, pp. 302–312.

[4] M. FREIMER AND P. L. YU, *Some new results on compromise solutions for group decision problems*, Management Sci., 22 (1976), pp. 688–693.

[5] W. B. GEARHART, *Compromise solutions and estimation of the noninferior set*, J. Optim. Theory Appl., 28 (1979), pp. 29–47.

[6] P. L. YU AND G. LEITMANN, *Compromise solutions, domination structures, and Salukvadze's solution*, Ibid., 13 (1974), pp. 362–378.

[7] P. L. YU, *A class of solutions for group decision problems*, Management Sci. 19 (1973), pp. 936–946.

[8] M. ZELENY, *Compromise programming*, Multiple Criteria Decision Making, J. L. Cochrane and M. Zeleny, eds., University of South Carolina Press, Columbia, South Carolina, 1973), pp. 262–301.

[9] ———, *A concept of compromise solutions and the method of the displaced ideal*, Comput. Oper. Res., 1 (1974), pp. 479–496.

[10] ———, *The theory of the displaced ideal*, Multiple Criteria Decision Making, Kyoto, 1975, M. Zeleny, ed., Springer Verlag, New York, 1975, pp. 151–205.

[11] H. M. WAGNER, *Linear programming techniques for regression analysis*, J. Amer. Stat. Assoc., 54 (1959), pp. 206–212.

[12] M. S. HEGNER, *Multiple objective linear programming*, Ph.D thesis, Rensselaer Polytechnic Institute, Troy, NY, 1977.

[13] J. PHILIP, *Algorithms for the vector maximization problem*, Math. Prog., 2 (1972), pp. 207–229.

[14] T. GAL, *A general method for determining the set of all efficient solutions to a linear vectormaximum problem*, European J. Oper. Res., to appear.

[15] H. ISERMANN, *The enumeration of the set of all efficient solution for a linear multiple objective program*, Oper. Res. Quarterly, 28 (1977), pp. 711–725.

[16] J. G. ECKER, N. S. HEGNER AND I. A. KOUADA, *Generating all maximal efficient faces for multiple objective linear programs*, JOTA, 30 (1980), no. 3, pp. 353–381.

[17] P. L. YU AND M. ZELENY, *The set of all nondominated solutions in linear cases and a multicriteria simplex method*, J. Math, Anal. Appl., 49 (1975), pp. 430–468.

[18] H. ISERMANN, *The enumeration of the set of all efficient solutions for a linear multiple objective program*, Rep. B7601, University of Saarland, D6600 Saarbrucker, March 1976.

[19] J. P. EVANS AND R. E. STEUER, *A revised simplex method for linear multiple objective programs*, Math. Prog., 5 (1973), pp. 54–72.

[20] J. G. ECKER AND I. A. KOUADA, *Finding all efficient extreme points for multiple objective linear programs*, Math. Prog., 14 (1978), pp. 249–261.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 4, July 1981

0363-0129/81/1904-0007 $01.00/0

NECESSARY AND SUFFICIENT CONDITIONS OF APPROXIMATE CONTROLLABILITY FOR GENERAL LINEAR RETARDED SYSTEMS*

A. MANITIUS†

Abstract. Necessary and sufficient conditions of approximate controllability, in the space $R^n \times L_2([-h, 0], R^n)$, of general linear retarded systems are obtained. It is shown that approximate controllability is equivalent to two conditions: a) spectral controllability, and b) the existence of linear feedback which transforms the original system into a system with a complete set of generalized eigenfunctions. Both conditions are expressed in algebraic form. The proof of this result is based on recently obtained criteria of completeness of generalized eigenfunctions associated with retarded systems and on an algebraic approach to functional differential equations. Practical verifiability of the new conditions is demonstrated on several examples.

1. Introduction. This paper gives necessary and sufficient conditions of approximate controllability, in the space $R^n \times L_2([-h, 0], R^n)$, for linear autonomous retarded functional equations (FDE) of the general form

$$\dot{z}(t) = \int_{-h}^{0} d\eta(\theta) z(t+\theta) + B_0 u(t), \tag{1.1}$$

where $z(t) \in R^n$, $u(t) \in R^m$, $h < \infty$, $\eta(\cdot)$ is an $n \times n$ matrix of functions of bounded variation, consisting of an absolutely continuous part of a finite number of jump discontinuities, and B_0 is an $n \times m$ matrix.

The approximate function space controllability of simple retarded systems of the form

$$\dot{z}(t) = A_0 z(t) + A_1 z(t-h) + B_0 u(t), \tag{1.2}$$

where A_0, A_1 are $n \times n$ matrices, has been investigated by this author and R. Triggiani [16]. Starting from abstract operator type controllability conditions for differential equations in Banach spaces, we have obtained several verifiable conditions expressed in terms of the original system matrices. It remained, however, an open question whether such verifiable conditions could be obtained for general systems of type (1.1).

In the present paper such conditions are obtained by making use of several recent results, namely the criteria of completeness of the generalized eigenfunctions associated with equation (1.1) [14], the properties of the structural operator F induced by η [3], [6], and the algebraic approach to functional differential equations in the style of [2], [13].

As a corollary to these new conditions, a very simple controllability criterion is obtained for systems (1.2). This criterion extends the result [16, Prop. 7.6] to an arbitrary n, giving in this way a counterpart to the existing results for exact controllability of neutral systems [20], [21].

A somewhat surprising feature of these results is that the approximate controllability, which is a topological notion involving the closure of a subspace of a Hilbert space, can be verified via an algebraic criterion.

* Received by the editors January 1, 1980, and in revised form September 4, 1980. This research was supported in part by NSERC of Canada Grant A-9240 and in part under Grant FCAC 1978/79 of the Quebec Ministry of Education.

† Centre de Recherche de Mathématiques Appliquées, Université de Montréal, C.P. 6128, Montreal, Quebec, Canada H3C 3J7.

From a more general point of view, the results of this paper are of some relevance to the theory of abstract systems governed by differential equations in Banach spaces. As the paper shows, linear retarded systems provide an example of a class of infinite dimensional systems in which approximate controllability is *equivalent* to two conditions: a) spectral controllability, and b) the existence of linear feedback that transforms the original system into a system with a complete set of generalized eigenfunctions. It may be interesting to see whether there are other physically significant examples of systems of the same class.

Other recent contributions to the controllability of FDE's are described in [11], [12], [13], [15], [16], [19], [21], which contain many references to earlier research in this area. An up-to-date survey of controllability of other infinite dimensional systems is contained in [23].

As in [16], the investigation of approximate controllability of equation (1.1), in the space $R^n \times L_2([-h, 0], R^n)$, has been motivated by the existence of general controllability criteria for abstract differential equations in Banach spaces [7], [25]. Since these latter were usually expressed in an abstract operator form, their translation into concrete, practically verifiable criteria for various special classes of equations was a problem in itself (for more details see [16, § 1]). The present paper completes this task for systems governed by (1.1). However, other aspects of controllability of systems (1.1) still are worthy of further investigation; for instance, it is not known what the approximate controllability of (1.1) implies about the type of closed-loop system behavior that can be achieved by using state feedback; the relationship between this type of controllability and the one obtained via an algebraic approach [13] is also not known.

2. Notation and preliminaries. Let R and C denote the fields of real and complex numbers, respectively. The symbol R^n will denote n-dimensional Euclidean space. The letters $\mathbb{R}$ and J will denote rings to be specified later. $L_p(a, b)$, $L_p([a, b], R^n)$, $p = 1, 2$, will denote the customary Lebesgue spaces of scalar valued functions or n-vector valued functions, respectively, on $[a, b]$.

Let $\mathscr{X}$ denote the Hilbert space $R^n \times L_2([-h, 0], R^n)$ now often used in studies of FDE's (e.g., [1], [3], [5]). We assume that $h \in (0, \infty)$. For $x \in \mathscr{X}$, let x^0, x^1 denote its R^n and $L_2([-h, 0], R^n)$ components, respectively; i.e., $x = (x^0, x^1)$. If K is a subset of $\mathscr{X}$, $\bar{K}$ will denote its closure in the strong topology of $\mathscr{X}$. If H is a bounded linear operator from one Hilbert space to another, H^* will denote its adjoint. A centered asterisk $*$ will denote a certain convolution product. For vectors, elements of R^n, or for matrices, the superscript T will denote a transpose. The symbol χ_I will denote a characteristic function of a set I.

It is well known (e.g., [1], [3], [5]) that (1.1) induces a strongly continuous semigroup $\{S(t)\}_{t \geqq 0}$ on $\mathscr{X}$. Let $z(t)$ be a solution of (1.1) corresponding to some initial conditions $z(0) = \phi^0$, $z(\theta) = \phi^1(\theta)$, $\theta \in [-h, 0)$, $\phi = (\phi^0, \phi^1) \in \mathscr{X}$, and to some control $u(\cdot) \in L_1([0, T], R^m)$, $T > 0$. Let z_t denote the function $\theta \to z_t(\theta) = z(t + \theta)$, $\theta \in [-h, 0]$. It is well known (e.g., [1]) that $x(t) = (z(t), z_t) \in \mathscr{X}$ is the "mild solution" of the abstract differential equation

$$\dot{x}(t) = Ax(t) + Bu(t), \qquad x(0) = \phi, \quad t \geqq 0, \tag{2.1}$$

where $A : \mathscr{D}(A) \subset \mathscr{X} \to \mathscr{X}$ is the infinitesimal generator of $\{S(t)\}_{t \geqq 0}$, ($\mathscr{D}(A)$ is the domain of A), and $B : R^m \to \mathscr{X}$ is a bounded linear operator $Bu = (B_0 u, 0)$. For a given $t > 0$, let K_t denote the attainable set at time t of (2.1) corresponding to $\phi = 0$, [16], and let $K_\infty = \bigcup_{t > 0} K_t$.

System (2.1), or equivalently system (1.1), is said to be *approximately controllable* if $\bar{K}_\infty = \mathscr{X}$.

Throughout this paper we assume that $\eta(\cdot)$ in (1.1) is given by

$$\eta(\theta) = -A_0\chi_{(-\infty,0)}(\theta) - \sum_{i=1}^{N} A_i\chi_{(-\infty,-h_i)}(\theta) - \int_\theta^0 E(\alpha)\,d\alpha, \tag{2.2}$$

where N is a fixed positive integer, $0 = h_0 < h_1 < \cdots < h_N = h$, A_i are constant real $n \times n$ matrices and $E(\cdot)$ is an $n \times n$ matrix of functions in $L_2(-h, 0)$. With this $\eta(\cdot)$, (1.1) can also be written as

$$\dot{z}(t) = \sum_{i=0}^{N} A_i z(t-h_i) + \int_{-h}^{0} E(\theta)z(t+\theta)\,d\theta + B_0u(t). \tag{2.3}$$

Let $\tilde{\eta}(\theta) = \eta(\theta) + A_0\chi_{(-\infty,0)}(\theta)$, and define

$$(H\phi)(\theta) = \int_{-h}^{\theta} d\tilde{\eta}(s)\phi(s-\theta), \qquad \theta \in [-h, 0]. \tag{2.4}$$

It is known [3], that H is a linear continuous mapping from $L_2([-h, 0], R^n)$ into itself. By definition, H depends only on the strictly retarded part $\tilde{\eta}$ of η. The adjoint H^* of H has the same form as H except that the matrix $\tilde{\eta}$ is replaced by its transpose $\tilde{\eta}^T$. Note that the right-hand side of (2.4) is a convolution. H induces the mapping

$$F : \mathscr{X} \to \mathscr{X}, \qquad F = \begin{bmatrix} I & 0 \\ 0 & H \end{bmatrix},$$

(where I is the identity map on R^n) referred to as the "structural operator". For more details, see [3], [6], [15].

In this paper controllability will be related to the spectral properties of system (1.1). Let $\Delta(\lambda)$ be the characteristic matrix

$$\Delta(\lambda) = I\lambda - \int_{-h}^{0} d\eta(\theta)\, e^{\lambda\theta}. \tag{2.5}$$

We recall that the spectrum of A is $\sigma(A) = \{\lambda \,|\det \Delta(\lambda) = 0\}$. For $\lambda \in \sigma(A)$, let $\mathscr{M}_\lambda$ denote the (largest) generalized eigenspace of A corresponding to λ, that is $\mathscr{M}_\lambda = \operatorname{Ker}(I\lambda - A)^k$, where k is a positive integer such that $\operatorname{Ker}(I\lambda - A)^k = \operatorname{Ker}(I\lambda - A)^{k+j}$, $j = 1, 2, \cdots$ (for a characterization of elements of $\mathscr{M}_\lambda$ see [14]). Let $\mathscr{M} = \operatorname{span}\{\mathscr{M}_\lambda \,|\lambda \in \sigma(A)\}$.

The generalized eigenfunctions of A are said to be *complete* in $\mathscr{X}$ if $\bar{\mathscr{M}} = \mathscr{X}$. By [14, Thm. 5.1], the completeness holds if and only if $\operatorname{Ker} H^* = \{0\}$.

We recall that (2.1) is said to be *spectrally controllable*, if for each $\lambda \in \sigma(A)$ the canonical projection of (2.1) on $\mathscr{M}_\lambda$ is completely controllable [16]. By [20] a necessary and sufficient condition for spectral controllability of a retarded system is

$$\operatorname{rank}[\Delta(\lambda), B_0] = n \quad \forall \lambda \in C. \tag{2.6}$$

Finally, we recall an abstract characterization of controllability [7]. System (2.1) is approximately controllable if and only if

$$B^*S^*(t)x = 0 \quad \forall t \geqq 0 \Rightarrow x = 0, \tag{2.7}$$

where B^* and $S^*(t)$ are adjoints of B and $S(t)$, respectively.

3. General necessary conditions. For simple systems (1.2), conditions of controllability were given in [16] in a number of different forms. A necessary condition [16,

Thm. 3.1] is that a rank of certain $n \times nm$ polynomial matrix $P(\lambda)$ be equal to n. This implies another, simpler, necessary condition

$$\operatorname{rank}[A_1, B_0] = n. \tag{3.1}$$

Independently of (3.1), spectral controllability is also necessary for approximate controllability [16, Prop. 7.1].

We now give a corresponding result for general systems (1.1). Let us define the following linear mapping, $D^*: L_2([-h, 0], R^n) \to L_2([-h, 0], R^m)$

$$(D^*\phi)(\theta) = B_0^T \phi(\theta) \quad \text{a.e.}, \qquad \theta \in [-h, 0], \tag{3.2}$$

where B_0^T is the transpose of B_0.

THEOREM 1. *If system* (1.1) *is approximately controllable, then*:

(i) $\operatorname{rank}[\Delta(\lambda), B_0] = n$, $\forall \lambda \in C$; *and*

(ii) $\operatorname{Ker} H^* \cap \operatorname{Ker} D^* = \{0\}$.

Proof. The necessity of condition (i) is known (see e.g., [16, Prop. 7.1], or [7, Cor. 3.2] combined with (2.6))[1]. It remains to prove (ii).

Suppose that (ii) is not satisfied; i.e., there exists $\phi \neq 0$, $\phi \in L_2([-h, 0], R^n)$ such that $D^*\phi = 0$ and $H^*\phi = 0$; i.e.,

$$\int_{-h}^{\alpha} d\tilde{\eta}^T(\theta)\phi(\theta - \alpha) = 0 \quad \text{a.e.}, \qquad \alpha \in [-h, 0].$$

Let

$$\psi^1(\theta) = \int_{\theta}^{0} \phi(\alpha)\, d\alpha, \qquad \theta \in [-h, 0],$$

and let $\psi = (0, \psi^1)$. Then ψ^1 is absolutely continuous, its derivative is in L^2, and $\psi^1(0) = 0$. Hence $0 \neq \psi \in \mathcal{D}(A)$. Furthermore,

$$(D^*\psi^1)(\theta) = \int_{\theta}^{0} B_0^T \phi(\alpha)\, d\alpha = 0 \quad \forall \theta \in [-h, 0]. \tag{3.3}$$

By [6, Lemma 3.2] the function $\theta \to (H^*\psi^1)(\theta)$ is absolutely continuous, $(H^*\psi^1)(-h) = 0$, and

$$\frac{d}{d\alpha}\int_{-h}^{\alpha} d\tilde{\eta}^T(\theta)\psi^1(\theta - \alpha) = -\int_{-h}^{\alpha} d\tilde{\eta}^T(\theta)\frac{d}{d\alpha}\psi^1(\theta - \alpha) = \int_{-h}^{\alpha} d\eta^T(\theta)\phi(\theta - \alpha) = 0.$$

Hence $\psi^1 \in \operatorname{Ker} H^* \cap \operatorname{Ker} D^*$ and $\psi \in \operatorname{Ker} F^* \cap \mathcal{D}(A)$.

Let $G^*: \mathcal{X} \to \mathcal{X}$ be the linear operator

$$[G^*\xi]^1(\theta) = X^T(h + \theta)\xi^0 + \int_{-h}^{0} X^T(h + \theta + s)\xi^1(s)\, ds, \qquad \theta \in [-h, 0],$$

$$[G^*\xi]^0 = [G^*\xi]^1(0)$$

where $X(\cdot)$ is the fundamental matrix solution of the homogeneous part of equation (1.1). By [14, Props. 3.1, 3.2] G^* is one-to-one, continuous and its image in $\mathcal{X}$ coincides with $\mathcal{D}(A)$. Since $\psi \in \mathcal{D}(A)$, there exists a nonzero $\xi \in \mathcal{X}$ such that $\psi = G^*\xi$. But $F^*\psi = 0$; hence $F^*G^*\xi = 0$. Since, by [14, Prop. 3.5], $F^*G^* = S^*(h)$, where $S^*(t)$ is the

[1] See also § 7.3.

adjoint semigroup of $S(t)$, we have, by the semigroup property, that

$$S^*(t)\xi = 0 \quad \forall t \geqq h,$$

hence also

$$B^*S^*(t)\xi = 0 \quad \forall t \geqq h. \tag{3.4}$$

For $t \in [0, h]$

$$B^*S^*(t)\xi = (B_0^T, 0)S^*(t)\xi = B_0^T[S^*(t)\xi]^0. \tag{3.5}$$

Let $t = \theta + h$, $\theta \in [-h, 0]$. The explicit representation of $S^*(t)$ [3, (5.18)] yields

$$\begin{aligned} B_0^T[S^*(\theta+h)\xi]^0 &= B_0^T\left[X^T(h+\theta)\xi^0 + \int_{-h}^{0} X^T(h+\theta+s)\xi^1(s)\,ds\right] \\ &= B_0^T[G^*\xi]^1(\theta) = B_0^T\psi^1(\theta) \\ &= (D^*\psi^1)(\theta) = 0 \quad \forall \theta \in [-h, 0]. \end{aligned}$$

Hence $B^*S^*(t)\xi = 0$ for all $t \geqq 0$, while $\xi \neq 0$. By condition (2.7) the system is not approximately controllable. □

Remark 1. Condition (ii) of Theorem 1 contains (3.1) as a special case. For system (1.2)

$$(H^*\phi)(\theta) = A_1^T\phi(-h-\theta), \qquad \theta \in [-h, 0].$$

Hence (ii) can be written

$$A_1^T\phi(-h-\theta) = 0 \quad \text{a.e.} \quad \text{and } B_0^T\phi(\theta) = 0 \quad \text{a.e.} \Rightarrow \phi(\theta) = 0 \quad \text{a.e.} \quad \text{in } [-\text{h}, 0].$$

This yields (3.1) at once. One can also prove that for differential-difference equations ($E(\cdot) \equiv 0$) the analogous necessary condition is rank $[A_N, B_0] = n$. This will be further discussed in § 5.

4. An algebraic characterization of controllability. In this section condition (ii) of Theorem 1 is transformed into an equivalent algebraic form which has two advantages:

a) It is more easily verifiable than the original condition (ii);

b) It will enable us to prove the sufficiency of the condition (ii).

It will be more convenient to work on the interval $[0, h]$ than $[-h, 0]$. By making the substitutions: $s = \tau - h$, $\tau \in [0, h]$, $t = h + \theta$, $t \in [0, h]$, $\hat{\eta}(\tau) = \tilde{\eta}(\tau - h)$, $\tau \in [0, h]$, $\psi(t) = \phi(-t)$, $t \in [0, h]$, in (2.4) one has that (2.4) gives rise to a mapping $\hat{H} : L_2([0, h], R^n) \to L_2([0, h], R^n)$ defined by

$$\begin{aligned} (\hat{H}\psi)(t) &= \int_0^t d\hat{\eta}(\tau)\psi(t-\tau), \qquad t \in [0, h], \\ &= A_N\psi(t) + \sum_{i=1}^{N-1} A_i\psi(t-b_i)\chi_{[b_i,h]} + \int_0^t \hat{E}(\tau)\psi(t-\tau)\,d\tau, \end{aligned} \tag{4.1}$$

where $\hat{E}(\tau) = E(\tau - h)$, $\tau \in [0, h]$, $b_i = h - h_i$, $i = 1, \cdots, N-1$. The value of $\tilde{\eta}$ at $-h$ becomes the value of $\hat{\eta}$ at 0. Analogous transformations performed on H^* yield $\hat{H}^*$ which differs from $\hat{H}$ only by a transposition of all the matrices.

At this point we shall employ an interesting idea due to Bartosiewicz [2], which consists of representing $\hat{H}$ as a certain module homomorphism. Similar algebraic techniques had been first applied to FDE's by Kamen [13].

It is well known [8, § 1, § 16] that the space $L_1(0, h)$ endowed with standard addition and with the convolution product

$$(f*g)(t)=\int_0^t f(t-\tau)g(\tau)\,d\tau \qquad 0\leqq t\leqq h \tag{4.2}$$

is a *commutative ring*. If we extend f and g to $[0, \infty)$, by putting both $f(t)$ and $g(t)$ equal to zero for $t>h$, then the product (4.2) in $L_1(0, h)$ is a *restriction* to the interval $[0, h]$ of the convolution product of the corresponding functions in $L_1(0, \infty)$.

Let δ_b denote the Dirac distribution concentrated at the point $\{b\}$. Let J denote the set consisting of all formal sums of the form

$$\mu=\sum_{i=0}^{N} a_i\delta_{b_i}+f, \tag{4.3}$$

where $N\in\{0, 1, 2, \cdots\}$, a_i, b_i are real numbers, $0=b_0<b_1<\cdots<b_N\leqq h$ and $f\in L_1(0, h)$. μ is a distribution with a support contained in $[0, h]$.

The set J can be regarded as a subspace of a space $\mathscr{D}'_+$ of distributions over R, with support bounded on the left. It is well known [24], [22] that the convolution of distributions $\in\mathscr{D}'_+$ is associative and commutative. If $\mu, \nu\in J$, let $\mu*\nu$ denote the *restriction* to $[0, h]$ of the standard convolution of distributions $\in\mathscr{D}'_+$. Hence if $\mu=\sum_i a_i\delta_{b_i}+f$, $\nu=\sum_j c_j\delta_{d_j}+g$ then, by [24, Chapt. VI, Thm. VII],

$$\mu*\nu=\sum_i\sum_j (a_i\delta_{b_i})*(c_j\delta_{d_j})+\sum_i (a_i\delta_{b_i})*g+\sum_j (c_j\delta_{d_j})*f+f*g, \tag{4.4}$$

where

$$(a\delta_b)*(c\delta_d)=\begin{cases} ac\delta_{b+d} & \text{if } b+d\leqq h,\\ 0 & \text{otherwise,}\end{cases} \tag{4.5}$$

$$(\delta_d*f)\,(t)=f(t-d)\chi_{[d,h]}, \qquad t\in[0, h], \tag{4.6}$$

and $f*g$ is given by (4.2). It follows that $\mu*\nu\in J$ and the product $*$ is associative and commutative. Hence J is a commutative ring. It has a unit element given by δ_0. As opposed to the ring $\mathscr{D}'_+$, the ring J has divisors of zero, characterized by the following statement [2].

PROPOSITION 2. *$\mu\in J$ given by* (4.3) *is a divisor of zero in the ring J if and only if $a_0=0$ and $f(t)=0$* a.e. *on* $[0, \varepsilon)$ *for some $\varepsilon>0$.*

One can equivalently say that the support of a divisor of zero, μ, satisfies $\operatorname{supp}\mu\in[\varepsilon, h]$ for some $\varepsilon>0$. The proof of this result depends on Titchmarsh's theorem [18] and on the fact that $a_0\delta_0+f$ with $a_0\neq 0$, is invertible [8, § 18, *2].

The following two corollaries are crucial for the proof of the sufficient condition.

COROLLARY 3. *If μ and ν are divisors of zero in J, the element $\mu+\nu$ is a divisor of zero in J.*

Proof. One has $\operatorname{supp}\mu\in[\varepsilon_1, h]$, $\varepsilon_1>0$ and $\operatorname{supp}\nu\in[\varepsilon_2, h]$, $\varepsilon_2>0$. Hence $\mu+\omega$ is null on $[0, \min(\varepsilon_1, \varepsilon_2)]$. □

COROLLARY 4. *If a finite subset $\{\mu_1, \cdots, \mu_k\}$ of J does not have a nonzero annihilator, then at least one of its elements is not a divisor of zero.*

Proof. If all μ_j, $j=1, \cdots, k$ are divisors of zero, then their supports are contained respectively in $[\varepsilon_j, h]$, $h\geqq\varepsilon_j>0$, $j=1, \cdots, k$. Let $\varepsilon_0=\min\varepsilon_j$; then $\varepsilon_0>0$. Let $0<\beta<\varepsilon_0$. The element $\delta_{h-\beta}\in J$ is nonnull and $\delta_{h-\beta}*\mu_j=0$, $j=1, \cdots, k$. □

Remark 2. Properties given by Corollaries 3 and 4 do not hold in some well-known rings. Take the ring $\mathscr{C}$, of continuous functions on $[0, 1]$ with pointwise multiplication [8, § 10]. Divisors of zero in $\mathscr{C}$ are functions that vanish on a subinterval of $[0, 1]$ of positive measure. Let $f, g \in C$ be given by

$$f(t) = \begin{cases} 0, & t \in [0, \frac{1}{3}], \\ t - \frac{1}{3}, & t \in [\frac{1}{3}, 1], \end{cases} \qquad g(t) = \begin{cases} \frac{2}{3} - t, & t \in [0, \frac{2}{3}], \\ 0 & t \in [\frac{2}{3}, 1]. \end{cases}$$

Both f and g are divisors of zero; however, $f+g$ is not a divisor of zero, nor do f and g have a common nonzero annihilator. Another example of such a ring is given in § 7.2.

Let M be an $n \times n$ matrix over J with elements m_{ij}. We recall that the determinant of M over J is defined by (see, e.g., [4], [10])

$$\det_J M = \sum_P (-1)^\sigma m_{1j_1} * m_{2j_2} * \cdots * m_{nj_n}, \tag{4.7}$$

(where the sum is taken over all permutations P of indices $j_1, \cdots, j_n$, and $(-1)^\sigma$ is a signature of the permutation). Let J^n denote the standard free finitely generated module $J \times J \times \cdots \times J$, [10]. Let $u = (u_1, \cdots, u_n)$, $v = (v_1, \cdots, v_n)$, $u_i, v_i \in J$. The mapping $u \to v = M * u$ defined by

$$v_i = \sum_{j=1}^{n} m_{ij} * u_j, \qquad i = 1, \cdots, n,$$

is a *module homomorphism* $J^n \to J^n$.

Let $\mathscr{A}$ denote the $n \times n$ matrix over J

$$\mathscr{A} = A_N \delta_0 + \sum_{i=1}^{N-1} A_i \delta_{b_i} + \hat{E}, \tag{4.8}$$

where A_i, $\hat{E}$ and b_i are as (4.1). It follows from the form of $\hat{H}$ and from the multiplication rules (4.2) and (4.6) that for $\psi = (\psi_1, \cdots, \psi_n)$ with $\psi_i \in L_2(0, h)$

$$\hat{H}\psi = \mathscr{A} * \psi. \tag{4.9}$$

The following result combines two earlier results given by [14, Thm. 5.1] and [2, Thm. 1]. Here $\mathscr{A}^T$ is the transpose of $\mathscr{A}$.

THEOREM 5. *The following statements are equivalent*:

(i) *the generalized eigenfunctions of A are complete in* $\mathscr{X}$;

(ii) $\operatorname{Ker} H^* = \{0\}$;

(iii) $\mathscr{A}^T * u = 0 \Rightarrow u = 0$ *for* $u \in J^n$;

(iv) $\det_J \mathscr{A}$ *is not a divisor of zero in* J.

Proof. See § 7. □

COROLLARY 6. $\operatorname{Ker} F = \{0\} \Leftrightarrow \operatorname{Ker} F^* = \{0\}$.

Proof. This follows from the fact that $\det_J \mathscr{A} = \det_J \mathscr{A}^T$. □

We are now ready to recast statement (ii) of Theorem 1 into algebraic form. Let $\mathscr{B}$ be an $n \times m$ matrix over J defined by

$$\mathscr{B} = B_0 \delta_0. \tag{4.10}$$

One easily verifies that the mapping $\phi \to D^*\phi$ given by (3.2) can be represented by the module homomorphism $J^n \to J^m$ given by $\psi \to \mathscr{B}^T * \psi$, where $\psi = (\psi_1, \cdots, \psi_n)$, $\psi_i \in L_2(0, h)$, $\psi_i(t) = \phi_i(-t)$, $t \in [0, h]$, with similar identifications for $D^*\phi$ and $\mathscr{B}^T * \psi$.

PROPOSITION 7. *A necessary condition for approximate controllability of system* (1.1) *is*

$$\mathscr{A}^T * u = 0 \quad \text{and} \quad \mathscr{B}^T * u = 0 \Rightarrow u = 0. \tag{4.11}$$

Proof. This follows from condition (ii) of Theorem 1 and the formalism introduced above. In particular, the equation $D^*\phi = 0$ becomes $(B_0^T\phi)(\theta) = 0$ a.e. for $\theta \in [-h, 0]$, that is $(B_0^T\psi)(t) = 0$ a.e. $t \in [0, h]$ where $\psi \in L_2([0, h], R^n)$; that is $B_0^T\delta_0 * \psi = 0$; hence $\mathscr{B}^T * \psi = 0$. This, (4.9) and Remark 1 of § 7 conclude the proof. □

The interpretation of (4.11) *is that the n rows of the* $n \times (n+m)$ *matrix* $[\mathscr{A}, \mathscr{B}]$ *over J should be linearly independent.*

Given a rectangular matrix M with elements in some commutative ring $\mathbb{R}$, its rank, denoted by $\operatorname{rank}_{\mathbb{R}} M$ is defined [17] as the greatest positive integer r such that the set of all determinants (over $\mathbb{R}$) of square minors of M of order r, does not have a nonzero annihilator. If all the elements of M have a common nonzero annihilator, then $\operatorname{rank}_{\mathbb{R}} M$ is 0.

We now state the main result.

THEOREM 8. *A necessary and sufficient condition for approximate controllability of* (1.1) *is*

(i) $\operatorname{rank}[\Delta(\lambda), B_0] = n \quad \forall \lambda \in C$ *and*

(ii) $\operatorname{rank}_J [\mathscr{A}, \mathscr{B}] = n$.

Proof. See § 5. □

Condition (ii) means that the set of all the determinants $\det_J$ of square minors of order n of the matrix $[\mathscr{A}, \mathscr{B}]$ does not have a common annihilator. By recalling Corollary 4, condition (ii) can be stated alternatively as:

(ii)′ *Among all the determinants* (4.7) *of square minors of order n of the matrix* $[\mathscr{A}, \mathscr{B}]$ *there is at least one which is not a divisor of zero.*

COROLLARY 9. *Consider a differential difference equation*

$$\dot{z}(t) = \sum_{i=0}^{N} A_i z(t - h_i) + B_0 u(t). \tag{4.12}$$

A necessary and sufficient condition of approximate controllability is

(i) $\operatorname{rank}[\Delta(\lambda), B_0] = n \quad \forall \lambda \in C$ *and*

(ii) $\operatorname{rank}[A_N, B_0] = n$.

Proof. For system (4.12) the matrix $[\mathscr{A}, \mathscr{B}]$ is given by

$$[\mathscr{A}, \mathscr{B}] = [A_N\delta_0 + \sum_{i=1}^{N-1} A_i\delta_{b_i},\ B_0\delta_0],$$

where $b_i = h_N - h_i$, $i = 1, \cdots, N-1$. Any determinant of a minor of $[\mathscr{A}, \mathscr{B}]$ is a sum of products $*$ of δ_0 and δ_{b_i} with appropriate coefficients. It is not a divisor of zero if and only if it contains the term $\delta_0 * \delta_0 * \cdots * \delta_0$ with a nonzero coefficient. For nth order determinants this occurs if and only if the real matrix $[A_N, B_0]$ has rank n. □

For $N = 1$, this corollary contains the necessary condition (3.1) and, moreover, says that (3.1) along with spectral controllability is also *sufficient.* This enables us to generalize [16, Prop. 7.6] to arbitrary n.

COROLLARY 10. *For system* (1.2) *with* $m = 1$ *and arbitrary n approximate controllability holds if and only if*

$$\det P(\lambda) \not\equiv 0 \quad \textit{and} \quad P(\lambda)v(e^{-\lambda h}) \neq 0 \quad \forall \lambda \in C \tag{4.13}$$

where $P(\lambda)$ *is defined as in* [16], *and* $v(z) = [1, z, \cdots, z^{n-1}]^T$.

Proof. The necessity is proved in [16]. By [16, Thm. 3.6] $\det P(\lambda) \not\equiv 0$, implies $\operatorname{rank}[A_1, B_0] = n$, while $P(\lambda)v(e^{-\lambda h}) \neq 0$ for all $\lambda \in C$ is equivalent to spectral controllability [16, Prop. 7.1]. Hence, by Corollary 9, (4.13) implies approximate controllability. □

This result gives the easiest way to verify the approximate controllability in the case of $m = 1$ and one delay (see [16] for further discussion). An analogue of this result for the exact controllability in $W_2^{(1)}$ of neutral systems has been proved in [21].

Examples and further comments on the applicability of Theorem 8 are given in § 6.

5. Proof of Theorem 8. Before proceeding to the main part of the proof, we establish two technical results concerning rectangular matrices over J. These two results depend on the properties of determinants over a commutative ring with identity, and on Corollaries 3 and 4.

Let M be a rectangular $k \times l$ matrix of elements m_{ij} of some commutative ring $\mathbb{R}$ with identity. Let 0 be the zero element of $\mathbb{R}$ and let $\cdot$ denote multiplication in $\mathbb{R}$. [17, Thm. 51] says that the system of linear equations over $\mathbb{R}$

$$\sum_{j=1}^{l} m_{ij} \cdot x_j = 0, \qquad i = 1, \cdots, k, \tag{5.1}$$

has a nonzero solution if and only if $\operatorname{rank}_{\mathbb{R}} M < l$. It follows [4, Chapt. III, § 8] that the set of columns of M is linearly independent (as a subset of the module $\mathbb{R}^k$) if and only if $\operatorname{rank}_{\mathbb{R}} M = l$. This statement about the columns is nontrivial if $k \geqq l$. If $k < l$, by considering the transposed matrix M^T, one has that the set of rows of M is linearly independent (as a subset of $\mathbb{R}^l$) if and only if $\operatorname{rank}_{\mathbb{R}} M = k$. In general, if $k < l$ the linear independence of k rows of M *does not* imply that among the l columns of M there are k linearly independent ones, although the implication does hold if $\mathbb{R}$ is a field. An example of a ring in which the implication does not hold, is the ring of diagonal matrices, see § 7.2.

PROPOSITION 11. *Let M be a $k \times l$ ($k \leqq l$) matrix over the ring J. Then*

$$\begin{aligned} \operatorname{rank}_J M &= \textit{number of linearly independent rows} \\ &= \textit{number of linearly independent columns.} \end{aligned} \tag{5.2}$$

Proof. Let $\operatorname{rank}_J M = r$. By the definition of rank, the set of all determinants D_i^r of $r \times r$ minors of M, does not have a nonzero annihilator (but all the determinants D_i^{r+1} of $(r+1) \times (r+1)$ minors of M do have a common nonzero annihilator). Hence, by Corollary 4, at least one of the D_i^r, say D_1^r is not a divisor of zero. The r rows as well as the r columns of D_1^r are linearly independent (see [4, Chapt III, § 8 or Chapt. IV, § 2]). Enlarging the length of these rows or columns, by adding the remaining elements of corresponding rows or columns of M, respectively, does not destroy the linear independence. At the same time, by [17, Thm. 51] no set of $r+1$ rows or columns of M can be linearly independent, because all the D_i^{r+1} have a common nonzero annihilator. □

LEMMA 12. *Let $\mathbb{R}$ be a commutative ring with a unit $1 \neq 0$. Suppose that $\mathbb{R}$ has divisors of zero, and, moreover, a sum of any two divisors of zero in $\mathbb{R}$ is a divisor of zero in $\mathbb{R}$. Let $\mathcal{A}$, $\mathcal{B}$ be n, $n \times m$ matrices over $\mathbf{R}$, respectively. Suppose that the $n \times (n+m)$ matrix $[\mathcal{A}, \mathcal{B}]$ has n linearly independent columns. Then there is an $m \times n$ matrix $\mathcal{K}$ over $\mathbb{R}$, whose entries are either 0's or 1's, such that $\operatorname{rank}_{\mathbb{R}}[\mathcal{A} + \mathcal{B} \cdot \mathcal{K}] = n$.*

Proof. The idea is to add some columns of $\mathcal{B}$ to appropriate columns of $\mathcal{A}$ to obtain an $n \times n$ matrix with n linearly independent columns. Let $a_1, \cdots, a_n$, $b_1, \cdots, b_m$, denote the columns of $\mathcal{A}$ and $\mathcal{B}$, respectively. Suppose that $\mathcal{A}$ has $s < n$ linearly

independent columns, but any set of $s+1$ columns of $\mathscr{A}$ is linearly dependent (the case $s=n$ is trivial); so without loss of generality we can suppose that the linearly independent columns of $\mathscr{A}$ are $a_1, \cdots, a_s$. In fact, by the properties of determinants over a ring $\mathbb{R}$ [4], [10], the rank of $\mathscr{A}+\mathscr{B}\cdot\mathscr{K}$ is invariant with respect to the ordering of the columns; we can simultaneously reorder the columns of both $\mathscr{A}$ and $\mathscr{B}\cdot\mathscr{K}$; changing the order of columns of $\mathscr{B}\cdot\mathscr{K}$ amounts to changing the order of columns of $\mathscr{K}$ alone. Also, without loss of generality we can suppose that the columns of $\mathscr{B}$ completing $\{a_1, \cdots, a_s\}$ to a linearly independent set are $\{b_1, \cdots, b_{n-s}\}$; otherwise, we can always reorder the columns of B with a simultaneous corresponding reordering of the rows of $\mathscr{K}$, keeping the product $\mathscr{B}\cdot\mathscr{K}$ unchanged. Hence we assume that $\{a_1, \cdots, a_s, b_1, \cdots, b_{n-s}\}$ is a linearly independent set of n columns of $[\mathscr{A}, \mathscr{B}]$.

We now form an $n\times n$ matrix $\mathscr{C}$ with columns c_i given by

$$c_i = a_i, \qquad i=1, \cdots, s,$$
$$c_i = a_i + b_{i-s} \qquad i=s+1, \cdots, n.$$

A simple verification, via the matrix multiplication rule, yields that $\mathscr{C}=\mathscr{A}+\mathscr{B}\cdot\mathscr{K}$, where

$$\mathscr{K}=\left[\begin{array}{ccccccc} 0 & \cdots & 0 & 1 & & & \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \\ \underbrace{0 \quad \cdots \quad 0}_{s} & & & \underbrace{0 \quad \cdots \quad 0}_{n-s} & & & \end{array}\right] \Bigg\} m, \tag{5.3}$$

where $0=0_{\mathbb{R}}$, 1 is the identity in $\mathbb{R}$.

Consider the determinant as a function of its columns, $\mathscr{D}: \mathbb{R}^n \times \cdots \times \mathbb{R}^n \to \mathbb{R}$. Then, since $\mathscr{D}$ is multilinear, we have

$$\begin{aligned} \det \mathscr{C} &= \mathscr{D}(a_1, \cdots, a_s, a_{s+1}+b_1, \cdots, a_n+b_{n-s}) \\ &= \mathscr{D}(a_1, \cdots, a_s, a_{s+1}, a_{s+2}, \cdots, a_n) \\ &\quad + \mathscr{D}(a_1, \cdots, a_s, b_1, a_{s+2}, \cdots, a_n) \\ &\quad + \mathscr{D}(a_1, \cdots, a_s, a_{s+1}, b_2, a_{s+3}, a_n) \\ &\quad + \mathscr{D}(a_1, \cdots, a_s, b_1, b_2, a_{s+3}, \cdots, a_n) \\ &\quad \vdots \\ &\quad + \mathscr{D}(a_1, \cdots, a_s, b_1, \cdots, b_{n-s}). \end{aligned} \tag{5.4}$$

Since $\mathscr{A}$ has at most $s(<n)$ linearly independent columns, by [17, Thm. 51], all the determinants on the right-hand side except the last one are divisors of zero in $\mathbb{R}$. Let $\mathscr{S}$ denote the sum of all determinants except the last one. It follows from the assumption, that $\mathscr{S}$ is a divisor of zero. But then $\det \mathscr{C}$ cannot be a divisor of zero, because otherwise $\mathscr{D}(a_1, \cdots, a_s, b_1, \cdots, b_{n-s}) = \det \mathscr{C} - \mathscr{S}$ would also be a divisor of zero, a contradiction. Hence $\operatorname{rank}_{\mathbb{R}} \mathscr{C} = n$. □

Proof of Theorem 8.

a) *Necessity.* By Proposition 7, approximate controllability implies,

$$\begin{bmatrix} \mathscr{A}^T \\ \mathscr{B}^T \end{bmatrix} * u = 0 \Rightarrow u = 0.$$

This means that the n columns of the matrix $[\mathscr{A}, \mathscr{B}]^T$, that is the n rows of $[\mathscr{A}, \mathscr{B}]$, are linearly independent (as elements of the J module J^{n+m}). Hence rank $[\mathscr{A}, \mathscr{B}] = n$.

b) *Sufficiency.* Condition (ii) implies that the n rows of $[\mathscr{A}, \mathscr{B}]$ are linearly independent. By Proposition 11, the matrix $[\mathscr{A}, \mathscr{B}]$ has n linearly independent columns. Since, by Corollary 3, the ring J and the matrix $[\mathscr{A}, \mathscr{B}]$ satisfy the assumptions of Lemma 12, we have that there exists an $m \times n$ matrix $\mathscr{K}$ over J whose elements are either 0 or δ_0 such that

$$\operatorname{rank}_J [\mathscr{A} + \mathscr{B} * \mathscr{K}] = n. \tag{5.6}$$

Actually, the matrix $\mathscr{B}$ is equal to $B_0\delta_0$, while $\mathscr{K}$ can be written as $\mathscr{K} = K\delta_0$, where K is an $m \times n$ matrix whose entries are real numbers, either 0 or 1. Then $\mathscr{B} * \mathscr{K} = B_0\delta_0 * K\delta_0 = B_0K\delta_0$.

Condition (5.6) allows us now to introduce into the original system (1.1), feedback which transforms the system into an *equivalent control system with a complete set of generalized eigenfunctions.* Let

$$u(t) = Kz(t-h) + v(t), \tag{5.7}$$

where $v(\cdot)$ is another control function. By substituting (5.6) into (1.1) we have

$$\dot{z}(t) = \int_{-h}^{0} d\eta(\theta) z(t+\theta) + B_0Kz(t-h) + B_0v(t), \tag{5.8}$$

the transformation being, of course, reversible. For every $u(\cdot) \in L_1$ there exists a $v(\cdot) \in L_1$ such that the solution of (1.1) corresponding to $u(\cdot)$ is also a solution to (5.8) corresponding to $v(\cdot)$. The operator $\hat{H}$ defined by (4.1) corresponding to (5.8) is now

$$(\hat{H}\psi)(t) = \int_0^t d\hat{\eta}(\tau)\psi(t-\tau) + B_0K\psi(t), \qquad t \in [0, h].$$

Now, this new $\hat{H}$ induces the following matrix $\hat{\mathscr{A}}$ over J:

$$\hat{\mathscr{A}} = A_N\delta_0 + B_0K\delta_0 + \sum_{i=1}^{N-1} A_i\delta_{b_i} + \hat{E} = \mathscr{A} + \mathscr{B} * \mathscr{K}.$$

By (5.6), $\det_J \hat{\mathscr{A}}$ is not a divisor of zero; hence, by Theorem 5, generalized eigenfunctions associated with the modified system are complete in $\mathscr{X}$.

The characteristic matrix of the modified system is

$$\hat{\Delta}(\lambda) = \Delta(\lambda) - B_0Ke^{-\lambda h}.$$

This implies that rank $[\hat{\Delta}(\lambda), B_0] = n$ for all $\lambda \in C$; for otherwise, there would exist a $\lambda_0 \in C$ and a nonzero vector $y \in C^n$ such that

$$y^T[\Delta(\lambda_0) - B_0K e^{-\lambda_0 h}] = 0 \quad \text{and} \quad y^TB_0 = 0;$$

hence $y^TB_0K e^{-\lambda_0 h} = 0$ and $y^T\Delta(\lambda_0) = 0$, yielding rank $[\Delta(\lambda_0), B_0] < n$, contrary to hypothesis (i) of Theorem 8.

Hence, the modified system (5.8) is spectrally controllable. This, by a result of Fattorini [7, Prop. 3.1, Cor. 3.2] gives for the modified system

$$\bar{\mathscr{M}}^0 \subset \bar{K}_\infty^0,$$

where $\mathscr{M}^0$ and K_∞^0 denote the linear space spanned by the generalized eigenfunctions, and the attainable set, respectively, of the modified system. By completeness $\bar{\mathscr{M}}^0 = \mathscr{X}$, hence $\bar{K}_\infty^0 = \mathscr{X}$. But the attainable sets of (1.1) and (5.8) coincide. Hence $\bar{K}_\infty = \mathscr{X}$. □

6. Comments and examples. As the proof of Theorem 8 shows, the following statement is true.

COROLLARY 13. *System* (1.1) *is approximately controllable in* $\mathscr{X}$ *if and only if*: (i) *it is spectrally controllable, and* (ii) *there exists feedback* (*given in fact by* (5.7)) *such that the closed loop system has a complete set of generalized eigenfunctions.*

We note that spectral controllability is invariant under feedback, while completeness is not. Therefore, the class of approximately controllable systems is feedback equivalent (via feedback (5.7)) to spectrally controllable systems with a complete set of generalized eigenfunctions. We note that the feedback operator is not bounded when its domain is taken as the whole space $\mathscr{X}$.

Verification of conditions (i), (ii) of Theorem 8 might appear to be difficult, but actually it is not, at least on reasonable examples. At first glance (i) might require the computation of all the eigenvalues. This can actually be avoided by the following device. Condition (ii) (i.e., (2.6)) is for $m = 1$, equivalent [16, Thm. 7.2] to

$$[\operatorname{adj} \Delta(\lambda)]B_0 \neq 0 \quad \forall \lambda \in C, \tag{6.1}$$

where adj means the matrix adjoint. The implication (6.1) $\Rightarrow$ (2.6) is true for any $m \geqq 1$. Verifying (6.1) amounts to checking that a system of n transcendental equations in λ does not have a solution.

Example 1.

$$\frac{d}{dt}\begin{bmatrix} z_1(t) \\ z_2(t) \end{bmatrix} = \int_{-1}^{0} \begin{bmatrix} 1 & h+\theta \\ h+\theta & (h+\theta)^2/2 \end{bmatrix} \begin{bmatrix} z_1(t+\theta) \\ z_2(t+\theta) \end{bmatrix} d\theta + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t).$$

As pointed out in [2], the homogeneous part of this system fails to satisfy the completeness criterion. We check the conditions of Theorem 8.

(i)

$$\Delta(\lambda) = \begin{bmatrix} \lambda - \dfrac{1-e^{-\lambda}}{\lambda} & -\dfrac{\lambda - 1 + e^{-\lambda}}{\lambda^2} \\ -\dfrac{\lambda - 1 + e^{-\lambda}}{\lambda^2} & \lambda - \dfrac{\lambda^2 - 2\lambda + 2 - 2e^{-\lambda}}{2\lambda^3} \end{bmatrix}.$$

To avoid solving $\det \Delta(\lambda) = 0$, we check (6.1). Suppose

$$\operatorname{adj} \Delta(\lambda) B_0 = \begin{bmatrix} \dfrac{\lambda - 1 + e^{-\lambda}}{\lambda^2} \\ \lambda - \dfrac{1-e^{-\lambda}}{\lambda} \end{bmatrix} = 0$$

for some λ. From the first row we have $e^{-\lambda} = 1 - \lambda$. Substituting this into the second row we have $\lambda - ((1 - 1 + \lambda)/\lambda) = \lambda - 1 = 0$, hence λ must be 1, but then $e^{-1} \neq 0$. Consequently (6.1) is true for all $\lambda \in C$, and spectral controllability holds.

(ii) $\hat{H}$ given by (4.1) is represented by $A_i = 0$, $i = 1, \cdots, N$ and

$$\hat{E}(\tau) = \begin{bmatrix} 1 & \tau \\ \tau & \tau^2/2 \end{bmatrix}, \qquad \tau \in [0, h].$$

Hence, by putting $f_0(\tau) \equiv 1$, $f_1(\tau) = \tau$, $f_2(\tau) = \tau^2/2$ we have

$$[\mathscr{A}, \mathscr{B}] = \begin{bmatrix} f_0 & f_1 & 0 \\ f_1 & f_2 & \delta_0 \end{bmatrix},$$

$\det_J \mathscr{A} = f_0 * f_2 - f_1 * f_1 = 0(\int_0^t (\tau^2/2)\, d\tau - \int_0^t (t-\tau)\tau\, d\tau \equiv 0$ on $[0, 1])$, hence the eigenfunctions are not complete. However,

$$\det_J \begin{bmatrix} f_0 & 0 \\ f_1 & \delta_0 \end{bmatrix} = f_0 * \delta_0 = f_0.$$

Since f_0 is not a divisor of zero, $\operatorname{rank}_J [\mathscr{A}, \mathscr{B}] = 2$. Condition (ii) is satisfied. Actually, the feedback $u(t) = z_1(t-1) + v(t)$ causes the closed loop system to have a complete set of generalized eigenfunctions, because

$$\det_J \begin{bmatrix} f_0 & f_1 \\ f_1 - \delta_0 & f_2 \end{bmatrix} = \delta_0 * f_1 \quad \text{is not a divisor of zero.}$$

The system is approximately controllable. □

Let us now consider differential-difference systems (4.12) with commensurate delays $h_i = id$, $i = 0, 1, \cdots, N$, $d \in (0, \infty)$, $h = Nd$. We briefly sketch how to verify the spectral controllability condition (6.1); the other condition (Corollary 9 (ii)) is easy to check. Suppose $m = 1$ and $P(\lambda)$ is a polynomial matrix defined by the identity $P(\lambda)v(e^{-\lambda d}) = [\operatorname{adj} \Delta(\lambda)]B_0$, where $v(\mu) = (1, \mu, \mu^2, \cdots, \mu^{(n-1)N})^T$. $P(\lambda)$ has dimension $n \times r$, $r = (n-1)N + 1$. By (6.1), system (4.12) with delays $h_i = id$ is spectrally controllable if and only if

$$P(\lambda)v(e^{-\lambda d}) \neq 0 \quad \forall \lambda \in C. \tag{6.2}$$

This is analogous to condition (4.13) except that now $P(\lambda)$ is rectangular, $n \leqq r$. Define an augmented matrix $\tilde{P}(\lambda)$ at dimension $nN \times nN$

$$\tilde{P}(\lambda) = \begin{bmatrix} P(\lambda) & 0 & \cdots & \cdots & 0 \\ 0 & P(\lambda) & 0 & \cdots & 0 \\ 0 & 0 & P(\lambda) & 0 & 0 \\ & & & \ddots & \\ 0 & & & 0 & P(\lambda) \end{bmatrix}.$$

PROPOSITION 14. *A necessary condition of approximate controllability of system* (4.12) *with commensurate delays* $h_i = id$, $i = 1, \cdots, N$ *is* $\det \tilde{P}(\lambda) \not\equiv 0$. *If this condition holds, spectral controllability of* (4.12) *is equivalent to* (6.2) *being satisfied for all* λ *such that* $\det \tilde{P}(\lambda) = 0$.

The proof of this proposition is an extension of the proof of [16, Thm. 3.1] and will be omitted.

Example 2. Consider a two-dimensional system with four delays, $n = 2$, $N = 4$. Let $d = 1$, $A_0 = 0$, and

$$A_1 = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix}, \quad A_3 = \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix}, \quad A_4 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad B_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

We have

$$\Delta(\lambda) = \begin{bmatrix} \lambda + e^{-3\lambda} & -e^{-4\lambda} \\ -e^{-\lambda} & \lambda + e^{-2\lambda} \end{bmatrix},$$

$\det \Delta(\lambda) = (\lambda + e^{-3\lambda})(\lambda + e^{-2\lambda}) - e^{-5\lambda}$.

The spectrum is infinite and a direct verification of spectral controllability via (2.6) seems hopeless. However, by using the method suggested above, we have

$$P(\lambda) = \begin{bmatrix} \lambda & 0 & 1 & 0 & 1 \\ \lambda & 1 & 0 & 1 & 0 \end{bmatrix},$$

$\tilde{P}(\lambda)$ is an 8×8 matrix with $\det \tilde{P}(\lambda)=\lambda^3(\lambda+2)$. Hence we verify that

$$P(\lambda)v(e^{-\lambda})=\begin{bmatrix}\lambda+e^{-2\lambda}+e^{-4\lambda}\\ \lambda+e^{-\lambda}+e^{-3\lambda}\end{bmatrix}\neq 0$$

for $\lambda=0$ and $\lambda=-2$. Since the condition is satisfied, the system is spectrally controllable. We also have rank $[A_4, B_0]=2$. Hence the system is approximately controllable.

Example 3. Consider the system:

$$\dot{z}_1(t)=\int_{-2}^{0} z_2(t+\theta)\,d\theta,$$

$$\dot{z}_2(t)=\int_{-2}^{0} z_1(t+\theta)\,d\theta+z_3(t-2),$$

$$\dot{z}_3(t)=\int_{-1}^{0} z_1(t+\theta)\,d\theta+z_2(t-2)+z_3(t-1)+u(t).$$

We have

$$[\operatorname{adj}\Delta(\lambda)]B_0=\left[e^{-2\lambda}\frac{1-e^{-2\lambda}}{\lambda},\ \lambda e^{-2\lambda},\ \lambda^2-\left(\frac{1-e^{-2\lambda}}{\lambda}\right)^2\right]^T,$$

so that (6.1) is satisfied. Now,

$$[\mathscr{A},\mathscr{B}]=\begin{bmatrix}0 & e & 0 & 0\\ e & 0 & \delta_0 & 0\\ f & \delta_0 & \delta_1 & \delta_0\end{bmatrix},$$

where $e(t)\equiv 1$, $t\in[0,2]$, $f(t)=\chi_{[1,2]}(t)$, $t\in[0,2]$, $\det\mathscr{A}=e*f*\delta_0-e*e*\delta_1$. $\det\mathscr{A}$ is a divisor of zero because both f and δ_1 are divisors of zero. However, condition (ii) of Theorem 8 is satisfied, because the last three columns of $[\mathscr{A},\mathscr{B}]$ are linearly independent. The system is approximately controllable.

As an example of a system which satisfies the spectral controllability condition but does not satisfy condition (ii) of Theorem 8, one can take a scalar nth order differential-difference equation

$$z^{(n)}+\sum_{i=1}^{n-1}\sum_{j=1}^{N} a_{ij}z^{(i)}(t-h_j)+u(t)=0.$$

rewritten as a system of n first order equations.

When a system is spectrally controllable but does not satisfy condition (ii) of Theorem 8, it may still be approximately controllable in some weaker sense; for instance it may be F-controllable [15]. One can show that most of the ideas presented in this paper carry over the the F-controllability case [27].

7. Appendix. The material below is intended to make the paper self-contained.

7.1. Proof of Theorem 5. By [14, Thm. 5.1], (i)$\Leftrightarrow$(ii). The equivalence of (ii) and (iii), (iv) is given by [2, Thm. 1] the proof of which is adapted here with some modifications and clarifications.

The set of all $L_2(0,h)$ functions is a proper ideal I in J (by [9, 21.32] and the fact that $\delta_a * w\in I$ if $w\in I$). Let I^n denote the corresponding proper subset of J^n. By (4.1),

$\hat{H}^*$ maps I^n into itself, and

$$(\hat{H}^*\psi)(t) = A_N^T\psi(t) + \sum_{i=1}^{N-1} A_i^T\psi(t-b_i)\chi_{[b_i,h]} + \int_0^t \hat{E}^T(\tau)\psi(t-\tau)\,d\tau$$

$$= \left[\left(A_N^T\delta_0 + \sum_{i=1}^{N-1} A_i^T\,\delta_{b_i} + \hat{E}^T\right) * \psi\right](t), \qquad t\in[0,h].$$

Hence

$$\hat{H}^*\psi = \mathscr{A}^T * \psi \quad \text{for } \psi \in I^n.$$

But Ker $H^* = \{0\} \Leftrightarrow$ Ker $\hat{H}^* = \{0\}$, and the latter is equivalent to

$$\forall \psi \in I^n, \quad \mathscr{A}^T * \psi = 0 \Rightarrow \psi = 0. \tag{7.1}$$

Remark 3. The restriction $\psi \in I^n$ can be removed. In fact, suppose that (7.1) holds, but there exists $u \neq 0$, $u \in J^n \backslash I^n$, such that $\mathscr{A}^T * u = 0$. By taking $w \in I$ which is not a divisor of zero (e.g., $w(t) \equiv 1\ t \in [0, h]$), and taking a diagonal $n \times n$ matrix $\Omega =$ diag $(w, \cdots, w)$, we have $0 = \mathscr{A}^T * u = \Omega * \mathscr{A}^T * u = \mathscr{A}^T * \Omega * u = \mathscr{A}^T * \psi$, where $\psi \in I^n$ because $\psi_i = w * u_i$ and I is an ideal. Now $0 = \mathscr{A}^T * \psi$ implies $\psi = 0$, hence $w * u_i = 0$ $i = 1, \cdots, n$, hence $u_i = 0$, $i = 1, \cdots, n$.

Now (7.1) and Remark 3 prove that Ker $\hat{H}^* = \{0\} \Leftrightarrow$ Ker $\mathscr{A}^T = \{0\}$. This proves the equivalence of (ii) and (iii).

$\mathscr{A}^T$ represents a J^n module endomorphism. By [4, Chapt. III, § 8, Prop. 3, p. 524] or [17, Thm. 51] $\mathscr{A}^T$ is injective if and only if $\det_J \mathscr{A}^T$ is not a divisor of zero. □

7.2. A comment on Proposition 11. Relation (5.2) depends crucially on the property of the ring J expressed by Corollary 4. With another ring, if Corollary 4 does not hold, (5.2) may be false. As an example take the ring DM_3 of 3×3 real diagonal matrices, with standard addition and multiplication operations. Since diagonal matrices commute, the ring is commutative. The matrices can be represented by triples of their diagonal elements. Hence $x \in DM_3$ has a representation $\{\xi_1, \xi_2, \xi_3\}$, $\xi_i \in R$. Now

$$x + y = \{\xi_1, \xi_2, \xi_3\} + \{\eta_1, \eta_2, \eta_3\} = \{\xi_1 + \eta_1, \xi_2 + \eta_2, \xi_3 + \eta_3\},$$

$$x \cdot y = \{\xi_1\eta_1, \xi_2\eta_2, \xi_3\eta_3\}.$$

Obviously, x is a divisor of zero iff one of the ξ_i is zero, $i = 1, 2, 3$. Corollaries 3 and 4 do not hold in this ring, as the following counterexamples show:

a) $\{1, 0, 1\} + \{0, 1, 0\} = \{1, 1, 1\}$.

b) The set $\{0, 1, 1\}$, $\{1, 0, 1\}$, $\{1, 1, 0\}$ does not have a nonzero annihilator even though all its members are divisors of zero. Now take a 2×3 matrix over DM_3 given by

$$M = \begin{bmatrix} \{1,1,1\} & \{1,2,0\} & \{0,1,1\} \\ \{1,0,1\} & \{1,2,3\} & \{1,1,1\} \end{bmatrix}.$$

The three 2×2 determinants over DM_3 taken out of this matrix are $\{0, 2, 3\}, \{1, 0, -3\}$, $\{1, 1, 0\}$. They do not have a common nonzero annihilator. By [17, Thm. 51], the rows of M are linearly independent (this can also be checked by elementary calculations, taking $\alpha, \beta \in DM_3$ as "coefficients" multiplying the rows, and proving that if a linear combination of rows is null, then $\alpha = \beta = \{0, 0, 0\}$). However, there is no set of 2 linearly independent columns (over DM_3) of this matrix, because every 2×2 determinant is a divisor of zero. (For instance, the first two columns of M are linearly dependent with coefficients $\alpha = \{1, 0, 0\}$, $\beta = \{-1, 0, 0\}$.)

7.3. A simple proof of condition (i) of Theorem 8. We show that the condition

$$\text{Im}\,(I\lambda - A) + \text{Im}\,B = \mathscr{X} \quad \forall \lambda \in \sigma(A), \tag{7.2}$$

is necessary for approximate controllability. Suppose it does not hold. Since the left-hand side is a closed subspace of $\mathscr{X}$, there exist $\lambda \in \sigma(A)$, $x \in \mathscr{X}$, $x \neq 0$ such that $\langle x, (I\lambda - A)\xi\rangle = 0$ for all $\xi \in \mathscr{D}(A)$ and $\langle x, Bu\rangle = 0$ for all $u \in R^m$, where $\langle \cdot, \cdot \rangle$ is the scalar product in $\mathscr{X}$. Hence $(I\bar{\lambda} - A^*)x = 0$ and $B^*x = 0$, $\bar{\lambda}$ being a complex conjugate of λ. Take any $\mu \in \rho(A) = \rho(A^*)$ (the resolvent set of A). Hence $\mu \neq \lambda$, because $\bar{\lambda}$ is in $\sigma(A^*)$ and $\sigma(A^*)$ is symmetric with respect to the real axis. Take $y = (\bar{\mu} - \bar{\lambda})^{-1}x$. But $x = (\bar{\mu} - \bar{\lambda})^{-1}(\bar{\mu} - \bar{\lambda})x = (\bar{\mu} - \bar{\lambda})^{-1}(I\bar{\mu} - A^*)x = (I\bar{\mu} - A^*)y$. Now $B^*x = 0$, implies $0 = B^*y = B^*(I\bar{\mu} - A^*)^{-1}x$, or $\langle x, (I\mu - A)^{-1}Bu\rangle = 0$ for all $u \in R^m$, $\mu \in \rho(A)$. By [16, Thm. 2.1] the system is not approximately controllable.

The equivalence between (7.2) and (2.6) (as pointed out by Bhat and Wonham [26]) can be obtained via the following calculations. (7.2) holds if and only if for all $y \in \mathscr{X}$, there exist $u \in R^m$ and $x \in \mathscr{D}(A)$ such that $(I\lambda - A)x + Bu = y$. Substituting the expressions for A and B we obtain

$$\Delta(\lambda)x^0 + B_0u = y^0 - \int_{-h}^{0} d\eta(\theta) \int_0^{\theta} e^{h(\theta-\alpha)}y^1(\alpha)\,d\alpha.$$

The last statement holds for all $(y^0, y^1) \in \mathscr{X}$ if and only if rank $[\Delta(\lambda), B_0] = n$. □

REFERENCES

[1] H. T. BANKS AND J. A. BURNS, *Hereditary control problems: numerical methods based on averaging approximations*, this Journal, 16 (1978), pp. 169–208.

[2] Z. BARTOSIEWICZ, *Density of images of semigroup operators for linear neutral functional differential equations*, J. Differential Equations, to appear.

[3] C. BERNIER AND A. MANITIUS, *On semigroups in $R^n \times L_2$ corresponding to differential equations with delays*, Canad. J. Math. 30 (1978), pp. 897–914.

[4] N. BOURBAKI, *Elements of Mathematics, Algebra I*, Hermann–Addison Wesley, Reading, MA, 1974, Chapters 1–3.

[5] M. C. DELFOUR, *The largest class of hereditary systems defining a C_0-semigroup on the product space*, Canad. J. Math., to appear.

[6] M. C. DELFOUR AND A. MANITIUS, *The structural operator F and its role in the theory of retarded systems, I and II*, J. Math. Anal. Appl., 73 (1980), pp. 466–490, and 74 (1980), pp. 359–381.

[7] H. O. FATTORINI, *Some remarks on complete controllability*, SIAM J. Control, 4 (1966), pl. 686–694.

[8] I. GELFAND, D. RAIKOV AND G. SHILOV, *Commutative Normed Rings*, Chelsea, New York, 1964.

[9] E. HEWITT AND K. STROMBERG, *Real and Abstract Analysis*, Springer–Verlag, New York, 1965.

[10] T. W. HUNGERFORD, *Algebra*, Holt, Rinehart and Winston, New York, 1974.

[11] M. Q. JACOBS AND C. E. LANGENHOP, *Criteria for function space controllability of linear neutral systems*, this Journal, 14 (1976), pp. 1009–1048.

[12] B. JAKUBCZYK, *A classification of attainable sets of linear differential-difference systems*, Preprint # 134, Institute of Mathematics, Polish Academy of Sciences, Warsaw, 1978.

[13] E. KAMEN, *An operator theory of functional differential equations*, J. Differential Equations, 27 (1978), pp. 274–296.

[14] A. MANITIUS, *Completeness and F-completeness of eigenfunctions associated with retarded functional differential equations*, J. Differential Equations, 35 (1980), pp. 1–29.

[15] ———, *Controllability, observability and stabilizability of retarded systems*, Proc. 1976 IEEE Conference on Decision and Control, IEEE Publications, New York, 1976, pp. 752–758.

[16] A. MANITIUS AND R. TRIGGIANI, *Function space controllability of linear retarded systems: a derivation from abstract operator conditions*, this Journal, 16 (1978), pp. 599–645.

[17] N. H. MCCOY, *Rings and Ideals*, Carus Mathematical Monographs, no. 8, Mathematical Association of America, 1948.

[18] J. G. MIKUSINSKI, *A new proof of Titchmarsh's theorem on convolution*, Studia Math., 13 (1953), pp. 56–58.

[19] D. A. O'CONNOR, *State Controllability and Observability for Linear Neutral Systems*, PhD Thesis, Washington University, St. Louis, MO, 1978.
[20] L. PANDOLFI, *On feedback stabilization of functional differential equations*, Bolletino U.M.I. 4, 11, Supplemento al fascicolo 3, Giugno 1975, Serie IV, vol. XI, pp. 626–635.
[21] H. R. RODAS AND C. E. LANGENHOP, *A sufficient condition for function space controllability of a linear neutral system*, this Journal, 16 (1978), pp. 429–435.
[22] W. RUDIN, *Functional Analysis*, McGraw-Hill, New York, 1973.
[23] D. L. RUSSELL, *Controllability and stability theory of linear partial differential equations: recent progress and open questions*, SIAM Rev., 20 (1978), pp. 639–739.
[24] L. SCHWARTZ, *Théorie des distributions*, Hermann, Paris, 1966.
[25] R. TRIGGIANI, *Extensions of rank conditions for controllability and observability to Banach spaces and unbounded operators*, this Journal, 14 (1976), pp. 313–338.
[26] K. P. M. BHAT AND W. M. WONHAM, *Stabilizability and detectability for evolution systems on Banach spaces*, Proc. 1976 IEEE Conference on Decision and Control, IEEE Publications, New York, 1976, pp. 1240–1243.
[27] A. MANITIUS, *F-controllability and observability of linear retarded systems*, report CRMA-921, Centre de Recherche de Mathématiques Appliquées, Université de Montréal, 1980.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 4, July 1981

0363-0129/81/1904-0008 $01.00/0

LOWER SEMICONTINUITY OF INTEGRAL FUNCTIONALS WITH NONCONVEX INTEGRANDS BY RELAXATION-COMPACTIFICATION*

E. J. BALDER†

Abstract. A new approach to the lower semicontinuity of integral functionals is presented. By a topological embedding of the "control" and "state" spaces in the Hilbert cube and a simultaneous relaxation of the "control functions," a powerful approach emerges whose main features include:

(i) A generalized convexity condition is imposed upon the integrand of which the classical convexity condition is a special case.

(ii) In the embedded setting the integrand can be supposed Lipschitz-continuous in "control" and "state" arguments without loss of generality.

(iii) Convergence in measure of the "trajectories," metamorphoses into L_1-norm convergence in the embedded setting.

1. Introduction. This paper will deal with the following problem: Let $(T, \mathcal{T}, \mu)$ be a finite measure space and X_1, X_2 metrizable Lusin spaces, equipped with given metrics d_1, d_2 respectively. Let f be a nonnegative normal integrand on $T \times (X_1 \times X_2)$; i.e., f is a functional from $T \times X_1 \times X_2$ into $[0, \infty]$, which is product-measurable and (jointly) lower semicontinuous in its second and third argument. Here the Lusin space $X_1 \times X_2$ is equipped with the product topology and—as will be our understanding for all other topological spaces to be met in the sequel—with its Borel σ-algebra. Given two collections $\mathcal{L}$ and $\mathcal{M}$, consisting of equivalence classes (with respect to equality μ-a.e.) of measurable functions from T into X_1 and X_2 respectively, and two sequences $\{y_k\}$, $\{u_k\}$ converging to y_0 and u_0 in $\mathcal{L}$ and $\mathcal{M}$, respectively, we address the lower semicontinuity question for the functional I from $\mathcal{L} \times \mathcal{M}$ into $[0, \infty]$ defined by

$$I(y, u) \equiv \int_T f(t, y(t), u(t))\mu(dt), \qquad y \in \mathcal{L}, \quad u \in \mathcal{M}.$$

That is to say, we shall address the problem of formulating sufficient conditions so that

$$\gamma \equiv \liminf_k I(y_k, u_k) \geqq I(y_0, u_0). \tag{1}$$

Here the convergence of the "trajectories" in $\mathcal{L}$ is of a strong type (to be specified later) whereas the convergence of the "control functions" in $\mathcal{M}$ is of a weak type.

The literature on the above lower semicontinuity question is quite extensive, cf. [1], [2], [3], [4], [5] and the references given there. For the purposes of this paper, it seems enough to mention here that the problem comes forth quite naturally from the existence question posed in the calculus of variations and optimal control theory (cf. [3], [6]). It should also be pointed out that all of the classical sufficient conditions for lower semicontinuity contain a convexity assumption about the integrand f (viz., $f(t, x, \cdot)$ is assumed convex for every $t \in T$, $x \in X_1$). That such a convexity assumption can be quite unnatural on occasion is clearly illustrated by the "simplest lower semicontinuity problem," where μ is taken to be a Dirac measure on T. It is the purpose of this paper to present a new, very general way of studying the lower semicontinuity problem for which no intrinsic convexity assumptions are needed. We cannot, however, abandon the convexity notion altogether, for the relaxation of the control functions will play a crucial role. In order to obtain a high level of generality, we shall use the technique introduced

* Received by the editors November 8, 1979, and in revised form August 15, 1980.
† Mathematical Institute, University of Utrecht, Utrecht, the Netherlands.

in [7], of "compactifying" the lower semicontinuity problem, by embedding the space $X_1 \times X_2$ in the Hilbert cube. The separate treatment of topological and geometrical aspects of the problem thus afforded leads also to great analytical efficiency in switching the limit and the integral sign: namely, by the embedding procedure and the associated change of metric, the "ugly ducklings" consisting of the normal integrand f and the convergence in measure of the sequence $\{y_k\}$ undergo a metamorphosis whereby:

(i) The normal integrand f can be supposed wlog Lipschitz-continuous on $X_1 \times X_2$ (in the new metric), with a Lipschitz constant not depending upon $t \in T$.

(ii) The convergence in measure of the sequence $\{y_k\}$ (with respect to the original metric) transforms into convergence in L_1-norm with respect to the new metric.

As can be gathered from the above, the switch of limit and integral sign has now become a fairly straightforward procedure. We obtain the inequality

$$\gamma \geqq I(y_0, \bar{\delta}).$$

Here $\bar{\delta}$ denotes a relaxed control function from T to X_2, which figures as the generalized limit of a subsequence of $\{u_k\}$. Next, the generalized convexity assumption will pave the way to the final conclusion. Major sources of inspiration for the relaxation–compactification approach introduced here have been the very interesting works by Berliocchi–Lasry, McShane and Warga [8], [9], [10]. Particularly in [8], where a one-point (i.e., Alexandrov) compactification has been applied to the control space, it has been made quite transparent that in the compactified set-up, the normal integrands of the original problem can be regarded as the supremum of a collection of very nice functionals on the extended product space.

2. Fundamentals. Let S be a metrizable Lusin space, equipped with a given metric d, and let $\{s_j\}$ denote a countable, dense subset of S. It is well known that the mapping ϕ from S to the Hilbert cube $\hat{S} \equiv [0, 1]^{\mathbb{N}}$, defined by

$$\phi(s) \equiv \left\{ \frac{d(s, s_j)}{1 + d(s, s_j)} \right\}_{j=1}^{\infty}, \qquad s \in S,$$

establishes a topological homeomorphism between S and its image $\phi(S)$, when we equip this image with the relative topology of pointwise convergence on $\hat{S}$. Moreover, since S is Lusin, $\phi(S)$ is known to be Borel measurable in $\hat{S}$ [11, III.20]. Now $\hat{S}$ is obviously compact and metrizable. In particular, we shall equip it from now on with the metric ρ defined by

$$\rho(s, s') \equiv \sum_{j=1}^{\infty} 2^{-j} |s_j - s'_j|, \qquad s, s' \in \hat{S}, \quad s = \{s_j\}, \quad s' = \{s'_j\}.$$

We may identify S with $\phi(S)$ for all topological purposes. Consequently, for every $s, s' \in S$

$$\rho(s, s') \leqq \min(d(s, s'), 1). \tag{2}$$

Let $(T, \mathcal{T}, \mu)$ be as introduced in § 1. The following proposition results trivially from (2).

PROPOSITION 1. *Let $\{z_k\}$ be a sequence of measurable functions from T into S, converging to another such function z_0 in measure, with respect to d (i.e., for every $\varepsilon > 0$, $\lim_k \mu\{t \in T : d(z_k(t), z_0(t)) > \varepsilon\} = 0$). Then $\{z_k\}$ converges to z_0 in L_1-norm with respect to ρ* (i.e., $\lim_k \int_T \rho(z_k(t), z_0(t)) \mu(dt) = 0$).

As we know from topology, once a space has been embedded, it usually pays to extend all relevant functionals from the original to the new setting. Naturally, in our

case these functionals are the normal integrands. In constructing such extensions a valuable property of the normal integrands comes to light.

The following result was proven in [7] for the case where $(T, \mathcal{T}, \mu)$ is a complete measure space and in [12] for the general case. For the sake of completeness, we include a proof here; it is slightly simpler than that given in [12].

LEMMA 2. *Let g be a nonnegative normal integrand on $T \times S$. Then there exists a nondecreasing sequence $\{\hat{g}_n\}$ of nonnegative normal integrands on $T \times \hat{S}$ and a μ-null set $N \in \mathcal{T}$ such that:*

$$\text{for every } t \notin N, \quad s \in S, \quad \lim_n \uparrow \hat{g}_n(t, s) = g(t, s); \tag{3}$$

$$\text{for every } t \in T, \quad s, s' \in \hat{S}, \quad n \in \mathbb{N}, \quad |\hat{g}_n(t, s) - \hat{g}_n(t, s')| \leqq n\rho(s, s'). \tag{4}$$

Proof. For $n \in \mathbb{N}$, $t \in T$, $s \in \hat{S}$ we first define

$$g_n(t, s) \equiv \inf \{n\rho(s, s') + g(t, s') : s' \in S\},$$

in the case where $g(t, \cdot)$ is not identically equal to $+\infty$ on S; otherwise, we set $g_n(t, \cdot) \equiv n$ and (3), (4) will hold trivially. Note that $\{g_n\}$ is nondecreasing and satisfies (4) by the triangle inequality. To show that (3) holds for $\{g_n\}$, it would suffice to invoke [13, Thm. 1], since for every $s \in S$, the family $\{-n\rho(s, \cdot) : n \in \mathbb{N}\}$ is of needle type and g is bounded below. However, an ad hoc proof is also possible here. Given $s \in S$, $t \in T$, consider the case where $g(t, s) < +\infty$. Given $\varepsilon > 0$ (arbitrary), there exists $\eta > 0$ such that for every $s' \in S$, $\rho(s, s') < \eta$, implies $g(t, s') \geqq g(t, s) - \varepsilon$. Clearly, for large enough $n \in \mathbb{N}$, $g_n(t, s)$ is equal to the infimum of $n\rho(s, s') + g(t, s')$ over all $s' \in S$, such that $\rho(s, s') < \eta$ and so not smaller than $g(t, s) - \varepsilon$ (note that here the boundedness below of g is essential). In case $g(t, s) = +\infty$, a simple modification of the above argument will do. Thus far, we cannot assert that the g_n are normal integrands, since we know only that $g_n(\cdot, s)$ is universally measurable for every $n \in \mathbb{N}$, $s \in \hat{S}$. (This follows from the fact that for every $\beta \in \mathbb{R}$, the set $\{t \in T : g_n(t, s) < \beta\}$ is the projection onto T of the product-measurable set $\{(t, s') \in T \times S : n\rho(s, s') + g(t, s') < \beta\}$ in $T \times \hat{S}$ [14, III.23].) We shall now modify g_n. Let $\{\hat{s}_j\}$ denote a countable, dense subset of $\hat{S}$. For every $n, j \in \mathbb{N}$ there exists a μ-null set $N_{n,j} \in \mathcal{T}$ and a $\mathcal{T}$-measurable functional $g_{n,j}$ on T such that $g_{n,j}(t) = g_n(t, \hat{s}_j)$ for all $t \notin N_{n,j}$ [11, III.32]. Set $N \equiv \cup_{n,j} N_{n,j}$ and define for $t \notin N$, $s \in S$, $n \in \mathbb{N}$

$$\hat{g}_n(t, s) \equiv \lim_{j'} g_{n,j'}(t),$$

where $\{\hat{s}_{j'}\}$ is an arbitrary subsequence of $\{\hat{s}_j\}$ converging to s. It is easy to see that this definition does not depend upon the particular subsequence taken (since the g_n have property (4)). Also, for $t \in N$, $n \in \mathbb{N}$, we shall set $\hat{g}_n(t, \cdot) \equiv 0$. By construction, $\hat{g}_n(\cdot, s)$ is $\mathcal{T}$-measurable for every $s \in \hat{S}$, $n \in \mathbb{N}$ and one checks easily that (3), (4) also hold for the $\hat{g}_n$. Finally, since $\hat{g}_n(t, \cdot)$ is continuous for every $t \in T$, $n \in \mathbb{N}$, $\hat{g}_n$ is product-measurable [14, III.14].

The fundamental results of this section will see useful service in § 3. Proposition 1, for instance, will be applied to $\{y_k\}$, i.e., with $S = X_1$. Lemma 2 will be used for the situation where $g = f$, i.e., with $S = X_1 \times X_2$. On the other hand, the following result will be applied to the case where $S = X_2$.

Let $M(\hat{S})$ denote the set of Radon measures on $\hat{S}$ equipped with the vague topology, and let $M_1^+(\hat{S})$ stand for the set of probability measures in $M(\hat{S})$, and $M_1^+(S)$ for the set of those elements of $M_1^+(\hat{S})$ that are carried by the set S; cf. [11, III.58]. Further, let $\mathscr{C}(\hat{S})$ denote the collection of continuous functionals on $\hat{S}$ equipped with

the supremum norm topology. For the sake of notational convenience, we shall denote, for any measurable functional e on $\hat{S}$ and any measure $\nu \in M_1^+(\hat{S})$, the integral of e with respect to ν over the set $\hat{S}$, by $e(\nu)$. It is well known that the set $L_\infty(T; M(\hat{S}))$ is the dual of the space $L_1(T; \mathscr{C}(\hat{S}))$, the usual L_1-space of (equivalent classes of) measurable functions from $(T, \mathscr{T}, \mu)$ into $\mathscr{C}(\hat{S})$. We equip $L_\infty(T; M(\hat{S}))$ with the weak star topology and denote by $\hat{\mathscr{R}}$ and $\mathscr{R}$ the subsets of $L_\infty(T; M(\hat{S}))$ consisting of the $M_1^+(\hat{S})$-valued and $M_1^+(S)$-valued elements of $L_\infty(T; M(\hat{S}))$. The elements of $\hat{\mathscr{R}}(\mathscr{R})$ are known in control theory under the name "relaxed control functions" [8], [9], [10], [14], [15].

LEMMA 3. *The set $\hat{\mathscr{R}}$ is a compact subset of $L_\infty(T; M(\hat{S}))$. Moreover, if the σ-algebra $\mathscr{T}$ is countably generated or the completion of a countably generated σ-algebra, then $\hat{\mathscr{R}}$ is sequentially compact.*

Proof. Compactness is a well-known consequence of the Alaoglu–Bourbaki theorem and we shall not repeat the proof; cf., e.g., [14, V.2]. Sequential compactness follows from an application of [16, III.12F], since under the additional hypothesis the space $L_1(T; \mathscr{C}(\hat{S}))$ is separable.

LEMMA 4. *Let g and h be normal integrands on $T \times S$. Suppose that h is nonnegative and that for every $\varepsilon > 0$ there exists $f_\varepsilon \in L_1(T; \mathbb{R})$ such that on $T \times X_2$*

$$\max(-g, 0) \leqq \varepsilon h + f_\varepsilon.$$

Then the functional $\delta \mapsto \int_T g(t, \delta(t))\mu(dt)$ is well defined and lower semicontinuous on the subset $\mathscr{R}(h)$ of $L_\infty(T; M(\hat{S}))$, consisting of those $\delta \in \mathscr{R}$ satisfying

$$\int_T h(t, \delta(t))\mu(dt) \leqq 1.$$

Proof. Suppose first that g is also nonnegative. By Lemma 2, there exists a sequence $\{\hat{g}_n\}$ satisfying (3), (4). By (4), for every $n \in \mathbb{N}$ the bounded functional $\tilde{g}_n \equiv \min(\hat{g}_n, n)$ is a representation of an element in $L_1(T; \mathscr{C}(\hat{S}))$. By (3) and the monotone convergence theorem for every $\delta \in \mathscr{R}$

$$\int_T g(t, \delta(t))\mu(dt) = \lim_n \uparrow \int_T \tilde{g}_n(t, \delta(t))\mu(dt).$$

So lower semicontinuity has been proven in this case. In general, we are given that for every $\varepsilon > 0$ the functional $g_\varepsilon \equiv g + \varepsilon h + f_\varepsilon$ is a nonnegative normal integrand on $T \times S$. Therefore, by the above, the functional $\pi_\varepsilon : \delta \mapsto \int_T g_\varepsilon(t, \delta(t))\mu(dt)$ is l.s.c. on $\mathscr{R}$. It remains to convince ourselves that for every $\delta \in \mathscr{R}(h)$,

$$\int_T g(t, \delta(t))\mu(dt) = \sup\left\{\pi_\varepsilon(\delta) - \int_T f_\varepsilon \, d\mu - \varepsilon : \varepsilon > 0\right\}.$$

3. Main results. Consider the lower semicontinuity question for the integral functional I as formulated in § 1. Let us introduce the following assumptions on the nature of the convergence of $\{y_k\}$, $\{u_k\}$:

(A1) The sequence $\{y_k\}$ converges to y_0 in measure.

(A2) There exist a nonnegative normal integrand h on $T \times X_2$, inf-compact in its second argument, and an at most countable collection $\mathscr{A}$ of normal integrands on $T \times X_2$ such that:

(A2a) For every $k \in \mathbb{N} \cup \{0\}$ $\int_T h(t, u_k(t))\mu(dt) \leqq 1$.

(A2b) For every $a \in \mathscr{A}$, $\varepsilon > 0$ there exists $f_\varepsilon \in L_1(T; \mathbb{R})$ such that on $T \times X_2$ $\max(-a, 0) \leqq \varepsilon h + f_\varepsilon$.

(A2c) For every $a \in \mathcal{A}$, $B \in \mathcal{T}$,

$$\limsup_k \int_B a(t, u_k(t))\mu(dt) \leqq \int_B a(t, u_0(t))\mu(dt).$$

(A2d) For every $\eta > 0$ for μ-a.e. $t \in T$ and every $x \in X_2$,

$$f(t, y_0(t), x) = \sup\{-a(t, x): a \in \mathcal{A}, -a(t, \cdot) \leqq f(t, y_0(t), \cdot) + \eta h(t, \cdot)\}.$$

(A3) The σ-algebra $\mathcal{T}$ is countably generated or the completion of a countably generated σ-algebra.

Remark 1. The integrals in (A2c) make sense in view of (A2a, b).

THEOREM 5. *Under* (A1)–(A3),

$$\liminf_k I(y_k, u_k) \geqq I(y_0, u_0).$$

Proof. In the case $\gamma = +\infty$ there is nothing to prove. Rather than taking a suitable subsequence, we may suppose without loss of generality that all of $\{I(y_k, u_k)\}$ are finite and converge to γ. By measurability of the inclusion $X_2 \subset \hat{X}_2$, (where $\hat{X}_2$ is defined as in § 2 for $S = X_2$) for every $k \in \mathbb{N}$, ε_{u_k} is in $\hat{\mathcal{R}}$. Here $\varepsilon_{u_k}(t)$ is defined as the Dirac probability measure at $u_k(t)$, $t \in T$, and $\hat{\mathcal{R}}$ is as in § 2 for $S = X_2$. By (A3) and Lemma 3 the set $\hat{\mathcal{R}}$ is sequentially compact. Rather than extracting a convergent subsequence, we suppose without loss of generality that $\{\varepsilon_{u_k}\}$ converges to a certain $\bar{\delta} \in \hat{\mathcal{R}}$. We claim first:

$$\bar{\delta} \in \mathcal{R}. \tag{5}$$

To see this, we define the functional $\hat{h}$ on $T \times \hat{X}_2$, by setting $\hat{h} \equiv h$ on $T \times X_2$ and $\hat{h} \equiv +\infty$ on $T \times (\hat{X}_2 \backslash X_2)$. It is easy to see that $\hat{h}$ is product-measurable and nonnegative. Also, for every $t \in T$, $\beta \in \mathbb{R}$ the set $\{x \in \hat{X}_2: \hat{h}(t, x) \leqq \beta\}$ is identical to the set $\{x \in X_2: h(t, x) \leqq \beta\}$. By (A2) the latter is a compact subset of X_2, hence homeomorphic to (i.e., identified with) a compact subset of $\hat{X}_2$. So we can conclude that $\hat{h}$ is a nonnegative normal integrand on $T \times \hat{X}_2$. By Lemma 2, as explained in the first part of the proof of Lemma 4, the functional $\delta \mapsto \int_T \hat{h}(t, \delta(t))\mu(dt)$ is lower semicontinuous on $\hat{\mathcal{R}}$. Therefore, (A2a) entails

$$\int_T \hat{h}(t, \bar{\delta}(t))\mu(dt) \leqq 1. \tag{6}$$

In view of the nature of $\hat{h}$, this implies that for μ-a.e. $t \in T$ the probability measure $\bar{\delta}(t)$ is supported by X_2; i.e., (5) holds. Next, we will demonstrate that

$$\gamma \geqq I(y_0, \bar{\delta}), \tag{7}$$

where we use the from now on self-evident notation:

$$I(y_0, \bar{\delta}) \equiv \int_T f(t, y_0(t), \bar{\delta}(t))\mu(dt).$$

To prove (7), note that $I = \lim_n \uparrow \hat{I}_n$ on $\mathcal{L} \times \mathcal{M}$, by an application of Lemma 2 for $S = X_1 \times X_2$, $g = f$, and the monotone convergence theorem. Here for $n \in \mathbb{N}$

$$\hat{I}_n(y, u) \equiv \int_T \hat{f}_n(t, y(t), u(t))\mu(dt), \qquad y \in \mathcal{L}, \quad u \in \mathcal{M},$$

with $\hat{f}_n$ defined as $\hat{g}_n$ in Lemma 2. By (3), (5) and the monotone convergence theorem, it is therefore enough to show that for every $n \in \mathbb{N}$

$$\gamma_n \equiv \liminf_k \hat{I}_n(y_k, u_k) \geqq \hat{I}_n(y_0, \bar{\delta}). \tag{8}$$

Fix $n \in \mathbb{N}$, arbitrarily. For every $k \in \mathbb{N}$

$$\hat{I}_n(y_k, u_k) = (\hat{I}_n(y_k, u_k) - \hat{I}_n(y_0, u_k)) + \hat{I}_n(y_0, u_k). \tag{9}$$

By (4) the first term in (9) can be majorized in absolute value as follows:

$$|\hat{I}_n(y_k, u_k) - \hat{I}_n(y_0, u_k)| \leqq n \int_T \rho_1(y_k(t), y_0(t)) \mu(dt).$$

Here ρ_i is defined as ρ in § 2 for $S = X_i$, equipped with the metric d_i, $i = 1, 2$. To apply (4) we note that ρ_0, which is defined in analogy to ρ for the Lusin space $X_1 \times X_2$ with metric $d_1 + d_2$, satisfies $\rho_0 \leqq \rho_1 + \rho_2$.

By (A1) and Proposition 1, the first term in (9) converges to zero as k goes to infinity. Thus,

$$\gamma_n = \liminf_k \hat{I}_n(y_0, u_k).$$

The functional $(t, x) \mapsto \hat{f}_n(t, y_0(t), x)$ is a nonnegative normal integrand on $T \times \hat{X}_2$, so (8) holds, by the lower semicontinuity of the functional $\delta \mapsto \int_T \hat{f}_n(t, y_0(t), \delta(t)) \mu(dt)$ on $\hat{\mathcal{R}}$. This finishes the proof of (7). We conclude the proof of the theorem by showing that

$$I(y_0, \bar{\delta}) \geqq I(y_0, u_0); \tag{10}$$

together with (7) this will yield the desired result. First, we fix $B \in \mathcal{T}$, $a \in \mathcal{A}$ arbitrarily. Note that $g : (t, x) \mapsto 1_B(t) a(x)$ defines a normal integrand on $T \times X_2$ which, by (A2b), has the property that for every $\varepsilon > 0$ there exists $f_\varepsilon \in L_1(T; \mathbb{R})$ such that on $T \times X_2$

$$\max(-g, 0) \leqq \varepsilon h + f_\varepsilon.$$

It follows from Lemma 4 and (5) that

$$\liminf_k \int_B a(t, u_k(t)) \mu(dt) \geqq \int_B a(t, \bar{\delta}(t)) \mu(dt).$$

So by (A2c) we have that

$$\int_B a(t, u_0(t)) \mu(dt) \geqq \int_B a(t, \bar{\delta}(t)) \mu(dt).$$

In view of the countability of the set $\mathcal{A}$ we conclude that there exists a μ-null set N such that for every $t \notin N$

$$a(t, u_0(t)) \geqq a(t, \bar{\delta}(t)) \quad \text{for every } a \in \mathcal{A}. \tag{11}$$

Fix $\eta > 0$. By (11), for every $t \notin N$, $a \in \mathcal{A}$ such that $-a(t, \cdot) \leqq f(t, y_0(t), \cdot) + \eta h(t, \cdot)$,

$$f(t, y_0(t), \bar{\delta}(t)) + \eta h(t, \bar{\delta}(t)) \geqq -a(t, u_0(t)).$$

It thus follows from (A2d) and (6) that

$$I(y_0, \bar{\delta}) + \eta \geqq I(y_0, u_0).$$

Hence the remaining statement, (10), has also been proven.

Remark 2. In the case where the functional f also takes negative values, Theorem 5 remains valid, of course if f is bounded below by a μ-integrable functional on T, but also under the following more general assumption:

(A4) For every subsequence $\{(y_{k'}, u_{k'})\}$ of $\{(y_k, u_k)\}$ for which $\{I(y_{k'}, u_{k'})\}$ is bounded above, the sequence $\{\max(-f(\cdot, y_{k'}(\cdot), u_{k'}(\cdot)), 0)\}$ is weakly precompact in $L_1(T; \mathbb{R})$.

To see this, it is enough to refer to [4], where this "lower compactness property" was introduced. It is shown there that under (A4) one can suppose without loss of generality that the integrand f is bounded below in the lower semicontinuity proof. (The argument is carried over to the present more general setting.)

4. Some applications. We hardly need to mention that the usefulness of Theorem 5 hinges on the satisfaction of assumption (A2). The remaining assumptions are quite transparent, so the formulation of cases in which they are satisfied can be left to the reader. In this section we shall formulate a few cases in which assumption (A2) holds.

In the first case we encounter a generalized version of the classical situation, where the integrand f has a (classical) convexity property. Together with Theorem 5 and Remark 2 one can thus obtain generalizations of the lower semicontinuity results of [2], [3], [4].

Case 1. Suppose that $X_2 = \mathbb{R}^m$, $m \in \mathbb{N}$ and that for every $k \in \mathbb{N}$ $u_k \in L_1(T; \mathbb{R}^m)$. Then (A2) holds if:

(C1a) the sequence $\{\mu_k\}$ converges to u_0 in $\sigma(L_1(T; \mathbb{N}^m), L_\infty(T; \mathbb{R}^m))$.

(C1b) for μ-a.e. $t \in T$, $f(t, y_0(t), \cdot)$ is convex.

Proof. By the theorem of de la Vallée Poussin [11, II.22, 25] (taking into account the minor misprint in [11, line 3, p. 39]) and (C1a) there exists a nonnegative lower semicontinuous functional h' on $\mathbb{R}_+$ such that $\lim_{\beta\to\infty} h'(\beta)/\beta = +\infty$ and for every $k \in \mathbb{N}$

$$\int_T h'(|u_k(t)|)\mu(dt) \leqq 1.$$

If we set $h(t, x) \equiv h'(|x|)$, $t \in T$, $x \in \mathbb{R}^m$, then (A2a) holds. Let us define $\mathscr{A}$ to be the collection of all functionals $a_{z,\zeta}$, $\zeta \in \mathbb{Q}$ (rationals), $z \in \mathbb{Q}^m$, where

$$a_{z,\zeta}(t, x) \equiv z \cdot x + \zeta, \qquad t \in T, \quad x \in \mathbb{R}^m,$$

(here $\cdot$ denotes the usual inner product). The properties of h' imply that (A2b) is valid, and (C1a) implies (A2c). Finally, (A2d) holds by [17, Corollary], since for every $t \in T$ and $\eta > 0$ the epigraph of $f(t, y_0(t), \cdot) + \eta h(t, \cdot)$ does not contain a straight line by inf-compactness of $h(t, \cdot)$.

Case 2. Suppose that $X_2 = E$, a separable Banach space with norm $\|\cdot\|$, separable dual E^* and dual norm $\|\cdot\|^*$, that $(T, \mathscr{T}, \mu)$ is a complete measure space, and that there exists a measurable multifunction Γ from T into E with convex, compact values such that for every $k \in \mathbb{N} \cup \{0\}$:

(C2a) for μ-a.e. $t \in T$, $u_k(t) \in \Gamma(t)$.

Also, suppose that there exists $r \in L_1(T; \mathbb{R})$ such that for every $t \in T$, $x \in \Gamma(t)$:

(C2b) $\|x\| \leqq r(t)$.

Finally, suppose that:

(C2c) The sequence $\{u_k\}$ converges to u_0 in $\sigma(L_1(T; E), L_\infty(T; E^*))$.

(C2d) For μ-a.e. $t \in T$, $f(t, y_0(t), \cdot)$ is convex on $\Gamma(t)$.

Then (A2) is satisfied.

Proof. Define the functional h on $T \times E$ by $h(t, x) \equiv 0$ if $x \in \Gamma(t)$, $\equiv +\infty$ if $x \notin \Gamma(t)$, $t \in T$; then h is a nonnegative normal integrand on $T \times E$, inf-compact in its second argument and (A2a) holds by virtue of (C2a). Further, for every $t \in T$, we define the convex lower semicontinuous functional $f_0(t, \cdot)$ by $f_0(t, x) \equiv f(t, y_0(t), x)$ if $x \in \Gamma(t)$, $\equiv +\infty$ if not; note that we may assume without loss of generality that the convex functional $f_0(t, \cdot)$ is proper [14, I.3]. We define the collection $\mathscr{A}$ by repeating the argument of [3] as follows. By [14, VII.2], the functional $(t, x^*) \mapsto f_0^*(t, x^*) \equiv$

$\sup\{\langle x, x^*\rangle - f_0(t, x): x \in X\}$ is a normal integrand on $T \times E^*$ (as usual, $\langle \cdot, \cdot \rangle$ denotes the duality between E and E^*). Therefore, by [14, III.22], the multifunction $t \mapsto \{(x^*, \beta): \beta \geqq f_0^*(t, x^*)\}$ is measurable from T into the separable Banach space $E^* \times \mathbb{R}$ and must have a Castaing representation. That is to say, there exist countable collections $\{v_i\}$, $\{\beta_i\}$, consisting of measurable functions from T into E^* and $\mathbb{R}$, respectively, such that for every $t \in T$

$$cl\{(v_i(t), \beta_i(t)): i \in \mathbb{N}\} = \{(x^*, \beta) \in E^* \times \mathbb{R}: \beta \geqq f_0^*(t, x^*)\} \tag{12}$$

and

$$\beta_i(t) \geqq f_0^*(t, v_i(t)) \quad \text{for every } i \in \mathbb{N}. \tag{13}$$

By [14, I.3], $f_0(t, \cdot)$ equals its biconjugate for μ-a.e. $t \in T$; one easily checks that therefore for μ-a.e. $t \in T$ and every $x \in E$

$$f_0(t, x) = \sup\{\langle x, v_i(t)\rangle - \beta_i(t): i \in \mathbb{N}\}. \tag{14}$$

For $t \in T$, $i, j \in \mathbb{N}$, define $(v_{i,j}(t), \beta_{i,j}(t)) \equiv (v_i(t), \beta_i(t))$ if $\|v_i(t)\|^* \leqq j$, and $|\beta_{i,j}(t)| \leqq j$, $\equiv (0, 0)$ otherwise. Then the collection $\mathcal{A}$ consisting of all functionals $(t, x) \mapsto \beta_{i,j}(t) - \langle x, v_{i,j}(t)\rangle$, $i, j \in \mathbb{N}$, satisfies, in view of (12), (13), (14) and the nonnegativity of f,

$$f_0(t, x) = \sup\{-a(t, x): -a(t, \cdot) \leqq f_0(t, \cdot),\ a \in \mathcal{A}\}$$

for every $t \in T$ and $x \in \Gamma(t)$. Therefore, (A2d) holds. It is also easy to verify that, by (C2b) for every $i, j \in \mathbb{N}$ and every $t \in T$, $x \in \Gamma(t)$,

$$|\beta_{i,j}(t) - \langle x, v_{i,j}(t)\rangle| \leqq j + jr(t),$$

and this means that (A2b) holds. Finally, (A2c) follows immediately from (C2c) and the above.

The next case confronts us with a situation that is also presented by the "simplest lower semicontinuity problem" (where μ is a Dirac measure).

Case 3. As Case 1, but without the convexity assumption (C1b). Then (A2) is valid if:

(C3) the sequence $\{u_k\}$ converges to u_0 in L_1-norm.

Proof. Let h be introduced as in Case 1; then (A2a) is satisfied. Define $\mathcal{A}$ to be the collection of functionals $a_{n,\zeta,z}$, $n \in \mathbb{N}$, $\zeta \in \mathbb{Q}$, $z \in \mathbb{Q}^m$, where

$$a_{n,\zeta,z}(t, x) \equiv n|x - z| + \zeta, \qquad t \in T, \quad x \in \mathbb{R}^m.$$

Then $\mathcal{A}$ has the following needle type property at every point $x \in \mathbb{R}^m$ [13]: for every $\hat{a} \in \mathcal{A}$, ε, $\hat{\eta} > 0$ there exist $a \in \mathcal{A}$, $\eta > 0$, $\eta \leqq \hat{\eta}$, such that for every $x' \in \mathbb{R}^m$

$$|x' - x| \geqq \eta, \quad \text{implies } -a(x') \leqq -\hat{a}(x'),$$
$$|x' - x| < \eta, \quad \text{implies } -a(x') \leqq \varepsilon.$$

It follows from [13, Thm. 1] (where this result was proven for $\varepsilon = 0$), as extended in [18], that (A2d) is satisfied, since $f(t, y_0(t), \cdot)$ is nonnegative and lower semicontinuous for every $t \in T$ (cf. the proof of Lemma 2). From the properties of h, as described in the proof of Case 1, it follows that (A2b) is valid. Finally, it is not hard to see that (C3) is

equivalent to the following: for every $a \in \mathcal{A}$, $B \in \mathcal{T}$,

$$\lim_k \int_B a(t, u_k(t))\mu(dt) = \int_B a(t, u_0(t))\mu(dt).$$

(Consider step functions with values in $\mathbb{Q}^m$ and use the fact that these are dense in $L_1(T; \mathbb{R}^m)$.) Moreover, since for every $B \in \mathcal{T}$, $z \in \mathbb{Q}$ the functional $u \mapsto \int_B |u(t) - z|\mu(dt)$ is convex and lower semicontinuous on $L_1(T; \mathbb{R}^m)$, assumption (C3) is in fact equivalent to (A2c) in the presence of (C1a). Thus, the convergence of $\{u_k\}$ is of necessity of a strong nature, although the sequence $\{a(\cdot, u_k(\cdot))\}$ does converge weakly.

Remark 3. Let us point out here that the metrizability condition imposed upon the Lusin space X_2 stands in the way of a number of potentially interesting applications (e.g., the case where X_2 is a separable Hilbert space, equipped with its weak topology). In notable contrast to some results on the generalized biconjugate of an integral functional (which form the subject of a forthcoming paper by the present author) we have not been able to rid the space from the metrizability condition. This explains the relatively simple choices made for X_2 in the above case studies.

Note. In [19] the present author has demonstrated that a large number of lower closure problems with weak convergence conditions, can be reduced to the lower semicontinuity problem that was solved in this paper. A quite different approach to the existence problem along the lines of relaxed control theory can be found in [20] (and other references mentioned there), where "control functions" and "trajectories" are relaxed.

REFERENCES

[1] C. B. MORREY, *Multiple integral problems in the calculus of variations and related topics*, Univ. California Publ. Math., 1 (1943), pp. 1–130.

[2] L. D. BERKOVITZ, *Lower semicontinuity of integral functionals*, Trans. Amer. Math. Soc., 192 (1974), pp. 51–57.

[3] C. OLECH, *Weak lower semicontinuity of integral functionals*, J. Optimization Theory Appl., 19 (1976), pp. 3–16.

[4] A. D. IOFFE, *On lower semicontinuity of integral functionals* I, this Journal, 15 (1977), pp. 521–538.

[5] ———, *On lower semicontinuity of integral functionals* II, Ibid., 15 (1977), pp. 991–1000.

[6] C. OLECH, *Existence theory in optimal control problems—the underlying ideas*, International Conference on Differential Equations, H. A. Antosiewicz, ed., Academic Press, New York, 1975, pp. 612–629.

[7] E. J. BALDER, *On a useful compactification for optimal control problems*, J. Math. Anal. Appl., 72 (1979), pp. 391–398.

[8] H. BERLIOCCHI AND J.-M. LASRY, *Intégrandes normales et mesures paramétrées en calcul des variations*, Bull. Soc. Math. France, 131 (1973), pp. 129–184.

[9] E. J. MCSHANE, *Relaxed controls and variational problems*, this Journal, 5 (1967), pp. 438–458.

[10] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

[11] C. DELLACHERIE AND P. A. MEYER, *Probabilités et Potentiel*, Hermann, Paris, 1975.

[12] E. J. BALDER, *Relaxed inf-compactness for variational problems by Hilbert cube compactification*, J. Math. Anal. Appl., to appear.

[13] ———, *An extension of duality-stability relations to non-convex optimization problems*, this Journal, 15 (1977), pp. 329–343.

[14] C. CASTAING AND M. VALADIER, *Convex Analysis and Measurable Multifunctions*, Springer Lecture Notes in Mathematics, 580, Springer, Berlin, 1977.

[15] L. C. YOUNG, *Generalized curves and the existence of an attained absolute minimum in the calculus of variations*, C. R. Acad. Sci. Lett. Varsovie C III, 30 (1937), pp. 212–234.

[16] R. B. HOLMES, *Geometric Functional Analysis and its Applications*, Springer, Berlin, 1975.

[17] V. KLEE AND C. OLECH, *Characterizations of a class of convex sets*, Math. Scand., 20 (1967), pp. 290–296.

[18] P. O. LINDBERG, *A generalization of Fenchel conjugation giving generalized Lagrangians and symmetric nonconvex duality*, Survey of Mathematical Programming, A. Prékopa, ed., North Holland, Amsterdam, 1980, pp. 249–269.

[19] E. J. BALDER, *Lower closure problems with weak convergence conditions in a new perspective*, this Journal, to appear.

[20] R. M. LEWIS AND R. B. VINTER, *Relation of optimal control problems to equivalent convex programs*, J. Math. Anal. Appl., 74 (1980), pp. 475–493.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 4, July 1981

0363-0129/81/1904-0009 $01.00/0

OPTIMAL PLAY IN A STOCHASTIC DIFFERENTIAL GAME*

R. J. ELLIOTT† AND M. H. A. DAVIS‡

Abstract. This paper considers play in a two-person zero-sum differential game where the dynamics are given by a differential equation with additive white noise. Feedback strategies are employed. Standard results from control theory show that the maximizing player has an optimal response to any pre-announced strategy of the minimizing player. Here it is shown that the minimizing player can achieve the upper value of the game by playing a strategy which is constructed by performing a pointwise *min–max* on a certain fixed Hamiltonian function.

1. Introduction. In reference [7] the martingale methods of Davis and Varaiya [4] were applied to two-person zero-sum differential games, and in a later paper [9] the existence of optimal strategies in a game was discussed. Roughly, the method of [9] and [7] consisted of two steps: it was first assumed that the minimizing player J_2 would announce what control z he was going to use throughout the game and the optimal reply $y^*(z)$ for the maximizing player J_1 was shown to exist. Then, secondly, the optimal control z^* for J_2 was investigated, given that J_2 was to play first in the above manner. However, on re-reading [7] and [9] it eventually became clear that, because the optimal reply $y^*(z)$ depends on the knowledge of the future behavior of z, the upper value function W_t^+ (V_t^+ in [9]), is not a submartingale under the measure constructed from $y^*(z)$ and z.

In this paper we show that W_t^+ satisfies a certain dynamic programming identity, Theorem 3.5. For each control z a reply $\hat{y}(z)$ is then constructed by maximizing a certain fixed Hamiltonian, so that $\hat{y}(z)$ does not depend on information about future behavior of z. When the infimum of the payoffs, (or costs), is taken over all the controls z, the maximizing player J_1 can do as well by using the reply $\hat{y}(z)$. Finally, we show that the player J_2 should choose a control z which minimizes the maximum of the Hamiltonian mentioned above. Thus, the final results of [7] and [8] are correct.

Most results from previous publications that we use below are summarized in [3].

2. Dynamics and payoff. Suppose the dynamics of a stochastic system are described by a differential equation of the form

$$dx_t = f(t, x, y, z)\,dt + dw,$$

with initial condition $x(0) = 0 \in R^m$. Here $t \in [0, 1]$ and w is an m-dimensional Brownian motion. Write $\mathscr{C}$ for the space of continuous functions from $[0, 1]$ to R^m, $\mathscr{F}_t^0$ for the σ-field in $\mathscr{C}$ generated by the coordinate functions $\{x_s, s \leqq t\}$ and $\mathscr{F}_t$ for the completion of $\mathscr{F}_t^0$ with all null sets of Wiener measure on $(\mathscr{C}, \mathscr{F}_1^0)$. Finally, let $\mathscr{P}$ denote the $\mathscr{F}_t$-predictable σ-field on $[0,1] \times \mathscr{C}$.

DEFINITION 2.1. The *drift function* f maps $[0, 1] \times \mathscr{C} \times Y \times Z$ into R^m. Here Y and Z are compact subsets of Euclidean spaces R^k and R^l, and are the sets where the control functions take values. Furthermore, f is supposed to satisfy the following conditions:

(i) f is jointly measurable with respect to the product σ-field of $\mathscr{P}$ on $[0, 1] \times \mathscr{C}$ and the Borel fields on Y and Z.

(ii) For each $t \in [0, 1]$, $x \in \mathscr{C}$, $f(t, x, \cdots)$ is continuous on $Y \times Z$.

(iii) $|f(t, x, y, z)| \leqq M(1 + \|x\|_t)$, where $\|x\|_t = \sup_{0 \leqq s \leqq t} |x(s)|$.

* Received by the editors November 16, 1979, and in revised form August 1, 1980.

† Department of Pure Mathematics, University of Hull, Hull HU5 2DW, England.

‡ Department of Electrical Engineering, Imperial College of Science and Technology, London SW7 2BT, England.

A player J_1 controls the parameter $y \in Y$ as the game evolves. Similarly, a player J_2 controls $z \in Z$.

DEFINITION 2.2. For $0 \leqq s \leqq t \leqq 1$ the *admissible control strategies* $\mathcal{M}_s^t$ for J_1 ($\mathcal{N}_s^t$ for J_2) are all Y-valued (Z-valued) $\mathcal{P}$-measurable functions on $[s, t] \times \mathcal{C}$.

Write $\mathcal{M} = \mathcal{M}_0^1$ and $\mathcal{N} = \mathcal{N}_0^1$. For $y \in \mathcal{M}$ and $z \in \mathcal{N}$ write

$$f^{y,z}(t, x) = f(t, x, y(t, x), z(t, x)).$$

Let μ denote Wiener measure on $(\mathcal{C}, \mathcal{F}_1)$, so that $\{x_t\}$ is a standard Brownian motion on $(\mathcal{C}, \mathcal{F}_1, \mu)$. E denotes integration with respect to μ. The conditions on f ensure that

$$E[\exp \zeta_s^t(f^{y,z}) \,|\, \mathcal{F}_s] = 1 \quad \text{a.s. } \mu,$$

where

$$\zeta_s^t(f^{y,z}) = \int_s^t f^{y,z}(\tau, x)' \, dx_\tau - \frac{1}{2} \int_s^t |f^{y,z}(\tau, x)|^2 \, d\tau.$$

Therefore, for each $y \in \mathcal{M}$, $z \in \mathcal{N}$, a probability measure $\mu_{y,z}$ can be defined on Ω by putting

$$\frac{d\mu_{y,z}}{d\mu} = \rho_0^1(y, z),$$

where

$$\rho_0^1(y, z) = \exp \zeta_0^1(f^{y,z}).$$

Girsanov's theorem then states the following result.

THEOREM 2.3. *Under the measure $\mu_{y,z}$ the process $w_t^{y,z}$ is a Brownian motion, where*

$$dw_t^{y,z} = dx_t - f^{y,z}(t, x) \, dt.$$

That is, under the measure $\mu_{y,z}$,

$$dx_t = f^{y,z}(t, x) \, dt + dw_t^{y,z},$$

so that x_t is a solution of the dynamical equation when admissible controls y and z are used.

Payoff. The cost, or payoff, is supposed to be just of terminal form

$$P(y, z) = E_{y,z}[\phi],$$

where $E_{y,z}$ denotes expectation with respect to $\mu_{y,z}$ and ϕ is a given bounded $\mathcal{F}_1$-measurable random variable.

We are considering a zero sum differential game, so J_1 wishes to choose y so that $P(y, z)$ is maximized and J_2 wishes to choose z so that $P(y, z)$ is minimized.

3. The upper value. Suppose that J_2 announces at the beginning of the game that he (or she) is going to use control $z \in \mathcal{N}$ throughout the game. Then, because $f(t, x, y, z(t, x))$ is continuous in $y \in Y$ we can apply the principal theorem of [2] as reformulated in [3] to deduce that, with the knowledge of z, there is an optimal reply $y^*(z) \in \mathcal{M}$ for J_1 such that

$$\sup_{y \in \mathcal{M}} E_{y,z}[\phi] = E_{y^*(z),z}[\phi].$$

Note that y^* is a *strategy-to-strategy* mapping, $y^* : \mathcal{N} \to \mathcal{M}$, and that the optimal reply at time t depends on the *future* strategy $\mathcal{N}_t^1$ of J_2.

DEFINITION 3.1. The quantity

$$W_0^+ = \inf_{z \in \mathcal{N}} \sup_{y \in \mathcal{M}} P(y, z) = \inf_z P(y^*(z), z)$$

is called the *upper value* of the game.

Remarks 3.2. The optimum reply $y^*(z)$ to z, described above, is unrealistic because it supposes that J_2 discloses in advance what he is going to do. The discussion below shows that J_1 can attain the upper value by playing an admissible control which does not depend on future information about the intentions of J_2. Furthermore, it indicates how J_2 should play optimally.

DEFINITION 3.3. Similarly to Definition 3.1, we define the *upper value process*

$$W_t^+ = \inf_{z \in \mathcal{N}_t^1} \sup_{y \in \mathcal{M}_t^1} E_{y,z}[\phi \mid \mathcal{F}_t].$$

Here the infimum and supremum are taken in the complete lattice $L^1(\Omega, \mu, \mathcal{F}_t)$ (see [6, p. 302]). For any $z \in \mathcal{N}_t^1$ there is, again applying the result of [2], an optimal reply $\tilde{y}^*(z)$ such that

$$\sup_y E_{y,z}[\phi \mid \mathcal{F}_t] = E_{\tilde{y}^*(z),z}[\phi \mid \mathcal{F}_t] \quad \text{a.s.} \tag{3.1}$$

Therefore,

$$W_t^+ = \inf_z E_{\tilde{y}^*(z),z}[\phi \mid \mathcal{F}_t].$$

The following lemma shows that we may take $\tilde{y}^*(z) = y^*(z)$.

LEMMA 3.4. *If $y^*(z)$ is an optimal to $z \in \mathcal{N}$ over the whole time interval $[0, 1]$, then $y^*(z)$ is an optimal reply to z over $[t, 1]$.*

Proof. This is a direct consequence of the "principle of optimality" [3, Prop. 2.8] which states that W_t^z is a martingale under $\mu_{y^*(z),z}$ and a supermartingale under any other admissible measure. Thus for $z \in \mathcal{N}$, since $W_1^z = \phi$ a.s.,

$$W_t^z = E_{y^*(z),z}[\phi \mid \mathcal{F}_t] \quad \text{a.s.},$$

and for any $y \in \mathcal{M}$

$$W_t^z \geqq E_{y,z}[\phi \mid \mathcal{F}_t] \quad \text{a.s.}$$

Thus (3.1) holds with $\tilde{y}^* = y^*$.

We next prove a fundamental dynamic programming identity for W_t^+.

THEOREM 3.5. *For each $t \in [0, 1]$ and $h \leqq 1 - t$*

$$W_t^+ = \inf_{z \in \mathcal{N}_t^{t+h}} \sup_{y \in \mathcal{M}_t^{t+h}} E_{y,z}[W_{t+h}^+ \mid \mathcal{F}_t].$$

Remarks 3.6. Again, the infimum and supremum are taken in the lattice $L^1(\Omega, \mu, \mathcal{F}_t)$, and implicit in the definition is that for each $z \in \mathcal{N}_t^{t+h}$ the supremum is first taken over $y \in \mathcal{M}_t^{t+h}$. In other words, y knows in advance what control J_2 is going to use over $[t, t+h]$. The infimum is then taken over $z \in \mathcal{N}_t^{t+h}$. Furthermore, again by the arguments of [2] quoted above, for any $z \in \mathcal{N}_t^{t+h}$ there is an optimal reply $y^*(t, h, z) = y^*$ such that

$$\sup_{y \in \mathcal{M}_t^{t+h}} E_{y,z}[W_{t+h}^+ \mid \mathcal{F}_t] = E_{y^*,z}[W_{t+h}^+ \mid \mathcal{F}_t].$$

We now give the proof of Theorem 3.5.

Proof. Write

$$I(t, h) = \inf_z \sup_y E_{y,z}[W_{t+h}^+ \mid \mathcal{F}_t].$$

If $\rho_s^t(y, z) = \exp \zeta_s^t(f^{y,z})$ we know that $\rho_r^s(y, z)\rho_s^t(y, z) = \rho_r^t(y, z)$ and $\rho_s^t(y, z)$ is a martingale on $(\Omega, \mathscr{F}_t, \mu)$. Also, for $y \in \mathscr{M}_t^{t+h}$ and $z \in \mathscr{N}_t^{t+h}$

$$E_{y,z}[W_{t+h}^+ \mid \mathscr{F}_t] = E[\rho_t^{t+h}(y, z) W_{t+h}^+ \mid \mathscr{F}_t] \quad \text{a.s.}$$

By the same argument as used in the proof of [4, Lemma 3.1] $\mathscr{N}_{t+h}^1$ is "relatively complete" in the sense that for any $\varepsilon > 0$ there is a $z_\varepsilon \in \mathscr{N}_{t+h}^1$ such that

$$\sup_{y \in \mathscr{M}_{t+h}^1} E_{y,z_\varepsilon}[\phi \mid \mathscr{F}_{t+h}] \leqq W_{t+h}^+ + \varepsilon \quad \text{a.s.}$$

Furthermore, there is a $z'_\varepsilon \in \mathscr{N}_t^{t+h}$ such that for all $y \in \mathscr{M}_t^{t+h}$

$$E_{y,z'_\varepsilon}[W_{t+h}^+ \mid \mathscr{F}_t] \leqq I(t, h) + \varepsilon \quad \text{a.s.}$$

Concatenating z'_ε and z_ε to give a control $z_\varepsilon^* \in \mathscr{N}_t^1$ we have that for all $y \in \mathscr{M}_t^1$

$$E_{y,z_\varepsilon^*}[\phi \mid \mathscr{F}_t] \leqq I(t, h) + 2\varepsilon.$$

Taking the supremum over $y \in \mathscr{M}_t^1$ and the infimum over $z \in \mathscr{N}_t^1$ we see that $W_t^+ \leqq I(t, h) + 2\varepsilon$.

Now consider any $z \in \mathscr{N}_t^1$. For the restriction of z to $[t+h, 1]$ there is an optimal reply $y^* = y^*(z) \in \mathscr{M}_{t+h}^1$ and certainly

$$E_{y^*,z}[\phi \mid \mathscr{F}_{t+h}] \geqq W_{t+h}^+ \quad \text{a.s.}$$

Therefore, for any $y \in \mathscr{M}_t^{t+h}$ concatenated with y^*

$$E_{y,z}[E_{y^*,z}[\phi \mid \mathscr{F}_{t+h}]\mathscr{F}_t] \geqq E_{y,z}[W_{t+h}^+ \mid \mathscr{F}_t]$$

so, for all $z \in \mathscr{N}_t^1$

$$\begin{aligned} \sup_{y \in \mathscr{M}_t^1} E_{y,z}[\phi \mid \mathscr{F}_t] &\geqq \sup_{y \in \mathscr{M}_t^{t+h}} E_{y,z}[E_{y^*,z}[\phi \mid \mathscr{F}_{t+h}]\mathscr{F}_t] \\ &\geqq \sup_{y \in \mathscr{M}_t^{t+h}} E_{y,z}[W_{t+h}^+ \mid \mathscr{F}_t]. \end{aligned}$$

Therefore

$$W_t^+ = \inf_z \sup_y E[\phi \mid \mathscr{F}_t] \geqq I(t, h)$$

and the result follows.

Notation 3.7. Write Φ for all the set predictable functions ϕ from $[0, 1] \times \mathscr{C} \to \boldsymbol{R}^m$ such that

$$|\varphi(t, x)| \leqq M(1 + \|x\|_t).$$

Then, if

$$\zeta_0^1(\varphi) = \int_0^1 \varphi(t, x)' \, dx - \frac{1}{2} \int_0^1 |\varphi(t, x)|^2 \, dt,$$

we have

$$E\rho_0^1(\varphi) = 1,$$

where

$$\rho_0^1(\varphi) = \exp \zeta_0^1(\varphi). \tag{3.2}$$

From [5, Thm. 2] we quote the following result.

THEOREM 3.8. *Define* $\mathscr{D} = \{\exp \zeta_0^1(\varphi)\colon \varphi \in \Phi\}$; *then* $\mathscr{D}$ *is a weakly compact subset of* $L^1(\Omega, \mathscr{F}_1, \mu)$.

Note that $f^{y,z} \in \Phi$, for $y \in \mathscr{M}$.[1] Consequently, for any sequence $\{(y_n, z_n)\} \subset \mathscr{M} \times \mathscr{N}$ there is a subsequence $\{(y_{n_k}, z_{n_k})\}$ and an element $h \in \Phi$ such that $\rho_0^1(y_{n_k}, z_{n_k}) \to \rho_0^1(h)$ weakly in L^1 as $k \to \infty$.

Suppose now that $\{z_n\}$ is a sequence in $\mathscr{N}$ such that

$$E_{y^*(z_n), z_n}[\phi]$$

decreases to W_0^+, where, as above, $y^*(z_n) \in \mathscr{M}$ is the optimum reply to z_n.

Then, by Theorem 3.8, there is a subsequence, again denoted by $\{z_n\}$, and a function $f^0 \in \Phi$ such that $\rho_n = \rho_0^1(y^*(z_n), z_n)$ converges weakly to $\rho_0^1(f^0)$. Write μ^* for the probability measure defined by

$$\frac{d\mu^*}{d\mu} = \rho_0^1(f^0) = \rho^*.$$

LEMMA 3.9. *For the above sequence of controls* $\{z_n\}$ *we have that*

$$\lim_n \int_F \rho_n (W_t^{z_n} - W_t^+)\, d\mu = 0,$$

for any $F \in \mathscr{F}_t$.

Proof. From Lemma 3.4 we know that $y^*(z_n)$ is also the optimum reply on $[t, 1]$ to z_n, so that for each n,

$$W_t^{z_n} = E_{y^*(z_n), z_n}[\phi \mid \mathscr{F}_t] \geqq \inf_{z \in \mathscr{N}_t^1} E_{y^*(z), z}[\phi \mid \mathscr{F}_t] = W_t^+ \quad \text{a.s.}$$

Also, for any $F \in \mathscr{F}_t$,

$$\begin{aligned} \lim_n \int_F \rho_n W_t^{z_n}\, d\mu &= \lim_n \int_F (\rho_n)_0^t E[(\rho_n)_t^1 \phi \mid \mathscr{F}_t]\, d\mu \\ &= \lim_n \int_F \rho_n \phi\, d\mu \\ &= \int_F \rho^* \phi\, d\mu = \int_F (\rho^*)_0^t E^*[\phi \mid \mathscr{F}_t]\, d\mu. \end{aligned}$$

Thus

$$0 \leqq \lim_n \int_F \rho_n (W_t^{z_n} - W_t^+)\, d\mu \leqq \lim_n \int_\Omega \rho_n (W_t^{z_n} - W_t^+)\, d\mu = \int_\Omega (\rho^*)_0^t (W_t^* - W_t^+)\, d\mu,$$

where $W_t^* = E^*[\phi \mid \mathscr{F}_t] \geqq W_t^+$. We shall show that

$$\lim_n \int_\Omega \rho_n (W_t^{z_n} - W_t^+)\, d\mu = 0.$$

[1] There is a slight abuse of notation, in that we previously wrote $\rho_0^1(y, z)$ for what should, according to (3.2), be denoted $\rho_0^1(f^{y,z})$. This will continue below.

Indeed, suppose to the contrary that

$$\lim_n \int_\Omega \rho_n(W_t^{z_n} - W_t^+)\, d\mu = \int_\Omega \rho^*(W_t^* - W_t^+)\, d\mu$$
$$= 2\varepsilon > 0.$$

Then there is a set $A \in \mathcal{F}_t$, with $\mu(A) > 0$, such that $W_t^* - W_t^+ > 2\varepsilon$ on A; that is,

$$W_t^+ + 2\varepsilon I_A < W_t^*.$$

Suppose $y'(0, t, z) = y'(z)$ is the optimum reply to $z \in \mathcal{N}_0^t$ in the game played over $[0, t]$ with payoff W_t^+, i.e.,

$$\sup_{y \in \mathcal{M}_0^t} E_{y,z}[W_t^+] = E_{y'(z),z}[W_t^+].$$

Then in particular,

$$E_{y'(z_n),z_n}[W_t^+] + 2\varepsilon \mu_{y'(z_n),z_n}(A) \leqq E_{y'(z_n),z_n}[W_t^*]$$
$$= \int_\Omega \rho_0^t(y'(z_n), z_n)\rho_t^1(f^0)\phi\, d\mu. \tag{3.3}$$

For any n,

$$\int_\Omega \rho_0^t(y'(z_n)z_n)\rho_t^1(y^*(z_n), z_n)\phi\, d\mu \leqq \int_\Omega \rho_0^t(y^*(z_n), z_n)\rho_t^1(y^*(z_n), z_n)\phi\, d\mu. \tag{3.4}$$

Again from Theorem 3.8 there is a subsequence, still denoted by $\rho_0^t(y'(z_n), z_n)$, and a function $h \in \Phi$, such that

$$\lim_n E_{y'(z_n),z_n}[W_t^+] = \inf_n E_{y'(z_n),z_n}[W_t^+],$$

and the densities $\rho_0^t(y'(z_n), z_n)$ converge weakly to $\rho_0^t(h)$. Write μ' for the measure defined by $d\mu'/d\mu = \rho_0^t(h)$. There is a sequence of convex combinations of the densities which converges strongly in L^1 (see [6, V.3.14]), so by considering the corresponding convex combinations of inequality (3.4) and taking the limit we have that

$$\int_\Omega \rho_0^t(h)\rho_t^1(f^0)\phi\, d\mu \leqq \int_\Omega \rho_0^1(f^0)\phi\, d\mu = W_0^+.$$

Therefore, taking the limit in (3.3), we have $\inf_n E_{y'(z_n),z_n}[W_t^+] + 2\varepsilon\mu'(A) \leqq W_0^+$. But this is a contradiction because $\mu'(A) > 0$, and by Theorem 3.5

$$W_0^+ = \inf_z E_{y'(z),z}[W_t^+].$$

This completes the proof.

THEOREM 3.10. *The process $\{W_t^+\}$ is a martingale under the measure μ^*.*

Proof. Consider $t \in [0, 1]$, $0 \leqq h \leqq 1 - t$ and $F \in \mathcal{F}_t$. We shall show that

$$\int_F (W_t^+ - W_{t+h}^+)\, d\mu^* = 0,$$

where, as above, $d\mu^*/d\mu = \rho_0^1(f^0) = \rho^*$ and $\rho_0^1(f^0)$ is the weak limit of the densities $\rho_n = \rho_0^1(y^*(z_n), z_n)$. Recall that $W_t^{z_n} = E_{y^*(z_n),z_n}[\phi \mid \mathscr{F}_t]$. Consider any $\varepsilon > 0$.

$$\int_F (W_t^+ - W_{t+h}^+)\rho^* \, d\mu = \int_F (\rho^* - \rho_n)(W_t^+ - W_{t+h}^+) \, d\mu$$

$$+ \int_F \rho_n (W_t^+ - W_t^{z_n}) \, d\mu$$

$$+ \int_F \rho_n (W_{t+h}^{z_n} - W_{t+h}^+) \, d\mu$$

$$+ \int_F \rho_n (W_t^{z_n} - W_{t+h}^{z_n}) \, d\mu.$$

The final term is zero because $W_t^{z_n}$ is a martingale under μ_n, (where $d\mu_n/d\mu = \rho_n$). Because the payoff is bounded, $I_F(W_t^+ - W_{t+h}^+) \in L^\infty$, so there is an n_1 such that, if $n > n_1$,

$$\left| \int_F (\rho^* - \rho_n)(W_t^+ - W_{t+h}^+) \, d\mu \right| < \frac{\varepsilon}{3}.$$

From Lemma 3.9 above there is an n_2 such that if $n > n_2$:

$$\left| \int_F \rho_n (W_t^{z_n} - W_t^+) \, d\mu \right| < \frac{\varepsilon}{3}$$

and

$$\left| \int_F \rho_n (W_{t+h}^{z_n} - W_{t+h}^+) \, d\mu \right| < \frac{\varepsilon}{3}.$$

Therefore, if $n > \max(n_1, n_2)$,

$$\left| \int_F \rho^* (W_t^+ - W_{t+h}^+) \, d\mu \right| \leqq \varepsilon,$$

where ε is arbitrary. Consequently,

$$E^*[W_{t+h}^+ \mid \mathscr{F}_t] = W_t^+,$$

where E^* denotes the expectation with respect to μ^*.

4. The maximizing strategy. In the above section we have shown that the upper value process satisfies a dynamic programming identity

$$W_t^+ = \inf_z \sup_y E_{y,z}[W_{t+h}^+ \mid \mathscr{F}_t],$$

and that there is a measure μ^*, given by $d\mu^*/d\mu = \rho_0^1(f^0)$, such that W_t^+ is a martingale under the measure μ^*.

Notation 4.1. For $t \in [0, 1]$, $h_n = n^{-1} \wedge (1-t)$ and $z \in \mathscr{N}$ write y_n^* for $y^*(t, h_n, z)$ (the optimal reply to z over $[t, t+h_n]$ in the game with payoff $W_{t+h_n}^+$). That is:

$$E_{y_n^*,z}[W_{t+h_n}^+ \mid \mathscr{F}_t] = \sup_y E_{y,z}[W_{t+h_n}^+ \mid \mathscr{F}_t].$$

By Girsanov's theorem, the process w^0 is a Brownian motion under μ^*, where $dw^0 = dx - f^0 \, ds$. Therefore, by the representation theorem (see [4, Thm. 2.3]) there is a

predictable process $\{g_t^*\}$ such that

$$E^*\left[\int_0^1 (g_t^*)^2\, dt\right] < \infty$$

and

$$W_t^+ = W_0^+ + \int_0^t g^*(dx - f^0\, ds).$$

It follows from [8, Lemma 6.6] that

$$E_{y,z}\left[\int_0^1 (g_t^*)^2\, dt\right] < \infty$$

for all $y \in \mathcal{M}$, $z \in \mathcal{N}$, and hence that

$$\int_0^t g_s^*\, dw_s^{y,z}$$

is a martingale under the measure $\mu_{y,z}$ (see Theorem 2.3 above). We use this fact below without further comment.

For any $z \in \mathcal{N}$, let $\hat{y}(z) \in \mathcal{M}$ be a control for J_1 such that

$$g_t^* f_t^{\hat{y}(z),z} \geqq g_t^* f^{y,z} \quad \text{a.s.}$$

for all other $y \in \mathcal{M}$. Such a control exists from [1, Lemma 1].[2]

Write

$$\Delta(t, h, y, z) = \int_t^{t+h} g_s^*(f_s^{y,z} - f_s^0)\, ds.$$

THEOREM 4.2. (i) *For all* n: $E_{y_n^*,z}[\Delta(t, h_n, y_n^*, z)\,|\,\mathcal{F}_t] \geqq 0$ a.s.

(ii) *For all* n:

$$E_{y_n^*,z}[\Delta(t, h_n, \hat{y}(z), z)\,|\,\mathcal{F}_t] \geqq E_{y_n^*,z}[\Delta(t, h_n, y_n^*, z)\,|\,\mathcal{F}_t]$$
$$\geqq E_{\hat{y}(z)z}[\Delta(t, h_n, \hat{y}(z), z)\,|\,\mathcal{F}_t] \quad \text{a.s.}$$

Proof. For $y \in \mathcal{M}$, $z \in \mathcal{N}$ we have from the above representation that

$$W_{t+h}^+ = W_t^+ + \int_t^{t+h} g_s^*(dx - f_s^{y,z}\, ds) + \int_t^{t+h} g_s^*(f_s^{y,z} - f_s^0)\, ds.$$

Therefore,

$$E_{y,z}[W_{t+h}^+ - W_t^+\,|\,\mathcal{F}_t] = E_{y,z}[\Delta(t, h, y, z)\,|\,\mathcal{F}_t].$$

Suppose now that y_n^* is the optimum reply to z, as above. Then

$$E_{y_n^*,z}[\Delta(t, h_n, y_n^*, z)\,|\,\mathcal{F}_t] \geqq E_{y,z}[\Delta(t, h_n, y, z)\,|\,\mathcal{F}_t],$$

and, because $\min_z E_{y_n^*,z}[\Delta(t, h_n, y_n^*, z)\,|\,\mathcal{F}_t] = 0$, by Theorem 3.5,

$$E_{y_n^*,z}[\Delta(t, h_n, y_n^*, z)\,|\,\mathcal{F}_t] \geqq 0,$$

so establishing inequality (i).

[2] Note that (in contrast to $y^*(z)$) $\hat{y}(z)_t$ depends on z only through the value z_t.

From the definition of $\hat{y}(z)$ we have that $\Delta(t, h_n, \hat{y}(z), z) \geqq \Delta(t, h_n, y_n^*, z)$

$$E_{y_n^*,z}[\Delta(t, h_n, \hat{y}(z), z) | \mathscr{F}_t] \geqq E_{y_n^*,z}[\Delta(t, h_n, y_n^*, z) | \mathscr{F}_t]$$
$$\geqq E_{\hat{y}(z),z}[\Delta(t, h_n, \hat{y}(z), z) | \mathscr{F}_t] \quad \text{a.s.}$$

Rephrasing the above inequalities we have the following corollary.

COROLLARY 4.3. *For any set* $F \in \mathscr{F}_t$,

$$\int_F \rho_t^{t+h_n}(y_n^*, z_n)\Delta(t, h_n, y_n^*, z)\, d\mu \geqq 0$$

and

$$\int_F \rho_t^{t+h_n}(y_n^*, z)\Delta(t, h_n, \hat{y}(z), z)\, d\mu \geqq \int_F \rho_t^{t+h_n}(y_n^*, z)\Delta(t, h_n, y_n^*, z)\, d\mu$$
$$\geqq \int_F \rho_t^{t+h_n}(\hat{y}(z), z)\Delta(t, h_n, \hat{y}(z), z)\, d\mu.$$

Remarks 4.4. By multiplying by n and letting n tend to infinity, we wish to differentiate these inequalities. However, as it is pointed out in [8, § 7] (see also [10]), care must be taken to show that there is a single subset $T \subset [0, 1]$ of measure zero such that the results hold for all $t \notin T$. This can be done because both the σ-fields and the spaces of controls are countably generated (see [8] and [10]).

Because the trajectories in $\mathscr{C}$ are continuous almost surely, $\mathscr{F}_r$ is countably generated for every rational number $r \in [0, 1]$ by sets $\{A_{ir}\}$, $i = 1, 2, \cdots$. Suppose $\mathscr{G}_t$ is the set of measurable functions $\{y\}$ from $(\mathscr{C}, \mathscr{F}_t)$ to $Y \subset R^k$. Because Y is compact, $E|y| < \infty$ if $y \in \mathscr{G}_t$, and if $y(\cdot, \cdot)$ is an admissible control for J_1 $y(t, x(\cdot))$ is in $\mathscr{G}_t$. There is a countable dense subset $G_r = \{y_{jr}\}$ of $\mathscr{G}_r$ and, if $G_t = \bigcup_{r \leqq t} G_r$, then G_t is a countable dense subset of $\mathscr{G}_t$.

Similarly, if $\mathscr{H}_t$ is the set of measurable functions $\{z\}$ from $(\mathscr{C}, \mathscr{F}_t)$ to $Z \subset R^l$ there is a countable dense subset H_t of $\mathscr{H}_t$ for each t, where $H_t = \bigcup_{r \leqq t} H_r$. Suppose, $z_{jr} \in H_r$. Then, as a function constant in time, z_{jr} can be considered as belonging to $\mathscr{N}_t^{t+h}$ for any $t \geqq r$.

As above, write $y_n^*(z_{jr})$ for the optimal reply to z_{jr} over $[t, t+h]$. Then for each i, j, r

$$\lim_{n\to\infty} n \int_t^{t+h_n} \int_{A_{ir}} \rho_t^{t+h_n}(\hat{y}(z_{jr}), z_{jr}) g_s^*(f_s^{\hat{y}(z_{jr}),z_{jr}} - f_s^0)\, d\mu\, ds$$

exists and equals

$$\int_{A_{ir}} g^*(f_t^{\hat{y}(z_{jr}),z_{jr}} - f_t^0)\, d\mu$$

for almost all $t \in [0, 1]$. Consequently, there is a set $T_1 \subset [0, 1]$ of zero measure, such that for $t \notin T_1$ and all i, j, r

$$\lim_{n\to\infty} n \int_t^{t+h_n} \int_{A_{ir}} \rho_t^{t+h} n(\hat{y}(z_{jr}), z_{jr}) g^*(f_s^{\hat{y}(z_{jr}),z_{jr}} - f_s^0)\, d\mu\, ds = \int_{A_{ir}} g^*(f^{\hat{y}(z_{ir}),z_{jr}} - f_t^0)\, d\mu.$$

Also, there is a set $T_2 \subset [0, 1]$, of zero measure, such that if $t \notin T_2$,

$$\lim_n n \int_t^{t+h_n} E[g_s^{*2}]\, ds = E[g_t^{*2}].$$

Because all the densities $\rho_t^{t+h}(y, z)$ form a uniformly integrable set, and because

$$\lim_n \rho_t^{t+h_n}(y_n^*(z_{jr}), z_{jr}) = 1 \quad \text{a.s.},$$

we can then prove, as in [8, Lemma 7.2], the following result.

LEMMA 4.5. *For $t \notin T = T_1 \cup T_2$, all $r \leqq t$ and all i, j,*

$$\lim_n n \int_t^{t+h_n} \int_{A_{ir}} \rho_t^{t+h_n}(y_n^*(z_{jr}), z_{jr}) g^*(f^{\hat{y}(z_{jr}), z_{jr}} - f_s^0)\, d\mu\, ds$$

$$= \int_{A_{ir}} g^*(f^{\hat{y}(z_{jr}), z_{jr}} - f_t^0)\, d\mu.$$

Therefore, from the inequalities of Corollary 4.3 we have:

THEOREM 4.6. *For $t \notin T$*

$$\lim_n n \int_t^{t+h_n} \int_{A_{ir}} \rho_t^{t+h_n}(y_n^*(z_{jr}), z_{jr}) g^*(f^{y_n^*, z_n} - f_s^0)\, d\mu\, ds$$

$$= \int_{A_{ir}} g^*(f^{\hat{y}(z_{jr}), z_{jr}} - f_t^0)\, d\mu.$$

However, each term in the limit above is nonnegative, so by the monotone class theorem we have

$$\int_A g^*(f^{\hat{y}(z_{jr}), z_{jr}} - f_t^0)\, d\mu \geqq 0,$$

for all $A \in \mathcal{F}_t$. Because the integrands are $\mathcal{F}_t$-measurable,

$$g^* f^{\hat{y}(z_{jr}), z_{jr}} \geqq g^* f^0,$$

for all j and all $r \leqq t$.

By approximating the value $z(t, x)$ of an arbitrary control $z \in \mathcal{N}$ by functions z_{jr} we have finally, because f is continuous in the control variables, the following result.

THEOREM 4.7. *For all $z \in \mathcal{N}$ and $t \notin T$*

$$g^* f^{\hat{y}(z), z} \geqq g^* f_t^0 \quad a.s.$$

5. Optimal controls. In this section we show that J_1 can attain the upper value by playing, in reply to control $z \in \mathcal{N}$, the maximizing control $\hat{y}(z)$. This control, $\hat{y}(z)$, is more reasonable to consider as a reply to z than, say, $y^*(z)$, because $\hat{y}(z)$ depends at any time t only on the values of z up to and including t, and does not depend on any information about future behavior of z.

Recall that $y^*(z)$ is the optimal control for J_1 when J_2 announces at the beginning of the game that he is going to play control $z \in \mathcal{N}$ throughout the time interval $[0, 1]$.

THEOREM 5.1.

$$\inf_{z \in \mathcal{N}} P(\hat{y}(z), z) = \inf_{z \in \mathcal{N}} P(y^*(z), z) = W_0^+.$$

Proof. From Theorem 3.10 we have that W_t^+ is a martingale under measure μ^* and has a representation

$$W_t^+ = W_0^+ + \int_0^t g^*(dx - f^0\,ds).$$

In particular,

$$\begin{aligned}\phi &= W_0^+ + \int_0^1 g^*(dx - f^0\,ds)\\ &= W_0^+ + \int_0^1 g^*(dx - f^{\hat{y}(z),z}\,ds) + \int_0^1 g^*(f^{\hat{y}(z),z} - f^0)\,ds\end{aligned}$$

for any $z \in \mathcal{N}$. Therefore,

$$P(\hat{y}(z), z) = W_0^+ + \int_0^1 E_{\hat{y}(z),z}(g^*(f^{\hat{y}(z),z} - f^0))\,ds. \tag{5.1}$$

From Theorem 4.7 the integrand above is always nonnegative, so for all $z \in \mathcal{N}$

$$P(\hat{y}(z), z) \geqq W_0^+.$$

By the definition of $y^*(z)$,

$$P(y^*(z), z) \geqq P(\hat{y}(z), z),$$

so from the definition of W_0^+,

$$\inf_z P(y^*(z), z) = \inf_z P(\hat{y}(z), z) = W_0^+.$$

Remarks 5.2. It follows from property (ii) of Definition 2.1 and the compactness of the control spaces Y and Z that, for each (t, x), $f(t, x, \hat{y}(z), z)$ is a continuous function of $z \in Z$. The selection theorem of [1] therefore implies the existence of a control $\hat{z} \in \mathcal{N}$ such that

$$g^* f(t, x, \hat{y}(\hat{z}), \hat{z}) \leqq g^* f(t, x, \hat{y}(z), z) \quad \text{a.e.}$$

for all $z \in \mathcal{N}$.

THEOREM 5.3.

$$P(\hat{y}(\hat{z}), \hat{z}) = W_0^+.$$

Proof. Write

$$\begin{aligned}\phi(x(1)) &= W_0^+ + \int_0^1 g^*(dx - f^0\,ds)\\ &= W_0^+ + \int_0^1 g^*(dx - f^{\hat{y}(z),z}\,ds)\\ &\qquad + \int_0^1 g^*(f^{\hat{y}(z),z} - f^{\hat{y}(\hat{z}),\hat{z}})\,ds + \int_0^1 g^*(f^{\hat{y}(\hat{z}),\hat{z}} - f^0)\,ds.\end{aligned}$$

Then taking expectations with respect to $\mu_{\hat{y}(z),z}$ we have

$$P(\hat{y}(z), z) \geqq W_0^+ + E_{\hat{y}(z),z} A_1^{\hat{z}}, \tag{5.2}$$

where

$$A_t^{\hat{z}} = \int_0^t g^*(f^{\hat{y}(\hat{z}),\hat{z}} - f^0)\, ds.$$

From Theorem 4.7,

$$A_1^{\hat{z}} \geqq 0 \quad \text{a.s.} \tag{5.3}$$

Suppose $\{z_n\}$ is a sequence in $\mathcal{N}$ such that $P(\hat{y}(z_n), z_n) \downarrow W_0^+$. Then from (5.2)

$$E_{\hat{y}(z_n),z_n}[A_1^{\hat{z}}] \to 0. \tag{5.4}$$

Because the set of densities is weakly compact there is a function h such that $\rho_0^1(\hat{y}(z_n), z_n)$ converges to $\rho_0^1(h) = \rho$, and $\rho > 0$ a.s. In particular, for any positive integer N

$$E_{\hat{y}(z_n),z_n}[A_1^{\hat{z}} \wedge N] = E[\rho_0^1(\hat{y}(z_n), z_n)(A_1^{\hat{z}} \wedge N)]$$

converges to $E[\rho(A_1^{\hat{z}} \wedge N)]$. In view of (5.3) and (5.4) this implies that

$$A_1^{\hat{z}} \wedge N = 0 \quad \text{a.s.}$$

and hence that

$$A_1^{\hat{z}} = 0 \quad \text{a.s.}$$

Recalling (5.2), this shows that

$$\begin{aligned} W_0^+ &= P(\hat{y}(\hat{z}), \hat{z}) \\ &\leqq P(\hat{y}(z), z), \quad \text{for all } z \in \mathcal{N}, \end{aligned}$$

so that the control $\hat{z}$ obtained by minimizing the Hamiltonian as in (5.2) is the optimal strategy for the minimizing player to announce.

REFERENCES

[1] V. E. BENEŠ, *Existence of optimal strategies based on specific information for a class of stochastic decision problems*, SIAM J. Control, 8 (1970), pp. 179–188.

[2] M. H. A. DAVIS, *On the existence of optimal policies in stochastic control*, SIAM J. Control, 11 (1973), pp. 587–594.

[3] ———, *Martingale methods in stochastic control*, in Stochastic Control Theory and Stochastic Differential Systems, M. Kohlmann, ed., Lecture Notes in Control and Information Sciences 16, Springer-Verlag, Berlin, 1979.

[4] M. H. A. DAVIS AND P. P. VARAIYA, *Dynamic programming conditions for partially observable stochastic systems*, SIAM J. Control, 11 (1973), pp. 226–261.

[5] T. E. DUNCAN AND P. P. VARAIYA, *On the solutions of a stochastic control system*, SIAM J. Control, 9 (1971), pp. 354–371.

[6] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Part I*, Interscience, New York, 1958.

[7] R. J. ELLIOTT, *The existence of value in stochastic differential games*, SIAM J. Control, 14 (1976), pp. 85–94.

[8] ———, *The optimal control of a stochastic system*, this Journal, 15 (1977), pp. 756–778.

[9] ———, *The existence of optimal strategies and saddle points in stochastic differential games*, in Differential Games and Applications, P. Hagedorn, ed., Lecture Notes in Control and Information Sciences 3, Springer-Verlag, Berlin, 1977.

[10] R. J. ELLIOTT AND M. KOHLMANN, *The variational principle and stochastic optimal control*, Stochastics, 3 (1980), pp. 229–241.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 4, July 1981

0363-0129/81/1904-0010 $01.00/0

ON THE EXISTENCE OF OBSERVABLE SINGLE-OUTPUT SYSTEMS OF A SIMPLE TYPE*

MARTIN PHILIP BENDSØE†

Abstract. It is shown that on every paracompact, smooth (real analytic) manifold M of finite dimension there exists an observable control system with a smooth (analytic) vector field as the dynamics and a smooth (analytic) real function as the output. The result is a generalization of a similar result for Euclidean space, easily obtained from the observability rank condition for linear systems.

1. Introduction.

In this paper we shall consider systems of the form

$$\Sigma: \quad \begin{aligned} \frac{dx}{dt} &= X(x), \\ y &= f(x), \end{aligned}$$

on a paracompact smooth manifold M of dimension n, where X is a smooth vector field and $f: M \to \mathbb{R}^k$ a smooth function. Throughout the paper "smooth" stands for differentiability of class C^∞ and "paracompactness" includes the Hausdorff axiom. Although no control variables are present, we think of the system as a simple type of control system. In that spirit $f: M \to \mathbb{R}^k$ is an output function with k outputs. As usual we say that the system is observable if two different points in M, after some positive time, have flowed, along the trajectories of X, to points with different output values. Using the well-known rank condition for observability of linear systems on Euclidean space E^n [4, p. 61], it is easy to construct an observable single-output system of the form Σ on E^n, for each $n \in \mathbb{N}$ (Lemma 1). The purpose of the present paper is to prove the following nonlinear version of this result.

THEOREM 1. *Let M denote a paracompact n-dimensional, smooth (analytic) manifold. Then there exist a smooth (analytic) vector field X on M, and a smooth (analytic) real valued function $f: M \to \mathbb{R}$, such that the system*

$$\Sigma: \quad \frac{dx}{dt} = X(x), \qquad y = f(x)$$

is observable.

The proof of the theorem in the smooth case is based on the following idea: If X is the gradient field of $-\tilde{f}$ (with respect to a suitable Riemannian metric), where $\tilde{f}$ is a proper Morse function, X can be used to push points in M towards the critical points of $\tilde{f}$; it is then possible, using a theorem of Whitney, to construct functions defined locally around the critical points, so that these functions when added to $\tilde{f}$ provide a function f for which the pair (X, f) has the desired property.

Theorem 1 is a sort of observability result analogous to the following theorem of N. Levitt and H. J. Sussmann [2].

THEOREM 2. *On every connected, paracompact, n-dimensional, smooth (analytic) manifold M, there exists a completely controllable pair $\{X, Y\}$ of smooth (analytic) vector fields.*

Here "completely controllable" means that any two points in M can be connected by a continuous curve that piecewise is an integral curve for one of the vector fields.

* Received by the editors April 2, 1980.
† Department of Mathematics, Technical University of Denmark, DK-2800 Lyngby, Denmark.

2. Some definitions. In the following, M denotes a paracompact, smooth manifold of dimension n. For a smooth vector field X on M we denote by $\{X_t\}$ the one-parameter family of local diffeomorphisms generated by X. For each point $x \in M$ $t \to X_t(x)$ is therefore the maximal integral curve for X through x. If $\{\varphi_i : U_i \to \mathbb{R}^k | i \in I\}$ is a family of smooth functions defined on open sets U_i in M, we say that the family separates two different points $x, x' \in M$ if there exists an index $i \in I$, so that $x, x' \in U_i$ and $\varphi_i(x) \neq \varphi_i(x')$. If the family separates all pairs of different points in M we call it a separating family of functions.

A system of the form Σ as defined in § 1 is said to be *observable*, if $\{f \circ X_t | t \geqq 0\}$ is a separating family of functions on M.

3. The linear case.

LEMMA 1. *For any dimension $n \geqq 1$ there exists a real $n \times n$ matrix $\mathbf{A}$ and a real $1 \times n$ matrix $\mathbf{C}$, so that the system*

$$\frac{dx}{dt} = \mathbf{A}x, \qquad y = \mathbf{C}x$$

defined on Euclidean n-space E^n is observable ($x \in E^n$ is considered a column matrix).

Proof. Let $\mathbf{A} = \{a_{ij}\}$ be the real $n \times n$ matrix given by

$$a_{ij} = \begin{cases} 1 & \text{for } i = j-1, \\ 0 & \text{else,} \end{cases}$$

and let $\mathbf{C} = (1, 0, \cdots, 0)$. By a simple computation it follows that $\mathbf{CA}^{p-1}$ for $1 \leqq p \leqq n$ has the form $(0, \cdots, 0, 1, 0, \cdots, 0)$ where the number 1 appears in the pth place. Thus the rank of the matrix

$$\begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \vdots \\ \mathbf{CA}^p \\ \vdots \\ \mathbf{CA}^{n-1} \end{bmatrix}$$

is n, and the lemma follows from the observability rank condition for linear systems (see, e.g., [4]).

4. Proof of the general (smooth) case. Let $f : M \to \mathbb{R}$ be a smooth Morse function on M, so that:

(i) For every real a, the set $M_a = \{x \in M | f(x) \leqq a\}$ is compact.

(ii) If c', c'' are critical points with $c' \neq c''$, then $f(c') \neq f(c'')$.

It is proved in [3, Cor. 6.7] that there exists a Morse function for which (i) holds, and this function is easily modified so that (ii) is also satisfied. The conditions imply that the critical points can be arranged in a sequence $c_0, c_1, c_2, \cdots$ with $f(c_0) < f(c_1) < \cdots$. By the Morse lemma [3, Lemma 2.2] and (ii) we can for every c_j choose a coordinate chart $\{x_1^j, \cdots, x_n^j\}$ defined on a neighborhood U_j of c_j, such that

$$x_1^j(c_j) = \cdots = x_n^j(c_j) = 0;$$

$$f(x) = f(c_j) - (x_1^j)^2 - \cdots - (x_{\lambda_j}^j)^2 + (x_{\lambda_j+1}^j)^2 + \cdots + (x_n^j)^2 \text{ for } x \in U_j,$$

U_j is mapped diffeomorphically by $\{x_1^j, \cdots, x_n^j\}$ onto the ball of radius α_j; and $f(x) < f(x')$ for $x \in U_j$, $x' \in U_{j+1}$. λ_j is the index of the critical point c_j.

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 $(x_1^j)^2+\cdots+(x_n^j)^2<\beta_j^2$. By standard methods using a partition of unity construct a Riemannian metric g on M so that

$$g\left(\frac{\partial}{\partial x_k^j},\frac{\partial}{\partial x_l^j}\right)\equiv\delta_{kl},\quad \text{on } V_j,$$

and let $X=-\text{grad } f$, where the gradient is taken with respect to the chosen metric. X is our choice of vector field.

It follows from condition (i) for f, that $X_t(x)$ is defined for all $x\in M$ and all times $t\geqq 0$, and moreover that the limit $\lim_{t\to\infty} X_t(x)$ exists for every $x\in M$ and is a critical point of f. We let, for $j=0,1,\cdots$, $C_j=\{x\in M|\lim_{t\to\infty} X_t(x)=c_j\}$. If in V_j we work with the coordinates $\{x_1^j,\cdots,x_n^j\}$, we have

$$X=\sum_1^{\lambda_j} 2x_i^j\frac{\partial}{\partial x_i^j}-\sum_{\lambda_j+1}^{n} 2x_i^j\frac{\partial}{\partial x_i^j},$$

and consequently,

$$X_t(x)=(x_1^j e^{2t},\cdots,x_{\lambda_j}^j e^{2t},x_{\lambda_j+1}^j e^{-2t},\cdots,x_n^j e^{-2t}),$$

so that $C_j\cap V_j$ is the set of points in V_j for which $x_1^j=\cdots=x_{\lambda_j}^j=0$.

Finally, for a point $x\in C_j\cap V_j$,

$$f\circ X_t(x)=e^{-4t}[(x_{\lambda_j+1}^j)^2+\cdots+(x_n^j)^2]+f(c_j).$$

Thus it is clear that $\{f\circ X_t|t\geqq 0\}$ is a family of real functions with the following properties:

(a) For $x\in C_i$, $y\in C_j$, $i\neq j$, there exists a $T\geqq 0$ such that $X_T(x)\in V_i$ and $X_T(t)\in V_j$, and such that $f\circ X_T(x)\neq f\circ X_T(y)$.

(b) For x, $y\in C_j$ there exists a $T\geqq 0$ such that $x'=X_T(x)$ and $y'=X_T(y)$ both lie in $C_j\cap V_j$. Then $f(x')=f(y')$ if and only if $f\circ X_t(x')=f\circ X_t(y')$ for all $t\geqq 0$, if and only if

$$(x_{\lambda_j+1}^j(x'))^2+\cdots+(x_n^j(x'))^2=(x_{\lambda_j+1}^j(y'))^2+\cdots+(x_n^j(y'))^2.$$

Thus $\{f\circ X_t|t\geqq 0\}$ separates many points in M and we only have to be concerned with points in the sets $C_j\cap V_j$, $j=0,1,\cdots$. We will now, on every V_j, construct functions $f_j\colon V_j\to\mathbb{R}$, where f_j in the coordinates $\{x_1^j,\cdots,x_n^j\}$ is analogous to the following function g defined on the open unit ball in $\mathbb{R}^n$.

To construct g first define for each natural number $p\in\mathbb{N}$ the following sets in $\mathbb{R}$:

$$A_p=[\alpha_p;\beta_p],\qquad \alpha_p=\frac{1}{2p+1},\qquad \beta_p=\frac{1}{2p},$$

$$B_p=\left[\frac{1}{2p};\frac{1}{2p-1}\right],$$

$$\tilde{A}_p^i=\left[\alpha_p+\frac{\beta_p-\alpha_p}{2n-1}(2i-1);\alpha_p+\frac{\beta_p-\alpha_p}{2n-1}2i\right],\qquad i=1,\cdots,n-1,$$

$$A_p^i=\left]\alpha_p+\frac{\beta_p-\alpha_p}{2n-1}(2i-2);\alpha_p+\frac{\beta_p-\alpha_p}{2n-1}(2i-1)\right[,\qquad i=1,\cdots,n.$$

Then let

$$A=]-\infty,0]\cup\left(\bigcup_1^{\infty}B_p\right)\cup\left(\bigcup_1^{\infty}\left(\bigcup_1^{n-1}\tilde{A}_p^i\right)\right)\cup[1,\infty[.$$

Now A is closed, and by a theorem of Whitney (see Bröcker [1, p. 24]), there exists a smooth, nonnegative function $\tilde{g}: \mathbb{R} \to \mathbb{R}$, so that A is the zero set for $\tilde{g}$. The $\tilde{g}$ could be called a multi-ripple bump-function. See Fig. 1.

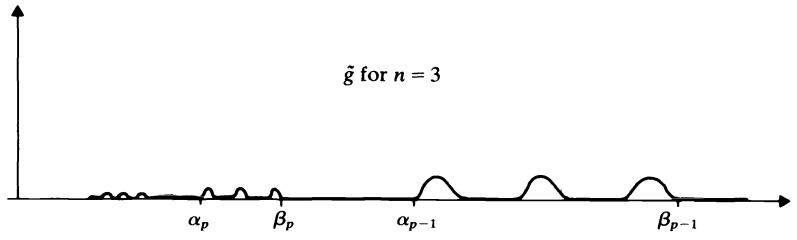


FIG. 1

In coordinates $\{y_1, \cdots, y_n\}$ on $\mathbb{R}^n$ and the usual norm $\|y\|^2 = y_1^2 + \cdots + y_n^2$, the function $g: \mathbb{R}^n \to \mathbb{R}$ is defined as

$$g(y) = \begin{cases} y_i \cdot \tilde{g}(\|y\|^2) & \text{if } \|y\|^2 \in A_p^i, \quad \text{some } p \in \mathbb{N}, \text{ some } i = 1, \cdots, n. \\ 0 & \text{elsewhere.} \end{cases}$$

From the construction of $\tilde{g}$ it follows that g is smooth.

As the sets V_j in M are disjoint it is possible to add all the functions f_j, $j = 0, 1, \cdots$, to the function f, and the resulting function $\tilde{f}$ gives us, in conjunction with X, an observable system. The form of the f_j's ensures that the family $\{\tilde{f} \circ X_t | t \geqq 0\}$ separates at least the same points as the family $\{f \circ X_t | t \geqq 0\}$. Points x, y in a set $C_j \cap V_j$, for which $(x^i_{\lambda_j+1}(x))^2 + \cdots + (x^i_n(x))^2 = (x^i_{\lambda_j+1}(y))^2 + \cdots + (x^i_n(y))^2$ can now be separated, since the curves $t \to X_t(x)$ and $t \to X_t(y)$, $t \geqq 0$, will pass through a set in V_j where the function f_j can compare the coordinates of x and y because $X_t(x)$ and $X_t(y)$ have equal f-values (and thus the same "distance" to c_j).

5. The real analytic case. We now assume that M is a real analytic manifold. We have already shown that there exists a pair (X, f) consisting of a smooth vector field X and a smooth real function f for which the corresponding system of type Σ is observable. We want to show that the the pair can be taken to be real analytic. We shall use the fact that the set of real analytic vector fields on M is dense in the space $\Gamma(TM)$ of C^∞-vector fields on M when $\Gamma(TM)$ is given the fine C^∞-topology (for a proof, see [2, § 8]). Similarly, the set of real analytic functions on M is dense in the space $C^\infty(M, \mathbb{R})$ of real C^∞-functions on M, when $C^\infty(M, \mathbb{R})$ is given the fine C^∞-topology.

Now consider the property of observability of systems of the form Σ. A pair (X, f) in $\Gamma(TM) \times C^\infty(M, \mathbb{R})$ gives an observable system, if for all x, y in M with $x \neq y$ the map $t \to f \circ X_t(x) - f \circ X_t(y)$, $t \geqq 0$, is not identically zero. For analytic systems the condition with nonnegative time is equivalent to a condition where all times $t \in \mathbb{R}$ are allowed. Using the fact that the flow of a vector field depends continuously on the vector field (in the fine C^∞-topology), one has that the set of pairs (X, f) in $\Gamma(TM) \times C^\infty(M, \mathbb{R})$ for which the corresponding system Σ is observable is an open set in $\Gamma(TM) \times C^\infty(M, \mathbb{R})$ provided with the fine C^∞-topology. This proves the real analytic case.

Acknowledgments. The author would like to thank Henrik Pedersen, Jens Gravesen, Bodil Branner-Jørgensen and Vagn Lundsgaard Hansen for inspiring discussions.

REFERENCES

[1] TH. BRÖCKER AND L. LANDER, *Differentiable Germs and Catastrophes*, Cambridge University Press, Cambridge, 1975.

[2] N. LEWITT AND H. J. SUSSMANN, *On controllability by means of two vector fields*, this Journal, 13 (1975), pp. 1271–1281.

[3] J. MILNOR, *Morse Theory*, Princeton University Press, Princeton, NJ, 1963.

[4] W. M. WONHAM, *Linear Multivariable Control: a Geometric Approach*, second ed., Springer-Verlag, New York, 1979.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 5, September 1981

0363-0129/81/1905-0001 $01.00/0

ANALYSIS OF THE ASYMPTOTIC BEHAVIOR OF OPTIMAL CONTROL TRAJECTORIES: THE IMPLICIT PROGRAMMING PROBLEM*

C. D. FEINSTEIN† AND D. G. LUENBERGER‡

Abstract. The asymptotic behavior of the optimal trajectories of the infinite horizon control problem with discounting, is characterized by a static optimization problem. In the undiscounted case, the limit point of the optimal dynamic trajectory is the steady-state that minimizes the kernel of the objective functional. The corresponding static characterization of the limit point in the discounted case, called the implicit programming problem, is derived. The implicit programming problem is a mathematical programming problem with the special feature that part of the solution is contained in the definition of the problem. All results are achieved in the context of a sufficient maximum principle, which is shown to be equivalent to the other approaches taken in the literature to perform the dynamic analysis. The equivalence is based on convexity conditions assumed in the current dynamic theory. The class of problems that satisfy such convexity conditions is characterized in terms of a property of vector-valued mappings conceptually related to monotonicity.

1. Introduction. The objective of this paper is to characterize the asymptotic behavior of the solutions of the optimal control problem defined on an infinite horizon:

$$\text{minimize} \qquad \int_0^\infty L(x, u, t)\, dt \tag{1.1a}$$

$$\text{subject to} \qquad \dot{x}(t) = f(x(t), u(t), t), \tag{1.1b}$$

$$x(0) = x_0, \tag{1.1c}$$

$$(x(t), u(t)) \in X \times U \subseteq \mathbb{R}^n \times \mathbb{R}^m \quad \text{for each } t \in [0, \infty). \tag{1.1d}$$

The variable $x \in \mathbb{R}^n$ is the state variable, and the variable $u \in \mathbb{R}^m$ is the control variable. The set $U \subseteq \mathbb{R}^m$ may depend on $x(t)$ and t, explicitly. In that case, we shall write $U = U(x, t)$, where $U(x, t)$ is a set-valued mapping from $X \times [0, \infty)$ to $2^{\mathbb{R}^m}$, the set of all subsets of $\mathbb{R}^m$. L is a real-valued function and f is n-dimensional.

In this paper, we present a static characterization of the optimal steady-state trajectories of a subclass of problem (1.1), the optimal control problem with discounting. In this problem, a familiar model in mathematical economics, the kernel of the objective functional is $L(x, u, t) = e^{-\rho t} l(x, u)$ where ρ is the discount rate, and the system is autonomous, $f(x, u, t) = f(x, u)$.

It is known [26], [27] that the optimal trajectories of the discounted problem converge to a steady-state, under certain conditions. What is interesting about this property is that it follows from essentially static, geometric conditions about the data of the problem. We exploit this fact and characterize the optimal steady-state trajectory by a static optimization problem, the implicit programming problem. With this approach, we are able to determine the asymptotic properties of the optimal dynamic trajectories without having to solve the full dynamic problem itself.

It is often the case that precise knowledge of individual trajectories of a model is of less importance than the information that the optimal dynamic trajectory converges,

* Received by the editors February 8, 1980. This research was supported in part by the National Science Foundation under grant NSF-ENG-76-18748.

† Xerox Corporation, Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, California 94305, and Department of Engineering-Economic Systems, Stanford University, Stanford, California 94305.

‡ Department of Engineering-Economic Systems, Stanford University, Stanford, California 94305.

and converges to a particular point. Such limit points provide important information about the construction of approximate optimal trajectories. Moreover, the static characterization offers a convenient method for investigating the sensitivity of the optimal trajectory to various modelling assumptions; in particular, the sensitivity of the optimal steady-state to the discount rate is relatively easy to analyze using the implicit programming problem.

In the next section, we recall two approaches to the dynamic analysis of problem (1.1), given by the maximum principle, and the Hamiltonian dynamic system. In § 3 we present a third, equivalent characterization of the optimal dynamic trajectories that is given by sufficient conditions for dynamic optimality. The equivalence of these three approaches is a result of the basic convexity assumption invoked in the dynamic theory. This equivalence is discussed in § 4. The sufficient conditions provide a somewhat different perspective on the problem, and suggest a decomposition of the analysis into static and dynamic parts; the static aspect is interpreted as a supporting hyperplane result. In § 5 we give conditions under which the basic convexity assumption holds. In the last section, based on the decomposition perspective, we formulate the implicit programming problem and present a series of theorems that verify that the solutions to the implicit programming problem are the optimal steady-states of the optimal control problem with discounting.

2. Characterization of the optimal dynamic trajectory.

2.1. The maximum principle. The most familiar approach to the analysis of optimal control problems is based upon the necessary conditions for dynamic optimality that are expressed by the Pontryagin maximum principle [21]. The maximum principle has been extended by Halkin [12] to problems defined on an infinite horizon. In the process of that extension, Halkin proposed a relaxed concept of optimality, which is able to distinguish between trajectories even if the objective functional diverges. This extension of the maximum principle provides a set of necessary conditions that must be satisfied by a weakly overtaking-optimal trajectory. We include the following definitions for completeness, and then state the main result.

DEFINITION 2.1. A *trajectory* of problem (1.1) is a pair (x, u) such that:

(i) x is a continuous, piecewise continuously differentiable function from $[0, \infty)$ into $X \subseteq \mathbb{R}^n$;

(ii) u is a piecewise continuous function from $[0, \infty)$ into $U \subseteq \mathbb{R}^m$;

(iii) $\dot{x}(t) = f(x(t), u(t), t)$ for almost every $t \in [0, \infty)$;[1] and

(iv) $x(0) = x_0$.

DEFINITION 2.2. A trajectory (x^*, u^*) is said to be *weakly overtaking optimal* if for any feasible trajectory (x, u) and any $T \in [0, \infty)$ and any $\varepsilon > 0$ there exists a $t \geqq T$ such that

$$\int_0^t L(x^*, u^*, t)dt - \varepsilon \leqq \int_0^t L(x, u, t)\, dt,$$

or

$$\limsup_{t \to \infty} \int_0^t [L(x, u, t) - L(x^*, u^*, t)]dt \geqq 0. \tag{2.1}$$

[1] Halkin defines the words "for almost every $t \in [0, \infty)$" to mean "for all $t \in [0, \infty)$ with the possible exception of a set I such that $I \cap [a, b]$ is finite for every closed bounded interval $[a, b] \subset [0, \infty)$" [12, p. 268n]. Similarly, piecewise continuity is qualified with respect to finite sets. Thus (i) implies that x is absolutely continuous on $[0, \infty)$. We shall employ Halkin's terminology. (See [12] for further details; alternate definitions are given in [17], [5], and [2].)

Other relaxed concepts of optimality have been proposed for the infinite horizon problem. In particular, weakly overtaking optimality is a modification of the concept of *overtaking optimality*, which was employed by Gale [10] and may be expressed as

$$\liminf_{t\to\infty} \int_0^t [L(x, u, t) - L(x^*, u^*, t)]\, dt \geqq 0. \tag{2.2}$$

A review of some of these concepts may be found in [15]. Brock and Haurie [5] gave conditions for the existence of overtaking- and weakly-overtaking optimal trajectories. We shall not be directly concerned with existence theories in this paper.

THEOREM 2.1. (Maximum principle). *Let U be a closed subset of $\mathbb{R}^m$. Let (f, L) be a continuous function from $\mathbb{R}^n \times U \times [0, \infty)$ into $\mathbb{R}^n \times \mathbb{R}$ whose first derivatives with respect to the first n arguments exist and are continuous over $\mathbb{R}^n \times U \times [0, \infty)$.*

If a trajectory (x^*, u^*) is weakly overtaking-optimal for problem (1.1) then there exist a nonnegative number p_0 and a continuous, piecewise continuously differentiable function p from $[0, \infty)$ into $\mathbb{R}^n$ such that:

(i) $\|(p_0, p(0))\| = 1$;

(ii) $\dot{p}(t) = -\partial/\partial\xi[H(\xi, u^*(t), t, p_0, p(t))]_{\xi = x^*(t)}$ *for almost every* $t \in [0, \infty)$;

(iii) $H(x^*(t), u^*(t), t, p_0, p(t) \geqq H(x^*(t), u, t, p_0, p(t)) \forall u \in U, \forall t \in [0, \infty)$;

where the Hamiltonian, H, is defined as

(iv) $H(x, u, t, p_0, p(t)) = -p_0 L(x, u, t) + \langle p(t), f(x, u, t)\rangle$.

Proof. See Halkin [12]. □

It is important to note that the maximum principle for the infinite horizon problem does not contain a transversality condition describing the behavior of the costate variable, $p(t)$, as t approaches infinity. In particular the boundary condition $\lim_{t\to\infty} p(t) = \theta$ is not generally satisfied.

It is also not proper to assume that $p_0 > 0$ (hence taken as 1), in general. This is the so-called normality assumption. Bliss [4] has given necessary and sufficient conditions for normality in problems in the calculus of variations; Berkovitz [3] has given a sufficient condition for normality in the control problem formulated on a finite horizon. Conditions for normality on an infinite horizon are not known at present. However, since the theory presented in this paper is essentially a sufficiency theory, our conditions will be written for the normal case.

2.2. Convexity theory and the Hamiltonian dynamic system. Another approach to the analysis of the optimal control problem has been studied by Rockafellar ([23], [25]–[27]). The main theme of this approach is the replacement of differentiability assumptions by convexity assumptions. The trajectories of the optimal control problem are then characterized by the Hamiltonian dynamic system. In his development of the application of convexity theory to the optimal control problem (1.1), Rockafellar [27] considers the problem of Lagrange, for an infinite time horizon,

$$\text{minimize} \qquad \int_0^\infty L(t, x(t), \dot{x}(t))\, dt \tag{2.3a}$$

$$\text{subject to} \qquad x(0) = x_0. \tag{2.3b}$$

The essential assumption that is made in the analysis of (2.3) is that $L(t, \cdot, \cdot)$ is a lower semicontinuous convex function on $\mathbb{R}^n \times \mathbb{R}^n$ with values in $(-\infty, +\infty]$, not identically $+\infty$; i.e., a lower semicontinuous *proper* convex function [25, p. 24].

The relationship between this problem and the optimal control problem (1.1) is effected by the formulation of the image function, $L^*(x, v, t)$, defined by

$$L^*(x, v, t) = \begin{cases} \inf_u \{L(x, u, t) : v = f(x, u, t),\ u \in U(x, t)\} \\ +\infty \quad \text{if } x \notin X \text{ or } v \neq f(x, u, t) \forall u \in U(x, t). \end{cases} \tag{2.4}$$

The problem (2.3), in which the image $L^*(x, \dot{x}, t)$ replaces $L(t, x, \dot{x})$ is known as the deparametrized problem, in which the control parameter u has been eliminated. The effect of the control variable is felt as an infinite penalty, through the definition of $L^*(x, v, t)$ (2.4). Young [29] discusses the idea of control as a parameter in calculus of variations problems. Zachrisson [30] investigated the role of convexity theory in deparametrized problems. More recently, Goodman [11] analyzed a deparametrized formulation of the control problem.

The main results that can be achieved using this problem structure ([23], [25]) are that the optimal state-costate trajectories are solutions of the Hamiltonian dynamic system, defined by the subdifferential equations,

$$\dot{x}(t) \in \partial_p H^*(t, x(t), p(t)) \tag{2.5a}$$

and

$$\dot{p}(t) \in -\partial_x H^*(t, x(t), p(t)), \tag{2.5b}$$

with the Hamiltonian defined by the conjugacy formula

$$H^*(t, x, p) = \sup_v \{\langle v, p\rangle - L^*(t, x, v) : v \in \mathbb{R}^n\}. \tag{2.6}$$

The operator "∂_ξ" is the subdifferential operator. The subdifferential of a function at a point is the set of all subgradients of the function at that particular point. If the function happens to be differentiable at the point x, the subdifferential reduces to the gradient of the function at x [24, Thm. 25.1]. For completeness, we include

DEFINITION 2.3. A vector s is said to be a *subgradient* of a convex function h at a point x if

$$h(z) \geqq h(x) + \langle s, z - x\rangle \quad \forall z. \tag{2.7}$$

The relationship of the state-costate trajectories that satisfy the Hamiltonian dynamic system (2.5) and the state-control-costate trajectories that satisfy the necessary conditions of the maximum principle (Thm. 1.1) was discussed in [23, Example 12]. It was shown that if a state-costate trajectory is a solution to the Hamiltonian dynamic system, where l and f also satisfy the smoothness assumptions of the maximum principle, then there exists a control trajectory corresponding to the state-costate trajectory, such that all the conditions of the maximum principle are satisfied for the normal case ($p_0 = 1$). Further, in terms of the data of the control problem, the Hamiltonian H^* is equal to the optimal value Hamiltonian,

$$H^*(x, p, t) = \sup_u \{\langle p, f(x, u, t)\rangle - L(x, u, t)\}. \tag{2.8}$$

However, if the image function is indeed a lower semicontinuous proper convex function, there is yet another possible characterization of the optimal trajectories of the control problem. This third characterization is equivalent, in this case, to both the maximum principle and the Hamiltonian dynamic system, but is of a somewhat different nature, since it follows from sufficient conditions for dynamic optimality. It is discussed below.

3. A sufficient maximum principle and the support property. The next theorem is an extension of a result established by Peterson [20]. The sufficient maximum principle provides a context for all our subsequent developments. As we shall show, the more familiar characterizations of the optimal trajectories of problem (1.1) (given by the maximum principle or the Hamiltonian dynamic system), are actually equivalent to this sufficiency theorem, under the particular convexity assumptions that are generally invoked in the analysis of the optimal control problem.

THEOREM 3.1. (Sufficient maximum principle). *Let* (x^*, u^*) *be a trajectory* (*Definition* 1.1) *of problem* (1.1). *Let* p^* *be a continuous, piecewise continuously differentiable function from* $[0, \infty)$ *into* $\mathbb{R}^n$. *Define the Hamiltonian,* $H(x, u, t, p) = -L(x, u, t) + \langle p, f(x, u, t)\rangle$. *Suppose that*:

(i) $H(x^*(t), u^*(t), t, p^*(t)) + \langle \dot{p}^*(t), x^*(t)\rangle \geqq H(x, u, t, p^*(t)) + \langle \dot{p}^*(t), x\rangle \forall (x, u) \in X \times U$, *for almost every* $t \in [0, \infty)$;

(ii) $\lim_{t\to\infty} \langle p^*(t), x^*(t)\rangle$ *exists, and there holds*

$$-\infty < \lim_{t\to\infty} \langle p^*(t), x^*(t)\rangle \leqq \liminf_{t\to\infty} \langle p^*(t), x(t)\rangle < +\infty,$$

for any feasible state trajectory.

Then (x^*, u^*) *is overtaking-optimal for the optimal control problem* (1.1).

Proof. Since (x^*, u^*) is a trajectory, $\dot{x}^*(t) = f(x^*(t), u^*(t), t)$ holds for almost every $t \in [0, \infty)$. Let (x, u) be any other trajectory of problem (1.1). Then assumption (i) implies that, for any $T < +\infty$,

$$\int_0^T (-L(x^*(t), u^*(t), t) + d/dt[\langle p^*(t), x^*(t)\rangle])\, dt$$

$$\geqq \int_0^T (-L(x(t), u(t), t) + d/dt[\langle p^*(t), x(t)\rangle])\, dt.$$

Then

$$\int_0^T [L(x(t), u(t), t) - L(x^*(t), u^*(t), t)]\, dt \geqq \langle p^*(T), x(T) - x^*(T)\rangle.$$

Hence,

$$\liminf_{T\to\infty} \int_0^T [L(x, u, t) - L(x^*, u^*, t)]\, dt \geqq \liminf_{T\to\infty} \langle p^*(T), x(T) - x^*(T)\rangle$$

$$\geqq 0, \quad \text{by (ii).} \qquad \square$$

As Peterson [20] observed, no differentiability or continuity assumptions are invoked on L or f, directly, in Theorem 3.1. Moreover, condition (ii) of Definition 2.1 may be replaced by the simple inclusion, $u^*: [0, \infty) \to U$, since the continuity properties of u^* are irrelevant.

The assumption (i) of Theorem 3.1 occurs frequently in economic analysis, and is called the support property.

DEFINITION 3.1. A trajectory $(x^*(t), u^*(t))$ is said to be *supported* if there exists a continuous, piecewise continuously differentiable function $p^*: [0, \infty) \to \mathbb{R}^n$ such that

$$-L(x^*(t), u^*(t), t) + d/dt[\langle p^*(t), x^*(t)\rangle] \geqq -L(x, u, t) + \langle p^*(t), f(x, u, t)\rangle + \langle \dot{p}^*(t), x\rangle, \tag{3.1}$$

$$\forall (x, u) \in X \times U, \text{ for almost every } t \in [0, \infty).$$

The support property was defined by Gale [10] (he referred to a supported trajectory as "competitive"). More recently, Haurie [14] generalized the support property to suggest an approach to nonconvex problems of the form (1.1). Following Cass and Shell [7], the support functional

$$H(x, u, t, p^*(t)) + \langle \dot{p}^*(t), x \rangle = -L(x, u, t) + \langle p^*(t), f(x, u, t) \rangle + \langle \dot{p}^*(t), x \rangle, \tag{3.2}$$

may be interpreted as the profit rate given by the state control pair (x, u) at the "price" p^*, at time t. The sufficient maximum principle indicates the optimality of a "greedy" solution, one maximizing the profit rate at each instant and minimizing the asymptotic worth of the state variable, given by the inner product of the state and the supporting function, $\langle p^*(t), x^*(t) \rangle$.

It is evident that the concept of optimality provided by the sufficiency theorem is completely dependent upon the asymptotic behavior of the inner product $\langle p^*(T), x(T) - x^*(T) \rangle$. This suggests a decomposition of the analysis of the infinite horizon problem into separate components. One part of the analysis would determine conditions under which a trajectory is supported, and the other aspect would investigate the asymptotic properties of the inner product. We shall find this decomposition perspective useful in characterizing the optimal steady-states of the dynamic problem.

We shall now determine the source of the supporting function p^*. We show that the supporting function is the costate trajectory determined by the maximum principle, if a particular convexity assumption is satisfied. The convexity assumption we invoke is actually equivalent to the convexity assumption on the image function L^*. Moreover, under that convexity assumption, the maximum principle, the Hamiltonian dynamic system, and the support property are equivalent characterizations of the optimal dynamic trajectory.

4. The support theorem: equivalent characterizations of optimality. We relate the necessary conditions to the sufficient conditions through a convexity assumption. To motivate the assumption, we introduce some standard terminology. Suppose that the set of admissible controls, given that the system is in state x at time t, is described by the set $U(x, t)$. Then the *velocity set*, $F(x, t)$, is composed of all possible cost kernels and state velocities, as u ranges over the allowable set $U(x, t)$:

$$F(x, t) = \{(-L(x, u, t), f(x, u, t)) : u \in U(x, t)\}. \tag{4.1}$$

In his formulation of the optimal control problem, Filippov [9], in studying the system $\dot{x}(t) = f(x, u, t)$, where $u \in U(x, t)$, defined the set

$$R(x, t) = \{f(x, u, t) : u \in U(x, t)\} \tag{4.2}$$

and analyzed the contingent equation $\dot{x}(t) \in R(x, t)$ (which, by Filippov's lemma, is equivalent to the original dynamic system). Filippov made the important convexity assumption: $R(x, t)$ is convex for every pair (x, t). This assumption leads to several important existence results ([9], [28], [17]).

Baum [2] made assumptions about the set

$$Q(x, t) = \{(z^0, z) : z^0 \geqq L(x, u, t),\ z = f(x, u, t),\ u \in U(x, t)\}, \tag{4.3}$$

a set that is related to the velocity set $F(x, t)$. Baum required that $Q(x, t)$ be convex and closed for each pair (x, t) in order to achieve existence results for the infinite horizon problem.

For each state-time pair (x, t), $F(x, t)$ is a collection of points in $\mathbb{R} \times \mathbb{R}^n$ determined by all allowable controls, $u \in U(x, t)$. If, at time t, $(x^*(t), u^*(t))$ is on the optimal

trajectory, then $(-L(x^*(t), u^*(t), t), f(x^*(t), u^*(t), t))$ is the optimal "velocity" vector of the joint cost-state dynamic system. Hence, any other velocity vector in $F(x^*(t), t)$ would be suboptimal. This essentially static property admits of a characterization in terms of supporting hyperplanes to convex sets, the convex set being $Q(x^*(t), t)$ and the hyperplane defined by the normal vector $(-1, p^*(t))$. Hence, Baum's convexity assumption expresses the maximum principle as a supporting hyperplane property of the costate variable, $p^*(t)$.

We will impose a stronger convexity assumption, extending Baum's assumption.

Assumption 4.1. The set

$$\Omega(t) = \{(x, v, \gamma): x \in X, v = f(x, u, t), \gamma \geqq L(x, u, t), u \in U(x, t)\}, \tag{4.4}$$

is convex and closed for each $t \in [0, \infty)$.

An immediate consequence of Assumption 4.1 is that Baum's convexity assumption follows. That is, the set

$$\begin{aligned}\Omega(x^*(t), t) &= \{(x^*(t), v, \gamma): \gamma \geqq L(x^*(t), u, t), v = f(x^*(t), u, t), u \in U(x^*(t), t)\} \\ &= \{(x^*(t), v, \gamma): (v, \gamma) \in Q(x^*(t), t)\}\end{aligned}$$

is also convex and closed, since it is the intersection of two closed, convex sets:

$$\Omega(x^*(t), t) = \Omega(t) \cap \{(x^*(t), v, \gamma): (v, \gamma) \in \mathbb{R}^n \times \mathbb{R}\}.$$

The main result that follows from this convexity assumption is that every trajectory that satisfies the conditions of the maximum principle, or is a solution of the Hamiltonian dynamic system, is actually supported. Just as the convexity of the set $Q(x, t)$ permits the maximum principle to be characterized by a supporting hyperplane, the convexity of the set $\Omega(t)$ permits the support property (Definition 3.1) to be characterized by a supporting hyperplane defined by the costate variable $p^*(t)$. We now state and prove the support theorem.

THEOREM 4.1 (Support theorem). *Let X be a subset of $\mathbb{R}^n$. Let $U(x, t)$ be a mapping* defined on $X \times [0, \infty)$ into $2^{\mathbb{R}^m}$. *Let (f, L) be a continuous function from the set*

$$D = \{(x, u, t): x \in X, t \in [0, \infty), u \in U(x, t)\},$$

into $\mathbb{R}^n \times \mathbb{R}$, whose first derivatives with respect to the first n arguments exist and are continuous over the set D.

Let (x^*, u^*) be a trajectory (Definition 2.1) *and suppose further that there exists a continuous, piecewise continuously differentiable function p^* from $[0, \infty)$ into $\mathbb{R}^n$ such that the triple (x^*, u^*, p^*) satisfies the conditions of the maximum principle* (Theorem 2.1) *for the normal case. Suppose, in addition, that $x^*(t) \in X^0$ (i.e., an interior point) for almost every $t \in [0, \infty)$, and that the set*

$$\Omega(t) = \{(x, v, \gamma): x \in X, v = f(x, u, t), \gamma \geqq L(x, u, t), u \in U(x, t)\}$$

is convex and closed $\forall t \in [0, \infty)$.

Then the trajectory (x^*, u^*) is supported by p^*.

Proof. Since $\Omega(t)$ is a closed, convex set, it follows that

$$\Omega(x^*(t), t) = \{(x^*(t), v, \gamma): v = f(x^*(t), u, t), \gamma \geqq L(x^*(t), u, t), u \in U(x^*(t), t)\}$$

is closed and convex for each $t \in [0, \infty)$. Consider the restriction of $\Omega(x^*(t), t)$ to $\mathbb{R}^n \times \mathbb{R}$, the set

$$\omega(x^*(t), t) = \{(v, \gamma): (x^*(t), v, \gamma) \in \Omega(x^*(t), t)\}.$$

Let $H(x, u, t, p^*(t)) = -L(x, u, t) + \langle p^*(t), f(x, u, t)\rangle$ and define the hyperplane in $\mathbb{R}^n \times \mathbb{R}$,

$$\pi(x^*(t), u^*(t), t, p^*(t))$$
$$= \{(\eta, \gamma): \langle (p^*(t), -1), (\eta, \gamma)\rangle = H(x^*(t), u^*(t), t, p^*(t)), (\eta, \gamma) \in \mathbb{R}^n \times \mathbb{R}\}.$$

By the maximum principle (Thm. 2.1 (iii), the maximization of the Hamiltonian), $(\eta^*(t), \gamma^*(t)) = (f(x^*(t), u^*(t), t), L(x^*(t), u^*(t), t))$ is a boundary point of the convex set $\omega(x^*(t), t)$ and $\pi(x^*(t), u^*(t), t, p^*(t))$ is a supporting hyperplane of $\omega(x^*(t), t)$ at $(\eta^*(t), \gamma^*(t))$.

Then, the point $(x^*(t), \eta^*(t), \gamma^*(t))$ is a boundary point of the convex set $\Omega(t)$ (since all ε-spheres centered at $(x^*(t), \eta^*(t), \gamma^*(t))$ contain points of the form $(x^*(t), \eta^*(t), \gamma)$, $\gamma < \gamma^*(t)$, and hence, not in $\Omega(t)$). By the support theorem for convex sets [19, Thm. 2, p. 133] (note that for finite dimensional space, the requirement that the supported set contain an interior point may be eliminated) there exists a closed hyperplane containing the point $(x^*(t), \eta^*(t), \gamma^*(t))$ such that $\Omega(t)$ is on one side of the hyperplane. Let the normal to the hyperplane be given by the functional $(-r(t), p^*(t), -1): \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$. Hence, the hyperplane

$$\Pi(x^*(t), u^*(t), t, p^*(t))$$
$$= \{(\xi, \eta, \gamma): \langle (-r(t), p^*(t), -1), (\xi, \eta, \gamma)\rangle$$
$$= \langle -r(t), x^*(t)\rangle + H(x^*(t), u^*(t), t, p^*(t)), (\xi, \eta, \gamma) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}\}$$

contains $(x^*(t), \eta^*(t), \gamma^*(t))$ and supports the convex set $\Omega(t)$. Therefore, the functional $\langle (-r(t), p^*(t), -1), (\xi, \eta, \gamma)\rangle$ is maximized over the set $(\xi, \eta, \gamma) \in \Omega(t)$ at the point $(x^*(t), \eta^*(t), \gamma^*(t))$. Equivalently, we consider the maximization of the functional

$$-L(x, u, t) + \langle p^*(t), f(x, u, t)\rangle + \langle -r(t), x\rangle,$$

subject to the constraints $x \in X$, $u \in U(x, t)$. The maximum is attained at $(x^*(t), u^*(t))$. Clearly, the maximization of the Hamiltonian (Thm. 2.1 (iii)) is a necessary condition for this maximization. Another necessary condition follows from the differentiability assumptions on the function (f, L) and the interior point assumption, $x^*(t) \in X^0$. Thus the gradient, with respect to x, of the functional vanishes at $(x^*(t), u^*(t))$. We have

$$\partial/\partial\xi[-L(\xi, u^*(t), t) + \langle p^*(t), f(\xi, u^*(t), t)\rangle]_{\xi = x^*(t)} - r(t) = \theta.$$

By the costate differential equation (Thm. 2.1 (ii)), we conclude that $-r(t) = \dot{p}^*(t)$, for almost every $t \in [0, \infty)$. Therefore,

$$-L(x^*(t), u^*(t), t) + \frac{d}{dt}\langle p^*(t), x^*(t)\rangle \geqq -L(x, u, t) + \langle p^*(t), f(x, u, t)\rangle + \langle \dot{p}^*(t), x\rangle,$$
$$\forall (x, u), \quad x \in X, u \in U(x, t),$$
$$\text{for almost every } t \in [0, \infty).$$

(A similar result was given in [14].) □

We will now show that the convexity assumption on the set $\Omega(t)$ is equivalent to assuming that the image function is a lower semi-continuous convex function. This follows from the fact that $\Omega(t)$ is the epigraph of the image function.

LEMMA 4.1. *The set* $\Omega(t) = \{(x, v, \gamma): x \in X, v = f(x, u, t), \gamma \geqq L(x, u, t), u \in U(x, t)\}$ *is convex and closed if and only if the function*

$$L^*(x, v, t) = \begin{cases} \inf_u \{L(x, u, t): v = f(x, u, t), u \in U(x, t)\} \\ +\infty \quad \text{if } x \notin X \quad \text{or} \quad v \neq f(x, u, t) \quad \forall u \in U(x, t), \end{cases} \tag{4.5}$$

is convex and lower semi-continuous (hence, with effective domain the convex set

$$\operatorname{dom} L^*(t) = \{(x, v): x \in X, \exists u \in U(x, t) \ni v = f(x, u, t)\}).$$

Proof. The epigraph of $L^*(x, v, t)$, for fixed $t \in [0, \infty)$, is the set

$$\begin{aligned}\operatorname{epi} L^*(t) &= \{(x, v, \gamma): \gamma \geqq L^*(x, v, t), (x, v) \in \operatorname{dom} L^*(t)\} \\ &= \{(x, v, \gamma): \gamma \geqq \inf_u \{L(x, u, t): v = f(x, u, t), u \in U(x, t)\}\} \\ &= \{(x, v, \gamma): \gamma \geqq L(x, u, t), v = f(x, u, t), u \in U(x, t)\} \\ &= \Omega(t).\end{aligned}$$

By definition, a function is convex if its epigraph is convex. The equivalence between lower semi-continuity of a function and closedness of the epigraph of the function is given by Rockafellar [24, Thm. 7.1.]. □

Lemma 4.1 implies the conclusion that every trajectory generated by the Hamiltonian dynamic system is supported. An indirect proof of this claim follows from Theorem 4.1 and the equivalence between the Hamiltonian dynamic system and the maximum principle, as noted in § 2. Directly, we observe that the Hamiltonian dynamic system and the Euler-Lagrange conditions for the image function L^*, the subdifferential equations $(\dot{p}^*(t), p^*(t)) \in \partial L^*(x^*(t), \dot{x}^*(t), t)$, for almost every t, are equivalent. The Euler–Lagrange conditions are themselves equivalent to the support property if L^* is convex. (The equivalence between the Hamiltonian dynamic system and the Euler–Lagrange conditions is based on the conjugacy properties of L^* and H^*, which we state explicitly in § 6 (see Lemma 6.1).)

Therefore, if Assumption 4.1 holds, we may characterize the optimal trajectories by the support property. The importance of this convexity assumption may be understood by observing the prominent place it occupies in the literature; the dynamic theory of Rockafellar ([23], [25]–[27]), the existence theory of Brock and Haurie [5], and the turnpike theory of Haurie [15] all are based on this essential assumption. It is significant to note that the difference between Assumption 4.1 and Baum's assumption, the convexity and closedness of $Q(x, t)$ for each (x, t), is that the latter is equivalent to assuming that $L^*(x, v, t)$ is convex as a function of v only. The joint variation in (x, v) is unspecified under Baum's assumption. Conditions under which Assumption 4.1 holds are presented in the immediately following section.

5. Convexity assumptions and the *M*-property. This section presents conditions under which the set $\Omega(t)$ (4.4) is closed and convex; or equivalently, (Lemma 4.1) that the image function $L^*(x, v, t)$ (4.5) is convex and lower semi-continuous. To proceed, we require the following assumptions:

Assumption 5.1. The set $X \subseteq \mathbb{R}^n$ is convex.

Assumption 5.2. The mapping $U: X \times [0, \infty) \to 2^{\mathbb{R}^m}$ satisfies the convexity property: if $u_i \in U(x_i, t)$, $x_i \in X$, $i = 1, 2$, then, for each $t \in [0, \infty)$, $\lambda u_1 + (1-\lambda) u_2 \in U(\lambda x_1 + (1-\lambda) x_2, t), \forall \lambda \in [0, 1]$.

Observe that Assumption 5.2 is satisfied if $U(x, t) = U(t)$, for all $x \in X$, with $U(t)$ a convex set for each $t \in [0, \infty)$. Assumption 5.2 also holds if $U(x, t) = \{u: u \in \mathbb{R}^m, g(x, u, t) \leqq \theta\}$, where $g: \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}^s$ is a convex mapping (jointly in (x, u)) for each $t \in [0, \infty)$.

Assumption 5.3. The non-empty set $\Delta(t) = \{(x, v): x \in X, v = f(x, u, t), u \in U(x, t)\}$ is convex and closed, for each $t \in [0, \infty)$.

Assumption 5.3 is stronger than Filippov's convexity assumption. Clearly, $\Delta(t)$ convex implies that $R(x, t) = \{f(x, u, t): u \in U(x, t)\}$ is convex for each $(x, t) \in X \times [0, \infty)$.

This assumption may be interpreted as a condition on the inverse mapping, $f^{-1}(v; x, t)$, that assigns to each v in $\mathbb{R}^n$ the set of points ω in $\mathbb{R}^m$ such that $v = f(x, \omega, t)$. If $v \in R(x, t)$, then there exists at least one control, u, in $U(x, t)$ such that $v = f(x, u, t)$. That is, for $v \in R(x, t)$ the intersection $\{f^{-1}(v; x, t) \cap U(x, t)\} \neq \varnothing$. Assumption 5.3 requires that for all $\lambda \in [0, 1]$, for all $v_i \in R(x_i, t)$, $i = 1, 2$, the intersection of the sets $f^{-1}(\lambda(x_1, v_1, t) + (1-\lambda)(x_2, v_2, t))$ and $U(\lambda x_1 + (1-\lambda)x_2, t)$ is not empty.

Observe that the set $\Delta(t)$ is the effective domain of the function $L^*(x, v, t)$, defined by (4.5), for each $t \in [0, \infty)$. One might conjecture that since the domain is assumed to be convex, the convexity of $L^*(x, v, t)$ would follow from the convexity of the function $L(x, u, t)$, for each $t \in [0, \infty)$. This is not the case. In fact, more than convexity of $L(x, u, t)$ is required to assert the convexity of the function $L^*(x, v, t)$ or, equivalently, the convexity of the set $\Omega(t)$. The technical problem is the presence of the equality constraint in the definition of $\Omega(t)$ and the minimization problem that defines $L^*(x, v, t)$. Indeed if $\Omega(t)$ were defined by the inequality $v \geqq f(x, u, t)$ instead of the equality, convexity would follow from the assumption that (f, L) is a convex mapping for each $t \in [0, \infty)$. The actual definition of $\Omega(t)$ necessitates a further assumption.

Functions defined by equality-constrained minimization operations have been studied by Rockafellar [24]; he called them images, hence our use of the term to describe the function L^*. For a linear transformation A from $\mathbb{R}^n$ to $\mathbb{R}^m$, the image of the convex function h under A is defined by $(Ah)(y) = \inf\{h(x): Ax = y\}$. The image (Ah) is easily shown to be convex [24, Thm. 5.7]. For given (x, t), the function $L^*(x, v, t)$ is then the image of $L(x, \cdot, t): U(x, t) \to \mathbb{R}$ under $f(x, \cdot, t): U(x, t) \to \mathbb{R}^n$.

For nonlinear mappings A, no general theorems relating to the convexity of the image (Ah) are known. We propose a condition to be satisfied by the function f that will aid in the characterization of images. The condition is called the M-property, since it is a property conceptually related to the monotonicity of mappings. As shown in Lemma 5.1, below, the M-property enables equality constraints to be treated as inequality constraints for certain nonlinear programming problems.

DEFINITION 5.1 A mapping $f: D \subseteq \mathbb{R}^s \to \mathbb{R}^n$ is said to satisfy the *M-property* on D if, for any v in the range of f and for any $x \in D$ such that $f(x) \geqq v$, there exists $y \in D$, $y \leqq x$, such that $f(y) = v$.

The M-property is concerned with positive cones. Let v belong to the range of f; i.e., $v \in f(D)$. At each element $y \in D$ that is in the level set $\{y: y \in D, f(y) = v\}$ erect a positive cone. The M-property is satisfied if all $x \in D$ such that $f(x) \geqq v$ are contained in the union of all positive cones with vertices in the level set.

A simple example of a mapping that satisfies the M-property is given by

$$f: \mathbb{R}^2 \to \mathbb{R}, \quad f(x_1, x_2) = x_1^2 + x_2^2, \quad D = \{(x_1, x_2): x_i \geqq 0, i = 1, 2\}.$$

The level sets are quarter circles in the positive quadrant of the $x_1 - x_2$ plane, centered about (0, 0). The point $y = (0, 0)$ validates the M-property for $v = 0$. It is easy to see that the property holds for any $v > 0$.

The first result that can be proven using the M-property is an equivalence relationship between equality- and inequality-constrained nonlinear programming problems.

LEMMA 5.1 *Let $l: D \subseteq \mathbb{R}^s \to \mathbb{R}$ be nondecreasing on D. That is, if $x_1, x_2 \in D$, $x_1 \leqq x_2$, then $l(x_1) \leqq l(x_2)$. Let $f: D \subseteq \mathbb{R}^s \to \mathbb{R}^n$ satisfy the M-property on D. Suppose that the set*

$\{x: x \in D,\ f(x) = \theta\} \neq \varnothing$. *Then the following problems have equal optimal objective values*:

$$\begin{array}{lll} \min l(x), & \text{and} & \min l(x), \\ \text{s.t. } f(x) \geqq \theta, & & \text{s.t. } f(x) = \theta, \\ x \in D, & & x \in D, \end{array}$$

or

$$\min \{l(x): f(x) \geqq \theta,\ x \in D\} = \min \{l(x): f(x) = \theta,\ x \in D\}.$$

Proof. Since $f: D \to \mathbb{R}^n$ satisfies the M-property, for all $x \in D$ such that $f(x) \geqq \theta$, there exists $y \in D$, $y \leqq x$, $f(y) = \theta$. Since l is nondecreasing on D, $l(y) \leqq l(x)$. □

Using Lemma 5.1., we may now prove that the image $L^*(x, v, t)$ is convex and lower semicontinuous in (x, v) for each $t \in [0, \infty)$.

PROPOSITION 5.1. *Let Assumptions* 5.1, 5.2 *and* 5.3 *hold for* X, U, *and* f. *Let* $D = \{(x, u, t): x \in X,\ t \in [0, \infty),\ u \in U(x, t)\} \subseteq \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}$.

Suppose further that:

(i) $L: D \to \mathbb{R}$ *is convex and continuous with respect to* (x, u) *and non-decreasing with respect to* u, *for each* $t \in [0, \infty)$;

(ii) $f: D \to \mathbb{R}^n$ *is a continuous concave mapping with respect to* (x, u) *for each* $t \in [0, \infty)$; *and*

(iii) *for each* $(x, t) \in X \times [0, \infty)$ *the function* $f(x, \cdot, t): U(x, t) \to \mathbb{R}^n$ *satisfies the* M-*property on* $U(x, t)$.

Then $L^*(x, v, t)$ *is convex and lower semicontinuous in* (x, v) *for each* $t \in [0, \infty)$.

Proof. Fix $t \in [0, \infty)$. Let $(x_i, v_i) \in \Delta(t)$. That is, there exist $u_i \in U(x_i, t)$ such that $v_i = f(x_i, u_i, t)$, $i = 1, 2$. By definition,

$$\begin{aligned} &L^*(\lambda(x_1, v_1) + (1-\lambda)(x_2, v_2), t) \\ &\quad = \inf_u \{L(\lambda x_1 + (1-\lambda)x_2, u, t): \lambda v_1 + (1-\lambda)v_2 \\ &\qquad\qquad = f(\lambda x_1 + (1-\lambda)x_2, u, t),\ u \in U(\lambda x_1 + (1-\lambda)x_2, t)\}. \end{aligned}$$

By Assumption 5.3, the set

$$\{u: u \in U(\lambda x_1 + (1-\lambda)x_2, t),\ \lambda v_1 + (1-\lambda)v_2 = f(\lambda x_1 + (1-\lambda)x_2, u, t)\} \neq \varnothing.$$

By Assumption 5.1, $\lambda x_1 + (1-\lambda)x_2 \in X$.

Since the functions

$$L(\lambda x_1 + (1-\lambda)x_2, \cdot, t): U(\lambda x_1 + (1-\lambda)x_2, t) \to \mathbb{R},$$

and

$$f(\lambda x_1 + (1-\lambda)x_2, \cdot, t): U(\lambda x_1 + (1-\lambda)x_2, t) \to \mathbb{R}^n,$$

satisfy the assumptions of Lemma 5.1, and the feasible set is not empty, the equality constraints may be replaced by inequalities. Hence

$$\begin{aligned} &L^*(\lambda(x_1, v_1) + (1-\lambda)(x_2, v_2), t) \\ &\quad = \inf_u \{L(\lambda x_1 + (1-\lambda)x_2, u, t): \lambda v_1 + (1-\lambda)v_2 \\ &\qquad\qquad \leqq f(\lambda x_1 + (1-\lambda)x_2, u, t),\ u \in U(\lambda x_1 + (1-\lambda)x_2, t)\}. \end{aligned}$$

Let $u = \lambda\omega_1 + (1-\lambda)\omega_2$, with $\omega_i \in U(x_i, t)$, $i = 1, 2$. By Assumption 5.2, $u \in U(\lambda x_1 + (1-\lambda)x_2, t)$. However, restricting u to be of this form increases the value of the infimum. Moreover, by convexity of L and concavity of f, it follows that

$$L^*(\lambda(x_1, v_1) + (1-\lambda)(x_2, v_2), t)$$
$$\leqq \lambda \inf_{\omega_1} \{L(x_1, \omega_1, t): v_1 \leqq f(x_1, \omega_1, t), \omega_1 \in U(x_1, t)\}$$
$$+ (1-\lambda) \inf_{\omega_2} \{L(x_2, \omega_2, t): v_2 \leqq f(x_2, \omega_2, t), \omega_2 \in U(x_2, t)\}.$$

By Lemma 5.1, we can replace the inequalities by equality constraints since the sets $\{\omega : v_i = f(x_i, \omega, t), \omega \in U(x_i, t)\}$ are not empty, $i = 1, 2$. Hence,

$$L^*(\lambda(x_1, v_1) + (1-\lambda)(x_2, v_2), t) \leqq \lambda L^*(x_1, v_1, t) + (1-\lambda)L^*(x_2, v_2, t).$$

The lower semicontinuity of L^* follows from Assumption 5.3 and the continuity of (f, L). □

Thus, Proposition 5.1 provides a set of conditions that are sufficient for the conclusion that the set $\Omega(t)$ is convex and closed for each $t \in [0, \infty)$, or equivalently, that the image function is convex and lower semicontinuous. That L^* is a proper convex function follows if L is never $\pm\infty$.

6. The static characterizations of the asymptotic behavior of the optimal trajectories of the discounted optimal control problem: the implicit programming problem.

6.1 The optimal control problem with discounting. We now apply the theory to a subclass of problems (1.1) that is of interest in mathematical economics, the so-called problem with discounting. Here a discount rate ρ is introduced and the kernel of the objective functional becomes $L(x, u, t) = e^{-\rho t} l(x, u)$. The dynamic system will be autonomous, $f(x, u, t) = f(x, u)$. The basic control problem (1.1) then becomes the discounted problem, with discount rate ρ:

(6.1a) minimize $$\int_0^\infty e^{-\rho t} l(x, u)\, dt$$

(6.1b) subject to $\dot{x}(t) = f(x, u)$,

(6.1c) $x(0) = x_0$,

(6.1d) $(x(t), u(t)) \in X \times U(x(t)) \subseteq \mathbb{R}^n \times \mathbb{R}^m$, for each $t \in [0, \infty)$.

We seek a characterization of the optimal steady-state trajectories of the discounted optimal control problem, pairs $(x^*, u^*) \equiv (x^0, u^0)$, for all t. We have seen that if the basic convexity property, given by Assumption 4.1, holds for the data of the problem, we may characterize the optimal trajectories of the problem by any one of three approaches. We shall describe the optimal steady-state trajectories in terms of the support property and formulate a static optimization problem, based on that description, that has as its solution the optimal steady-state. Then, we shall show that the dynamic theory that has been developed to analyze problem (6.1) may be interpreted as a means of establishing the boundary condition contained in Theorem 3.1, the sufficient conditions for dynamic optimality.

6.2 Static characterization of the optimal steady-state for the undiscounted problem. There is a complete body of literature analyzing problem (6.1) when $\rho = 0$.

Some of the more recent references are Rockafellar [26], Haurie [15] and Brock and Haurie [5]; the classic reference is Ramsey [22]. It is known that, under sufficiently strict convexity conditions (including the convexity of the image L^*) the optimal state-costate trajectory converges to the saddlepoint (x^0, p^0) of the optimal value Hamiltonian [26, Thm. 1.2]. This saddlepoint may be characterized in terms of a static optimization problem, as noted by Brock and Haurie [5]. This result is given by the next theorem.

The theorem is motivated by the intuitively appealing idea that if the optimal trajectory converges toward a steady-state, or if a steady-state is an optimal trajectory for the dynamic problem, then that steady-state should minimize the kernel of the objective functional, $l(x, u)$. To estabish this result, as well as the other results of this section, we require the convexity assumption on $\Omega(t)$ (Assumption 4.1) or equivalently (Lemma 4.1).

Assumption 6.1. The image function, defined by

$$L^*(x, v) = \begin{cases} \inf_u \{l(x, u): v = f(x, u), u \in U(x)\} \\ +\infty \quad \text{if } x \notin X \quad \text{or} \quad v \neq f(x, u) \quad \forall u \in U(x), \end{cases} \tag{6.2}$$

is a lower semicontinuous proper convex function. (Recall that Proposition 5.1 provides a set of conditions under which Assumption 6.1 holds.)

THEOREM 6.1. *Suppose that Assumption* 6.1 *holds. Define the optimal value Hamiltonian*

$$H^*(x, p) = \sup_v \{\langle p, v\rangle - L^*(x, v): v \in \mathbb{R}^n\}.$$

Let (x^0, p^0) *be a saddlepoint of* H^*. *Assume* $H^*(x^0, p^0)$ *is finite. Then* (x^0, u^0) *is a solution to the mathematical programming problem*

$$\text{(6.4a)} \qquad \text{minimize} \qquad l(x, u)$$

$$\text{(6.4b)} \qquad \text{subject to} \qquad f(x, u) = \theta,$$

$$x \in X, u \in U(x)$$

where $u^0 = \inf_u^{-1} \{l(x^0, u): \theta = f(x^0, u), u \in U(x^0)\}$.

Proof. Apply the proof of Theorem 6.2, below, with $\rho = 0$ □

Theorem 6.1 provides a static characterization of the optimal steady state trajectory of the undiscounted optimal control problem. We relate the static problem (6.4) to the dynamic theory through the support property (Definition 3.1). For a steady-state trajectory, $(x^*(t), u^*(t), p^*(t)) \equiv (x^0, u^0, p^0)$, the support property becomes

$$l(x^0, u^0) \leqq l(x, u) - \langle p^0, f(x, u)\rangle, \tag{6.5}$$

$$\forall (x, u) \in D = \{(x, u): x \in X, u \in U(x)\}.$$

It is clear that a supported steady-state trajectory is necessarily a solution of the static problem (6.4). This observation suggests a similar characterization for the optimal steady-state trajectories of the discounted optimal control problem.

6.3 The implicit programming problem. The static characterization of the optimal steady-state trajectories of the discounted problem will be derived from the support property. The support property for the discounted problem is, for the trajectory

(x^*, u^*) and supporting function $p^*(t) = e^{\rho t} q^*(t)$ (see § 6.4), the inequality

$$-e^{-\rho t} l(x^*(t), u^*(t)) + \frac{d}{dt}[\langle e^{-\rho t} q^*(t), x^*(t)\rangle]$$

(6.6a) $$\geqq -e^{-\rho t} l(x, u) + \langle e^{-\rho t} q^*(t), f(x, u)\rangle + \langle e^{-\rho t}(\dot{q}^*(t) - \rho q^*(t)), x\rangle$$

$$\forall (x, u) \in D = \{(x, u) : x \in X, u \in U(x)\}, \text{ for almost every } t \in [0, \infty).$$

For a steady-state trajectory, $(x^*, u^*, q^*) \equiv (x^0, u^0, q^0)$ with $f(x^0, u^0) = \theta$, the support property becomes the inequality

(6.6b) $$l(x^0, u^0) \leqq l(x, u) - \langle q^0, f(x, u) - \rho(x - x^0)\rangle, \quad \forall (x, u) \in D.$$

A necessary condition for (x^0, u^0) to be supported is that (x^0, u^0) is a solution to the well-defined nonlinear programming problem

(6.7a) minimize $l(x, u)$

(6.7b) subject to $f(x, u) - \rho(x - x^0) = \theta,$

(6.7c) $x \in X, \quad u \in U(x).$

Since x^0 is a fixed vector, the constraint (6.7b) is specified precisely and problem (6.7) is well defined. We now wish to consider the problem

(6.8a) minimize $l(x, u)$

(6.8b) subject to $f(x, u) - \rho(x - x^*) = \theta,$

(6.8c) $x \in X, \quad u \in U(x)$

where x^* is not fixed in advance. Indeed, x^*, as it appears in the constraint (6.8b), indicates the value of the x-component of the solution to the problem (6.8). In other words, the constraint is defined implicitly by the solution to the problem itself.

We call problem (6.8) an implicit programming problem, and we claim that it is a well-defined mathematical programming problem. Furthermore, we claim that under certain conditions, to be described below, the solution to the implicit programming problem is an optimal steady-state trajectory for the optimal control problem with discounting.

The most natural way to interpret the implicit programming problem is that it actually defines a mapping from $\mathbb{R}^n$ to $\mathbb{R}^n \times \mathbb{R}^m$. To highlight this interpretation, let us replace x^* in the constraint (6.8b) with a parameter $c \in \mathbb{R}^n$. As c varies over $\mathbb{R}^n$, a family of nonlinear programming problems is created. Moreover, for any c, the problem

(6.9a) minimize $l(x, u)$

(6.9b) subject to $f(x, u) - \rho(x - c) = \theta,$

(6.9c) $x \in X, \quad u \in U(x)$

defines a mapping that takes $c \in \mathbb{R}^n$ into the solution of problem (6.9), $(x^*(c), u^*(c)) \in D$.

The implicit programming problem (6.8) may then be written:

(6.10a) minimize $l(x^*(c), u^*(c))$

(6.10b) subject to $x^*(c) = c,$

(6.10c) $c \in X,$

where the minimization is taken over the fixed-points of the mapping $c \to x^*(c)$. The fact that this mapping is defined implicitly by the mathematical programming problem (6.9) suggests the terminology "implicit programming problem." One would expect the minimization defined by (6.10) to be over a discrete set (although there is a possibility of degeneracy in the data of the problem, in which case the fixed-points of the mapping form a continuum). Indeed, the discrete nature of the "feasible" set for (6.10) indicates that what is of primary importance in the analysis of the implicit programming problem is the set of local solutions; i.e., all the fixed-point of the implicit mapping defined by (6.9). We will designate such points as feasible points for the implicit programming problem (6.8).

It is clear that every feasible point of the implicit programming problem (i.e., a fixed-point of the implicit mapping) is a steady-state of the dynamic system $\dot{x} = f(x, u)$, since the term $\rho(x - c)$ in the constraint (6.9b) vanishes identically at the solution $x^*(c)$ whenever $x^*(c) = c$. Moreover, the solution to the implicit programming problem will not be the same as the solution of the undiscounted static problem (6.4); in general, the (globally) optimal value of the implicit programming problem will be higher than that of the undiscounted problem. This follows from the fact that only a subset of the steady-states correspond to fixed-points of the implicit mapping defined by (6.9), while the entire null space of f is feasible for (6.4). Thus, the optimal steady-state for the discounted problem is inferior compared with that for the undiscounted problem. This behavior is a direct result of the discounting of the objective; since later performance is valued less than earlier performance, the optimal trajectory exploits the dynamic possibilities in the structure of the problem at the beginning of the period to converge to what appears to be a suboptimal steady-state at the end. In fact, even if the initial condition were specified as the (global) solution to (6.4), the optimal trajectory would not remain at that point, but instead converge to a local solution of the implicit programming problem.

6.4 Determination of stationary points of the Hamiltonian dynamic system. We will now verify the claim that every local solution to the implicit programming problem is an optimal steady-state trajectory for the optimal control problem with discounting. To perform the analysis, we formulate the Hamiltonian dynamic system for the discounted problem, and express the steady-state trajectories in terms of the subdifferential equations of the dynamic system. We first show that the implicit programming problem characterizes the steady-states of the Hamiltonian dynamic system.

The Hamiltonian dynamic system is formulated subject ot the basic assumption that the image function L^* is a lower semicontinuous proper convex function (Assumption 6.1). The Hamiltonian is defined by:

$$H^*(t, x, p) = \sup_v \{\langle p, v\rangle - e^{-\rho t} L^*(x, v) : v \in \mathbb{R}^n\}. \tag{6.11}$$

It is convenient to introduce the change of costate variables, $p(t) = e^{-\rho t} q(t)$, which defines the current-value Hamiltonian, $H^*(x, q)$, such that $H^*(t, x, p) = e^{-\rho t} H^*(x, q)$, where

$$\begin{aligned} H^*(x, q) &= \sup_u \{\langle q, f(x, u)\rangle - l(x, u) : u \in U(x)\} \\ &= \sup_v \{\langle q, v\rangle - L^*(x, v) : v \in \mathbb{R}^n\}. \end{aligned} \tag{6.12}$$

The optimal state-costate trajectory of problem (6.1) is a solution to the Hamiltonian dynamic system, which due to the change of variables, becomes the autonomous

system

$$(-\dot{q}(t)+\rho q(t), \dot{x}(t)) \in \partial H^*(x(t), q(t)), \tag{6.13}$$

the so-called *modified* Hamiltonian dynamic system. The steady-state trajectories of the modified Hamiltonian dynamic system are pairs (x^0, q^0) such that

$$(\rho q^0, \theta) \in \partial H^*(x^0, q^0). \tag{6.14}$$

We begin by proving the counterpart to Theorem 6.1, which indicates that every stationary point of the modified Hamiltonian dynamic system determines a feasible point of the implicit programming problem. The proof requires the following lemma, which relates the Hamiltonian dynamic system to the Euler-Lagrange Equations.

LEMMA 6.1 (Conjugate duality (partial conjugates)). *Let* $L(x, v): \mathbb{R}^n \times R^n \to (-\infty, +\infty]$ *be a lower semicontinuous proper convex function. Define*

$$H(x, p) = \sup_v \{\langle p, v\rangle - L(x, v): v \in \mathbb{R}^n\}.$$

Then $H(x, p): \mathbb{R}^n \times \mathbb{R}^n \to [-\infty, +\infty]$ *is concave in* x *and convex in* p, *and the following conditions on* $(x, p) \in \mathbb{R}^n \times \mathbb{R}^n$ *and* $(x^*, v^*) \in \mathbb{R}^n \times \mathbb{R}^n$ *are equivalent*:

(i) $-x^* \in \partial_x H(x, p)$, $\quad v^* \in \partial_p H(x, p)$;
(ii) $x^* \in \partial_x L(x, v^*)$, $\quad p \in \partial_v L(x, v^*)$.

Proof. These results are all given by Rockafellar [24, see Thms. 33.1 and 37.5]. □

THEOREM 6.2. *Suppose that Assumption* 6.1 *holds. Let* $(x^0, q^0) \in X \times \mathbb{R}^n$ *be a stationary point of the modified Hamiltonian dynamic system.*

Then (x^0, u^0) *is a solution to the mathematical programming problem* (6.9), *with* $c = x^0$,

$$\begin{aligned} &\text{minimize} && l(x, u) \\ &\text{subject to} && f(x, u) - \rho(x - x^0) = \theta, \\ & && x \in X, \quad u \in U(x), \end{aligned}$$

where $u^0 = \min_u{}^{-1}\{l(x^0, u): \theta = f(x^0, u), u \in U(x^0)\}$.

Proof. The stationary point condition (6.14), is equivalent to the subdifferential condition on L^* (Lemma 6.1), $(-\rho q^0, q^0) \in \partial L^*(x^0, \theta)$. By definition of a subgradient of a convex function (Definition 2.2), it follows that

$$L^*(x^0 + \xi, v) \geqq L^*(x^0, \theta) + \langle q^0, v - \rho\xi\rangle, \quad \forall (\xi, v) \in \mathbb{R}^n \times \mathbb{R}^n.$$

Thus, for $v = \rho\xi$, we have

$$L^*(x^0 + \xi, \rho\xi) \geqq L^*(x^0, \theta) = l(x^0, u^0), \quad \text{where}$$
$$u^0 = \min_u{}^{-1}\{l(x^0, u): \theta = f(x^0, u), u \in U(x^0)\}.$$

By definition of $L^*(x^0 + \xi, \rho\xi)$, letting $\xi = x - x^0$,

$$\begin{aligned} l(x^0, u^0) &\leqq \inf_u \{l(x, u): \rho(x - x^0) = f(x, u), u \in U(x), x \in \mathbb{R}^n\} \\ &= \min_{(x,u)} \{l(x, u): \rho(x - x^0) = f(x, u), u \in U(x), x \in X\}. \end{aligned}$$

□

We are far more interested in a converse result, which would indicate how the implicit programming problem may be used as an analytic tool in conjunction with the established dynamic theory. Our objective is to pose sufficient conditions on the solution of the implicit programming problem to ensure that an optimal steady-state has been determined. We first show that every local solution of the implicit programming problem (i.e., a fixed-point of the implicit mapping $x^*(c)$) determines the state component of a stationary point of the modified Hamiltonian dynamic system.

PROPOSITION 6.1. *Let* (x^*, u^*) *be a solution to the mathematical programming problem* (6.9) *with* $c = x^*$, *hence feasible for* (6.8). *Suppose that Assumption* 6.1 *holds. Suppose further that the point* (x^*, θ) *is the interior of the effective domain of* L^*. *Then there exists a stationary point of the modified Hamiltonian dynamic system at* (x^*, q^*), *for some* $q^* \in \mathbb{R}^n$.

Proof. Since (x^*, u^*) is a solution to the mathematical programming problem (6.9) with $c = x^*$ we have

$$L^*(x^*, \theta) = \inf_u \{l(x^*, u): \theta = f(x^*, u), u \in U(x^*), x^* \in X\} = l(x^*, u^*).$$

By assumption, $L^*(x, v)$ is a proper convex function, hence, if (x^*, θ) is in the interior of the effective domain of L^*, then the subdifferential of L^* at (x^*, θ) is not empty [24, Thm. 23.4]. Hence, for some $(\zeta, q^*) \in \mathbb{R}^n \times \mathbb{R}^n$,

$$L^*(\xi + x^*, v) \geqq L^*(x^*, \theta) + \langle q^*, v\rangle + \langle \zeta, \xi\rangle \quad \forall (\xi, v) \in \mathbb{R}^n \times \mathbb{R}^n.$$

Now let $v = \rho\xi$. We have

$$L^*(\xi + x^*, \rho\xi) \geqq L^*(x^*, \theta) + \langle \rho q^* + \zeta, \xi\rangle \quad \forall \xi \in \mathbb{R}^n.$$

However, since (x^*, u^*) is the solution to problem (6.9), we also have

$$\begin{aligned} l(x^*, u^*) &= L^*(x^*, \theta) \\ &= \min_{(x, u)} \{l(x, u): \rho(x - x^*) = f(x, u), u \in U(x), x \in X\} \\ &\leqq L^*(x, \rho(x - x^*)) \quad \forall x \in X. \end{aligned}$$

Letting $\xi = x - x^*$, this implies $L^*(\xi + x^*, \rho\xi) - L^*(x^*, \theta) \geqq 0$, $\forall \xi \in \mathbb{R}^n$. Hence, it follows that $\rho q^* + \zeta = \theta$, which implies $(-\rho q^*, q^*) \in \partial L^*(x^*, \theta)$. By Lemma 6.1, we conclude that the subdifferential condition holds:

$$(\rho q^*, \theta) \in \partial H^*(x^*, q^*). \qquad \square$$

We will now characterize the costate component of the stationary point of the modified Hamiltonian dynamic system in terms of the Lagrange multiplier of the implicit programming problem. Recall the following results from the theory of nonlinear programming.

We consider the nonlinear programming problem with equality constraints:

(6.15a) minimize $l(x)$

(6.15b) subject to $f(x) = \theta$,

(6.15c) $x \in D \subseteq \mathbb{R}^s$.

DEFINITION 6.1. Let $f: \mathbb{R}^s \to \mathbb{R}^n$ be continuously differentiable. A point x^0 satisfying the constraints $f(x^0) = \theta$ is said to be a *regular point* of the constraints if the gradient vectors $\nabla_x f_1(x^0), \cdots, \nabla_x f_n(x^0)$ are linearly independent.

Definition 6.1 is equivalent to the fact that the $n \times s$ Jacobian matrix $[\nabla_x f(x)]$ is full rank at x^0.

The idea of a regular point is essential for characterizing the solution of problem (6.15), as indicated in the next theorem.

THEOREM 6.3. (First-order necessary conditions for a minimum; Lagrange multipliers). *Let (l, f) be $\mathbb{C}^1$ functions. Let x^0 be a solution to* (6.15), *that is a regular point of the constraints. Suppose further that $x^0 \in D^0$, an interior point of the feasible set. Then there exists $\lambda^0 \in \mathbb{R}^n$ such that $\nabla_x l(x^0) - \langle \lambda^0, \nabla_x f(x^0) \rangle = \theta$.*

Proof. See [18, Chpt. 10]. □

We shall define the Lagrangian of the implicit programming problem as follows: let

$$L_\rho(x, u, \lambda; c) = \begin{cases} l(x, u) - \langle \lambda, f(x, u) - \rho(x - c) \rangle, \\ +\infty \quad \text{if } x \notin X \quad \text{or} \quad u \notin U(x), \end{cases} \tag{6.16}$$

where $c \in \mathbb{R}^n$ is a fixed vector. This is the Lagrangian of a member of the class of problems (6.9).

With this definition of the Lagrangian, we may express the optimal value Hamiltonian as

$$H^*(x, q) = \begin{cases} \sup_u \{-L_\rho(x, u, q; c) : u \in U(x)\} + \rho\langle q, x - c \rangle, \\ -\infty \quad \text{if } x \notin X. \end{cases} \tag{6.17}$$

We now relate the saddlepoints of the Lagrangian to the saddlepoints of the (perturbed) optimal value Hamiltonian,

$$H_\rho^*(x, q; c) = H^*(x, q) - \rho\langle q, x - c \rangle. \tag{6.18}$$

LEMMA 6.2. *If $L_\rho(x, u, q; x^0)$, the Lagrangian* (6.16) *with $c = x^0$, possesses a saddlepoint (x^0, u^0, q^0), with $x^0 \in X$ and $u^0 \in U(x^0)$, then (x^0, q^0) is a saddlepoint of $H_\rho^*(x, q; x^0)$ where u^0 furnishes the supremum in the definition of $H^*(x^0, q^0)$.*

Further, if $H^*(x, q)$ is a concave-convex function, then (x^0, q^0) is a stationary point of the modified Hamiltonian dynamic system.

Proof. The saddlepoint condition on L_ρ is equivalent to the inequalities

$$-L_\rho(x, u, q^0; x^0) \leqq -L_\rho(x^0, u^0, q^0; x^0) \leqq -L_\rho(x^0, u^0, q; x^0) \quad \forall (x, u, q).$$

By definition, for $x^0 \in X$,

$$H_\rho^*(x^0, q; x^0) = \sup_u \{-L_\rho(x^0, u, q; x^0) : u \in U(x^0)\}$$

$$\geqq -L_\rho(x^0, u^0, q; x^0) \quad \forall q.$$

Further, $-L_\rho(x, u, q^0; x^0) \leqq -L_\rho(x^0, u^0, q^0; x^0)$ implies that

$$H_\rho^*(x, q^0; x^0) = \sup_u \{-L_\rho(x, u, q^0; x^0) : u \in U(x)\}$$

$$\leqq -L_\rho(x^0, u^0, q^0; x^0) \quad \forall x.$$

Hence,

$$H_\rho^*(x, q^0; x^0) \leqq -L_\rho(x^0, u^0, q^0, x^0) \leqq H_\rho^*(x^0, q; x^0) \quad \forall (x, q).$$

Thus,

$$H_\rho^*(x^0, q^0; x^0) = -L_\rho(x^0, u^0, q^0; x^0).$$

Therefore u^0 furnishes the supremum in the definition of $H_\rho^*(x^0, q^0; x^0)$, and (x^0, q^0) is a saddlepoint of H_ρ^*.

If $H^*(x, q)$ is a concave-convex function, it follows that $H_\rho^*(x, q; x^0)$ is also concave-convex. Hence, the saddlepoint condition on $H_\rho^*(x, q; x^0)$ is equivalent to the subgradient condition

$$(\theta, \theta) \in \partial H_\rho^*(x^0, q^0; x^0) \quad \text{or} \quad (\rho q^0, \theta) \in \partial H^*(x^0, q^0). \qquad \square$$

The search for stationary points of the modified Hamiltonian dynamic system becomes, by Lemma 6.2, a search for saddlepoints of the Lagrangian $L_\rho(x, u, q; c)$, with the special property that the state-component of the saddlepoint, x^0, is equal to the parameter c that defines the Lagrangian. This bears out the observation of Cass and Shell that the optimal steady-state is "something like a saddlepoint for the modified current value Hamiltonian $H^*(x, q) - \rho\langle q, x\rangle$" [7, p. 54]. The implicit programming formulation suggests consideration of the perturbed Hamiltonian $H_\rho^*(x, q; c)$; the saddlepoint behavior of this function is clearly a restatement of the support property. The main result of Lemma 6.2 is that every stationary point of the modified Hamiltonian dynamic system determines a stationary trajectory that is supported. Moreover, a necessary condition for such a trajectory to exist is that (x^0, u^0) is a solution to problem (6.9), with $c = x^0$, hence (x^0, u^0) is feasible for the implicit programming problem.

The next result provides sufficient conditions under which a feasible point of the implicit programming problem completely characterizes the stationary point of the modified Hamiltonian dynamic system. The result is a corollary to Proposition 6.1 and Lemma 6.2.

COROLLARY 6.1. *Let (x^*, u^*, λ^*) be a solution to the mathematical programming problem* (6.9) *with $c = x^*$. Assume that f and l are $\mathbb{C}^1$ functions. Suppose (x^*, u^*) is a regular point of the constraints $f(x, u) - \rho(x - x^*) = \theta$, and $x^* \in X^0$, $u^* \in [U(x^*)]^0$. Suppose further that Assumption* 6.1 *holds and that the point (x^*, θ) is in the interior of the effective domain of L^*. Then (x^*, λ^*) is a stationary point of the modified Hamiltonian dynamic system.*

Proof. The assumptions that L^* is a proper convex function and $(x^*, \theta) \in [\text{dom } L^*]^0$ imply, as in the proof of Proposition 6.1, that for some $q^* \in \mathbb{R}^n$ there holds

$$L^*(\xi + x^*, v) \geqq L^*(x^*, \theta) + \langle q^*, v - \rho\xi\rangle, \quad \forall (\xi, v) \in \mathbb{R}^n \times \mathbb{R}^n.$$

By definition of L^*, letting $\xi = x - x^*$, this is equivalent to

$$l(x^*, u^*) \leqq \inf_u \{l(x, u) : v = f(x, u),\ u \in U(x)\} - \langle q^*, v - \rho(x - x^*)\rangle, \quad \forall (x, v)$$

$$\leqq \inf_u \{l(x, u) - \langle q^*, f(x, u) - \rho(x - x^*)\rangle,\ u \in U(x)\}, \quad \forall x.$$

Now let $x = x^*$. The infimum is then attained at $u = u^* \in [U(x^*)]^0$. Since u^* is an interior point, and (x^*, u^*) is a regular point of the constraints $f(x, u) - \rho(x - x^*) = \theta$, the necessary conditions hold,

$$\nabla_u [l(x^*, u) - \langle q^*, f(x^*, u)\rangle]_{u = u^*} = \theta,$$

and imply that $q^* = \lambda^*$, by Theorem 6.3. It follows that

$$l(x^*, u^*) \leqq l(x, u) - \langle \lambda^*, f(x, u) - \rho(x - x^*)\rangle \quad \forall x \in X, \quad u \in U(x).$$

Then the Langrangian $L_\rho(x, u, q; x^*)$ possesses a saddlepoint at (x^*, u^*, λ^*). Lemma 6.1 indicates that the optimal value Hamiltonian, $H^*(x, q)$, is a concave-convex

function, given the assumptions on L^*. Hence, by Lemma 6.2, (x^*, λ^*) is a stationary point of the modified Hamiltonian dynamic system. □

Thus, we have established conditions under which a local solution to the implicit programming problem, with Lagrange multiplier, completely determines a stationary point of the modified Hamiltonian dynamic system. To assert the dynamic optimality of a local solution to the implicit programming problem, we must apply the sufficient maximum principle. This requires an investigation of the relationship of the solutions of the implicit programming problem to the dynamic theory.

To this point, our analysis has presented new results of a purely static nature. What we seek to demonstrate below is the extent to which the static approach can substitute for the dynamic analysis, for the purpose of characterizing the complete dynamic trajectory. The novel aspects of these subsequent results are contained in their relationship to the formulation of the implicit programming problem itself, and to the decomposition perspective suggested by the sufficient maximum principle. However, this portion of the theory is derived from the results of the current dynamic theory; we are, in this last section, indicating how implicit programming complements the dynamic analysis, rather than presenting any new results of a dynamic nature.

6.5 Application of the dynamic theory. The most complete dynamic analysis of the discounted problem is due to Rockafellar [27]. The theory developed rests on the following curvature assumption.

Assumption 6.2 (*curvature assumption*). We assume that for certain values $\alpha > 0$ and $\beta > 0$ the Hamiltonian H^* is locally α-concave and β-convex near the stationary point of the modified Hamiltonian dynamic system, (x^0, q^0), or in other words, that there exists a convex neighborhood $S \times T$ of (x^0, q^0) in $\mathbb{R}^n \times \mathbb{R}^n$ such that $H^*(x, q)$ is (finite and) α-concave in $x \in S$ for each $q \in T$ and β-convex in $q \in T$ for each $x \in S$. Moreover, the discount rate $\rho > 0$ is small enough so that $\rho^2 < 4\alpha\beta$.

The notion of α-convexity is a measure of strict convexity.

DEFINITION 6.2. A finite function h on a convex set $C \subseteq \mathbb{R}^n$ is said to be *α-convex*, where $\alpha \in \mathbb{R}$, if for all $x \in C$, $x' \in C$ and $\lambda \in [0, 1]$ it is true that

$$h((1-\lambda)x + \lambda x') \leqq (1-\lambda)h(x) + \lambda h(x') - \tfrac{1}{2}\alpha\lambda(1-\lambda)\|x - x'\|^2. \tag{6.19a}$$

α-concavity is defined by replacing the inequality above with

$$h((1-\lambda)x + \lambda x') \geqq (1-\lambda)h(x) + \lambda h(x') + \tfrac{1}{2}\alpha\lambda(1-\lambda)\|x - x'\|^2. \tag{6.19b}$$

The optimality of a steady state trajectory is based on the following convergence lemma. The convergence properties of the optimal dynamic trajectories are established subject to the curvature assumption (see [27]). It is this convergence property that indicates the important role that the optimal steady-state trajectories play in the dynamic theory.

LEMMA 6.3. *Suppose $x : [0, \infty) \to X \subseteq \mathbb{R}^n$ is a continuous, piecewise continuously differentiable function such that the objective functional converges, where L^* satisfies Assumption* 6.1:

$$\int_0^\infty e^{-\rho t} L^*(x, \dot{x})\, dt < +\infty.$$

Let (x^0, q^0) be a stationary point of the modified Hamiltonian dynamic system such that Assumption 6.2 *holds. Suppose further that*

$$\liminf_{t \to \infty} e^{-\rho t}\langle q^0, x(t)\rangle > -\infty.$$

Then, if Assumptions 6.1 *and* 6.2 *hold,* $\lim_{t\to\infty} e^{-\rho t}(x(t)-x^0)=\theta$.

Proof. See [27, Propositions 1 and 2]. □

We may now apply the sufficient maximum principle to claim that a local solution to the implicit programming problem is an optimal steady-state trajectory.

THEOREM 6.4. *Let* (x^*, u^*, λ^*) *be a local solution to the implicit programming problem. Assume that f and l are* $\mathbb{C}^1$ *functions. Suppose* (x^*, u^*) *is a regular point of the constraints* $f(x, u)-\rho(x-x^*)=\theta$, *and* $x^*\in X^0$, $u^*\in[U(x^*)]^0$. *Suppose that Assumptions* 6.1 *and* 6.2 *hold, the latter in a neighborhood of* (x^*, λ^*). *Then the stationary trajectory* (x^*, u^*) *is optimal in the class of trajectories with initial condition* $x(0)=x^*$ *such that* $e^{-\rho t}x(t)$ *remains bounded as* $t\to\infty$.

Proof. It follows, from Corollary 6.1, that (x^*, u^*) is a stationary trajectory that is supported by λ^*. (The interior point condition on (x^*, θ) follows from the curvature assumption, (see [27, Proposition 1]).) Lemma 6.3 indicates that the boundary condition

$$\lim_{t\to\infty} e^{-\rho t}\langle\lambda^*, (x(t)-x^*)\rangle\geqq 0,$$

is satisfied for all state trajectories x such that $e^{-\rho t}x(t)$ remains bounded as $t\to\infty$, that also provide finite cost sums. Since the support property and the boundary condition are satisfied, the optimality of the stationary trajectory follows from Theorem 3.1. □

Theorem 6.4 indicates conditions for a local solution of the implicit programming problem to be an optimal stationary trajectory. Moreover, the proof of the theorem is based on the decomposition perspective that is suggested by the sufficient maximum principle. The steady-state that is characterized by the implicit programming problem is a supported trajectory, because of the convexity assumption on L^* (Assumption 6.1). The optimality results from the boundary condition describing the asymptotic behavior of other trajectories. The boundary condition is established by the dynamic theory, as indicated by Lemma 6.3. In fact, the curvature assumption is actually a strengthening of the basic convexity assumption. By Lemma 6.1, the convexity of L^* implies that the Hamiltonian is concave-convex; Assumption 6.2 strengthens that property sufficiently to allow the boundary condition to be established.

If we view the theory from this perspective, there is another point to be mentioned. The curvature assumption is sufficiently powerful that we are able to conclude that it is actually the limit of the inner product that satisfies the nonnegativity condition required in the sufficiency theorem. Yet the relaxed concepts of optimality (e.g., weakly overtaking) require weaker asymptotic conditions. This suggests that one direction in which to proceed would be to weaken the curvature assumption and aim the dynamic theory towards establishing the boundary condition of Theorem 3.1.

There are two other kinds of theorems that can be proven about the solutions to the implicit programming problem. The first is a uniqueness theorem that indicates sufficient conditions for (at least) the state-component of the implicit programming problem to be unique. Corollary 6.1 indicates that every solution of the implicit programming problem determines a stationary point of the modified Hamiltonian dynamic system. Hence, conditions that are sufficient for the stationary point to be unique also imply that the solution to the implicit programming problem is unique (if the control that defines the optimal value Hamiltonian is unique). It is straightforward to establish that if the curvature assumption holds globally, then the stationary point is unique. (This result follows from an argument similar to that of [27, Proposition 5].) It is not true, however, that the stationary point is unique if the curvature assumption happens to hold at a particular point (locally). Further, even if the Hamiltonian H^* is

globally strictly concave-convex, there may be multiple stationary points if the inequality $\rho^2 < 4\alpha\beta$ is violated (see [16, Example 1]).

The other kind of theorem indicates how the implicit programming problem can be used to analyze the asymptotic behavior of other optimal trajectories, with initial conditions sufficiently close to the steady-state. This is essentially a stability question, particularly when it is viewed in the context of the Hamiltonian dynamic system. The dynamic theory has established that, if Assumption 6.2 holds, the stationary point of the Hamiltonian dynamic system behaves like a saddlepoint of a differential equation. That is, there exist, locally, stable and unstable manifolds, which intersect only at the stationary point, that are comprised of the trajectories that converge to the stationary point as t approaches plus or minus infinity, respectively [27, Thms. 1 and 1']. Then, for any initial condition on the manifold, the optimal trajectory remains in the manifold and converges to the stationary point [27, Thm. 2]. It is this result that indicates the important role played by the optimal steady-state trajectories; they are limit points of other optimal trajectories. Moreover, if Assumption 6.2 holds globally, then the solution of the implicit programming problem is unique, the stationary point of the modified Hamiltonian dynamic system is unique, and all optimal trajectories converge to this stationary point. In any event, if the curvature assumption holds in a neighborhood of a local solution to the implicit programming problem, the dynamic theory indicates that the solution is the limit point of other optimal trajectories, as well as an optimal steady-state trajectory.

Since the curvature assumption is a static property of the Hamiltonian, additional conditions may be imposed on the solution of the implicit programming problem that are sufficient to conclude that the curvature assumption holds. One way of establishing the curvature assumption would be to investigate the Hessian matrices of the Hamiltonian. This is the approach taken in the next theorem. We add smoothness conditions to the functions l and f, and a regularity condition on the mapping U. We also add two special assumptions, the local duality assumption [18, Ch. 12], that guarantees that the Lagrangian is locally convex at the solution of the implicit programming problem, and, a local controllability assumption that takes the form of a rank condition of the u-Jacobian of the function f. This latter assumption implies that the dimension of the control space is at least as great as the dimension of the state space. This condition is met in the calculus of variations, where control is identified with the derivative of the arc; in general control problems, it amounts to a restriction. However, in economic models, such a condition is generally satisfied. The main result of the theorem is that the Hamiltonian is strictly concave-convex in the neighborhood of a local solution to the implicit programming problem.

THEOREM 6.5. *Let* (x^*, u^*, λ^*) *be a solution to the implicit programming problem* (6.8), *such that* (x^*, u^*) *is a regular point of the constraints* $f(x, u) - \rho(x - x^*) = \theta$, *and* $x^* \in X^0$, $u^* \in [U(x^*)]^0$. *Let Assumption* 6.1 *hold and suppose further that*:

(i) *the functions f and l possess continuous second derivatives with respect to all arguments*;

(ii) *the mapping* $U : X \to 2^{\mathbb{R}^m}$ *is lower semi-continuous*;

(iii) *the Hessian of the Lagrangian*

$$\nabla^2_{(x,u)} L_\rho(x, u, \lambda^*; x^*) = \nabla^2_{(x,u)}(l(x, u) - \langle \lambda^*, f(x, u)\rangle)$$

evaluated at (x^*, u^*) *is positive-definite on all of* $\mathbb{R}^n \times \mathbb{R}^n$; *and*

(iv) *the matrix* $[\nabla_u f(x, u)]_{(x^*,u^*)}$ *is an* $n \times m$ *matrix of rank* n (hence $m \geqq n$).

Then the concave-convex optimal value Hamiltonian $H^*(x, q)$ is strictly concave-convex in an open convex neighborhood, $S \times T$, of the stationary point (x^*, λ^*) of the modified Hamiltonian dynamic system.

In addition, if $H^*(x, \lambda^*)$ is α-concave for $x \in S$ and if $H^*(x^*, q)$ is β-convex for $q \in T$, such that $\rho^2 < 4\alpha\beta$, then there exists an open neighborhood $S_+ \subset S$, $x^* \in S_+$, such that for any initial condition $x(0) \in S_+$, the optimal state-control-costate trajectory converges to (x^*, u^*, λ^*).

Outline of proof. Defining $u^*(x, q)$ as the minimizer in the definition of the optimal value Hamiltonian (6.12), the first-order necessary conditions for the implicit programming problem contain a system of m equations in $(n+m+n)$ variables that defines $u^*(x, q)$ implicitly. The local duality assumption (iii) indicates that the Jacobian of this system is nonsingular at (x^*, u^*, λ^*), so that the implicit function theorem applies to $u^*(x, q)$ in a neighborhood of the solution (x^*, λ^*). Calculating the Hessians of the optimal value Hamiltonian, $H^*(x, q) = \langle q, f(x, u^*(x, q))\rangle - l(x, u^*(x, q))$, it is easy to see that they are definite. The lower semi-continuity assumption on the mapping U is required to assert that the implicit function $u^*(x, q)$ actually belongs to the set $U(x)$ for x in a neighborhood S of x^*. The assumption on the relative sizes of the state and control spaces is required to express the Hessian of the Hamiltonian, with respect to q, as an inner product of full-rank matrices, hence definite. The last statement of the theorem is precisely Theorem 2, [27]. For further details, see [8]. □

6.6 Summary. We have shown that, under the basic convexity assumption describing the image function L^* (Assumption 6.1):

(i) every stationary point of the modified Hamiltonian dynamic system is a feasible point of the implicit programming problem (Theorem 6.2);

(ii) conversely, every feasible point of the implicit programming problem determines a stationary point of the modified Hamiltonian dynamic system (Proposition 6.1);

(iii) every stationary point of the modified Hamiltonian dynamic system determines a stationary trajectory that is supported (Lemma 6.2);

(iv) if the solution to the implicit programming problem is a regular point, then the stationary point of the Hamiltonian dynamic system is determined by the Lagrange multiplier (Corollary 6.1);

and, complementing the dynamic theory,

(v) if the curvature assumption holds in a neighborhood of the solution of the implicit programming problem, then the solution of the implicit programming problem is an optimal steady-state trajectory in the class of dynamic trajectories such that $e^{-\rho t}x(t)$ is bounded as $t \to \infty$ (Theorem 6.4);

(vi) the solution to the implicit programming problem is the limit point of other optimal trajectories, under additional (local) assumptions designed to establish the curvature assumption (Theorem 6.5).

The implicit programming problem also remedies the lack of a transversality condition in the maximum principle for the infinite horizon problem. The Lagrange multiplier of the implicit programming problem determines the boundary condition at infinity of the costate trajectory, and the familiar two-point boundary-value problem determines the optimal trajectory.

7. Conclusions. We have formulated a static optimization problem, the implicit programming problem, that characterizes the optimal steady-states of the optimal control problem with discounting; all results hold for the undiscounted case as well. This

characterization does not require the solution of the full dynamic problem to determine the asymptotic behavior of the optimal dynamic trajectories. The formulation of the problem is based on an application of sufficient conditions for dynamic optimality, which have been shown to be equivalent to the more familiar approaches found in the literature. The sufficient conditions suggest a decomposition of the analysis of optimal control problems defined on an infinite horizon. The implicit programming problem responds to one aspect of that decomposition, the support property. The current dynamic theory may be interpreted as an approach to the other aspect of this decomposition, the boundary condition. This perspective suggests a possible direction in which to proceed, that of investigating ways in which to weaken the assumptions of the current dynamic theory.

REFERENCES

[1] A. V. BALAKRISHNAN AND L. W. NEUSTADT ed., *Mathematical Theory of Control*, Academic Press, New York, 1967.

[2] R. F. BAUM, *Existence theorems for Lagrange control problems with unbounded time domain*, J. Optim. Theory Appl., 19 (1976), pp. 89–116.

[3] L. D. BERKOVITZ, *Variational methods in problems of control and programming*, J. Math. Anal. Appl., 3 (1961), pp. 145–169.

[4] G. A. BLISS, *Lectures on the Calculus of Variations*, University of Chicago Press, Chicago, 1946.

[5] W. A. BROCK AND A. HAURIE, *On existence of overtaking optimal trajectories over an infinite time horizon.* Math. of Oper. Res. 1 (1976), pp. 337–346.

[6] D. CASS AND K. SHELL, ed., *The Hamiltonian Approach to Dynamic Economics*, Academic Press, New York, 1976.

[7] ———. *The structure and stability of competitive dynamical systems*, Essay III in [6].

[8] C. D. FEINSTEIN, *Implicit programming: A method for characterizing the asymptotic behavior of optimal control trajectories*, Doctoral dissertation, Stanford Univ., Stanford, CA, 1980.

[9] A. F. FILIPPOV, *On certain questions in the theory of optimal control.* SIAM J. Control, ser. A, 1 (1962), pp. 76–84.

[10] D. GALE, *On optimal development in a multi-sector economy*, Rev. of Economic Studies, 34 (1967), pp. 1–18.

[11] G. S. GOODMAN, *The duality of convex functions and Cesari's property (Q)*, J. Optim. Theory Appl., 19 (1976) pp. 17–27.

[12] H. HALKIN, *Necessary conditions for optimal control problems with infinite horizons*, Econometrica, 42 (1974), no. 2, pp. 267–272.

[13] ———, *On the necessary conditions for optimal control of nonlinear systems*, J. Analyse Math. 12 (1964), pp. 1–82.

[14] A. HAURIE, *Existence and global asymptotic stability of optimal trajectories for a class of infinite horizon non-convex systems*, Rep. no. 77–14, Ecole des Hautes Etudes Commerciales, Montreal, Canada, September 1977.

[15] ———, *Optimal control on an infinite time horizon: The turnpike approach.* J. Math. Econ. 3 (1976), pp. 81–102.

[16] M. KURZ, *Optimal economic growth and wealth effects*, Internat. Econom. Rev. 9 (1968) pp. 348–357.

[17] E. B. LEE AND L. MARKUS, *Optimal control for nonlinear processes*, Arch. Rat. Mech. Anal. 8 (1961), pp. 36–58.

[18] D. G. LUENBERGER, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA, 1973.

[19] ———, *Optimization by Vector Space Methods*, John Wiley, New York 1969.

[20] D. W. PETERSON, *A sufficient maximum principle*, IEEE Trans Automat. Control, February 1971, pp. 85–86.

[21] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISCHENKO, *The Mathematical Theory of Optimal Processes*, K. N. Trirogoff, trans, L. W. Neustadt, ed., Interscience, John Wiley, New York, 1962.

[22] F. P. RAMSEY, *A mathematical theory of saving*, Econom. J. 38 (1928), pp. 543–549.

[23] R. T. ROCKAFELLAR, *Conjugate convex functions in optimal control and the calculus of variations*, J. Math. Anal. Appl., 32 (1970), pp. 174–222.

[24] ———, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[25] ———, *Generalized Hamiltonian equations for convex problems of Lagrange*, Pacific J. Math. 33 (1970), pp. 411–427.
[26] ———, *Saddle points of Hamiltonian systems in convex problems of Lagrange*, J. Optim. Theory Appl., 12 (1973), pp. 367–390.
[27] ———, *Saddle points of Hamiltonian systems in convex Lagrange problems having a non-zero discount rate*, Essay IV in [6].
[28] P. P. VARAIYA, *On the trajectories of a differential system*, pp. 115–128, in [1].
[29] L. C. YOUNG, *Lectures on the Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, 1969.
[30] L. E. ZACHRISSON, *Deparametrization of the Pontryagin maximum principle*, pp. 234–245, in [1].

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 5, September 1981

0363-0129/81/1905-0002 $01.00/0

OBSERVER DESIGN FOR LINEAR CONTRACTIVE CONTROL SYSTEMS ON HILBERT SPACES*

TERUO HAMATSUKA,† ABDUL-AZIZ MO'OMEN‡ AND HAJIME AKASHI†

Abstract. This paper is concerned with constructing observers for infinite dimensional linear systems characterized by semigroups on Hilbert spaces. Two types of observers (identity-type and general-type) are considered. Sufficient conditions are given for both types to exist for infinite dimensional systems. The basic tools of our approach make use of the properties of invariant and reducing subspaces of Hilbert spaces and the canonical decomposition of contraction semigroups.

1. Introduction. This paper presents an approach for the design of observers for infinite dimensional linear systems characterized by semigroups on Hilbert spaces. In the finite dimensional case, observers approximately reconstruct missing state-variable information necessary for control [6], [7]. For systems with infinite dimensional state spaces, observers have been considered for some classes of systems, as in [9]. However, the results obtained either involve approximations reducing the problem to the finite dimensional case, or are restricted in application, since they need many assumptions. Gressang and Lamont [4], followed Luenberger [7] and Gopinath [3] in extending the theory of observers to linear systems whose state space is an abstract Banach space. But they made five assumptions which are equivalent to assuming that all unstable systems behavior is restricted to a finite dimensional subspace. The treatment allows one to extend the pole placement property of Luenberger observers to infinite dimensional systems only in a limited sense. However, their assumptions permit one to avoid a crucial problem facing all attempts to generalize results to the infinite dimensional case, that the spectrum of an operator on an infinite dimensional space consists, generally, of something more than eigenvalues and that the number of the spectrum of the operator which lies in the half plane $\operatorname{Re} s > -\varepsilon$, $\varepsilon > 0$, may not even be countable.

Some basic notions about observers are defined in § 2 interpreting the essential feature of the Luenberger observer, that is, the norm of the error in the estimate of the states approaches zero as time increases; for this reason, the observer is also referred to here as an asymptotic state estimator. The design of the identity observer is presented in § 3 and the design of the general-type observer is given in § 4. The technique is based on some rather simple properties of invariant and reducing subspaces of Hilbert spaces, and on a canonical decomposition of contraction semigroups due to B. Sz-Nagy and C. Foias [8] and its applications to the strong stabilizability problem studied by N. Levan et al. [5]. In order to highlight the new technique, a simple example is given at the end of § 4.

2. The definition of an observer. Consider the linear system governed by the equations

$$\frac{d}{dt}x(t) = Ax(t) + Bu(t), \qquad x(0) \in \mathcal{D}(A), \tag{1a}$$

$$y(t) = Cx(t), \tag{1b}$$

* Received by the editors January 29, 1980, and in revised form October 16, 1980.

† Department of Precision Mechanics, Faculty of Engineering, Kyoto University, Kyoto 606, Japan.

‡ Department of Mechanical Engineering, Faculty of Engineering, Ain-Shams University, Abbasia, Cairo, Egypt; on leave at Department of Precision Mechanics, Faculty of Engineering, Kyoto University, Kyoto 606, Japan.

where x, u and y belong to Hilbert spaces $\mathscr{X}$ (the state space), $\mathscr{U}$ (the control space) and $\mathscr{Y}$ (the observation space), respectively. The operator A is closed with dense domain $\mathscr{D}(A)$ in $\mathscr{X}$, and it is always taken to be the infinitesimal generator of a C_0 contraction semigroup, denoted by $T(t)$, $t \geqq 0$, over $\mathscr{X}$. $B: \mathscr{U} \to \mathscr{X}$ and $C: \mathscr{X} \to \mathscr{Y}$ are bounded linear operators. The solution (1a) can be represented by the integral equation

$$x(t) = T(t)x(0) + \int_0^t T(t-s)Bu(s)\,ds, \qquad t > 0, \tag{2}$$

if $u(s)$ is sufficiently smooth in $[0, \infty)$, (for instance continuously differentiable) and it is always so assumed. From the available outputs of system (1) we derive a related system,

$$\frac{d}{dt}z(t) = Fz(t) + Gy(t) + Hu(t), \qquad z(0) \in \mathscr{D}(F), \tag{3}$$

where z belongs to a Hilbert space $\mathscr{Z}$, F is an infinitesimal generator of a C_0-semigroup $S(t)$, $G: \mathscr{Y} \to \mathscr{Z}$ and $H: \mathscr{U} \to \mathscr{Z}$ are bounded operators. For $P \in L(\mathscr{X}, \mathscr{Z})$ (the set of linear bounded operators from $\mathscr{X}$ to $\mathscr{Z}$), we say that (3) is an *asymptotic state estimator* of $Px(t)$ if and only if

$$\lim_{t \to \infty} [z(t) - Px(t)] = 0$$

and P maps $\mathscr{D}(A)$ into $\mathscr{D}(F)$, where $x(t)$ is the solution of (1). Moreover, we can state that (3) is an *observer* of system (1) if the following hold:

i) Equation (3) is an asymptotic state estimator of $Px(t)$.
ii) $PA - FP = GC$.
iii) $H = PB$.
iv) There exist $M \in L(\mathscr{Y}, \mathscr{X})$ and $N \in L(\mathscr{Z}, \mathscr{X})$ such that $MC + NP = I$.

If $P = I$ and $\mathscr{Z} = \mathscr{X}$, then (3) is called an *identity observer*. The *observer error* $e(t)$ is defined by

$$e(t) = z(t) - Px(t).$$

Then the asymptotic estimator requires that

$$\lim_{t \to \infty} e(t) = 0.$$

The purpose of an observer is to provide an approximation to the state of the original system. This approximation is given by

$$\hat{x}(t) = My(t) + Nz(t).$$

By a simple calculation, the error in the approximation is found to be

$$\hat{x}(t) - x(t) = Ne(t).$$

Thus,

$$\|\hat{x}(t) - x(t)\| \leqq \|N\| \|e(t)\| \to 0 \quad \text{as } t \to \infty,$$

since N is a bounded operator. It is easy to show that $e(t)$ satisfies the homogeneous abstract differential equation

$$\frac{d}{dt}e(t) = Fe(t), \qquad e(0) \in \mathscr{D}(F), \tag{4}$$

from the requirement of the asymptotic state estimator, (4) is asymptotically stable.

3. Identity observer. Recall that an observer is an identity observer if the operator P relating the state of the observer to the state of the original system is the identity operator. The operator equation, $PA - FP = GC$, is then reduced to $F = A - GC$. The problem of designing the identity observer can be reduced to one of determining a bounded linear operator G such that $u(t) = -Cx(t)$ is a stabilizing control in

$$\frac{d}{dt}x(t) = Ax(t) + Gu(t), \qquad x(0) \in \mathscr{D}(A).$$

We need the following definitions to prove that the stabilizing operator G exists.

By a *system* (A, B), we mean (1a). A state x in $\mathscr{X}$ is called (*approximately*) *controllable* if for any $\varepsilon > 0$, there is a $u \in \mathscr{L}$ such that the solution of (1a) with $x(0) = 0$ satisfies

$$\|x(t) - x\| < \varepsilon \quad \text{for some } t > 0,$$

where $\mathscr{L}$ is the control set which is continuously differentiable. Then, the set of all controllable states of (A, B) denoted by $\mathscr{X}_c(A, B)$, is said to be the (*approximately*) *controllable subspace*, and

$$\mathscr{X}_c(A, B) = \overline{\bigcup_{t \geqq 0} T(t)B\mathscr{U}}$$

where $\overline{\ }$ denotes closure. The orthogonal complement of $\mathscr{X}_c(A, B)$ on $\mathscr{X}$, denoted by $\mathscr{X}_c^{\perp}(A, B)$, is then,

$$\mathscr{X}_c^{\perp}(A, B) = \bigcap_{t \geqq 0} \operatorname{Ker}[B^* T(t)^*],$$

where the operator with superscript * is the adjoint operator [1].

DEFINITION 1. Let $\mathscr{H}$ be a Hilbert space, and V be a bounded operator in $\mathscr{H}$. We say that a subspace $\mathscr{M}$ *reduces* V if

$$V\mathscr{M} \subset \mathscr{M} \quad \text{and} \quad V^*\mathscr{M} \subset \mathscr{M}.$$

DEFINITION 2. A semigroup $T(t)$, $t \geqq 0$, on $\mathscr{X}$ is called *completely nonunitary* (cnu) if for each nonzero x in $\mathscr{X}$ there is some $t > 0$ such that either $\|T(t)x\| \neq \|x\|$ or $\|T(t)^*x\| \neq \|x\|$.

The following decomposition theorem of contraction semigroups is due to Sz-Nagy and Foias [8].

THEOREM 1. [Sz-Nagy and Foias]. *Let $T(t)$, $t \geqq 0$, be a C_0-contraction semigroup with infinitesimal generator A in a Hilbert space $\mathscr{X}$. Then $\mathscr{X}$ can be decomposed into an orthogonal sum*

$$\mathscr{X} = \mathscr{X}_u(A) \oplus \mathscr{X}_{\text{cnu}}(A),$$

(uniquely) where $\mathscr{X}_u(A)$ and $\mathscr{X}_{\text{cnu}}(A)$ are reducing subspaces for $T(t)$, such that $T(t)$, $t \geqq 0$, admits the unique decomposition

$$T(t) = T_u(t) \oplus T_{\text{cnu}}(t), \qquad t \geqq 0,$$

where the restriction $T_u(t) = T(t)/\mathscr{X}_u(A)$ is unitary and the restriction $T_{\text{cnu}}(t) = T(t)/\mathscr{X}_{\text{cnu}}(A)$ is cnu. *$\mathscr{X}_u(A)$ is the maximal reducing subspace such that the semigroup is unitary, and*

$$\mathscr{X}_u(A) = \{x \in \mathscr{X};\ \|T(t)x\| = \|x\| = \|T(t)^*x\|,\ t \geqq 0\}.$$

It is obvious from this theorem that the subspaces $\mathscr{D}(A) \cap \mathscr{X}_u(A)$ and $\mathscr{D}(A^*) \cap \mathscr{X}_u(A)$ are dense in $\mathscr{X}_u(A)$ (see [5]). Let $\mathscr{M}_s(A)$ and $\mathscr{M}_s(A^*)$ be the sets of strongly stable

states of the system such that

$$\mathcal{M}_s(A)=\{x\in\mathcal{X};\ T(t)x\to 0,\ t\to\infty\},$$

$$\mathcal{M}_s(A^*)=\{x\in\mathcal{X};\ T(t)^*x\to 0,\ t\to\infty\},$$

respectively. It is evident that $\mathcal{M}_s(A)$ and $\mathcal{M}_s(A^*)$ are closed invariant subspaces of $T(t)$ and $T(t)^*$, respectively. It follows from Theorem 1 that $\mathcal{M}_s(A)\cap\mathcal{M}_s(A^*)\subset\mathcal{M}_s(A)\subset\mathcal{X}_{\mathrm{cnu}}(A)$ and $\mathcal{M}_s(A)\cap\mathcal{M}_s(A^*)\subset\mathcal{M}_s(A^*)\subset\mathcal{X}_{\mathrm{cnu}}(A)$. The following theorem is due to Levan and Rigby [5].

THEOREM 2. [Levan and Rigby]. *If the conditions*

i) $\mathcal{X}_{\mathrm{cnu}}(A)=\mathcal{M}_s(A)=\mathcal{M}_s(A^*)$,

and

ii) $\mathcal{X}_u(A)$ *is controllable for the system* (A, B) *or* (A^*, B),

hold, then the system is stabilizable by the feedback $-B^*$.

If we apply Theorem 2 to the system

$$\frac{d}{dt}x(t)=A^*x(t)-C^*y(t),\qquad x(0)\in\mathcal{D}(A^*),$$

then the solution of the design problem of the observer is obtained as follows:

THEOREM 3. *An identity observer can be constructed for the system* (1a) *and* (1b) *if the following conditions hold*:

i) $\mathcal{X}_{\mathrm{cnu}}(A)=\mathcal{M}_s(A)=\mathcal{M}_s(A^*)$,

ii) $\mathcal{X}_u(A)$ *is controllable for the system* (A, C^*) *or* (A^*, C^*).

The stabilizing operator G *is equal to* C^*.

4. General-type observer. The following theorem is the main result of this paper.

THEOREM 4. *Suppose that there exists a closed subspace* $\mathcal{P}$ *of* $\mathcal{X}$ *such that*

i) $\mathcal{P}$ *reduces* $T(t)$, $t\geqq 0$;

ii) $\mathcal{P}^\perp+\mathrm{Range}\ C^*=\mathcal{X}$;

iii) $\mathrm{Ker}\,(C/\mathcal{P}^\perp)^*\supset C\mathcal{P}$.

Let $\tilde{\mathcal{X}}$ *be a quotient space such that* $\tilde{\mathcal{X}}=\mathcal{X}/\mathcal{P}$. *Let* $\tilde{T}(t)$ *and* $\tilde{T}(t)^*$ *be the induced semigroups on* $\tilde{\mathcal{X}}$ *of* $T(t)$ *and* $T(t)^*$, *respectively. Then the observer design problem for system* (1) *can be solved if the following conditions hold*:

iv) $\tilde{\mathcal{X}}_{\mathrm{cnu}}(\tilde{A})=\tilde{\mathcal{M}}_s(\tilde{A})=\tilde{\mathcal{M}}_s(\tilde{A}^*)$,

v) $\tilde{\mathcal{X}}_u(\tilde{A})$ *is controllable for the system* $(\tilde{A}, \widetilde{(C/\mathcal{P}^\perp)})$ *or* $(\tilde{A}, \widetilde{(C/\mathcal{P}^\perp)}^*)$,

where $\tilde{A}$ *and* $\tilde{A}^*$ *are the infinitesimal generators of* $\tilde{T}(t)$ *and* $\tilde{T}(t)^*$, *respectively.*

Remark. The norm on $\tilde{\mathcal{X}}$ is defined in the usual way as

$$\|\tilde{x}\|=\inf_{x\in\tilde{x}}\|x\|,\qquad \tilde{x}\in\tilde{\mathcal{X}},$$

$\tilde{\mathcal{X}}$ is a Hilbert space with inner product

$$(\tilde{x}, \tilde{x}')=(x, x'),$$

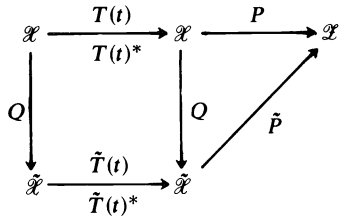
where x and x' are the elements of the cosets $\tilde{x}$ and $\tilde{x}$, respectively, which realize the norm defined above.

Before proceeding to the proof of Theorem 4 we need the following notations and lemmas.

Let $\mathcal{Z}=\mathcal{P}^\perp$, and let P be an orthogonal projection on $\mathcal{Z}$. Let $\tilde{P}$ be an operator from $\tilde{\mathcal{X}}$ to $\mathcal{Z}$ such that $\tilde{P}\tilde{x}=Px$. Then $\tilde{P}$ is one-to-one with the same range as P and a bounded inverse $\tilde{P}^{-1}:\mathcal{Z}\to\tilde{\mathcal{X}}$ exists. Let Q be a canonical projection from $\mathcal{X}$ to $\tilde{\mathcal{X}}$. Since $\mathcal{P}$ reduces $T(t)$, i.e., $T(t)\mathcal{P}\subset\mathcal{P}$ and $T(t)^*\mathcal{P}\subset\mathcal{P}$, there exist operators $\tilde{T}(t)$ and $\tilde{T}(t)^*$ on $\tilde{\mathcal{X}}$ such

that $QT(t) = \tilde{T}(t)Q$, $QT(t)^* = \tilde{T}(t)^*Q$, and $P = \tilde{P}Q$; that is, the following diagram commutes.

(5)



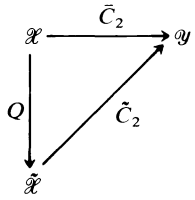
It is easy to prove that $\tilde{T}(t)$ is a contraction semigroup.

Let us decompose the state space $\mathscr{X}$ into subspace $\mathscr{P}$ and its orthogonal complement $\mathscr{P}^\perp$, i.e., $\mathscr{X} = \mathscr{P} \oplus \mathscr{P}^\perp$. Then, the operator $C: \mathscr{X} \to \mathscr{Y}$ can be represented in matrix form by

$$C = [C_1 \quad C_2], \tag{6}$$

where $C_1 = C/\mathscr{P}$ and $C_2 = C/\mathscr{P}^\perp$. Let $\bar{C}_2: \mathscr{X} \to \mathscr{Y}$ be an extension of C_2 to $\mathscr{X}$ such that $\bar{C}_2 = C_2 x_2$ for $x = x_1 + x_2$, $x_1 \in \mathscr{P}$, $x_2 \in \mathscr{P}^\perp$. It is obvious that there exists an operator $\tilde{C}_2: \tilde{\mathscr{X}} \to \mathscr{Y}$ such that $\bar{C}_2 = \tilde{C}_2 Q$, i.e., the diagram

(7)



commutes, since $\mathscr{P} \subset \operatorname{Ker} \bar{C}_2$. G in (3) can be regarded as an operator from $\mathscr{Y}$ to $\mathscr{X}$ such that Range $G \subset \mathscr{P}^\perp = \mathscr{Z}$. Let $\tilde{G} = QG$.

LEMMA 1. *Let* $F = \tilde{P}(\tilde{A} - \tilde{G}\tilde{C}_2)\tilde{P}^{-1}$, *and* $G = C_2^*$. *Then condition* ii) *in* § 2 (*i.e.*, $FP = PA - GC$) *holds*, P *maps* $\mathscr{D}(A)$ *into* $\mathscr{D}(F)$, *and* F *generates a* C_0*-semigroup, say*, $S(t)$, *if* $\operatorname{Ker} C_2^* \supset C\mathscr{P}$.

Proof. From the fact that $\tilde{P}$, $\tilde{P}^{-1}$, $\tilde{G}$ and $\tilde{C}_2$ are bounded and $\tilde{A}$ is an infinitesimal generator of a C_0-semigroup $\tilde{T}(t)$, it is obvious that F generates a C_0-semigroup. Note that the commutative diagram (5) implies $QAx = \tilde{A}Qx$ for $x \in \mathscr{D}(A)$. Thus $Px \in \mathscr{D}(F)$ for $x \in \mathscr{D}(A)$, since $Px = \tilde{P}\tilde{x}$, $\tilde{x} \in \mathscr{D}(\tilde{A})$ means $\tilde{P}\tilde{x} \in \mathscr{D}(F)$ by the definition of F, and $x \in \mathscr{D}(A)$ means $\tilde{x} \in \mathscr{D}(\tilde{A})$. Since $\operatorname{Ker} C_2^* \supset C\mathscr{P}$, $\mathscr{P}$ is GC-invariant. Therefore,

$$\begin{aligned} FP &= \tilde{P}\tilde{A}\tilde{P}^{-1}P - \tilde{P}\tilde{G}\tilde{C}_2\tilde{P}^{-1}P = \tilde{P}\tilde{A}Q - \tilde{P}QG\tilde{C}_2Q \\ &= \tilde{P}QA - \tilde{P}QG\bar{C}_2 = PA - PGC = PA - GC, \end{aligned}$$

since $P = \tilde{P}Q$, $\tilde{G} = QG$, $\tilde{A}Q = QA$, and $PG = G$. This completes the proof. □

LEMMA 2. *Suppose that* $x(t)$ *and* $z(t)$ *are solutions of* (1a) *and* (3) *with* $G = C_2^*$, $F = \tilde{P}(\tilde{A} - \tilde{G}C_2)\tilde{P}^{-1}$, *and* $H = PB$. *Let* $\tilde{x}(t) = Qx(t)$ *and* $\tilde{z}(t) = \tilde{P}^{-1}z(t)$. *Then* $\tilde{x}(t)$ *and* $\tilde{z}(t)$ *satisfy the following abstract differential equations, if* $\operatorname{Ker} C_2^* \supset C\mathscr{P}$:

$$\frac{d}{dt}\tilde{x}(t) = \tilde{A}\tilde{x}(t) + \tilde{B}u(t), \tag{8}$$

$$\frac{d}{dt}\tilde{z}(t) = \tilde{F}\tilde{z}(t) + \tilde{G}\tilde{y}(t) + \tilde{B}u(t), \tag{9}$$

where

$$\tilde{y}(t) = \tilde{C}_2 \tilde{x}(t), \tag{10}$$

$$\tilde{F} = \tilde{A} - \tilde{G}\tilde{C}_2, \tag{11}$$

and

$$\tilde{B} = QB. \tag{12}$$

Proof. Equation (8) is obtained easily by operating with Q on (1a). Note that

$$\tilde{P}^{-1}G = \tilde{P}^{-1}PG = \tilde{P}^{-1}\tilde{P}QG = \tilde{G} \quad \text{(see (5))}$$

and

$$\tilde{P}^{-1}H = \tilde{P}^{-1}PB = \tilde{P}^{-1}\tilde{P}QB = \tilde{B}.$$

Thus, for $\tilde{z}(t)$,

$$\frac{d}{dt}\tilde{z}(t) = \tilde{P}^{-1}(Fz(t) + Gy(t) + Hu(t))$$

$$= \tilde{F}\tilde{z}(t) + \tilde{G}y(t) + \tilde{B}u(t).$$

The condition $\operatorname{Ker} C_2^* \supset C\mathscr{P}$ implies that $\mathscr{P}$ is GC-invariant. Therefore,

$$\tilde{G}y = QGCx = QGC_2x_2 = QG\bar{C}_2x = \tilde{G}\tilde{C}_2\tilde{x} = \tilde{G}\tilde{y}$$

for $x = x_1 + x_2$, $x_1 \in \mathscr{P}$, $x_2 \in \mathscr{P}^{\perp}$ (see (7)). This completes the proof. □

The observer design problem of general-type is then reduced to the problem of identity-type. We summarize the above results as follows:

System equation

$$\frac{d}{dt}\tilde{x}(t) = \tilde{A}\tilde{x}(t) + \tilde{B}u(t),$$

$$\tilde{y}(t) = \tilde{C}_2\tilde{x}(t).$$

Observer equation

$$\frac{d}{dt}\tilde{z}(t) = \tilde{F}\tilde{z}(t) + \tilde{G}\tilde{y}(t) + \tilde{B}u(t).$$

Conditions which must be satisfied

(a) $\tilde{F} = \tilde{A} - \tilde{G}\tilde{C}_2$;

(b) $\tilde{F}$ is an infinitesimal generator of a stable semigroup.

We are ready now to prove Theorem 4.

Proof of Theorem 4. Noting Lemmas 1 and 2, it is enough to prove that $\lim_{t\to\infty}[z(t) - Px(t)] = 0$ and $MC + NP = I$ for some $M \in \mathscr{L}(\mathscr{Y}, \mathscr{X})$ and $N \in \mathscr{L}(\mathscr{Z}, \mathscr{X})$. Since $Q/\mathscr{P}^{\perp} = \tilde{P}^{-1}$, $(\tilde{C}_2)^* = (\widetilde{C_2^*})$. Therefore, applying Theorem 3 to the reduced system (8), (9) and (10), we have $\tilde{z}(t) - \tilde{x}(t) \to 0$, as $t \to \infty$ with $G = C_2^*$. This implies $\tilde{P}^{-1}z(t) - Qx(t) \to 0$, i.e., $z(t) - \tilde{P}Qx(t) \to 0$. Thus we obtain $\lim_{t\to\infty}[z(t) - Px(t)] = 0$. Let us introduce the space $\mathscr{Y} \oplus \mathscr{Z}$, and operators $M \oplus N$ and $C \oplus P$ such that $M \oplus N : \mathscr{Y} \oplus \mathscr{Z} \to \mathscr{X}$; $y \oplus z \mapsto My + Nz$, and $C \oplus P : \mathscr{X} \to \mathscr{Y} \oplus \mathscr{Z}$; $x \mapsto Cx \oplus Px$. The condition $MC + NP = I$ can then be rewritten as

$$(M \oplus N)(C \oplus P) = I,$$

$$\mathscr{X} \xrightarrow{C\oplus P} \mathscr{Y} \oplus \mathscr{Z} \xrightarrow{M\oplus N} \mathscr{X},$$

and the adjoint condition will be

$$(C^*\oplus P^*)(M^*\oplus N^*)=I, \tag{13}$$

$$\mathscr{X}\xleftarrow{C^*\oplus P^*}\mathscr{Y}\oplus\mathscr{Z}\xleftarrow{M^*\oplus N^*}\mathscr{X}.$$

According to the theorem presented by Douglas [2], a necessary and sufficient condition for the existence of the operator M^*+N^* such that the operator equation (13) holds is that

$$\mathscr{X}=\text{Range }(C^*\oplus P^*). \tag{14}$$

Equation (14) implies (ii). This completes the proof. □

Remark. If $\dim \mathscr{Y}<\infty$, then the condition (ii) in Theorem 4 is equivalent to

$$\text{Ker } C\cap\mathscr{P}=0, \tag{ii'}$$

since Range C is closed.

COROLLARY. *Under the conditions of Theorem* 4, *the observer equation will be as follows.*

$$\frac{d}{dt}z(t)=Fz(t)+(C/\mathscr{P}^{\perp})^*y(t)+PBu(t), \tag{15}$$

where

$$F=\tilde{P}(\tilde{A}-\widetilde{(C/\mathscr{P}^{\perp})}^*(\widetilde{C/\mathscr{P}^{\perp}}))\tilde{P}^{-1}.$$

Proof. This is obvious from Lemmas 1 and 2. □

In the following we give an illustrative example.

Example. Let us consider $\mathscr{X}=L_2(0, 2\pi)$ and the heat equation

$$\frac{\partial}{\partial t}x(t,\xi)=\frac{\partial^2}{\partial\xi^2}x(t,\xi),\qquad 0\leqq\xi\leqq 2\pi,\quad t>0,$$

subject to the boundary conditions $x(0)=x(2\pi)$, $x'(0)=x'(2\pi)$. Let the output of the system be given by

$$y=Cx,\qquad Cx=(x,c),\qquad c\in\mathscr{X}.$$

$\mathscr{D}(A)$, the domain of A, is

$$\mathscr{D}(A)=\{x\in\mathscr{X};\ x\in C^2(0,2\pi),\ x, x'\in\mathscr{X},\text{ and } x(0)=x(2\pi),\ x'(0)=x'(2\pi)\},$$

where $A=\partial^2/\partial\xi^2$. Then A is self-adjoint and generates a contraction semigroup $T(t)$ such that

$$T(t)x=\sum_{-\infty}^{\infty}e^{-n^2t}(x,\phi_n)\phi_n,\qquad \phi_n(\xi)=\frac{e^{in\xi}}{\sqrt{2\pi}}.$$

First, let us consider the identity observer. Note that

$$\mathscr{X}_u(A)=\{\phi_0\},$$

$$\mathscr{X}_{\text{cnu}}(A)=\overline{\text{span}}\,\{\phi_n,\ n=\pm1,\pm2,\cdots\}=\mathscr{M}_s(A)=\mathscr{M}_s(A^*),$$

and

$$\mathscr{X}_c^{\perp}(A^*,C^*)=\mathscr{X}_c^{\perp}(A,C^*)=\{x\in\mathscr{X};\ (T(t)x,c)=0,\ t\geqq 0\}.$$

Thus the identity observer is constructable if $(T(t)\phi_0, c) \neq 0, t \geqq 0$. The observer equation is given by

$$\frac{\partial}{\partial t} z(t, \xi) = \frac{\partial^2}{\partial \xi^2} z(t, \xi) - (z, c)c(\xi) + c(\xi)y(t).$$

Next, we will consider the general-type observer. Let $\mathscr{P} = \{\phi_0\}$, then Range $C^* + \mathscr{P}^\perp = \mathscr{X}$ if and only if $(c, \phi_0) \neq 0$, since Range $C^* = \{c(\xi)\}$. Note that

$$Qx = \tilde{x} = \sum_{\substack{-\infty \\ n \neq 0}}^{\infty} (x, \phi_n) + \mathscr{P},$$

$$\mathscr{X} = \overline{\text{span}}\,\{\phi_n, n = \pm 1, \pm 2, \cdots\},$$

$$Px = \tilde{P}\tilde{x} = \sum_{n \neq 0} (x, \phi_n)\phi_n, \qquad \tilde{P}^{-1} z = \sum_{n \neq 0} (z, \phi_n)\phi_n + \mathscr{P},$$

$$\tilde{T}(t)\tilde{x} = \sum_{n \neq 0} (x, \phi_n)\, e^{-n^2 t}\phi_n + \mathscr{P},$$

$$\tilde{\mathscr{X}}_u(\tilde{A}) = 0, \qquad \tilde{\mathscr{X}}_{\text{cnu}}(\tilde{A}) = \tilde{\mathscr{X}} = \tilde{\mathscr{M}}_s(\tilde{A}) = \tilde{\mathscr{M}}_s(\tilde{A}^*).$$

This implies that the observer error approaches zero as time increases, without any feedback. it follows that $F = \partial^2/\partial\xi^2$, since $S(t)z = \tilde{P}\tilde{T}(t)\tilde{P}^{-1}z = \sum_{n \neq 0} (z, \phi_n)\, e^{-n^2 t}\phi_n$. Thus the observer equation is

$$\frac{\partial}{\partial t} z(t, \xi) = \frac{\partial^2}{\partial \xi^2} z(t, \xi), \qquad z \in \mathscr{X},$$

and the estimate of the state of the original system is

$$\hat{x}(t) = (c, \phi_0)^{-1} y(t)\phi_0 + \sum_{n \neq 0} \{1 - (c, \phi_n)\}(z, \phi_n)\phi_n.$$

5. Conclusion. Since the pioneering work of Luenberger, who proposed the so-called observer of asymptotic state estimator, there has been a heavy emphasis on the use of this device for generating missing state information. This paper has considered the problem of the design of the identity type and general type observers for linear systems whose state space is a Hilbert space. Sufficient conditions are given for the existence of both types. As shown, the approach in this paper is based on the properties of invariant and reducing subspaces of Hilbert spaces and on a canonical decomposition of contraction semigroups due to B. Sz-Nagy and C. Foias [8]. Unlike techniques used before, neither approximations reducing the problem to the finite dimensional case, nor assumptions removing difficulties arising from the special characteristics of the spectrum of operators on infinite dimensional spaces are required. Indeed, the only requirement which must be imposed, is the existence of a reducing subspace $\mathscr{P}$. Using the quotient space $\mathscr{X}/\mathscr{P}$, the general type observer was reduced to the identity observer. We believe that our approach can be applied to several design problems of contractive systems.

REFERENCES

[1] A. V. BALAKRISHNAN, *Applied Functional Analysis*, Applications of Mathematics, vol. 3, Springer-Verlag, New York, 1976.

[2] R. G. DOUGLAS, *On majorization, factorization, and range inclusion of operators on Hilbert space*, Proc. Amer. Math. Soc., 17 (1966), pp. 413–415.

[3] B. GOPINATH, *On the control of linear multiple input-output systems*, Bell Syst. Tech. J., 50 (1971), pp. 1063–1081.

[4] R. V. GRESSANG AND G. B. LAMONT, *Observers for systems characterized by semigroup*, IEEE Trans. Automat. Control, AC-20 (1975), pp. 523–528.

[5] N.LEVAN AND L. RIGBY, *Strong stabilizability of linear contractive control systems on Hilbert space*, this Journal, 17 (1979), pp. 23–35.

[6] D. G. LUENBERGER, *Observing the state of a linear system*, IEEE Trans. Mil. Electron., MIL-8 (1964), pp. 74–80.

[7] ———, *Observers for multivariable systems*, IEEE Trans. Automat. Control, AC-11 (1966), pp. 190–197.

[8] B. SZ-NAGY AND C. FOIAS, *Harmonic Analysis of Operators on Hilbert Space*, American Elsevier, New York, 1970.

[9] P. STAVROULAKIS AND P. E. SARACHIK, *Design of optimal controllers for distributed systems using finite dimensional state observers*, presented at the 1973 Decision and Control Conference, San Diego, Calif., December 1973, pp. 105–109.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 5, September 1981

0363-0129/81/1905-0003 $01.00/0

GENERIC OBSERVABILITY OF DIFFERENTIABLE SYSTEMS*

DIRK AEYELS†

Abstract. A dynamical system consists of a smooth vectorfield defined on a differentiable manifold, and a smooth mapping from the manifold to the real numbers. The vectorfield represents the dynamics of a physical system. The mapping stands for a measuring device by which experimental information on the dynamics is made available. The information itself is modeled as a sampled version of the image of the state trajectory under the smooth mapping. In this paper the observability of this set-up is discussed from the viewpoint of genericity. First the observability property is expressed in terms of transversality conditions. Then the theory of transversal intersection is called upon to yield the desired results. It is shown that almost any measuring device will combine with a given physical system to form an observable dynamical system, if $(2n+1)$ samples are taken and not fewer, where n is the dimension of the manifold. Dually, it is shown that almost any physical system will combine with a given measuring device to form an observable dynamical system, if $(2n+1)$ samples are taken and not fewer. The analysis leads to the corollary that for nonlinear systems observability is a generic property, a fact well known for linear systems.

The relation of the theory to the study of turbulence and to control theory is explained.

1. Introduction.

Consider a physical system with dynamics modeled by a smooth vectorfield defined on its state space, viz., a finite dimensional second countable smooth manifold with no boundary. An investigation of the qualitative behavior of the flow of smooth vectorfields has been carried out over the last two decades. In general, from a practical standpoint, the phase portrait cannot be observed directly. Instead, by means of a measuring device—modeled as a smooth function from the configuration space to the reals—experimental information on the dynamical process can be made available. This information is modeled as the image of the state trajectory under the output function, mentioned above. In this manner a mapping has been defined which assigns an output trajectory to each state trajectory. If this mapping is bijective, it makes sense to undertake a study of the state dynamics of the system, starting from experimental evidence. In control theory, a system is a pair consisting of a smooth vectorfield and a smooth output function. A system is observable if the mapping mentioned above is one-to-one.

This paper is concerned with the question whether observability is a natural assumption. Two types of problems will be considered, which in some sense are dual to each other. In the first problem an almost arbitrarily chosen smooth vectorfield is given. The question is whether the choice of a measuring device is "critical" in order to be able to investigate the flow from the output, i.e., in order to achieve observability. In the second problem an almost arbitrarily chosen measuring device is available. The question here is whether the vectorfields that can be investigated with this apparatus, i.e., the vectorfields pairing with the measuring device to form an observable system, constitute a "big" subset of the set of all smooth vectorfields. For both problems, observability turns out to be a generic property if one is allowed to sample the output trajectories $(2n+1)$ times *and not fewer.*

The notion of observability stands central in control theory. It is a necessary assumption in the reconstruction problem, i.e., the problem of recovering the state trajectory corresponding to the sampled output trajectory. Observability is also of interest in the study of turbulence and chaos, as pointed out by an anonymous referee who communicated the following problem. Assume that a dynamical system has a global attractor. Here, instead of finding the state trajectory corresponding to the

* Received by the editors May 20, 1980, and in revised form January 9, 1981.
† Department of System Dynamics, State University of Gent, Gent, Belgium.

observed output trajectory, one is asked to recover a homeomorphic picture of the global attractor or some characteristic properties of the attractor. We will return to this in the final section.

Within the context of linear, algebraic and analytic systems, some aspects of observability as a generic property have been treated in the literature [1], [2], [3]. Here we will be concerned with the general class of smooth nonlinear systems. As opposed to the more algebraic orientation in the above-mentioned references, the results in this paper are obtained as an application of parametric transversality theory.

The organization of the paper is as follows. In the second section the technical definition of observability is given and formulated as a transversality property. In § 3, the two problems mentioned above are considered. The answer to the first question is a rather immediate consequence of a parametric transversality theorem. To obtain the answer to the second question, more work is needed. As usual, verifying the transversality of the evaluation mapping, defined later, is responsible for the main part of the proof. This is harder for the second problem, where transversality has to be achieved by manipulating the vectorfields.

All this manipulation is "filtered" through the fixed output mapping. In fact, one has to show that this filtering process does not affect the transversality of the evaluation mapping. This accounts for the greater length of the proof of the result of the second problem as compared with the first problem, where this type of difficulty does not arise. In § 4 it is shown, by means of a counterexample, that in general at least $2n+1$ samples are necessary in order to achieve observability. In § 5 the paper is concluded with a few additional remarks on turbulence and control theory.

2. Mathematical preliminaries. The observability property. Let X be a C^k differentiable paracompact manifold with k sufficiently high. Let $\xi \in \mathscr{X}^r(X)$, $r \geqq 1$, the space of all C^r vectorfields defined on X. Let $h \in \mathscr{C}^r(X, \mathbb{R})$, $r \geqq 1$, the space of all C^r functions mapping into $\mathbb{R}$. Both $\mathscr{X}^r(X)$ and $\mathscr{C}^r(X, \mathbb{R})$ are endowed with the Whitney C^r topology. The function h is called the output mapping. The results derived in this paper can be easily extended to output functions mapping into more general spaces. The proofs remain virtually unaltered. The classical set theoretic definition of observability goes as follows.

DEFINITION A. The pair (ξ, h) is *observable over a time interval* $[0, T]$, T a positive number, if and only if for each pair $(x, y) \in X \times X \backslash \Delta(X \times X)$ there exists a time $t \in [0, T]$ such that $h \circ \phi_t(x) \neq h \circ \phi_t(y)$. Here $\Delta(X \times X)$ is the diagonal of $X \times X$. It is remarked that in this definition, to each pair of points (x, y) there might correspond a different time instant t.

Transversality theory is not particularly well-adapted for a discussion of Definition A, because of the infinite dimensionality of the output trajectory space. Also, considerations from engineering practice suggest the placement of a sampling device into the system description. Let P be a *sample program*, i.e., a finite set of points $t_i \in [0, T]$, with T given a priori; see also [1].

DEFINITION B. A system (ξ, h) is *observable with respect to P*, or shortly, *P-observable*, if and only if for each $(x, y) \in X \times X \backslash \Delta(X \times X)$, there exists a $t_i \in P$ such that $h \circ \phi_{t_i}(x) \neq h \circ \phi_{t_i}(y)$.

Although stronger in general, Definition B is equivalent to Definition A for linear systems, when considering sample programs with at least n points. For easy reference recall the following density theorem [4] basic to this paper. Let A, X, Y be C^r-manifolds; $\mathscr{C}^r(X, Y)$ is the set of C^r mappings from X to Y and $\rho: A \to \mathscr{C}^r(X, Y)$ is a map which is called a C^r *representation* if and only if the *evaluation map* $\mathrm{ev}_\rho : A \times X \to Y$

with $\mathrm{ev}_\rho(a, x) = \rho(a)(x)$ is a C^r map from $A \times X$ to Y. In the following, we write ρ_a instead of $\rho(a)$ (i.e., $\rho_a : X \to Y$ is a C^r map).

DENSITY THEOREM. *Let W be a submanifold of Y. Define $\mathscr{A}_W \subset \mathscr{A}$ by $\mathscr{A}_W = \{a \in \mathscr{A} : \rho_a \pitchfork W\}$ ($\pitchfork$ is the symbol for transversality),*

Assume that:

(1) *X has finite dimension n and W has finite codimension q in Y.*

(2) *$\mathscr{A}$ and X are second countable.*

(3) *$r > \max(0, n - q)$.*

(4) *$\mathrm{ev}_\rho \pitchfork W$ (this will be called transversality of the evaluation mapping).*

Then $\mathscr{A}_W$ is residual (and hence dense) in $\mathscr{A}$.

As for the connection of observability with transversality theory, it is remarked that observability is equivalent to injectiveness of a particular mapping. Notice also that $f : X \to Y$ is injective if and only if $(f \times f) : X \times X \to Y \times Y$ does not intersect $\Delta(Y \times Y)$ when restricted to $X \times X \backslash \Delta(X \times X)$. Under the right differentiability conditions and when $\dim(X \times X) < \operatorname{codim} \Delta(Y \times Y)$ this is equivalent to $(f \times f) \pitchfork_x \Delta(Y \times Y)$, $x \in X \times X \backslash \Delta(X \times X)$.

Given an observation interval $[0, T]$, by a sample program P we mean a set with at least $(2n+1)$ different points t_i, $0 \leqq t_i \leqq T$ with $n = \dim X$. From now on, P will be assumed to have exactly $2n+1$ points. Given P, define the evaluation mapping

$$\mathrm{ev} : \Delta(\mathscr{A} \times \mathscr{A}) \times (X \times X) \to \mathbb{R}^{2n+1} \times \mathbb{R}^{2n+1},$$

with $\mathscr{A} = \mathscr{X}^r(X) \times \mathscr{C}^r(X, \mathbb{R})$ and

$$\mathrm{ev}(\xi, h, \xi, h, x, y) = \begin{Bmatrix} h \circ \phi_{t_1}(x) \\ \vdots \\ h \circ \phi_{t_{2n+1}}(x) \end{Bmatrix}, \begin{Bmatrix} h \circ \phi_{t_1}(y) \\ \vdots \\ h \circ \phi_{t_{2n+1}}(y) \end{Bmatrix}.$$

Here $\xi \in \mathscr{X}^r(X)$, $h \in \mathscr{C}^r(X, \mathbb{R})$, $x \in X$, $y \in X$. $\phi : X \times \mathbb{R} \to X$ denotes the flow corresponding to ξ.

It is clear that the natural bijection $\delta : \mathscr{A} \to \Delta(\mathscr{A} \times \mathscr{A})$ induces a C^r-manifold structure on $\Delta(\mathscr{A} \times \mathscr{A})$. When X is compact, ev is class C^r [4]. Compactness of X will always be understood in the following. In the concluding remarks, the noncompact case will be treated. In the following section our main concern is showing that the evaluation mapping—with appropriate restrictions on $\mathscr{A}$—is transversal to $\Delta(\mathbb{R}^{2n+1} \times \mathbb{R}^{2n+1})$. This will provide an answer to the problems mentioned in the introduction.

3. Observability is generic. In this section two results on observability as a generic property are proved. In the first result, an almost arbitrarily chosen smooth vectorfield is given. It is shown that almost all smooth output functions pair with the vectorfield to form an observable system if $(2n+1)$ samples are taken. For a motivation, one is referred to the introduction. Before announcing the theorem, we state two lemmas whose proofs are direct consequences of transversality theory.

LEMMA 1. *The subset $\mathscr{A} \subset \mathscr{X}^r(X)$, $r \geqq 1$ of vectorfields with a finite number of equilibrium points and a finite number of closed orbits with period $\leqq T$, constitutes an open and dense set of $\mathscr{X}^r(X)$.*

LEMMA 2. *Let $\xi \in \mathscr{A}$ have a finite number of equilibrium points x_i. The subset $\mathscr{B} \subset \mathscr{C}^r(X, \mathbb{R})$, $r \geqq 1$ of functions h with $h(x_i) \neq h(x_j)$, $i \neq j$, constitutes an open and dense set of $\mathscr{C}^r(X, \mathbb{R})$.*

THEOREM 1. *Given a vectorfield $\xi \in \mathscr{A}$ and a positive real number T, then the set of functions h, belonging to $\mathscr{C}^r(X, \mathbb{R})$ such that (ξ, h) is P-observable, is open and dense in $\mathscr{C}^r(X, \mathbb{R})$. This is true for almost any sample program P of $(2n+1)$ points t_i, $0 \leqq t_i \leqq T$.*

Proof. For the proof of the density part, we will consider output functions belonging to $\mathcal{B}$. If density can be shown with respect to $\mathcal{B}$ then it is also shown with respect to $\mathcal{C}^r(X, \mathbb{R})$.

Let $\phi_t(x)$ denote the flow corresponding to ξ, and let $\{t_i : i = 1, \cdots, 2n+1\}$ denote $(2n+1)$ different sample instants chosen from the interval $[0, T]$. Consider the evaluation mapping

$$\text{ev}: \Delta(\mathcal{B} \times \mathcal{B}) \times (X \times X) \to \mathbb{R}^{2n+1} \times \mathbb{R}^{2n+1},$$

$$\text{ev}(h, h, x, y) = \begin{Bmatrix} h \circ \phi_{t_1}(x) \\ \vdots \\ h \circ \phi_{t_{2n+1}}(x) \end{Bmatrix}, \begin{Bmatrix} h \circ \phi_{t_1}(y) \\ \vdots \\ h \circ \phi_{t_{2n+1}}(y) \end{Bmatrix}.$$

This mapping is class $\mathcal{C}^r$. A pair (ξ, h) is observable if and only if $\text{ev}(h, h, X \times X \backslash \Delta(X \times X)) \cap \Delta(\mathbb{R}^{2n+1} \times \mathbb{R}^{2n+1}) = \varnothing$, which is equivalent to $\text{ev} \pitchfork_{x,y} \Delta(\mathbb{R}^{2n+1} \times \mathbb{R}^{2n+1})$ on $X \times X \backslash \Delta(X \times X)$, since $\text{codim}\, \Delta(\mathbb{R}^{2n+1} \times \mathbb{R}^{2n+1}) = 2n+1 > 2n = \dim(X \times X)$. The application of $2n+1$ samples (or more) is fundamental to this equivalence. Notice that in order to apply the transversality density theorem, the finiteness of the codimension of W in Y is required (for notation, see § 2). It is by considering P-observality that the a priori infinite dimensional space of output curves defined on $[0, T]$ is replaced by a finite dimensional sample space. Conditions (1), (2) and (3) in the density theorem of § 2 are satisfied. In order to satisfy condition (4), we have to show that if $\text{ev}(h^*, h^*, x^*, y^*) =: (w, w) \in \Delta(\mathbb{R}^{2n+1} \times \mathbb{R}^{2n+1})$, then range $(D\ \text{ev}(h^*, h^*, x^*, y^*))$ contains a complement of $T_{w,w}\Delta(\mathbb{R}^{2n+1} \times \mathbb{R}^{2n+1})$ in $T_w\mathbb{R}^{2n+1} \times T_w\mathbb{R}^{2n+1}$. D denotes the derivative, and T denotes the tangent space. When we show, by picking appropriate functions $g \in \mathcal{B}$, that

$$\frac{d}{d\lambda}\bigg|_{\lambda=0} (\text{ev}(h^* + \lambda g, h^* + \lambda g, x^*, y^*)), \qquad \lambda \in \mathbb{R}$$

can span $T_w\mathbb{R}^{2n+1} \times \{0\}$ or $\{0\} \times T_w\mathbb{R}^{2n+1}$, the proof will be finished. Now

$$\frac{d}{d\lambda}\bigg|_{\lambda=0} (\text{ev}(h^* + \lambda g, h^* + \lambda g, x^*, y^*))$$

$$= \begin{Bmatrix} g \circ \phi_{t_1}(x^*) \\ \vdots \\ g \circ \phi_{t_{2n+1}}(x^*) \end{Bmatrix}, \begin{Bmatrix} g \circ \phi_{t_1}(y^*) \\ \vdots \\ g \circ \phi_{t_{2n+1}}(y^*) \end{Bmatrix}.$$

Therefore if x^* is not an equilibrium point, then it is possible to pick a $g \in \mathcal{B}$ such that

$$\begin{Bmatrix} g \circ \phi_{t_1}(x^*) \\ \vdots \\ g \circ \phi_{t_{2n+1}}(x^*) \end{Bmatrix}$$

equals any vector $\alpha \in T_w\mathbb{R}^{2n+1} \approx \mathbb{R}^{2n+1}$, a priori given, and we are done. Problems could arise if both x^* and y^* were equilibrium points. This cannot occur by assumption, since then $\text{ev}(h^*, h^*, x^*, y^*) \notin \Delta(\mathbb{R}^{2n+1} \times \mathbb{R}^{2n+1})$. When x^* and y^* are both on closed orbits with the same period (or x^* is an equilibrium and y^* is on a closed orbit) a periodic sample program (with period equal to the period of the closed orbits involved) is an obstruction to the proof of the transversality of the evaluation mapping. But this phenomenon cannot occur, due to the phrase "almost any sample program P" in the statement of the

theorem. Indeed, by this phrase we mean to exclude the above-mentioned periodic sample programs.

The openness part of the theorem is an immediate consequence of the openness-of-transversality theorem [4]. □

Remarks. 1. The theorem is valid for $\xi \in \mathscr{A}$ and thus for vectorfields belonging to an open and dense subset of $\mathscr{X}^r(X)$.

2. Theorem 1 implies that the set of P-observable (or observable) pairs (ξ, h) is open and dense in $\mathscr{X}^r(X) \times \mathscr{C}^r(X, \mathbb{R})$. This is a well-known result for linear systems. It is remarked that for linear systems the genericity of observability can be shown in a direct way by considering the algebraic characterization of observability in terms of the Kalman matrix.

We now proceed to prove a theorem dual to Theorem 1. Here an almost arbitrarily chosen smooth output function is given. It is shown that almost all smooth vectorfields pair with the output function to form an observable system if $2n+1$ samples are taken. Although the statement of this theorem is dual to the previous theorem, the proof is not. Indeed, showing that the appropriate evaluation mapping is transversal is somewhat involved. For the intuitive reason behind this, and also for a motivation of the problem one is again referred to the introduction. Before stating the theorem, we state two lemmas whose proofs follow directly from transversality theory.

LEMMA 3. *The set of functions $\mathscr{D} \subset \mathscr{C}^r(X, \mathbb{R})$, $r \geqq 1$, with a finite number of nondegenerate critical points x_i, and with $h(x_i) \neq h(x_i)$, $i \neq j$ constitutes an open and dense set of $\mathscr{C}^r(X, \mathbb{R})$.*

LEMMA 4. *Given $h \in \mathscr{D}$, consider the set of vectorfields $\mathscr{E} \subset \mathscr{X}^r(X)$ which is the intersection of the sets $\mathscr{E}^1$, $\mathscr{E}^2$ and $\mathscr{E}^3$ defined, respectively, by the following.*

1) *No two equilibrium points belong to the same level surface of h.*

2) *No two equilibrium points coincide with critical points of h.*

3) *No integral curve contains two (or more) critical points of h.*

The set $\mathscr{E}$ is an open and dense set of $\mathscr{X}^r(X)$.

THEOREM 2. *Given a function $h \in \mathscr{D}$ and a positive real number T, then the set of vectorfields ξ belonging to $\mathscr{X}^r(X)$ such that (ξ, h) is P-observable is open and dense in $\mathscr{X}^r(X)$. This is true for almost any sample program of $2n+1$ points t_i, $0 \leqq t_i \leqq T$.*

Proof. For the proof of the density part of Theorem 2, we will consider vectorfields belonging to $\mathscr{E}$. If density can be shown with respect to $\mathscr{E}$ then it is also shown with respect to $\mathscr{X}^r(X)$.

Consider the evaluation mapping

$$\operatorname{ev}: \Delta(\mathscr{E} \times \mathscr{E}) \times X \times X \to \mathbb{R}^{2n+1} \times \mathbb{R}^{2n+1},$$

$$\operatorname{ev}(\xi, \xi, x, y) = \begin{Bmatrix} h \circ \phi_{t_1}(x) \\ \vdots \\ h \circ \phi_{t_{2n+1}}(x) \end{Bmatrix}, \begin{Bmatrix} h \circ \phi_{t_1}(y) \\ \vdots \\ h \circ \phi_{t_{2n+1}}(y) \end{Bmatrix},$$

with $h \in \mathscr{D}$, $\phi_t(x)$ the flow corresponding with ξ, and all t_i different. The only difficulty in applying the transversality density theorem, is in proving the transversality of the evaluation mapping. The other conditions are satisfied. We have to show that, given

$$(w, w) := \operatorname{ev}(\xi^*, \xi^*, x^*, y^*) \in \Delta(\mathbb{R}^{2n+1} \times \mathbb{R}^{2n+1}),$$

range $(D \operatorname{ev}(\xi^*, \xi^*, x^*, y^*))$ contains a complement of

$$T_{w,w}\Delta(\mathbb{R}^{2n+1} \times \mathbb{R}^{2n+1}) \quad \text{in } T_w\mathbb{R}^{2n+1} \times T_w\mathbb{R}^{2n+1}.$$

We evaluate the derivative with the partial derivative rule [4]. With $\eta \in \mathbb{R}$, $\eta \in \mathscr{E}$, ϕ^{λ} the flow of $\xi^* + \lambda\eta$, we obtain

$$\begin{aligned}
&\frac{d}{d\lambda}\Big|_{\lambda=0} \operatorname{ev}(\xi^* + \lambda\eta, \xi^* + \lambda\eta, x^*, y^*) \\
&\quad = \frac{d}{d\lambda}\Big|_{\lambda=0} \left(\begin{Bmatrix} h \circ \phi^{\lambda}_{t_1}(x^*) \\ \vdots \\ h \circ \phi^{\lambda}_{t_{2n+1}}(x^*) \end{Bmatrix}, \begin{Bmatrix} h \circ \phi^{\lambda}_{t_1}(y^*) \\ \vdots \\ h \circ \phi^{\lambda}_{t_{2n+1}}(y^*) \end{Bmatrix} \right) \\
&\quad = \begin{Bmatrix} Dh|_{\phi_{t_1}(x^*)} \cdot \int_0^{t_1} D\phi^0_s \cdot \eta(\phi^0_{-s+t_1}(x^*))\, ds \\ \vdots \\ Dh|_{\phi_{t_{2n+1}}(x^*)} \cdot \int_0^{t_{2n+1}} D\phi^0_s \cdot \eta(\phi^0_{-s+t_{2n+1}}(x^*))\, ds \end{Bmatrix}, \qquad (1) \\
&\qquad \begin{Bmatrix} Dh|_{\phi_{t_1}(y^*)} \cdot \int_0^{t_1} D\phi^0_s \cdot \eta(\phi^0_{-s+t_1}(y^*))\, ds \\ \vdots \\ Dh|_{\phi_{t_{2n+1}}(y^*)} \cdot \int_0^{t_{2n+1}} D\phi^0_s \cdot \eta(\phi^0_{-s+t_{2n+1}}(y^*))\, ds \end{Bmatrix}.
\end{aligned}$$

Recall that $x^* \neq y^*$, since transversality of the evaluation mapping has to be verified on $X \times X \backslash \Delta(X \times X)$. We consider three cases.

Case 1 Neither x^* nor y^* is an equilibrium point of ξ^*.

For each $i \in \{1, 2, \cdots, (2n+1)\}$, at least one of $Dh|_{\phi_{t_i}(x^*)}$ or $Dh|_{\phi_{t_i}(y^*)}$ is not equal to zero. Otherwise h is critical in $\phi_{t_i}(x^*)$ and $\phi_{t_i}(y^*)$, $i = 1, \cdots, 2n+1$, and since $h \in \mathscr{D}$, this implies that $h(\phi_{t_i}(x^*)) \neq h(\phi_{t_i}(y^*))$, $i = 1, 2, \cdots, 2n+1$, and thus

$$\operatorname{ev}(\xi^*, \xi^*, x^*, y^*) \notin \Delta(\mathbb{R}^{2n+1} \times \mathbb{R}^{2n+1}).$$

Without loss of generality, we assume

$$Dh|_{\phi_{t_i}(x^*)} \neq 0, \qquad i = 1, 2, \cdots, 2n+1. \qquad (2)$$

We show that for any vector $\alpha \in T_w\mathbb{R}^{2n+1}$, we can find a C^r vectorfield η_α in $\mathscr{E}$ such that

$$\frac{d}{d\lambda}\Big|_{\lambda=0} \operatorname{ev}(\xi^* + \lambda\eta_\alpha, \xi^* + \lambda\eta_\alpha, x^*, y^*) = \alpha \times \{0\}.$$

The vectorfield η_α will first be defined on the range of the integral curve from x^* to $\phi_{t_{2n+1}}(x^*)$ and then extended to the whole manifold by a partition of unity argument. We denote by α_k the kth component of the vector α. We now show how to define η_α such that the first component of

$$\frac{d}{d\lambda}\Big|_{\lambda=0} \operatorname{ev}(\xi^* + \lambda\eta_\alpha, \xi^* + \lambda\eta_\alpha, x^*, y^*)$$

equals $(\alpha_1, 0)$.

Choose η_α such that $\eta_\alpha(\phi^0_{-s+t_1}(x^*)) = g(s)D\phi^0_{-s} \cdot \dot{x}$, $0 \leqq s \leqq t_1$. The function $g : [0, t_1] \to \mathbb{R}$ is such that $\int_0^{t_1} g(s)\, ds = 1$, and has the property that its first r derivatives evaluated at 0 and t_1, are equal to zero. Taking into account (1), and taking $\eta_\alpha = 0$ on the range of the integral curve from y^* to $\phi_{t_1}(y^*)$, one finds that the first component of the

derivative of the evaluation map equals $(Dh|_{\phi_{t_1}(x^*)} \cdot \dot{x}, 0)$, which is equal to $(\alpha_1, 0)$ for a good choice of $\dot{x}$. The proof will be finished by induction. Let the vectorfield η_α be defined on the range of the integral curve connecting x^* to $\phi_{t_{k-1}}(x^*)$.

The vectorfield η_α has also been defined simultaneously on the range of the integral curve connecting y^* to $\phi_{t_{k-1}}(y^*)$, where, because of (2), it has been taken equal to zero. We show how to define η_α such that the kth component of

$$\left.\frac{d}{d\lambda}\right|_{\lambda=0} \operatorname{ev}(\xi^* + \lambda\eta_\alpha, \xi^* + \lambda\eta_\alpha, x^*, y^*)$$

equals $(\alpha_k, 0)$.

Therefore, consider

$$\begin{aligned} &Dh|_{\phi_{t_k}(x^*)} \int_0^{t_k} D\phi_s^0 \cdot \eta(\phi^0_{-s+t_k}(x^*))\, ds \\ &\quad = Dh|_{\phi_{t_k}(x^*)} \left(\left(\int_0^{t_{k-1}} + \int_{t_{k-1}}^{t_k} \right) (D\phi_s^0 \cdot \eta(\phi^0_{-s+t_k}(x^*))\, ds) \right) \\ &\quad = Dh|_{\phi_{t_k}(x^*)}(\dot{y} + \dot{w}). \end{aligned}$$

Here $\dot{y}$ has been chosen already. The vector $\dot{w}$ can be chosen arbitrarily outside $\ker(Dh|_{\phi_{t_k}(x^*)})$ as indicated above. Therefore transversality of the evaluation mapping has been shown in Case 1.

Case 2 One of x^* or y^*, say x^*, is an equilibrium point of ξ^*.

In this case

$$\left.\frac{d}{d\lambda}\right|_{\lambda=0} (\operatorname{ev}(\xi^* + \lambda\eta, \xi^* + \lambda\eta, x^*, y^*)) = \left\{ \begin{array}{l} Dh(x^*) \cdot \eta(x^*) \cdot t_1 \\ \vdots \\ Dh(x^*) \cdot \eta(x^*) \cdot t_{2n+1} \end{array} \right\}, \quad \left\{ \begin{array}{l} Dh|_{\phi_{t_1}(y^*)} \cdot \int_0^{t_1} D\phi_s^0 \cdot \eta(\phi^0_{-s+t_1}(y^*))\, ds \\ \vdots \\ Dh|_{\phi_{t_{2n+1}}(y^*)} \cdot \int_0^{t_{2n+1}} D\phi_s^0 \cdot \eta(\phi^0_{-s+t_{2n+1}}(y^*))\, ds \end{array} \right\}.$$

Since $\xi \in \mathscr{C}^2$, η can be defined in x^* such that the left column has an arbitrary component with arbitrary value (the other values of the components are forced). Since $\xi \in \mathscr{C}^3$, $Dh|_{\phi_{t_i}(y^*)} = 0$ for at most one particular $i \in \{1, \cdots, 2n+1\}$, say for $i = m$. Assume $Dh|_{\phi_{t_m}(y^*)} = 0$. By constructing η on the range of the integral curve y^* to $\phi_{t_{2n+1}}(y^*)$, as in Case 1, and taking $Dh(x^*) \cdot \eta(x^*) t_m$ appropriately, it is shown that the evaluation mapping is transversal.

Case 3 Both x^* and y^* are equilibrium points of ξ^*.

Since $\xi \in \mathscr{C}^1$, this case need not be considered because

$$h \circ \phi_{t_i}(x^*) = h(x^*) \neq h(y^*) = h \circ \phi_{t_i}(y^*), \qquad i = 1, \cdots, 2n+1.$$

Hence $\operatorname{ev}(\xi^*, \xi^*, x^*, y^*) \notin \Delta(\mathbb{R}^{2n+1} \times \mathbb{R}^{2n+1})$.

For the openness part we need only refer to the openness-of-transversality theorem [4]. □

Remark. The theorem is valid for $h \in \mathscr{D}$ and thus for functions belonging to an open and dense subset of $\mathscr{C}^r(X, \mathbb{R})$.

4. Counterexample. In this section we provide a counterexample to speculations that $2n$ or fewer samples might suffice to achieve observability generically. The counterexample will be constructed on S^1.

Given two arbitrary sample instants 0 and t_1 (without loss of generality, 0 is taken as the first sample instant) we will construct a vectorfield on S^1 with flow $\phi_t(x)$ and an output mapping h from S^1 to $\mathbb{R}$ such that the mapping from S^1 to $\mathbb{R}^2$ defined by $x \to (h(x), h \circ \phi_{t_1}(x))(*)$ is not injective and such that small perturbations in the flow and the output mapping do not destroy the noninjectiveness. Notice first that there exists a smooth mapping from S^1 to $\mathbb{R}^2$ which twists the circle into a figure eight. The idea of the counterexample is to find a vectorfield and an output mapping which does something similar, i.e., such that the mapping (*) maps S^1 into a figure with transversal self-intersections.

Define a vectorfield on S^1 (with coordinate θ denoting the angle), by $\dot{\theta} = \pi/t_1$. The output is defined on S^1 considered as the closed interval $[0, 1]$ with 0 and 1 identified. It is given by the following function, C^∞-smoothed at the discontinuities

$$\begin{aligned} h &= 1, \quad 0 < t < \tfrac{1}{4}, \quad \tfrac{3}{4} < t < 1, \\ h &= 2, \quad \tfrac{1}{4} < t < \tfrac{3}{8}, \quad \tfrac{5}{8} < t < \tfrac{3}{4}, \\ h &= 0, \quad \tfrac{3}{8} < t < \tfrac{5}{8}. \end{aligned}$$

The mapping (*) with ϕ and h as defined is not injective, since it maps S^1 into a figure with transversal crossings. The crossings are stable and thus noninjectiveness is preserved under perturbations of the vectorfield, the output mapping, and the sample times. Therefore Theorems 1 and 2 are no longer valid for sample programs consisting of $2n$ samples.

5. Concluding remarks.

1. In the previous theorems we have been confined to compact manifolds. The results can be extended to noncompact manifolds if "open and dense" is replaced by "residual", i.e., a countable intersection of open and dense sets, with respect to the Whitney topology. The procedure by which the extension is carried out is standard. For more details one is referred to [5].

2. This remark is related to the definition of observability. As defined in § 2 observability or P-observability is expressed in set theoretic terms. It is natural, when considering differentiable systems, to require not just one-to-one-ness of the relevant mapping in the definition of observability, but also some differentiability condition on this functional relationship. As far as P-observability is concerned, we propose to call a system (ξ, h) P-observable over $[0, T]$ if and only if the mapping $x \to (h \circ \phi_{t_1}(x), \cdots, h \circ \phi_{t_{2n+1}}(x))$ *embeds* X into $\mathbb{R}^{2n+1}$.

This definition also takes into account remarks put forward by Kalman and Sontag [6], to the effect that the inverse relation (from output curve to initial condition) should be differentiable.

Theorems 1 and 2 remain true with this stronger notion of observability. Indeed, transversality theory again provides the right framework. A proof amounts to producing the right evaluation mapping (with a codomain involving jet spaces) and then going through steps similar to those in this paper.

3. We have shown that given $2n+1$ time instants l_i with $0 < l_1 < l_2 < \cdots < l_{2n+1}$, the mapping $(h \circ \phi_{l_1}, \cdots, h \circ \phi_{l_{2n+1}})$ generically embeds X into $\mathbb{R}^{2n+1}$. Notice that a state trajectory $\phi_t(p)$, $t \geqq 0$, starting in $p \in X$ at time zero is mapped into $(h \circ \phi_{t+l_1}(p), \cdots, h \circ \phi_{t+l_{2n+1}}(p))$, $t > 0$. Therefore the ω-limit set associated with p is homeomorphic with the limit set of the curve $t \to (h \circ \phi_{t+l_1}(p), \cdots, h \circ \phi_{t+l_{2n+1}}(p))$.

Assume that the state dynamics has a global attractor; then by studying the limit set of $t \to (h \circ \phi_{t+l_1}(p), \cdots, h \circ \phi_{t+l_{2n+1}}(p))$, which is available from experiment, one obtains

information on the global attractor. F. Takens, in a forthcoming paper, constructs algorithms based on experimental evidence which provide estimates of the dimension and the topological entropy of the attractor. This gives, in principle, a strategy for testing the Ruelle–Takens picture [7], where the onset of turbulence is caused by the presence of strange attractors.

4. Given $h \in \mathscr{D}$, and also a parametrized set of vectorfields of $\mathscr{E}$, i.e., a smooth path defined on $[0, 1]$ (or a higher dimensional interval) into $\mathscr{E}$, a similar argument to that in Theorem 2 implies that for an open and dense set of paths, the vectorfield corresponding with an arbitrary but a priori given point in $[0, 1]$ and the function h are an observable pair. Translated to a control context, this means that for a given output function and a control system, any constant control almost always distinguishes pairs of points on the manifold. Thus, in a sense, the question whether couples of points are distinguishable or not with respect to an observed dynamical system has no bearing upon the controls available in the dynamics of the system.

Acknowledgments. The author is grateful to D. L. Elliott for numerous discussions and for raising the questions which lead to this paper. The author also acknowledges several interesting comments made by an anonymous referee.

REFERENCES

[1] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
[2] E. SONTAG, *On the observability of polynomial systems*, I: *Finite-time problems*, this Journal, 17 (1979), pp. 139–151.
[3] H. SUSSMANN, *Generic single-input observability of continuous-time polynomial systems*, IEEE Trans. Automat. Control, to appear.
[4] R. ABRAHAM AND J. W. ROBBIN, *Transversal Mappings and Flows*, W. A. Benjamin, New York, 1967.
[5] M. M. PEIXOTO, *On an approximation theorem of Kupka and Smale*, J. Differential Equations, 3 (1966), pp. 214–227.
[6] E. SONTAG, *Polynomial Response Maps*, Springer-Verlag, New York, 1979.
[7] D. RUELLE AND F. TAKENS, *On the nature of turbulence*, Comm. Math. Phys., 20 (1971), pp. 167–192; 23 (1971), pp. 343–344.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 5, September 1981

0363-0129/81/1905-0004 $01.00/0

A CHARACTERIZATION OF WELL-POSED OPTIMAL CONTROL SYSTEMS*

T. ZOLEZZI†

Abstract. A constrained optimal regulator problem is considered. Continuous dependence of the optimal control on the desired trajectory (Hadamard well-posedness) or convergence toward the optimal control of any minimizing sequence (Tykhonov well-posedness) are proved when the dynamics are affine (linear plus constant). Dense well-posedness is obtained in the non-affine case. A necessary and sufficient condition for well-posedness for all desired trajectories is shown to be the affine structure of the plant.

Introduction. We consider well-posedness properties of quadratic optimal control problems described by ordinary differential systems. We wish to minimize the quadratic performance

$$(1) \qquad \int_0^T [(u-u^*)'Q(u-u^*)+(x-x^*)'P(x-x^*)]\,dt+[x(T)-y^*]'E[x(T)-y^*]$$

on the trajectories (v, u, x) of the ordinary differential system

$$(2) \qquad \begin{aligned} &\dot{x}(t)=g(t, x(t), u(t)), \quad \text{a.e. in } (0, T),\\ &x(0)=v, \end{aligned}$$

subject to constraints of the general form

$$(3) \qquad (v, u, x)\in K.$$

Here (v, u) is the control and x is the state.

The precise assumptions about the matrices Q, P, E, the function g and the set K will be listed in the next section.

We assume that the Cauchy problem (2) uniquely defines the state variable x on the whole time interval $[0, T]$ for every $v\in R^m$ and $u\in L^2=L^2(0, T)$ of the appropriate dimension. To each desired trajectory

$$z^*=(y^*, u^*, x^*)\in R^m\oplus L^2\oplus L^2$$

there corresponds a (nonlinear) constrained optimal regulator problem (1), (2), (3). From a theoretical as well as practical standpoint it is useful to know in advance which functions g define well behaved optimization problems (1), (2), (3) for as wide a class of desired trajectories as possible.

We consider two properties of well-posedness of the optimal control problem (1), (2), (3). The first one is Tykhonov well-posedness [1]. We require existence and uniqueness of the optimal control $(\bar{v}, \bar{u})$ and strong convergence in $R^m\oplus L^2$ of any minimizing sequence to $(\bar{v}, \bar{u})$. Therefore Tykhonov well-posedness implies unique solvability together with convergence of the numerical methods of minimization for the optimal control problems (1), (2), (3).

The optimal control problem described above will be said to be well-posed in the sense of Hadamard (see [2, remark 1, p. 164]) provided unique solvability holds along with continuous dependence of the optimal control on the desired trajectory between the strong topologies involved. It is easily seen that Hadamard well-posedness implies

* Received by the editors May 16, 1980, and in revised form July 28, 1980. This work was supported in part by Laboratorio per la Matematica Applicata del C.N.R. Genova.
† Istituto Matematico, via L. B. Alberti, 4-16132, Genova, Italy.

continuous dependence on z^* of the optimal control, the optimal state and the value under mild regularity conditions on g.

The practical meaning of such a well-posedness is clear: suitably small changes of the desired trajectory result in arbitrarily small deviations of the corresponding optimal control, state and value. In particular, an a priori knowledge of Hadamard well-posedness is useful in connection with numerical methods of solution of the optimal control problems, which require approximations of the data. We refer to [12] for a characterization of continuous dependence on the coefficients in the plant of optimal controls of linear regulator problems. A further motivation for the study of this type of well-posedness is given by optimal control problems with desired trajectories either depending on parameters or only approximately known.

From a more general point of view (see [3]) the mathematical structure of an optimization problem can be revealed by finding its variational stability properties under data perturbations.

There exist some relationships between Tykhonov and Hadamard well-posedness, in particular the former implies the latter for some problems of best approximation [2, theorem, p. 164].

Let us remark that if K is a closed convex set and g is an affine function, that is

$$g(t, x, u) = A(t)x + B(t)u + C(t) \tag{4}$$

with appropriate matrices $A(t)$, $B(t)$, $C(t)$, then the corresponding affine regulator problem (1), (2), (3) can be naturally viewed as a convex best approximation problem in a Hilbert space, whenever the matrices P, Q, E in (1) are positive (semi) definite in a suitable way. This geometrical interpretation automatically gives Tykhonov well-posedness for all desired trajectories in this affine case as a consequence of the classical Riesz projection theorem.

This paper is devoted to a study of well-posedness of (1), (2), (3) mainly in the nonlinear case. The existence of optimal controls may not hold in general, as is well known; hence well-posedness may fail also.

In § 1 of this paper the well-posedness of the affine regulator problem with convex constraints is considered, both in Tykhonov and Hadamard sense.

In § 2 it is shown that well-posedness obtains for a dense set of desired trajectories under mild regularity assumptions about the nonlinear function g. These results are direct outcomes of [4] or [5]. Unfortunately, except in particular cases, nothing seems to be known about the explicit structure of this dense set of trajectories, except its existence and topological properties. We refer to [4], [5], [6], [7], [8], [9], [10], [11] for information on generic properties of existence of solutions and well-posedness for optimization problems (with some applications to optimal control theory).

In § 3 the main result is given as follows. Tykhonov or Hadamard well-posedness of problem (1), (2) without constraints for all desired trajectories is equivalent to (4), under some (possibly superfluous) regularity assumptions on g. Therefore, under such assumptions, the class of control systems (2) such that every quadratic optimal control problem (1) is Tykhonov or Hadamard well-posed, is that of affine control systems. This gives a variational characterization that seems to be of interest and capable of extensions to other control systems.

In the last section of the paper explicit conditions about the function g are given which yield the (previously required) continuous dependence of the state on the control.

A preliminary version of these results was presented at the 9th IFIP Conference on Optimization Techniques held in Warsaw (September, 1979).

Notation, statement of the problem and assumptions. Given a positive number T we shall denote briefly by L^p the usual Lebesgue space of (classes of) p-summable (essentially bounded when $p=\infty$) functions defined on $[0, T]$, with values in a real euclidean space whose dimension will be clear from the context. Moreover we denote by $\to$ the strong convergence, by $\rightharpoonup$ the weak one, by $\langle\cdot,\cdot\rangle$ either the inner product in any Hilbert space or the pairing between the space and its dual, by $H_1 \oplus H_2$ the direct sum of the Hilbert spaces H_1 and H_2 equipped with the scalar product

$$\langle (x_1, x_2), (y_1, y_2)\rangle = \langle x_1, y_1\rangle_1 + \langle x_2, y_2\rangle_2$$

where $\langle\cdot,\cdot\rangle_i$ is the scalar product in H_i, by $DI(u)$ the Gateaux derivative of the real-valued function I evaluated at u. A prime denotes transpose. A mapping f between real vector spaces is called affine whenever

$$f(au+(1-a)v) = af(u)+(1-a)f(v)$$

for every real a and all u, v.

We are given a positive number T, a function

$$g:[0, T]\times R^m \times R^k \to R^m,$$

two symmetric matrices P, Q with entries belonging to L^∞, of respective dimensions $m\times m$, $k\times k$, a constant symmetric $m\times m$ matrix E, two nonempty subsets

$$K \subset R^m \oplus L^2 \oplus L^2, \qquad K^* \subset L^2 \oplus L^2.$$

The problem we shall consider is the following. Given the desired trajectory

$$z^* = (y^*, u^*, x^*) \in R^m \oplus L^2 \oplus L^2$$

we wish to minimize the performance

$$(1)\quad I(v, u) = \int_0^T [(u-u^*)'Q(u-u^*)+(x-x^*)'P(x-x^*)]\,dt + [x(T)-y^*]'E[x(T)-y^*]$$

on the trajectories (v, u, x) of the control system

$$(2)\quad \begin{aligned} &\dot{x}(t) = g(t, x(t), u(t)), \quad \text{a.e. in } (0, T),\\ &x(0) = v, \end{aligned}$$

subject to one of the following constraints

$$(3)\qquad (v, u, x) \in K,$$

$$(3^*)\qquad (u, x) \in K^*.$$

We remark that constrained affine regulator problems arise if $g(t,\cdot,\cdot)$ is an affine function for every $t\in[0, T]$. This obtains if and only if for every $t\in[0, T]$ there exist matrices $A(t)$, $B(t)$, $C(t)$ of the appropriate dimensions such that for every $x\in R^m$ and $u\in R^k$

$$(4)\qquad g(t, x, u) = A(t)x + B(t)u + C(t).$$

The pair $(v, u)\in R^m\oplus L^2$ will be referred to as a control, and any absolutely continuous solution x of (2) as a state (corresponding to (v, u)). The control (v, u) is called admissible if there exists a state x corresponding to (v, u) such that (v, u, x) satisfies (3) or (3*). Given z^*, the corresponding optimal control problem will be referred to briefly as problem (1), (2), (3) or (3*), and as problem (1), (2) whenever it is unconstrained.

For any problem (1), (2), (3) or (3*) considered in the following, we will assume that some admissible control exists.

The following definitions will be used throughout the paper. Given a nonempty subset $D \subset R^m \oplus L^2 \oplus L^2$, any of the problems (1), (2), (3) or (3*) will be called:

Tykhonov well-posed in D (see [1]) if and only if for every $z^* \in D$ there exists an unique optimal control $(\bar{v}, \bar{u}) \in R^m \oplus L^2$, and for every minimizing sequence of admissible controls $(v_n, u_n) \in R^m \oplus L^2$, that is

$$I(v_n, u_n) \to \inf \{I(v, u) : (u, v) \text{ admissible}\}$$

we have

$$v_n \to \bar{v} \text{ in } R^m, \qquad u_n \to \bar{u} \text{ in } L^2;$$

Hadamard well-posed in D iff for every $z^* \in D$ there exists an unique optimal control $(\bar{v}, \bar{u})$, and the mapping

$$z^* \to (\bar{v}, \bar{u})$$

is continuous in D between the strong topologies of $R^m \oplus L^2 \oplus L^2$ and of $R^m \oplus L^2$. When D is the whole space, we shall say briefly that the problem is well-posed for every desired trajectory.

The following assumptions will be used in the next sections.

(5) For every $t_0 \in [0, T]$, $v \in R^m$, $u \in L^2$ there exists a unique absolutely continuous function x, defined in the whole interval $[0, T]$ such that $x(t_0) = v$, $\dot{x} = g(t, x, u)$ a.e. in $(0, T)$. Such a unique solution will be denoted by

$$x = z(t_0, v, u).$$

(6) Given $v_n \to v$ in R^m, $u_n \to u$ in L^2 then $z(0, v_n, u_n) \to z(0, v, u)$ uniformly in $[0, T]$.

Explicit assumptions about g such that (6) holds will be briefly considered in the last section.

(7) There exists a positive constant a such that for every vector c of the appropriate dimension and a.e. $t \in (0, T)$

$$c'Q(t)c \geqq a|c|^2, \qquad c'P(t)c \geqq 0, \qquad c'Ec \geqq 0.$$

(8) There exists a closed set $L \subset [0, T]$ of Lebesgue measure 0 such that for every $u \in R^k$ the function $g(\cdot, \cdot, u)$ is continuous at every $(t, x) \in [0, T] x R^m$ if $t \notin L$. The following stronger form of (7) will be needed.

(9) There exists a positive constant a such that for every vector c of the appropriate dimension and a.e. $t \in (0, T)$

$$c'Q(t)c \geqq a|c|^2, \qquad c'P(t)c \geqq a|c|^2, \qquad c'Ec \geqq a|c|^2.$$

Results. Hadamard well-posedness by definition entails the continuous dependence of the optimal control on the desired trajectory. If global existence, uniqueness and continuous dependence of the state on the control is assumed, then the optimal state and the value are continuous too. We state this (obvious) stronger form of well-posedness formally as

PROPOSITION 1. *If problems* (1), (2), (3) *or* (3*) *are Hadamard well-posed on a given set D and assumptions* (5), (6) *hold, then the optimal state and the value are continuous functions on D of the desired trajectory when continuity is with respect to the*

topology on D induced by the strong $R^m \oplus L^2 \oplus L^2$ topology, the topology of uniform convergence on the space of states, and the usual topology on R.

1. The affine case. The simplest instance of the above optimal control problem is the constrained affine regulator problem. In such a case we have the following theorem.

THEOREM 1. *Let $v \in R^m$ be fixed. Assume* (7) *and let g be an affine function of the form* (4) *with A, $C \in L^1$ and $B \in L^2$. Let K^* be closed and convex. Then for every desired trajectory the problem* (1), (2), (3*) *is both Tykhonov and Hadamard well-posed.*

Proof. Given any desired trajectory, the existence and uniqueness of the corresponding optimal control $\bar{u}$ is well known (and easily proved by standard means). Consider the nonempty projection of the set K^*

$$K_0 = \{u \in L^2 : (u, z(0, v, u)) \in K^*\}.$$

Since v is fixed, we denote simply by $I(u)$, $u \in K_0$, the performance (1), and we write briefly

$$I(u) = \langle u - u^*, Q(u - u^*)\rangle + \langle x - x^*, P(x - x^*)\rangle + [x(T) - y^*]' E[x(T) - y^*], \qquad u \in L^2,$$

where $x = z(0, v, u)$. Now define

$$(Lu)(t) = F(t) \int_0^t F^{-1}(s) B(s) u(s)\, ds,$$

$$y(t) = F(t)\left[v + \int_0^t F^{-1}(s) C(s)\, ds\right],$$

so that

$$x = Lu + y,$$

where F is the fundamental matrix of $\dot{x} = A(t)x$, principal at 0. Therefore L is a bounded linear operator between L^2 spaces.

From [13, Cor. 1, p. 212] Tykhonov well-posedness obtains if for every $u \in K_0$

$$I(u) \geqq I(\bar{u}) + a\|u - \bar{u}\|^2. \tag{10}$$

It is easy to see that, denoting by $\bar{x}$ the optimal state,

$$\begin{aligned} I(u) - I(\bar{u}) = {} & \langle u - \bar{u}, Q(u - \bar{u})\rangle \\ & + \langle x - \bar{x}, P(x - \bar{x})\rangle \\ & + [x(T) - \bar{x}(T)]' E[x(T) - \bar{x}(T)] \\ & + 2\langle \bar{u} - u^*, Q(u - \bar{u})\rangle + 2\langle P(\bar{x} - x^*), x - \bar{x}\rangle \\ & + 2[\bar{x}(T) - y^*]' E[x(T) - \bar{x}(T)], \end{aligned}$$

moreover for every u and w in L^2

$$\langle DI(u), w\rangle = 2\langle u - u^*, Qw\rangle + 2\langle P(x - x^*), Lw\rangle + 2[x(T) - y^*]' E(Lw)(T), \tag{11}$$

but we have the variational inequality

$$\langle DI(\bar{u}), u - \bar{u}\rangle \geqq 0, \qquad u \in K_0,$$

giving (10).

Let us show Hadamard well-posedness. Consider any sequence of desired trajectories

$$(y_n^*, u_n^*, x_n^*) \to (y_0^*, u_o^*, x_0^*) \quad \text{in } R^m \oplus L^2 \oplus L^2.$$

Denote by I_n the corresponding performance index (1) and by $\bar{u}_n$ the corresponding optimal control. Then, from the form of (1), for each fixed $u \in K_0$, $I_n(u) \to I_0(u)$. By (7),

$$a\|u_n^* - \bar{u}_n\|^2 \leqq I_n(\bar{u}_n) \leqq I_n(u) \leqq \text{constant}$$

for every fixed $u \in K_0$. Thus some $w \in L^2$ exists such that for a subsequence

$$\bar{u}_n \rightharpoonup w \text{ in } L^2, \qquad w \in K_0, \quad \text{since } K_0 \text{ is closed and convex.}$$

Since $I_n(\bar{u}_n) \leqq I_n(u)$ for every n and $u \in K_0$, letting $n \to +\infty$ and by weak sequential lower semicontinuity we see that w is the unique optimal control for I_0, and for the original sequence

$$\bar{u}_n \rightharpoonup \bar{u}_0.$$

Remembering (11) it is easily checked that

$$\langle DI_n(\bar{u}_0), \bar{u}_0 - \bar{u}_n \rangle \geqq a\|\bar{u}_0 - \bar{u}_n\|^2.$$

Since

$$DI_n(\bar{u}_0) \to DI_0(\bar{u}_0)$$

we get

$$\bar{u}_n \to \bar{u}_0. \qquad \square$$

Remark. With the terminology of [13], Hadamard well-posedness has been obtained from equiwell-posedness of (K_0, I_n).

When the initial state v is controlled, too, a well-posed problem is obtained under positive definiteness of E as shown in the following theorem.

THEOREM 2. *Assume g an affine function of the form* (4) *with $A, C \in L^1$ and $B \in L^2$, let K be closed and convex, and assume* (7). *Suppose E is a positive definite matrix. Then for every desired trajectory the corresponding problem* (1), (2), (3) *is both Tykhonov and Hadamard well-posed.*

Proof. Given any desired trajectory, let (v_n, u_n) be a minimizing sequence for the corresponding optimal control problem, with corresponding states x_n. The uniformly positive definiteness of Q gives boundedness of u_n in L^2. Moreover (notations as in the proof of Theorem 1)

$$v_n = F^{-1}(T)x_n(T) - \int_0^T F^{-1}(s)B(s)u_n(s)\,ds - \int_0^T F^{-1}(s)C(s)\,ds. \tag{12}$$

Since E is positive definite, $x_n(T)$ is bounded, therefore (12) implies boundedness of v_n. Then there exist $\bar{u} \in L^2$ and $\bar{v} \in R^m$ such that for a subsequence

$$u_n \rightharpoonup \bar{u} \text{ in } L^2, \qquad v_n \to \bar{v}.$$

By standard estimates, we see that a continuous $\bar{x}$ exists such that for a subsequence $x_n \to \bar{x}$ uniformly in $[0, T]$, moreover $\bar{x} = z(0, \bar{v}, \bar{u})$. A routine argument shows optimality of $(\bar{v}, \bar{u})$. By positive definiteness of E, Q and nonsingularity of $F(T)$ it is easy to check the strict convexity of I. Therefore existence and uniqueness of the optimal control $(\bar{v}, \bar{u})$ with state $\bar{x}$ is guaranteed. Moreover for any minimizing sequence (v_n, u_n) the above conclusions show for the original sequences $u_n \rightharpoonup u$ in L^2, $v_n \to \bar{v}$, $x_n \to \bar{x}$ uniformly in $[0, T]$, $I(v_n, u_n) \to I(\bar{v}, \bar{u})$.

Therefore

$$\int_0^T (x_n - x^*)'P(x_n - x^*)\,dt \to \int_0^T (\bar{x} - x^*)'P(\bar{x} - x^*)\,dt;$$

$$[x_n(T) - y^*]'E[x_n(T) - y^*] \to [\bar{x}(T) - y^*]'E[\bar{x}(T) - y^*].$$

This implies

$$\int_0^T (u_n - u^*)'Q(u_n - u^*)\,dt \to \int_0^T (\bar{u} - u^*)'Q(\bar{u} - u^*)\,dt,$$

and finally by (7), $u_n \to \bar{u}$ in L^2, thus showing Tykhonov well-posedness. The Hadamard well-posedness is proved in a similar way. □

Remarks. (1) An indirect proof of Theorem 2 (based on results of [13]) can be given analogous to that of Theorem 1. (2) By remembering Proposition 1, we see that under the assumptions of Theorem 1 or 2, continuity of the optimal state and of the value obtains with respect to the desired trajectory, so giving a stronger form of Hadamard well-posedness. (3) The assumption of positive definiteness of E cannot be dispensed with in Theorem 2, as we see by taking $A = P = E = O$, $B = Q = 1$.

2. Dense well-posedness. In this section we consider the optimal regulator problem (1), (2), (3) or (3*) without the assumption that g is affine.

THEOREM 3. *Fix $v \in R^m$ and assume* (5), (6), (7). *Let K^* be a closed set. Then for every $y^* \in R^m$ and $x^* \in L^2$ there exists a dense subset $G \subset L^2$ such that for every $u^* \in G$ the corresponding problem* (1), (2), (3*) *is Tykhonov well-posed. Moreover $u_n^*, u^* \in G$ and $u_n^* \to u^*$ in L^2 imply that the optimal controls $\bar{u}_n \to \bar{u}$ in L^2, the optimal states $\bar{x}_n \to \bar{x}$ uniformly and the values $V_n \to V$.*

Proof. Given $u \in L^2$, write briefly

$$x(u) = z(0, v, u).$$

For any $y^* \in R^m$, $x^* \in L^2$ consider

$$f(u) = \int_0^T (x(u) - x^*)'P(x(u) - x^*)\,dt + (x(u)(T) - y^*)'E(x(u)(T) - y^*).$$

By (6) and (7) we see that f is bounded from below by zero and is lower semicontinuous with respect to the strong convergence in L^2 (since f is continuous). Define

$$K_0 = \{u \in L^2;\ (u, x(u)) \in K^*\}.$$

Then K_0 is closed. The inner product

$$\langle u_1, u_2 \rangle = \int_0^T u_1' Q u_2\,dt \tag{13}$$

induces on L^2 a Hilbert space structure equivalent to the usual one, with corresponding norm $\|\cdot\|$. Then

$$I(v, u) = f(u) + \|u - u^*\|^2, \qquad u \in K_0,$$

and the conclusions follow from [4, Prop. 4, the remarks there and Thm. 3]. □

With stronger assumptions about P and E we get dense Hadamard well-posedness as shown in the following theorem.

THEOREM 4. *Assume that K is a closed set with bounded projection on R^m. Let* (5), (6) *and* (9) *hold. Then there exists a dense subset D of $R^m \oplus L^2 \oplus L^2$ such that the problem* (1), (2), (3) *is both Tykhonov and Hadamard well-posed in D.*

Proof. Given $z_1, z_2 \in R^m \oplus L^2 \oplus L^2$, $z_i = (y_i, u_i, x_i)$ we define

$$\langle z_1, z_2 \rangle = \int_0^T (u_1' Q u_2 + x_1' P x_2)\, dt + y_1' E y_2. \tag{14}$$

The Hilbert space structure induced by the inner product (14) on $R^m \oplus L^2 \oplus L^2$ is equivalent to the usual one by the uniform positive definiteness of Q, P and E. Consider now the corresponding Hilbert space norm $\|\cdot\|$ and define

$$W = \{(y, u, x) \in R^m \oplus L^2 \oplus L^2 : u \in L^2, x = z(0, v, u) \text{ for some } v \in R^m, y = x(T), (v, u, x) \in K\}.$$

If $(y_n, u_n, x_n) \in W$ with $y_n = x_n(T) \to y$, $u_n \to u$ and $x_n = z(0, v_n, u_n) \to x$, then v_n is bounded; therefore $v_n \to v$ for some $v \in R^m$ and some subsequence. By (6) $x = z(0, v, u)$ and $y = x(T)$, thus showing closedness of W. Furthermore for any desired trajectory z^*, the pair $(\bar{v}, \bar{u})$ is an optimal control for the corresponding problem (1), (2), (3) if and only if there exists some $\bar{y} \in R^m$ such that (writing $\bar{x} = z(0, \bar{v}, \bar{u})$), $\bar{z} = (\bar{y}, \bar{u}, \bar{x}) \in W$ and $\|\bar{z} - z^*\| \leqq \|z - z^*\|$ for every $z \in W$. Then the conclusion follows by [4, Proposition 4 and Theorem 3]. □

Remarks. 1) No positive definiteness assumption about E is needed to get the conclusions in Theorem 2 or 4 if the constraints force the final state $x(T)$ to be fixed.

2) By Proposition 1, the optimal state and value are continuous functions in D of the desired trajectory under the assumptions of Theorem 4.

3) If $v_n \to v$ in R^m and $u_n \to u$ in L^2 imply $z(t_0, v_n, u_n) \to z(t_0, v, u)$ uniformly in $[0, T]$ for every $t_0 \in \{0, T\}$, then the assumption of bounded projection of K on R^m can be omitted in Theorem 4 by the following modification in the proof. Given $(y_n, u_n, x_n) \in W$, with $y_n \to y$, $u_n \to u$ and $x_n \to x$, then by (5)

$$x_n = z(0, v_n, u_n) = z(T, y_n, u_n)$$

since $y_n = x_n(T)$. Then by continuous dependence at T we get $x = z(T, y, u)$. Therefore $x = z(0, z(T, y, u)(0), u)$, $y = x(T)$ and

$$v_n = x_n(0) \to v = x(0)$$

thus showing closedness of W.

3. A variational characterization of affine control systems. In this section we study well-posedness of problem (1), (2) for every desired trajectory. We consider only problems without constraints, that is $K = R^m \oplus L^2 \oplus L^2$. We shall work within the Hilbert space

$$H = R^m \oplus L^2 \oplus L^2. \tag{15}$$

The following lemma is easily proved.

LEMMA 1. *Assume* (5), (6), (9). *Then the inner product* (14) *defines on* H *a Hilbert space structure equivalent to the natural one. For every* $z^* \in H$, $(\bar{v}, \bar{u})$ *is an optimal control for the corresponding problem* (1), (2) *if and only if* $(\bar{x}(T), \bar{u}, \bar{x})$, *where* $\bar{x} = z(0, \bar{v}, \bar{u})$, *minimizes the distance of* z^* *from*

$$G = \{(x(T), u, x) \in H : u \in L^2, x = z(0, v, u) \text{ for some } v \in R^m\} \tag{16}$$

in the sense of the norm induced by (14) *on* H.

LEMMA 2. *Assume* (5), (6), (9). *Then* G *defined by* (16) *is convex if for every desired trajectory the problem* (1), (2) *is Tykhonov or Hadamard well-posed.*

Proof. Let us denote by H the Hilbert space (15) equipped by the inner product (14). Suppose that Hadamard well-posedness obtains. Then G is a Chebyshev set (see [14, p. 235] for the definition) by Lemma 1. Given a convergent sequence $z_n^* \to z_0^*$ in H, denote by $(\bar{v}_n, \bar{u}_n)$ the corresponding optimal control. Then by Hadamard well-posedness $\bar{v}_n \to \bar{v}_0$ in R^m and $\bar{u}_n \to \bar{u}_0$ in L^2, giving $z(0, \bar{v}, \bar{u}_n) \to z(0, \bar{v}_0, \bar{u}_0)$ in L^2 and $z(0, \bar{v}_n, \bar{u}_n)(T) \to z(0, \bar{v}_0, \bar{u}_0)(T)$ by (6). Then the convexity of G follows from [14, corollary, p. 237]. Assume now Tykhonov well-posedness. Then G is a Chebyshev set.

Given $z^* \in H$ with corresponding optimal control $(\bar{v}, \bar{u})$ let $(y_n, u_n, x_n) \in G$ such that

$$\|(y_n, u_n, x_n) - z^*\| \to \inf \{\|z^* - z\| : z \in G\}. \tag{17}$$

Then $x_n = z(0, v_n, u_n)$ for some $v_n \in R^m$ and $y_n = x_n(T)$. The Tykhonov well-posedness yields

$$v_n \to \bar{v} \text{ in } R^m, \qquad u_n \to \bar{u} \text{ in } L^2$$

and $x_n \to z(0, \bar{v}, \bar{u})$ uniformly in $[0, T]$, so that $y_n \to z(0, \bar{v}, \bar{u})(T)$. The convexity of G is now given by [14, Cor. 3, p. 238]. □

The following lemma is quite elementary and its proof is therefore omitted.

LEMMA 3. *Let X and Y be real vector spaces, and $f: X \to Y$ a mapping such that its graph is convex. Then f is affine.*

LEMMA 4. *Suppose that assumption* (5) *holds. Then the following are equivalent properties*:

$z(0, \cdot, \cdot\rangle$ is affine (between $R^m \oplus L^2$ and L^2);
there exists some $t_0 \in [0, T]$ such that $z(t_0, \cdot, \cdot)$ is affine;
$z(t_0, \cdot, \cdot)$ is affine for every $t_0 \in [0, T]$.

Proof. Let $z(0, \cdot, \cdot)$ be affine. Given any $t_0 \in [0, T]$, v_1 and $v_2 \in R^m$, u_1 and $u_2 \in L^2$, $b \in R$, consider

$$z_1 = z(t_0, v_1, u_1), \qquad z_2 = z(t_0, v_2, u_2),$$

$$z_3 = z(t_0, bv_1 + (1-b)v_2, bu_1 + (1-b)u_2), \qquad \bar{z}_i = z_i(0).$$

Let us show that

$$z_3 = bz_1 + (1-b)z_2. \tag{18}$$

By uniqueness in the large (assumption (5))

$$\begin{aligned} bz_1 + (1-b)z_2 &= bz(0, \bar{z}_1, u_1) + (1-b)z(0, \bar{z}_2, u_2) \\ &= z(0, b\bar{z}_1 + (1-b)\bar{z}_2, bu_1 + (1-b)u_2), \end{aligned}$$

moreover

$$z_3(t_0) = bv_1 + (1-b)v_2 = bz_1(t_0) + (1-b)z_2(t_0);$$

consequently (18) follows.

The proof is completed by exchanging the roles of 0 and t_0. □

LEMMA 5. *Assume* (5), (6) *and* (9). *Then Tykhonov or Hadamard well-posedness of problem* (1), (2) *for every desired trajectory implies that for every $t_0 \in [0, T]$ the mapping*

$$z(t_0, \cdot, \cdot): R^m \oplus L^2 \to L^2$$

is affine.

Proof. By (5) the graph of $z(T, \cdot, \cdot)$ is the same as the set G given by (16). But G is convex by Lemma 2. By Lemmas 3 and 4, $z(t_0, \cdot, \cdot)$ is affine for every $t_0 \in [0, T]$. □

THEOREM 5. *Assume* (5), (6), (8) *and* (9). *If problem* (1), (2) *is Tykhonov or Hadamard well-posed for every desired trajectory, then there exist matrix-valued functions* A, B, C, *continuous in* $[0, T]\backslash L$, *such that for every* $t \notin L$, $x \in R^m$ *and* $u \in R^k$

$$g(t, x, u) = A(t)x + B(t)u + C(t). \tag{4}$$

Proof. Fix any $t_0 \in [0, T]\backslash L$. Given $v, w \in R^m$, $p, q \in R^k$, $b \in R$ let us consider $x_1 = z(t_0, v, p)$, $x_2 = z(t_0, w, q)$, $x_3 = z(t_0, bv + (1-b)w, bp + (1-b)q)$. By Lemma 5, for every $t \in [0, T]$ we get

$$b \int_{t_0}^{t} g(s, x_1(s), p)\, ds + (1-b) \int_{t_0}^{t} g(s, x_2(s), q)\, ds = \int_{t_0}^{t} g(s, x_3(s), bp + (1-b)q)\, ds. \tag{19}$$

Dividing (19) by $t - t_0$ and then letting $t \to t_0$, by (8) we obtain

$$bg(t_0, v, p) + (1-b)g(t_0, w, q) = g(t_0, bv + (1-b)w, bp + (1-b)q).$$

This shows that $g(t, \cdot, \cdot)$ is affine for every $t \notin L$, thus proving (4). Continuity of A, B, C in $[0, T]\backslash L$ is implied by (4) and (8). □

Summarizing, we state the main result of this paper as follows from Theorems 2 and 5.

THEOREM 6. *Suppose that assumptions* (5), (6), (8) *and* (9) *hold. Moreover let* $g(\cdot, x, 0) \in L^1$ *for every* $x \in R^m$, *and* $g(\cdot, 0, u) - g(\cdot, 0, 0) \in L^2$ *for every* $u \in R^k$. *Then a necessary and sufficient condition such that* $g(t, \cdot, \cdot)$ *is affine for almost every* $t \in [0, T]$ *is that the optimal control problem* (1), (2) *is Tykhonov or Hadamard well-posed for every desired trajectory.*

Remark. If v is fixed in (2), so that the control is effected through u alone, then well-posedness does not imply that g is affine, as we see by taking $g(t, x, u) = \sin x$.

Let us compare Theorems 4 and 6 under the assumption of continuous dependence at every $t_0 \in [0, T]$ of the state on the control (see Remark 3 of § 2). Roughly speaking, for many nonlinear control systems (2), given any desired trajectory we can modify it by an arbitrary small amount (in the L^2 sense) to obtain a well-posed problem. But suppose that no changes of the desired trajectories are allowed. Then if well-posedness is required on the whole space, the nonaffine control systems (2) are necessarily ruled out.

Let us now consider a simple semilinear control system. Here the affine character of the dynamics can be shown without assumption (8) as a byproduct of well-posedness. We take no control of the initial state.

THEOREM 7. *Let* $v \in R^m$ *be fixed,* $E = 0$,

$$g(t, x, u) = A(t)x + B(t, u),$$

where $A \in L^1$ *and* B *is a Carathéodory function on* $[0, T] \times R^k$. *Assume* (5) *and* P, Q *uniformly positive definite as in* (9). *Then Tykhonov or Hadamard well-posedness for every desired trajectory of the problem* (1), (2) *implies that for almost every* t *and all* $u \in R^k$

$$B(t, u) = b(t)u + c(t)$$

for some $b \in L^2$ *and* $c \in L^1$.

Proof. Assumption (5) and the summability of A imply that for every control $u \in L^2$ we have

$$t \to B(t, u(t)) \in L^1.$$

Then [15, Thm. 191] there exist $p \in L^1$, $q \geqq 0$ such that for a.e. t and every u

$$|B(t, u)| \leqq p(t) + q|u|^2. \tag{20}$$

Write $x(u) = z(0, v, u)$ and consider the set $G = \{(u, x(u)): u \in L^2\}$. By [15, Thm. 19.1] we see that (6) holds (with v fixed).

Using [14, Cor. 3 and corollary, p. 237] we get the convexity of G.

The proof of this is similar to the corresponding proof of Theorem 5. By Lemma 3 this implies affinity of $u \to x(u)$. Denote by F the fundamental matrix of $\dot{x} = A(t)x$, principal at 0. Then the mapping

$$u \to \int_0 F^{-1}(s)B(s, u(s))\, ds$$

that assigns to every $u \in L^2$ the function $t \to \int_0^t F^{-1}(s)B(s, u(s))\, ds$, $0 \leqq t \leqq T$, turns out to be affine. Therefore for every $u, v \in R^k$, $q \in R$ and $t \in [0, T]$

$$\int_0^t F^{-1}(s)B(s, qu + (1-q)v)\, ds = q \int_0^t F^{-1}(s)B(s, u)\, ds + (1-q) \int_0^t F^{-1}(s)B(s, v)\, ds.$$

By taking derivatives we see that given q, u, v there exists a set of full measure in $(0, T)$ such that for every t in this set

$$B(t, qu + (1-q)v) = qB(t, u) + (1-q)B(t, v).$$

By considering a countable dense set in $R \times R^k$ we get that $B(t, \cdot)$ is affine for a.e. $t \in (0, T)$. The properties of b, c the follow from (20). □

Remark. If E is positive definite then an analogous statement holds in the semilinear case considered in Theorem 7 for problem (1), (2).

4. Continuous dependence of the state on the control. In this section we introduce some explicit conditions about the function g such that assumption (6) is satisfied.

PROPOSITION 2. *Assume* (5) *and the following conditions*:

(21) *g is a Carathéodory function, that is, $g(t, \cdot, \cdot)$ is continuous for every t and $g(\cdot, x, u)$ is Lebesgue measurable for every x and u;*

(22) *for every $A > 0$ and nonnegative $p \in L^2$ there exists $q \in L^1$ such that if $|x| \leqq A$ and $|u| \leqq p(s)$ then*

$$|g(s, x, u)| \leqq q(s) \quad \text{a.e. in } (0, T);$$

(23) *for every nonnegative $p \in L^2$ there exists $B > 0$ and $r \in L^1$ such that if $|x| \geqq B$ and $|u| \leqq p(s)$ then*

$$x'g(s, x, u) \leqq r(s)(1 + |x|^2).$$

Then property (6) *is true.*

Proof. Given $v_n \to v$ in R^m, $u_n \to u$ in L^2, a nonnegative $p \in L^2$ exists such that, for some subsequence,

$$|u_n(s)| \leqq p(s), \qquad u_n(s) \to u(s), \quad \text{a.e. in } (0, T). \tag{24}$$

(See [16, Lemma 3.9].) Corresponding to p we consider the constant B given by (23). Write $x_n = z(0, v_n, u_n)$ and assume that some n and $s^* \in (0, T)$ exist such that $|x_n(s^*)| > B$. Consider then

$$I_n = \{s \in [0, T]: |x_n(s)| > B\}$$

and for any $t \in [0, T]$

$$E_n = [0, t] \cap I_n; \qquad F_n = [0, t] \setminus I_n.$$

Then for every $t \in [0, T]$, using (24), (23) and (22),

$$\frac{1}{2}(|x_n(t)|^2 - |v_n|^2) = \left(\int_{E_n} + \int_{F_n}\right) x_n'(s) g(s, x_n(s), u_n(s))\, ds$$

$$\leqq \int_0^t r(s)|x_n(s)|^2\, ds + \int_0^T [r(s) + Bq(s)]\, ds.$$

By Gronwall's lemma we get equiboundedness of x_n. This implies equicontinuity of x_n by (22).

Thus the Ascoli-Arzelà theorem yields the existence of a further subsequence of x_n and a continuous function y such that $x_n \to y$ uniformly in $[0, T]$. By (21) and (24)

$$g(s, x_n(s), u_n(s)) \to g(s, y(s), u(s)) \quad \text{a.e.}$$

Since this convergence is dominated by (22), we get $x_n \to y = z(0, v, u)$ uniformly (for the original sequence). □

Remarks. (1) Assumption (22) is satisfied if the Nemytskij operator defined by the Caratheodory function g,

$$(u, x) \to g(\cdot, x(\cdot), u(\cdot))$$

is strongly continuous from $L^2 \oplus L^2$ into L^1 ([15, Theorem 19.1]). By using the continuous dependence results of [17, Thm. 3.1], we see that, assuming (5), (6) holds whenever the following is true. For every nonnegative $p \in L^2$ and $A > 0$ there exists $q \in L^1$ such that for a.e. $s \in (0, T)$, if $|u| \leqq p(s)$, $|x| \leqq A$, $|y| \leqq A$ then $|g(s, x, u) - g(s, y, u)| \leqq q(s)|x - y|$; moreover

$$g(\cdot, x, u(\cdot)) \in L^1$$

for every $x \in R^m$ and $u \in L^2$. This last condition in fact implies continuity between L^2 and L^1 of every Nemytskij operator $u \to g(\cdot, x, u(\cdot))$ [15, Thm. 19.1]. (3) Given a Tykhonov unstable optimization problem, by suitable modifications (regularizations) of the cost functional we obtain in some cases a well-posed problem (see [18, especially Chapter 7]).

Acknowledgment. We wish to thank the referees for many suggestions in writing style that improved the presentation of this paper.

REFERENCES

[1] A. N. TYKHONOV, *On the stability of the functional optimization problem*, USSR Computational Math. and Math. Phys., 64 (1966), pp. 28–34.

[2] R. B. HOLMES, *A Course on Optimization and Best Approximation*, Lecture Notes in Mathematics, 257, Springer-Verlag, New York, 1972.

[3] T. ZOLEZZI, *Approximations and Perturbations of Minimum Problems*, book in preparation.

[4] G. LÉ BOURG, *Perturbed optimization problems in Banach spaces*, Bull. Soc. Math. France, Mémoire 60 (1979), pp. 95–111.

[5] M. F. BIDAUT, *Théorèmes d'existence et d'existence "en général" d'un contrôle optimal pour des systèmes règis par des équations aux dérivées partielles non linéaires*, Thèse, Université de Paris VI, 1973.

[6] J. BARANGER, *Existence de solutions pour des problèmes d'optimisation non convexes*, J. Math. Pures Appl., 52 (1973), pp. 377–406.

[7] J. BARANGER AND R. TEMAM, *Non convex optimization problems depending on a parameter*, this Journal, 13 (1975), pp. 146–152.
[8] I. EKELAND AND G. LEBOURG, *Generic Fréchet differentiability and perturbed optimization problems in Banach spaces*, Trans. Amer. Math. Soc., 224 (1976), pp. 193–216.
[9] M. F. BIDAUT, *Existence theorems for usual and approximate solutions of optimal control problems*, J. Optim. Theory Appl., 15 (1975), pp. 393–411.
[10] G. LEBOURG, *Solutions "en densité" de problèmes d'optimisation paramétrés*, C.R. Acad. Sci. Paris, 289 (1979), pp. 79–82.
[11] R. LUCCHETTI AND F. PATRONE, *Sulla densità e genericità di alcuni problemi di minimo ben posti*, Boll. Un. Mat. Ital., 15B (1978), pp. 225–240.
[12] T. ZOLEZZI, *Characterizations of some variational perturbations of the abstract linear quadratic problem*, this Journal, 16 (1978), pp. 106–121.
[13] ———, *On equiwell-set minimum problems*. Appl. Math. Optim., 4 (1978), pp. 209–223.
[14] E. ASPLUND, *Čebišev sets in Hilbert space*, Trans. Amer. Math. Soc., 144 (1969), pp. 235–240.
[15] M. M. VAJNBERG, *Variational methods for the study of nonlinear operators*, Holden-Day, San Francisco, 1964.
[16] P. R. HALMOS AND V. S. SUNDER, *Bounded integral operators on L^2 spaces*, Ergebnisse der Mathematik und ihrer Grenzgebiete, 96, Springer-Verlag, New York, 1978.
[17] Z. ARTSTEIN, *Topological dynamics of an ordinary differential equation*, J. Differential Equations, 23 (1977), pp. 216–223.
[18] A. TIKHONOV AND V. ARSENINE, *Méthodes de résolutions de problèmes mal posés*, Editions MIR, Moscow, 1976.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 5, September 1981

0363-0129/81/1905-0005 $01.00/0

EXISTENCE OF VALUE AND RANDOMIZED STRATEGIES IN ZERO-SUM DISCRETE-TIME STOCHASTIC DYNAMIC GAMES*

P. R. KUMAR† AND T. H. SHIAU†

Abstract. Two players with conflicting objectives are simultaneously controlling a discrete-time stochastic system. The goal of this paper is to analyze such zero-sum, discrete-time, stochastic systems when the two players are allowed to use randomized strategies.

Previous results have been restricted to systems with finite or compact state spaces. Such restrictions are usually untenable from the point of view of applications, since many applications frequently use either the integers or $\mathbb{R}^n$ as a state space. Our results are proved for complete, separable, metric spaces which are very useful for applications.

All previously known results emerge as special cases of our results. In addition, a variety of conjectures and open problems are resolved regarding the existence of a value function, its properties such as Borel measurability or continuity, and the existence for either or both players of optimal or ε-optimal stationary strategies.

1. Introduction. The goal of this paper is to analyze zero-sum discrete-time stochastic games where the two players are allowed to use randomized strategies (i.e., two players with conflicting objectives simultaneously controlling a stochastic system).

Starting with Shapley [1], many researchers, e.g., Everett [2], Maitra and Parthasarathy [3], [4] have treated such systems. However, all these treatments suffer from the fact that they restrict the state space to be either finite or compact. Many useful models of dynamic games however use a state space which is often the integers or $\mathbb{R}^n$—both of which fail to satisfy these restrictions. Even in some special cases which do satisfy these restrictions, earlier results are not applicable if the value function is not continuous. (We give such an example in § 7.) Also nonstationary (i.e., time-varying) systems, when made stationary by adjoining an additional state variable to count time, possess noncompact state spaces. The restriction of the state space to be either finite or compact is therefore untenable from the point of view of applications.

We provide in this paper a general theory of zero-sum discrete-time stochastic games which overcomes these restrictions. We consider two alternative models for such problems—a Borel model and a continuous model, both of which have complete, separable, metric state spaces and therefore include both integer state spaces and Euclidean state spaces. These two models are very useful in applications.

Our results include previously known results and also solve a number of open problems. The models and our results are stated in the next section.

2. Problem statement. We consider a system evolving in a state space X according to:

$$x_{k+1} = f(x_k, u_k, v_k, w_k). \tag{1}$$

Player I (the maximizer) controlling u wishes to maximize the expected cost

$$E\left[\sum_{k=0}^{\infty} \alpha^k c(x_k, u_k, v_k, x_{k+1})\right]. \tag{2}$$

At each time k, player I is allowed to choose u_k from some set $U_{x_k} \subset U$. Player II (the minimizer) controlling v wishes to minimize (2) by choosing, at each instant k, v_k from $V_{x_k} \subset V$. $\{w_k\}$ is a random disturbance.

* Received by the editors December 26, 1979, and in revised form September 6, 1980.

† Department of Mathematics, University of Maryland Baltimore County, 5401 Wilkens Ave., Baltimore, Maryland 21228. This research was supported by the U.S. Army Research Office.

We assume that X, U, and V are complete, separable, metric spaces. $\alpha \in (0, 1]$ is called the discount factor. By $P(A)$ we shall denote the set of all Borel probability measures on the complete, separable, metric space A. Instead of working with the state equation (1), we prefer to deal with the transition kernel, $q(B|x, u, v) = \text{Probability}(\{w: f(x, u, v, w) \in B\}|x, u, v)$, which we assume to be well defined for every Borel set $B \subset X$.

Since even elementary games need not possess a value, we need to impose additional conditions to make the problem meaningful. We impose two alternative sets of conditions—the Borel model and the continuous model. These two models are very useful from the point of view of applications.

Borel model.

(i) q is a Borel measurable stochastic kernel, i.e., $q(B|\cdot)$ is Borel measurable in the second argument for every fixed Borel subset $B \subset X$.

(ii) U_x and V_x are *finite* for each $x \in X$, and

$$\begin{aligned} \Gamma_1 &= \{(x, u): x \in X \text{ and } u \in U_x\} \subset X \times U, \\ \Gamma_2 &= \{(x, v): x \in X \text{ and } v \in V_x\} \subset X \times V, \\ Q_1 &= \{(x, T): x \in X \text{ and } T \in P(U_x)\} \subset X \times P(U), \\ Q_2 &= \{(x, R): x \in X \text{ and } R \in P(V_x)\} \subset X \times P(V) \end{aligned} \tag{3}$$

are all Borel subsets of the corresponding spaces.

Continuous model.

(4)
(i) $U_x \equiv U$, $V_x \equiv V$ for all $x \in X$ where U and V are compact. (Note. This will be generalized in § 3 to allow U_x and V_x to depend on x as in the Borel model, but to preserve clarity of exposition, we prefer to state the simpler version first.)

(ii) c is continuous on $X \times U \times V \times X$.

(iii) $q: X \times U \times V \to P(X)$ is weakly continuous, i.e., continuous with respect to the weak topology on $P(X)$.

In both models we assume

$$0 \leqq c(x, u, v, y) \leqq \theta < \infty$$

(even this will be relaxed in § 3, Remark 1).

Players I and II are allowed to use randomized strategies. We define a randomized strategy for player I to be a sequence $F = \{F^0, F^1, F^2, F^3, \cdots\}$ where each $F^k = F^k(du_k|x_0, u_0, x_1, u_1, x_2, u_2, \cdots, x_k)$ is a Borel measurable stochastic kernel on U_{x_k}. Player I chooses u_k according to this probability distribution F^k which utilizes the past history $(x_0, u_0, x_1, u_1, x_2, u_2, \cdots, x_k)$ known to him in a Borel measurable way. A randomized strategy $F = \{F^0, F^1, F^2, \cdots\}$ is said to be Markovian if each F^k depends only on x_k, i.e., $F^k(du_k|x_k)$. A Markovian strategy is said to be stationary if all the F^k's are identical. The different types of randomized strategies for player II are defined similarly. For player i, the sets of randomized, Markovian and stationary strategies are denoted by D_i, M_i and S_i respectively.

The cases $0 < \alpha < 1$ and $\alpha = 1$ will be called the discounted and positive cases respectively and will be identified by the letters D and P. Similarly, B and C will denote the Borel model and the continuous model. A combination of letters such as BD will refer to the discounted cost case of the Borel model. When no letters are used, a result holds for all models and cases.

For every initial state x_0 and randomized strategy pair $(F, G) \in D_1 \times D_2$ adopted by the two players, the cost incurred is

$$J(x_0; F, G) := E_{F,G}\Big[\sum_{k=0}^{\infty} \alpha^k c(x_k, u_k, v_k, x_{k+1}) \,|\, x_0\Big],$$

where $E_{F,G}$ denotes the expectation under the probability measure induced on the future evolution of the system by (F, G) and the random disturbance. It is allowed to be $+\infty$, as are all functions throughout this paper.

Our main results are the following:

(i) For every $x_0 \in X$,

$$\inf_{G \in D_2} \sup_{F \in D_1} J(x_0; F, G) = \sup_{F \in D_1} \inf_{G \in D_2} J(x_0; F, G) =: J^*(x_0). \tag{5}$$

$J^*(\cdot)$ is called the value function (Theorem 2).

(ii) (6)

(B) $J^*(\cdot)$ is a Borel measurable function.

(CD) $J^*(\cdot)$ is a bounded continuous function.

(CP) $J^*(\cdot)$ is a lower-semicontinuous function (Lemma 3).

(iii) (D) $J^*(\cdot)$ is the *unique* solution of

$$\begin{aligned} J^*(x) &= \min_{R \in P(V_x)} \max_{T \in P(U_x)} \int_{V_x} \int_{U_x} \int_X [\alpha J^*(y) + c(x, u, v, y)] q(dy|x, u, v) T(du) R(dv) \\ &= \max_{T \in P(U_x)} \min_{R \in P(V_x)} \int_{U_x} \int_{V_x} \int_X [\alpha J^*(y) + c(x, y, v, y)] q(dy|x, u, v) R(dv) T(du) \end{aligned} \tag{7}$$

for every $x \in X$ (Theorem 1).

(P) $J^*(\cdot)$ satisfies

$$J^*(x) = \min_{R \in P(V_x)} \sup_{T \in P(U_x)} \int_{V_x} \int_{U_x} \int_X [J^*(y) + c(x, u, v, y)] q(dy|x, u, v) T(du) R(dv) \tag{8}$$

for every $x \in X$. Furthermore, if any nonnegative function $J(\cdot)$ satisfies

$$J(x) \geqq \inf_{R \in P(V_x)} \sup_{T \in P(U_x)} \int_{V_x} \int_{U_x} \int_X [J(y) + c(x, u, v, y)] q(dy|x, u, v) T(du) R(dv) \tag{9}$$

for every $x \in X$, then $J(x) \geqq J^*(x)$ for every $x \in X$ (Theorem 1).

(iv) If $G^* = \{G_0, G_0, G_0, \cdots\} \in S_2$ is such that $G_0(dv|x)$ attains the outer minimum in (7) or (8) then $J(x_0, F, G^*) \leqq J^*(x_0)$ for all $F \in D_1$ and all $x_0 \in X$. There always exists such a G^*. In words, player II has a stationary strategy G^* which is optimal irrespective of the initial state (Theorem 2).

(v) (D) If $F^* = \{F_0, F_0, F_0, \cdots\} \in S_1$ is such that $F_0(du|x)$ attains the outer maximum in (7), then $J(x_0; F^*, G) \geqq J^*(x_0)$ for all $G \in D_2$ and all $x_0 \in X$. There always exists such an F^*. In words, player I has a stationary strategy F^* which is optimal irrespective of the initial state (Theorem 3).

(P) (a) For any probability measure λ on X and every $\varepsilon > 0$, there exists $F_{\lambda,\varepsilon} \in S_1$ such that

$$\lambda\left(\left\{x \in X: \inf_{G \in D_2} J(x; F_{\lambda,\varepsilon}, G) \geqq \begin{cases} J^*(x) - \varepsilon & \text{if } J^*(x) < \infty, \\ \dfrac{1}{\varepsilon} & \text{if } J^*(x) = \infty \end{cases}\right\}\right) \geqq 1 - \varepsilon.$$

(b) In particular, for every finite $S \subset X$, there exists an $F_{S,\varepsilon} \in S_1$ such that

$$\inf_{G \in D_2} J(x; F_{S,\varepsilon}, G) \geqq \begin{cases} J^*(x) - \varepsilon & \text{if } J^*(x) < \infty \text{ and } x \in S, \\ \dfrac{1}{\varepsilon} & \text{if } J^*(x) = \infty \text{ and } x \in S. \end{cases}$$

(c) If X is compact and $J^*(\cdot)$ is continuous, then for every $\varepsilon > 0$, player I has an ε-optimal stationary strategy; i.e., there exists $F_\varepsilon \in S_1$ such that (Theorem 3)

$$\inf_{G \in D_2} J(x; F_\varepsilon, G) \geqq J^*(x) - \varepsilon \quad \text{for all } x \in X.$$

Some comments about our results are in order.

(i) For the case where X is not necessarily compact (or finite) all the above mentioned results are new. In particular we call attention to our result (6) that $J^*(\cdot)$ is Borel measurable. This result is striking since it is even stronger than the previous *conjectures*. In [5, Open Problem 2, p. 253] it is conjectured that $J^*(\cdot)$ would be universally measurable. Our result goes beyond this conjecture and proves that $J^*(\cdot)$ is Borel measurable.

(ii) In the continuous model with undiscounted cost (CP), even when X is compact, our results are considerably stronger than earlier results [4]. In [4] an additional assumption regarding the equicontinuity of the family of value functions of the corresponding discounted games is made. A statement is also made [4, Remark 3.2] that the authors are unaware if the results hold when such an assumption is not made. Besides being a restrictive assumption which is not a priori verifiable, this assumption also results in a continuous value function. As evidenced by an example in § 7, there do exist simple games where $J^*(\cdot)$ is not continuous. We therefore eliminate this assumption. Our result stated in (6) is that $J^*(\cdot)$ is always lower-semicontinuous.

3. The truncated games. We start with a well-known result which we repeat here for convenience.

LEMMA 1. *Let $K: U \times V \to \mathbb{R}$ be continuous with U and V compact. Then*

$$\min_{R \in P(V)} \max_{T \in P(U)} \int_V \int_U K(u, v) T(du) R(dv) = \max_{T \in P(U)} \min_{R \in P(V)} \int_U \int_V K(u, v) R(dv) T(du).$$

Throughout this paper, by measurable we shall mean Borel measurable.

LEMMA 2.

(i) (B) *Let $J^n: X \to [0, M]$ be measurable. Then*

$$\begin{aligned} J^{n+1}(x) &:= \max_{T \in P(U_x)} \min_{v \in V_x} \int_{U_x} \int_X [\alpha J^n(y) + c(x, u, v, y)] q(dy \mid x, u, v) T(du) \\ &:= \min_{R \in P(V_x)} \max_{u \in U_x} \int_{V_x} \int_X [\alpha J^n(y) + c(x, u, v, y)] q(dy \mid x, u, v) R(dv) \end{aligned} \tag{10}$$

is well defined, measurable, nonnegative and bounded by $\alpha M + \theta$.

(C) *Let* $J^n: X \to [0, M]$ *be continuous. Then* $J^{n+1}(\cdot)$ *defined by* (10) *is well defined, continuous, nonnegative and bounded by* $\alpha M + \theta$.

(ii) *There exist measurable stochastic kernels* $T^n(du|x)$ *and* $R^n(dv|x)$ *which achieve the outer maximum and outer minimum respectively in* (10) *for each* $x \in X$.

Proof. Let $K(x, u, v) := \int_X [\alpha J^n(y) + c(x, u, v, y)] q(dy|x, u, v)$.

(B) $K(\cdot)$ is measurable [6, Prop. 7.29] and $K(x, u, v) \leqq \alpha M + \theta$. For fixed x, U_x and V_x are finite; hence, from von Neumann [7], the two expressions in (10) are equal and therefore $J^{n+1}(\cdot)$ is well defined, nonnegative and bounded by $\alpha M + \theta$. To show it is measurable, let $l(x, T, v) = \int_U K(x, u, v) T(du)$. Then since l can be rewritten as $l(x, T, v) = \int_U \bar{K}(x, T, u, v) \varphi(du|x, T, v)$ where $\bar{K}(x, T, u, v) = K(x, u, v)$ and $\varphi(\cdot|x, T, v) \equiv T$ are both measurable, it follows from [6, Prop. 7.29] that l is measurable. To show that the map $(x, T) \mapsto \min_{v \in V_x} l(x, T, v)$ is measurable, let $D = \{(x, T, v): x \in X,\ T \in P(U),\ v \in V_x\}$. Then $D \subset X \times P(U) \times V$ is a Borel set since by assumption (3), Γ_2 is a Borel set. The (x, T)-section of D, $D_{(x,T)} = \{v \in V: (x, T, v) \in D\}$ is just V_x which is finite. From [8, Cor. 1] it follows that the map $m: (x, T) \mapsto \min_{v \in V_x} l(x, T, v)$ is measurable. For fixed x and $v \in V_x$, $l(x, \cdot, v)$ is linear on $P(U_x)$. Since U_x is finite, the set of all probability measures on U_x, i.e., $P(U_x)$ is a standard simplex in a finite dimensional space. Since a linear mapping defined on a subset of a finite dimensional space is continuous, it follows that $l(x, \cdot, v)$ is continuous on $P(U_x)$. $m(x, \cdot): T \mapsto \min_{v \in V_x} l(x, T, v)$ is the minimum of a finite number of such functions and therefore $m(x, \cdot)$ is continuous on $P(U_x)$. A repeated application of [8, Cor. 1] to $-m$ then shows that $J^{n+1}(x) = \max_{T \in P(U_x)} m(x, T)$ is also measurable and that there exists a $T^n(du|x)$ maximizing (10) which is a Borel measurable stochastic kernel. A similar proof holds for $R_n(dv|x)$.

(C) From Lemma 1, the two expressions in (10) are equal and therefore $J^{n+1}(\cdot)$ is well defined, nonnegative and bounded by $\alpha M + \theta$. Define K and l as in the proof of case (B), immediately above. From [6, Prop. 7.30], l is continuous since it can be rewritten as in the proof of case (B) and $\bar{K}$, φ are continuous. By a repeated application, $J^{n+1}(\cdot)$ is also continuous. The existence of Borel-measurable stochastic kernels $T^n(du|x)$ *and* $R^n(dv|x)$ follows as in the proof of case (B).

We now define $J^0(x) \equiv 0$ for all $x \in X$. $J^0(\cdot)$ satisfies the condition of Lemma 2 and by induction it follows that $J^n(\cdot)$ inductively defined by (10) also satisfies the condition of Lemma 2. Moreover, since $J^1(x) \geqq J^0(x) \equiv 0$ for all $x \in X$, it follows by induction and (10) that $J^{n+1}(x) \geqq J^n(x)$ for all $x \in X$ and all n. Define

$$J^*(x) = \lim_{n \to \infty} J^n(x) \quad \text{for every } x \in X. \tag{11}$$

From Lemma 2, it easily follows that

(12)
- (BP) $J^*(\cdot)$ is a Borel measurable function.
- (CP) $J^*(\cdot)$ is a lower semicontinuous function (since it is the increasing limit of continuous functions).

LEMMA 3.

(13)
- (CD) *$J^*(\cdot)$ is a nonnegative continuous function bounded by $\theta/(1-\alpha)$.*
- (BD) *$J^*(\cdot)$ is a nonnegative Borel measurable function bounded by $\theta/(1-\alpha)$.*

Proof.

(CD) Let $K_n(x) := \int_X [\alpha J^n(y) + c(x, u, v, y)] q(dy|x, u, v)$. Since

$$K_n(x, u, v) - K_{n-1}(x, u, v) = \int_X \alpha [J^n(y) - J^{n-1}(y)] q(dy|x, u, v)$$
$$\leqq \alpha \|J^n - J^{n-1}\|_\infty,$$

we obtain

$$\int_{U_x} K_n(x, u, v) T(du) \leqq \int_{U_x} K_{n-1}(x, u, v) T(du) + \alpha \|J^n - J^{n-1}\|_\infty.$$

Hence

$$\max_{T \in P(U_x)} \min_{v \in V_x} \int_{U_x} K_n(x, u, v) T(du)$$
$$\leqq \max_{T \in P(U_x)} \min_{v \in V_x} \int_{U_x} K_{n-1}(x, u, v) T(du) + \alpha \|J^n - J^{n-1}\|_\infty.$$

It follows that $J^{n+1}(x) - J^n(x) \leqq \alpha \|J^n - J^{n-1}\|_\infty$ for every $x \in X$, and hence $\|J^{n+1} - J^n\|_\infty \leqq \alpha \|J^n - J^{n-1}\|_\infty$ for all n. Therefore, $\|J^{n+p} - J^n\|_\infty \leqq (\alpha^{n+p-1} + \cdots + \alpha^n) \|J^1 - J^0\|_\infty = (\alpha^{n+p-1} + \cdots + \alpha^n) \|J^1\|_\infty$. Hence $J^n(\cdot)$ converges uniformly to $J^*(\cdot)$. Let $n = 0$, $p \to +\infty$ we have $\|J^*\|_\infty \leqq (1/(1-\alpha)) \|J^1\|_\infty \leqq \theta/(1-\alpha)$, proving the lemma.

(BD) The proof is similar.

Remark 1. The condition (6) defining the Borel and the continuous models can be replaced by the weaker condition $\|J^1\|_\infty \leqq \theta < \infty$. All the results of this paper will continue to hold under this weaker condition.

We now wish to deal with the generalization of the continuous model (5) to allow U_x and V_x to depend on x. To ensure the equality of the two expressions in (10) we need the continuity of $J^n(\cdot)$. Before stating the most general version, we consider the following one-dimensional situation.

LEMMA 4. (C) *Let*

$$U_x = \{u: a_1(x) \leqq u \leqq b_1(x)\} \subset U \subset \mathbb{R},$$
$$V_x = \{v: a_2(x) \leqq v \leqq b_2(x)\} \subset V \subset \mathbb{R},$$

where a_i and b_i are continuous functions on X satisfying $a_i(x) \leqq b_i(x)$ for all $x \in X$ and $i = 1, 2$. Let X be locally compact. Then the results of Lemma 2 hold.

Proof. Given $x_0 \in X$, by the local compactness of X, there exists an open neighborhood $N(x_0)$ of x_0 with $\overline{N(x_0)}$ compact. Let

$$A_i = \min \{a_i(x): x \in \overline{N(x_0)}\}, \qquad B_i = \max \{b_i(x): x \in \overline{N(x_0)}\}.$$

Define $U(x_0) =: [A_1, B_1]$ and $V(x_0) = [A_2, B_2]$; then $U_x \subset U(x_0)$ and $V_x \subset V(x_0)$ for all $x \in \overline{N(x_0)}$. Define $\varphi_1 : X \times U \times V \to \Gamma_1 \times V$ by

$$\varphi_1(x, u, v) = \begin{cases} (x, u, v), & \text{if } a_1(x) \leqq u \leqq b_1(x), \quad \text{i.e., } (x, u) \in \Gamma_1, \\ (x, a_1(x), v) & \text{if } u < a_1(x), \\ (x, b_1(x), v) & \text{if } u > b_1(x). \end{cases}$$

Also define $\varphi_2\colon \Gamma_1 \times V \to D := \{(x, u, v)\colon u \in U_x, v \in V_x, x \in X\}$ by

$$\varphi_2(x, u, v) = \begin{cases} (x, u, v), & \text{if } a_2(x) \leqq v \leqq b_2(x), \\ (x, u, a_2(x)) & \text{if } v < a_2(x), \\ (x, u, b_2(x)) & \text{if } v > b_2(x). \end{cases}$$

Let $\varphi = \varphi_2 \circ \varphi_1$. Then since φ_1 and φ_2 are continuous, it follows that φ also is continuous. Furthermore, $\varphi(x, u, v) = (x, u'(x, u), v)$ for $v \in V_x$ and $\varphi(x, u, v) = (x, u, v'(x, v))$ for $u \in U_x$. Define $\hat{K}(x, u, v) = K(\varphi(x, u, v))$ where K is defined as in the proof of Lemma 2. For fixed $x \in \overline{N(x_0)}$, let $T^* \in P(U_x)$, $R^* \in P(V_x)$ be the saddle point of the static "game on the unit square" $K(x, u, v)$ on $U_x \times V_x$. For $v \in V_x$ and $u \in U(x_0)$, $\hat{K}(x, u, v) = K(\varphi(x, u, v)) = K(x, u', v)$ where $u' = u'(x, u) \in U_x$. Hence

$$\int_{V(x_0)} \hat{K}(x, u, v) R^*(dv) = \int_{V_x} \hat{K}(x, u, v) R^*(dv) = \int_{V_x} K(x, u', v) R^*(dv)$$

and, therefore,

$$\max_{u \in U(x_0)} \int_{V(x_0)} \hat{K}(x, u, v) R^*(dv) = \max_{u' \in U_x} \int_{V_x} K(x, u', v) R^*(dv) = J^{n+1}(x).$$

Hence

$$\min_{R \in P(V(x_0))} \max_{u \in U(x_0)} \int_{V(x_0)} \hat{K}(x, u, v) R(dv) \leqq J^{n+1}(x).$$

Similarly, we can prove

$$\max_{T \in P(U(x_0))} \min_{v \in V(x_0)} \int_{U(x_0)} \hat{K}(x, u, v) T(du) \geqq J^{n+1}(x).$$

Therefore

$$\min_{R \in P(V(x_0))} \max_{u \in U(x_0)} \int_{V(x_0)} \hat{K}(x, u, v) R(dv) = \min_{R \in P(V_x)} \max_{u \in U_x} \int_{V_x} K(x, u, v) R(dv) =: J^{n+1}(x)$$

for every $x \in \overline{N(x_0)}$. However, $\hat{K}$ does not feature state-constrained control sets, and therefore the proof of Lemma 2 shows that $J^{n+1}(x)$ is continuous at x_0. Since x_0 was arbitrary this proves the lemma.

Remark 2. Condition (i) of the continuous model can be generalized to:

(i') For each $x \in X$, U_x and V_x are compact subsets of U and V such that:

(a) There exists a $\varphi : X \times U \times V \to D = \{(x, u, v)\colon x \in X, u \in U_x, v \in V_x\}$ such that φ is continuous, $\varphi(x, u, v) = (x, u'(x, u), v)$ for $v \in V_x$ and $\varphi(x, u, v) = (x, u, v'(x, v))$ for $u \in U_x$.

(b) For every $x_0 \in X$ there exists an open neighborhood of x_0 such that $\cup\{X_x : x \in \overline{N(x_0)}\}$ and $\cup\{V_x : x \in \overline{N(x_0)}\}$ are compact.

The proof is the same as Lemma 4.

4. The value function. We now show that $J^*(\cdot)$ is a lower bound for the lower value of the game.

LEMMA 5. *For every* $x_0 \in X$, $\sup_{F \in M_1} \inf_{G \in D_2} J(x_0; F, G) \geqq J^*(x_0)$.

Proof. Let $F_{n+1} \in M_1$ be defined by

$$F_{n+1} = \{T^n, T^{n-1}, \cdots, T^2, T^1, T^0, T^0, T^0, \cdots\}$$

where $T^n(du|x)$ is defined as in Lemma 2. Then

$$
\begin{aligned}
&E_{F_{n+1},G}\Big[\sum_{k=0}^{n}\alpha^k c(x_k,u_k,v_k,x_{k+1})|x_0\Big]\\
&=E_{F_{n+1},G}\Big[\sum_{k=0}^{n-1}\alpha^k c(x_k,u_k,v_k,x_{k+1})|x_0\Big]+E_{F_{n+1},G}[\alpha^n c(x_n,u_n,v_n,x_{n+1})|x_0]\\
&=E_{F_{n+1},G}\Big[\sum_{k=0}^{n-1}\alpha^k c(x_k,u_k,v_k,x_{k+1})|x_0\Big]\\
&\quad+\alpha^n E_{F_{n+1},G}\Big\{E_{F_{n+1},G}\Big[\int_{V_{x_n}}\int_{U_{x_n}}\int_X c(x_n,u_n,v_n,y)q(dy|x_n,u_n,v_n)T^0(du_n|x_n)\\
&\qquad\qquad\cdot G^n(dv_n|x_0,v_0,x_1,v_1,\cdots,x_n)|x_0,v_0,x_1,v_1,\cdots,x_n\Big]\Big|x_0\Big\}\\
&\geqq E_{F_{n+1},G}\Big[\sum_{k=0}^{n-1}\alpha^k c(x_k,u_k,v_k,x_{k+1})|x_0\Big]+\alpha^n E_{F_{n+1},G}[J^1(x_n)|x_0]\\
&=E_{F_{n+1},G}\Big[\sum_{k=0}^{n-2}\alpha^k c(x_k,u_k,v_k,x_{k+1})|x_0\Big]\\
&\quad+\alpha^{n-1}E_{F_{n+1},G}[\alpha J^1(x_n)+c(x_{n-1},u_{n-1},v_{n-1},x_n)|x_0]\\
&=E_{F_{n+1},G}\Big[\sum_{k=0}^{n-2}\alpha^k c(x_k,u_k,v_k,x_{k+1})|x_0\Big]\\
&\quad+\alpha^{n-1}E_{F_{n+1},G}\Big\{E_{F_{n+1},G}\Big[\int_{V_{x_{n-1}}}\int_{U_{x_{n-1}}}\int_X[\alpha J^1(y)+c(x_{n-1},u_{n-1},v_{n-1},y)]\\
&\qquad\qquad\cdot q(dy|x_{n-1},u_{n-1},v_{n-1})T^1(du_{n-1}|x_{n-1})\\
&\qquad\qquad\cdot G^{n-1}(dv_{n-1}|x_0,v_0,x_1,v_1,\cdots,x_{n-1})|\\
&\qquad\qquad\cdot x_0,v_0,\cdots,x_{n-1}\Big]\Big|x_0\Big\}\\
&\geqq E_{F_{n+1},G}\Big[\sum_{k=0}^{n-2}\alpha^k c(x_k,u_k,v_k,x_{k+1})|x_0\Big]+\alpha^{n-1}E_{F_{n+1},G}[J^2(x_{n-1})|x_0]\\
&\qquad\qquad\vdots\\
&\geqq \alpha^0 E_{F_{n+1},G}[J^{n+1}(x_0)|x_0]\\
&=J^{n+1}(x_0).
\end{aligned}
$$

Hence $J(x_0;F_{n+1},G)\geqq J^{n+1}(x_0)$ for any $G\in D_2$. Hence $\inf_{G\in D_2}J(x_0;F_{n+1},G)\geqq J^{n+1}(x_0)$ and therefore, $\sup_{F\in M_1}\inf_{G\in D_2}J(x_0;F,G)\geqq J^{n+1}(x_0)$ for arbitrary n, proving that $\sup_{F\in M_1}\inf_{G\in D_2}J(x_0;F,G)\geqq\sup_n J^{n+1}(x_0)=J^*(x_0)$.

We now provide the following characterization of $J^*(\cdot)$. This characterization is exceedingly useful in many applications.

THEOREM 1.

(D) $J^*(\cdot)$ *is the* unique *solution of* (7).

(P) $J^*(\cdot)$ *satisfies* (8). *If any nonnegative* $J(\cdot)$ *satisfies* (9) *then* $J^*(x)\leqq J(x)$ *for every* $x\in X$.

Proof. We first show that $J^*(\cdot)$ satisfies (8). Since $J^n(y) \leqq J^*(y)$ for every $y \in X$, from (10) we obtain

$$J^{n+1}(x) \leqq \inf_{R \in P(V_x)} \sup_{u \in U_x} \int_{V_x} \int_X [\alpha J^*(y) + c(x, u, v, y)] q(dy|x, u, v) R(dv).$$

Since this is true for every n, we obtain

$$J^*(x) \leqq \inf_{R \in P(V_x)} \sup_{u \in U_x} \int_{V_x} \int_X [\alpha J^*(y) + c(x, u, v, y)] q(dy|x, u, v) R(dv).$$

To prove the reverse inequality consider $R^n(\cdot|x) \in P(V_x)$, which achieves the outer minimum in (10). Then

$$J^*(x) \geqq J^{n+1}(x) = \max_{u \in U_x} \int_{V_x} \int_X [\alpha J^n(y) + c(x, u, v, y)] q(dy|x, u, v) R^n(dv|x). \tag{14}$$

Since V_x is compact, $P(V_x)$ is compact [12, II.6.4] for fixed x, and hence the sequence $\{R^n(\cdot|x)\} \subset P(V_x)$ has a subsequence $\{R^{n_k}(\cdot|x)\}$ which converges to $\tilde{R} \in P(V_x)$. Fix x, relabel the sequence $\{R^{n_k}(\cdot|x)\}$ as $\{R^k\}$ and define

$$L^k(x, u, v) := \int_X [\alpha J^{n_k}(y) + c(x, u, v, y)] q(dy|x, u, v).$$

By (14)

$$J^*(x) \geqq \max_{u \in U_x} \int_{V_x} L^k(x, u, v) R^k(dv) \quad \text{for all } k.$$

Hence

$$J^*(x) \geqq \int_{V_x} L^k(x, u, v) R^k(dv) \quad \text{for all } k \text{ and } u \in U_x,$$

and

$$J^*(x) \geqq \sup_k \int_{V_x} L^k(x, u, v) R^k(dv) \quad \text{for all } u \in U_x. \tag{15}$$

Clearly, $0 \leqq L^1(x, u, v) \leqq L^2(x, u, v) \leqq \cdots \leqq +\infty$, and

$$\begin{aligned} \lim_{k \to \infty} L^k(x, u, v) &= \lim_{k \to \infty} \int_X [\alpha J^{n_k}(y) + c(x, u, v, y)] q(dy|x, u, v) \\ &= \int_X [\alpha J^*(y) + c(x, u, v, y)] q(dy|x, u, v), \end{aligned} \tag{16}$$

by the monotone convergence theorem.

Denote

$$L(x, u, v) := \int_X [\alpha J^*(y) + c(x, u, v, y)] q(dy|x, u, v) \leqq +\infty.$$

We now proceed to show that

$$\sup_k \int_{V_x} L^k(x, u, v) R^k(dv) \geqq \int_{V_x} L(x, u, v) \tilde{R}(dv). \tag{17}$$

By the monotone convergence theorem again, (16) implies

$$\int_{V_x} L(x, u, v)\tilde{R}(dv) = \lim_{k\to\infty} \int_{V_x} L^k(x, u, v)\tilde{R}(dv|x).$$

Hence given $\varepsilon > 0$, there exists N so large that

$$(18)\quad \int_{V_x} L^N(x, u, v)\tilde{R}(dv) \geqq \begin{cases} \displaystyle\int_{V_x} L(x, u, v)\tilde{R}(dv) - \varepsilon & \text{if } \displaystyle\int_{V_x} L(x, u, v)\tilde{R}(dv) < \infty, \\ \dfrac{1}{\varepsilon} & \text{if } \displaystyle\int_{V_x} L(x, u, v)\tilde{R}(dv) = \infty. \end{cases}$$

Now fix N and $u \in U_x$, $L^N(x, u, \cdot)$ is bounded and continuous on V_x and $R^k \to \tilde{R}$ as $k \to \infty$. From [6, Prop. 7.21] we obtain

$$\int_{V_x} L^N(x, u, v)R^k(dv|x) \to \int_{V_x} L^N(x, u, v)\tilde{R}(dv|x) \quad \text{as } k \to \infty.$$

Hence there exists a K such that for all $k \geqq K$

$$(19)\qquad \int_{V_x} L^N(x, u, v)R^k(dv) \geqq \int_{V_x} L^N(x, u, v)\tilde{R}(dv) - \varepsilon.$$

We now consider two cases.

Case 1. Suppose $N \geqq K$. Replacing k by N in (19), we obtain

$$\int_{V_x} L^N(x, u, v)R^N(dv) \geqq \int_{V_x} L^N(x, u, v)\tilde{R}(dv) - \varepsilon.$$

This together with (18) implies that

$$\int_{V_x} L^N(x, u, v)R^N(dv) \geqq \begin{cases} \displaystyle\int_{V_x} L(x, u, v)\tilde{R}(dv) - 2\varepsilon & \text{if } \displaystyle\int_{V_x} L(x, u, v)\tilde{R}(dv) < \infty, \\ \dfrac{1}{\varepsilon} - \varepsilon & \text{otherwise.} \end{cases}$$

Hence

$$(20)\quad \sup_k \int_{V_x} L^k(x, u, v)R^k(dv) \geqq \begin{cases} \displaystyle\int_{V_x} L(x, u, v)\tilde{R}(dv) - 2\varepsilon & \text{if } \displaystyle\int_{V_x} L(x, u, v)\tilde{R}(dv) < \infty, \\ \dfrac{1}{\varepsilon} - \varepsilon & \text{otherwise.} \end{cases}$$

Since ε was arbitrary, (17) is proved.

Case 2. Suppose $N < K$. Then $L^N(x, u, v) \leqq L^K(x, u, v)$,

$$\int_{V_x} L^K(x, u, v)R^K(dv) \geqq \int_{V_x} L^N(x, u, v)R^K(dv),$$

and the latter in turn is greater than or equal to $\int_{V_x} L^N(x, u, v)\tilde{R}(dv) - \varepsilon$ by (19). We thus obtain

$$\sup_k \int_{V_x} L^k(x, u, v)R^k(dv) \geqq \int_{V_x} L^N(x, u, v)\tilde{R}(dv) - \varepsilon.$$

Again, this together with (18) implies (20) and then (17) follows.

Now (15) and (17) together imply

$$J^*(x) \geqq \int_{V_x} L(x, u, v)\tilde{R}(dv) \quad \text{for every } u \in U_x.$$

Hence

$$J^*(x) \geqq \sup_{u \in U_x} \int_{V_x} L(x, u, v)\tilde{R}(dv),$$

i.e.,

$$J^*(x) \geqq \sup_{u \in U_x} \int_{V_x} \int_X [\alpha J^*(y) + c(x, u, v, y)]q(dy|x, u, v)\tilde{R}(dv).$$

Hence we obtain the reverse inequality and $\tilde{R}$ achieves the minimum. (Note that $\tilde{R}$ depends on x).

(P) Hence $J^*(\cdot)$ satisfies (8). Suppose $J(\cdot)$ is nonnegative and satisfies (9). Then $J(x) \geqq J^0(x) \equiv 0$ for all $x \in X$. Hence, by induction, using (9) and (10), $J(x) \geqq J^n(x)$ for all $x \in X$ and all n, therefore $J(x) \geqq J^*(x)$.

(D) As earlier, $J^*(\cdot)$ satisfies (8). From (13), $J^*(\cdot)$ is bounded by $\theta/(1-\alpha)$. In case B, U_x and V_x are finite for each x and by von Neumann's Fundamental Theorem of Matrix Games [7], (8) and (7) are equivalent. In Case C, by Lemma 3, $J^*(\cdot)$ is continuous and since U_x and V_x are compact for each x, it follows from Lemma 1 that (8) and (7) are equivalent. Hence, in summary, $J^*(\cdot)$ satisfies (7) in case D. To show it is the unique solution of (7), we proceed as follows:

Let $J(\cdot)$ be a nonnegative solution of (7). Given x, let

$$K(T, R) = \int_{V_x} \int_{U_x} \int_X [\alpha J(y) + c(x, u, v, y)]q(dy|x, u, v)T(du)R(dv),$$

$$K^*(T, R) = \int_{V_x} \int_{U_x} \int_X [\alpha J^*(y) + c(x, u, v, y)]q(dy|x, u, v)T(du)R(dv).$$

Clearly,

$$K(T, R) - K^*(T, R) = \int_{V_x} \int_{U_x} \int_X \alpha(J(y) - J^*(y))q(dy|x, u, v)T(du)R(dv)$$

$$\leqq \alpha\|J - J^*\| \quad \text{for all } T, R, \text{ where } \|\cdot\| \text{ is the sup norm.}$$

Hence $K(T, R) \leqq K^*(T, R) + \alpha\|J - J^*\|$, and therefore

$$\max_{T \in P(U_x)} K(T, R) \leqq \max_{T \in P(U_x)} K^*(T, R) + \alpha\|J - J^*\|,$$

and therefore

$$\min_{R \in P(V_x)} \max_{T \in P(U_x)} K(T, R) \leqq \min_{R \in P(V_x)} \max_{T \in P(U_x)} K^*(T, R) + \alpha\|J - J^*\|;$$

i.e., $J(x) \leqq J^*(x) + \alpha\|J - J^*\|$. We obtain $J(x) - J^*(x) \leqq \alpha\|J - J^*\|$. Similarly, we can show $J^*(x) - J(x) \leqq \alpha\|J - J^*\|$, therefore $|J(x) - J^*(x)| \leqq \alpha\|J - J^*\|$. Since x was arbitrary, we have $\|J - J^*\| \leqq \alpha\|J - J^*\|$. Because $\alpha < 1$, we must have $\|J - J^*\| = 0$, i.e., $J(x) \equiv J^*(x)$, proving that $J^*(x)$ is unique.

5. Optimal stationary strategy for minimizer. In this section we show that $J^*(\cdot)$ is in fact the value of the game. Additionally we show that player II has an optimal

stationary strategy and show how such a strategy can be obtained from knowledge of the value function $J^*(\cdot)$.

THEOREM 2.

(i) *$J^*(\cdot)$ is the value function of the game, i.e.,*

$$\inf_{G\in D_2}\sup_{F\in D_1} J(x_0; F, G) = \sup_{F\in D_1}\inf_{G\in D_2} J(x_0; F, G) = J^*(x_0) \quad \textit{for every } x_0 \in X.$$

(ii) *Any $G^* = \{G^0, G^0, G^0, \cdots\} \in S_2$ such that $G^0(dv|x)$ achieves the outer minimum in* (8) *for every x is an optimal stationary strategy for player* II, *i.e.,*

$$J(x_0; F, G^*) \leqq J^*(x_0) \quad \textit{for every } F \in D_1 \textit{ and } x_0 \in X.$$

There always exists such a G^*.

(iii) *Denoting by $J_\alpha^*(x)$ the value function of the game with discount factor α, we have $\lim_{\alpha\uparrow 1} J_\alpha^*(x) = J_1^*(x)$.*

Proof. (i and ii). We begin by showing the existence of a $G^* = \{G^0, G^0, \cdots\} \in S_2$ such that $G^0(dv|x)$ achieves the outer minimum in (8) for each $x \in X$.

(B) Since $J^*(\cdot)$ is measurable, the proof of existence of a *measurable* $R^n(dv|x)$ achieving the outer minimum in (10) given in Lemma 2 is applicable.

(C) $J^*(\cdot)$ is lower-semicontinuous (l.s.c.), and hence the map

$$k(x, u, v) := \int_X [\alpha J^*(y) + c(x, u, v, y)] q(dy|x, u, v)$$

is l.s.c. [6, Prop. 7.31]. The map

$$l(x, u, R) := \int_V k(x, u, v) R(dv)$$

is therefore l.s.c. (see proof of Lemma 2). The map $m(x, R) := \sup_{u\in U} l(x, u, R)$ is therefore l.s.c. since $\{(x, r)|m(x, r) > \gamma\} = U_{u\in U}\{(x, R)|l(x, u, R) > \gamma\}$ is open for all $\gamma \in \mathbb{R}$. Now [8, Corollary 1] implies the existence of a Borel measurable stochastic kernel, say G^0, which achieves the minimum in $\min_{R\in P(Vx)}(m(x, R)$.

To show $G^* = \{G^0, G^0, \cdots\}$ is optimal, let $F \in D$ be arbitrary. Then

$$\begin{aligned}
&E_{F,G^*}\Big[\sum_{k=0}^{n} \alpha^k c(x_k, u_k, v_k, x_{k+1})|x_0\Big] + \alpha^{n+1} E_{F,G^*}[J^*(x_{n+1})|x_0] \\
&= E_{F,G^*}\Big[\sum_{k=0}^{n-1} \alpha^k c(x_k, u_k, v_k, x_{k+1})|x_0\Big] + \alpha^n E_{F,G^*}[\alpha J^*(x_{n+1}) + c(x_n, u_n, v_n, x_{n+1})|x_0] \\
&= E_{F,G^*}\Big[\sum_{k=0}^{n-1} \alpha^k c(x_k, u_k, v_k, x_{k+1})|x_0\Big] \\
&\quad + \alpha^n E_{F,G^*}\Big\{E_{F,G^*}\Big[\int_{U_{x_n}}\int_{V_{x_n}}\int_X [\alpha J^*(y) + c(x_n, u_n, v_n, y)] \\
&\qquad \cdot q(dy|x_n, u_n, v_n) G^0(dv_n|x_n) \\
&\qquad \cdot F^n(du_n|x_0, u_0, x_1, u_1, \cdots, x_n)|x_0, u_0, x_1, u_1, \cdots, x_n\Big]\Big|x_0\Big\} \\
&\leqq E_{F,G^*}\Big[\sum_{k=0}^{n-1} \alpha^k c(x_k, u_k, v_k, x_{k+1})|x_0\Big]
\end{aligned}$$

$$(21)\qquad +\alpha^n E_{F,G^*}\Big\{E_{F,G^*}\Big[\max_{u_n\in U_{x_n}}\int_{V_{x_n}}\int_X [\alpha J^*(y)+c(x_n,u_n,v_n,y)]$$

$$\cdot\, q(dy|x_n,u_n,v_n)G^0(dv_n|x_n)|x_0,u_0,x_1,u_1,\cdots,x_n\Big]|x_0\Big\}$$

$$=E_{F,G^*}\Big[\sum_{k=0}^{n-1}\alpha^k c(x_k,u_k,v_k,x_{k+1})|x_0\Big]+\alpha^n E_{F,G^*}[J^*(x_n)|x_0]$$

$$\leqq E_{F,G^*}\Big[\sum_{k=0}^{n-2}\alpha^k c(x_k,u_k,v_k,x_{k+1})|x_0\Big]$$

$$+\alpha^{n-1}E_{F,G^*}[\alpha J^*(x_n)+c(x_{n-1},u_{n-1},v_{n-1},x_n)|x_0]$$

$$\leqq E_{F,G^*}\Big[\sum_{k=0}^{n-3}\alpha^k c(x_k,u_k,v_k,x_{k+1})|x_0\Big]$$

$$+\alpha^{n-2}E_{F,G^*}[\alpha J^*(x_{n-1})+c(x_{n-2},u_{n-2},v_{n-2},x_{n-1})|x_0]$$

$$\vdots$$

$$\leqq \alpha^0 E_{F,G^*}[J^*(x_0)|x_0]=J^*(x_0).$$

Hence $E_{F,G^*}[\sum_{k=0}^n \alpha^k c(x_k,u_k,v_k,x_{k+1})|x_0]\leqq J^*(x_0)$. Since this holds true for every n, by the monotone convergence theorem as $n\to\infty$ we obtain $J(x_0;F,G^*)\leqq J^*(x_0)$. Since $F\in D_1$ was arbitrary, we obtain $\sup_{F\in D_1}J(x_0;F,G^*)\leqq J^*(x_0)$. This combined with Lemma 5 shows that $J^*(x_0)$ is the value and also that G^* is optimal for every $x_0\in X$, and completes the proof of (i) and (ii).

(iii) Let $\mathrm{Val}\,(x)=\lim_{\alpha\uparrow 1}J^*(x)$; then $\mathrm{Val}\,(x)\leqq J_1^*(x)$. To show the reverse inequality, we first show

$$(8')\qquad \mathrm{Val}\,(x)=\min_{R\in P(V_x)}\sup_{u\in U_x}\int_{V_x}\int_X [\mathrm{Val}\,(y)+c(x,u,v,y)]q(dy|x,u,v)R(dv).$$

This inequality is similar to (8) and the proof is the same; hence we use the notation (8'), (9'), (10'), $\cdots$, etc.

From (8),

$$(22)\qquad J_\alpha^*(x)=\min_{R\in P(V_x)}\sup_{u\in U_x}\int_{V_x}\int_X [\alpha J_\alpha^*(y)+c(x,u,v,y)]q(dy|x,u,v)R(dv),$$

and since $\alpha J_\alpha^*(y)\leqq \mathrm{Val}\,(y)$ for all $\alpha\in(0,1)$, we obtain that the RHS of (22) is less than or equal to the RHS of (8'). Hence

$$J_\alpha^*(x)\leqq \min_{R\in P(V_x)}\sup_{u\in U_x}\int_{V_x}\int_X [\mathrm{Val}\,(y)+c(x,u,v,y)]q(dy|x,u,v)R(dv).$$

Since this is true for all $\alpha\in(0,1)$, as $\alpha\uparrow 1$ we obtain

$$\mathrm{Val}\,(x)\leqq \min_{R\in P(V_x)}\sup_{u\in U_x}\int_{V_x}\int_X [\mathrm{Val}\,(y)+c(x,u,v,y)]q(dy|x,u,v)R(dv).$$

To prove the reverse inequality, let $\{\alpha_n\}$ be a monotonically increasing sequence with $\alpha_n<1$ and $\lim_{n\to\infty}\alpha_n=1$. Fix x, let $\tilde{R}^n\in P(V_x)$ be such that $\tilde{R}^n$ achieves the outer minimum in (22) when $\alpha=\alpha_n$. Then

$$(14')\quad \mathrm{Val}\,(x)\geqq J_{\alpha_n}^*(x)=\max_{u\in U_x}\int_{V_x}\int_X [\alpha_n J_{\alpha_n}^*(y)+c(x,u,v,y)]q(dy|x,u,v)\tilde{R}^n(dv).$$

Now the same argument as in the proof of Theorem 1 shows that (15′)–(20′) are true. Hence (8′) is also true. From Theorem 1 we know that $J_1^*(\cdot)$ is the minimum (at each x) among all the nonnegative functions which satisfy (8). Hence $\text{Val}(x) \geqq J_1^*(x)$ for each $x \in X$, proving (iii).

6. Optimal and near optimal strategies for the maximizer. Player I does not always have optimal strategies, stationary or nonstationary, in the positive case. We give an example in § 7. In the positive case we therefore prove the existence of near optimal strategies, appropriately defined in Theorem 3 below. In the discounted case, however, player I always has an optimal stationary strategy.

THEOREM 3.

(D) *Let* $F^* = \{F^0, F^0, F^0, \cdots\} \in S_1$ *be such that* $F^0(du|x)$ *achieves the outer maximum in* (7). *Then* F^* *is an optimal stationary strategy, i.e.,*

$$J(x_0; F^*, G) \geqq J^*(x_0) \quad \text{for every } G \in D_2 \text{ and } x_0 \in X.$$

There always exists such an F^*.

(P)(i) *For any* $\lambda \in P(X)$ *and any* $\varepsilon > 0$, *there exists a stationary strategy* $F \in S_1$ *such that*

$$\lambda\left(\left\{\begin{array}{ll} x: \inf\limits_{G \in D_2} J(x; F, G) \geqq J^*(x) - \varepsilon & \text{if } J^*(x) < \infty \\ \qquad\qquad \geqq \dfrac{1}{\varepsilon} & \text{if } J^*(x) = \infty \end{array}\right\}\right) \geqq 1 - \varepsilon.$$

In particular, for every finite subset $S \subset X$ *and* $\varepsilon > 0$, *there exists an* $F \in S_1$ *such that*

$$J(x; F, G) \geqq \begin{cases} J^*(x) - \varepsilon & \text{if } J^*(x) < \infty \text{ and } x \in S, \\ \dfrac{1}{\varepsilon} & \text{if } J^*(x) = \infty \text{ and } x \in S \end{cases}$$

for every $G \in D_2$.

(ii) *If* X *is compact and* $J^*(\cdot)$ *is continuous, then for every* $\varepsilon > 0$, *player* I *has an* ε*-optimal stationary stationary strategy; i.e., there exists an* $F \in S_1$ *such that* $J(x; F, G) \geqq J^*(x) - \varepsilon$ *for every* $x \in X$.

Remark 3. To see that in (P)(i) the λ measure of the set is well defined, we show that $\inf_{G \in D_2} J(\cdot\,; F, G)$ is universally measurable. This is true by Strauch [9], since for fixed F the problem reduces to a dynamic programming problem with maximization of a "negative" cost as the objective.

Proof. (D) From Lemma 3, $J^*(\cdot)$ is bounded; hence, as in the proof of Lemma 2, there exists an F^* satisfying the conditions of the theorem. In Theorem 2, we have shown that G^* is optimal both for $\alpha < 1$ and $\alpha = 1$ by (21). Replacing (F, G^*) by (F^*, G), $\max_{u \in U_x}$ by $\min_{v \in V_x}$ and reversing all the inequalities in (21), we obtain

$$E_{F^*,G}\left[\sum_{k=0}^{n} \alpha^k c(x_k, u_k, v_k, x_{k+1}) | x_0\right] + \alpha^{n+1} E_{F^*,G}[J^*(x_{n+1}) | x_0]$$

$$\geqq J^*(x_0) \quad \text{for all } x_0 \in X, \quad \text{all } G \in D_2.$$

In our case since $\alpha < 1$, $\|J^*\|_\infty < \infty$, letting $n \to \infty$, we obtain

$$E_{F^*,G}\left[\sum_{k=0}^{\infty} \alpha^k c(x_k, u_k, v_k, x_{k+1}) | x_0\right] \geqq J^*(x_0) \quad \text{for all } x_0 \in X, \quad \text{all } G \in D_2.$$

Hence F^* is optimal, proving (D).

(P)(i) From Theorem 2, we have $J_\alpha^*(x) \uparrow J_1^*(x)$ as $\alpha \uparrow 1$. Let $\alpha_n := 1-(1/n)$, $I_n(x) := J_{\alpha_n}^*(x)$; then as $n \to \infty$, $\alpha_n \uparrow 1$ and $I_n(x) \uparrow J_1^*(x)$. Given $\varepsilon > 0$, let

$$E_n = \{x \in X : I_n(x) < J_1^*(x) - \varepsilon, \quad \text{if } J_1^*(x) < \infty, \quad \text{or } I_n(x) < \frac{1}{\varepsilon} \quad \text{if } J_1^*(x) = \infty\}.$$

Then $E_1 \supset E_2 \supset E_3 \supset \cdots$ and also $\cap_{n-1}^{\infty} E_n = \varnothing$. Hence for any $\lambda \in P(X)$, $\lim_{n\to\infty} \lambda(E_n) = 0$. Choose N so large that $\lambda(E_N) < \varepsilon$. We have for $x \notin E_N$,

$$I_N(x) \geqq \begin{cases} J^*(x) - \varepsilon & \text{if } J^*(x) < \infty, \\ \dfrac{1}{\varepsilon} & \text{if } J^*(x) = \infty. \end{cases}$$

Let F_N be the optimal stationary strategy in the game with discount factor α_N given above in the proof of (D). Then

$$\min_G J(x; F_N, G) \geqq I_N(x) \geqq \begin{cases} J^*(x) - \varepsilon & \text{if } J^*(x) < \infty, \\ \dfrac{1}{\varepsilon} & \text{if } J^*(x) = \infty. \end{cases}$$

This proves the first assertion. For the second, if $S \subset X$ is a finite set with m elements, let λ be the uniform distribution on S. Now let $\bar{\varepsilon} < \min(\varepsilon, 1/m)$. Then the stationary strategy F such that

$$\lambda\left(\left\{x : \inf_G J(x; F, G) \geqq \begin{cases} J_{\text{opt}}(x) - \bar{\varepsilon} & \text{if } J^*(x) < \infty \\ \dfrac{1}{\bar{\varepsilon}} & \text{if } J^*(x) = \infty \end{cases}\right\}\right) \geqq 1 - \bar{\varepsilon}$$

is ε-optimal for $x \in S$.

(ii) Let J_α^*, J_α, F_α^*, G_α^* be J^*, J, F^*, G^* respectively in the game with discount factor $\alpha \leqq 1$. Since X is compact and $J_1^*(x)$ is a continuous real function, Dini's theorem implies $J_1^*(x) - J_\alpha^*(x) \downarrow 0$ uniformly on X as $\alpha \uparrow 1$, i.e., $J_\alpha^*(x) \uparrow J_1^*(x)$ uniformly on X as $\alpha \uparrow 1$. Given $\varepsilon > 0$, choose α close to 1 so that $J_\alpha^*(x) \geqq J_1^*(x) - \varepsilon$ for all $x \in X$. Then F_α^* is ε-optimal in the Positive ($\alpha = 1$) Case, since $J(x; F_\alpha^*, G) \geqq J_\alpha(x; F_\alpha^*, G) \geqq J_\alpha^*(x) \geqq J_1^*(x) - \varepsilon$ for all $x \in X$ and for all $G \in D_2$.

Remark 4. In the proof of Theorem 3, it is clear that if $J_\alpha^*(\cdot) \uparrow J_1^*(\cdot)$ uniformly as $\alpha \uparrow 1$, then player I has an ε-optimal stationary strategy, even if X is not compact. On the other hand, an example in § 7 shows that even if the convergence of $J_\alpha^*(\cdot)$ to $J_1^*(\cdot)$ is not uniform, player I can have an ε-optimal stationary strategy for every $\varepsilon > 0$. Whether player I always has an ε-optimal stationary strategy is an open problem. Some related papers are Ornstein [10] and Bertsekas and Shreve [11].

Remark 5. (C) If $J_1^*(\cdot)$ is a continuous real function, then for any compact subset $Y \subset X$, $J_\alpha^* \uparrow J_1^*$ uniformly on Y (Dini's theorem). Hence for any $\varepsilon > 0$, player I has a stationary strategy which is ε-optimal on Y. The proof is similar.

7. Some examples. Our first example, adapted from [2], shows that even in particularly simple problems, player I need not have an optimal strategy—stationary or nonstationary.

Example 1. Let

$$
\begin{aligned}
&X = \{x_0, x_1, x_2\},\\
&U_x = \{1, 2\} && \text{for all } x \in X,\\
&V_x = \{1, 2\} && \text{for all } x \in X,\\
&f(x_0, u, v) = x_0 && \text{for all } (u, v),\\
&f(x_1, u, v) = x_1 && \text{for all } (u, v),\\
&f(x_2, u, v) = x_0 && \text{if } u = v,\\
&f(x_2, 1, 2) = x_2,\\
&f(x_2, 2, 1) = x_1,\\
&c(x, u, v, y) = \begin{cases} 1 & \text{if } x \neq x_0,\ y = x_0,\\ 0 & \text{otherwise.} \end{cases}
\end{aligned}
$$

x_0 and x_1 are absorption points of the system. Hence the only nonzero cost-transition is the transition from x_2 to x_0 which yields a cost $= 1$. Clearly $J^*(x_0) = 0$, $J^*(x_1) = 0$. Hence we need consider only x_2. By Theorem 1, $J^*(x_2)$ is the smallest nonnegative number satisfying

$$J^*(x_2) = \text{value} \begin{bmatrix} 1 & J^*(x_2) \\ 0 & 1 \end{bmatrix}.$$

Since

$$\text{value} \begin{bmatrix} 1 & J^*(x_2) \\ 0 & 1 \end{bmatrix} = \frac{1}{2 - J^*(x_2)},$$

the equation $J^*(x_2) = 1/(2 - J^*(x_2))$ has a unique solution $J^*(x_2) = 1$, which must be the value at x_2. However there is no optimal strategy for player I. To show this consider two cases.

Case 1. Player I always plays $u = 1$. Then if player II always plays $v = 2$, the system stays indefinitely in x_2 and the total cost is only 0.

Case 2. At some time n, player I plays $u = 1$ with probability $1 - \varepsilon$, and $u = 2$ with probability $\varepsilon > 0$. But then if player II at time n chooses $v = 1$, the system ends in state x_1 with probability at least $\varepsilon > 0$, hence the total cost is less than or equal to $1 - \varepsilon$.

Hence player I has no optimal strategy. Note however that if player I chooses $u = 1$ with probability $1 - \varepsilon$ and $u = 2$ with probability ε, then this stationary strategy is ε-optimal.

Our second example shows that in the continuous model, $J^*(\cdot)$ need not be continuous even when X is compact. This example therefore cannot be solved by the results of [4], but can be by our results.

Example 2. Let

$$X = \left\{ \frac{1}{n} \middle| n \text{ positive integer} \right\} \cup \{0\},$$

$$U_x \equiv V_x \equiv \{1, 2\} \quad \text{for all } x \in X,$$

$$f\left(\frac{1}{n}, u, v\right) = \begin{cases} \dfrac{1}{n-1}, & \text{if } u = v, \\ \dfrac{1}{n+1}, & \text{if } u \neq v, \end{cases} \qquad \text{for } n \geqq 2,$$

$$f(1, u, v) = \begin{cases} 0, & \text{if } u = v, \\ \frac{1}{2}, & \text{if } u \neq v, \end{cases}$$

$$f(0, u, v) = 0, \qquad \text{for all } u, v,$$

$$c(x, u, v, y) = \begin{cases} 1, & \text{if } x = 1,\ y < \frac{1}{2}, \\ 0, & \text{otherwise.} \end{cases}$$

Briefly, the game can be interpreted as follows. Both players play a matrix game at $1/n$, with player I moving one step either towards or away from his goal $x = 0$ depending on whether $u = v$ or $u \neq v$. When player I reaches his goal $x = 0$ from $x = 1$, he gets a reward of 1 unit and the game ends.

To solve the game, we first observe that $J^*(0) = 0$. By Theorem 1, $J^*(\cdot)$ is the smallest nonnegative solution of

$$J^*\left(\frac{1}{n}\right) = \text{value of matrix game } \left[J^*\left(f\left(\frac{1}{n}, i, j\right)\right) + c\left(\frac{1}{n}, i, j, f\left(\frac{1}{n}, i, j\right)\right)\right]. \tag{23}$$

Since the value of the symmetric matrix game $\begin{bmatrix} a & b \\ b & a \end{bmatrix}$ is $\frac{1}{2}(a+b)$, (23) is equivalent to

$$J^*(1) = \tfrac{1}{2}[1 + J^*(\tfrac{1}{2})],$$

$$J^*\left(\frac{1}{n}\right) = \frac{1}{2}\left[J^*\left(\frac{1}{n-1}\right) + J^*\left(\frac{1}{n+1}\right)\right] \quad \text{for } n \geqq 2,$$

i.e.,

$$\begin{aligned} J^*(\tfrac{1}{2}) &= 2J^*(1) - 1, \\ J^*\left(\frac{1}{n+1}\right) &= 2J^*\left(\frac{1}{n}\right) - J^*\left(\frac{1}{n-1}\right) \quad \text{for } n \geqq 2. \end{aligned} \tag{24}$$

By induction, it follows that

$$J^*\left(\frac{1}{n+1}\right) = (n+1)J^*(1) - n.$$

Hence

$$J^*(1) = \frac{n}{n+1} + \frac{1}{n+1}J^*\left(\frac{1}{n+1}\right) \geqq \frac{n}{n+1} \quad \text{for all } n,$$

therefore $J^*(1) \geqq 1$. Also,

$$J^*\left(\frac{1}{n+1}\right) = (n+1)J^*(1) - n \geqq 1.$$

On the other hand $J^*(1/n) \equiv 1$ is a solution of (24), and hence from Theorem 1 we obtain $J^*(1/n) \equiv 1$.

Note now that X is compact with the relative Euclidean topology, $q(\cdot \mid x, u, v)$ and $c(x, u, v, y)$ are continuous, (we only need to check at $x = 0$), yet [4] does not work

because J^* is not continuous at $x = 0$ (hence J_α^* does not converge uniformly to J^* as $\alpha \uparrow 1$). However, we note that player I actually has an optimal stationary strategy which consists of choosing the probability vector $(\frac{1}{2}, \frac{1}{2})$ always.

REFERENCES

[1] L. S. SHAPLEY, *Stochastic games*, Proc. Nat. Acad. Sci., 39 (1953), pp. 1095–1100.
[2] H. EVERETT, *Recursive games*, Ann. Math. Stud., 39 (1957), pp. 47–78.
[3] A. MAITRA AND T. PARTHASARATHY, *On stochastic games*, J. Opt. Theor. Appl., 5 (1970), pp. 289–300.
[4] A. MAITRA AND T. PARTHASARATHY, *On stochastic games* II, J. Opt. Theor. Appl., 8 (1971), pp. 154–160.
[5] T. PARTHASARATHY AND T. E. S. RAGHAVAN, *Some Topics in Two Person Games*, American Elsevier, New York, 1971.
[6] D. P. BERTSEKAS AND S. E. SHREVE, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, 1978.
[7] J. VON NEUMANN AND O. MORGENSTERN, *Theory of Games and Economic Behavior*, University Press, Princeton, NJ, 1944.
[8] L. D. BROWN AND R. PURVES, *Measurable selections of extrema*, Ann. Stat., 1 (1973), pp. 902–912.
[9] R. E. STRAUCH, *Negative dynamic programming*, Ann. Math. Stat., 3 (1966), pp. 871–890.
[10] D. ORNSTEIN, *On the existence of stationary optimal strategies*, Proc. Am. Math. Soc., 20 (1969), pp. 563–569.
[11] D. P. BERTSEKAS AND S. E. SHREVE, *Existence of optimal stationary policies in deterministic optimal control*, J. Math. An. Appl., 69 (1979), pp. 607–620.
[12] K. R. PARTHASARATHY, *Probability Measures on Metric Spaces*, Academic Press, New York, 1967.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, September 1981

0363-0129/81/1905-0006 $01.00/0

AVERAGING METHODS FOR THE ASYMPTOTIC ANALYSIS OF LEARNING AND ADAPTIVE SYSTEMS, WITH SMALL ADJUSTMENT RATE*

H. J. KUSHNER† AND HAI HUANG‡

Abstract. Recently proven theorems concerning weak convergence of nonMarkovian processes to diffusions, together with an averaging and a stability method, are applied to two (learning or adaptive) processes of current interest: (1) an automata model for route selection in telephone traffic routing; (2) an adaptive quantizer for use in the transmission of random signals in communication theory. The models are chosen because they are prototypes of a large class to which the methods can be applied. The technique of application of the basic theorems to such processes is developed. Suitably interpolated and normalized "learning or adaptive" processes converge weakly to a diffusion, as the "learning or adaptation" rate goes to zero. For small learning rates, the qualitative properties (e.g., asymptotic (large-time) variances and parametric dependence) of the processes can be determined from the properties of the limit.

1. Introduction. References [1], [7] develop a useful method of studying the asymptotic properties, as $\varepsilon \to 0$ and $n\varepsilon \leqq T < \infty$ for any real T, of solutions to stochastic difference equations of the form

$$Y_{n+1}^{\varepsilon} = Y_n^{\varepsilon} + \varepsilon h_{\varepsilon}(Y_n^{\varepsilon}, \xi_n^{\varepsilon}) + \sqrt{\varepsilon} g_{\varepsilon}(Y_n^{\varepsilon}, \xi_n^{\varepsilon}) + o(\varepsilon), \qquad Y_n^{\varepsilon} \in R^r, \tag{1.1}$$

where the distributions of the random sequence $\{\xi_n^{\varepsilon}\}$ might depend on the $\{Y_n^{\varepsilon}\}$. Such equations occur frequently in applications. The methods in [1] also work when ε is replaced by a sequence $\varepsilon_n \to 0$ as $n \to \infty$, from which asymptotic properties (rates of convergence) of various forms of stochastic approximations can be obtained.

The emphasis in [1] (an application of [7]) concerned the case where the h_{ε} and g_{ε} are smooth, and no details for the nonsmooth case or its applications were given, nor was the asymptotic case where $n \to \infty$, then $\varepsilon \to 0$ treated. This is a deficiency, since in many applications in communication, control and automata theory, the h_{ε} and g_{ε} might simply be indicator functions and the noise $\{\xi_n\}$ depend on $\{Y_n^{\varepsilon}\}$, and the asymptotic properties (as $n \to \infty$, then $\varepsilon \to 0$) desired. Here, we apply the basic results of [7] to two such problems. The two problems have current technological importance in their own right and each has been the subject of a great deal of work. Our method often yields a complete analysis of the asymptotic properties under realistic conditions. The two problems are typical of a wide class, and they illustrate the power and applicability of the general technique, as well as the method of applying it to concrete problems. In a sense the method is an extension with more complex memory structure of the sort of "slow learning" results obtained by Norman [9], and should have broad applications to the areas cited above.

The basic type of result is the following. Define $Y^{\varepsilon}(\cdot)$, $t \in [0, \infty)$, by $Y^{\varepsilon}(0) = Y_0^{\varepsilon}$ and $Y^{\varepsilon}(t) = Y_i^{\varepsilon}$, on $[i\varepsilon, i\varepsilon + \varepsilon)$. Under appropriate conditions, Theorem 1 gives weak convergence of $\{Y^{\varepsilon}(\cdot)\}$, in $D^r[0, \infty)$, to a particular diffusion process, as $\varepsilon \to 0$. Now, let $\{n_{\varepsilon}\}$ denote a sequence of integers tending to ∞ as $\varepsilon \to 0$. For $t \geqq 0$, define $\tilde{Y}^{\varepsilon}(t) \equiv Y^{\varepsilon}(t + \varepsilon n_{\varepsilon})$. The tilde $\tilde{}$ always denotes a shift by n_{ε} (discrete parameter) or $\varepsilon n_{\varepsilon}$ (continuous parameter). By using Theorem 1 but starting $\{Y_n^{\varepsilon}\}$ at time n_{ε} instead of at

* Received by the editors April 24, 1980, and in revised form January 12, 1981.

† Divisions of Applied Mathematics and Engineering, Brown University, Providence, RI 02912. The work of this author was supported by the Air Force Office of Scientific Research AF-76-3063, the National Science Foundation Eng. 77-12946, and the Office of Naval Research N00014-76-C-0279-P003.

‡ Division of Applied Mathematics, Brown University, Providence, RI 02912. The work of this author was supported by the National Science Foundation Eng. 77-12946 and the Office of Naval Research N00014-76-C-0279-P003.

time 0, we will get a great deal of information on the asymptotic properties (large n, small ε). The next section gives some background material from [7]. Sections 3–6 treat a learning automata approach to certain problems in adaptive routing of telephone calls [2]–[3]. The second problem, in §§ 7 and 8, concerns the asymptotic theory of an adaptive quantizer from communications applications [4], [5].

2. Some background material. $D^r[0, \infty)$ denotes the space of R^r-valued functions on $[0, \infty)$ which are right continuous and have left-hand limits; it is endowed with the Skorokhod topology [6]. $\hat{\mathscr{C}}_0$ denotes the continuous functions on $R^r \times [0, \infty)$ with compact support and $\hat{\mathscr{C}}_0^{\alpha,\beta}$ the subset whose mixed partial derivatives up to order α in t and β in the components of x, are continuous. Let $b_i(\cdot, \cdot)$, $a_{ij}(\cdot, \cdot)$, $i, j \leqq r$, be continuous functions on $R^r \times [0, \infty)$. Let the operator

$$A = \sum_i b_i(x, t)\frac{\partial}{\partial x_i} + \frac{1}{2}\sum_{i,j} a_{ij}(x, t)\frac{\partial^2}{\partial x_i \partial x_j},$$

be the infinitesimal operator of a diffusion process $X(\cdot)$. Assume that the solution to the martingale problem (on $D^r[0, \infty)$) of Stroock and Varadhan [8], corresponding to A, has *a unique nonexplosive solution for each initial condition.*

Let $b_N(\cdot)$ denote a function with values in $[0, 1]$, equal to 1 on $S_N = \{x: |x| \leqq N\}$, equal to zero in $R^r - S_{N+1}$ and with second derivatives bounded uniformly in x and N. Define $\{Y_n^{\varepsilon,N}, n \geqq 0\}$ by

$$Y_{n+1}^{\varepsilon,N} = Y_n^{\varepsilon,N} + [\varepsilon h_\varepsilon(Y_n^{\varepsilon,N}, \xi_n^\varepsilon) + \sqrt{\varepsilon} g_\varepsilon(Y_n^{\varepsilon,N}, \xi_n^\varepsilon) + o(\varepsilon)]b_N(Y_n^{\varepsilon,N}),$$

$$Y_0^{\varepsilon,N} = \begin{cases} Y_0^\varepsilon & \text{if } |Y_0^\varepsilon| \leqq N, \\ 0 & \text{otherwise,} \end{cases} \tag{2.1}$$

and define $Y^{\varepsilon,N}(\cdot)$ analogously to $Y^\varepsilon(\cdot)$. For purely technical reasons, it is convenient to state the theorem in terms of $\{Y_n^{\varepsilon,N}\}$. Let A^N be the infinitesimal operator of a (not necessarily unique) diffusion process, denoted by $X^N(\cdot)$, and suppose that its coefficients $a_N(\cdot, \cdot)$, $b_N(\cdot, \cdot)$ are continuous, bounded, have compact support and equal $a(\cdot, \cdot)$, $b(\cdot, \cdot)$ in S_N. Suppose that $\{Y^{\varepsilon,N}(\cdot)\}$ converges weakly to some such $X^N(\cdot)$, as $\varepsilon \to 0$ for each N. Then [7] $\{Y^\varepsilon(\cdot)\}$ converges weakly to $X(\cdot)$ as $n \to \infty$. The following theorem is a restatement of [7, Thm. 3], with $\tau_\varepsilon = \varepsilon$. [7, Thm. 2] provides a very convenient method of proving tightness, and we will use it in the sequel. Let $E_n^{\varepsilon,N}$ denote expectation conditioned on $\{Y_j^{\varepsilon,N}, j \leqq n, \xi_j^\varepsilon, j < n\}$.

THEOREM 1. *Assume the conditions stated above on the solution to the martingale problem on $D^r[0, \infty)$ corresponding to operator A, and on A^N and $X^N(\cdot)$. For each N and $f(\cdot, \cdot) \in \mathscr{D}$, a dense set* (sup *norm*) *in $\hat{\mathscr{C}}_0$, let there be a sequence $\{f^{\varepsilon,N}(\cdot)\}$ satisfying the following conditions: it is constant on each interval $[n\varepsilon, n\varepsilon + \varepsilon)$; at $n\varepsilon$ it is measurable with respect to the σ-algebra induced by $\{Y_j^{\varepsilon,N}, j \leqq n, \xi_j^\varepsilon, j < n\}$ and*

$$\sup_{n,\varepsilon} E|f^{\varepsilon,N}(n\varepsilon)| + \sup_{n,\varepsilon} \frac{1}{\varepsilon} E|E_n^{\varepsilon,N} f^{\varepsilon,N}(n\varepsilon + \varepsilon) - f^{\varepsilon,N}(n\varepsilon)| < \infty; \tag{2.2}$$

and as $\varepsilon \to 0$ and for each t as $n\varepsilon \to t$,

$$E|f^{\varepsilon,N}(n\varepsilon) - f(Y_n^{\varepsilon,N}, n\varepsilon)| \to 0, \tag{2.3}$$

$$E\left|\frac{E_n^{\varepsilon,N} f^{\varepsilon,N}(n\varepsilon + \varepsilon) - f^{\varepsilon,N}(n\varepsilon)}{\varepsilon} - \left(\frac{\partial}{\partial t} + A^N\right) f(Y_n^{\varepsilon,N}, n\varepsilon)\right| \to 0. \tag{2.4}$$

Then, if $\{Y^{\varepsilon,N}(\cdot), \varepsilon_0 > \varepsilon > 0\}$ is tight in $D^r[0, \infty)$ for each N, where ε_0 does not depend

on N and $Y^\varepsilon(0)$ converges weakly to $X(0)$, $\{Y^\varepsilon(\cdot)\}$ converges weakly to $X(\cdot)$, the unique solution to the martingale problem with initial condition $X(0)$.

3. An automata problem—introduction. Narendra [2], [3] and others have studied the application of automata and learning theory to problems in the routing of telephone calls through a multinode network and have suggested a variety of interesting automata models for this application. Under various assumptions (both explicit and implicit) they have stated convergence results in a number of cases. Generally, their results are applications of Norman's [9] results on slow learning. Here, we take one of their models and show how to apply Theorem 1 to get a much more complete asymptotic theory (large time) for small rate of change of the automata behavior (ε), under more realistic conditions. The case dealt with here can readily be generalized, as will be commented on below. The example illustrates the power and usefulness of the approximation techniques used here. The algorithm should be considered as a prototype. It might not be the best, but it well serves to illustrate the method.

The problem formulation. Calls arrive at a transmitting or switching terminal, at random, at discrete time instants $n = 0, 1, 2, \cdots$, with $P\{\text{one call arrives at } n\text{th instant}\} = \mu$, $\mu \in (0, 1)$, $P\{>1 \text{ call arrives at } n\text{th instant}\} = 0$. From the terminal, there are two possible routings to the destination, route 1 and route 2; the ith route has N_i independent lines and can thus handle up to N_i calls simultaneously. Let $[n, n+1)$ denote the nth interval of time. The duration of each call is a random variable with a geometric distribution: $P\{\text{call completed in the } (n+1)\text{st interval} \mid \text{uncompleted at end of } n\text{th interval, route } i \text{ used}\} = \lambda_i$, $\lambda_i \in (0, 1)$. The members of the double sequence of the interarrival times and call durations are mutually independent. It is possible to work with more general Markovian arrival processes, but we retain a simple structure in order to emphasize the main points. In practice, a more complex network would occur; and perhaps cycles might exist, and a vector routing parameter would be used, one component per node. But the main idea is similar. As in Theorem 4, the average dynamics are used for the stability analysis. From that point on, the proof of the appropriate generalization of Theorem 5 would be quite similar to the proof of Theorem 5.

The parameter ε will be used for the rate of adjustment of the routing automaton, the device which selects the route. The adjustment mechanism will be defined later. The routing automaton operates as follows. For each fixed ε, let $\{y_n^\varepsilon\}$ denote a sequence of random variables with values in $[0, 1]$. In order to have an unambiguous sequencing of events, suppose that the calls terminating in the nth interval actually terminate at time $n + \frac{1}{2}$, and arrivals and route assignments are at the instants $0, 1, 2, \cdots$ precisely. *Thus the state at time $(n+1)^-$ does not include the calls just terminated or calls arriving at $(n+1)$.* Define the "route occupancy process" $X_n^\varepsilon = (X_n^{\varepsilon,1}, X_n^{\varepsilon,2})$, where $X_n^{\varepsilon,i}$ is the number of lines of route i occupied at time n^+. Thus, $X_n^{\varepsilon,i} \leqq N_i$. If a call arrives at instant $n+1$, the automaton "flips a coin," choosing route 1 with probability y_n^ε and route 2 with probability $(1 - y_n^\varepsilon)$. If all lines of the chosen route i are occupied at instant $(n+1)^-$, then the call is switched to route j $(j \neq i)$. If all lines of route j are also occupied at instant $(n+1)^-$, then the call is rejected, and disappears from the system.

In a more realistic situation, the network would have many nodes—not simply 2, and many possibilities of routing from node to node. The adjustment algorithm might be different, but the problem would be handled in exactly the same way. The object is to adjust the $\{y_n^\varepsilon\}$ sequentially (based on the system behavior) so that some desired behavior occurs. In order to be specific, we use the following "linear-reward" algorithm [3]. Let J_{in}^ε denote the indicator of the event {call arrives at $n+1$, is assigned first to

route i and is accepted by route i}. Let $0 < y_l < y_u < 1$. We use the algorithm (3.1), where $|_{y_l}^{y_u}$ denotes truncation at y_u or y_l, and $\alpha(y) = 1 - y$, $\beta(y) = -y$.

$$y_{n+1}^{\varepsilon} = [y_n^{\varepsilon} + \varepsilon\alpha(y_n^{\varepsilon})J_{1n}^{\varepsilon} + \varepsilon\beta(y_n^{\varepsilon})J_{2n}^{\varepsilon}]|_{y_l}^{y_u}. \tag{3.1}$$

There are $\alpha_\varepsilon(\cdot), \beta_\varepsilon(\cdot)$ such that $\alpha(\cdot) = \alpha_\varepsilon(\cdot)$ in $[y_l, y_u - \varepsilon]$ and $\beta(\cdot) = \beta_\varepsilon(\cdot)$ in $[y_l + \varepsilon, y_u]$ and

$$y_{n+1}^{\varepsilon} = y_n^{\varepsilon} + \varepsilon[\alpha_\varepsilon(y_n^{\varepsilon})J_{1n}^{\varepsilon} + \beta_\varepsilon(y_n^{\varepsilon})J_{2n}^{\varepsilon}]. \tag{3.2}$$

We will study the asymptotics of the behavior of a centered and normalized $\{y_n^{\varepsilon}\}$ for small ε.

Some definitions. If the choice probabilities y_n^{ε} are held fixed at some value y for all n, then the route choice mechanism still makes sense, although there is no learning. For *fixed* route choice probability $y \in (0, 1)$, let $\{X_n(y)\} = \{X_n^1(y), X_n^2(y), 0 \leqq n < \infty\}$ denote the corresponding route occupancy process, on the state space $Z = \{(i, j): i \leqq N_1, j \leqq N_2\}$ (whose points are supposed ordered in some fixed way). This Markov chain is a single ergodic class, and the probability transition matrix, denoted by $A'(y)$, has infinitely differentiable components and

$$P_{n+1}(y) = A(y)P_n(y), \quad \text{with } P_0(y) \text{ given}, \tag{3.3}$$

where $P_n(y) = \{P_n(\alpha \mid y), \alpha \in Z\}$ and $P_n(\alpha \mid y) = P\{X_n(y) = \alpha\}$.

The pair $\{X_n^{\varepsilon}, y_n^{\varepsilon}, n \geqq 0\}$ is a Markov process on $Z \times [y_l, y_u]$. Define the vector $P_n^{\varepsilon} = \{P_n^{\varepsilon}(\alpha), \alpha \in Z\}$, where $P_n^{\varepsilon}(\alpha) = P\{X_n^{\varepsilon} = \alpha \mid y_l^{\varepsilon}, l < n, X_0^{\varepsilon}\}$. Then

$$P_{n+1}^{\varepsilon} = A(y_n^{\varepsilon})P_n^{\varepsilon}. \tag{3.4}$$

Define $P^i(N_i \mid y) = \lim_{n\to\infty} P\{X_n^i(y) = N_i\}$. Finally, define the marginal transition probability

$$P^i(\alpha, j, k \mid y) = P\{X_j^i(y) = k \mid X_0(y) = \alpha\},$$

and let E_n^{ε} denote the expectation conditioned on $\{X_l^{\varepsilon}, y_l^{\varepsilon}, l \leqq n\}$.

A relationship of (3.1) *to a differential equation.* Define $\nu_i = (1 - \lambda_i)^{N_i}$. Note that

$$E_n^{\varepsilon}J_{1n}^{\varepsilon} = \mu y_n^{\varepsilon}[1 - \nu_1 I\{X_n^{\varepsilon,1} = N_1\}], \tag{3.5a}$$

$$E_n^{\varepsilon}J_{2n}^{\varepsilon} = \mu(1 - y_n^{\varepsilon})[1 - \nu_2 I\{X_n^{\varepsilon,2} = N_2\}]. \tag{3.5b}$$

For small ε, the behavior of $\{y_n^{\varepsilon}\}$ is approximated by $\{y(\varepsilon n)\}$, where $y(\cdot)$ is defined by (3.6), and $\hat{F}(y)$ is just $E[\alpha(y)J_{1n}^{\varepsilon} + \beta(y)J_{2n}^{\varepsilon}]$, with the stationary process $\{X_n(y)\}$ used in the definition of J_{1n}^{ε}.

$$\begin{aligned} \dot{y} &= \mu\alpha(y)y[1 - \nu_1 P^1(N_1 \mid y)] - \mu(1 - y)\beta(y)[1 - \nu_2 P^2(N_2 \mid y)] \\ &= \mu y(1 - y)[\nu_2 P^2(N_2 \mid y) - \nu_1 P^1(N_1 \mid y)] \equiv \hat{F}(y). \end{aligned} \tag{3.6}$$

As y increases, $P^1(N_1 \mid y)$ increases (and $P^2(N_2 \mid y)$ decreases) monotonically. Thus, there is a unique point $\bar{y} \in (0, 1)$ such that $\hat{F}(\bar{y}) = 0$. Also, $\hat{F}(y) > 0$ for $y < \bar{y}$ and $\hat{F}(y) < 0$ for $y > \bar{y}$. *We assume that* $\bar{y} \in (y_l, y_u)$ and that $\hat{F}_y(\bar{y}) \neq 0$. We actually will study the asymptotic properties of $U_n^{\varepsilon} \equiv (y_n^{\varepsilon} - \bar{y})/\sqrt{\varepsilon}$ for large n and small ε. In particular, let n_ε be a sequence of integers tending to ∞ as $\varepsilon \to 0$, and define the processes $\tilde{U}^{\varepsilon}(\cdot)$ by $\tilde{U}^{\varepsilon}(0) = U_{n_\varepsilon}^{\varepsilon}$ and $\tilde{U}^{\varepsilon}(t) = U_{n_\varepsilon + i}^{\varepsilon}$ on $[i\varepsilon, i\varepsilon + \varepsilon)$. We show weak convergence of $\tilde{U}^{\varepsilon}(\cdot)$ to the Gauss–Markov diffusion $u(\cdot)$, defined by (6.3). If $n_\varepsilon \to \infty$ fast enough as $\varepsilon \to 0$, then the limit $u(\cdot)$ is stationary. The general method can be applied to many other problems in learning, automata and systems theory.

4. Some preliminary results. In this section, we prove some auxiliary results concerning uniform convergence of $P_n(y)$ and its derivatives, to $P(y)$ and its derivatives.

THEOREM 2. *For each $y \in [y_l, y_u]$, let $A'(y)$ denote a Markov transition matrix (continuous in y) such that the corresponding Markov chain $\{X_n(y)\}$ is ergodic with invariant measure $P(y)$. Then $P(\cdot)$ is also continuous and there is a $\delta > 0$, such that the eigenvalues of $A(y)$, except for the single eigenvalue unity, are bounded in absolute value by $1-\delta$ for all $y \in [y_l, y_u]$. $P_n(y)$ converges to $P(y)$ uniformly (and at a geometric rate) in $y \in [y_l, y_u]$ and in $P_0(y)$.*

Proof. The last sentence follows from the penultimate sentence. The continuity of $P(\cdot)$ is a consequence of the uniqueness for each y, of the eigenvector of $A(y)$ corresponding to the eigenvalue unity (the invariant measure). *Next, suppose that there is no such δ.* Let $A(y)$ be a $q \times q$ matrix and let $\lambda_1(y), \cdots, \lambda_q(y)$ denote the eigenvalues. Order them such that $\lambda_1(y) \equiv 1$. Then there is a $\tilde{y}$ and a sequence $\{y_n\} \subset [y_l, y_u]$ such that as $y_n \to \tilde{y}$, at least one eigenvalue (other than the one which is always unity) approaches the unit circle. In particular, suppose that the ordering is such that $|\lambda_2(y_n)| \to 1$ and that (choosing a subsequence if necessary) the $\lambda_i(y_n)$ converge to some $\tilde{\lambda}_i$ as $n \to \infty$ for $i = 1, \cdots, q$. The $\{\tilde{\lambda}_i\}$ must be the eigenvalues of $A(\tilde{y})$. But then $A'(y)$ is not the transition matrix of an ergodic process, a contradiction. □

DEFINITION. Let $\Sigma(y)$ denote the span of the eigenvectors and generalized eigenvectors of $A(y)$, except for the eigenvector which corresponds to the eigenvalue unity.

THEOREM 3. *Assume the situation of Theorem 1, but let $A(\cdot)$ be continuously differentiable on $[y_l, y_u]$ (at the endpoints, take the left- or right-hand derivatives, as appropriate); then so is $P(\cdot)$, and $P_y(y)$ is the unique solution in $\Sigma(y)$ to the equation*

$$P_y(y) = A(y)P_y + A_y(y)P(y). \tag{4.1}$$

Furthermore, the derivative $P_{n,y}(y)$ given by

$$P_{n+1,y}(y) = A(y)P_{n,y} + A_y(y)P_n(y), \tag{4.2}$$

converges geometrically to $P_y(y)$, uniformly in $y \in [y_l, y_u]$ and in the initial condition $P_0(y)$, if we set $P_{0,y}(y) = 0$.

If $A(\cdot)$ has continuous second derivatives on $[y_l, y_u]$, then so do $P(\cdot)$ and $P_n(\cdot)$, and $P_{yy}(y)$ is the unique solution in $\Sigma(y)$ to

$$P_{yy}(y) = A(y)P_{yy}(y) + 2A_y(y)P_y(y) + A_{yy}(y)P(y). \tag{4.3}$$

Also, $P_{n,yy}(y)$ converges geometrically to $P_{yy}(y)$, uniformly in $y \in [y_l, y_u]$ and in the initial conditions, if $P_{0,y}(y) = P_{0,yy}(y) = 0$.

Proof. Fix y. Since $(I - A(y))V = 0$ for $V \in \Sigma(y)$ implies that $V = 0$, in order for (4.1) to have a unique solution in $\Sigma(y)$ it is necessary and sufficient that $A_y(y)P(y) \perp \mathcal{N}(I - A'(y))$, where $\mathcal{N}$ denotes the null space of the matrix. $\mathcal{N}(I - A'(y))$ is the set of vectors Q such that $A'(y)Q = Q$. Since there is a unique eigenvalue of value unity and since the row *sums* of $A'(y)$ are all unity, the components of Q must all have the same value. Thus, the necessary and sufficient condition reduces to $A_y(y)P(y) \perp$ constant vectors. For any constant vector, $C = (c, c, \cdots)'$, $C'A(y) = C'$. Thus, $C'A_y(y) = 0$, and hence $A_y(y)D \perp$ constant vectors for *any* vector D. Consequently (4.1) has a unique solution $\bar{P}_y(y)$ in $\Sigma(y)$.

Next, we show that $\bar{P}_y(y)$ is the desired derivative. Write (for $y \in (y_l, y_u)$, otherwise $\delta > 0$ or $\delta < 0$, as appropriate)

$$A(y+\delta)P(y+\delta) - A(y)P(y) = P(y+\delta) - P(y).$$

Thus,

$$\frac{[A(y+\delta)-A(y)]}{\delta}P(y+\delta)=(I-A(y))\frac{[P(y+\delta)-P(y)]}{\delta}. \tag{4.4}$$

The left-hand side of (4.4) is uniformly bounded and is in $\Sigma(y)$ for each $\delta>0$ (since $(I-A(y))V\in\Sigma(y)$ for any V), and it converges to $A_y(y)P(y)$ as $\delta\to 0$. When considered as an operator from $\Sigma(y)$ to $\Sigma(y)$, $[I-A(y)]$ has a bounded inverse. Thus, as $\delta\to 0$, $[P(y+\delta)-P(y)]/\delta$ converges to $P_y(y)$, which must equal $\bar{P}_y(y)$, by the uniqueness proved above.

We now turn to the convergence (4.2). By Theorem 1, $P_n(y)$ converges geometrically to $P(y)$ uniformly in $[y_l, y_u]$ and in $P_0(y)$. Also, since we use $P_{0,y}(y)=0$,

$$P_{n+1,y}(y)=\sum_{j=0}^{n}A^{n-i}(y)A_y(y)P_i(y).$$

But $A_y(y)P_i(y)$ is a bounded sequence in $\Sigma(y)$, and as $i\to\infty$ it converges geometrically and uniformly to $A_y(y)P(y)$. Also $A(y)$ is a contraction, uniformly in $y\in[y_l, y_u]$, when acting in $\Sigma(y)$. These facts imply the desired convergence of $P_{n,y}(y)$. The limit must be a solution to (4.1).

The assertions concerning P_{yy} are proved in the same way and we omit the details. □

5. Tightness of $\{U_n^\varepsilon$, small ε, large $n\}$. By "ε small" and "n large" we mean that there are $\varepsilon_0>0$, $N_\varepsilon<\infty$, such that the assertion holds for $\varepsilon\leqq\varepsilon_0$, $n\leqq N_\varepsilon$. The actual value of ε_0 will be unimportant. Basic to the proof of weak convergence of $\{\tilde{U}^\varepsilon(\cdot)\}$ is the tightness of $\{U_n^\varepsilon$, small ε, large $n\}$.

THEOREM 4. *For each small $\varepsilon>0$, there is an $N_\varepsilon<\infty$ such that the doubly indexed sequence $\{U_n^\varepsilon$, ε small, $n\geqq N_\varepsilon\}$ is tight, where $U_n^\varepsilon=(y_n^\varepsilon-\bar{y})/\sqrt{\varepsilon}$.*

Proof. The proof uses an "averaged" Lyapunov function. Define $V(y)=(y-\bar{y})^2$. We have

$$\begin{aligned}E_n^\varepsilon(y_{n+1}^\varepsilon-y_n^\varepsilon)=\mu\varepsilon[&\alpha_\varepsilon(y_n^\varepsilon)y_n^\varepsilon(1-\nu_1 I\{X_n^{\varepsilon,1}=N_1\})\\&+\beta_\varepsilon(y_n^\varepsilon)(1-y_n^\varepsilon)(1-\nu_2 I\{X_n^{\varepsilon,2}=N_2\})],\end{aligned} \tag{5.1}$$

For small ε,

$$E_n^\varepsilon(y_n^\varepsilon-\bar{y})[\alpha_\varepsilon(y_n^\varepsilon)J_{1,n}^\varepsilon+\beta_\varepsilon(y_n^\varepsilon)J_{2,n}^\varepsilon]\leqq E_n^\varepsilon(y_n^\varepsilon-\bar{y})[\alpha(y_n^\varepsilon)J_{1,n}^\varepsilon+\beta(y_n^\varepsilon)J_{2,n}^\varepsilon],$$

since $0\leqq\alpha_\varepsilon(y)\leqq\alpha(y)$ and $\alpha_\varepsilon(y)\neq\alpha(y)$ only if $y_n^\varepsilon-\bar{y}\geqq 0$ (for small ε), and conversely for the β_ε term. Using the above inequality, (5.1a) and $|y_{n+1}^\varepsilon-y_n^\varepsilon|=O(\varepsilon)$,

$$\begin{aligned}&E_n^\varepsilon V(y_{n+1}^\varepsilon)-V(y_n^\varepsilon)\\&\leqq 2\mu\varepsilon(y_n^\varepsilon-\bar{y})[\alpha(y_n^\varepsilon)y_n^\varepsilon(1-\nu_1 I\{X_n^{\varepsilon,1}=N_1\})\\&\qquad+\beta(y_n^\varepsilon)(1-y_n^\varepsilon)(1-\nu_2 I\{X_n^{\varepsilon,2}=N_2\})]+O(\varepsilon^2).\end{aligned} \tag{5.2}$$

Define the "perturbation" $V_1^\varepsilon(n)$ to the Lyapunov function $V(n)=V(y_n^\varepsilon)$ by

$$\begin{aligned}V_1^\varepsilon(n)=&2\mu\varepsilon(y_n^\varepsilon-\bar{y})\alpha(y_n^\varepsilon)y_n^\varepsilon\nu_1\sum_{j=n}^{\infty}[P^1(N_1|y_n^\varepsilon)-P^1(X_n^\varepsilon,j-n,N_1|y_n^\varepsilon)]\\&+2\mu\varepsilon(y_n^\varepsilon-\bar{y})\beta(y_n^\varepsilon)(1-y_n^\varepsilon)\nu_2\sum_{j=n}^{\infty}[P^2(N_2|y_n^\varepsilon)-P^2(X_n^\varepsilon,j-n,N_2|y_n^\varepsilon)].\end{aligned} \tag{5.3}$$

Note that $P^i(X_n^\varepsilon,0,N_i|y_n^\varepsilon)=I\{X_n^{\varepsilon,i}=N_i\}$. By Theorem 2, the sums converge absolutely

(the summands go to zero at a geometric rate) uniformly in n, y_n^ε, X_n^ε. Thus $|V_1^\varepsilon(\cdot)| = O(\varepsilon)$, uniformly in all the variables.

Next, evaluate

$$
\begin{aligned}
&E_n^\varepsilon V_1^\varepsilon(n+1) - V_1^\varepsilon(n) \\
&\quad = -2\mu\varepsilon(y_n^\varepsilon - \bar{y})\alpha(y_n^\varepsilon)y_n^\varepsilon\nu_1[P^1(N_1|y_n^\varepsilon) - I\{X_n^{\varepsilon,1} = N_1\}] \\
&\qquad - 2\mu\varepsilon(y_n^\varepsilon - \bar{y})\beta(y_n^\varepsilon)(1-y_n^\varepsilon)\nu_2[P^2(N_2|y_n^\varepsilon) - I\{X_n^{\varepsilon,2} = N_2\}] \\
&\qquad + \sum_{j=n+1}^{\infty} 2\mu\varepsilon\nu_1\{E_n^\varepsilon(y_{n+1}^\varepsilon - \bar{y})\alpha(y_{n+1}^\varepsilon)y_{n+1}^\varepsilon \\
&\qquad\qquad \cdot [P^1(N_1|y_{n+1}^\varepsilon) - P^1(X_{n+1}^\varepsilon, j-n-1, N_1|y_{n+1}^\varepsilon)] \\
&\qquad\qquad - (y_n^\varepsilon - \bar{y})\alpha(y_n^\varepsilon)y_n[P^1(N_1|y_n^\varepsilon) - P^1(X_n^\varepsilon, j-n, N_1|y_n^\varepsilon)]\} \\
&\qquad + \text{a similar sum for route 2.}
\end{aligned} \tag{5.4}
$$

We next show that the sums in (5.4) $= O(\varepsilon^2)$ uniformly in all the variables n, y_n^ε, X_n^ε. Using $|y_{n+1}^\varepsilon - y_n^\varepsilon| = O(\varepsilon)$, the smoothness of $\alpha(\cdot)$ and $\beta(\cdot)$, Theorem 2,

$$E_n^\varepsilon P^i(X_{n+1}^\varepsilon, j-n-1, N_i|y_n^\varepsilon) = P^i(X_n^\varepsilon, j-n, N_i|y_n^\varepsilon),$$

(the Markov property for $\{X_j(y), j \geqq n\}$ with $y = y_n^\varepsilon$ and initial condition $X_n(y_n^\varepsilon) = X_n^\varepsilon$), we can rewrite the sums as

$$
\begin{aligned}
&O(\varepsilon)\sum_{j=n+1}^{\infty} E_n^\varepsilon\{[P^i(N_i|y_{n+1}^\varepsilon) - P^i(N_i|y_n^\varepsilon)] \\
&\qquad\qquad - [P^i(X_{n+1}^\varepsilon, j-n-1, N_i|y_{n+1}^\varepsilon) - P^i(X_{n+1}^\varepsilon, j-n-1, N_i|y_n^\varepsilon)]\} \\
&\qquad\qquad + O(\varepsilon^2).
\end{aligned} \tag{5.5}
$$

Now, the smoothness of the $P^i(N_i|\cdot)$, $P^i(X, j, N_i|\cdot)$ and Theorem 3 imply that (5.5) $= O(\varepsilon^2)$.

Define $V^\varepsilon(n) = V(y_n^\varepsilon) + V_1^\varepsilon(n)$. By (5.2) and (5.4) and the fact that the sums in (5.4) are $O(\varepsilon^2)$,

$$
\begin{aligned}
E_n^\varepsilon V^\varepsilon(n+1) - V^\varepsilon(n) \leqq O(\varepsilon^2) &+ 2\mu\varepsilon(y_n^\varepsilon - \bar{y})[\alpha(y_n^\varepsilon)y_n^\varepsilon(1-\nu_1 P^1(N_1|y_n^\varepsilon)) \\
&+ \beta(y_n^\varepsilon)(1-y_n^\varepsilon)(1-\nu_2 P^2(N_2|y_n^\varepsilon))].
\end{aligned}
$$

Owing to the fact that $y_n^\varepsilon \in [y_l, y_u]$, the bracketed term has its unique zero at $y_n^\varepsilon = \bar{y}$ and it is positive (negative, resp.) for $y_n^\varepsilon < \bar{y}$ ($y_n^\varepsilon > \bar{y}$, resp.). Thus, there is a $\gamma > 0$ such that

$$E_n^\varepsilon V^\varepsilon(n+1) - V^\varepsilon(n) \leqq O(\varepsilon^2) - \varepsilon\gamma V(y_n^\varepsilon). \tag{5.6}$$

By $|V_1^\varepsilon(n)| = O(\varepsilon)$ uniformly in n, $E_n^\varepsilon V^\varepsilon(n+1) - V^\varepsilon(n) \leqq O(\varepsilon^2) - \varepsilon\gamma V^\varepsilon(n)$, and hence

$$EV(n) \leqq (\exp -\varepsilon\gamma n)EV(0) + O(\varepsilon), \tag{5.7}$$

which implies the tightness. Finally, let $0 < K_0$ be arbitrary and let N_ε be the smallest integer n such that $(\exp -\varepsilon n\gamma) \leqq K_0\varepsilon$. □

6. Weak convergence of $\{\tilde{U}^\varepsilon(\cdot)\}$.

DEFINITIONS. Recall the definition of N_ε given at the end of the proof of Theorem 4. For any sequence of integers $n_\varepsilon > N_\varepsilon$, define $Q_\varepsilon = n_\varepsilon - N_\varepsilon$. Define $\tilde{y}_n^\varepsilon = y_{n_\varepsilon+n}^\varepsilon$ and

similarly define the "shifted" sequences $\tilde{U}_n^\varepsilon$, $\tilde{X}_n^\varepsilon$ and $\tilde{J}_{in}^\varepsilon$. Then

$$\tilde{U}_{n+1}^\varepsilon = \tilde{U}_n^\varepsilon + \sqrt{\varepsilon}\,[\alpha_\varepsilon(\tilde{y}_n^\varepsilon)\tilde{J}_{1n}^\varepsilon + \beta_\varepsilon(y_n^\varepsilon)\tilde{J}_{2n}^\varepsilon]. \tag{6.1}$$

By Theorem 4, $\{\tilde{U}_n^\varepsilon, \varepsilon \text{ small}\}$ is tight. To use Theorem 1, we must truncate $\{\tilde{U}_n^\varepsilon\}$. For each integer N, define $\tilde{U}_n^{\varepsilon,N}$, $\tilde{y}_n^{\varepsilon,N}$, $\tilde{J}_{in}^{\varepsilon,N}$ via

$$\tilde{U}_{n+1}^{\varepsilon,N} = \tilde{U}_n^{\varepsilon,N} + \sqrt{\varepsilon}\,[\alpha(\tilde{y}_n^{\varepsilon,N})\tilde{J}_{1n}^{\varepsilon,N} + \beta(\tilde{y}_n^{\varepsilon,N})\tilde{J}_{2n}^{\varepsilon,N}]b_N(\tilde{U}_n^{\varepsilon,N}), \tag{6.2}$$

and let $\tilde{U}_n^{\varepsilon,N} = (\tilde{y}_n^{\varepsilon,N} - \bar{y})/\sqrt{\varepsilon}$, *define* $\tilde{y}_n^{\varepsilon,N}$. $\tilde{J}_{in}^{\varepsilon,N}$ *is simply the indicator function of the set* {route i is tried first and the call accepted}, when the sequence of choice probabilities is $\{\tilde{y}_n^{\varepsilon,N}\}$. Since $|\tilde{y}_n^{\varepsilon,N} - \bar{y}| \leqq \sqrt{\varepsilon}(N+1)$ for small ε, it is irrelevant whether we use α_ε, β_ε or α, β in (6.2), and we use α, β. To simplify the notation, we drop the superscript N until Part 4 of the proof of Theorem 5.

We now define an auxiliary process which is used in the averaging method employed in the proof. Let $\bar{P}$ denote the measure defined by the stationary process $\{X_j(\bar{y}), \infty > j > -\infty\}$, with corresponding expectation operator $\bar{E}$. For each n, it is necessary to introduce the process $\{X_j(\bar{y}), j \geqq n\}$, but with "initial" condition $X_n(\bar{y}) = \tilde{X}_n^\varepsilon$ (i.e., after time n, the route choice probability is $\bar{y}$). The operator $\bar{E}_n^\varepsilon$ denotes the expectation of functions of this process $\{X_j(\bar{y}), j \geqq n\}$ conditional on the "initial" condition $X_n(\bar{y}) = \tilde{X}_n^\varepsilon$. Let $J_{ij}(\bar{y})$ denote the indicator function $I\{$call arives at $j+1$, is assigned to and accepted by route $i\}$, when the route choice variable is $\bar{y}$ and the route occupancy process is $\{X_j(\bar{y})\}$. Whether we intend the ergodic process or the process $\{X_j(\bar{y}), j \geqq n\}$ starting at time n with $X_n(\bar{y}) = \tilde{X}_n^\varepsilon$, will be made clear by use of either $\bar{E}$ or $\bar{E}_n^\varepsilon$. Define

$$\delta u_j(\bar{y}) = [\alpha(\bar{y})J_{1j}(\bar{y}) + \beta(\bar{y})J_{2j}(\bar{y})].$$

Under $\bar{P}$, the right side has zero expectation.

THEOREM 5. *For any sequence $n_\varepsilon \geqq N_\varepsilon$, $\{\tilde{U}^\varepsilon(\cdot)\}$ is tight in $D[0,\infty)$. All weakly convergent subsequences converge to a Gauss–Markov diffusion satisfying* (6.3). *If $\varepsilon Q_\varepsilon \to \infty$ as $\varepsilon \to \infty$, then the limiting diffusion $u(\cdot)$ is stationary in that $u(0)$ has the stationary distribution (in all cases $u(0)$ is independent of $B(\cdot)$), and*

$$du = Gu\,dt + \sigma\,dB, \qquad B(\cdot) = \text{standard Brownian motion}, \tag{6.3}$$

$$G = \hat{F}_y(\bar{y}) = \frac{\partial}{\partial y}\mu y(1-y)[\nu_2 P^2(N_2 \mid y) - \nu_1 P^1(N_1 \mid y)]\big|_{y=\bar{y}}, \tag{6.4}$$

$$\sigma^2 = \bar{E}(\delta u_0(\bar{y}))^2 + 2\sum_{n=1}^{\infty} \bar{E}\delta u_0(\bar{y})\delta u_n(\bar{y}). \tag{6.5}$$

Proof. Part 1. By (5.1),

$$\tilde{E}_n^\varepsilon(\tilde{U}_{n+1}^\varepsilon - \tilde{U}_n^\varepsilon) = \sqrt{\varepsilon}\mu\tilde{y}_n^\varepsilon(1-\tilde{y}_n^\varepsilon)[\nu_2 I\{\tilde{X}_n^{\varepsilon,2} = N_2\} - \nu_1 I\{\tilde{X}_n^{\varepsilon,1} = N_1\}]b_N(\tilde{U}_n^\varepsilon). \tag{6.6}$$

Let $f(\cdot,\cdot) \in \mathscr{D} = \mathscr{C}_0^{2,3}$, the space of bounded (x,t) functions with compact support, whose mixed partial derivatives, up to order 2 in t and 3 in x, are continuous. If we apply Theorem 1 to $\{\tilde{U}^\varepsilon(\cdot)\}$, we will get an $f^\varepsilon(\cdot)$ of the form

$$f^\varepsilon(n\varepsilon) = f(\tilde{U}_n^\varepsilon, n\varepsilon) + f_0^\varepsilon(n\varepsilon) + f_1^\varepsilon(n\varepsilon) + f_2^\varepsilon(n\varepsilon),$$

where the $f_i^\varepsilon(n\varepsilon)$ will be defined in the sequel. For each N, all $o(\cdot)$ or $O(\cdot)$ are uniform

in all variables except their argument. We have

$$\tilde{E}_n^\varepsilon f(\tilde{U}_{n+1}^\varepsilon, n\varepsilon+\varepsilon)-f(\tilde{U}_n^\varepsilon, n\varepsilon)=\tilde{E}_n^\varepsilon[f(\tilde{U}_{n+1}^\varepsilon, n\varepsilon)-f(\tilde{U}_n^\varepsilon, n\varepsilon)]+f_t(\tilde{U}_n^\varepsilon, n\varepsilon)\varepsilon+o(\varepsilon),$$

$$\begin{aligned}
&\tilde{E}_n^\varepsilon[f(\tilde{U}_{n+1}^\varepsilon, n\varepsilon)-f(\tilde{U}_n^\varepsilon, n\varepsilon)]\\
&\quad=\tilde{E}_n^\varepsilon f_u(\tilde{U}_n^\varepsilon, n\varepsilon)(\tilde{U}_{n+1}^\varepsilon-\tilde{U}_n^\varepsilon)+\tfrac{1}{2}\tilde{E}_n^\varepsilon f_{uu}(\tilde{U}_n^\varepsilon, n\varepsilon)(\tilde{U}_{n+1}^\varepsilon-\tilde{U}_n^\varepsilon)^2+o(\varepsilon)\\
&\quad=\sqrt{\varepsilon}\mu f_u(\tilde{U}_n^\varepsilon, n\varepsilon)\tilde{y}_n^\varepsilon(1-\tilde{y}_n^\varepsilon)b_N(\tilde{U}_n^\varepsilon)[\nu_2 I\{\tilde{X}_n^{\varepsilon,2}=N_2\}-\nu_1 I\{\tilde{X}_n^{\varepsilon,1}=N_1\}]\\
&\qquad+\frac{f_{uu}}{2}(\tilde{U}_n^\varepsilon, n\varepsilon)\tilde{E}_n^\varepsilon(\tilde{U}_{n+1}^\varepsilon-\tilde{U}_n^\varepsilon)^2+o(\varepsilon).
\end{aligned}\tag{6.7}$$

By the differentiability result of Theorem 3, we can rewrite the term before the $o(\varepsilon)$ as follows:[1]

$$\begin{aligned}
&\varepsilon b_N^2(\tilde{U}_n^\varepsilon)\frac{f_{uu}(\tilde{U}_n^\varepsilon, n\varepsilon)}{2}\tilde{E}_n^\varepsilon[\alpha(\tilde{y}_n^\varepsilon)J_{1n}^\varepsilon+\beta(\tilde{y}_n^\varepsilon)\tilde{J}_{2n}^\varepsilon]^2\\
&\qquad=\varepsilon b_N^2(\tilde{U}_n^\varepsilon)\frac{f_{uu}(\tilde{U}_n^\varepsilon, n\varepsilon)}{2}\bar{E}_n^\varepsilon[\alpha(\bar{y})J_{1n}(\bar{y})+\beta(\bar{y})J_{2n}(\bar{y})]^2+o(\varepsilon).
\end{aligned}$$

Part 2. We will "average out" the terms in (6.7) one by one. Define $f_1^\varepsilon(n\varepsilon)$ (analogous to the definition of $V_1^\varepsilon(n)$ in the last section)

$$\begin{aligned}
f_1^\varepsilon(n\varepsilon)&=\sqrt{\varepsilon}\mu b_N(\tilde{U}_n^\varepsilon)\tilde{y}_n^\varepsilon(1-\tilde{y}_n^\varepsilon)f_u(\tilde{U}_n^\varepsilon, n\varepsilon)\\
&\quad\cdot\sum_{j=n}^{\infty}[\nu_2(P^2(\tilde{X}_n^\varepsilon, j-n, N_2\,|\,\tilde{y}_n^\varepsilon)-P^2(N_2\,|\,\tilde{y}_n^\varepsilon))\\
&\qquad-\nu_1(P^1(\tilde{X}_n^\varepsilon, j-n, N_1\,|\,\tilde{y}_n^\varepsilon)-P^1(N_1\,|\,\tilde{y}_n^\varepsilon))].
\end{aligned}\tag{6.8}$$

Using expressions similar to those used in Theorem 4 in treating $V_1^\varepsilon(n)$ and writing $P^i(\tilde{X}_n^\varepsilon, j-n, N_i|\tilde{y}_n^\varepsilon)$ in the more convenient form $\tilde{E}_n^\varepsilon P^i(\tilde{X}_{n+1}^\varepsilon, j-n-1, N_i|\tilde{y}_n^\varepsilon)$ for $j>n$ (see above (5.5)), we can get (details are in [10])

$$\tilde{E}_n^\varepsilon f_1^\varepsilon(n\varepsilon+\varepsilon)-f_1^\varepsilon(n\varepsilon)=S_1+S_2+S_3+o(\varepsilon),$$

$$S_1=-\sqrt{\varepsilon}\mu\tilde{y}_n^\varepsilon(1-\tilde{y}_n^\varepsilon)b_N(\tilde{U}_n^\varepsilon)f_u(\tilde{U}_n^\varepsilon, n\varepsilon)[\nu_2 I\{\tilde{X}_n^{\varepsilon,2}=N_2\}-\nu_1 I\{\tilde{X}_n^{\varepsilon,1}=N_1\}],\tag{6.9}$$

$$\begin{aligned}
S_2&=\sqrt{\varepsilon}\mu\tilde{y}_n^\varepsilon(1-\tilde{y}_n^\varepsilon)b_N(\tilde{U}_n^\varepsilon)f_u(\tilde{U}_n^\varepsilon, n\varepsilon)[\nu_2 P^2(N_2\,|\,\tilde{y}_n^\varepsilon)-\nu_1 P^i(N_1\,|\,\tilde{y}_n^\varepsilon)]\\
&=\varepsilon\mu b_N(\tilde{U}_n^\varepsilon)\frac{\partial}{\partial y}\{y(1-y)[\nu_2 P^2(N_2\,|\,y)-\nu_1 P^1(N_1\,|\,y)]\}_{y=\bar{y}}+o(\varepsilon),
\end{aligned}\tag{6.10}$$

$$\begin{aligned}
S_3&=o(\varepsilon)+\sqrt{\varepsilon}\mu\tilde{y}_n^\varepsilon(1-\tilde{y}_n^\varepsilon)(b_N(\tilde{U}_n^\varepsilon)f_u(\tilde{U}_n^\varepsilon, n\varepsilon))_u\\
&\quad\tilde{E}_n^\varepsilon(\tilde{U}_{n+1}^\varepsilon-\tilde{U}_n^\varepsilon)\sum_{j=n+1}^{\infty}[\nu_2(P^2(\tilde{X}_{n+1}^\varepsilon, j-n-1, N_2\,|\,\tilde{y}_n^\varepsilon)-P^2(N_2\,|\,\tilde{y}_n^\varepsilon))\\
&\qquad-\nu_1(P^1(\tilde{X}_{n+1}^\varepsilon, j-n-1, N_1\,|\,\tilde{y}_n^\varepsilon)-P^1(N_1\,|\,\tilde{y}_n^\varepsilon))]\\
&=o(\varepsilon)+\varepsilon b_N(\tilde{U}_n^\varepsilon)[b_N(\tilde{U}_n^\varepsilon)f_u(\tilde{U}_n^\varepsilon, n\varepsilon)]_u\sum_{j=n+1}^{\infty}\bar{E}_n^\varepsilon\delta u_n(\bar{y})\delta u_j(\bar{y}).
\end{aligned}\tag{6.11}$$

[1] The terms $\bar{E}_n^\varepsilon J_{1n}(\bar{y})$ and $\tilde{E}_n^\varepsilon J_{1n}^\varepsilon(\tilde{y}_n^\varepsilon)$ differ only in that in the first case $\bar{y}$ is used as the choice variable to get the successor state to $\tilde{X}_n^\varepsilon$, and $\tilde{y}_n^\varepsilon$ is used in the second case.

Part 3. Now, we "average out" the last sum in (6.11). Define $f_2^\varepsilon(n\varepsilon)$ by

$$f_2^\varepsilon(n\varepsilon) = \varepsilon b_N(\tilde{U}_n^\varepsilon)[b_N(\tilde{U}_n^\varepsilon)f_u(\tilde{U}_n^\varepsilon, n\varepsilon)]_u \sum_{j=n}^{\infty}\sum_{k=j+1}^{\infty}[\bar{E}_n^\varepsilon \delta u_j(\bar{y})\delta u_k(\bar{y}) - \bar{E}\delta u_j(\bar{y})\delta u_k(\bar{y})].$$

By the (uniform) geometric convergence result of Theorem 2, the sum converges absolutely and $|f_2^\varepsilon(n\varepsilon)| = O(\varepsilon)$. By a straightforward calculation using the stationarity of $\{\delta u_n(\bar{y})\}$ under $\bar{P}$, we can show that

$$\tilde{E}_n^\varepsilon f_2^\varepsilon(n\varepsilon+\varepsilon) - f_2^\varepsilon(n\varepsilon) = -(6.12) + o(\varepsilon)$$
$$+ \varepsilon b_N(\tilde{U}_n^\varepsilon)[b_N(\tilde{U}_n^\varepsilon)f_u(\tilde{U}_n^\varepsilon, n\varepsilon)]_u \sum_{j=1}^{\infty}\bar{E}(\delta u_0(\bar{y})\,\delta u_j(\bar{y})).$$

Finally, we treat the term before the $o(\varepsilon)$ of (6.7)—in the form in which it is written below (6.7). Define $f_0^\varepsilon(n\varepsilon)$ by

$$f_0^\varepsilon(n\varepsilon) = \varepsilon\frac{f_{uu}(\tilde{U}_n^\varepsilon, n\varepsilon)}{2} b_N^2(\tilde{U}_n^\varepsilon)\sum_{j=n}^{\infty}[\bar{E}_n^\varepsilon(\delta u_j(\bar{y}))^2 - \bar{E}(\delta u_j(\bar{y}))^2].$$

By a procedure similar to that used [10] for $f_1^\varepsilon(n\varepsilon)$, it can readily be shown that

$$\tilde{E}_n^\varepsilon f_0^\varepsilon(n\varepsilon+\varepsilon) - f_0^\varepsilon(n\varepsilon) = o(\varepsilon) + \varepsilon\frac{f_{uu}(\tilde{U}_n^\varepsilon, n\varepsilon)}{2} b_N^2(\tilde{U}_n^\varepsilon)\bar{E}(\delta u_0(\bar{y}))^2$$
$$- \varepsilon\frac{f_{uu}(\tilde{U}_n^\varepsilon, n\varepsilon)}{2} b_N^2(\tilde{U}_n^\varepsilon)\bar{E}_n^\varepsilon[\alpha(\bar{y})J_{1n}(\bar{y}) + \beta(\bar{y})J_{2n}(\bar{y})]^2.$$

Summarizing the previous calculations, we have

$$\tilde{E}_n^\varepsilon f^\varepsilon(n\varepsilon+\varepsilon) - f^\varepsilon(n\varepsilon) = o(\varepsilon) + \varepsilon f_t(\tilde{U}_n^\varepsilon, n\varepsilon) + \varepsilon f_u(\tilde{U}_n^\varepsilon, n\varepsilon)G\tilde{U}_n^\varepsilon b_N(\tilde{U}_n^\varepsilon)$$
$$+ \varepsilon f_u(\tilde{U}_n^\varepsilon, n\varepsilon)b_{N,u}(\tilde{U}_n^\varepsilon)b_N(\tilde{U}_n^\varepsilon)\sum_{j=1}^{\infty}\bar{E}\,\delta u_0(\bar{y})\,\delta u_j(\bar{y}) \tag{6.12}$$
$$+ \varepsilon\frac{f_{uu}(\tilde{U}_n^\varepsilon, n\varepsilon)}{2} b_N^2(\tilde{U}_n^\varepsilon)\left[\bar{E}(\delta u_0(\bar{y}))^2 + 2\sum_{j=1}^{\infty}\bar{E}\,\delta u_0(\bar{y})\,\delta u_j(\bar{y})\right].$$

Part 4. *Conclusion.* Reintroduce the superscript N. Fix N. All the $f_i^{\varepsilon,N}$ are bounded and of order $O(\sqrt{\varepsilon})$ and $\{\tilde{U}_0^{\varepsilon,N}\} = \{\tilde{U}^{\varepsilon,N}(0)\}$ is tight. Also $\tilde{E}_n^{\varepsilon,N}f^{\varepsilon,N}(n\varepsilon+\varepsilon) - f^{\varepsilon,N}(n\varepsilon) = O(\varepsilon)$. Thus, by [7, Thm. 2] the bounded sequence $\{\tilde{U}^{\varepsilon,N}(\cdot)\}$ is tight in $D[0,\infty)$. Let ε index a weakly convergent subsequence with limit $U^N(\cdot)$. Since A is defined to be the infinitesimal operator of (6.3), by (6.14) and Theorem 1, we see that $U^N(\cdot)$ solves the martingale problem corresponding to an infinitesimal operator A^N whose coefficients equal those of A in S_N. Thus, by Theorem 1, $\{\tilde{U}^\varepsilon(\cdot)\}$ converges weakly to a solution $u(\cdot)$ of (6.3).The independence of $B(\cdot)$ and $u(0)$ is a consequence of the fact that $u(\cdot)$ is the unique solution to the martingale problem. The stationarity assertion is not hard to prove, but we omit the details. □

7. Asymptotic theory of an adaptive quantizer: introduction. In recent years there has been a great deal of research concerning the efficient quantization of signals in telecommunications systems, e.g., of voice signals in telephone transmission systems.

Let $z(\cdot)$ denote the actual signal process and Δ a sampling interval. In the problem of interest, the signal is sampled at moments $\{n\Delta, n = 0, 1, \cdots\}$, then the samples $\{z(n\Delta)\}$ are quantized, and it is only the quantized samples which are transmitted. Let $0 = \xi_0 < \xi_1 < \cdots \xi_{L-1} < \xi_L = \infty$, $0 = \eta_1 < \eta_2 < \cdots < \eta_L$, where ξ_i, η_{i+1}, $i = 0, \cdots, L-1$, are real numbers. Let the quantization function $Q(\cdot)$ be defined as follows: there is a $y > 0$ such that for $z(n\Delta) > 0$, $Q(z(n\Delta)) = y\eta_i$ if $z(n\Delta) \in [y\xi_{i-1}, y\xi_i)$, and set $Q(-z) = -Q(z)$. The parameter y is a scaling parameter. As the signal power increases (decreases), y should increase (decrease) for efficient reconstruction of the signal from the sequence of quantizations.

The problem of choosing appropriate values of y when the signal powers can vary by an order of magnitude or more has led to the study of adaptive quantizers. We give only a brief description in order to formulate the problem. For more detail and discussion of the engineering considerations, the reader is referred to the references [4], [5]. Let ε denote a "rate of adjustment" parameter for the scale parameter y and let y_n^ε denote the value of the adapted scale parameter at the nth sampling instant. Set $\beta \in (0, 1]$, and let $0 < M_1^\varepsilon < M_2^\varepsilon < \cdots < M_L^\varepsilon < \infty$ with $M_1^\varepsilon < 1$, $M_L^\varepsilon > 1$. We study an adaptive quantizer which is a truncated form of the (typical in such an application) adaptive system

$$y_{n+1}^\varepsilon = (y_n^\varepsilon)^\beta B_n^\varepsilon, \quad \text{where } B_n^\varepsilon = M_i^\varepsilon \text{ if } |z(n\Delta)| \in [y_n^\varepsilon \xi_{i-1}, y_n^\varepsilon \xi_i). \tag{7.1}$$

Goodman and Gersho [4] did a thorough analysis of (7.1) for the case $\beta = 1$ and $\{z(n\Delta)\}$ independent and identically distributed (i.i.d). With $\beta < 1$, the system has some desirable robustness properties and this case, together with simulations, is discussed by Mitra [5] and others. The latter reference is concerned more with reconstruction of the process $z(\cdot)$ from $\{Q(Z(n\Delta))\}$ and does not give an asymptotic analysis.

Generally, with non i.i.d. $\{z(n\Delta)\}$, it is hard to get concrete information on $\{y_n^\varepsilon\}$ for large n. If the signal power varies over time or if (as is realistic for moderate values of Δ) $\{x(n\Delta)\}$ is not i.i.d., then techniques such as used in [4] fail, but for small rates of adjustment (ε) an asymptotic analysis can still shed light on the process behavior. At the present time, it seems that little more can be done for the general case. Here, we scale the problem so that an asymptotic analysis is possible. For mathematical as well as practical purposes, it is useful to confine y_n^ε to some finite positive interval $[y_l, y_u]$. Now, we define the truncated form of (7.1) which will be studied. Let $\alpha > 0$, $0 < \alpha\varepsilon < 1$ and let $\{l_i\}$ be real numbers such that $l_1 < l_2 < \cdots < l_L$ and $l_1 < 0$, $l_L > 0$. Then we use

$$y_{n+1}^\varepsilon = (y_n^\varepsilon)^{1-\varepsilon\alpha} B_n^\varepsilon \Big|_{y_l}^{y_u}, \tag{7.2}$$

where $|$ denotes truncation and

$$B_n^\varepsilon = (1 + \varepsilon l_i) \quad \text{if } |z(n\Delta)| \in [y_n^\varepsilon \xi_{i-1}, y_n^\varepsilon \xi_i).$$

The asymptotic results can be used to get information on the effects of $\{l_i\}$, Δ, structure of $z(\cdot)$ and α, on the performance for small ε. For notational convenience below, let $y_l < 1$ and $y_u > 1$. Rewrite (7.2) in the form (7.3), where $y^{1-\varepsilon\alpha} = y[1 - \varepsilon\alpha \log y] + O(\varepsilon^2)$ and $(1 + \varepsilon b_n^\varepsilon) \equiv B_N^\varepsilon$ are used, and F and b_n^ε have the obvious definitions:

$$y_{n+1}^\varepsilon = [y_n^\varepsilon (1 + \varepsilon b_n^\varepsilon) - \varepsilon\alpha y_n^\varepsilon \log y_n^\varepsilon + O(\varepsilon^2)]\Big|_{y_l}^{y_u} \equiv [y_n^\varepsilon + \varepsilon F(y_n^\varepsilon, z(n\Delta)) + O(\varepsilon^2)]\Big|_{y_l}^{y_u}. \tag{7.3}$$

In [4], the process $\{\log y_n^\varepsilon\}$ rather than $\{y_n^\varepsilon\}$ is discussed.

We proceed in very much the same way that we did for the automata problem. The main difference arises from the unboundedness of $\{z(n\Delta)\}$, under assumption (7.6).

By definition,

$$b_n^\varepsilon = \sum_{i=1}^{L} l_i I\{|z(n\Delta)| \in [y_n^\varepsilon \xi_{i-1}, y_n^\varepsilon \xi_i)\}.$$

There are continuous functions $l_i^\varepsilon(\cdot)$ such that (7.4) and the properties stated below it hold.

$$\begin{aligned} y_{n+1}^\varepsilon &= y_n^\varepsilon(1+\varepsilon\beta_n^\varepsilon(y_n^\varepsilon)) - \varepsilon\alpha y_n^\varepsilon \log y_n^\varepsilon + O(\varepsilon^2) \\ &\equiv y_n^\varepsilon + \varepsilon F_\varepsilon(y_n^\varepsilon, z(n\Delta)) + O(\varepsilon^2), \end{aligned} \tag{7.4}$$

where

$$\beta_n^\varepsilon(y) = \sum_{i=1}^{L} l_i^\varepsilon(y) I\{|z(n\Delta)| \in [y\xi_{i-1}, y\xi_i)\}. \tag{7.5}$$

Also, $l_i^\varepsilon(\cdot)$ can be chosen such that $l_i^\varepsilon(\cdot) = l_i$ out of an $O(\varepsilon)$ neighborhood of y_l (resp. y_u) if $l_i < 0$ (resp. $l_i > 0$), and $0 \geqq l_i^\varepsilon(y) \geqq l_i$ for $l_i < 0$ and $0 \leqq l_i^\varepsilon(y) \leqq l_i$ for $l_i > 0$.

Some assumptions. For specificity, $z(\cdot)$ is assumed to be a *stationary Gaussian process* with a rational spectral density. Thus there are an asymptotically stable matrix M, a matrix C, a row vector D and a process $v(\cdot)$, such that

$$\begin{aligned} &dv = Mvdt + Cdw, \qquad z = Dv, \\ &w(\cdot) = \text{vector-valued standard Brownian motion.} \end{aligned} \tag{7.6}$$

This assumption is not essential—only certain smoothness properties of the multivariate density are used, together with the exponential rate of decrease of the effects of the initial conditions.

Define $\hat{F}_\varepsilon(y) = EF_\varepsilon(y, z(n\Delta))$ and $\hat{F}(y) = EF(y, z(n\Delta))$. Let $\sigma_0^2 = \operatorname{var} z(t)$. We have

$$\begin{aligned} \frac{d}{dy}\left(\frac{\hat{F}(y)}{y}\right) &= \frac{2}{\sqrt{2\pi}\sigma_0} \sum_{i=1}^{L} l_i \left[\xi_i \exp -\frac{\xi_i^2 y^2}{2\sigma_0^2} - \xi_{i-1} \exp -\frac{\xi_{i-1}^2 y^2}{2\sigma_0^2}\right] - \frac{\alpha}{y} \\ &= \frac{2}{\sqrt{2\pi}\sigma_0} \sum_{i=1}^{L-1} (l_i - l_{i+1})\xi_i \exp -\frac{\xi_i^2 y^2}{2\sigma_0^2} - \frac{\alpha}{y}. \end{aligned} \tag{7.7}$$

We can see from the terms in (7.7) that $\hat{F}(y)/y$ is the sum of two strictly convex functions, the first being bounded and having a negative slope, and the second going to ∞ as $y \to 0$ and to $-\infty$ as $y \to \infty$. Thus there is a unique $\bar{y} \in (0, \infty)$ such that $\hat{F}_y(\bar{y}) = 0$. Also $\hat{F}(y) > 0$ for $0 < y < \bar{y}$ and $\hat{F}(y) < 0$ for $y > \bar{y}$ and $\hat{F}_y(\bar{y}) \neq 0$. We *assume* that $y \in (y_l, y_u)$. For small ε, the assertions in the last sentence hold with $\hat{F}_\varepsilon$ replacing $\hat{F}$. Define $U_n^\varepsilon = (y_n^\varepsilon - \bar{y})/\sqrt{\varepsilon}$ and let E_n denote expectation conditioned on $\{v(j\Delta), j < n\}$.

8. Tightness of $\{U_n^\varepsilon$, small ε, large $n\}$. The proof is similar to that of Theorem 4 in § 5 and we only set it up and indicate how to deal with the fact that $\{z(n\Delta)\}$ is unbounded.

THEOREM 6. *Under the conditions in* § 7, *the conclusions of Theorem* 4 *hold.*

Proof. Define $V(y)=(y-\bar{y})^2$. There is a $\gamma>0$ such that $(y-\bar{y})\hat{F}(y)\leqq-\gamma V(y)$. For all $\varepsilon>0$ and $y\in[y_l, y_u]$, we have

$$(y_{n+1}^{\varepsilon}-y_n^{\varepsilon})^2=O(\varepsilon^2),\qquad y_{n+1}^{\varepsilon}=y_n^{\varepsilon}+\varepsilon\hat{F}_{\varepsilon}(y_n^{\varepsilon})+\varepsilon[F_{\varepsilon}(y_n^{\varepsilon}, z(n\Delta))-\hat{F}_{\varepsilon}(y_n^{\varepsilon})]+O(\varepsilon^2),$$

$$\hat{F}_{\varepsilon}(y)=y\sum_{i=1}^{L}l_i^{\varepsilon}(y)P\{y\xi_{i-1}\leqq|z(n\Delta)|<y\xi_i\}-\alpha(y\log y),$$

(8.1)

$$E_n^{\varepsilon}(y_{n+1}^{\varepsilon}-y_n^{\varepsilon})=\varepsilon\hat{F}_{\varepsilon}(y_n^{\varepsilon})+\varepsilon y_n^{\varepsilon}\sum_{i=1}^{L}l_i^{\varepsilon}(y_n^{\varepsilon})[P\{y\xi_{i-1}\leqq|z(n\Delta)|<y\xi_i|v(n\Delta-\Delta)\}$$
$$-P\{y\xi_{i-1}\leqq|z(n\Delta)|<y\xi_i\}]|_{y=y_n^{\varepsilon}}+O(\varepsilon^2).$$

As done in connection with (5.2) (where α_{ε}, β_{ε} were replaced by α, β), we get an upper bound for the second moment by replacing $l_i(y_n^{\varepsilon})$ by l_i (hence $\hat{F}_{\varepsilon}$ by $\hat{F}$). Thus

$$E_n^{\varepsilon}V(y_{n+1}^{\varepsilon})-V(y_n^{\varepsilon})\leqq O(\varepsilon^2)+2\varepsilon(y_n^{\varepsilon}-\bar{y})\hat{F}(y_n^{\varepsilon})$$
(8.2)
$$+2(y_n^{\varepsilon}-\bar{y})\,[\text{sum in (8.1) with } l_i^{\varepsilon}(\cdot) \text{ replaced by } l_i].$$

Next, define $V_1^{\varepsilon}(n)$ by $V_1^{\varepsilon}(n)=V_1^{\varepsilon}(n, y_n^{\varepsilon})$, where

$$V_1^{\varepsilon}(n, y)=2\varepsilon(y-\bar{y})\sum_{j=n}^{\infty}\sum_{i=1}^{L}yl_i\Big[P\{y\xi_{i-1}\leqq|z(j\Delta)|<y\xi_i|v(n\Delta-\Delta)\}$$
(8.3)
$$-P\{y\xi_{i-1}\leqq|z(n\Delta)|<y\xi_i\}\Big].$$

$|V_1^{\varepsilon}(n)|$ can be estimated by use of the following fact. *There are $K_0<\infty$ and $a>0$ such that $|e^{Mt}|\leqq K_0e^{-at}$. There is an $a_1>0$ and a $K_1<\infty$ such that for $\tau_2>\tau_1>0$ and on the set $\{v(t)\colon|v(t)|e^{-a\tau_1/2}\leqq1\}$,*

$$|P\{v(t+\tau_i)\in B_i,\ i=1,2|v(t)\}-P\{v(t+\tau_i)\in B_i,\ i=1,2\}|\leqq K_1e^{-a_1\tau_1}.\tag{8.4}$$

for all B_1, B_2.

In order to use (8.4) (in this application we set $B_2=$ range space of $v(t)$, and write the *sum* in (8.3) as

$$\sum_{j=n}^{H}+\sum_{j=H+1}^{\infty},\tag{8.5}$$

where $H=\min\{m\colon e^{-(m-n)\Delta a/2}|v(n\Delta-\Delta)|\leqq1\}=O(1+\max(0,\log|v(n\Delta-\Delta)|))^{+n}$. Then the first sum in (8.5) is $O(1+\max(0,\log|v(n\Delta-\Delta)|))$, and the second is $O(1)$ by (8.4) and the summability of $\sum_{j\geqq0}\exp(-a_1j\Delta)$. Thus $|V_1^{\varepsilon}(n)|=O(\varepsilon)[1+\max(0,\log|v(n\Delta-\Delta)|)]\leqq O(\varepsilon)(1+|v(n\Delta-\Delta)|)$. From this point on, the proof is exactly the same as that for Theorem 4. □

9. The limit theorem. We continue to use the tilde ~ terminology of § 6, and define $\tilde{U}_n^{\varepsilon}$, $\tilde{y}_n^{\varepsilon}$, $\tilde{E}_n^{\varepsilon}$, etc., as there. Also, set $\tilde{z}(n\Delta)=z(n_{\varepsilon}\Delta+n\Delta)$ and $\tilde{v}(n\Delta)=v(n_{\varepsilon}\Delta+n\Delta)$. The idea now is still to prove weak convergence of $\tilde{U}^{\varepsilon}(\cdot)$. We use $\tilde{E}_n^{\varepsilon}$ for expectation conditional on $\{v(j\Delta), j<n+n_{\varepsilon}\}$. We have ((9.1b) defines $\tilde{y}_n^{\varepsilon,N}$ by $\tilde{U}_n^{\varepsilon,N}=(\tilde{y}_n^{\varepsilon,N}-\bar{y})/\sqrt{\varepsilon}$)

$$\tilde{U}_{n+1}^{\varepsilon}=\tilde{U}_n^{\varepsilon}+\sqrt{\varepsilon}\hat{F}_{\varepsilon}(\tilde{y}_n^{\varepsilon})+\sqrt{\varepsilon}(F_{\varepsilon}(\tilde{y}_n^{\varepsilon}, \tilde{z}(n\Delta))-\hat{F}_{\varepsilon}(\tilde{y}_n^{\varepsilon}))+O(\varepsilon^{3/2}),\tag{9.1a}$$

$$\tilde{U}_{n+1}^{\varepsilon,N}=\tilde{U}_n^{\varepsilon,N}+\sqrt{\varepsilon}[\hat{F}_{\varepsilon}(\tilde{y}_n^{\varepsilon,N})+(F_{\varepsilon}(\tilde{y}_n^{\varepsilon,N}, \tilde{z}(n\Delta))-\hat{F}_{\varepsilon}(\tilde{y}_n^{\varepsilon,N}))+O(\varepsilon^{3/2})]b_N(\tilde{U}_n^{\varepsilon,N}).\tag{9.1b}$$

THEOREM 7. *Under the conditions of* § 7, *the conclusions of Theorem* 5 *hold, but, now,* $G = \hat{F}_y(\bar{y})$ *and (stationary process* $z(\cdot)$ *used)*

$$\sigma^2 = EF^2(\bar{y}, z(0)) + 2\sum_{n=1}^{\infty} EF(\bar{y}, z(n\Delta))F(\bar{y}, z(0)).$$

Remark. If M, C or D is time-varying, then an extension of the technique is possible, provided that the time variation per step is $O(\varepsilon)$. The limit diffusion yields information on the dependence of the performance on the parameters α, $\{l_i\}$, Δ, $\{\xi_i\}$, as well as an estimate of the asymptotic variance and correlation function for small ε.

Proof. Except for the unboundedness of the noise $\{z(n\Delta)\}$, the proof would be essentially the same as that of Theorem 5, and only an outline will be given.

Owing to the truncation $|\tilde{U}_n^{\varepsilon,N}| \leqq N+1$, the F_ε, $\hat{F}_\varepsilon$ in (9.1b) *can be replaced* by F and $\hat{F}$, respectively, without changing the values, for small ε. Let us make the replacement. Fix $f(\cdot,\cdot) \in \hat{\mathcal{C}}_0^{2,3}$. *Drop the superscript N on all variables for notational convenience, as done in Theorem* 5. Then, by a Taylor expansion,

$$\begin{aligned}
&\tilde{E}_n^\varepsilon f(\tilde{U}_{n+1}^\varepsilon, n\varepsilon+\varepsilon) - f(\tilde{U}_n^\varepsilon, n\varepsilon) \\
&\quad = o(\varepsilon) + \varepsilon f_t(\tilde{U}_n^\varepsilon, n\varepsilon) + \varepsilon f_u(\tilde{U}_n^\varepsilon, n\varepsilon)\hat{F}_y(\bar{y})\tilde{U}_n^\varepsilon b_N(\tilde{U}_n^\varepsilon) \\
&\qquad + \sqrt{\varepsilon} f_u(\tilde{U}_n^\varepsilon, n\varepsilon)\tilde{E}_n^\varepsilon[F(\bar{y}+\sqrt{\varepsilon}\tilde{U}_n^\varepsilon, \tilde{z}(n\Delta)) - \hat{F}(\bar{y}+\sqrt{\varepsilon}\tilde{U}_n^\varepsilon)]b_N(\tilde{U}_n^\varepsilon) \\
&\qquad + \frac{\varepsilon}{2} f_{uu}(\tilde{U}_n^\varepsilon, n\varepsilon)\tilde{E}_n^\varepsilon[F(\bar{y}+\sqrt{\varepsilon}\tilde{U}_n^\varepsilon, \tilde{z}(n\Delta)) - \hat{F}(\bar{y}+\sqrt{\varepsilon}\tilde{U}_n^\varepsilon)]^2 b_N^2(\tilde{U}_n^\varepsilon).
\end{aligned} \tag{9.2}$$

Since the second derivative of $\tilde{E}_n^\varepsilon F(y, \tilde{z}(n\Delta))$ with respect to y is bounded by constant $[1+|\tilde{v}(n\Delta-\Delta)|]$, the next-to-last term of (9.2) can be written as

$$\begin{aligned}
&\sqrt{\varepsilon} f_u(\tilde{U}_n^\varepsilon, n\varepsilon)\tilde{E}_n^\varepsilon[F(\bar{y}, \tilde{z}(n\Delta)) - \hat{F}(\bar{y})]b_N(\tilde{U}_n^\varepsilon) + \varepsilon f_u(\tilde{U}_n^\varepsilon, n\varepsilon) \\
&\quad \cdot \frac{\partial}{\partial y}\tilde{E}_n^\varepsilon[F(y, \tilde{z}(n\Delta)) - \hat{F}(y)]|_{y=\bar{y}}\tilde{U}_n^\varepsilon b_N(\tilde{U}_n^\varepsilon) + o(\varepsilon)[1+|\tilde{v}(n\Delta-\Delta)|].
\end{aligned} \tag{9.3}$$

The last term of (9.2) can be written as (recall that $\hat{F}(\bar{y}) = 0$)

$$\frac{\varepsilon}{2} f_{uu}(\tilde{U}_n^\varepsilon, n\varepsilon)\tilde{E}_n^\varepsilon[F(\bar{y}, \tilde{z}(n\Delta)) - \hat{F}(\bar{y})]^2 b_N^2(\tilde{U}_n^\varepsilon) + o(\varepsilon). \tag{9.4}$$

Now, we use the method of Theorem 5 in order to average out the terms of (9.2). We use $f^\varepsilon(n\varepsilon) = f(\tilde{U}_n^\varepsilon, n\varepsilon) + \sum_{i=3}^6 f_i^\varepsilon(n\varepsilon)$. Define $f_3^\varepsilon(n\varepsilon)$ by (to average out the second term of (9.3))

$$f_3^\varepsilon(n\varepsilon) = \varepsilon f_u(\tilde{U}_n^\varepsilon, n\varepsilon)b_N(\tilde{U}_n^\varepsilon)\tilde{U}_n^\varepsilon \sum_{j=n}^{\infty} \frac{\partial}{\partial y}\tilde{E}_n^\varepsilon[F(y, \tilde{z}(j\Delta)) - \hat{F}(y)]\Big|_{y=\bar{y}}.$$

By an argument similar to that used below (8.5), together with the derivative bound stated above (9.3), it can be shown that $\tilde{E}_n^\varepsilon f_3^\varepsilon(n\varepsilon+\varepsilon) - f_3^\varepsilon(n\varepsilon) = -$(second term of (9.3)) $+ o(\varepsilon)[1+|\tilde{v}(n\Delta-\Delta)|^2]$ and that $|f_3^\varepsilon(n\varepsilon)| \leqq O(\varepsilon)[1+|\tilde{v}(n\Delta-\Delta)|]$. Next, introduce $f_4^\varepsilon(n\varepsilon)$ (to average out (9.4)):

$$f_4^\varepsilon(n\varepsilon) = \frac{\varepsilon}{2} f_{uu}(\tilde{U}_n^\varepsilon, n\varepsilon)b_N^2(\tilde{U}_n^\varepsilon) \sum_{j=n}^{\infty} [\tilde{E}_n^\varepsilon F^2(\bar{y}, \tilde{z}(j\Delta)) - EF^2(\bar{y}, \tilde{z}(j\Delta))].$$

Then, as for f_3^ε, we have $|f_4^\varepsilon(n\varepsilon)| \leqq O(\varepsilon)[1+|\tilde{v}(n\Delta-\Delta)|]$. Using this, it is not hard to

show via a small amount of manipulation that

$$\tilde{E}_n^\varepsilon f_4^\varepsilon(n\varepsilon+\varepsilon)-f_4^\varepsilon(n\varepsilon)=-\frac{\varepsilon}{2}f_{uu}(\tilde{U}_n^\varepsilon, n\varepsilon)b_N(\tilde{U}_n^\varepsilon)[\tilde{E}_n^\varepsilon F^2(\bar{y}, \tilde{z}(n\Delta))-EF^2(\bar{y}, \tilde{z}(n\Delta))]$$
$$+o(\varepsilon)[1+|\tilde{v}(n\Delta-\Delta)|].$$

Next, introduce $f_5^\varepsilon(n\varepsilon)$ in order to average out the first term of (9.3):

$$f_5^\varepsilon(n\varepsilon)=\sqrt{\varepsilon}f_u(\tilde{U}_n^\varepsilon, n\varepsilon)b_N(\tilde{U}_n^\varepsilon)\sum_{j=n}^{\infty}\tilde{E}_n^\varepsilon F(\bar{y}, \tilde{z}(j\Delta)).$$

Then, again, $|f_5^\varepsilon(n\varepsilon)|=o(\sqrt{\varepsilon})(1+|\tilde{v}(n\Delta-\Delta)|)$ and we can write

$$\tilde{E}_n^\varepsilon f_5^\varepsilon(n\varepsilon+\varepsilon)-f_5^\varepsilon(n\varepsilon)=-(\text{first term of (9.3)})$$
$$+\varepsilon\tilde{E}_n^\varepsilon[f_u(\tilde{U}_{n+1}^\varepsilon, n\varepsilon)b_N(\tilde{U}_{n+1}^\varepsilon)-f_u(\tilde{U}_n^\varepsilon, n\varepsilon)b_N(\tilde{U}_n^\varepsilon)]$$
$$\cdot\sum_{j=n+1}^{\infty}\tilde{E}_{n+1}^\varepsilon F(\bar{y}, \tilde{z}(j\Delta)). \tag{9.5a}$$

With a small amount of manipulation, we can show that the last term of (9.5a) equals

$$\varepsilon b_N(\tilde{U}_n^\varepsilon)[f_u(\tilde{U}_n^\varepsilon, n\varepsilon)b_N(\tilde{U}_n^\varepsilon)]_u\sum_{j=n+1}^{\infty}\tilde{E}_n^\varepsilon F(\bar{y}, \tilde{z}(j\Delta))F(\bar{y}, \tilde{z}(n\Delta))+o(\varepsilon)$$
$$\cdot[1+|\tilde{v}(n\Delta-\Delta)|]. \tag{9.5b}$$

Finally, $f_6^\varepsilon(n\varepsilon)$ is introduced in order to average out the sum term in (9.5b) in the same way that $f_2^\varepsilon(n\varepsilon)$ was used to average out (6.11) in Theorem 5. Define

$$f_6^\varepsilon(n\varepsilon)=\varepsilon[f_u(\tilde{U}_n^\varepsilon, n\varepsilon)b_N(\tilde{U}_n^\varepsilon)]_u b_N(\tilde{U}_n^\varepsilon)$$
$$\cdot\sum_{j=n}^{\infty}\sum_{k=j+1}^{\infty}[\tilde{E}_n^\varepsilon F(\bar{y}, \tilde{z}(k\Delta))F(\bar{y}, \tilde{z}(j\Delta))-EF(\bar{y}, \tilde{z}(k\Delta))F(\bar{y}, \tilde{z}(j\Delta))]. \tag{9.6}$$

By (8.4), $f_6^\varepsilon(n\varepsilon)$ is well defined and is $O(\varepsilon)[1+|\tilde{v}(n\Delta-\Delta)|^2]$. Also,

$$\tilde{E}_n^\varepsilon f_6^\varepsilon(n\varepsilon+\varepsilon)-f_6^\varepsilon(n\varepsilon)=-(\text{sum term in (9.5)})$$
$$+\varepsilon b_N(\tilde{U}_n^\varepsilon)[f_u(\tilde{U}_n^\varepsilon, n\varepsilon)b_N(\tilde{U}_n^\varepsilon)]_u\sum_{n=1}^{\infty}EF(\bar{y}, \tilde{z}(n\Delta))F(\bar{y}, \tilde{z}(0))$$
$$+o(\varepsilon)[1+|v(n\Delta-\Delta)|^2].$$

Summarizing, with $f^\varepsilon(n\varepsilon)$ defined by $f^\varepsilon(n\varepsilon)\equiv f(\tilde{U}_n^\varepsilon, n\varepsilon)+\sum_{i=3}^{6}f_i^\varepsilon(n\varepsilon)$, we have

$$\tilde{E}_n^\varepsilon f^\varepsilon(n\varepsilon+\varepsilon)-f^\varepsilon(n\varepsilon)=o(\varepsilon)[1+|\tilde{v}(n\Delta-\Delta)|^2]+\varepsilon f_t(\tilde{U}_n^\varepsilon, n\varepsilon)$$
$$+\varepsilon f_u(\tilde{U}_n^\varepsilon, n\varepsilon)\hat{F}_y(\bar{y})\tilde{U}_n^\varepsilon b_N(\tilde{U}_n^\varepsilon)+\varepsilon b_N(\tilde{U}_n^\varepsilon)$$
$$\cdot[f_u(\tilde{U}_n^\varepsilon, n\varepsilon)b_N(\tilde{U}_n^\varepsilon)]_u\sum_{n=1}^{\infty}EF(\bar{y}, z(n\Delta))F(\bar{y}, z(0))$$
$$+\varepsilon b_N(\tilde{U}_N^\varepsilon)\frac{f_{uu}(\tilde{U}_n^\varepsilon, n\varepsilon)}{2}EF^2(\bar{y}, z(0))+o(\varepsilon)$$
$$\cdot[1+|\tilde{v}(n\Delta-\Delta)|^2]. \tag{9.7}$$

Now, if the $\{\tilde{U}^{\varepsilon,N}(\cdot)\}$ (*returning to the use of superscript N*) are tight for each N, then (9.7) and Theorem 1 imply that any weakly convergent subsequence of $\{\tilde{U}^{\varepsilon,N}(\cdot)\}$ converges to a diffusion with operator A^N, whose coefficients equal those of A in S_N and, hence, that the original $\{\tilde{U}^{\varepsilon}(\cdot)\}$ converge weakly to the solution of (6.3) with the G and σ defined in Theorem 7.

But (*dropping the superscript N again*) $|\sum_{i=3}^{6} f_i^{\varepsilon}(n\varepsilon)| = O(\sqrt{\varepsilon})[1+|\tilde{v}(n\Delta-\Delta)|^2]$ and $|\tilde{E}_n^{\varepsilon} f^{\varepsilon}(n\varepsilon+\varepsilon)-f^{\varepsilon}(n\varepsilon)| = O(\varepsilon)+o(\varepsilon)[1+|\tilde{v}(n\Delta-\Delta)|^2]$ and for any $T<\infty$, $K>0$, the Gaussian property implies that

$$\lim_{\varepsilon\to 0} P\{\sup_{n\leqq T/\varepsilon} \varepsilon|v(n\varepsilon)|^4, \geqq K\} = 0.$$

Thus, since the above $o(\varepsilon)$ satisfies $o(\varepsilon) = O(\varepsilon^{3/2})$, tightness follows by [1, Thm. 2] or [7, Thm. 2] as it did for the case of Theorem 1. □

REFERENCES

[1] H. J. KUSHNER AND HAI HUANG (1979), *On the weak convergence of a sequence of general stochastic difference equations to a diffusion*, SIAM J. Applied Math., 40 (1981), pp. 528–541.

[2] K. S. NARENDRA, E. A. WRIGHT AND L. E. MASON (1977), *Application of learning automata to telephone traffic routing and control*, IEEE Trans. Systems, Man Cyber., SMC-7, pp. 785–792.

[3] K. S. NARENDRA AND M. A. L. THATHACHAR (1979), *On the behavior of a learning automaton in a changing environment with application to telephone traffic routing*, preprint, Dept. of Engineering, Yale University.

[4] D. J. GOODMAN AND A. GERSHO (1974), *Theory of an adaptive quantizer*, IEEE Trans. Comm., COM-22, pp. 1037–1045.

[5] D. MITRA (1979), *A generalized adaptive quantization system with a new reconstruction method for noisy transmission*, IEEE Trans. Comm., COM-27, pp. 1681–1689.

[6] P. BILLINGSLEY (1968), *Convergence of Probability Measures*, John Wiley, New York.

[7] H. J. KUSHNER (1979), *A martingale method for the convergence of a sequence of processes to a jump-diffusion process*, Z. Wahrsch., 53, pp. 207–219.

[8] D. W. STROOK AND S. R. S. VARADHAN (1979), *Multidimensional Diffusion Processes*, Springer, Berlin.

[9] M. F. NORMAN (1974), *Markovian learning processes*, SIAM Rev., 16, pp. 143–162.

[10] H. J. KUSHNER AND HAI HUANG (1980), *Averaging methods for the asymptotic analysis of learning and adaptive systems with small adjustment rate*, Brown Univ., Rep. 80–1.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 5, September 1981

0363-0129/81/1905-0007 $01.00/0

SOLUTION OF SOME STOCHASTIC QUADRATIC NASH AND LEADER-FOLLOWER GAMES*

G. P. PAPAVASSILOPOULOS†

Abstract. The linear quadratic Gaussian static Nash and Stackelberg two-player games are considered and completely solved. Necessary and sufficient conditions for existence and uniqueness of the solutions are presented as well as the procedure for finding all the solutions. For the Nash game, in particular, it is shown that if there exists a solution there will exist a solution affine in the information, and that the solution will be nonunique if (intuitively) the coupling of the information of the two players equals some power of the inverse of the coupling of their costs. Many interesting dynamic cases with nested information structures can be reduced to static ones and are essentially covered by the analysis presented.

1. Introduction. It has been recognized that the single objective optimization problem cannot capture all the aims of a decision procedure. Usually there are many conflicting objectives that a decision maker has to meet, and the formulation of a single objective as a weighted sum of the several objectives is not necessarily the only way to go. Also, there might exist many decision makers with conflicting objectives who do not agree on an overall average objective. On the other hand, an existing hierarchy among the several decision makers in a certain organization should not be ignored when one creates the mathematical model. Such considerations make game theory a natural vehicle for studying multiobjective hierarchical decision procedures. In particular, the so-called Nash and leader-follower (or Stalkelberg) games offer themselves for studying such situations. For definitions and some properties of these games see [2], [3]. (See also [13] for some recent results concerning leader-follower games and their relation to the theory of incentives in economics.)

There are several results concerning Nash and leader-follower games, but there are still many open problems. In this paper we study, and solve completely the static Nash and leader-follower games, where the players have quadratic costs and linear measurements of a random variable which enters linearly into the costs, see (1)–(7). Several dynamic cases (where there is time evolution), are included in the static formulation, as long as appropriate nestedness conditions [4] are imposed on the information of the players. For example, the stochastic linear quadratic discrete time Gaussian Nash game, where the players share at each stage all the past information with one step delay, belongs to the class of dynamic games that can be reduced to static ones. Another example is the stochastic linear quadratic, discrete time Gaussian leader-follower game, where the leader has information only at the first stage, whereas the follower in addition to having his own information acquires the information of the leader with one step delay. Although such dynamic problems can be handled by the methods developed here, we will focus on the static case only. We will nonetheless provide in § 5 the procedure, which reduces dynamic problems of this type to static ones.

The only existing results concerning such types of stochastic games are in [6], [7]. There, sufficient conditions for existence and uniqueness of solutions are found by imposing a contraction assumption. As a consequence, the results of [6], [7], in addition

* Received by the editors March 17, 1980, and in revised form January 20, 1981. This work was supported in part by the U.S. Air Force Office of Scientific Research under grant AFOSR-80-0171.

† Dept. of Electrical Engineering, University of Southern California, University Park, Los Angeles, CA 90007.

to being extremely conservative, cannot answer the important question of how the interplay among the information and the costs of the players affects the solution.

The structure of the present paper is the following. In § 2 we pose the problems and in § 3 we give the complete solution of the Nash game. It is shown that the solution will be nonunique if some numbers, which are products of even powers of the canonical correlation coefficients of the information of the two players, are inverses of eigenvalues of a matrix (R_1R_2), which represents the strength of the coupling of the costs. Intuitively, the solution will be nonunique if the coupling of the information is equal to some power of the inverse of the coupling of the costs. It is also shown that if there exists a solution there will exist a solution linear in the information. The way to construct all the solutions is also given. In § 4 we solve the leader-follower game. In § 5 we derive a sufficient condition for the existence and uniqueness of a solution of an equation which is a generalization of an equation playing a central role in § 3. A special case of this condition was presented in [5], but our result, in addition to being more powerful, is proven in a much easier fashion. This condition can be used to guarantee existence and uniqueness of solutions of Nash and leader-follower games, if there are many players and one is not willing to generalize the exact results of §§ 3 and 4 to the many player case. In this section we also sketch a way to reduce dynamic problems with nested information to static ones. Finally, § 6 presents the solution of a simple one-dimensional Nash static stochastic game which can serve to illustrate some aspects of the whole analysis. § 7 is a conclusions section. The proofs of two lemmas used in § 3 are given in Appendix A and B. Appendix C contains some results concerning an operator which turns out to be of importance when solving the Nash or leader-follower games.

In our analysis, we will consider two players only, but generalization to the many player case is possible.

2. Statement of the problems. Let $x : \Omega \to R^n$ be a Gaussian random variable with respect to a probability space $(\Omega, \mathcal{F}, P)$, which, without loss of generality, is assumed to have zero mean and unit covariance matrix. Let

$$y_i = C_i x, \qquad i = 1, 2, \tag{1}$$

where the C_i's are real matrices of dimension $n_i \times n$. The random variable $y_i : \Omega \to R^{n_i}$ generates a minimal sub σ-field $\mathcal{F}_i$ of $\mathcal{F}$ in Ω. Let U_i, $i = 1, 2$ denote the space of functions $u_i : \Omega \to R^{m_i}$, which are $\mathcal{F}_i$ measurable and for which $\|u_i\|^2 = \langle u_i, u_i \rangle < +\infty$, where the inner product in U_i is defined by

$$\langle u_i, v_j \rangle = \int_\Omega u_i' v_j \, dP, \qquad u_i, v_i \in U_i. \tag{2}$$

U_i is a separable Hilbert space and u_i can be considered as a function of y_i; see [8]. For a given pair $(u_1, u_2) \in U_1 \times U_2$ consider

$$J_1(u_1, u_2) = E[\tfrac{1}{2}u_1'u_1 + u_1'R_1u_2 + u_1'S_1x + u_2'Q_{12}u_2 + u_2'F_1x], \tag{3}$$

$$J_2(u_1, u_2) = E[\tfrac{1}{2}u_2'u_2 + u_2'R_2u_1 + u_2'S_2x + u_1'Q_{21}u_1 + u_1'F_2x], \tag{4}$$

where E denotes total expectation, and $Q_{ij} = Q_{ij}'$, R_i, S_i, F_i are real matrices of appropriate dimensions. We want to solve the problems N and S.

Problem N. Find the pairs $(u_1^*, u_2^*) \in U_1 \times U_2$ for which

$$J_1(u_1^*, u_2^*) \leqq J_1(u_1, u_2^*) \quad \forall u_1 \in U_1, \tag{5}$$

$$J_2(u_1^*, u_2^*) \leqq J_2(u_1^*, u_2) \quad \forall u_2 \in U_2. \tag{6}$$

Problem S. Find the pairs $(u_1^*, u_2^*) \in U_1 \times U_2$ which solve:

$$\begin{aligned} &\text{minimize} && J_1(u_1, u_2) \\ &\text{subject to} && (u_1, u_2) \in U_1 \times U_2 \quad \text{and} \quad J_2(u_1, u_2) \leqq J_2(u_1, \bar{u}_2) \quad \forall \bar{u}_2 \in U_2. \end{aligned} \tag{7}$$

The formalism (1)–(6) describes a two-player nonzero sum, static, stochastic, Nash game, where player i has information y_i, chooses u_i and wants to minimize J_i. The formalism (1)–(4), (7) describes a two-player, nonzero sum, static, stochastic, leader-follower game, where player i has information y_i, chooses u_i and wants to minimize J_i, and, in addition, player 1 (leader) decides and announces his decision u_1, first, before player 2 (follower) decides on u_2.

Without loss of generality, we make the following assumption, which is assumed to hold throughout the present paper:

Assumption.

$$\text{rank } C_i = n_i, \qquad i = 1, 2.$$

The formula $E[x \mid y_i] = C_i'(C_i C_i')^{-1} C_i x$, will be used repeatedly in the later sections.

3. Solution of the problem N. In this section we solve problem N. For fixed $u_2 \in U_2$, the problem

$$\text{minimize} \qquad J_1(u_1, u_2), \qquad u_1 \in U_1, \tag{8}$$

is a quadratic minimization problem in the Hilbert space $U_{m_1} = \{u : \Omega \to R^{m_1},\ u \text{ is } \mathcal{F} \text{ measurable and } \|u\| < +\infty\}$, which has U_1 as a closed subspace. Use of the projection theorem yields (9) as a necessary and sufficient condition for u_1 to solve (8):

$$u_1 + E[R_1 u_2 \mid y_1] + E[S_1 x \mid y_1] = 0; \tag{9}$$

see [1] for details. $E[\cdot \mid y_i]$ denotes conditional expectation. Similarly for fixed $u_1 \in U_1$, the problem

$$\text{minimize} \qquad J_2(u_1, u_2), \qquad u_2 \in U_2, \tag{10}$$

has u_2 as a solution if and only if

$$u_2 + E[R_2 u_1 \mid y_2] + E[S_2 x \mid y_2] = 0. \tag{11}$$

Substituting u_2 from (11) into (9), we conclude that the study of Problem N is equivalent to the study of the equation

$$u_1 - R_1 R_2 E[E[u_1 \mid y_2] \mid y_1] = R_1 S_2 E[E[x \mid y_2] \mid y_1] - S_1 E[x \mid y_1], \tag{12}$$

on which we will concentrate from now on.

We will need the following lemma.

LEMMA 1. *There exist nonsingular square matrices T_1, T_2 so that the matrices*

$$\bar{C}_1 = T_1 C_1, \qquad \bar{C}_2 = T_2 C_2,$$

have the following properties:

(1) $\bar{C}_1 \bar{C}_1' = I$, $\bar{C}_2 \bar{C}_2' = I$.

(2)

$$\bar{C}_1 = \begin{bmatrix} \bar{C}_{111} \\ \bar{C}_{112} \\ \bar{C}_{12} \end{bmatrix}, \qquad \bar{C}_2 = \begin{bmatrix} \bar{C}_{222} \\ \bar{C}_{221} \\ \bar{C}_{21} \end{bmatrix}, \qquad \bar{C}_{12} = \bar{C}_{21}, \qquad \bar{C}_1 \bar{C}_2' = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \bar{C}_{112} \bar{C}_{221}' & 0 \\ 0 & 0 & I \end{bmatrix}.$$

(3) *$\bar{C}_{112}$ and $\bar{C}_{221}$ have the same dimensions $k \times n$ and*

$$\bar{C}_{112}\bar{C}'_{221} = \begin{bmatrix} \sqrt{\mu_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\mu_k} \end{bmatrix} \quad \text{where } 0 \leqq \sqrt{\mu_i} \leqq 1.$$

(4) *The dimensions of all the component matrices are uniquely determined.*

The proof of this lemma can be found in standard books on statistics [12], where the elements of $\bar{C}_1\bar{C}'_2$ are called the canonical correlation coefficients of y_1, y_2 and algorithms for finding T_1, T_2 are described. For the sake of completeness we present a proof of Lemma 1 in Appendix A.

The importance of Lemma 1 for our problem is that since T_i is nonsingular, the minimal σ-fields generated by y_i and T_iy_i are the same, and thus, we can consider equivalently that u_i is a function of T_iy_i. *In the following we will assume that C_1, C_2 have been brought into the form suggested by Lemma* 1 (and drop the bars from $\bar{C}_1$, $\bar{C}_2$).

In terms of the information structure of the game, Lemma 1 allows us to consider y_1 and y_2 as normal Gaussian vectors which can be decomposed into independent components as $y_1 = (y_{112}, y_{111}, y_{12})$, $y_2 = (y_{221}, y_{222}, y_{21})$, where $y_{12} = y_{21}$ is the common information, y_{111} is known only to player 1, y_{222} is known only to player 2 and y_{112}, y_{221} represent the nontrivial coupling of the information.

Let us introduce the following notation:

$$\begin{gathered} y_i = C_ix, \quad y_{ij} = C_{ij}x, \quad y_{ijl} = C_{ijl}x, \\ P_i = E[\cdot \mid y_i], \quad P_{ij} = E[\cdot \mid y_{ij}], \quad P_{ijl} = E[\cdot \mid y_{ijl}], \qquad i, j, l = 1, 2. \end{gathered} \tag{13}$$

P_i, P_{ij}, P_{ijl} and E are projections in the Hilbert space U_{m_1}. An equivalent interpretation of Lemma 1 is that, if without loss of generality, we impose $E[u] = 0$ in U_{m_1}, then P_1 and P_2 can be decomposed into sums of orthogonal projections; i.e., $P_1 = P_{111} + P_{112} + P_{12}$, $P_2 = P_{222} + P_{221} + P_{21}$, where $P_{12} = P_{21}$, $P_{111}P_2 = 0$, $P_{222}P_1 = 0$, $P_{112}P_{221} \neq P_{221}P_{112}$ and $\|P_{221}P_{112}\| = \|P_{112}P_{221}\| < 1$ (see Lemma 2). One can verify that $P_{12} = \lim (P_1P_2)^n$ as $n \to +\infty$, $P_{111} = \lim (P_1(I - P_2))^n$, $P_{222} = \lim (P_2(I - P_2))^n$, $P_{112} = P_1 - P_{12} - P_{111}$, $P_{221} = P_2 - P_{21} - P_{222}$ (see [10, problem 96]).

We can write (12) as

$$u_1 - R_1R_2P_1P_2u_1 = Sy_1, \tag{14}$$

where

$$S = R_1S_2C'_2C_2C'_1 - S_1C'_1, \qquad S = [s_1 | \cdots | s_{n_1}]. \tag{15}$$

We will construct an orthonormal complete set for U_1. Let

$$p_n(z) = \frac{(-1)^n}{\sqrt{n!}} e^{\frac{1}{2}z^2} \frac{d^n}{dz^n} e^{-\frac{1}{2}z^2}, \qquad n = 0, 1, 2, \cdots, \tag{16}$$

be the Hermite polynomials which constitute an orthonormal complete set with respect to the Gaussian measure $\mu(z)$; i.e.,

$$\int_{-\infty}^{+\infty} p_n(z)p_l(z)\, d\mu(z) = \begin{cases} 1 & \text{if } n = l, \\ 0 & \text{if } n \neq l, \end{cases} \tag{17}$$

where

$$\mu(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}w^2}\, dw \tag{18}$$

(see [9, p. 217]). Let $y_1 = (z_1, \cdots, z_{n_1})'$. Since y_1 is a normal Gaussian random vector,

$$\{p_{k_1}(z_1)p_{k_2}(z_2)\cdots p_{k_{n_1}}(z_{n_1})\}, \qquad k_1, \cdots, k_{n_1} \in \{0, 1, 2, \cdots\} \tag{19}$$

constitute an orthonormal complete set with which we can express every component of u_1, see [11, p. 56]. Let us enumerate this set and denote it by

$$\{\bar{p}_0(y_1), \bar{p}_1(y_1), \cdots\}. \tag{20}$$

We will assume, without loss of generality, that $\bar{p}_0 = 1$, $(\bar{p}_1(y_1), \cdots, \bar{p}_{n_1}(y_1))' = y_1$, and that as n increases, the power of each component of y_1 in $\bar{p}_n(y_1)$ goes to infinity. Each $u_1 \in U_1$ can be expressed as

$$u_1(y_1) = \sum_{n=0}^{\infty} c_n \bar{p}_n(y_1), \tag{21}$$

where $c_n \in R^{m_1}$ and

$$\sum_{n=0}^{\infty} \|c_n\|^2 < +\infty. \tag{22}$$

Equation (14) can be written equivalently as

$$\sum_{n=0}^{\infty} c_n \bar{p}_n(y_1) - R_1 R_2 \sum_{n=0}^{\infty} c_n P_1 P_2 \bar{p}_n(y_1) = S y_1. \tag{23}$$

We will need the following lemma, the proof of which is given in Appendix B.

LEMMA 2. *Let*

$$a_{nl} = E[\bar{p}_n(y_1) P_1 P_2 \bar{p}_l(y_1)], \qquad a_n = a_{nn}.$$

Then the following hold:

(1) $a_{nl} = a_{ln}$.

(2) $a_{nl} = 0$ *if* $n \neq l$.

(3) $a_n = 0$ *if* $\bar{p}_n$ *depends on* y_{111}.

(4) $a_n = 1$ *if* $\bar{p}_n$ *depends only on* y_{12}, *or if* $n = 0$.

(5) *If* $\bar{p}_n(y_1) = \bar{p}_{n_3}(y_{12}) \cdot p_{m_1}(z_1) \cdots p_{m_k}(z_k)$, *where* $y_{112} = (z_1, \cdots, z_k)'$ *then* $a_n = \mu_1^{m_1} \cdots \mu_k^{m_k}$, a_n *is independent of* n_3, $0 < a_n < 1$ *and these* a_n*'s constitute a sequence decreasing strictly to zero.*

(6) *The operators* $P_{112}P_{221}$, $P_{221}P_{112}$, *restricted on the domain of the* u*'s with* $E[u] = 0$, *have norm equal to* $\max\{\mu_1, \cdots, \mu_k\} < 1$.

Multiplying both sides of (23) by $\bar{p}_n(y_1)$, taking expectation and using Lemma 2 (2) yields

$$c_n - R_1 R_2 c_n a_n = \begin{cases} s_n & \text{if } n = 1, \cdots, n_1, \\ 0 & \text{otherwise,} \end{cases} \tag{24}$$

or more concisely

$$[c_0, c_1, \cdots] - R_1 R_2 [c_0, c_1, \cdots] \begin{bmatrix} a_0 & & 0 \\ & a_1 & \\ 0 & & \ddots \end{bmatrix} = [\underbrace{0}_{m_1 \times 1} | S | 00 \cdots]. \tag{25}$$

The conditions for solvability of (25) are apparent.

Let us now state formally all the previous analysis, in the form of a theorem.

THEOREM 1. *Consider equation* (12):

(1) *It has a solution if and only if there exist* $c_0, c_1, \cdots \in \mathbb{R}^{m_1}$ *which satisfy* (25).

(2) *If there exists at least one solution, then there exists a solution linear in (the information)* y_1.

(3) *The general solution, if it exists, has the form*

$$u_1 = c_0 + [c_1, \cdots, c_{n_1}]y_1 + \sum_{k=1}^{q} c_{\bar{n}_k}\bar{p}_{\bar{n}_k}(y_{112})\phi_k(y_{12}) + [c_{\bar{m}_1}, \cdots, c_{\bar{m}_r}]\phi(y_{12}), \tag{26}$$

where $c_0, c_1, \cdots, c_{n_1}, c_{\bar{n}_1}, \cdots, c_{\bar{n}_q}$ *satisfy* (24), $\bar{n}_1, \cdots, \bar{n}_q > n_1$, $c_{\bar{m}_1}, \cdots, c_{\bar{m}_r}$ *constitute a basis for the null space of* $(I - R_1R_2)$, ϕ *and* ϕ_k *are arbitrary measurable functions of* y_{12} *taking values in* R^r *and* R *respectively,* $E[\phi_k^2] = 1$, $E[\phi'\phi] < +\infty$, *and* ϕ *contains no affine term in* y_{12} *(i.e.,* $E[\phi] = 0$, $E[y_{12} \cdot \phi'] = 0$).

Proof. The proof is immediate from the previous analysis. It need only be pointed out that the $c_{\bar{n}_k}$'s will be finite in number since as n increases the a_n's which correspond to $\bar{p}_n(y_{112})$ decrease to zero, and R_1R_2 has a finite number of eigenvalues. c_0, $c_{\bar{m}_1}, \cdots, c_{\bar{m}_r}$ are all eigenvectors of R_1R_2 corresponding to the eigenvalue 1. The appearance of the product $\bar{p}_{\bar{n}_k}(y_{112})\phi_k(y_{12})$ is a consequence of Lemma 2(5). □

The procedure suggested by Theorem 1 for solving (12) is the following.

Step 1. Try to find $u_L = c_0 + L_1y_1$, which solves (12). ($c_0 \in R^{m_1}$ and L_1 will equal $[c_1, \cdots, c_{n_1}]$ see (26).) This is equivalent to solving the equations

$$(I - R_1R_2)c_0 = 0,$$

$$L_1 - R_1R_2L_1C_1C_2'(C_2C_2')^{-1}C_2C_1'(C_1C_1')^{-1} = R_1S_2C_2'(C_2C_2')^{-1}C_2C_1'(C_1C_1')^{-1} - S_1C_1'(C_1C_1')^{-1}.$$

(Notice that we do not need to find $\bar{C}_i = T_iC_i$ in order to carry out this step. If this has been done and we use $\bar{C}_i$ in place of C_i then it is easily seen that the two equations above are equivalent to (25) with $n = 0, 1, \cdots, n_1$.) If there exists no such L_1, stop and conclude that there is no solution. Otherwise go to step 2.

Step 2. Solve for L_2 the equation

$$(I - R_1R_2)L_2 = 0,$$

where L_2 is an $m_1 \times r$ matrix and where r is the dimension of the null space of $I - R_1R_2$. Set $u_0 = L_2\varphi(y_{12})$, where φ is any $r \times 1$ vector function of y_{12}, which satisfies $E[\varphi'\varphi] < +\infty$, $E[\varphi] = 0$, $E[y_{12}\varphi'] = 0$ (L_2 will equal $[c_{\bar{m}_1}, \cdots, c_{\bar{m}_r}]$, see (26)).

Step 3. Calculate the eigenvalues of R_1R_2, the μ_i's of Lemma 1, and consider T_iy_i in place of y_i in accordance with Lemma 1. Check whether for some nonnegative integers $m_1, \cdots, m_k$, not all zero $\mu_1^{m_1}, \cdots, \mu_k^{m_k}$ is the inverse of some eigenvalue of R_1R_2. This check will stop in a finite number of steps, since $\mu_1^{m_1} \cdots \mu_k^{m_k}$ goes to zero because at least one of the m_i's increases. Let $\mu_1^{m_1} \cdots \mu_k^{m_k}$ be an inverse eigenvalue of R_1R_2 with corresponding eigenvectors $c_{\bar{n}_1}, \cdots, c_{\bar{n}_f}$. If μ_i corresponds to the ith component of y_{112}, z_i (see Lemma 2 (5)) consider

$$u_d = \sum_{k=1}^{f} c_{\bar{n}_k}p_{m_1}(z_1) \cdots p_{m_k}(z_k)\varphi_{\bar{n}_k}(y_{12}),$$

and set

$$u_c = \sum_d u_d \quad \text{(finite sum)},$$

(u_c corresponds to the third term of (26)). Then the solution of (12) is $u_1 = u_L + u_0 + u_c$.

It should be noticed that the part of u_1 which depends on y_{111} is determined uniquely and is linear in y_{111}, because every $\bar{p}_n$ which depends only on y_{111} has $a_n = 0$, and thus the corresponding c_n is either equal to s_n or 0.

Let us now consider the impact of the possible nonunique solutions for (u_1, u_2) to the costs J_1, J_2 of the Nash game. Let $Q_{ij} = 0$, $F_i = 0$, $i \neq j$. If (u_1, u_2) is a solution, then using (9), we obtain

$$E[u_1' R_1 u_2] = E[P_1(u_1' R_1 u_2)] = E[u_1' R_1 P_1 u_2]$$
$$= E[u_1'(-u_1 - S_1 P_1 x)] = -E[u_1' u_1] - E[u_1' S_1 x].$$

Thus,

$$J_1^*(u_1, u_2) = -\tfrac{1}{2} E[u_1' u_1],$$

and from (26)

$$J_1^*(u_1, u_2) = -\tfrac{1}{2}\{\|c_0\|^2 + \|c_1\|^2 + \cdots + \|c_{n_1}\|^2$$
$$+ \|c_{\bar{n}_1}\|^2 + \cdots + \|c_{\bar{n}_q}\|^2 + E[\phi' C' \cdot C\phi]\}, \quad \text{where } C = [c_{\bar{m}_1} \cdots c_{\bar{m}_r}].$$

If Q_{12}, F_1 are not zero, we can again calculate J_1^*. J_1^* will have some more quadratic terms in the c_i's, but it will also include the term $E[\phi' C' R_2' Q_{12} R_2 C \phi]$. It is clear that in the case of nonuniqueness of solutions, by choosing the c_n's, and ϕ appropriately, we can vary J_1^*, J_2^*. If Q_{12} and F_1 are not set equal to zero, but are chosen so as to convexify J_1^* as a function of the c_i's and ϕ, then the possibility of arbitrarily small J_1^* will be ruled out. It will then necessarily hold that $C'(-I + 2R_2' Q_{12} R_2) C \geqq 0$ (i.e., $-I + 2R_2' Q_{12} R_2 \geqq 0$ on the null space of $I - R_1 R_2$) and thus player 1 will choose $\phi = 0$. Thus, if $C'(-I + 2R_2' Q_{12} R_2) C \geqq 0$, a second level deterministic problem can be introduced where player 1 will determine the c_i's to find the best (for himself) out of the many Nash solutions. Of course additional restrictions will have to be imposed in order to guarantee the convexity of the second level deterministic problem.

It should also be noticed that Theorem 1 reveals the dependence of the solution of the Nash game upon the relation between the matrices which determine the information and the matrices which determine the cost. Obviously only C_1, C_2, $R_1 R_2$ have an impact as far as it concerns the existence and uniqueness of the solution. Nonetheless, the matrices Q_{ij}, F_i have an influence in the choice of the best out of the many Nash solutions when the second level optimization problem is solved and this influence can be very drastic (see § 6).

In Appendix C we present some useful results concerning the operator $I - R_1 R_2 P_1 P_2$, which is obviously of central importance in our analysis.

4. Solution of problem S. The purpose of this section is to solve Problem S. Our development will be brief in view of the analysis of § 3.

For a given u_1 the follower solves problem (10), finds u_2 in terms of u_1 (see (11)), and u_2 is substituted in J_1 so that the leader has to solve

$$\underset{u_1 \in U_1}{\text{minimize}}\, J(u_1) = E[\tfrac{1}{2} u_1'(I - (R_1 R_2 + R_2' R_1' + R_2' Q_{12} R_2) P_2) u_1$$
$$+ u_1'(S_1 + P_2(-R_1 S_2 + R_2' Q_{12} S_2 - R_2 F_1)) x \tag{27}$$
$$+ x'(\tfrac{1}{2} S_2' Q_{12} S_2 - S_2' F_1) P_2 x].$$

To guarantee $\inf_{u_1} J(u_1) > -\infty$, we assume that J_1 is convex in u_1 for $u_1 = P_1 u_1$, i.e., we assume that

$$P_1 \geqq R P_1 P_2 P_1, \tag{28}$$

where

$$R = R_1 R_2 + R_2' R_1' + R_2' Q_{12} R_2. \tag{29}$$

If (28) holds, then u_1 is a solution if and only if

$$u_1 - RP_1P_2u_1 + P_1(S_1 + P_2(-R_1S_2 + R_2'Q_{12}S_2 - R_2F_1))x = 0, \tag{30}$$

which is of exactly the same type as (12) and thus all the analysis of § 3 carries over. The only question pertaining to Problem S in particular is present in assumption (28). Using (21) and Lemma 2 we obtain the following relation equivalent to (28):

$$\sum_{n=0}^{\infty} \|c_n\|^2 \geqq \sum_{n=0}^{\infty} c_n'(a_nR)c_n, \quad \forall c_0, c_1, \cdots, \quad \text{with } \sum_{n=0}^{\infty} \|c_n\|^2 < +\infty,$$

or

$$a_nR \leqq I, \qquad n = 0, 1, 2, \cdots, \tag{31}$$

and since $a_0 = 1$, we conclude that (28) is equivalent to

$$R \leqq I. \tag{32}$$

(32) could have been deduced directly from (28) by allowing u = constant, but (31) can be useful if we decide to restrict u. For example, if u is restricted to being a nonlinear function of y_{112}, only, then using (B-5) and (37), a weaker condition which will involve the μ_i's, can be substituted for (32).

In light of the discussion above and the analysis of § 3 we can easily conclude the following concerning the leader-follower game:

(i) If $R < I$, then there is a unique solution and it is linear in the information (since if $R < I$ no inverse eigenvalue of R can be equal to some $\mu_1^{m_1} \cdots \mu_k^{m_k}$).

(ii) If there exists a solution there will exist a solution affine in the information.

(iii) If R has some eigenvalue equal to 1 then the general solution of (30), if it exists, will be of the form $l_0 + L_1y_1 + L_2\varphi(y_{12})$, where $l_0 + L_1y_1$ is a solution and the columns of L_2 constitute a basis for the null space of $I - R$.

(iv) If (32) does not hold then $\inf J_1 = -\infty$ (since if $u = c \in R^{m_1}$, where $(I - R)c = \lambda c$, $\lambda < 0$, then $J_1 \to -\infty$ as $\|c\| \to +\infty$).

Solving the leader-follower game is less demanding than solving the Nash game since the calculation of $\mu_1, \cdots, \mu_k$ is not necessary. Of course, one needs to find y_{12} if $I - R$ is singular.

It is clear that the discussion in § 3 about different values of the cost induced by different solutions, convexification and a second level game, carry over to the leader-follower game as well.

5. A sufficient condition, extensions to dynamic cases. It is an immediate consequence of the analysis of § 3 that, if R_1R_2 has no eigenvalues in $[1, +\infty)$, then the Nash problem admits a unique solution which will be linear in the information. This can be proved independently as a consequence of Theorem 2 below. Theorem 2 is related to [5, Thm. 1], which is a special case of our Theorem 2. In addition our proof of Theorem 2 is much simpler than the one given in [5].

THEOREM 2. *Let H be a Hilbert space over the complex numbers and P an orthogonal projection in H. Let $Q : H \to H$ be a continuous linear operator (P and Q do not necessarily commute) and v an element of H. Then, a sufficient condition that the equation*

$$PQu + Pv = 0, \qquad Pu = u \tag{33}$$

have a unique solution $u \in H$ is that there exist a continuous linear invertible operator

$E: H \to H$ which commutes with P, and that the following holds:

$$QE^* + EQ^* \geqq I \quad \text{on } PH. \tag{34}$$

If (34) *holds, then the solution of* (33) *is given by*

$$u = P \sum_{n=0}^{\infty} [(I - \bar{E}^{-1}Q)P]^n \bar{E}^{-1} v, \tag{35}$$

where $\bar{E} = \delta E$ for any $\delta > \|QQ^*\|$.

Proof. The requirement that u solve (33) is equivalent to the requirement that u solve

$$PQu - PEu + Eu + Pv = 0 \tag{36}$$

for some E, as in the statement of the theorem. Obviously, (33) implies (36). Conversely, if (36) holds, applying P to both sides of (36) yields

$$PQu + Pv = 0, \tag{37}$$

and (36) together with (37) implies $-PEu + Eu = 0$, i.e., $Pu = u$. (36) can be written as

$$[I - P[I - E^{-1}Q]] + u + PE^{-1}v = 0. \tag{38}$$

A sufficient condition that (38) have a unique solution given by (35) is that

$$\|P[I - E^{-1}Q]\| < 1,$$

or equivalently

$$P[E^{-1}Q - I][Q^*E^{-1^*} - I]P \leqq (1 - \varepsilon)I \quad \text{on } H, \tag{39}$$

for some ε, $0 < \varepsilon < 1$. Taking into consideration the fact that we can multiply E by any $\delta > 0$, we can easily conclude that (39) is equivalent to (34). □

Condition (34) holds if $Q = Q' > 0$, is a real matrix, if we choose $E = \varepsilon I$, where ε is sufficiently large and positive. This special case was proved in [5] by more complicated arguments.

To apply Theorem 2 to the Nash game we first bring R_1R_2 into its Jordan form, $TR_1R_2T^{-1} = J$ and let $u = Tu_1$, $v = TSy_1$. If u_i is a component of u and ρ_i an eigenvalue of R_1R_2, it suffices to be able to invert the operator $1 - \rho_i P_1 P_2$. The role of Q in Theorem 2 will be now played by $1 - \rho_i P_2$. Taking E to be any complex number $\neq 0$ and using (34), we conclude that if $\rho_i \notin [1, +\infty)$ then the solution of (12) exists, is unique and linear in the information. (It should be pointed out that J does not need to be diagonal: if a 2×2 block of J involves the eigenvalue ρ and has a 1 in the upper right corner, then we first invert the $1 - \rho P_1 P_2$ associated with the component of u, u_i corresponding to the bottom row of this block and move to the above row in order to solve for the other component of u, u_{i-1}; the 1 of the Jordan block multiplies u_i which is already known, and so we have to invert $1 - \rho P_1 P_2$ again.)

Another application of Theorem 2 is in the study of equations of the form

$$PQu + Pv = 0, \qquad Pu = u, \tag{40}$$

where $H = H_1 \oplus H_2 \oplus \cdots \oplus H_n$, $P = \operatorname{diag}[P_1, \cdots, P_n]$, Q is a real matrix, $v \in H$ and P_i is the projection of H onto H_i. Such an equation will appear if we consider the n-player Nash game instead of the 2-player game of § 2. It will also appear in the study of dynamic linear quadratic Nash games with noisy linear state measurements, a discrete time evolution equation and appropriate nestedness conditions (see [4]) on the information of the players. Application of Theorem 2 to (40) yields that if there exists a real

matrix $E = \text{diag}\,[E_1, \cdots, E_n]$, $E_i : H_i \to H_i$, E_i, with

$$EQ^* + QE^* > 0, \tag{41}$$

then (40) admits a unique solution. Of course, if the H_i's admit Hermite polynomials as a complete set of orthonormal eigenvectors, one can follow a procedure identical to the one of § 3, but the formulae derived will be quite complicated. Finally notice that if $n = 2$ and

$$Q = \begin{bmatrix} I & R_1 \\ R_2 & I \end{bmatrix},$$

then (40) represents another way of writing (9) and (11). Application of the condition (41) is possible, but the result will be weaker than the one derived by first transforming (12) into its Jordan form and then applying Theorem 2.

In the rest of this section we will show how some dynamic problems can be reduced to static ones (see also [4]), and how one can solve them. Let

$$x_{k+1} = A_k x_k + B_k^1 u_k^1 + B_k^2 u_k^2 + w_k, \qquad k = 0, 1, \cdots, N,$$

$$y_k^i = C_k^i x_k + v_k^i, \qquad i = 1, 2$$

$$J_i(u_1, u_2) = E\left[x_{N+1}' Q_{N+1}^i x_{N+1} + \sum_{k=0}^{N} x_k' Q_k^i x_k + u_k^{i'} u_k^i + u_k^{i'} R_k^{ij} u_k^j \right], \qquad i, j = 1, 2, \quad i \neq j.$$

x_0, w_k, v_k^i are independent Gaussian random variables with nonsingular covariance matrices. The real matrices $Q_k^i = Q_k^{i'} \geqq 0$, R_k^{ij}, A_k, B_k^i, C_k^i have appropriate dimensions, $x_k \in R^n$, $u_k^i \in R^{n_i}$, $y_k^i \in R^{m_i}$ and $u_i = (u_0^i, u_1^i, \cdots, u_k^i)'$. Player 1 chooses u_k^1 as a function of $(y_0^1, y_1^1, \cdots, y_k^1, y_0^2, \cdots, y_{k-1}^2)$, and player 2 chooses u_k^2 as a function of $(y_0^1, \cdots, y_{k-1}^1, y_0^2, \cdots, y_k^2)$. Using the evolution equation, we can express the J_i's as quadratic functions of x_0, w_k, v_k^i, u_k^1, u_k^2 and the y_k^1's as functions of x_0, $w_0, \cdots, w_{k-1}$, $v_0^1, \cdots, v_k^1$, $v_0^2, \cdots, v_{k-1}^2$ and $u_0^1, \cdots, u_{k-1}^1$, $u_0^2, \cdots, u_{k-1}^2$. Because of the nestedness of the information we can do away with the presence of the u_l^i's $0 \leqq l \leqq k-1$, in the expression for y_k^1, and similarly for y_k^2. Let $x = (x_0, w_0, \cdots, w_N, v_0^1, \cdots, v_N^1, v_0^2, \cdots, v_N^2)$. We have thus transformed our problem into the following:

$$J_1(u_1, u_2) = E[u_1' Q_{11} u_1 + u_1' R_{12} u_2 + u_1' S_1 x + u_2' Q_{12} u_2 + u_2' F_1 x + x' L_1 x],$$

$$J_2(u_1, u_2) = E[u_2' Q_{22} u_2 + u_2' R_{21} u_1 + u_2' S_2 x + u_1' Q_{21} u_1 + u_1' F_2 x + x' L_2 x],$$

$$y_{10} = C_{10} x, \qquad y_{11} = C_{11} x, \cdots, y_{1N} = C_{1N} x,$$

$$y_{20} = C_{20} x, \qquad y_{21} = C_{21} x, \cdots, y_{2N} = C_{2N} x.$$

Let

$$P_{1k} = E[\cdot \,|\, y_0^1, \cdots, y_k^1, y_0^2, \cdots, y_{(k-1)}^2],$$

$$P_{2k} = E[\cdot \,|\, y_0^1, \cdots, y_{(k-1)}^1, y_0^2, \cdots, y_k^2],$$

$$P_1 = \text{diag}\,[P_{10}, \cdots, P_{1N}], \qquad P_2 = \text{diag}\,[P_{20}, \cdots, P_{2N}].$$

($P_{1k} P_{2l} = P_{2l}$, $P_{2k} P_{1l} = P_{1l}$ if $l \leqq k-1$ and $P_{1k} P_{1l} = P_{1l}$, $P_{2k} P_{2l} = P_{2l}$ if $l \leqq k$.) If we are interested in the Nash solution we can write down the analogues of (9) and (11), which can be viewed together, as an equation of the type (40). We can thus either apply a generalization of the analysis of § 3 or settle for less and use Theorem 2. (If we are interested in the leader-follower solution, we have to assume that $y_k^1 = 0$ for $k \geqq 1$.)

6. Example of a Nash problem. In this section we will solve a one-dimensional Nash problem. Let

$$y_1 = x_1, \qquad y_2 = x_1 + ax_2.$$

x_1, x_2 are normally distributed, independent Gaussian random variables and $a \neq 0$. Let θ be the absolute value of the correlation coefficient of y_1, y_2, $\theta = (1+a^2)^{-1}$, $0 < \theta < 1$,

$$J_1(u_1, u_2) = E[\tfrac{1}{2}u_1^2 + r_1u_1u_2 + u_1(s_{11}x_1 + s_{12}x_2) + \tfrac{1}{2}q_1u_2^2 + u_2(t_{11}x_1 + t_{12}x_2)],$$
$$J_2(u_1, u_2) = E[\tfrac{1}{2}u_2^2 + r_2u_1u_2 + u_2(s_{21}x_1 + s_{22}x_2) + \tfrac{1}{2}q_2u_1^2 + u_2(t_{21}x_1 + t_{22}x_2)],$$

where r_i, s_{ij}, q_i are reals. (12) assumes the form

$$u_1 - \rho E[E[u_1 | x_1 + ax_2] | x_1] = sy_1, \tag{42}$$

where

$$\rho = r_1r_2, \qquad s = r_1(s_{21} + as_{22}) - s_{11}.$$

Let $u = \sum_{n=0}^{\infty} c_n p_n(y_1)$, where the p_n are the one-dimensional Hermite polynomials (see [9]), and $\sum c_n^2 < +\infty$. A straightforward application of Theorem 1 yields the following. Consider the equations for the c_n's,

$$c_n(1 - \rho\theta^n) = \begin{cases} s & \text{if } n = 1, \\ 0 & \text{if } n = 0, 2, 3, \cdots. \end{cases}$$

If:

(i) $1 \neq \rho\theta^n$, $n = 0, 1, 2, \cdots$, then the solution exists, is unique and is given by $u_1(y_1) = s(1 - \rho\theta)^{-1}y_1$.

(ii) $1 \neq \rho\theta$, but $\rho\theta^n = 1$ for some $n = 0, 2, 3, \cdots$, then the solution is $u_1(y_1) = s(1 - \rho\theta)^{-1}y_1 + cp_n(y_1)$, c arbitrary and real.

(iii) $1 = \rho\theta$ and $s = 0$, then the solution is $u_1 = (y_1) = ly_1$, l is arbitrary and real.

(iv) $1 = \rho\theta$ and $s \neq 0$, then there is no solution.

If $1 = \rho\theta^n$ for some $n \geqq 2$, then case (ii) holds and an easy calculation shows that

$$J_1 = \tfrac{1}{2}c^2[-1 + q_1\theta^n r_2^2] + \text{constant},$$

since $\theta^n = 1/(r_1r_2) > 0$. We conclude that if $r_1/r_2 > q_1$ player 1 can make his cost arbitrarily small for sufficiently large c. If $r_1/r_2 < q_1$ he will do well to choose $c = 0$. If both $r_1/r_2 < q_1$ and $r_2/r_1 < q_2$, hold then both players will agree on $c = 0$ (or on c sufficiently small if $r_1/r_2 > q_1$ and $r_2/r_1 > q_2$). If $r_1/r_2 = q_1$, then player 1 does not care about c. Conflict will arise about the choice of c if $r_1/r_2 > q_1$ and $r_2/r_1 < q_2$, in which case player 1 will want c as big as possible whereas player 2 will want $c = 0$. If player 1 is faster than 2, he calculates his u_1 through (42) first, realizes the possibility of choosing c arbitrarily and by declaring his decision he forces player 2 to use (11) to find his decision and thus player 1 imposes his choice of c. Therefore, the case of nonunique Nash solutions carries hidden in it the concept of the leader-follower game. Finally, notice that if J_1 is convex in u_1, u_2, i.e., $q_i \geqq r_1^2$, since $r_1r_2 = 1/\theta^n > 1$, then we obtain $1/r_2 < r_1$ and thus $r_1/r_2 < r_1^2 \leqq q_1$, i.e., $q_1 \geqq r_1/r_2$; therefore player 1 will prefer $c = 0$, in agreement with the fact that the convexity of J_1 cannot permit it to go to $-\infty$. Nonetheless, it might very well be that $r_1/r_2 < q_1 < r_1^2$ in which case player 1 will again prefer $c = 0$, although J_1 is not convex in u_1 and u_2, i.e., he cannot make J_1 arbitrarily small although J_1 is not convex in (u_1, u_2). This situation is due to the fact that what matters is the convexity of J_1 in $u_1 = P_1u_1$, $u_2 = P_2u_2$, i.e., convexity on some subspace and this convexity is guaranteed by $q_1 > r_1/r_2$.

The condition $1 = \rho\theta^n$, i.e., θ = correlation coefficient of y_1, $y_2 = (r_1 r_2)^{-n}$, is critical. $r_1 r_2$ can be interpreted as the coupling of J_1, J_2, whereas θ is the coupling of the information. We can thus interpret the condition $1 = \rho\theta^n$, as saying that if the coupling of the information equals the inverse of some power $n \neq 1$ of the coupling of the costs, the solution will be nonunique.

7. Conclusions. Here we will point out several directions in which the analysis presented can be generalized, or problems which suggest themselves for study within our framework.

The second level problems that have to be solved in the case of nonunique solutions, as discussed at the end of §§ 3 and 4 are of definite importance. In the Nash case, J_1^* is a quadratic function of the c_i's (we set $\phi = 0$) and the constraints on the c_i's are finite, since the c_i's involved are finite in number. Thus player 1 is faced with a classical quadratic deterministic optimization problem subject to linear constraints. Although it is an easy problem, it merits special attention because it will provide the best Nash decision to player 1.

To generalize our analysis to the many player case one needs to extend Lemma 1 and Lemma 2. One can go one step further and allow different components of u_1 to have different information or even more, one can study equations of the form $PQu + Pv = 0$, where $P = \text{diag}\,[P_1, \cdots, P_n]$, $\bar{P}u = u$, $\bar{P} = \text{diag}\,[\bar{P}_1, \cdots, \bar{P}_k]$, where P_i, $\bar{P}_i$ are projections. Such extensions are important in order to be able to handle dynamic games with nested information structures (although conceptually they are covered by the methods presented here).

Another interesting problem whose study lies within the capabilities of the methods presented, is the one where one leader is followed by two followers, which followers play Nash.

Appendix A: Proof of Lemma 1. Let $R(C)$ denote the range of a matrix C. All bases to be mentioned are orthonormal. Let the rows of $\bar{C}_{12} = \bar{C}_{21}$ be a basis for $R(C_1') \cap R(C_2')$. Let the rows of $\bar{C}_{111}$ be a basis for $R(C_1') \cap R(C_2')^{\perp}$. Let the rows of $\bar{C}_{222}$ be a basis for $R(C_1')^{\perp} \cap R(C_1')$. Choose $\bar{C}_{112}$ so that its rows together with those of $\bar{C}_{111}$ and $\bar{C}_{12}$ constitute a basis for $R(C_1')$. Choose $\bar{C}_{221}$ so that its rows together with those of $\bar{C}_{222}$ and $\bar{C}_{21}$ constitute a basis for $R(C_2')$. This construction proves (1), (2) and (4). Let us concentrate on $\bar{C}_{112}$, $\bar{C}_{221}$, which we will denote by D_1, D_2. If D_1 is $k_1 \times n$, D_2 is $k_2 \times n$ and $k_1 \neq k_2$, let $k_1 > k_2$, without loss of generality. Then there are nonsingular square matrices L_1, L_2 so that $L_1 D_1 D_2' L_2'$ will have its last row equal to zero, which means that the last row of $L_1 D_1$ is an element of $R(D_1')$ perpendicular to $R(D_2')$. Such elements, nonetheless, were put in $R(\bar{C}_{111}')$ and thus k_1 cannot be strictly greater than k_2. Reversing the roles of k_1 and k_2 we conclude that $k_1 = k_2 = k$ and that $D_1 D_2'$ is a square nonsingular matrix. Let Λ be the diagonal Jordan equivalent of $D_1 D_1'$, i.e.,

$$(D_1 D_1')U = U\Lambda,$$

where U is the matrix of the orthogonal eigenvectors. Let M be the diagonal Jordan equivalent for which

$$(\Lambda^{-1/2} U' D_1 D_2' (D_2 D_2')^{-1} D_2 D_1' U \Lambda^{-1/2}) V = VM,$$

where V is the matrix of the orthogonal eigenvectors. Let

$$\bar{D}_1 = V' \Lambda^{-1/2} U' D_1,$$

$$\bar{D}_2 = M^{-1/2} V' \Lambda^{-1/2} U' D_1 D_2' (D_2 D_2')^{-1} D_2.$$

It can be verified that

$$\bar{D}_1\bar{D}_1' = I, \qquad \bar{D}_2\bar{D}_2' = I, \qquad \bar{D}_1\bar{D}_2' = M^{1/2},$$

so we can use $\bar{D}_1, \bar{D}_2$ for $\bar{C}_{112}, \bar{C}_{221}$. $R(\bar{D}_1)$ and $R(\bar{D}_2)$ have no common elements, since if they had one, it should have been placed in $R(\bar{C}_{12}')$ from the beginning. M is diagonal has positive elements, and each $\sqrt{\mu_{ii}}$ is the product of two nonidentical unit length vectors (rows of $\bar{D}_1$, $\bar{D}_2$). Thus $0 < \sqrt{\mu_{ii}} < 1$. □

Appendix B: Proof of Lemma 2.

$$a_{nl} = E[\bar{p}_n P_1 P_2 \bar{p}_l] = E[\bar{p}_n P_2 \bar{p}_l] = E[(P_2\bar{p}_n)\cdot \bar{p}_l] = a_{ln},$$

since P_i is self-adjoint and $P_1\bar{p}_n = \bar{p}_n$. $\bar{p}_n$ and $\bar{p}_l$ have the form

$$\bar{p}_n(y_1) = \bar{p}_{n_1}(y_{111})\bar{p}_{n_2}(y_{112})\bar{p}_{n_3}(y_{12}),$$
$$\bar{p}_l(y_1) = \bar{p}_{l_1}(y_{111})\bar{p}_{l_2}(y_{112})\bar{p}_{l_3}(y_{12}).$$

Using the independence of some of the components of y_1, y_2, we have

$$\begin{aligned}
a_{nl} &= E[\bar{p}_n(y_1)P_1P_2\bar{p}_l(y_1)] \\
&= E[\bar{p}_n(y_1)P_2\bar{p}_l(y_1)] \\
&= E[\bar{p}_n(y_1)E[\bar{p}_{l_1}(y_{111})\bar{p}_{l_2}(y_{112})\bar{p}_{l_3}(y_{12})\,|\,y_{12}, y_{221}, y_{222}]] \\
&= E[\bar{p}_n(y_1)\bar{p}_{l_3}(y_{12})E[\bar{p}_{l_1}(y_{111})\bar{p}_{l_2}(y_{112})\,|\,y_{12}, y_{221}, y_{222}]] \\
&= E[\bar{p}_n(y_1)\bar{p}_{l_3}(y_{12})E[\bar{p}_{l_1}(y_{111})\bar{p}_{l_2}(y_{112})\,|\,y_{221}]] \\
&= E[\bar{p}_n(y_1)\bar{p}_{l_3}(y_{12})E[E[\bar{p}_{l_1}(y_{111})\bar{p}_{l_2}(y_{112})\,|\,y_{221}, y_{112}]\,|\,y_{221}]] \qquad \text{(B.1)} \\
&= E[\bar{p}_n(y_1)\bar{p}_{l_3}(y_{12})E[\bar{p}_{l_2}(y_{112})E[\bar{p}_{l_1}(y_{111})\,|\,y_{221}, y_{112}]\,|\,y_{221}]] \\
&= E[\bar{p}_n(y_1)\bar{p}_{l_3}(y_{12})E[\bar{p}_{l_2}(y_{112})E[\bar{p}_{l_1}(y_{111})]\,|\,y_{221}]] \\
&= E[\bar{p}_n(y_1)\cdot\bar{p}_{l_3}(y_{12})E[p_{l_1}(y_{111})]\cdot E[\bar{p}_{l_2}(y_{112})\,|\,y_{221}]] \\
&= E[\bar{p}_{n_1}(y_{111})\bar{p}_{n_2}(y_{112})\bar{p}_{n_3}(y_{12})\bar{p}_{l_3}(y_{12})E[\bar{p}_{l_1}(y_{111})]\cdot E[\bar{p}_{l_2}(y_{112})\,|\,y_{221}]] \\
&= E[\bar{p}_{n_1}(y_{111})]\cdot E[\bar{p}_{l_2}(y_{111})]\cdot E[\bar{p}_{n_3}(y_{12})\bar{p}_{l_3}(y_{12})]\cdot E[\bar{p}_{n_2}(y_{112})E[\bar{p}_{l_2}(y_{112})\,|\,y_{221}]].
\end{aligned}$$

It holds that

$$E[\bar{p}_{n_1}(y_{111})] = \begin{cases} 0 & \text{if } n_1 \neq 0, \\ 1 & \text{if } n_1 = 0, \end{cases}$$

(B.2)

$$E[\bar{p}_{n_3}(y_{12})\bar{p}_{l_3}(y_{12})] = \begin{cases} 0 & \text{if } n_3 \neq l_3, \\ 1 & \text{if } n_3 = l_3. \end{cases}$$

Let

$$\bar{p}_{n_2}(y_{112}) = p_{m_1}(z_1)\cdots p_{m_k}(z_k),$$
$$\bar{p}_{l_2}(y_{112}) = p_{s_1}(z_1)\cdots p_{s_k}(z_k),$$

where $y_{112} = (z_1, \cdots, z_k)'$. Let $y_{221} = (w_1, \cdots, w_k)'$.

Since z_i depends only on w_i and vice versa (Lemma 1, (3)) we have

$$\begin{aligned}
&E[p_{s_1}(z_1)\cdots p_{s_k}(z_k)\,|\,w_1,\cdots,w_k]\\
&\quad=E[E[p_{s_1}(z_1)\cdots p_{s_k}(z_k)\,|\,z_1,w_1,\cdots,w_k]\,|\,w_1,\cdots,w_k]\\
&\quad=E[p_{s_1}(z_1)E[p_{s_2}(z_2)\cdots p_{s_k}(z_k)\,|\,z_1,w_1,\cdots,w_k]\,|\,w_1,\cdots,w_k]\\
&\quad=E[p_{s_1}(z_1)\,|\,w_1]\cdot E[p_{s_2}(z_2)\cdots p_{s_k}(z_k)\,|\,w_2,\cdots,w_k]\\
&\quad=\cdots=E[p_{s_1}(z_1)\,|\,w_1]\cdot E[p_{s_2}(z_2)\,|\,w_2]\cdots E[p_{s_k}(z_k)\,|\,w_k].
\end{aligned}$$

Therefore,

$$\begin{aligned}
&E[\bar p_{n_2}(y_{112})E[\bar p_{l_2}(y_{112})\,|\,y_{221}]]\\
&\quad=E[p_{m_1}(z_1)\cdots p_{m_k}(z_k)E[p_{s_1}(z_1)\,|\,w_1]\cdots E[p_{s_k}(z_k)\,|\,w_k]]\\
&\quad=E[p_{m_1}(z_1)E[p_{s_1}(z_1)|w_1]]\cdots E[p_{m_k}(z_k)E[p_{s_k}(z_k)|w_k]].
\end{aligned}\tag{B.3}$$

Since $E[z_1w_1]=\sqrt{\mu_1}$, $E[p_{s_1}(z_1)\,|\,w_1]$ will be a polynomial of order s_1 in w_1 and the leading coefficient of this polynomial will be the leading coefficient of $p_{s_1}(z_1)$ multiplied by $(\sqrt{\mu_1})^{s_1}$. Similarly, $E[E[p_{s_1}(z_1)\,|\,w_1]\,|\,z_1]$ will be a polynomial in z_1 of order s_1, with the leading coefficient of p_{s_1} multiplied by $(\sqrt{\mu_1})^{2s_1}$. Thus $E[E[p_{s_1}(z_1)\,|\,w_1]\,|\,z_1]=\mu_1^{s_1}p_{s_1}(z_1)+$Hermite polynomials in z_1 of order strictly less than s_1. Thus we conclude that

$$\begin{aligned}
E[p_{m_1}(z_1)\cdot E[p_{s_1}(z_1)\,|\,w_1]]&=E[p_{m_1}(z_1)E[E[p_{s_1}(z_1)\,|\,w_1]\,|\,z_1]]\\
&=\begin{cases}0 & \text{if } s_1<m_1,\\ \mu^{m_1} & \text{if } m_1=s_1.\end{cases}
\end{aligned}$$

Since $E[p_{m_1}(z_1)E[p_{s_1}(z_1)\,|\,w_1]]=E[p_{s_1}(z_1)E[p_{m_1}(z_1)\,|\,w_1]]$, we conclude that

$$E[p_{m_1}(z_1)E[p_{s_1}(z_1)\,|\,w_1]]=\begin{cases}0 & \text{if } m_1\neq s_1,\\ \mu_1 & \text{if } m_1=s_1.\end{cases}\tag{B.4}$$

From (B.3), (B.4) we now obtain

$$E[p_{n_2}(y_{112})E[p_{l_2}(y_{112})\,|\,y_{221}]]=\begin{cases}0 & \text{if } n_2\neq l_2,\\ \mu_1^{m_1}\cdots\mu_k^{m_k} & \text{if } n_2=l_2 \text{ and}\\ & p_{n_2}(y_{112})=p_{m_1}(z_1)\cdots p_{m_k}(z_k).\end{cases}\tag{B.5}$$

Equations (B.1), (B.2) and (B.5) prove (2)–(5).

Let us now prove (6). To find $\|P_{112}P_{221}\|$ we will calculate $P_{112}P_{221}u$. u can be restricted to depend only on y_{112} and thus $u=\sum_{n=1}^{\infty}c_n\bar p_n(y_{112})$, where $\sum_{n=1}^{\infty}\|c_n\|^2<\infty$, ($c_0=0$ so that $E[u]=0$).

$$\begin{aligned}
\|P_{112}P_{221}u\|&=E[(P_{112}P_{221}u)'\cdot P_{112}P_{221}u]\\
&=E[u'\cdot P_{221}P_{112}P_{221}u]\\
&=\sum_{n,l\geqq 1}^{\infty}c_n'c_lE[\bar p_n(y_{112})P_{221}P_{112}P_{221}\bar p_l(y_{112})]\\
&=\sum_{n,l\geqq 1}c_n'c_l\bar a_{nl}.
\end{aligned}$$

An argument similar to the one used before shows that $\bar a_{nl}=0$ if $n\neq l$, and if $n=l$,

$$\bar a_{nn}=\bar a_n=\mu_1^{2m_1}\cdots\mu_k^{2m_k}=a_n^2.$$

Thus, $\|P_{112}P_{221}u\|^2 = \sum_{n=1}^{\infty} \|c_n\|^2 a_n^2$. Also,

$$\|u\|^2 = \sum_{n,l\geq 1} c_n' c_l E[\bar{p}_n \cdot p_l] = \sum_{n=1}^{\infty} \|c_n\|^2,$$

and therefore

$$\frac{\|P_{112}P_{221}u\|^2}{\|u\|^2} = \frac{\sum_{n=1}^{\infty} \|c_n\|^2 a_n^2}{\sum_{n=1}^{\infty} \|c_n\|^2}.$$

Obviously $\|P_{112}P_{221}\| = \sup a_n$, and since a_n decreases, because $0 < \mu_i < 1$, we conclude $\|P_{112}P_{221}\| = \max\{\mu_1, \cdots, \mu_k\}$ 1. For reasons of symmetry, $\|P_{221}P_{112}\| = \max\{\mu_1, \cdots, \mu_k\} = \|P_{112}P_{221}\|$. □

Appendix C: The operator $I - RP_1P_2$. From the analysis of § 3 it is obvious that if instead of having to solve (14) we had to solve

$$(I - RP_1P_2)u = v, \tag{C.1}$$

where $v \in U_1$ (and thus $v = \sum_{j=0}^{\infty} d_j \bar{p}_j$, $d_j \in R^{m_1}$, $\sum \|d_j\|^2 < +\infty$), we would end up with the equivalent system of linear equations

$$(I - a_i R)c_i = d_i, \qquad i = 0, 1, \cdots. \tag{C.2}$$

If (C.1) has a solution $c_0, c_1, \cdots$, with $c_j \in \mathbb{R}^{m_1}$ then

$$u = \sum_{j=1}^{\infty} c_j p_j, \qquad \sum_{j=1}^{\infty} \|c_j\|^2 < +\infty,$$

is a solution of (C.2). Therefore the R's for which $I - RP_1P_2$ is invertible are those which do not have any of the $1/a_n$'s (for $a_n \neq 0$) as eigenvalues.

To find $\|I - RP_1P_2\|$, let $u = \sum_{n=0}^{\infty} c_n \bar{p}_n$. Thus $\|u\|^2 = \sum_{n=0}^{\infty} c_n^2 < \infty$,

$$\begin{aligned}\|(I - RP_1P_2)u\|^2 &= E[u'u + u'R'RP_1P_2P_1u - 2u'RP_1P_2u] \\ &= \sum_{n=0}^{\infty} (\|c_n\|^2 + a_n^2 c_n' R'Rc_n - 2a_n c_n' Rc_n) \\ &= \sum_{n=0}^{\infty} c_n'(I - a_nR)'(I - a_nR)c_n \leqq \left(\sum_{n=0}^{\infty} \|c_n\|^2\right) \sup_n \|I - a_nR\|^2,\end{aligned}$$

and obviously $\|I - RP_1P_2\| = \sup_n \|I - a_nR\|$. If $(I - RP_1P_2)^{-1}$ exists, to find $\|(I - RP_1P_2)^{-1}\|$, let $v = \sum_{n=0}^{\infty} d_n \bar{p}_n$, $\|v\|^2 = \sum_{n=0}^{\infty} \|d_n\|^2 < +\infty$. Then $(I - RP_1P_2)^{-1}v = \sum_{n=0}^{\infty} (I - a_nR)^{-1} d_p \bar{p}_n = \sum_{n=0}^{\infty} c_n \bar{p}_n$, $\|(I - RP_1P_2)^{-1}v\|^2 = \sum_{n=0}^{\infty} \|C_n\|^2 = \sum_{n=0}^{\infty} d_n'(I - a_nR)^{-1\prime} \cdot (I - a_nR)^{-1} d_n$. Thus $\|(I - RP_1P_2)^{-1}\| = \sup_n \|(I - a_nR)^{-1}\|$. (It is easy to see that if $(I - a_nR)^{-1}$ exists for all a_n, then $\sup_n (\|(I - a_nR)^{-1}\|) < +\infty$.)

Let us formalize this discussion into a proposition.

PROPOSITION 1.

(1) spectrum $(RP_1P_2) = \{a_n r;\ r = \textit{eigenvalue of } R,\ n = 0, 1, 2, \cdots\}$.

(2) $\|I - RP_1P_2\| = \sup_n \|I - a_nR\|$.

(3) $(I - RP_1P_2)^{-1}$ *exists, if and only if* $1 \neq a_n r$ *for all n and then* $\|(I - RP_1P_2)^{-1}\| = \sup_n \|(I - a_nR)^{-1}\|$.

We can use (3) in the case where we have to solve for u the equation $(I - RP_1P_2)u + f(u) = v$, where $\|f(u) - f(\bar{u})\| \leqq L\|u - \bar{u}\|$ and $f: U_1 \to U_1$, $v \in U_1$. If $L < \inf_n [\|(I - a_nR)^{-1}\|^{-1}]$ the contraction mapping theorem is applicable and yields existence and uniqueness of a solution. Equations of this form can arise when the cost J_1 is nonlinear in u_1, u_2.

REFERENCES

[1] R. RADNER, *Team decision problems*, Ann. Math. Statis., 33 (1962), pp. 857–881.
[2] A. W. STARR AND Y. C. HO, *Nonzero-sum differential games*, J. Optim. Theory Appl., 3 (1969), pp. 184–206.
[3] J. B. CRUZ Jr., *Leader-follower strategies for multilevel systems*, IEEE Trans. Automat. Control, AC-23 (1978), pp. 244–255.
[4] Y. C. HO AND K. C. CHU, *Team decision theory and information structures in optimal control problems—part I*, IEEE Trans. Automat. Control, AC-17 (1972), pp. 15–22.
[5] J. F. RUDGE, *Series solutions to static team control problems*, Math. Oper. Res., 1 (1976), pp. 67–81.
[6] T. BASAR, *Equilibrium solutions in two-person quadratic decision problems with static information structures*, IEEE Trans. Automat. Control, AC-20 (1975), pp. 320–328.
[7] ———, *Decentralized multicriteria optimization of linear stochastic systems*, IEEE Trans. Automat. Control, AC-23 (1978), pp. 233–243.
[8] R. B. ASH, *Real Analysis and Probability*, Academic Press, New York, 1972.
[9] K. R. PARTHASARATHY, *Introduction to Probability and Measure*, Springer-Verlag, New York, 1977.
[10] P. R. HALMOS, *A Hilbert Space Problem Book*, Van Nostrand, Princeton, NJ, 1967.
[11] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, vol. I, Interscience Publishers, New York, 1953.
[12] T. W. ANDERSON, *An Introduction to Multivariate Statistical Analysis*, John Wiley, New York, 1958.
[13] Y. C. HO, P. B. LUH AND R. MURALIDHARAN, *Information structure: Stackelberg games and incentive controllability*, IEEE Trans. Automat. Control, (1981), to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 5, September 1981

0363-0129/81/1905-0008 $01.00/0

SUFFICIENT CONDITIONS FOR OBSERVATION-INNOVATION EQUIVALENCE IN WHITE GAUSSIAN CHANNELS WITH FEEDBACK*

KENKO UCHIDA†

Abstract. This paper deals with the problem of establishing conditions for the informational equivalence between observation processes and their innovation processes in white Gaussian channels with feedback. It is shown that the informational equivalence holds in either case that there is a nonzero time delay in the feedback loop or that a time integral operation is included in the feedback loop.

1. Introduction. The objective of this paper is to establish some sufficient conditions for the informational equivalence between observation processes and their innovation processes in white Gaussian channels with feedback.

Let the white Gaussian channel be described by $dy(t) = F(t)\,dt + dw(t)$, where the signal F is a stochastic process and the noise w is a Wiener process. Let us denote by $\hat{F}(t)$; the expected value of $F(t)$ given by the past of the observation up to time t, i.e., $\hat{F}(t) = E\{F(t) \mid y(s), 0 \leqq s \leqq t\}$. Thus, the innovation process corresponding to the observation y is $d\nu(t) = dy(t) - \hat{F}(t)\,dt$. It can be shown that, under weak conditions, ν is a Wiener process [6], [8]. The subject of this paper is the question whether the observation y and ν are informationally equivalent, i.e., whether y and ν generate the same families of σ-fields.

This question is called the "innovation problem", which was first posed by Frost [6]. Though it is now known that the answer to the general case is in the negative [3], various conditions for the positive answer have been reported. In the Gaussian case where F and w are jointly Gaussian, informational equivalence was proved in [1], [5], [9], [16]. Clark [4] obtained a positive answer for the case that F is not necessarily Gaussian, but independent of w and bounded. An extension of Clark's result was tried in [2]. The assumption of independence between F and w is adequate, indeed, for many problems related to one-way signal processing, but it is also true that this assumption excludes many problems of practical importance, e.g., multi-way communication and detection problems and estimation and control problems. Beneš [1] succeeded in eliminating this independence assumption in a control problem setup. Another result was obtained by Kallianpur [11].

In this paper our concern also lies in establishing informational equivalence without the independence assumption between F and w. We describe here the dependence between F and w explicitly as the presence of the feedback of the observations in the signal F similar to the model of [7]. First, it is shown that informational equivalence holds in the case that there is a nonzero time delay in the feedback loop. Secondly, we also establish the informational equivalence in the case that an integral operation with respect to time is included in the feedback loop. Examples of the control problem setup are presented for both cases. Finally, we discuss the case of additive type feedback.

2. Formulation and preliminaries. The model to be considered here is the white Gaussian channel with feedback, which is precisely written as

$$y(t) = \int_0^t F(s, x, y)\,ds + w(t), \qquad 0 \leqq t \leqq T, \tag{1}$$

* Received by the editors April 22, 1980, and in revised form February 12, 1981.
† Department of Electrical Engineering, Waseda University, Tokyo 160, Japan.

where $y(t)$ is the observation, $F(t, x, y)$ is the signal, x is the message and $w(t)$ is the noise. Let $(\Omega, \mathcal{B}, P)$ be an underlying probability space which carries x and the $w(t)$ process; x takes its value in a measurable space $(X, \mathcal{X})$, $w(t)$ is a Wiener process, and x and $w(t)$ are independent.

Let C be the space of continuous functions from $[0, T]$ to R, and $\mathcal{C}_t$ be the σ-field of C generated by $\{f(s), 0 \leqq s \leqq t, f \in C\}$. The signal function

$$F : [0, T] \times X \times C \to R$$

is a jointly measurable function with the following properties:

(A.1) $F(t, \cdot, \cdot)$ is $\mathcal{X} \otimes \mathcal{C}_t$-measurable for each $t \in [0, T]$.

(A.2) There exists a number $M > 0$ such that $|F(t, x, f)| \leqq M$ for all $(t, x, f) \in [0, T] \times X \times C$.

Assumption (A.1) means that the signal $F(t, x, y)$ consists of the message x and the past observation $\{y(s), 0 \leqq s \leqq t\}$.

Let $\mathcal{Y}_t = \sigma\{y(s), 0 \leqq s \leqq t\}$ denote the σ-field of Ω generated by the observation up to time t, and define $\hat{F}(t, y) = E\{F(t, x, y) \mid \mathcal{Y}_t\}$. Then, the innovation process is defined to be $\nu(t) = y(t) - \int_0^t \hat{F}(s, y)\, ds$. This gives another expression of the observation process

$$y(t) = \int_0^t \hat{F}(s, y)\, ds + \nu(t). \tag{2}$$

The problem is to determine whether $\mathcal{Y}_t = \mathcal{N}_t$ for all t, where $\mathcal{N}_t = \sigma\{\nu(s), 0 \leqq s \leqq t\}$. However, we know from the definition of $\nu(t)$ that $\nu(t)$ is a $\mathcal{Y}_t$-adapted process; i.e., $\mathcal{Y}_t \supset \mathcal{N}_t$ for each t. Therefore, in the following, we can focus our attention only to the question whether $\mathcal{Y}_t \subset \mathcal{N}_t$ for all t.

Our arguments are based on the Bayes formula for $\hat{F}(t, y)$. The following result is a special version of the Kallianpur–Striebel formula [10]. For the details of the proof, see [14].

LEMMA. *Let $y(t)$, $0 \leqq t \leqq T$ be a strong (i.e., $\mathcal{X} \vee \sigma\{w(s), 0 \leqq s \leqq t\}$-adapted) solution of the stochastic differential equation* (1). *Then,*

$$\hat{F}(t, y) = \frac{\int_X F(t, x, y)\alpha(t, x, y)\mu(dx)}{\int_X \alpha(t, x, y)\mu(dx)}, \tag{3}$$

where μ is the distribution function for x and $\alpha(t, x, y)$ is given by

$$\alpha(t, x, y) = \exp\left\{\int_0^t F(s, x, y)\, dy(s) - \frac{1}{2}\int_0^t |F(s, x, y)|^2\, ds\right\}.$$

3. Main results.

(a) *Delayed feedback.* Consider the case that there is a nonzero time delay $h > 0$ in the feedback loop. Specifically,

(A.3) $F(t, \cdot, \cdot)$ is $\mathcal{X} \otimes \mathcal{C}_{t-h}$-measurable for each $t \in [0, T]$, where $\mathcal{C}_s = \{\phi, C\}$, $-h \leqq s \leqq 0$.

Then, it is noted that in the interval $[0, h]$ there is no feedback loop and so the problem is the same as Clark's [4]. Under assumptions (A.2) and (A.3) the stochastic differential equation (1) has a unique, strong solution $y(t)$, $0 \leqq t \leqq T$. This is shown by dividing the interval $[0, T]$ into the subintervals $[0, h]$, $[h, 2h]$, $[2h, 3h]$, $\cdots$ and constructing the solution recursively on these subintervals.

THEOREM 1. *Under assumptions* (A.2) *and* (A.3), *it follows that $\mathcal{Y}_t = \mathcal{N}_t$ for all $t \in [0, T]$.*

Proof. The proof here is a modification of Clark's [4]. Let Z_t be progressively measurable with respect to $\mathcal{Y}_t$ and bounded as $|Z_t| \leqq M$. Writing

$$R_t^Z(x) = \int_0^t F\Big(s, x, \int_0^{\cdot} Z_r\, dr + \nu\Big)\, d\nu(s) - \frac{1}{2}\int_0^t \Big|F\Big(s, x, \int_0^{\cdot} Z_r\, dr + \nu\Big)\Big|^2 ds + \int_0^t F\Big(s, x, \int_0^{\cdot} Z_r\, dr + \nu\Big) Z_s\, ds,$$

set

$$T_t(Z) = \frac{\int_X F(t, x, \int_0^{\cdot} Z_s\, ds + \nu) \exp R_t^Z(x)\mu(dx)}{\int_X \exp R_t^Z(x)\mu(dx)}. \tag{4}$$

Then, since $\nu(t)$ is $\mathcal{Y}_t$-measurable, $T_t(Z)$ is $\mathcal{Y}_t$-measurable and has a progressively measurable version with respect to $\mathcal{Y}_t$. It is also shown that $|T_t(Z)| \leqq M$.

If we take $Z_t = \hat{F}_t$ ($\hat{F}_t = \hat{F}(t, y)$ for short) we obtain $\exp R_t^{\hat{F}}(x) = \alpha(t, x, y)$ by (2). Therefore, the Bayes formula (3) gives

$$\hat{F}_t = T_t(\hat{F}). \tag{5}$$

This means that $\hat{F}$ is a fixed point of T_t. In the following we will construct a sequence of $\mathcal{N}_t$-adapted processes converging to this fixed point $\hat{F}$.

Now consider $T_t(Z) - \hat{F}_t$, and rewrite it as

$$\begin{aligned} T_t(Z) - \hat{F}_t &= T_t(Z) - T_t(\hat{F}) \\ &= \Big[\int_X \exp R_t^Z(x)\mu(dx) \int_X \exp R_t^{\hat{F}}(x)\mu(dx)\Big]^{-1} \\ &\quad \times \int_X \int_X [\{F_t^Z(x) - F_t^{\hat{F}}(x)\} \exp R_t^Z(x) \exp R_t^{\hat{F}}(x') \\ &\qquad + F_t^{\hat{F}}(x)\{\exp R_t^Z(x) - \exp R_t^{\hat{F}}(x)\} \exp R_t^{\hat{F}}(x') \\ &\qquad + F_t^{\hat{F}}(x) \exp R_t^{\hat{F}}(x)\{\exp R_t^{\hat{F}}(x') - \exp R_t^Z(x')\}]\mu(dx)\mu(dx'), \end{aligned}$$

where for short we write

$$F_t^Z(x) = F\Big(t, x, \int_0^{\cdot} Z_s\, ds + \nu\Big),$$

$$F_t^{\hat{F}}(x) = F\Big(t, x, \int_0^{\cdot} \hat{F}_s\, ds + \nu\Big).$$

Using the inequality $|\exp \xi - \exp \zeta| \leqq \frac{1}{2}(\exp \xi + \exp \zeta)|\xi - \zeta|$, $\xi, \zeta \in R$ in the second and third terms of the numerator, and $|F_t^{\hat{F}}(x)| \leqq M$, we find

$$\begin{aligned} |T_t(Z) - \hat{F}_t| &\leqq \Big[\int_X \exp R_t^Z(x)\mu(dx)\Big]^{-1} \\ &\quad \times \Big[\int_X |F_t^Z(x) - F_t^{\hat{F}}(x)| \exp R_t^Z(x)\mu(dx) \\ &\qquad + M\int_X (\exp R_t^Z(x) + \exp R_t^{\hat{F}}(x))|R_t^Z(x) - R_t^{\hat{F}}(x)|\mu(dx)\Big], \end{aligned} \tag{6}$$

where

$$R_t^Z(x)-R_t^{\hat{F}}(x)=\int_0^t (F_s^Z(x)-F_s^{\hat{F}}(x))\,d\nu(s)$$

(7)

$$-\frac{1}{2}\int_0^t (|F_s^Z(x)|^2-|F_s^{\hat{F}}(x)|^2)\,ds+\int_0^t (F_s^Z(x)Z_s-F_s^{\hat{F}}(x)\hat{F}_s)\,ds.$$

Remember here that $F(t, x, y)$ is assumed to be dependent on the delayed feedback $\{y(s), 0\leqq s\leqq t-h\}$. Therefore, if Z_t is taken such that $Z_t=\hat{F}_t$ for all $t\in[0, k]$, $k\geqq 0$, it follows that

$$F_t^Z(x)=F_t^{\hat{F}}(x) \quad \text{for all } t\in[0, k+h], \tag{8}$$

and so (6) and (7) are reduced to

$$|T_t(Z)-\hat{F}_t|\leqq\left[\int_X \exp R_t^Z(x)\mu(dx)\right]^{-1}$$

(9)

$$\times M\int_X (\exp R_t^Z(x)+\exp R_t^{\hat{F}}(x))|R_t^Z(x)-R_t^{\hat{F}}(x)|\mu(dx),$$

$$R_t^Z(x)-R_t^{\hat{F}}(x)=\int_k^t F_s^{\hat{F}}(x)(Z_s-\hat{F}_s)\,ds \tag{10}$$

for all $t\in[k, k+h]$. Substituting (10) into (9) and using $|F_t^{\hat{F}}(x)|\leqq M$, we obtain

$$|T_t(Z)-\hat{F}_t|\leqq M^2\int_k^t |Z_s-\hat{F}_s|\,ds\left\{1+\frac{\int_X \exp R_t^{\hat{F}}(x)\mu(dx)}{\int_X \exp R_t^Z(x)\mu(dx)}\right\}.$$

Here, using $|F_t^{\hat{F}}(x)|\leqq M$, $|F_t^Z(x)|\leqq M$, $|\hat{F}_t|\leqq M$ and $|Z_t|\leqq M$, we have

$$\frac{\int_X \exp R_t^{\hat{F}}(x)\mu(dx)}{\int_X \exp R_t^Z(x)\mu(dx)}\leqq\frac{\exp\frac{3}{2}M^2t\cdot\int_X \exp\int_0^t F_s^{\hat{F}}(x)\,d\nu(s)\mu(dx)}{\exp(-\frac{3}{2}M^2t)\cdot\int_X \exp\int_0^t F_s^Z(x)\,d\nu(s)\mu(dx)}=\exp 3M^2t,$$

where the last equality follows from (8). Thus, for the particular Z_t such that $Z_t=\hat{F}_t$, $0\leqq t\leqq k$, we have

$$|T_t(Z)-\hat{F}_t|\leqq K\int_k^t |Z_s-\hat{F}_s|\,ds, \qquad K=M^2(1+\exp 3M^2T) \quad \text{for all } t\in[k, k+h]. \tag{11}$$

Now define the $\mathcal{N}_t$-adapted processes $\hat{F}_t^n$, $n=0, 1, 2, \cdots$ in the interval $[0, h]$ as follows. Set $\hat{F}_t^0\equiv 0$ and define

$$\hat{F}_t^n=T_t(\hat{F}^{n-1}), \qquad n=1, 2, \cdots \quad \text{for } t\in[0, h].$$

Then, setting $k=0$ and replacing Z by $\hat{F}^n$ in (11), we have from (11)

$$|\hat{F}_t^{n+1}-\hat{F}_t|\leqq M\frac{K^{n+1}}{(n+1)!}t^{n+1} \quad \text{for } t\in[0, h].$$

This implies that $\hat{F}_t^n$ converges to the fixed point $\hat{F}_t$ uniformly on the interval $[0, h]$, which is $\mathcal{N}_t$-measurable.

Next, using thus obtained fixed point $\hat{F}_t$, $0\leqq t\leqq h$, define the $\mathcal{N}_t$-adapted processes $\hat{F}_t^n$, $n=0, 1, 2, \cdots$ in the interval $[h, 2h]$ as follows. Set $\hat{F}_t^n=\hat{F}_t$, $n=0, 1, 2, \cdots$ for

$t \in [0, h]$, and define $\hat{F}_t^0 \equiv \hat{F}_h$ and

$$\hat{F}_t^n = T_t(\hat{F}^{n-1}), \qquad n = 1, 2, \cdots \quad \text{for } t \in [h, 2h].$$

Then, setting $k = h$ and replacing Z by $\hat{F}^n$ in (11), we have from (11)

$$|\hat{F}_t^{n+1} - \hat{F}_t| \leqq 2M \frac{K^{n+1}}{(n+1)!}(t-h)^{n+1} \quad \text{for } t \in [h, 2h].$$

This implies that $\hat{F}_t^n$ converges to $\hat{F}_t$ uniformly on $[h, 2h]$, which is $\mathcal{N}_t$-measurable.

In this manner, the fixed point $\hat{F}_t$ can be constructed recursively on $[0, h]$, $[h, 2h]$, $[2h, 3h]$, $\cdots$ and it is proved to be a $\mathcal{N}_t$-adapted process defined on the whole interval $[0, T]$. Then it follows from (2) that $y(t)$ is $\mathcal{N}_t$-measurable for each $t \in [0, T]$ and therefore $\mathcal{Y}_t \subset \mathcal{N}_t$ for all $t \in [0, T]$. □

(b) *Integrated-Lipschitzian feedback.* Consider another type of observation process which involves feedback loops in some smooth manner. Let $F(t, x, y)$ have the particular form

$$F(t, x, y) = \int_0^t F_0(s, x, y)\, ds + F_1(t, x), \tag{A.4}$$

satisfying the Lipschitz condition that there is a number $L > 0$ such that

$$|F_0(t, x, f) - F_0(t, x, g)| \leqq L\left(|f(t) - g(t)| + \int_0^t |f(s) - g(s)|\, ds\right)$$

$$\text{for all } (t, x) \in [0, T] \times X \quad \text{and} \quad f, g \in C.$$

Then, by modifying the standard argument slightly, it can be shown the stochastic differential equation (1) has a unique strong solution under the assumptions (A.1), (A.2) and (A.4). For this type of the observation we will establish again the informational equivalence between the observation and the innovation.

THEOREM 2. *Under the assumptions* (A.1), (A.2) *and* (A.4), *it follows that* $\mathcal{Y}_t = \mathcal{N}_t$ *for all* $t \in [0, T]$.

Proof. The proof is again based on the successive approximation for the fixed point of T_t which is defined by (4). First note that all the arguments up to the derivation of (6) and (7) in the proof of Theorem 1 are still valid here. So we will start from the estimate (6) with (7).

Applying the Lipschitz condition on F_0 to the first term of the right-hand side of (6), we have

$$\frac{\int_X |F_t^Z(x) - F_t^{\hat{F}}(x)| \exp R_t^Z(x) \mu(dx)}{\int_X \exp R_t^Z(x) \mu(dx)} \leqq L \int_0^t \int_0^s \left(|Z_r - \hat{F}_r| + \int_0^r |Z_u - \hat{F}_u|\, du\right) dr\, ds. \tag{12}$$

In order to estimate the second term, let us observe (7). Using the particular form of F assumed in (A.4), the first term of (7) is rewritten as

$$\int_0^t (F_s^Z(x) - F_s^{\hat{F}}(x))\, d\nu(s) = \int_0^t \int_0^s (F_{0r}^Z(x) - F_{0r}^{\hat{F}}(x))\, dr\, d\nu(s)$$

$$= \int_0^t (F_{0s}^Z(x) - F_{0s}^{\hat{F}}(x))\, ds\, \nu(t) - \int_0^t (F_{0s}^Z(x) - F_{0s}^{\hat{F}}(x))\nu(s)\, ds.$$

Therefore the Lipschitz condition on F_0 and Schwarz's inequality give

$$\left|\int_0^t (F_s^Z(x)-F_s^{\hat{F}}(x))\,d\nu(s)\right|$$

$$\leqq L|\nu(t)|\int_0^t\int_0^s\left(|Z_r-\hat{F}_r|+\int_0^r|Z_u-\hat{F}_u|\,du\right)dr\,ds \tag{13}$$
$$+L\left[\int_0^t|\nu(s)|^2\,ds\right]^{1/2}\left[\int_0^t\left(\int_0^s\left(|Z_r-\hat{F}_r|+\int_0^r|Z_u-\hat{F}_u|\,du\right)dr\right)^2 ds\right]^{1/2}.$$

The second and third terms of (7) are estimated by $|F_t^Z|\leqq M$, $|F_t^{\hat{F}}|\leqq M$, $|Z_t|\leqq M$, $|\hat{F}_t|\leqq M$ and the Lipschitz condition on F_0 as follows:

$$\left|\frac{1}{2}\int_0^t(|F_s^Z(x)|^2-|F_s^{\hat{F}}(x)|^2)\,ds\right|\leqq ML\int_0^t\int_0^s\left(|Z_r-\hat{F}_r|+\int_0^r|Z_u-\hat{F}_u|\,du\right)dr\,ds, \tag{14}$$

$$\left|\int_0^t(F_s^Z(x)Z_s-F_s^{\hat{F}}(x)\hat{F}_s)\,ds\right|$$
$$\leqq M\int_0^t|Z_s-\hat{F}_s|\,ds+ML\int_0^t\int_0^s\left(|Z_r-\hat{F}_r|+\int_0^r|Z_u-\hat{F}_u|\,du\right)dr\,ds. \tag{15}$$

Substituting (13), (14) and (15) into the second term of the right-hand side of (6), we have

$$\frac{\int_X(\exp R_t^Z(x)+\exp R_t^{\hat{F}}(x))|R_t^Z(x)-R_t^{\hat{F}}(x)|\mu(dx)}{\int_X\exp R_t^Z(x)\mu(dx)}$$
$$\leqq\left\{1+\frac{\int_X\exp R_t^{\hat{F}}(x)\mu(dx)}{\int_X\exp R_t^Z(x)\mu(dx)}\right\}\cdot\left\{\begin{matrix}\text{Right-hand side of}\\(13)+(14)+(15)\end{matrix}\right\}. \tag{16}$$

To bound the first factor on the right of (16), note that

$$\left|\int_0^t(F_s^{\hat{F}}(x)-F_s^Z(x))\,d\nu(s)\right|$$
$$\leqq 2ML|\nu(t)|\left(\frac{t^2}{2!}+\frac{t^3}{3!}\right)+2ML\left[\int_0^t|\nu(s)|^2\,ds\right]^{1/2}\left[\int_0^t\left(s+\frac{s^2}{2!}\right)^2 ds\right]^{1/2}$$

follows from (13). Denote the right-hand side of this inequality by $m(t,\nu)$. Then, we find

$$\frac{\int_X\exp R_t^{\hat{F}}(x)\mu(dx)}{\int_X\exp R_t^Z(x)\mu(dx)}\leqq\frac{\int_X\exp\left(\int_0^t F_s^Z(x)\,d\nu(s)+\frac{3}{2}M^2t+m(t,\nu)\right)\mu(dx)}{\int_X\exp\left(\int_0^t F_s^Z(x)\,d\nu(s)-\frac{3}{2}M^2t\right)\mu(dx)}$$
$$=\exp(3M^2t+m(t,\nu)). \tag{17}$$

It follows from (13)–(17) and (6) that

$$|T_t(Z)-\hat{F}_t|\leqq M_0(t,\nu)\int_0^t|Z_s-\hat{F}_s|\,ds$$
$$+M_1(t,\nu)\int_0^t\int_0^s\left(|Z_r-\hat{F}_r|+\int_0^r|Z_u-\hat{F}_u|\,du\right)dr\,ds \tag{18}$$
$$+M_2(t,\nu)\left[\int_0^t\left(\int_0^s\left(|Z_r-\hat{F}_r|+\int_0^r|Z_u-\hat{F}_u|\,du\right)dr\right)^2 ds\right]^{1/2},$$

where

$$M_0(t, \nu) = M(1+\exp(m(t, \nu)+3M^2t)),$$
$$M_1(t, \nu) = L(|\nu(t)|+2M)(1+\exp(m(t, \nu)+3M^2t))+L,$$
$$M_2(t, \nu) = L\left[\int_0^t |\nu(s)|^2\, ds\right]^{1/2}(1+\exp(m(t, \nu)+3M^2t)).$$

Thus, we find

$$|T_t(Z)-\hat{F}_t| \leqq N(t, \nu)\int_0^t |Z_s-\hat{F}_s|\, ds, \tag{19}$$

where

$$N(t, \nu) = M_0(t, \nu)+\left(t+\frac{t^2}{2!}\right)M_1(t, \nu)+\left[\int_0^t (1+s)^2\, ds\right]^{1/2} M_2(t, \nu).$$

Now, we define the $\mathcal{N}_t$-adapted processes $\hat{F}_t^n$, $n = 0, 1, 2, \cdots$ in the interval $[0, T]$ as follows. Set $\hat{F}_t^0 \equiv 0$ and define $\hat{F}_t^n = T_t(\hat{F}^{n-1})$, $n = 1, 2, \cdots$. Then, replacing Z by $\hat{F}^n$ in (19) and using a simple induction, we have

$$|\hat{F}_t^{n+1}-\hat{F}_t| \leqq N(t, \nu)\frac{[\int_0^t N(s, \nu)\, ds]^n}{n!}Mt. \tag{20}$$

This implies that $\hat{F}_t^n$ converges to the fixed point $\hat{F}_t$ uniformly in t. Thus, since $\hat{F}_t^n$ is $\mathcal{N}_t$-measurable, we see that $\hat{F}_t$ is $\mathcal{N}_t$-measurable, and so $y(t)$ by (2). Therefore, $\mathcal{Y}_t \subset \mathcal{N}_t$ for all $\in [0, T]$. □

As examples for both cases, we will present two particular types of observation processes involving feedback of observation via stochastic differential equations, which are regarded as setups of stochastic control problems (compare with the model of [1]).

Example 1. Let the observation process be given by

$$y(t) = \int_0^t H(s, z(s))\, ds + w(t). \tag{21}$$

Here $z(t)$ is the state variable satisfying

$$z(t) = \int_0^t G(s, z(s), y(r), -h \leqq r \leqq s-h)\, ds + v(t), \tag{22}$$

where $y(t) \equiv 0$ for $t \in [-h, 0]$, and $v(t)$ is a Wiener process independent of $w(t)$. This setup is a general one for control problems with delayed controls [17]. If G satisfies the Lipschitz condition and the growth condition in the second argument, i.e., there is a number $\kappa > 0$ such that

$$|G(t, \xi, y)-G(t, \zeta, y)| \leqq \kappa|\xi-\zeta|, \tag{23}$$

$$|G(t, \xi, y)| \leqq \kappa(1+|\xi|) \quad \text{for } \xi, \zeta \in R, \tag{24}$$

(with, of course, appropriate measurability conditions in all arguments), then the stochastic differential equation (22) has a unique solution; $z(t, v(s), y(s-h), 0 \leqq s \leqq t)$ for each $y \in C$. In this case, if H is bounded, the observation (21) becomes the delayed feedback type (where $x = v$).

Example 2. Next consider the following type of observation:

$$y(t) = \int_0^t \int_0^s H(r, z(r))\, dr\, ds + w(t), \tag{25}$$

where $z(t)$ is the state variable satisfying the same equation as (22) but $h = 0$:

$$z(t) = \int_0^t G(s, z(s), y(r), 0 \leqq r \leqq s)\, ds + v(t). \tag{26}$$

Similarly as before, if G satisfies the Lipschitz condition (23) and the growth condition (24), then (26) has a unique solution $z(t, v(s), y(s), 0 \leqq s \leqq t)$ for each $y \in C$. Furthermore, if the Lipschitz condition as used in the assumption (A.4) holds with respect to the third argument of G, it can be shown that the solution satisfies the similar Lipschitz condition in the third argument. Therefore, the observation (25) becomes the integrated-Lipschitzian feedback type (where $x = v$), if it is assumed additionally that H is bounded and satisfies the Lipschitz condition as (23) in the second argument.

4. Remarks. From the previous discussion it may be conjectured that the existence of a strong solution of (1) is sufficient to ensure informational equivalence between the observation and the innovation, though we have not found out any cue for the proof. The delayed feedback and the integrated-Lipschitzian respectively are sufficient conditions for such existence. The other special case is that with additive type feedback: $F(t, x, y) = \bar{F}(t, x) + \tilde{F}(t, y)$. For this type, the informational equivalence has been already established in [11]. We give here a simple proof. First note that

$$\begin{aligned} \nu(t) &= y(t) - \int_0^t E\{F(s, x, y) | \mathcal{Y}_s\}\, ds \\ &= \bar{y}(t) - \int_0^t E\{\bar{F}(s, x) | \mathcal{Y}_s\}\, ds, \end{aligned}$$

where $\bar{y}(t) = \int_0^t \bar{F}(s, x)\, ds + w(t)$. Let $\bar{\mathcal{Y}}_t = \sigma\{\bar{y}(s), 0 \leqq s \leqq t\}$ and $\bar{\nu}(t) = \bar{y}(t) - \int_0^t \times E\{\bar{F}(s, x) | \bar{\mathcal{Y}}_s\}\, ds$. Now it follows from Clark's result [4] that $\bar{\mathcal{Y}}_t = \sigma\{\bar{\nu}(s), 0 \leqq s \leqq t\}$ whenever $\bar{F}$ is bounded. Therefore, if we can show that $\mathcal{Y}_t = \bar{\mathcal{Y}}_t$, we have $\nu(t) = \bar{\nu}(t)$ and the desired result. However, from the definition of $\bar{y}(t)$,

$$y(t) = \bar{y}(t) + \int_0^t \tilde{F}(s, y)\, ds; \tag{27}$$

therefore the Lipschitz condition and the growth condition on $\tilde{F}$ give a one-to-one correspondence between y and $\bar{y}$, i.e., $\mathcal{Y}_t = \bar{\mathcal{Y}}_t$. It should be noted that the delayed feedback assumption, i.e., $\tilde{F}(t, y(s), -h \leqq s \leqq t - h)$ assures also such one-to-one correspondence in (27), (see [17]).

Finally we give additional remarks on the references. A published version of Clark's proof [4] appears in [15] (see also [14], [12]). After submitting this paper, the author learned of a recent result of Krylov [13] which assures observation-innovation equivalence when the signal arises from a diffusion process different from the one in [1] and is not independent of the noise. Krylov's result is summarized in [12]. The reviewers informed the author that Beneš [2] and Kallianpur [11] both have the same mistakes in the proof of their extensions.

REFERENCES

[1] V. E. BENEŠ, *On Kailath's innovations conjecture hold*, Bell System Tech. J., 55 (1976), pp. 981–1001.

[2] ———, *Extension of Clark's innovation equivalence theorem to the case of signal z independent of noise, with* $\int_0^t z_s^2\,ds < \infty$ *a.s.*, Math. Programming Stud., 5 (1976), pp. 2–7.

[3] B. S. TSIREL'SON, *An example of a stochastic differential equation having no strong solution*, Theory Prob. Appl., 20 (1975), pp. 416–418.

[4] J. M. C. CLARK, *Conditions for the one-to-one correspondence between an observation process and its innovation*, Tech. Report 1, Imperial College, London, (1969).

[5] M. H. A. DAVIS, *A direct proof of innovations/observations equivalence for Gaussian processes*, IEEE Trans. Inform. Theory, IT-24 (1978), pp. 252–254.

[6] P. FROST, *Estimation in continuous-time nonlinear systems*, Dissertation, Stanford University, Stanford, CA., 1968.

[7] T. T. KADOTA, M. ZAKAI AND J. ZIV, *Mutual information of the white Gaussian channel with and without feedback*. IEEE Trans. Inform. Theory, IT-17 (1971), pp. 368–371.

[8] T. KAILATH, *Some extensions of the innovation theorem*, Bell System Tech. J., 50 (1971), pp. 1487–1494.

[9] ———, *A note on linear squares estimation by the innovations method*, this Journal, 10 (1972), pp. 477–486.

[10] G. KALLIANPUR AND C. STRIEBEL, *Estimation of stochastic processes: arbitrary system process with additive white noise errors*, Ann. Math. Statist, 39 (1968), pp. 785–801.

[11] G. KALLIANPUR, *A linear stochastic system with discontinuous control*. Proc. of Intern. Symp. SDE, Kyoto, 1976, Kinokuniya Book-Store, Co., Ltd., Tokyo, 178.

[12] ———, *Stochastic Filtering Theory*. Springer-Verlag, Berlin, 1980.

[13] N. V. KRYLOV, *On the equivalence of* σ*-algebras in the filtering problem of diffusion processes*, Theor. Prob. Appl., 24 (1979), pp. 772–782.

[14] H. KUNITA, *Estimation of stochastic processes*, Sangyotosho, Tokyo, (1976), (in Japanese).

[15] P. A. MEYER, *Sur un problème de filtration*, Séminaire de Probabilités VII, Lecture Notes in Mathematics 321, Springer-Verlag, Berlin, 1973.

[16] E. MOSCA, *Weak conditions for innovations informational equivalence in the independent Gaussian case*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 63–69.

[17] K. UCHIDA, *On optimal control of the stochastic systems with delayed controls and delayed measurements*, J. Math. Anal. Appl., 75 (1980), pp. 454–464.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 5, September 1981

0363-0129/81/1905-0009 $01.00/0

RECURSIVE ESTIMATION IN DIFFUSION MODEL*

G. BANON† AND HUNG T. NGUYEN‡

Abstract. This paper is concerned with the nonlinear identification of dynamical systems disturbed by white noise, an important problem in control engineering. A nonparametric identification procedure for such systems which are usually described by a diffusion model is given in Banon (1977), (1978), where weak consistency of estimators has been obtained and simulation study has been carried out successfully. In this paper, we prove a stronger result concerning asymptotic properties of the estimators of the drift term, namely, strong consistency, and also related results.

1. Introduction. The diffusion model is usually used to describe the behavior of dynamical systems disturbed by white noise. Specifically, we are concerned with systems represented by a stochastic differential equation of the form

$$dX_t = m(X_t)\,dt + \sigma(X_t)\,dW_t, \qquad t \geqq 0 \tag{1.1}$$

where $(X_t, t \geqq 0)$ is the one-dimensional observation process, $(W_t, t \geqq 0)$ is the standard Wiener process, and $m(\cdot)$, $\sigma(\cdot)$ are unknown functions to be estimated.

In the case of linear systems, many identification techniques have been proposed, e.g., Kalman and Bucy (1961). The approach which is presented in Banon's work (Banon (1977), (1978)) to solve a class of nonlinear identification problems is based mainly on the kernel method in statistical estimation theory. The estimators obtained are recursive in the sense that they can be easily updated when more data are available.

Since the diffusion term $\sigma^2(x)$ can be expressed as

$$\lim_{t \to 0} \frac{1}{t} E((X_{t+s} - X_s)^2 | X_s = x), \quad x \in \mathbb{R}, \quad s \geqq 0, \tag{1.2}$$

the property of quadratic variation of diffusion processes (Wong and Zakai (1965)), can be used to obtain a recursive estimator for $\sigma^2(x)$, as shown in Banon (1978); we are led to focus on the nonparametric estimation of the drift term $m(\cdot)$, assuming that σ^2 is known or unknown but constant.

If we denote by f the common probability density of the X_t's (a stationary Markov process, solution of (1.1)), and f' its derivative, then it can be shown (Banon (1978)) that

$$m(x) = \frac{1}{2}\sigma^2 \frac{f'(x)}{f(x)}, \qquad f(x) > 0.$$

Therefore, we are led to consider the estimation of the logarithmic derivative of f based on a realization of the continuous-time, stationary Markov process $(X_t, t \leqq T)$. For this purpose, we shall use the results in Nguyen (1979) concerning the estimation of $f(x)$ at each point x such that $f(x) > 0$. (The values of $m(\cdot)$ on $\{f = 0\}$ are irrelevant, based on the observations $X_t, t \geqq 0$).

This paper is organized as follows. In § 2, we shall show that a stationary Markov process satisfying a certain mixing condition, namely the G_2 condition of Rosenblatt (1970), (1971) can be interpreted as an asymptotically uncorrelated process. This condition G_2 will be used throughout our work to obtain asymptotic properties of the estimators. In § 3, we shall extend various results in Nguyen (1979) to include recursive

* Received by the editors February 29, 1980, and in revised form January 23, 1981.

† Laboratoire d'Automatique et d'Analyse des Systèmes, Toulouse, France.

‡ Department of Mathematics and Statistics, University of Massachusetts, Amherst, Massachusetts 01003.

estimates of Yamato type (Yamato (1971)), and recursive estimates of the derivative of a probability density. In § 4, we specify the assumptions on our model of diffusion, describe in detail the estimation scheme and prove the strong consistency of the estimator of the drift term. This main result is stronger than the one in Banon (1978). Note that the asymptotic normality of all the estimators involved in these papers has been obtained in Nguyen (1978). A simulation study in order to compare our procedure of identification with others is under consideration.

2. Asymptotically uncorrelated processes. In density estimation from stationary Markov processes, e.g., Rosenblatt (1970), Banon (1978), asymptotic properties of estimators, e.g., consistency in quadratic mean, have been obtained under an additional condition on the process, namely the G_2 condition of Rosenblatt (1970). Strictly speaking, such a condition applies to the transition operator. We shall show that a stationary Markov process $(X_t, t \in \mathbb{R}^+)$ satisfying the condition G_2 can be interpreted as an asymptotically uncorrelated process (Rosenblatt (1971)).

From now on, we assume that the stationary Markov process $(X_t, t \in \mathbb{R}^+)$, the solution of (1.1), has first order density $f(\cdot)$ on the real line $\mathbb{R}$. Conditions for such a situation will be given in § 4.

For continuous-time processes, let us recall the definition of the condition G_2 (Banon (1978)) (for discrete-time processes, see Rosenblatt (1970)). For each $t \in (0, +\infty)$, let T_t be the transition operator of $(X_t, t \in \mathbb{R}^+)$ defined on the space L^∞ of bounded Borel measurable functions. (For our purpose, it is sufficient to consider bounded functions on $\mathbb{R}$.)

$$(T_t g)(x) = E(g(X_t) | X_0 = x), \quad g \in L^\infty, \quad x \in \mathbb{R}.$$

We consider the following norm of the operator T_t

$$\langle T_t \rangle_p = \sup_{g \perp 1} \frac{\|T_t g\|_p}{\|g\|_p}, \qquad 1 \leqq p \leqq +\infty,$$

where $\|g\|_p$ stands for $E^{1/p}$, $|g(X_0)|^p$ for $g \in L^\infty$, $g \perp 1$ means $E(g(X_0)) = 0$ and E means expectation with respect to the stationary density f defined previously.

The process $(X_t, t \in \mathbb{R}^+)$ is said to satisfy the condition $G_p(s, a)$ if there exists $s > 0$ such that

$$\langle T_s \rangle_p \leqq a < 1.$$

As for the discrete case, one can show that the above G_p conditions, $1 < p < +\infty$, are all equivalent to each other.

Rosenblatt (1971) has considered the L^p-norm condition for discrete-time Markov processes. We now extend his definition to continuous-time Markov processes. A stationary Markov process $(X_t, t \in \mathbb{R}^+)$ is said to satisfy the L^p-norm condition, $1 \leqq p \leqq +\infty$ if $\langle T_t \rangle_p \to 0$ as $t \to +\infty$.

It was shown in Rosenblatt (1971) that the L^p-norm conditions, $1 < p < +\infty$, are all equivalent to each other, and that the L^2-norm condition is equivalent to the fact that the Markov process is asymptotically uncorrelated. It is easy to see that these results remain true for continuous-time processes.

According to the above assumptions, a stationary process $(X_t, t \in \mathbb{R}^+)$ is said to be asymptotically uncorrelated if

$$\sup_{g_1, g_2 \perp 1} \frac{E g_1(X_0) g_2(X_t)}{\|g_1\|_2 \|g_2\|_2} \to 0 \quad \text{as } t \to +\infty.$$

The L^p-norm and the G_p conditions, $1 \leqq p \leqq +\infty$, are equivalent. Indeed, if the $G_p(s, a)$ condition is satisfied, then it is easy to check that $\langle T_t \rangle_p \leqq d^t$, where $d = a^{1/s} \in (0, 1)$, for $t = ks$, $k = 0, 1, 2, \cdots$ (using simply the fact that $(T_t, t \in \mathbb{R}^+)$ is a semigroup of contractions).

Now, if $t \in \mathbb{R}^+$, let k be an integer such that $ks \leqq t < (k+1)s$; since $(T_t, t \in \mathbb{R}^+)$ is a semigroup and $\langle T_u \rangle \leqq 1$, $\forall u \geqq 0$, we have

$$\langle T_t \rangle_p \leqq \langle T_{ks} \rangle_p \leqq d^{ks} \leqq d^{t-s} = \left(\frac{1}{a}\right) d^t.$$

Therefore $\lim_{t \to +\infty} \langle T_t \rangle_p = 0$; i.e., the process $(X_t, t \in \mathbb{R}^+)$ satisfies the L^p-norm condition.

The fact that the L^p-norm condition implies the G_p condition is obvious. We state the above results in

LEMMA 2.1. *Let $(X_t, t \in \mathbb{R}^+)$ be a stationary Markov process. If p, q $(1, +\infty)$, then the following statements are equivalent*:

(i) *The process satisfies the L^p-norm condition.*

(ii) *The process is asymptotically uncorrelated.*

(iii) *The process satisfies the G_q condition.*

Any one of (i) or (ii) or (iii) implies that $\langle T_t \rangle_r \leqq cd^t$ for $d \in (0, 1)$, $t \in \mathbb{R}^+$ and $r \in (1, +\infty)$.

If $p = q \in [1, +\infty]$ then (i) and (iii) are equivalent.

3. Strong consistent estimates of the density and its derivative.

3.1. Estimation scheme and assumptions. In the sequel, we shall make the following assumptions on the stationary Markov process $(X_t, t \in \mathbb{R}^+)$:

(i) $(X_t, t \in \mathbb{R}^+)$ is asymptotically uncorrelated.

(ii) $(X_t, t \in \mathbb{R}^+)$ is a measurable process.

(iii) The common probability distribution of the X_t's is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$. We denote by f its Radon-Nikodým derivative. We also assume that f is continuous and bounded.

The conditions for such a situation will be given in § 4.

We consider the following class of recursive estimators of $f(x)$, $x \in \mathbb{R}$, which have been investigated in Nguyen (1979). These estimators are the analogue of the ones in the case of independent identically distributed random variables, introduced in Deheuvels (1974), containing as a particular case the sequential estimators of Yamato (1971).

For $x \in \mathbb{R}$ and $t > 0$, the estimate of $f(x)$ based on the observation process up to time t is taken as

$$f_t(x) = \left(\int_0^t h(s) H(h(s))\, ds \right)^{-1} \int_0^t H(h(s)) K\left(\frac{X_s - x}{h(s)}\right) ds,$$

where $h(\cdot)$ is a mapping from $\mathbb{R}^+$ to $\mathbb{R}^+ - \{0\}$ and $H(\cdot)$ from $\mathbb{R}^+ - \{0\}$ to $\mathbb{R}^+$ such that:

(a) $h(s) \downarrow 0$ as $s \to +\infty$,

(b) $h(\cdot) H(h(\cdot))$ is locally integrable on $\mathbb{R}^+$, and

$$\int_0^t h(s) H(h(s))\, ds \to +\infty \qquad \text{as } t \to +\infty$$

and where the kernel K is a Borel measurable function on $\mathbb{R}$ such that:

(c_0) K is bounded and $\int_{-\infty}^{+\infty} K(y)\, dy = 1$, $K \geqq 0$.

Remark. In Banon's work (1978), $H(x) \equiv 1$. If $H(u) = 1/u$, we obtain the analogue of Yamato's estimates in the discrete case.

For the estimation of the derivative f' of f, we make an additional assumption on the process:

(iv) The density f has a continuous and bounded derivative.

As estimate of $f'(x)$, we use the derivative of $f_t(x)$:

$$f_t'(x) = \left[\int_0^t h(s)H(h(s))\,ds\right]^{-1} \int_0^t H(h(s))/h(s)K'[(X_s - x)/h(s)]\,ds,$$

where the kernel K is such that:

(c_1) It is a density of bounded variation and its derivative K' is bounded.

For simplicity, we will adopt the following notation. If u and v are functions on some subset of $\mathbb{R}$, and if there is some constant c such that $u(t) \leqq cv(t)$ for all t in that subset A, then we will write $u(t) = \hat{O}(v(t))$ for $t \in A$.

In Nguyen (1979), under the G_2 condition, sufficient conditions for almost sure convergence of $f_t(x)$ are given, for the case where $H(\cdot)$ is bounded or $H(t) = \hat{O}(1/t)$, for $t > 0$. In this section, we consider the case $H(t) = \hat{O}(t^k)$ for $t > 0$ and $k \in [-1, +\infty)$ which contains the estimates of Yamato type as a special case.

3.2. Preliminary lemmas. We start by improving some technical lemmas in Nguyen (1979).

LEMMA 3.2.1. *Let $(Z_t, t \in \mathbb{R}^+)$ be a measurable second-order process, with $EZ_t = 0$ for all $t \in \mathbb{R}^+$. Let g be a mapping from $\mathbb{R}^+$ to $\mathbb{R}^+$ such that $g(t) > 0$ for $t > 0$. If*

(i) *there exists $b \in (0, 1)$ such that $g(t) \sim t^b$ as $t \to +\infty$,*

(ii) *there exists a nonnegative constant u such that $C(t, t) = O(t^{2u})$ as $t \to +\infty$, where $C(\cdot, \cdot)$ denotes the covariance function of the process $(Z_t, t \in \mathbb{R}^+)$, and O is the usual notation, then*

$$\forall a \in \left(0, \frac{1}{2(u+1-b)}\right), \quad W_t - W_{m^a} \to 0,$$

almost surely as $t \to +\infty$, where

$$W_t = \frac{1}{g(t)} \int_0^t Z_s\,ds$$

and $(m^a, m \in N - \{0\})$ is a sequence of positive real numbers such that, for $t \geqq 1$, $m^a \leqq t < (m+1)^a$.

The proof of this lemma is similar to that of Lemma 5 in Nguyen (1979), which is, roughly speaking, an adaptation of the technique employed in Loeve (1960) for the almost-sure stability problem.

LEMMA 3.2.2. *Under the hypotheses of Lemma* 3.2.1 *and if, in addition*:

(iii) *there exists a $v > 0$ such that*

$$\frac{1}{t^{2b}} \int_0^t \int_0^t C(s, s')\,ds\,ds' = O(t^{2(b-1-u)-v}) \quad \text{as } t \to +\infty,$$

where u is the constant defined in (ii) *and b is the constant defined in* (i), *then $W_t \to 0$, almost surely, as $t \to +\infty$.*

Proof. Since $v > 0$, there exists an a such that

$$-\frac{1}{a} \in (2b - 2 - 2u - v, 2b - 2 - 2u).$$

For such an a, Lemma 3.2.1 tells us that $W_t - W_{m^a} \to 0$, almost surely, as $t \to +\infty$. The announced result is obtained if we prove that $W_{m^a} \to 0$ almost surely as $m \to +\infty$. But from assumption (iii) we have, for $n \to +\infty$,

$$\sum_{m=1}^{n} \frac{1}{m^{2ab}} \int_0^{m^a} \int_0^{m^a} C(s, s')\, ds\, ds' = O\Big(\sum_{m=1}^{n} m^{a(2b-2-2u-v)} \Big) = O(1)$$

since $a(2b-2-2u-v) < -1$, recalling that $g(t) \sim t^b$. This result implies that

$$\sum_{m=1}^{n} EW_{m^a}^2 = \sum_{m=1}^{n} \frac{1}{g^2(m^a)} \int_0^{m^a} \int_0^{m^a} C(s, s')\, ds\, ds' = O(1)$$

which completes the proof. □

We now specify the process $(Z_t, t \in \mathbb{R}^+)$.

LEMMA 3.2.3. *Let* $(Z_t^{(i)}, t \in \mathbb{R}^+)$, $i = 0, 1$, *be two processes defined for* $x \in \mathbb{R}$ *by*

$$Z_t^{(i)} = [H(h(t))/h(t)^i][K^{(i)}\{(X_t - x)/h(t)\} - EK^{(i)}\{(X_t - x)/h(t)\}],$$

where $(X_t, t \in \mathbb{R}^+)$ *is a stationary Markov process satisfying the conditions* (i) *and* (iii) *of* § 3.1, $h(\cdot)$, $H(\cdot)$, *and* $K(\cdot)$ *are functions defined in* § 3.1 *with* $K(\cdot)$ *satisfying* (c_0) *(resp.* (c_1)*) for* $i = 0$ *(resp.* $i = 1$*).*

If $C_i(\cdot, \cdot)$ *is the covariance function of the process* $(Z_t^{(i)}, t \in \mathbb{R}^+)$, *then there exists* $d \in (0, 1)$ *such that, for* $s, s' \in \mathbb{R}^+$, *one has*

$$C_i(s, s') = \hat{O}[H(h(s))H(h(s'))\{h(s)h(s')\}^{1/2-i} d^{|s-s'|}], \quad i = 0, 1.$$

Proof. By Lemma 2.1, the transition operator of $(X_t, t \in \mathbb{R}^+)$ satisfies $\langle T_t \rangle_2 \leqq d^t$ for all $t \in \mathbb{R}^+$ with $d \in (0, 1)$. Therefore, as in Banon (1978), we have, for any $s, s' \in \mathbb{R}^+$,

$$C_i(s, s') = EZ_s^{(i)} Z_{s'}^{(i)} = \hat{O}[d^{|s-s'|} E^{1/2}(Z_s^{(i)})^2 E^{1/2}(Z_{s'}^{(i)})^2].$$

Under the assumption (c_0) (resp. (c_1)) for $i = 0$, (resp. $i = 1$) and the fact that $f(\cdot)$ is bounded, we have, for any $s \in \mathbb{R}^+$,

$$\begin{aligned} E^{1/2}(Z_s^{(i)})^2 &\leqq H(h(s))h(s)^{-i} E^{1/2}\{K^{(i)}[(X_0 - x)/h(s)]\}^2 \\ &= \hat{O}(H(h(s))h(s)^{1/2-i}, \end{aligned}$$

which completes the proof. □

To prove the strong consistency of estimators in the following paragraph we need the following lemma on asymptotic unbiasedness:

LEMMA 3.2.4. *If* $(X_t, t \in \mathbb{R}^+)$ *is a stationary Markov process satisfying conditions* (ii) *and* (iii) *of* § 3.1, *and* $h(\cdot)$, $H(\cdot)$, $K(\cdot)$ *are functions satisfying conditions* (a), (b) *and* (c_1) *for* $i = 0$ *(resp.* $i = 1$*).*

$$Ef_t(x) \to f(x) \quad \text{as } t \to +\infty.$$

If in addition $(X_t, t \in \mathbb{R}^+)$ *satisfies condition* (iv) *in* § 3.1, *and* $K(\cdot)$ *satisfies the stronger condition* (c_1) *instead of* (c_0), *then* $Ef_t'(x) \to f'(x)$, $t \to +\infty$.

The first statement above was proved in Nguyen (1979); the second one follows from arguments similar to those used in Banon (1978) and Nguyen (1979).

3.3. Strong consistency of estimators. Let us call assumption A_0 the set of the conditions (i)–(iii), (a), (b) and (c_0) above and assumption A_1 the set of the conditions (i)–(iv), (a), (b) and (c_1) for simplicity. We will use the previous lemmas to prove the following theorem.

THEOREM 3.3.1. *If, in addition to assumption* A_0 *(resp.* A_1*), the functions* $h(\cdot)$ *and* $H(\cdot)$ *are such that either one of the following two conditions is satisfied:*

d) $H(t) = \hat{O}(t^k)$ *for* $t > 0$, *with* $k \in [-1, i - \frac{1}{2}]$,

$$\int_0^t h(s)H(h(s))\,ds \sim t^b \quad \text{and} \quad h(t)^{2k-2i+1} = O(t^q), \quad t \to +\infty,$$

with $b \in (q/2 + \frac{3}{4}, 1]$ *and* $q \in [0, \frac{1}{2})$,

d') $H(t) = \hat{O}(t^k)$ *for* $t > 0$, *with* $k \geqq i - \frac{1}{2}$,

$$\int_0^t h(s)H(h(s))\,ds \sim t^b \quad \text{and} \quad \int_0^t h(s)^{2k-2i+1}\,ds = O(t^{b'})$$

as $t \to +\infty$, *with* $b \in (b'/4 + \frac{1}{2}, 1)$ *and* $b' \in (0, 1]$ *for* $i = 0$ *(resp.* $i = 1$*), then, for any* $x \in \mathbb{R}$, $f_t^{(i)}(x) \to f^{(i)}(x)$, *almost surely,* $t \to +\infty$, *for* $i = 0, 1$.

Proof. Let $(Z_t^{(i)}, t \in \mathbb{R}^+)$ be the process defined in Lemma 3.2.3 and $g(\cdot)$ be the function defined by

$$g(t) = \int_0^t h(s)H(h(s))\,ds.$$

Then $f_t^{(i)}(x) - Ef_t^{(i)}(x) = W_t^{(i)}$, where

$$W_t^{(i)} = \left[\frac{1}{g(t)}\right] \int_0^t z_s^{(i)}\,ds.$$

Under the assumptions of the theorem, we first note that $(Z_t^{(i)}, t \in \mathbb{R}^+)$, $i = 0, 1$, are measurable, second-order processes with $EZ_t^{(i)} = 0$ for all $t \in \mathbb{R}^+$, and that $g(t) > 0$ for $t > 0$, and there exists $b \in (0, 1]$ for which $g(t) \sim t^b$ as $t \to +\infty$. Under assumption (d), i.e., $k \leqq i - 1/2$, from Lemma 3.2.3 we have, for $t \to +\infty$,

$$\begin{aligned} C_i(t, t) = O[H^2(h(t))h(t)^{1-2i}] &= O[h(t)^{2k-2i+1}] \\ &= O(t^q) = O(t^{2u}) \quad \text{for any } u \geqq q/2. \end{aligned}$$

Hence, for $u = q/2$, Lemma 3.2.1 is satisfied, and we also have, for $t \in \mathbb{R}^+$,

$$C_i(s, s') = O[\{h(s)h(s')\}^{k-i+1/2}\, d^{|s-s'|}],$$

$h(\cdot)$ being decreasing as $t \to +\infty$.

$$\begin{aligned} \frac{1}{t^{2b}} \int_0^t \int_0^t C_i(s, s')\,ds\,ds' &= O[f(t)^{2k-2i+1} t^{1-2b}] \\ &= O(t^{q+1-2b}) = O(t^{2(b-1-u)-v}) \quad \text{for any } v \leqq 4b - 3 - 2u - q, \end{aligned}$$

which proves that Lemma 3.2.2 is satisfied, since the conditions $b > q/2 + \frac{3}{4}$ and $u = q/2$ imply the existence of a $v > 0$.

In the same way, under (d'), i.e., $k \geqq 1 - \frac{1}{2}$, we have, for $t \to +\infty$, $C_i(t, t) = O(1) = O(t^{2u})$ for any $u \geqq 0$. Hence, for $u = 0$, Lemma 3.2.1 is satisfied and we also have for $t \to +\infty$,

$$\begin{aligned} \frac{1}{t^{2b}} \int_0^t \int_0^t C_i(s, s')\,ds\,ds' &= O\left[t^{-2b} \int_0^t h(s)^{2k-2i+1}\,ds\right] \\ &= O(t^{b'-2b}) \\ &= O(t^{2(b-1-u)-v}) \quad \text{for } v \leqq 4b - 2 - 2u - b' \end{aligned}$$

which proves that Lemma 3.2.2 is satisfied since the conditions $b > b'/4 + \frac{1}{2}$ and $u = 0$ imply the existence of a $v > 0$.

Therefore, under either (d) or (d') we have, for $i = 0, 1$, $W_t^{(i)} \to 0$ almost surely as $t \to +\infty$.

The last result together with the property of asymptotic unbiasedness (Lemma 3.2.4) imply that for any $x \in \mathbb{R}$ and $i = 0, 1$, $f_t^{(i)}(x) - f^{(i)}(x) = W_t^{(i)} + Ef_t^{(i)}(x) - f^{(i)}(x) \to 0$, almost surely as $t \to +\infty$, which completes the proof of the theorem. □

Remarks. We get the estimates of Yamato type by setting $H(t) = 1/t$, then $\int_0^t h(s)H(h(s))\,ds = t$, i.e., $k = -1$, $b = 1$. The corresponding sufficient condition for the almost-sure convergence, $i = 0$, is the existence of a $q \in [0, \frac{1}{2})$ such that for $t \to +\infty$, $1/h(t) = O(t^q)$.

The asymptotic bias and variance of these estimators, and also the conditions for asymptotic normality, are given in Nguyen (1978).

4. Strong consistent estimates of the drift term.

4.1. Nonlinear identification problem. Consider a stochastic dynamical system represented by the following stochastic differential equation

$$dX_t = m(X_t) + \sigma(X_t)\,dW_t, \qquad t \in \mathbb{R}^+,$$

with a second-order initial condition, $X_0 = X$, independent of $(W_t, t \in \mathbb{R}^+)$, which is a standard Wiener process defined on the same probability space as the observation process $(X_t, t \in \mathbb{R}^+)$, and $m(\cdot)$, $\sigma(\cdot)$ are two Borel measurable functions on the real line $\mathbb{R}$. We assume that the random variable X has a probability density on $\mathbb{R}$, denoted by $f_X(\cdot)$.

The problem is to estimate $m(\cdot)$ (and $\sigma(\cdot)$) from the observations X_t.

A restricted class of such identification problems is described as follows:

(i) The functions $m(\cdot)$ and $\sigma(\cdot)$ satisfy the Lipschitz condition $|m(x) - m(y)| + |\sigma(x) - \sigma(y)| \leqq c|x - y|$, for all $x, y \in \mathbb{R}$ where c is some constant. These two functions satisfy also the linear growth condition $|m(x)| + |\sigma(x)| \leqq c\sqrt{1 + x^2}$.

Note that under this assumption, the process $(X_t, t \in \mathbb{R}^+)$ is unique with probability one and is a measurable Markov process.

(ii) The function $\sigma(\cdot)$ is such that, for any $x \in \mathbb{R}$,

$$\sigma(x) \geqq \sigma_0 > 0.$$

Under (i) and (ii), the above process has a stationary transition density (Wong (1971)). We denote its value by $f_a(x, t)$, $x, a \in \mathbb{R}$ and $t \in \mathbb{R}^+ (f_a(x, t)$ is the density of X_t given that $X_0 = a)$.

(iii) The derivatives $m'(\cdot)$, $\sigma'(\cdot)$ and $\sigma''(\cdot)$ satisfy the Lipschitz and linear growth conditions of (i).

Under (i), (ii) and (iii), $f_a(\cdot, \cdot)$ is the unique fundamental solution of the forward equation of Kolmogorov.

(iv) The functions $m(\cdot)$ and $\sigma(\cdot)$ are such that the solutions of the equation $\frac{1}{2}d(\sigma^2(x)w(x))/dx = m(x)w(x)$ are bounded and integrable on $\mathbb{R}$.

Under (i)–(iv), $f_a(\cdot, \cdot)$ converges, as $t \to +\infty$, for any $a \in \mathbb{R}$, to a bounded and continuous limiting density on $\mathbb{R}$, $f(\cdot)$, say, which is solution of the above differential equation (Banon (1978)).

Since we are interested in asymptotic results, we assume from now on that $f_X(\cdot) = f(\cdot)$. In this case, the limiting density $f(\cdot)$ is the common density of the X_t's.

(v) The functions $m(\cdot)$ and $\sigma(\cdot)$ are such that

$$\min\left(\lim_{x \to -\infty} r(x), \lim_{x \to +\infty} r(x)\right) > 0,$$

where $r(\cdot)$ is the function defined by

$$r(x)=\frac{m^2(x)}{2}\sigma^2(x)+\frac{m'(x)}{2}-m(x)\frac{\sigma'(x)}{\sigma(x)}+\frac{\sigma'^2(x)}{8}-\frac{\sigma(x)\sigma''(x)}{4}, \qquad x\in\mathbb{R}.$$

Roughly speaking, this condition (v) is about the nature of the spectrum of a Sturm-Liouville problem; for details see Banon (1978).

Under (i)–(v), it is shown in Banon (1978) that the corresponding process $(X_t, t\in\mathbb{R}^+)$ satisfies the condition G_2, i.e., by Lemma 2.1, is asymptotically uncorrelated.

Now consider the case where $\sigma^2(\cdot)$ is known or unknown but constant (a recursive estimate of σ^2 in this latter case can be constructed, recalling that σ^2 may be characterized as a conditional expectation). Since $f(\cdot)$ is a solution of the differential equation given in (iv), we have, for $x\in\mathbb{R}$,

$$m(x)=\frac{1}{2}\left[\{\sigma^2(x)\}'+\frac{\sigma^2(x)f'(x)}{f(x)}\right].$$

Therefore, the estimation of $m(x)$ is reduced to the estimation of the quotient $Q(x)=f'(x)/f(x)$, for x such that $f(x)>0$.

4.2. Strong consistent estimates of $Q(x)$. Based on the previous investigation of the estimation of $f(x)$ and $f'(x)$, a natural estimate of $Q(x)$ from the observations up to time t is

$$Q_t(x)=\left[\int_0^t\left\{\frac{H(h(s))}{h(s)}\right\}K_1'\left\{\frac{X_s-x}{h(s)}\right\}ds\right]\cdot\left[\int_0^t H(h(s))K_0\left\{\frac{X_s-x}{h(s)}\right\}+\varepsilon\right]^{-1},$$

where $K_i(\cdot)$, $i=0,1$, are two kernels and ε is a positive constant.

THEOREM 4.1. *Let $(X_t, t\in\mathbb{R}^+)$ be the process defined in* § 4.1 *under the assumptions* (i)–(v), *with $f_X(\cdot)=f(\cdot)$; $h(\cdot)$ and $H(\cdot)$ the function defined in* § 3.1, *and satisfying* (a), (b); *$K_i(\cdot)$, $i=0,1$, the kernels satisfying (c_i), $i=0,1$, respectively. If in addition, $h(\cdot)$ and $H(\cdot)$ satisfy* (d) *or* (d') *of Theorem* 3.1 *with $i=1$, and if $f'(\cdot)$ is continuous and bounded, then, for any $x\in\mathbb{R}$ such that $f(x)\neq 0$, we have*

$$Q_t(x)\to Q(x), \quad \textit{almost surely, as } t\to+\infty.$$

Remark. Before proving the theorem, let us mention that the ε appearing in $Q_t(x)$ is added to make sure that the denominator will not vanish. On the other hand, examples of models which satisfy all conditions stated in § 4.1 are given in Banon's work (1977), (1978).

Proof of the theorem. Write

$$Q_t(x)=f_t'(x)\Big/\left[f_t(x)+\varepsilon\left\{\int_0^t h(s)H(h(s))\,ds\right\}^{-1}\right].$$

To prove the strong consistency of $Q_t(x)$, it is sufficient to have the strong consistency of $f_t^{(i)}(x)$, $i=0,1$, since under (b),

$$\int_0^t h(s)H(h(s))\,ds\to+\infty \quad \text{as } t\to+\infty.$$

Now under (i)–(v), we have seen that the process $(X_t, t \in \mathbb{R}^+)$ satisfies the assumptions (i)–(iii) of § 3.1. If in addition, $f'(\cdot)$ is continuous and bounded, then we are under all the sufficient conditions on the process to have strong consistent estimates of $f(x)$ and $f'(x)$.

Because $h(t)^{2k+1} = O(h(t)^{2k-1})$, as $t \to +\infty$, we note that:

(d) with $i = 1$ and $k \in [-1, -\frac{1}{2}] \Rightarrow$ (d) with $i = 0$,

(d) with $i = 1$ and $k \in [-\frac{1}{2}, \frac{1}{2}] \Rightarrow$ (d') with $i = 0$, $k \in [-\frac{1}{2}, \frac{1}{2}]$,

(d') with $i = 1 \Rightarrow$ (d') with $i = 0$.

Therefore, under the hypotheses of Theorem 4.1, we may apply Theorem 3.1 twice to obtain the announced result. □

Example. For $H(t) = t^k$ and $h(t) = t^{-a}$, we have $b = 1 - a(k+1)$. The corresponding sufficient condition for the almost-sure convergence of $Q_t(x)$ is that:

$$a \in (0, \tfrac{1}{6}) \quad \text{when } k \in [-1, \tfrac{1}{2}],$$

$$a \in \left(0, \frac{1}{2k+5}\right) \quad \text{when } k \geqq \tfrac{1}{2}.$$

This can be obtained by looking at (d) or (d'). For $k = \frac{1}{2}$, the strong consistency is obtained for any $a \in (0, \frac{1}{6})$.

COROLLARY (Strong consistent estimates of the drift term). *Under the conditions of Theorem* 4.1, *the recursive estimators* $m_t(x) = \frac{1}{2}(\{\sigma^2(x)\}' + \sigma^2(x) Q_t(x))$ *of the drift* $m(x)$, *at each point* x *such that* $f(x) \neq 0$, *is strongly consistent, i.e.,* $m_t(x) \to m(x)$, *almost surely as* $t \to +\infty$.

Remark. One could try to estimate $m(x)$ by using the fact that

$$m(x) = \lim_{t \to 0} \frac{1}{t} E(X_{t+s} - X_s | X_s = x).$$

The approach used in this paper is based on the relationship

$$m(x) = \tfrac{1}{2}(\sigma^2(x)' + \sigma^2(x) Q(x))$$

where $Q(x) = f'(x)/f(x)$.

Acknowledgment. We thank the referee for comments which led to considerable improvement of the presentation of this paper.

REFERENCES

G. BANON (1977), *Estimation nonparametrique de densité de probabilité pour les processus de Markov*, Thesis, Univ. Toulouse, France.

———, (1978), *Nonparametric identification for diffusion processes*, this Journal, 16, pp. 380–395.

P. DEHEUVELS (1974), *Conditions necessaires et suffisantes de convergence ponctuelle presque sure et uniforme des estimateurs de la densité*, C. R. Acad. Sci. Paris Sér. A, 287, pp. 1217–1220.

R. E. KALMAN AND R. C. BUCY (1961), *New results in linear filtering and prediction theory*, Trans. ASME Ser. D J. Basic Engrg., 83, pp. 95–108.

M. LOÈVE (1960), *Probability Theory*, Van Nostrand, New York.

H. T. NGUYEN (1978), *On estimation in Markov processes*, C. R. Acad. Sci. Paris Sér. A, 287, pp. 1129–1132.

———, (1979), *Density estimation in a continuous time, stationary Markov process*, Ann. Statist., 7, pp. 341–348.

M. ROSENBLATT (1970), *Density estimates and Markov sequences*, In Nonparametric Techniques in Statistical Inferences, M. Puri, ed., Cambridge University Press, London, pp. 199–210.

M. ROSENBLATT (1971), *Markov Processes, Structure and Asymptotic Behavior*, Springer-Verlag, New York.

H. YAMATO (1971), *Sequential estimation of a continuous probability density and mode*, Bull. Math. Stat. Jap., 14, pp. 1–12.

E. WONG AND M. ZAKAI (1965), *The oscillation of stochastic integrals*, Z. Wahrsch. Verw. Gebiete, 4, pp. 103–112.

E. WONG (1971), *Stochastic Processes in Information and Dynamical Systems*, McGraw-Hill, New York.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 5, September 1981

0363-0129/81/1905-0010 $01.00/0

AN ANALYSIS OF OPTIMAL MODAL REGULATION: CONVERGENCE AND STABILITY*

J. S. GIBSON†

Abstract. This paper treats the linear-quadratic regulator problem for infinite dimensional, second order (in time), linear oscillators. We solve the problem approximately by modeling a finite number of modes and obtaining a linear feedback control via the solution of a finite dimensional Riccati equation. This control we call the "modal control," and our analysis focuses on the convergence of the sequence of modal control laws corresponding to a sequence of models of increasing dimension. We seek conditions under which we can say that, as the number of modeled modes increases, the modal control law converges to a control law that is optimal for the full system, and that, if enough modes are modeled, the full closed-loop system that results from applying the modal control law to the actual system is stable. Roughly speaking, we have the convergence and stability we want if and only if we model enough damping to make the free system uniformly exponentially stable.

1. Introduction. The most common method for designing an active control scheme for a distributed system is to approximate the infinite dimensional system with a finite dimensional model and apply finite dimensional theory to an optimal control problem formulated for the model. The resulting control is then used for the actual system. In this paper, we investigate such a procedure for linear distributed systems represented by evolution equations of second order in time. Our finite dimensional optimal control problem is the standard linear-quadratic regulator problem, the solution of which follows from the solution of a Riccati matrix equation. Two closely related questions should be asked about a control scheme based on finite dimensional modeling of an infinite dimensional system: as the dimension of the model increases, does the control scheme somehow approach a control law which is optimal for the full system? and, if the model dimension is large enough, is the actual system guaranteed to perform satisfactorily? The first question is a question of convergence; the second, especially when posed about the closed-loop system resulting from a feedback control, is a question of stability. Answers to these questions are the goal of this paper, in which the results of [9] and [10] are applied to linear modal regulation.

Until recently, the common engineering philosophy and practice has been to assume that the answer to both questions is, in some sense, yes, for realistic systems and models. However, the increasing complexity and distinctly distributed nature of modern control systems such as electromagnetic and thermonuclear power plants and highly flexible aerospace structures have spawned a need for rigorous investigation of the convergence and stability questions stated above, particularly with regard to the modal regulation that has become so popular (see [3], [13], [17], [21] and their references). The need for such analysis in control theory for distributed systems is hardly surprising in view of the integral part that convergence analysis has come to play in the more developed theory of computational mechanics—for example, the finite element method for computing structural response to dynamic and static loading, or the characteristics of wave propagation in magnetic fields. Also as in computational mechanics, the purpose of the analysis is not just to say yes or no about the convergence of a particular approximation scheme, but studying the convergence

* Received by the editors April 16, 1979, and in revised form October 24, 1980. This research was supported by the National Science Foundation under grant ENG78-04753.

† Mechanics and Structures Department, School of Engineering and Applied Science, University of California, Los Angeles, California 90024.

properties of a modeling scheme can yield significant information about its desirability and improvement, in terms of both computational efficiency and design optimality.

This paper points to clear advantages of modeling inherent damping in modal regulation; the most important conclusion is that the answers to our convergence and stability questions are generally negative when no damping is modeled, and definitely positive when sufficient damping is modeled to provide a uniform decay rate for the free system. The difference results from the impossibility of using a compact linear feedback to give a uniform decay rate to an infinite dimensional linear oscillator when the free system does not already have a uniform decay rate (see [9]). As always in active control, the control vector here is finite dimensional. Hence the compactness of the feedback control laws.

It should help the reader to know at the start what is about to happen. In § 2 we define the distributed control system, lay some mathematical preliminaries concerning the semigroup which represents the free system response, and discuss the differences between the ways external damping and internal damping affect the free system response. Then, in § 3 we formulate the optimal control problem for the original infinite dimensional system and the sequence of finite dimensional approximate problems; in § 4 we give the pertinent results on the infinite dimensional regulator problem and convergence of sequences of approximate solutions; in § 5 we apply the results of § 4 to the approximation scheme of § 3 for systems without damping; and in § 6, for systems with damping.

The vectors spanning the finite dimensional space on which the approximate optimal control problems are defined we call "modes," and these may or may not be the natural modes of undamped, free vibration. From the solution of each approximate optimal control problem, we obtain a "modal control," which is a linear feedback control based on the modeled modes only, and our investigation focuses on the convergence of the sequence of modal control laws. In view of the central role played by the Riccati matrix equation in the finite dimensional regulator problem, it should not be surprising that the analysis of this paper hinges on an infinite dimensional Riccati equation corresponding to the optimal control problem for the full system and the approximate solution of this equation via the solutions to the sequence of finite dimensional Riccati equations corresponding to the approximate optimal control problems.

We obtain the finite dimensional optimal control problems by projecting the original infinite dimensional problem onto a sequence of finite dimensional spaces. In the previous literature on optimal control of distributed systems, a philosophical dichotomy has divided authors into a group with predominantly mathematical backgrounds and another group with predominately engineering backgrounds. The mathematicians have begun by defining an optimal control problem for the actual infinite dimensional system and then shown that its solution could be approximated by the solutions to a sequence of finite dimensional approximate problems, while the engineers have defined optimal control problems for their finite dimensional models only, without worrying much about what the solutions might approximate. It is true that the only optimization problems that actually will be solved for control parameters will be finite dimensional; however, if a control scheme based on finite dimensional modeling has any meaning with regard to optimal control of the actual system, then, as the dimension of the model increases, the control should converge to a control that is optimal for the full system, and the response of the full system should converge to the corresponding optimal response. Hence the usefulness of defining an optimal control problem for the infinite dimensional system and talking about convergence in

the full state space: the optimal control problem for the full system provides a single framework in which to study all the controls that are based on the sequence of models of increasing dimension. Even though the optimal control law for the full system may be somewhat abstract because it is defined in terms of infinite dimensional operators, by comparing, in the framework of the full state space, the finite dimensional control laws and the corresponding system responses to the ideal control law and system response, we are able to compare the finite dimensional control laws to one another and to address the convergence and stability questions that motivate our analysis.

Projecting the infinite dimensional control problem onto finite dimensional subspaces is a classical and fairly straightforward idea (for example, the Ritz method), which the technicalities of the procedure should not be allowed to obscure. In § 4, after projecting onto a finite dimensional subspace, the resulting optimal control problem is then artificially extended back to the full state space. This is done in order to satisfy the conditions of Theorem 4.2, which concerns convergence and is stated—as it must be—in terms of sequences of operators on the infinite dimensional state space. The thing to recognize is that the purpose of all this projecting and extending is to reconcile the finite dimensionality of the approximating models with the infinite dimensionality of the actual system. Once this reconciliation is accomplished, the theory of infinite dimensional Riccati equations sheds the light we need on the boundedness and convergence of the sequence of finite dimensional control laws and the stability of the full distributed system when these control laws are applied to it.

2. The distributed control system. This paper deals with control systems represented by the second order (in t) differential equation

$$\ddot{x}(t)+\mathscr{C}_0\dot{x}(t)+\mathscr{A}_0x(t)=\mathscr{B}_0u(t), \qquad t\geqq 0, \tag{2.1}$$

where $x(t)$ is in a real Hilbert space H and $u(t)$ is in a real, finite dimensional Hilbert space U; $\mathscr{A}_0$ is a self-adjoint linear operator from $D(\mathscr{A}_0)$, which is dense in H, onto H; $\mathscr{A}_0$ is coercive, i.e., there exists $\rho>0$ such that

$$\langle \mathscr{A}_0x, x\rangle_H \geqq \rho^2\|x\|_H^2, \qquad x\in D(\mathscr{A}_0); \tag{2.2}$$

and $\mathscr{A}_0^{-1}$ is compact. $\mathscr{C}_0$ is a nonnegative, symmetric linear operator from $D(\mathscr{C}_0)$, which contains $D(\mathscr{A}_0)$, to H, and there exists $\gamma\geqq 0$ such that

$$\|\mathscr{C}_0x\|_H \leqq \gamma^2\|\mathscr{A}_0x\|_H, \qquad x\in D(\mathscr{A}_0). \tag{2.3}$$

$\mathscr{B}_0$ is a bounded linear operator from U to H.

There are important examples where $\mathscr{C}_0$ is unbounded; for example, flexible mechanical systems with internal damping. However, we should have at least (2.3) in physical systems. We take U to be finite dimensional because any real active controller has only a finite number of control variables. Note that $\mathscr{B}_0\in\mathscr{L}(U,H)$ means that, if $\dim(U)=m$, so that $u(t)=(u^{(1)}(t), u^{(2)}(t), \ldots, u^{(m)}(t))$, then

$$\mathscr{B}_0u(t)=\sum_{i=1}^{m} b^{(i)}u^{(i)}(t), \tag{2.4}$$

where each $b^{(i)}$ is an element of H.

Now we will define the natural "energy space" and write (2.1) in first order form. Assuming that H is infinite dimensional, we know (see [12, pp. 187, 260], [25, p. 343]) that the spectrum of $\mathscr{A}_0$ is an infinitely increasing sequence of positive real eigenvalues ω_n^2, each of finite multiplicity, and that the corresponding mutually orthogonal eigenvectors ϕ_n comprise a complete basis in H. Of course, the ω_n's and ϕ_n's are, respectively,

the natural frequencies and mode shapes of free, undamped oscillations. As usual, we define the space $E = V \times H$, where $V = D(\mathscr{A}_0^{1/2})$ is a Hilbert space with inner product $\langle v_1, v_2\rangle_V = \langle \mathscr{A}_0^{1/2} v_1, \mathscr{A}_0^{1/2} v_2\rangle_H$; E has the energy inner product $\langle (v_1, h_1), (v_2, h_2)\rangle_E = \langle v_1, v_2\rangle_V + \langle h_1, h_2\rangle_H$. In defining E, we do not identify V with its natural image in H, so that we are free to identify E with its dual, as we do henceforth. The eigenvectors of $\mathscr{A}_0$ are also mutually orthogonal and complete in V, and the pairs $(\phi_n, 0)$ and $(0, \phi_n)$ are thus mutually orthogonal and complete in E.

Next we construct the generator of the semigroup that provides the homogeneous solution to (2.1). We begin with the operator $\mathring{\mathscr{A}}$ defined by

$$\mathring{\mathscr{A}} = \begin{bmatrix} & I \\ -\mathscr{A}_0 & -\mathscr{C}_0 \end{bmatrix}, \qquad D(\mathring{\mathscr{A}}) = D(\mathscr{A}_0) \times D(\mathscr{A}_0), \tag{2.5}$$

and seek an extension of $\mathring{\mathscr{A}}$ which generates a strongly continuous semigroup. (Strictly speaking, the I in $\mathring{\mathscr{A}}$ is the inverse of the natural injection of V into H—think identity.) We know (see [18], [24, p. 62]) that a linear operator $\mathscr{A}$ on a Hilbert space E generates a strongly continuous contraction semigroup if and only if $\mathscr{A}$ is densely defined and maximal dissipative. $\mathscr{A}$ is said to be dissipative if

$$\langle \mathscr{A}y, y\rangle_E \leqq 0, \qquad y \in D(\mathscr{A}), \tag{2.6}$$

and maximal dissipative if $\mathscr{A}$ is not a proper restriction of another dissipative operator. (All operators here are linear.) While the $\mathring{\mathscr{A}}$ of (2.5) is densely defined and dissipative, in general, it is not maximal dissipative.

Any dissipative operator has a maximal dissipative extension (see [18], [24, p. 20]). We will construct explicitly the unique maximal dissipative extension of the $\mathring{\mathscr{A}}$ of (2.5). Since a one to one correspondence exists between the class of semigroups on E and the class of generators, if $\mathring{\mathscr{A}}$ had more than one maximal dissipative extension, we would face a troublesome uncertainty about which semigroup, if any, provided the solution to the evolution equation we want to solve ((2.1) or (2.13) below). Fortunately, we have the following result.

THEOREM 2.1. *Let $\mathring{\mathscr{A}}$ be a densely defined, dissipative linear opeator on a Hilbert space E.*

(i) *Suppose that the range of $\mathring{\mathscr{A}}$, denoted by $R(\mathring{\mathscr{A}})$, is dense and $\mathring{\mathscr{A}}$ has a bounded inverse $\mathring{\mathscr{A}}^{-1}$. Let $\mathscr{A}^{-1}$ be the bounded extension of $\mathring{\mathscr{A}}^{-1}$ to all of E. Then $\mathscr{A} \equiv (\mathscr{A}^{-1})^{-1}$ is the unique maximal dissipative extension of $\mathring{\mathscr{A}}$, and $R(\lambda - \mathring{\mathscr{A}})$ is dense for $\lambda > 0$.*

(ii) *If $R(\lambda_0 - \mathring{\mathscr{A}})$ is dense for some $\lambda_0 > 0$, then $\mathring{\mathscr{A}}$ has a unique maximal dissipative extension, and $R(\lambda - \mathring{\mathscr{A}})$ is dense for $\lambda > 0$.*

Proof. (i) If a maximal dissipative operator is invertible, its inverse is maximal dissipative. Also, it is easy to show that a dissipative operator with dense range is one to one: Suppose $\tilde{\mathscr{A}}$ is such an operator, $\|x\| = 1$, and $\tilde{\mathscr{A}}x = 0$. Since $R(\tilde{\mathscr{A}})$ is dense, we can choose $y \in D(\tilde{\mathscr{A}})$ such that $\|\tilde{\mathscr{A}}y - x\| < 1/2$. Then, with α a positive real number, $\langle \tilde{\mathscr{A}}(\alpha x + y), (\alpha x + y)\rangle = \alpha\langle \tilde{\mathscr{A}}y, x\rangle + \langle \tilde{\mathscr{A}}y, y\rangle \geqq \alpha/2 + \langle \tilde{\mathscr{A}}y, y\rangle$, which is positive for α sufficiently large, contradicting the dissipativeness of $\tilde{\mathscr{A}}$. Hence, if a maximal dissipative operator has dense range, its inverse is maximal dissipative. Thus $\mathscr{A}^{-1}$ and $(\mathscr{A}^{-1})^{-1}$ are maximal dissipative.

Since any maximal dissipative linear operator with dense domain is closed (see [18], [24]) and since $R(\mathring{\mathscr{A}})$ is dense, any maximal dissipative extension of $\mathring{\mathscr{A}}$ must have an inverse which is a closed extension of $\mathring{\mathscr{A}}^{-1}$. But, since $\mathring{\mathscr{A}}^{-1}$ is densely defined and bounded, it has a unique closed extension, namely, $\mathscr{A}^{-1}$. Therefore, $\mathscr{A} = (\mathscr{A}^{-1})^{-1}$ is the unique maximal dissipative extension of $\mathring{\mathscr{A}}$.

Since $\mathring{\mathcal{A}}^{-1}$ is densely defined and bounded, the graph of $\mathring{\mathcal{A}}$ is dense in the graph of $\mathcal{A}$, and therefore the graph of $\lambda - \mathring{\mathcal{A}}$ is dense in the graph of $\lambda - \mathcal{A}$ for all (complex) λ. In particular, $R(\lambda - \mathring{\mathcal{A}})$ is dense in $R(\lambda - \mathcal{A})$. Since $R(\lambda - \mathcal{A}) = E$ for $\lambda > 0$, $R(\lambda - \mathring{\mathcal{A}})$ is dense in E for $\lambda > 0$.

(ii) Let $\mathcal{A}$ be a maximal dissipative extension of $\mathring{\mathcal{A}}$. Then $(\lambda - \mathcal{A})^{-1} \in \mathcal{L}(E, E)$, $\lambda > 0$. Since $(\lambda_0 - \mathring{\mathcal{A}})^{-1}$ is bounded and densely defined and $(\lambda_0 - \mathcal{A})^{-1}$ is a bounded extension of $(\lambda_0 - \mathring{\mathcal{A}})^{-1}$ to all of E, $(\lambda_0 - \mathcal{A})^{-1}$ is uniquely determied. Thus $\mathcal{A}$ is uniquely determined.

Since $(\lambda_0 - \mathring{\mathcal{A}})^{-1}$ is densely defined and bounded, the graph of $\lambda_0 - \mathring{\mathcal{A}}$ is dense in the graph of $\lambda_0 - \mathcal{A}$, so that, as before, $R(\lambda - \mathring{\mathcal{A}})$ is dense for any λ in the resolvent set of $\mathcal{A}$.

Now we will derive the maximal dissipative extension of the $\mathring{\mathcal{A}}$ defined in (2.5).

THEOREM 2.2. *Under our hypotheses on $\mathcal{A}_0$ and $\mathcal{C}_0$, $\mathcal{A}^{-1}\mathcal{C}_0$ has a bounded extension to all of V, which we will denote by $\widetilde{\mathcal{A}_0^{-1}\mathcal{C}_0}$.*

Proof. Let $v \in D(\mathcal{A}_0)$. Then $\mathcal{A}_0^{-1}\mathcal{C}_0 v \in D(\mathcal{A}_0)$, and

$$\begin{aligned} \|\mathcal{A}_0^{-1}\mathcal{C}_0 v\|_V^2 &= \langle \mathcal{C}_0 v, \mathcal{A}_0^{-1}\mathcal{C}_0 v\rangle_H \leqq \langle \mathcal{C}_0 v, v\rangle_H^{1/2} \langle \mathcal{C}_0\mathcal{A}_0^{-1}\mathcal{C}_0 v, \mathcal{A}_0^{-1}\mathcal{C}_0 v\rangle_H^{1/2} \\ &\leqq \gamma \langle \mathcal{A}_0 v, v\rangle_H^{1/2} \gamma \langle \mathcal{A}_0\mathcal{A}_0^{-1}\mathcal{C}_0 v, \mathcal{A}_0^{-1}\mathcal{C}_0 v\rangle_H^{1/2} = \gamma^2 \|v\|_V \|\mathcal{A}_0^{-1}\mathcal{C}_0 v\|_V. \end{aligned} \tag{2.7}$$

The first inequality follows from the generalized Schwarz inequality, and the second, from the fact that our hypotheses on $\mathcal{A}_0$ and $\mathcal{C}_0$ imply (see [12, p. 292, Thm. 4.12])

$$\langle \mathcal{C}_0 x, x\rangle_H \leqq \gamma^2 \langle \mathcal{A}_0 x, x\rangle_H, \qquad x \in D(\mathcal{A}_0). \tag{2.8}$$

The theorem follows from (2.7).

We can see easily that $R(\mathring{\mathcal{A}}) = D(\mathring{\mathcal{A}}^{-1}) = D(\mathcal{A}_0) \times H$, and that $\mathcal{A}^{-1}$ (the extension of $\mathring{\mathcal{A}}^{-1}$ to E) is

$$\mathcal{A}^{-1} = \begin{bmatrix} -\widetilde{\mathcal{A}_0^{-1}\mathcal{C}_0} & -\mathcal{A}_0^{-1} \\ I & 0 \end{bmatrix}. \tag{2.9}$$

Then, as in Theorem (2.1), we have $\mathcal{A} = (\mathcal{A}^{-1})^{-1}$, where $\mathcal{A}$ is the unique maximal dissipative extension of $\mathring{\mathcal{A}}$. $D(\mathcal{A})$ is just $R(\mathcal{A}^{-1})$. This $\mathcal{A}$ generates the semigroup $\mathcal{T}(\cdot)$ that represents the free response of the control system of (2.1), and in general we have dissipation of energy:

$$\|\mathcal{T}(t)y\|_E \leqq \|y\|_E, \qquad t \geqq 0, \qquad y \in E. \tag{2.10}$$

If $\mathcal{C}_0 = 0$, we have conservation of energy, i.e., equality in (2.10).

Remark 2.1. So far, the only place we have needed $\mathcal{C}_0$ to be symmetric is the proof of Theorem 2.2. Since any bounded perturbation of $\mathcal{A}$ results in a semigroup generator, we could require only that the unbounded part of $\mathcal{C}_0$ be nonnegative and symmetric (and $\mathcal{A}_0$-bounded) in order to extend $\mathring{\mathcal{A}}$ to a semigroup generator $\mathcal{A}$. As long as $\mathcal{C}_0$ is nonnegative, $\mathring{\mathcal{A}}$ and $\mathcal{A}$ will be dissipative and $\mathcal{T}(\cdot)$ will be a contraction semigroup. In subsequent sections, the only place where the symmetry of $\mathcal{C}_0$ itself, instead of its unbounded part only, seems essential is the proof of Theorem 6.1. For convenience, we will continue to assume that $\mathcal{C}_0$ is symmetric.

To solve the optimal control problem of this paper,we will need the adjoints of $\mathcal{A}$ and $\mathcal{T}(\cdot)$, which we denote by $\mathcal{A}^*$ and $\mathcal{T}^*(\cdot)$, respectively. To construct $\mathcal{A}^*$, we first observe that $(\mathcal{A}^{-1})^* = \mathcal{A}^{-*} \in \mathcal{L}(E, E)$, and

$$\mathcal{A}^{-*} = \begin{bmatrix} -\widetilde{\mathcal{A}_0^{-1}\mathcal{C}_0} & \mathcal{A}_0^{-1} \\ -I & 0 \end{bmatrix}. \tag{2.11}$$

(Remember that this is the adjoint of $\mathscr{A}^{-1}$ with respect to the energy inner product.) We have then $\mathscr{A}^* = (\mathscr{A}^{-*})^{-1}$. Also, $D(\mathscr{A}^*) = R(\mathscr{A}^{-*}) \supset D(\mathscr{A}_0) \times D(\mathscr{A}_0)$, and

$$\mathscr{A}^*|_{D(\mathscr{A}_0)\times D(\mathscr{A}_0)} = \begin{bmatrix} 0 & -I \\ \mathscr{A}_0 & -\mathscr{C}_0 \end{bmatrix}. \tag{2.12}$$

In general the restriction of $\mathscr{A}^*$ to $D(\mathscr{A}_0) \times D(\mathscr{A}_0)$ is not closed and is not equal to $\mathring{\mathscr{A}}^*$.

Clearly, $\mathscr{A}^*$ and $\mathscr{A}^{-*}$ are maximal dissipative; also we have the general result (see [18], [24, p. 21]) that the adjoint of a maximal dissipative operator with dense domain is maximal dissipative with dense domain. The operator $\mathscr{A}^*$ generates the semigroup $\mathscr{T}^*(\cdot)$.

With $y = (x, \dot{x}) \in E$, the first order form of (2.1) is

$$\dot{y}(t) = \mathscr{A}y(t) + \mathscr{B}u(t), \qquad t \geqq 0, \tag{2.13}$$

where

$$\mathscr{B} = \begin{bmatrix} 0 \\ \mathscr{B}_0 \end{bmatrix} \in \mathscr{L}(U, E). \tag{2.14}$$

Actually, since we want to allow $y(0)$ to be any element of E and $u(\cdot)$ to be any element of $L_2(0, \infty; U)$, in general we will have only the integral version of (2.13) (see (4.1) in § 4 below), which is written in terms of $\mathscr{T}(\cdot)$ rather than $\mathscr{A}$.

Because we allow $\mathscr{C}_0$ to be unbounded, $\mathscr{T}(\cdot)$ is, as we will see, in general only a semigroup of bounded linear operators on E. The difference between $\mathscr{C}_0$ being bounded and $\mathscr{C}_0$ being unbounded and the resulting differences in the spectra of $\mathscr{A}$ and $\mathscr{T}(\cdot)$ have both mathematical and physical significance. If $\mathscr{C}_0 \in \mathscr{L}(H, H)$, $D(\mathscr{A}) = D(\mathscr{A}_0) \times V$ and $\mathscr{A}^{-1}$ is compact (see [9]). But neither is the case in general; for example, take $\mathscr{C}_0 = \mathscr{A}_0$. Also, if $\mathscr{C}_0 \in \mathscr{L}(H, H)$, $\mathscr{T}(\cdot)$ is a group in $\mathscr{L}(H, H)$. The standard argument notes that, for $\mathscr{C}_0 = 0$, both $\mathscr{A}$ and $-\mathscr{A}$ generate strongly continuous semigroups, and that ([11, p. 390]) a bounded linear perturbation of a group generator yields a group generator.

To get an idea of how the type of boundedness of $\mathscr{C}_0$ affects the response of the system (2.1), consider the free vibration of a simply supported or cantilevered beam. The operator $\mathscr{A}_0$ is then a fourth order partial differential operator. First, suppose that the beam is subjected to no external damping, but is made of a material modeled by the Voigt–Kelvin (see [19]) model for linear viscoelasticity. Then (see [5, pp. 301–302]) $\mathscr{C}_0 = c_0\mathscr{A}_0$, where c_0 is a positive constant. The natural modes of free vibration remain uncoupled in the presence of the internal damping represented by $\mathscr{C}_0$, and the eigenvalues of $\mathscr{A}$ corresponding to the nth mode are

$$\lambda_n = (-c_0\omega_n^2 \pm \sqrt{c_0^2\omega_n^4 - 4\omega_n^2})/2, \qquad n \geqq 1. \tag{2.15}$$

(To get λ_n, set $u(t) = 0$ in (2.1), expand $x(t)$ in terms of the natural modes, and take the H-inner product of ϕ_n with the resulting equation.) Note that the eigenvalues are complex for only a finite number of modes. As ω_n approaches ∞, the values of λ_n approach $-\infty$ and $-1/c_0$. When $\mathscr{A}^{-1}$ is not compact, the spectrum of $\mathscr{A}$ may contain points other than eigenvaalues, and in the present example the spectrum of $\mathscr{A}$ consists precisely of the sequence of eigenvalues in (2.15) and the continuous spectrum $\{-1/c_0\}$.

Since $\mathscr{A}$ has a sequence of eigenvalues whose real parts approach $-\infty$, 0 turns up in the spectrum of $\mathscr{T}(t)$ for $t > 0$, so that $\mathscr{T}(\cdot)$ is not a group of bounded linear operators. (It is instructive here to think of 0 as $e^{-\infty t}$ and $\mathscr{T}(t)$ as $e^{\mathscr{A}t}$, for $t > 0$.) As a matter of fact, 0 is in the continuous spectrum of $\mathscr{T}(t)$ for $t > 0$.

Generally, (2.3) guarantees that the system damping is sufficiently bounded relative to the stiffness to keep the spectrum of $\mathcal{A}$ bounded away from zero. As the next examples illustates, this is necessary for $\mathcal{T}(\cdot)$ to be uniformly exponentially stable.

For an example where $\mathcal{C}_0$ is not bounded relative to $\mathcal{A}_0$, let $\mathcal{C}_0 = c_0\mathcal{A}_0^2$. While this $\mathcal{C}_0$ does not represent any physical damping of which the author is aware, the example is instructive because the resulting semigroup is not uniformly exponentially stable, even though $\mathcal{C}_0$ is positive definite. Our subsequent Theorem 6.1 says that, if $\mathcal{C}_0$ is $\mathcal{A}_0$-bounded and positive definite, the semigroup $\mathcal{T}(t)$ is uniformly exponentially stable, but of course $\mathcal{A}_0^2$ is not $\mathcal{A}_0$-bounded for the present example; i.e., $\mathcal{C}_0$ does not satisfy (2.3). Referring to (2.5) with $D(\mathring{\mathcal{A}}) = D(\mathcal{A}_0) \times D(\mathcal{A}_0^2)$, we could show that $R(\lambda - \mathring{\mathcal{A}})$ is dense for $\lambda > 0$, so that part (ii) of Theorem 2.1 says that $\mathring{\mathcal{A}}$ has a unique maximal dissipative extension, which generates a semigroup. Since the eigenvectors of $\mathcal{A}_0$ are also eigenvectors of $\mathcal{A}_0^2$, the natural modes again remain uncoupled, but this time the eigenvalues corresponding to the nth mode are

$$\lambda_n = (-c_0\omega_n^4 \pm \sqrt{c_0^2\omega_n^8 - 4\omega_n^2})/2, \qquad n \geqq 1. \tag{2.16}$$

As ω_n approaches ∞, these two values of λ_n approach $-\infty$ and 0. Therefore, $\mathcal{T}(\cdot)$ does not have a uniform decay rate.

Since $\mathcal{A}$ has a sequence of eigenvalues approaching 0, 1 is in the spectrum of $\mathcal{T}(t)$ for $t \geqq 0$. Actually, though we have not bothered to define $D(\mathcal{A})$ explicitly for this case, it is not difficult to see that 0 is in the continuous spectrum of $\mathcal{A}$, and hence 1 is in the continuous spectrum of $\mathcal{T}(t)$ for $t > 0$, Since the norm of $\mathcal{T}(t)$ is greater than or equal to the spectral radius, and since we already know that $\mathcal{T}(\cdot)$ is a contraction semigroup, we have $\|\mathcal{T}(t)\| = 1$, $t \geqq 0$.

Now suppose that our beam is made of a linearly elastic material, but is surrounded by a linearly viscious fluid. We have external damping and $\mathcal{C}_0 \in \mathcal{L}(H, H)$ (see [5, p. 301]). Hence, $\mathcal{T}(\cdot)$ is a strongly continuous group in $\mathcal{L}(E, E)$, and $\mathcal{A}^{-1}$ is compact. Since $\mathcal{A}$ has compact resolvent, the spectrum of $\mathcal{A}$ consists of eigenvalues with finite multiplicities, with no finite accumulation point (see [12, p. 187]). Since $\mathcal{T}^{-1}(t) = \mathcal{T}(-t) \in \mathcal{L}(E, E)$ for $-\infty < t < \infty$, the spectrum of $\mathcal{A}$ cannot contain a sequence whose real parts approach $-\infty$.

3. The optimal control problem and the modal approximation scheme. To the extent that we can cope with an infinite number of modes, the most natural optimal regulation problem for the control system here, in the case of unconstrained control, is: given $y(0)$ in E, choose the control $u \in L_2(0, \infty; U)$ which minimizes the cost functional

$$J(y(0), u) = \int_0^\infty (\langle \mathcal{D}y(t), y(t)\rangle_E + \langle Qu(t), u(t)\rangle_U)\, dt, \tag{3.1}$$

where $\mathcal{D} = \mathcal{D}^* \in \mathcal{L}(E, E)$, $Q = Q^* \in \mathcal{L}(U, U)$, and both $\mathcal{D}$ and Q are positive definite. Of course, (3.1) assumes that there is a control for which J is finite; we will discuss certain necessary and sufficient conditions later. It is important that $\mathcal{D}$ be positive definite so that, whenever J is finite, all the energy is driven out of the system, and, whenever J can be made finite for all $y(0)$, the optimal feedback sytem is asymptotically stable (see [8], [10], and Theorem 4.1 of the next section). The optimal control problem just stated will be referred to as "the optimal control problem on E."

While we can prove theorems about existence and uniqueness of an optimal control for this problem and stability of the resulting closed-loop system (see [6], [8], [10], [15], [16], and § 4 of this paper), the infinite dimensionality of the problem

prevents us from computing the optimal control scheme exactly, which we can do for the finite dimensional linear regulator. Therefore, we seek methods to approximate the optimal control scheme, and the most common approach, at least for flexible systems, is called modal control (see [3], [13], and their references). In modal control, a finite number of modes are modeled and a control problem for the approximate model is formulated and solved for the control scheme to be used for the original infinite dimensional system.

We will denote by ψ_n, $n \geqq 1$, the basis vectors for the approximation scheme, and we assume the following:

HYPOTHESIS 3.1. *The vectors ψ_n, $n = 1, 2, \ldots$, are linearly independent and are complete in $D(\mathscr{A}_0)$ when $D(\mathscr{A}_0)$ is a Hilbert space with $\| \cdot \|_{D(\mathscr{A}_0)} = \|\mathscr{A}_0 \cdot \|_H$.*

Certainly, the natural modes ϕ_n satisfy this hypothesis. While in modal control the vectors ψ_n are usually taken to be the natural modes, this is not necessary for our analysis. However, though the ψ_n's can be shape functions other than natural modes, in view of the current popularity of modal control schemes for flexible systems, we will refer to the ψ_n's as "modes," in an attempt to make the implications of our analysis as concrete as possible. Whenever it is important that the ψ_n's be the eigenvectors of $\mathscr{A}_0$, we will use the term "natural modes." The modal control problem is formulated by projecting the optimal control problem on E onto subspaces spanned by finite combinations of the ψ_n's.

Suppose then that we model the first n modes of the system (2.1). Let $H_n = \operatorname{span}\{\psi_j\}_{j \leqq n}$ with $\langle \cdot, \cdot \rangle_{H_n} = \langle \cdot, \cdot \rangle_H$, $V_n = \operatorname{span}\{\psi_j\}_{j \leqq n}$ with $\langle \cdot, \cdot \rangle_{V_n} = \langle \cdot, \cdot \rangle_V$, and $E_n = V_n \times H_n$ with $\langle \cdot, \cdot \rangle_{E_n} = \langle \cdot, \cdot \rangle_E$. Denote the (orthogonal) projection operator from H onto H_n by Λ_{H_n}, the projection operator from V onto V_n by Λ_{V_n}, the projection from E onto E_n by Λ_n. If the ψ_j's are the natural modes, then $\Lambda_{V_n} = \Lambda_{H_n}|_V$, but in general we have only $\Lambda_{V_n}|_{V_n} = \Lambda_{H_n}|_{H_n} = I$. Define the operators $\mathscr{A}_{0_n} = \Lambda_{H_n}\mathscr{A}_0\Lambda_{V_n}$, $\mathscr{C}_{0_n} = \Lambda_{H_n}\mathscr{C}_0\Lambda_{H_n}$, $\mathscr{B}_{0_n} = \Lambda_{H_n}\mathscr{B}_0$, $\mathscr{A}_n = \Lambda_n\mathscr{A}\Lambda_n$, $\mathscr{B}_n = \Lambda_n\mathscr{B}$ and $\mathscr{D}_n = \Lambda_n\mathscr{D}\Lambda_n$. Denote the restriction of $\mathscr{A}_{0_n}$ to H_n by A_{0_n}, and similarly set $C_{0_n} = \mathscr{C}_{0_n}|_{H_n}$, $B_{0_n} = \mathscr{B}_{0_n}$, $A_n = \mathscr{A}_n|_{E_n}$, $B_n = \mathscr{B}_n$, and $D_n = \mathscr{D}_n|_{E_n}$.

Of course, the finite dimensional operators A_{0_n}, C_{0_n}, B_{0_n}, A_n, B_n, and D_n can be identified with appropriate matrices; for example, if we denote by $\tilde{A}_{0_n}$ the matrix representing the operator A_{0_n} with respect to the basis vectors ψ_i, $1 \leqq i \leqq n$, we have

$$\tilde{A}_{0_n} = \Psi_{0_n}^{-1}\tilde{\tilde{A}}_{0_n}, \tag{3.2}$$

where $\tilde{\tilde{A}}_{0_n}$ and Ψ_{0_n} are $n \times n$ matrices whose elements are $\langle \mathscr{A}_0\psi_i, \psi_j \rangle_H$ and $\langle \psi_i, \psi_j \rangle_H$, respectively. In particular, if the ψ_n's are the natural modes ϕ_n, then $\tilde{A}_{0_n}$ is a diagonal matrix containing the first n eigenvalues ω_i^2 of $\mathscr{A}_0$, reapeated according to geometric multiplicity. For expressions and equations involving operators on E, like $e^{\mathscr{A}_n t} = e^{A_n t}\Lambda_n + I - \Lambda_n$, we must remember the difference between an operator and a matrix representing that operator, but we may interpret expressions and equations involving only finite dimensional operators, like $e^{A_n t}$ or the finite dimensional Riccati equation (3.13), in terms of the matrices representing the operators.

Note the identities

$$\mathscr{A}_n = \begin{bmatrix} 0 & I\Lambda_{H_n} \\ -\mathscr{A}_{0_n} & -\mathscr{C}_{0_n} \end{bmatrix}, \quad A_n = \begin{bmatrix} 0 & I \\ -A_{0_n} & -C_{0_n} \end{bmatrix}, \quad \text{and} \quad \mathscr{A}_n|_{E_n^\perp} = 0. \tag{3.3}$$

The I in $\mathscr{A}_n$ is the inverse of the natural injection of V into H, as in (2.5), and the I in A_n is the natural identification of V_n with H_n. The subspaces E_n and $E_n^\perp$ are invariant

under $\mathscr{A}_n$ and $e^{\mathscr{A}_n t}$, and we have

$$e^{\mathscr{A}_n t}|_{E_n} = e^{A_n t} \quad \text{and} \quad e^{\mathscr{A}_n t}|_{E_n^\perp} = I, \qquad -\infty < t < \infty. \tag{3.4}$$

We have similar identities and properties for $\mathscr{A}_n^* = (\Lambda_n \mathscr{A} \Lambda_n)^*$.

Also note

$$\|e^{\mathscr{A}_n t}\| = \|e^{\mathscr{A}_n^* t}\| \leqq 1, \qquad t \geqq 0. \tag{3.5}$$

As we will see in the next section, the extension of $e^{A_n t}$ to all of E, as $e^{\mathscr{A}_n t}$, is useful because it allows us to talk about convergence in E.

To deduce the convergence we will need for $e^{\mathscr{A}_n t}$ (as $n \to \infty$), we apply the following Trotter-Kato approximation result (see [12, p. 504, Theorem 2.16]):

THEOREM 3.1. *Let $\mathscr{A}$ and $\mathscr{A}_n$, $n = 1, 2, \ldots$, generate strongly continuous contraction semigroups $\mathscr{T}(\cdot)$ and $\mathscr{T}_n(\cdot)$, respectively, on E. If*

$$(\lambda - \mathscr{A}_n)^{-1} \to (\lambda - \mathscr{A})^{-1} \quad \text{strongly} \tag{3.6}$$

for some λ with $\operatorname{Re} \lambda > 0$, then

$$\mathscr{T}_n(t) \to \mathscr{T}(t) \quad \text{strongly}, \tag{3.7}$$

uniformly in any finite interval of $t \geqq 0$. Conversely, if (3.7) holds for all t in an interval of positive length, then (3.6) holds for every λ with $\operatorname{Re} \lambda > 0$.

THEOREM 3.2. *For the modal approximation scheme we have defined, we have*

$$e^{\mathscr{A}_n t} \to \mathscr{T}(t) \quad \text{strongly}, \qquad 0 \leqq t < \infty, \tag{3.8}$$

$$e^{\mathscr{A}_n^* t} \to \mathscr{T}^*(t) \quad \text{strongly}, \qquad 0 \leqq t < \infty, \tag{3.9}$$

and the convergence is uniform in t for t in bounded intervals; i.e., for each $y \in E$, $e^{\mathscr{A}_n t} y \overset{E}{\to} \mathscr{T}(t) y$ uniformly in each bounded t-interval, and similarly for $e^{\mathscr{A}_n^* t}$.

Proof. Let $\lambda > 0$. Writing $(\lambda - \mathscr{A}_n)^{-1} - (\lambda - \mathscr{A})^{-1} = (\lambda - \mathscr{A}_n)^{-1}(\mathscr{A}_n - \mathscr{A})(\lambda - \mathscr{A})^{-1}$ and noting $\|(\lambda - \mathscr{A}_n^{-1})\| \leqq 1/\lambda$, $n \geqq 1$, we see that (3.6) holds if $(\mathscr{A}_n - \mathscr{A})(\lambda - \mathscr{A})^{-1} y \to 0$ for y in a dense subset of E. Let $\hat{E} = \bigcup_{n \geqq 1} E_n$. For $\hat{y} \in \hat{E}$, we have $\mathscr{A}_n \hat{y} = \Lambda_n \mathscr{A} \hat{y}$ for n sufficiently large, and thus $\mathscr{A}_n \hat{y} \to \mathscr{A} \hat{y}$. Now we need only show that $(\lambda - \mathscr{A})(\hat{E})$ is dense in E to prove that $(\lambda - \mathscr{A}_n)^{-1} \to (\lambda - \mathscr{A})^{-1}$ strongly.

According to Theorem 2.1, $R(\lambda - \mathring{\mathscr{A}}) = (\lambda - \mathring{\mathscr{A}})(D(\mathscr{A}_0) \times D(\mathscr{A}_0))$ is dense in E. Let v, $h \in D(\mathscr{A}_0)$. Then, by Hypothesis 3.1, we can choose a sequence $(\hat{v}_n, \hat{h}_n)$ in $\hat{E}$ such that $\|\mathscr{A}_0(\hat{v}_n - v)\|_H + \|\mathscr{A}_0(\hat{h}_n - h)\|_H \to 0$. Then, since $\mathscr{A}_0^{1/2}$, $\mathscr{C}_0$, and the identity are bounded by $\mathscr{A}_0$ (as in (2.3)), $(\lambda - \mathscr{A})(\hat{v}_n, \hat{h}_n) \to (\lambda - \mathscr{A})(v, h)$. Thus, $(\lambda - \mathscr{A})(\hat{E})$ is dense in $R(\lambda - \mathring{\mathscr{A}})$, which is dense in E.

In the same manner, we can prove that $(\lambda - \mathscr{A}_n^*)^{-1} \to (\lambda - \mathscr{A}^*)^{-1}$ strongly, by using the restriction of $\mathscr{A}^*$ to $D(\mathscr{A}_0) \times D(\mathscr{A}_0)$ instead of $\mathring{\mathscr{A}}$.

At this point, we should comment on the classes of vectors ψ_n that satisfy our hypotheses. Actually, the requirement that the ψ_n's be in $D(\mathscr{A}_0)$ considerably restricts the choices for these shape functions. In applications, the Hilbert spaces H and V are spaces of functions on a finite dimensional region and $\mathscr{A}_0$ is a partial differential operator. For $\mathscr{A}_0$ to be coercive, enabling us to use the energy norm for E, $D(\mathscr{A}_0)$ must be restricted to those functions which, along with their spatial derivatives, satisfy boundary conditions that result from the physics of the particular problem. Thus, while we do not require the ψ_n's to be the natural modes, we do require them to satisfy the same natural boundary conditions that the natural modes must satisfy. Of course, the functions in $D(\mathscr{A}_0)$ must be sufficiently smooth.

It is possible for an approximation scheme like the modal scheme we have described to yield the convergence of (3.8) and (3.9) but not satisfy our hypothesis that the basis vectors form a basis for $D(\mathscr{A}_0)$—either because the basis vectors are not sufficiently smooth or because they do not satisfy the natural boundary conditions. For the convergence results of the subsequent sections pertaining to approximate solution of the infinite dimensional regulator problem, all we really need is the convergence of Theorem 3.2. (See Remark 4.1.).

However, the shape functions used in modal control schemes ordinarily should satisfy the hypotheses we have stated for the ψ_n's because, for a control law of a given finite order to be most effective, it should be based on mode shapes that represent the most significant motions of the system to be controlled. Most often, these motions are the free oscillations of the system. We allow shape functions other than the natural mode shapes (i.e., the eigenvectors of $\mathscr{A}_0$) because, for complex control systems, some of the ψ_n's are often taken to represent motion of certain components relative to the overall structure. For example, a satellite with a rigidly attached flexible boom might be modeled by the rigid body motion of the central body plus the natural vibration modes of a cantilevered beam for the boom. In this and similar examples the "modal coordinates" do not represent the natural modes of the composite system, but the shape functions ψ_n represent physically significant motions and satisfy the natural boundary conditions and smoothness requirements implicit in Hypothesis 3.1.

In view of the prominence of the finite element method, it is natural to ask whether the basis vectors of finite element approximation schemes satisfy our hypotheses on the ψ_n's. In general, the answer is no. As we have indicated, the most practical modal control schemes are based on "modes" that represent the most significant motions of the system. The role of the finite element method in modal control is to determine the mode shapes and corresponding frequencies; the basis vectors of the finite element scheme usually are not themselves the mode shapes we want and need not satisfy our hypotheses on the ψ_n's.

For the optimal modal control problem of this paper, the control u_n is chosen to minimize

$$J_n(y_n(0), u_n) = \int_0^\infty (\langle D_n y_n(t), y_n(t)\rangle_E + \langle Q u_n(t), u_n(t)\rangle_U)\, dt, \tag{3.10}$$

where

$$\dot{y}_n(t) = A_n y_n(t) + B_n u_n(t), \qquad t \geqq 0, \quad y_n(0) \in E_n. \tag{3.11}$$

The problem of choosing u_n to minimize J_n we call "the optimal control problem on E_n" or "the nth approximate problem;" (3.11) represents the model based on the first n modes. We assume the following:

HYPOTHESIS 3.2. *For each $y_n(0) \in E_n$, there is a $u \in L_2(0, \infty; U)$ such that $J_n(y_n(0), u) < \infty$; or, equivalently, the unstable states of* (3.11) *are controllable. (See Balas* [3] *for "modal" conditions for controllability).*

We know then from finite dimensional control theory (see [1], [14]) that the optimal control control u_n is the feedback control

$$u_n(t) = -Q^{-1} B_n^* P_n y_n(t), \tag{3.12}$$

where P_n is the unique real, positive definite, selfadjoint, $2n \times 2n$ matrix satisfying the Riccati (Kalman) equation

$$A_n^* P_n + P_n A_n - P_n B_n Q^{-1} B_n^* P_n + D_n = 0. \tag{3.13}$$

Remember that the adjoints in (3.12)–(3.14) are taken with respect to the energy inner product on E.

Based on (3.12), the *modal control* $\bar{u}_n$ for the full system (2.1) or (2.13) is given by

$$\bar{u}_n(t) = -Q^{-1}B_n^* P_n \Lambda_n y(t), \tag{3.14}$$

and this feedback control results in a closed-loop system which we will use the semigroup $\bar{\mathcal{T}}_n(\cdot)$ to represent. The generator of $\bar{\mathcal{T}}_n(\cdot)$ is $\mathcal{A} - \mathcal{B}Q^{-1}B_n^* P_n \Lambda_n$. Since the modal control law of (3.14) occupies the center of our attention, let us be as explicit as possible about it. The projection operator Λ_n picks out the first n modes (and their first derivatives with respect to time) from the full state vector $y(t)$, and then $\bar{u}_n(t)$ is taken to be the control that would be optimal if there were no other modes present. Now this modal control generally cannot be implemented exactly because of a limited number of measurements and "observation spillover" (see Balas [3]) of the unmodeled, or residual, modes into the sensor data; the hardware almost never can realize fully the projection Λ_n. However, linear modal regulators usually are designed to approximate (3.14) as closely as possible. A typical procedure (see Balas [2], [3] and Skelton and Likins [20], [21]) is to filter the sensor data to remove as much of the spillover from the truncated modes as possible, and then feed the filtered measurements into a Luenburger observer to estimate the right side of (3.14).

The main questions this paper addresses are (1) what are the convergence properties of P_n and of the control law of (3.14), as n increases? and (2) is the closed-loop system represented by $\bar{\mathcal{T}}_n(\cdot)$, i.e., the full system with the modal control $\bar{u}_n$, stable for n sufficiently large?

4. Approximation theory for the infinite dimensional regulator problem. We now state, in the form most covenient for present purposes, the results of [10] to be relied upon explicitly in this paper, and from these results we derive convergence properties for approximation schemes like the modal scheme of the previous section. Because this paper deals exclusively with optimal control on infinite time intervals, so that the resulting control laws are time-invariant, it must rely on [10] for the pertinent results concerning the infinite dimensional regulator problem and its approximate solution, rather than on earlier works such as [6], [8], [15] and [16]. Although these works did a lot to motivate [10], their results concerning steady-state solutions of infinite dimensional Ricatti equations are not adequate for the analysis of this paper. Especially important for this paper are (a) the fact that existence of a nonnegative, selfadjoint solution of the Riccati algebraic equation of our problem implies that there is a bounded (and, in our problem, compact) linear feedback which gives the control system a uniform decay rate (see Definition 4.2 and Theorem 4.1); and (b) Theorem 4.2, which concerns approximation of the solution of the Riccati algebraic equation. The earlier works contain neither these results nor anything that could be used in their place here.

Remark 4.1. For the results of this section, we need only the definitions stated in this section and Hypotheses 4.1 and 4.2. The control system of § 2 and the approximation scheme of § 3 satisfy these definitions and hypotheses.

Let E and U be real Hilbert spaces, identified with their respective duals, and take U to be finite dimensional. Let $\mathcal{A}$ generate a strongly continuous semigroup of bounded linear operators $\mathcal{T}(\cdot)$ on E, and let $\mathcal{B} \in \mathcal{L}(U, E)$, $\mathcal{D} \in \mathcal{L}(E, E)$, and $Q \in \mathcal{L}(U, U)$ with $\mathcal{D}$ and Q positive definite and selfadjoint. Define the state $y(t) \in E$ by

$$y(t) = \mathcal{T}(t)y(0) + \int_0^t \mathcal{T}(t-\eta)\mathcal{B}u(\eta)\, d\eta, \qquad t \geqq 0, \qquad y(0) \in E, \tag{4.1}$$

where the control $u(\cdot) \in L_2(0, \infty; U)$. The cost functional $J(y(0), u)$ is given as in (3.1) with $y(\cdot)$ given by (4.1). Solutions of (4.1) are called "mild solutions" of (2.13) (see [4]).

DEFINITION 4.1. A function $u \in L_2(0, \infty; U)$ is an *admissible control for the initial state* y, or simply an *admissible control for* y, if $J(y, u)$ is finite; i.e., if the state $y(\cdot)$ corresponding to the control $u(\cdot)$ and the initial condition $y(0) = y$ is in $L_2(0, \infty; E)$.

DEFINITION 4.2. Let the operators $\mathcal{A}$, $\mathcal{B}$, $\mathcal{D}$, and Q be as defined above. An operator $\mathcal{P}$ in $\mathcal{L}(E, E)$ is a *solution of the Riccati algebraic equation* if $\mathcal{P}$ maps the domain of $\mathcal{A}$ into the domain of $\mathcal{A}^*$ and satisfies the Riccati algebraic equation

$$\mathcal{A}^*\mathcal{P} + \mathcal{P}\mathcal{A} - \mathcal{P}\mathcal{B}Q^{-1}\mathcal{B}^*\mathcal{P} + \mathcal{D} = 0. \tag{4.2}$$

The following theorem gives a necessary and sufficient condition for a nonnegative, self-adjoint solution of (4.2) to exist. For such a solution, (4.2) is justified in [10] by showing that it holds on the domain of $\mathcal{A}$, which is dense, so that $\mathcal{A}^*\mathcal{P} + \mathcal{P}\mathcal{A}$ has a bounded extension to all of E.

THEOREM 4.1. (see [10, Thm. 4.11]). *Let the operators $\mathcal{A}$, $\mathcal{B}$, $\mathcal{D}$, and Q be as previously defined. There exists a nonnegative, self-adjoint solution of the Riccati algebraic equation if and only if, for each $y \in E$, there is an admissible control for the initial state y. When such a solution $\mathcal{P}$ exists, it is the unique nonnegative, self-adjoint solution of* (4.2); *the unique control $u(\cdot)$ which minimizes $J(y, \cdot)$ and the corresponding optimal trajectory $y(\cdot)$ are given by $u(t) = -Q^{-1}\mathcal{B}^*\mathcal{P}y(t)$ and $y(t) = \mathcal{S}(t)y$, where $\mathcal{S}(\cdot)$ is the strongly continuous semigroup generated by $\mathcal{A} - \mathcal{B}Q^{-1}\mathcal{B}^*\mathcal{P}$; and $\mathcal{S}(\cdot)$ is uniformly exponentially stable.*[1] *Furthermore,*

$$J(y, u) = \min_{v \text{ admissible}} J(y, v) = \langle \mathcal{P}y, y \rangle_E, \tag{4.3}$$

where u is the optimal control for the initial state y.

For the appropriate version of the most significant approximation theorem of [10], we need the following.

HYPOTHESIS 4.1. *There exists a sequence of C_0 semigroups $\mathcal{T}_n(\cdot)$ on E, with generators $\hat{\mathcal{A}}_n$, and $\|\mathcal{T}_n(t)\|$ is bounded uniformly in n and t for t in bounded intervals. Also, there exist sequences of operators $\mathcal{B}_n$ and $\hat{\mathcal{D}}_n$ in $\mathcal{L}(U, E)$ and $\mathcal{L}(E, E)$, respectively, with $\hat{\mathcal{D}}_n^* = \hat{\mathcal{D}}_n \geqq d > 0$ for $n \geqq 1$ and d independent of n. As $n \to \infty$,*

$$\mathcal{T}_n(t) \to \mathcal{T}(t) \quad \text{strongly}, \qquad t \geqq 0, \tag{4.4}$$

$$\mathcal{T}_n^*(t) \to \mathcal{T}^*(t) \quad \text{strongly}, \quad t \geqq 0, \tag{4.5}$$

$$\mathcal{B}_n \to \mathcal{B} \quad \text{strongly}, \tag{4.6}$$

$$\mathcal{B}_n^* \to \mathcal{B}^* \quad \text{strongly}, \tag{4.7}$$

$$\hat{\mathcal{D}}_n \to \mathcal{D} \quad \text{strongly}, \tag{4.8}$$

and, for each n, the Riccati algebraic equation corresponding to $\hat{\mathcal{A}}_n$, $\mathcal{B}_n$, $\hat{\mathcal{D}}_n$, and Q has a nonnegative, self-adjoint solution $\mathcal{P}_n$.

Note that, since $\dim(U) < \infty$, (4.6) implies (4.7). From [10, Thm. 5.3] we have:

THEOREM 4.2. *Assume Hypothesis* 4.1. *If $\|\mathcal{P}_n\|$ is bounded uniformly in n, then the Riccati algebraic equation* (4.2) *has a nonnegative, selfadjoint solution $\mathcal{P}$,*

$$\mathcal{P}_n \to \mathcal{P} \quad \text{strongly}, \tag{4.9}$$

[1] A semigroup $\mathcal{S}(\cdot)$ is said to be uniformly exponentially stable if there are positive constants M and α such that $\|\mathcal{S}(t)\| \leqq Me^{-\alpha t}$, $t \geqq 0$.

and

$$\mathcal{S}_n(t) \to \mathcal{S}(t) \quad \text{strongly}, \qquad t \geqq 0, \tag{4.10}$$

where $\mathcal{S}_n(\cdot)$ *and* $\mathcal{P}(\cdot)$ *are the semigroups generated by* $\hat{\mathcal{A}}_n - \mathcal{B}_n Q^{-1}\mathcal{B}_n^* \mathcal{P}_n$ *and* $\mathcal{A} - \mathcal{B}Q^{-1}\mathcal{B}^*\mathcal{P}$, *respectively, and there exist positive constants* M *and* β *such that*

$$\|\mathcal{S}_n(t)\| \leqq M e^{-\beta t} \quad \text{and} \quad \|\mathcal{S}(t)\| \leqq M e^{-\beta t}, \qquad t \geqq 0. \tag{4.11}$$

Furthermore, if the convergence in (4.4) *and* (4.5) *is uniform in* t *for* t *in bounded intervals, the same is true for the convergence in* (4.10); *in view of* (4.11), *the convergence in* (4.10) *is then uniform in* t *for* $0 \leqq t < \infty$, *and, for each* $y \in E$, $\mathcal{S}_n(\cdot)y$ *converges in* $L_2(0, \infty; E)$ *to* $\mathcal{S}(\cdot)y$.

Note that the nth approximate problem here, i.e., the optimal control problem corresponding to the operators $\hat{\mathcal{A}}_n$, $\mathcal{B}_n$, $\hat{\mathcal{D}}_n$, and Q, is defined on the whole space E, while in practice each approximate problem is stated on the finite dimensional subspace E_n (see (3.10)–(3.13)). We can bridge this gap by artificially extending the sequence of finite dimensional problems to all of E. For this procedure, we need an additional hypothesis:

HYPOTHESIS 4.2. *There exists an increasing sequence of finite dimensional subspaces* E_n, $n \geqq 1$, *whose union is dense in* E, *each* E_n *is in* $D(\mathcal{A})$, *and the projection operator from* E *onto* E_n *is* Λ_n. *For* $\mathcal{A}_n = \Lambda_n \mathcal{A} \Lambda_n$, $e^{\mathcal{A}_n t}$ *converges strongly to* $\mathcal{T}(t)$ *and* $e^{\mathcal{A}^*_n t}$ *converges strongly* $\mathcal{T}^*(t)$, *and this convergence is uniform in* t *for* t *in bounded subsets of* $[0, \infty)$.

Also, we take $\mathcal{B}_n = \Lambda_n \mathcal{B}$, and $\mathcal{D}_n = \Lambda_n \mathcal{D} \Lambda_n$.

The operators $\mathcal{A}_n$ and $\mathcal{D}_n$ are almost the $\hat{\mathcal{A}}_n$ and $\hat{\mathcal{D}}_n$ needed for Theorem 4.2, but not quite, because we need the nth approximate optimal control problem defined on E to admit a finite cost functional $\hat{J}_n$ with $\hat{\mathcal{D}}_n \geqq d \succ 0$. Define $\hat{\mathcal{A}}_n = \mathcal{A}_n + \Lambda_n - I$ and $\hat{\mathcal{D}}_n = \mathcal{D}_n + I - \Lambda_n$ and note that E_n and $E_n^\perp$ reduce $\hat{\mathcal{A}}_n$, $\hat{\mathcal{D}}_n$, and $e^{\hat{\mathcal{A}}_n t}$. Set $A_n = \hat{\mathcal{A}}_n|_{E_n}$, $D_n = \hat{\mathcal{D}}_n|_E = \mathcal{D}_n|_{E_n}$, and $B_n = \mathcal{B}_n$. For $u_n \in L_2(0, \infty; U)$, define the state vectors $y_n(t)$ and $\hat{y}_n(t)$ in E_n and E, respectively, by

$$y_n(t) = e^{A_n t} y_n(0) + \int_0^t e^{A_n(t-\eta)} B_n u_n(\eta)\, d\eta, \qquad t \geqq 0, \qquad y_n(0) \in E_n, \tag{4.12}$$

and

$$\begin{aligned} \hat{y}_n(t) &= e^{\hat{\mathcal{A}}_n t}\hat{y}_n(0) + \int_0^t e^{\hat{\mathcal{A}}_n(t-\eta)} B_n u_n(\eta)\, d\eta \\ &= y_n(t) + e^{-t}(\hat{y}_n(0) - y_n(0)), \qquad t \geqq 0, \qquad \hat{y}_n(0) \in E, \qquad y_n(0) = \Lambda_n \hat{y}_n(0). \end{aligned} \tag{4.13}$$

Here, $y_n(0)$ is an arbitrary vector in E_n and $\hat{y}_n(0)$ is any vector in E such that $y_n(0) = \Lambda_n \hat{y}_n(0)$. Hence $\hat{y}_n(t) - y_n(t) \in E_n^\perp$, $t \geqq 0$.

The nth optimal control problem defined on E_n is to choose u_n to minimize

$$J_n(y_n(0), u_n) = \int_0^\infty (\langle D_n y_n(\eta), y_n(\eta)\rangle_E + \langle Q u_n(\eta), u_n(\eta)\rangle_U)\, d\eta, \tag{4.14}$$

and the nth optimal control problem defined on E is to choose u_n to minimize

$$\begin{aligned} \hat{J}_n(\hat{y}_n(0), u_n) &= \int_0^\infty (\langle \hat{\mathcal{D}}_n \hat{y}_n(\eta), \hat{y}_n(\eta)\rangle_E + \langle Q u_n(\eta), u_n(\eta)\rangle_U)\, d\eta \\ &= J_n(y_n(0), u_n) + \tfrac{1}{2}\|\hat{y}_n(0) - y_n(0)\|^2. \end{aligned} \tag{4.15}$$

The nth problem on E then is a purely artificial extension of the nth problem on E_n, and the assumption in Hypothesis 4.1 that $\mathscr{P}_n$ exists is equivalent to assuming that the Riccati algebraic equation for the nth problem on E_n has a positive definite, selfadjoint solution $P_n \in \mathscr{L}(E_n, E_n)$. From Theorem 4.1 applied to the optimal control problems on E_n and on E, we see that the optimal control for both problems is $u_n(t) = -Q^{-1}B_nP_ny_n(t)$ and that $\mathscr{P}_n$, the solution of the Riccati algebraic equation for the nth problem on E, is (see (4.3), (4.14), (4.15))

$$\mathscr{P}_n = P_n\Lambda_n + \tfrac{1}{2}(I - \Lambda_n). \tag{4.16}$$

Also,

$$\Lambda_n\mathscr{P}_n = \Lambda_n\mathscr{P}_n\Lambda_n = \Lambda_nP_n\Lambda_n, \tag{4.17}$$

and

$$\|\mathscr{P}_n\| = \max\{\|P_n\|, \tfrac{1}{2}\}. \tag{4.18}$$

The semigroups $\mathscr{T}_n(\cdot)$ and $\mathscr{S}_n(\cdot)$ are given by

$$\mathscr{T}_n(t) = e^{\hat{\mathscr{A}}_nt} \quad \text{and} \quad \mathscr{S}_n(t) = e^{(\hat{\mathscr{A}}_n - \mathscr{B}_nQ^{-1}\mathscr{B}_n^*\mathscr{P}_n)t}, \qquad t \geqq 0. \tag{4.19}$$

Note that

$$\mathscr{B}_nQ^{-1}\mathscr{B}_n^*\mathscr{P}_n = B_nQ^{-1}B_n^*P_n\Lambda_n. \tag{4.20}$$

By Hypothesis 4.2,

$$\Lambda_n \to I \quad \text{strongly}, \tag{4.21}$$

so that (4.6)–(4.8) hold. Suppose that (4.9) holds also. Then (4.16) shows

$$P_n\Lambda_n \to \mathscr{P} \quad \text{strongly}. \tag{4.22}$$

We then have

$$(B_n^*P_n\Lambda_n)^* = \Lambda_nP_nB_n = \Lambda_nP_n\Lambda_n\mathscr{B} \to \mathscr{P}\mathscr{B} \quad \text{strongly}. \tag{4.23}$$

Since U is finite dimensional, $(B_n^*P_n\Lambda_n)^*$ must also converge in $\mathscr{L}(U, E)$, so that

$$B_n^*P_n\Lambda_n \to \mathscr{B}^*\mathscr{P} \quad \text{in } \mathscr{L}(E, U). \tag{4.24}$$

The significance of (4.24) is that it implies

$$\bar{\mathscr{S}}_n(t) \to \mathscr{S}(t) \quad \text{in } \mathscr{L}(E, E), \quad \text{uniformly for } 0 \leqq t < \infty, \tag{4.25}$$

where $\bar{\mathscr{S}}_n(\cdot)$ is the semigroup generated by $\mathscr{A} - \mathscr{B}Q^{-1}B_n^*P_n\Lambda_n$ (see (3.14)). Recall (see Theorem 4.1) that $\mathscr{S}(\cdot)$ represents the optimal system response and $\mathscr{S}(\cdot)$ is uniformly exponentially stable. Thus, when (4.25) holds, the response corresponding to the control $\bar{u}_n$ of (3.14) converges to the optimal response, in the sense of (4.25), and, for n sufficiently large, the closed-loop system represented by $\bar{\mathscr{S}}_n(\cdot)$ is uniformly exponentially stable.

Now the big question: under what conditions is $\|\mathscr{P}_n\|$ uniformly bounded in n? In view of (4.18), this is equivalent to asking the conditions under which $\|P_n\|$ is uniformly bounded in n. Note that, if $\mathscr{P}_n$ converges weakly to some $\mathscr{P} \in \mathscr{P}(E, E)$, then the Principle of Uniform Boundedness implies that $\|\mathscr{P}_n\|$ is uniformly bounded in n, so that, by Theorem 4.2, $\mathscr{P}_n$ converges strongly to the $\mathscr{P}$ of Theorem 4.1. Next we apply the results of this section to the problem formulated in § 3.

5. Systems without damping. In this section, we consider the modal control problem of § 3 for $\mathscr{C}_0 = 0$. To deduce the generally negative result for this case, we need the following theorem from [9].

THEOREM 5.1. *Let $\mathscr{A}$ be the operator defined in § 2 for $\mathscr{C}_0 = 0$, and let $\mathscr{C}$ be a compact operator in $\mathscr{L}(E, E)$. The semigroup generated by $\mathscr{A} + \mathscr{C}$ cannot be uniformly exponentially stable.*

The theorem follows from the fact (see [22, p. 204]) that, since $\mathscr{C}$ is compact, it can be approximated uniformly in norm by the sequence of operators $\mathscr{C}\Lambda_n$, where Λ_n is the projection onto the E_n spanned by pairs of the first n natural modes. Thus, if the semigroup geneated by $\mathscr{A} + \mathscr{C}$ were uniformly exponentially stable, so would be the semigroup generated by $\mathscr{A} + \mathscr{C}\Lambda_n$ for n sufficiently large. But this is impossible; just choose an initial condition in $E_n^{\perp}$, and the feedback control represented by $\mathscr{C}\Lambda_n$ is never activated.

From Theorems 4.1 and 5.1, we know that there can be no nonnegative, selfadjoint solution of the Riccati algebraic equation (4.2) for the optimal control problem on E of § 3 with $\mathscr{C}_0 = 0$; for, if there were such a solution, Theorem 4.1 would imply that the semigroup generated by $\mathscr{A} - \mathscr{B}Q^{-1}\mathscr{B}^*\mathscr{P}$ is uniformly exponentially stable, contradicting Theorem 5.1 because U is finite dimensional and therefore $\mathscr{B}Q^{-1}\mathscr{B}^*\mathscr{P}$ would be compact.

The implication here is quite negative for modal control models that neglect the inherent system damping represented by $\mathscr{C}_0$ in (2.1). In light of Theorem 4.2 and (4.18), we must have

$$\|\mathscr{P}_n\| \to \infty \quad \text{and} \quad \|P_n\| \to \infty \quad \text{as } n \to \infty. \tag{5.1}$$

By Hypothesis 3.2, P_n exists for $n \geqq 1$, so that $\mathscr{P}_n$ is defined by (4.16) for $n \geqq 1$—regardless of whether $\mathscr{P}$ exists.

While the operators $\mathscr{P}_n$ may seem rather abstract, the finite dimensional operators P_n and their matrix representations certainly are not. Specifically, (5.1) says that the maximum eigenvalue of P_n increases without bound as n increases. We know also from (5.1) that $\mathscr{P}_n$ cannot converge to anything in $\mathscr{L}(E, E)$, strongly or weakly.

It may be tempting then to dismiss the results of the modal scheme of § 3 with $\mathscr{C}_0 = 0$ as meaningless with regard to optimal control of the full system, but the following speculations should be considered first. We know from Theorem 4.1 that there exists some $y(0) \in E$ for which there is no admissible control (see Definition 4.1). However, with certain conditions on $\mathscr{B}_0$, Russell has shown in [23] that a bounded linear velocity feedback makes the system strongly stable[2] and that, for $y(0)$ in $D(\mathscr{A})$, this control is an admissible control. Russell's results and those in [10] suggest the following conjecture:

> Suppose that, for the system (2.13) with $\mathscr{C}_0 = 0$, there is an admissible control for each $y(0)$ in a dense subset of E. Let $\tilde{E}$ be the largest such subset. Then the Riccati algebraic equation has a nonnegative selfadjoint solution $\mathscr{P}$, which is a closed operator with domain $\tilde{E}$, and $\mathscr{A} - \mathscr{B}Q^{-1}\mathscr{B}^*\mathscr{P}$, defined on the maximal domain, generates a strongly stable semigroup. Furthermore, when $y(0) \in \tilde{E}$, the feedback control $u(t) = -Q^{-1}\mathscr{B}^*\mathscr{P}y(t)$ minimizes the cost functional of (3.1).

[2] A semigroup $\mathscr{S}(\cdot)$ is said to be strongly stable if, for each $y \in E$, $\|\mathscr{S}(t)y\| \to 0$ as $t \to \infty$.

After the present speculations about this $\mathscr{P}$, we will resume adherence to Definition 4.1 and say that an operator is a solution of the Riccati algebraic equation only if it is a bounded operator.

We might also suspect that $\tilde{E}$ contains each of the modal subspaces E_n, which is true when $D(\mathscr{A}) \subset \tilde{E}$, and that

$$\langle (\mathscr{P}_n - \mathscr{P})y, z\rangle_E \to 0, \qquad y, z \in \bigcup_{n\geq 1} E_n. \tag{5.2}$$

It is quite conceivable that we could have more; for example,

$$\mathscr{P}_n y \to \mathscr{P} y \quad \text{either strongly or weakly,} \qquad y \in \tilde{E}. \tag{5.3}$$

In most applications, including those where the ψ_n's are the natural modes, (5.2) says that the elements of the matrices representing the operators P_n converge individually (not uniformly, of course), which seems to be the least we could require in order to say that $\mathscr{P}_n$ converges in any meaningful sense. Also, if (5.3) holds weakly, we will have

$$B_n^* \mathscr{P}_n y \to \mathscr{B}^* \mathscr{P} y \quad \text{strongly,} \qquad y \in \tilde{E}, \tag{5.4}$$

and it seems possible that $B_n^* \mathscr{P}_n y$ could converge, and maybe for all $y \in E$, without (5.3)—but probably not without (5.2).

None of these speculations appear to be easy to check out, and finding sufficient conditions which are less restrictive than Russell's for the existence of an admissible control for each y in a dense subset of E appears quite difficult. However, from the mathematician's point of view, just making the theory more complete should be worth some effort, and the speculations should be relevant for the engineer who feels compelled by certain very lightly damped systems to design control schemes based on models without damping.

But (5.1) and its implications are inescapable. In practice the Riccati matrix equation must be solved numerically for P_n, and (5.1) suggests that P_n becomes increasingly ill-conditioned as n increases. Indeed, if (5.2) holds, then the greatest lower bound of P_n is bounded in n, so that $\|P_n^{-1}\|$ does not approach zero and

$$\|P_n\| \cdot \|P_n^{-1}\| \to \infty. \tag{5.5}$$

When this is the case, algorithms for computing P_n become less and less practical as n increases.

While there remains some possibility of meaningful convergence for $\mathscr{P}_n$ when damping is not modeled, we know that (5.1) holds and therefore $\mathscr{P}_n$ cannot converge as in Theorem 4.2. The full significance of this negative result will be seen only when it is contrasted with the positive results that are available, as the next section shows, when sufficient damping is modeled.

As for the stability of $\bar{\mathscr{T}}_n(\cdot)$ when $\mathscr{C}_0 = 0$, we cannot say much in general, except that there is no n for which $\bar{\mathscr{T}}_n(\cdot)$ is uniformly exponentially stable. In the case where the ψ_n's are the natural modes ϕ_n, which without damping are uncoupled, it is easy to see that the energy in the first n modes (i.e., the subspace E_n) decays exponentially, while the energy in the truncated modes remains bounded. When all of the initial energy is in the first n natural modes, all of the energy in the truncated modes results from the action of the control $\bar{u}_n(\cdot)$ on these modes. Balas has termed this action "control spillover," and, in [3], has given estimates of the resulting energy in the truncated modes.

6. Systems with damping. Now we require $\mathscr{C}_0 \neq 0$ in (2.1). First consider the case where the semigroup $\mathscr{T}(\cdot)$, which represents the free response of (2.1), is strongly

stable,[3] i.e.,

$$\|\mathcal{T}(t)y\|_E \to 0 \quad \text{as } t \to \infty, \qquad y \in E, \tag{6.1}$$

but not uniformly exponentially stable. We have then the physically realistic situation where, for any initial condition y, all the energy is eventually damped out in the free system ($u = 0$); however, there is no uniform decay rate. Theorem 2 of [9] says that, if $\mathcal{T}(\cdot)$ is a strongly stable contraction semigroup on E and is not uniformly exponentially stable, then no compact linear feedback can yield a uniformly exponentially stable closed-loop system. Thus, the reasoning of the preceding section also applies here, so that (5.1) holds.

To obtain positive convergence results, we must assume

$$\|e^{\mathcal{A}_n t}\| \leqq M e^{-\alpha t}, \qquad t \geqq 0, \qquad n \geqq 1, \tag{6.2}$$

where α and M are positive constants independent of n, and $\mathcal{A}_n = \Lambda_n \mathcal{A} \Lambda_n|_{E_n}$ as previously. Since $e^{\mathcal{A}_n t}$ converges strongly to $\mathcal{T}(t)$ for $t \geqq 0$, (6.2) implies that $\|\mathcal{T}(t)\| \leqq M e^{-\alpha t}$. According to the following theorem and its proof, a sufficient condition for (6.2) is that $\mathcal{C}_0$ be positive definite. Recall $\mathcal{A}_0$ and $\mathcal{C}_0$ from (2.1)–(2.3). As in (2.13), $\mathcal{A}$ is the maximal dissipative extension of the operator $\mathring{\mathcal{A}}$ defined in (2.5), and $\mathcal{A}$ generates the semigroup $\mathcal{T}(\cdot)$.

THEOREM 6.1. *Let $\mathcal{A}_0$, $\mathcal{C}_0$, $\mathcal{A}$, and $\mathcal{T}(\cdot)$ be as in § 2. If there exists a positive constant β such that*

$$\langle \mathcal{C}_0 x, x \rangle_H \geqq \beta^2 \|x\|_H^2, \qquad x \in \mathcal{D}(\mathcal{C}_0), \tag{6.3}$$

then there exist positive constants M and α, which depend only on β, ρ, and γ, such that

$$\|\mathcal{T}(t)\|_E \leqq M e^{-\alpha t}, \qquad t \geqq 0. \tag{6.4}$$

(Recall ρ and γ from (2.2) and (2.3).)

Proof. Referring to (2.1), let $(x(0), \dot{x}(0)) \in \mathcal{D}(\mathcal{A})$ and $\|(x(0), \dot{x}(0))\|_E = 1$. Also, assume for the moment that $\mathcal{C}_0$ is bounded. We have

$$\frac{d}{dt}\|(x(t), \dot{x}(t))\|_E^2 = -2\langle \mathcal{C}_0 \dot{x}(t), \dot{x}(t) \rangle_H. \tag{6.5}$$

Set $u(t) = 0$ in (2.1) and take the H-inner product of each term with $x(t)$. Then integrate the resulting equation over the interval (t_1, t_2), integrating the first term by parts. The result is

$$\begin{aligned} \int_{t_1}^{t_2} \|\dot{x}(t)\|_H^2 \, dt &= \langle \dot{x}(t), x(t) \rangle_H \Big|_{t_1}^{t_2} + \int_{t_1}^{t_2} \langle \mathcal{C}_0 \dot{x}(t), x(t) \rangle_H \, dt \\ &\quad + \int_{t_1}^{t_2} \langle \mathcal{A}_0 x(t), x(t) \rangle_H \, dt, \qquad 0 \leqq t_1 \leqq t_2 < \infty. \end{aligned} \tag{6.6}$$

According to (6.5), $\|(x(t), \dot{x}(t))\|_E^2 = \|x(t)\|_V^2 + \|\dot{x}(t)\|_H^2 \leqq 1$, $t \geqq 0$, so that

$$\langle \mathcal{A}_0 x(t), x(t) \rangle_H = \|x(t)\|_V^2 \leqq 1, \qquad t \geqq 0 \tag{6.7}$$

and

$$\|x(t)\|_H \leqq \frac{1}{\rho}, \qquad t \geqq 0. \tag{6.8}$$

[3] For a necessary and sufficient condition for the case where $\mathcal{C}_0$ is bounded, see Dafermos [7].

Next, we have (see the sentence following (2.7))

$$\left|\int_{t_1}^{t_2}\langle \mathscr{C}_0\dot{x}(t), x(t)\rangle_H\,dt\right|$$

$$\leqq\int_{t_1}^{t_2}\langle \mathscr{C}_0\dot{x}(t), \dot{x}(t)\rangle_H^{1/2}\langle \mathscr{C}_0 x(t), x(t)\rangle_H^{1/2}\,dt \tag{6.9}$$

$$\leqq\left(\int_{t_1}^{t_2}\langle \mathscr{C}_0\dot{x}(t), \dot{x}(t)\rangle_H\,dt\right)^{1/2}\left(\int_{t_1}^{t_2}\gamma^2\,dt\right)^{1/2}, \qquad 0\leqq t_1\leqq t_2<\infty.$$

Choose $c>0$ such that

$$\left(\frac{1}{2}+\frac{2}{\beta^2}\right)c^2+\left(\frac{4}{\beta\rho}+\gamma\right)c=\frac{1}{4}, \tag{6.10}$$

and suppose that

$$\int_0^1\langle \mathscr{C}_0\dot{x}(t), \dot{x}(t)\rangle_H\,dt<c^2. \tag{6.11}$$

Let $S=\{t:0\leqq t\leqq 1, \|\dot{x}(t)\|_H^2\geqq 4(c^2/\beta^2)\}$. Then measure $(S)<\frac{1}{4}$. (6.5) and (6.11) lead to

$$\begin{aligned}\langle \mathscr{A}_0 x(t), x(t)\rangle_H &= \|(x(0), \dot{x}(0))\|_E^2-2\int_0^t\langle \mathscr{C}_0\dot{x}, \dot{x}\rangle_H\,ds-\|\dot{x}(t)\|_H^2\\ &>1-\left(2c^2+4\frac{c^2}{\beta^2}\right), \qquad t\in[0,1]\sim S.\end{aligned} \tag{6.12}$$

Choose $t_1\in(0, \frac{1}{4})\sim S$ and $t_2\in(\frac{3}{4}, 1)\sim S$. Then (6.6), (6.8), (6.9), (6.11) and (6.12) yield

$$\int_{t_1}^{t_2}\|\dot{x}(t)\|_H^2\,dt>\left[1-\left(2c^2+4\frac{c^2}{\beta^2}\right)\right]\left(\frac{1}{4}\right)-\frac{4c}{\beta\rho}-c\gamma, \tag{6.13}$$

which, with (6.3) and (6.11), yields

$$\left(\frac{1}{2}+\frac{2}{\beta^2}\right)c^2+\left(\frac{4}{\beta\rho}+\gamma\right)c>\frac{1}{4}, \tag{6.14}$$

contradicting (6.10). Thus (6.11) cannot hold, so

$$\|(x(1), \dot{x}(1))\|_E^2\leqq 1-2c. \tag{6.15}$$

Since c depends only on β, γ, and ρ, and $D(\mathscr{A})$ is dense in E, the theorem is proved for $\mathscr{C}_0\in\mathscr{L}(H, H)$.

Next, we take advantage of our approximation of $\mathscr{T}(t)$ by $e^{\mathscr{A}_n t}$ to establish (6.2) and obtain the theorem for unbounded $\mathscr{C}_0$. Given the definitions of the bounded operators A_{0_n}, C_{0_n}, and A_n in § 2, we see that (2.2), (2.3) and (6.3) imply

$$\rho^2\leqq A_{0_n}, \qquad n\geqq 1, \tag{6.16}$$

and

$$\beta^2\leqq C_{0_n}\leqq\gamma^2 A_{0_n}, \qquad n\leqq 1. \tag{6.17}$$

Thus (6.4) holds with $\|\mathscr{T}(t)\|$ replaced by $\|e^{A_n t}\|_{E_n}$ for $n\geqq 1$. Hence, (6.2). Since $e^{\mathscr{A}_n t}$, which is an extension of $e^{A_n t}$ (see 3.4), converges strongly to $\mathscr{T}(t)$ for $t\geqq 0$, the theorem is proved.

Now, for the optimal control problem on E_n, we know from Theorem 4.1 (and from finite dimensional control theory; see [1] or [14]) that

$$\min_{u_n \text{ admissible}} J_n(y_n(0), u_n) = \langle P_n y_n(0), y_n(0)\rangle_{E_n}, \tag{6.18}$$

so that, when (6.2) holds, setting $u_n = 0$ shows

$$\|P_n\| \leqq \frac{M^2}{2\alpha}\|D_n\| \leqq \frac{M^2}{2\alpha}\|\mathscr{D}\|, \qquad n \geqq 1. \tag{6.19}$$

Therefore, (6.2) is sufficient for $\|P_n\|$ to be uniformly bounded in n, and hence for the convergence properties of (4.9), (4.10), and (4.22)–(4.25), and for $\bar{\mathscr{T}}_n(\cdot)$ to be uniformly exponentially stable for n sufficiently large.

Since we do not assume that the modes remain uncoupled in the presence of the damping represented by $\mathscr{C}_0$, the stability of $\bar{\mathscr{T}}_n(\cdot)$ for sufficiently large n is a nontrivial result. If the ψ_n's are the natural modes of the undamped system, which of course are uncoupled, and if the damping makes the free system uniformly exponentially stable without coupling the modes, then the closed-loop system resulting from the modal control $\bar{u}_n$ of (3.14) is obviously also uniformly exponentially stable, for any n; however, unless the eigenvectors of $\mathscr{A}_0$ are also eigenvectors of $\mathscr{C}_0$, the damping couples the modes and the stability of $\bar{\mathscr{T}}_n(\cdot)$ is not at all obvious. As the following example shows, it is sometimes possible, even in finite dimensions, for damping to couple the modes of an otherwise stable system in a way that results in instability.

Consider a finite dimensional version of our optimal control problem on E, with $H = \mathbb{R}^2$, $U = \mathbb{R}$,

$$\mathscr{A}_0 = \begin{bmatrix} \omega_1^2 & 0 \\ 0 & \omega_2^2 \end{bmatrix}, \quad \mathscr{C}_0 = \begin{bmatrix} c_1 & c_2 \\ c_2 & c_3 \end{bmatrix}, \quad \mathscr{B}_0 = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \qquad \mathscr{D} = I, \qquad Q = 1, \tag{6.20}$$

where $b_1 \neq 0$, $c_1 c_3 - c_2^2 \geqq 0$. Let $x(t)$ be $(x_1(t), x_2(t))$. We have then a system with two modes, $x_1(t)$ and $x_2(t)$. Suppose the control is chosen based on a model of the first mode only, i.e., we take $n = 1$ in the modal control scheme of § 3. The solution of the Riccati equation (3.13) yields the feedback control

$$\bar{u}_1(t) = -b_1(p_{12}x_1(t) + p_{13}\dot{x}_1(t)), \tag{6.21}$$

where

$$p_{12} = (\omega_1\sqrt{\omega_1^2 + b_1^2} - \omega_1^2)/b_1^2 \quad \text{and} \quad p_{13} = (\sqrt{c_1^2 + b_1^2(2p_{12}+1)} - c_1)/b_1^2.^4 \tag{6.22}$$

The resulting closed-loop system (the full system) has the characteristic equation

$$\begin{aligned} &\lambda^4 + (c_1 + c_3 + b_1^2 p_{13})\lambda^3 + (\omega_1^2 + \omega_2^2 + b_1^2 p_{12} + c_1 c_3 - c_2^2 + c_3 b_1^2 p_{13} - c_2 b_1 b_2 p_{13})\lambda^2 \\ &+ (c_1\omega_2^2 + b_1^2 p_{13}\omega_2^2 + c_3\omega_1^2 + c_3 b_1^2 p_{12} - c_2 b_1 b_2 p_{12})\lambda + (\omega_1^2\omega_2^2 + \omega_2^2 b_1^2 p_{12}) = 0. \end{aligned} \tag{6.23}$$

We know that, if $c_2 = 0$, $x_1(t)$ decays exponentially and $x_2(t)$ at least remains bounded; if $c_2 = 0$ and $c_3 > 0$, $x_2(t)$ also decays exponentially. However, if $c_2 \neq 0$, it is possible to select b_2 to make the coefficients of λ and λ^2 in (6.23) negative, so that the characteristic equation has at least one root with positive real part. For this, it does not

[4] The reader who cares to check (6.22) should remember that the adjoints in (3.13) are with respect to the energy inner product on $E_1 = \mathbb{R}^2$, and that this is the inner product for which the matrix P_1 is selfadjoint.

matter that p_{12} and p_{13} are given by (6.22) as long as one of them is nonzero. The only reason for (6.22) is to make this example—or counterexample—follow our optimal modal regulation scheme precisely. But, forgetting optimal control for a moment, we should realize that this example shows that it is possible to have a system which is uniformly exponentially stable, and becomes unstable when certain symmetric, positive definite (or semidefinite) damping is added. For instance, let c_3 and p_{13} initially be positive numbers and let c_2 initially be zero, so that the initial system is uniformly exponentially stable. Then add a symmetric, nonnegative increment to the $\mathscr{C}_0$ matrix so that the new c_2 is positive. If b_2 is sufficiently large, the coefficient of λ^2 in (6.23) will be negative for the new system.

Now, if the parameters of a two mode system like the example resulted in characteristic roots with positive real parts, the system probably would be contrived, and coupling of natural modes by damping, especially light damping, might not create stability problems in many real control systems. But whether it can appears to be a worthwhile question, the answer to which is not obvious. As for an infinite dimensional system of the type considered here, with the modal control $\bar{u}_n$, we know that the full closed-loop system will be uniformly exponentially stable if n is large enough.

7. Conclusions. The results of the previous sections indicate significant advantages to including inherent system damping in the modeling scheme on which a modal control for a distributed system is to be based. When the damping is sufficient for a single uniform decay rate, like that in (6.2), for all the finite dimensional models, we can say definitely that, as the number of modeled modes increases, the modal control of (3.14) approaches the optimal control for the full system, and the corresponding response of the full system approaches the optimal system response; also, if enough modes are modeled, the resulting closed-loop system is uniformly exponentially stable. Not only were we unable to show that these statements are true when no damping is modeled (or when the modeled damping suffices for strong stability but not uniform exponential stability), but we found that the norms of the solutions to the finite dimensional Riccati equations increase without bound as the model dimension increases (see 5.1), while this sequence of norms is bounded when (6.2) holds. As conjectured in § 5, a possibility remains for a kind of weak convergence for the P_n's (the solutions to the finite dimensional Riccati equations) when no damping is modeled, but the P_n's cannot converge in the strong sense in which they converge when the damping is sufficient for (6.2).

Note that the important thing for guaranteeing (6.2) is that all the modes be damped uniformly. The damping can be arbitrarily small as long as it provides some uniform decay rate. In particular, Theorem 6.1 shows that (6.2) holds if the modeled damping is positive definite. While we should expect a relationship between the amount of damping and the convergence rate of P_n, it is not at all clear how useful estimates of such a relationship could be obtained.

Returning to our discussion of the reasons for defining and proving theorems about infinite dimensional optimal control problems, we now can be more specific than we were in the Introduction. In particular, why an infinite dimensional Riccati equation? Philosophically, it is reassuring to know that the optimal regulator problem defined in terms of the actual infinite dimensional system has, under realistic conditions, a solution given by the solution of a Riccati equation involving operators on the full state space; if this were not the case, we would have to ask whether the results of finite dimensional control theory mean anything with regard to optimal regulation of distributed systems.

But really it is not surprising that we can generalize in some sense the finite dimensional regulator theory to infinite dimensional Hilbert spaces. The real value of the infinite dimensional Riccati algebraic equation (4.2) lies in the role that it played in answering our convergence and stability questions. The finite dimensional control laws were defined via the P_n's satisfying the sequence of finite dimensional Riccati equations, and these operators all had different dimensions. Extending the P_n's to the $\mathcal{P}_n$'s enabled us to deal with them all in terms of a common denominator, and that common denominator was the infinite dimensional Riccati equation. For each n, $\mathcal{P}_n$ was the solution of the Riccati equation corresponding to the nth approximate control problem extended to the full state space, so that the $\mathcal{P}_n$'s were all operators on the same space. Thus, the equation gave us a framework for analyzing the convergence and boundedness of the sequences $\{\mathcal{P}_n\}$ and $\{P_n\}$, and hence the convergence of the corresponding sequences of control laws and system responses.

It was especially important to know that, whenever it exists, a bounded nonnegative, selfadjoint solution of the Riccati algebraic equation for the actual control system (and for $\mathcal{D}$ positive definite) defines a feedback control that results in a uniformly exponentially stable closed-loop system (see Definition 4.2 and Theorem 4.1). This fact was essential both for the positive result that, when the modeled damping is sufficient for (6.2), the closed-loop system represented by $\bar{\mathcal{F}}_n(\cdot)$ is uniformly exponentially stable for n sufficiently large, and for the negative result that $\|P_n\| \to \infty$ when no damping is modeled.

Acknowledgment. The referees made many valuable suggestions concerning the revision of earlier versions of this paper.

REFERENCES

[1] M. ATHANS AND P. L. FALB, *Optimal Control*, McGraw-Hill, New York, 1966.
[2] M. J. BALAS, *Active control of flexible systems*, AIAA Symp. Dynamics and Control of Large Flexible Spacecraft, Blacksburg, VA, June 13–15, 1977.
[3] ———, *Modal control of certain flexible dynamic systems*, this Journal, 16 (1978), pp. 450–462.
[4] R. W. CARROLL, *Abstract Methods in Partial Differential Equations*, Harper and Row, New York, 1969.
[5] R. W. CLOUGH AND J. PENZIEN, *Dynamics of Structures*, McGraw-Hill, New York, 1975.
[6] R. F. CURTAIN AND A. J. PRITCHARD, *The infinite-dimensional Riccati equation for systems defined by evolution operators*, this Journal, 14 (1976). pp. 951–983.
[7] C. M. DAFERMOS, *Wave equations with weak damping*, SIAM J. Appl. Math., 18 (1970), pp. 759–767.
[8] R. DATKO, *A linear control problem in abstract Hilbert space*, J. Differential Equations, 9 (1971), pp.346–359.
[9] J. S. GIBSON, *A note on stabilization of infinite dimensional linear oscillators by compact linear feedback*, this Journal, 18 (1980) pp. 311–316.
[10] ———, *The Riccati integral equations for optimal control problems on Hilbert spaces*, this Journal, 17 (1979), pp. 537–565.
[11] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semigroups*, Colloquium Publications, 31, American Mathematical Society, Providence, RI, 1957.
[12] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.
[13] V. LARSON AND P. W. LIKINS, *Optimal estimation and control of elastic spacecraft*, Control and Dynamic Systems, vol. 13, C. T. Leondes, ed., Academic Press, New York, 1977.
[14] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
[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] L. MEIROVITCH, H. VAN LANDINGHAM AND H. OZ, *Control of spinning flexible spacecraft by modal synthesis*, Int. Aero. Fed. 27th Congress, Anaheim, CA, October 1976.
[18] R. S. PHILLIPS, *Dissipative operators and hyperbolic systems of partial differential equations*, Trans. Am. Math. Soc., 90 (1959), pp. 193–254.
[19] E. P. POPOV, *Introduction to Mechanics of Solids*, Prentice-Hall, Englewood Cliffs, NJ, 1968.
[20] R. E. SKELTON AND P. W. LIKINS, *Orthogonal filters for model error compensation in the control of nonrigid spacecraft*, AIAA J. Guidance and Control, 1 (1978).
[21] ———, *Techniques of modelling and model error compensation in linear regulator problems*, Control and Dynamic Systems, vol. 14, C. T. Leondes, ed., Academic Press, New York, 1978.
[22] F. RIESZ AND B. SZ.-NAGY, *Functional Analysis*, Ungar, New York, 1955.
[23] D. L. RUSSELL, *Decay rates for weakly damped systems in Hilbert space obtained with control-theoretic methods*, J. Differential Equations, 19 (1975), pp. 344–370.
[24] H. TANABE, *Equations of Evolution*, Pitman, London, 1979.
[25] A. E. TAYLOR, *Introduction to Functional Analysis*, Wiley, London, 1958.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 5, September 1981

0363-0129/81/1905-0001 $01.00/0

ADDENDUM: CONTROLLABILITY AND STABILIZABILITY IN MULTI-PAIR SYSTEMS*

DAVID P. STANFORD† AND LUTHER T. CONNER, JR.†

At the conclusion of [1] we listed three basic questions concerning convergence of sets of matrices. These questions are answered in this addendum, and we note that the answer to question (2) leads to the following results.

A. *The set* $\{H_1, H_2, \cdots, H_N\}$ *of* $n \times n$ *matrices is convergent if and only if some finite product of the* H_i*'s has spectral radius less than* 1.

B. *The set* $\{H_1, H_2, \cdots, H_N\}$ *of* $n \times n$ *matrices is exponentially convergent if and only if it is convergent.*

We begin by answering [1, question (2)] with the following theorem.

THEOREM 1. *If* $\{H_1, H_2, \cdots, H_N\}$ *is a convergent set of* $n \times n$ *matrices, then there is a single sequence* $\{p_i\}_{i=1}^{\infty}$ *from* $\bar{N}$ *such that*

$$\lim_{k\to\infty} \left(\prod_{i=k}^{1} H_{p_i} \right) = 0.$$

Proof. For each $q = \{q_i\}_{i=1}^{\infty}$ from $\bar{N}$, let

$$V_q = \left\{ x \in R^n \,\middle|\, \lim_{k\to\infty} \left(\prod_{i=k}^{1} H_{q_i} \right) x = 0 \right\}.$$

Then each V_q is a subspace of R^n, and, by the convergence of $\{H_1, H_2, \cdots H_N\}$,

$$R^n = \bigcup_q V_q.$$

If for all q we have $\dim(V_q) < n$, the Baire category theorem is contradicted, and so there is a sequence $p = \{p_i\}_{i=1}^{\infty}$ from $\bar{N}$ with $V_p = R^n$. Hence

$$\lim_{k\to\infty} \left(\prod_{i=k}^{1} H_{p_i} \right) = 0. \qquad \square$$

We observe that statement A above is easily justified using Theorem 1. Also, since contractiveness relative to any norm implies convergence (see [2]), we obtain the following result, which answers [1, question (3)].

COROLLARY. *If* $\{H_1, H_2, \cdots, H_N\}$ *is contractive relative to some norm on* R^n*, then there is a finite product of the* H_i*'s with spectral radius less than* 1.

It is well known that a single matrix H is convergent (i.e., $\lim_{k\to\infty} H^k x = 0$ for all x) if and only if $\rho(H) < 1$, and that this condition in turn implies that the convergence is exponential. The following theorem (statement B above) generalizes this result to sets of matrices. (Notice that the '$<$' in the definition of exponential convergence in [1] should be '$\leqq$'.)

THEOREM 2. *A set* $\{H_1, H_2, \cdots, H_N\}$ *of* $n \times n$ *matrices is exponentially convergent if and only if it is convergent. Moreover, in case of convergence, there is a single periodic sequence from* $\bar{N}$ *which produces exponential convergence for all* x.

---

* This Journal, 18 (1980), pp. 488–497.

† College of William and Mary, Williamsburg, Virginia 23185. This research was supported by NASA-Langley Research Center under grant NAS1-16042.

Proof. Clearly exponential convergence implies convergence.

Now suppose $\{H_1, H_2, \cdots, H_N\}$ is convergent. By statement A, there are $k \in Z^+$ and $\gamma \in \Gamma_k$ such that

$$\rho\left(\prod_{i=k}^{1} H_{\gamma(i)}\right) < 1.$$

We now show that $\bar{\gamma} = (\gamma(1), \gamma(2), \cdots, \gamma(k), \gamma(1), \gamma(2), \cdots, \gamma(k), \cdots)$ produces exponential convergence for all x. Let $\|\cdot\|$ be a matrix norm such that

$$\alpha = \left\|\prod_{i=k}^{1} H_{\gamma(i)}\right\| < 1,$$

and let

$$M = \max_{i=1}^{N} \|H_i\|.$$

Let β satisfy $\alpha < \beta^k \leqq \beta < 1$. Then there exists $P \in Z^+$ such that

$$\left(\frac{\alpha}{\beta^k}\right)^p < \beta^k \quad \text{for } p \geqq P.$$

Thus $\alpha^p < \beta^{pk+k} \leqq \beta^{pk+s}$ for $p \geqq P$ and $s \in \bar{k}$. Hence, for $p \geqq P$, $s \in \bar{k}$ and any nonzero $x \in R^n$,

$$\begin{aligned}\left\|\left(\prod_{i=pk+s}^{1} H_{\bar{\gamma}(i)}\right)x\right\|_0 &\leqq \left\|\prod_{i=pk+s}^{pk+1} H_{\bar{\gamma}(i)}\right\| \left\|\prod_{i=k}^{1} H_{\gamma(i)}\right\|^p \|x\|_0 \\ &\leqq M^k \alpha^p \|x\|_0 \\ &\leqq M^k \beta^{pk+s} \|x\|_0,\end{aligned}$$

where $\|\cdot\|_0$ denotes a vector norm subordinate to $\|\cdot\|$. For $j \in \{1, 2, \cdots, Pk\}$, let

$$B_j > \frac{\left\|\prod_{i=j}^{1} H_{\bar{\gamma}(i)}\right\|}{\beta^j}.$$

Then

$$B_j > \frac{\left\|\left(\prod_{i=j}^{1} H_{\bar{\gamma}(i)}\right)x\right\|_0}{\beta^j \|x\|_0}$$

for all nonzero $x \in R^n$. Letting $B = \max\{B_1, B_2, \cdots, B_{Pk}, M^k\}$, we obtain, for all nonzero $x \in R^n$,

$$\left\|\left(\prod_{i=j}^{1} H_{\bar{\gamma}(i)}\right)x\right\|_0 < B\beta^j \|x\|_0$$

for all $j \in Z^+$. □

[2, Example 4] presents a set of matrices which is convergent but not contractive relative to any norm. In view of Theorem 2, [1, question (1)] is answered in the negative.

REFERENCES

[1] D. P. STANFORD AND L. T. CONNER, JR., *Controllability and stabilizability in multi-pair systems*, this Journal, 18, (1980), pp. 488–497.

[2] D. P. STANFORD, *Stability for a multi-rate sampled-data system*, this Journal, 17 (1979), pp. 390–399.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 5, September 1981

ERRATUM: ON THE ADJOINT PROCESS FOR OPTIMAL CONTROL OF DIFFUSION PROCESSES*

U. G. HAUSSMANN*

Theorem 5.5 should read:
Assume A_1, A_5–A_7. *Then*

$$-p(t,x) = V_x(t,x)$$

$$= \hat{E}_{tx}\{[c_x(\tau, x(\tau)) + 1_{\{\tau<t\}}(\omega)\beta(\tau, x(\tau))n(x(\tau))]\Phi(\tau, t) + \int_t^\tau l_x(s, x(s), \hat{u}(s, x(s)))\Phi(s, t)\, ds\Big\}, \tag{5.6}$$

where $\beta(t,x) = [V_x(t,x) - c_x(t,x)]n(x)$ *and* $n(x)$ *is a unit normal to* ∂G *at* x.

The proof is correct as it stands, except that at the end it should be observed that for $t < T$ only the tangential component of V_x converges to the tangential component of c_x. The correction should also be made in (5.7) and (5.8). Now (5.6) is analogous to the deterministic case: n represents the gradient of the function describing the target set, i.e., $\partial^* Q$ for $t < T$.

* This journal, 19 (1981), pp. 221–243.

† Department of Mathematics, 2075 Wesbrook Mall, University of British Columbia, Vancouver, B.C., Canada V6T 1W5.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 6, November 1981

0363-0129/81/1906-0001 $01.00/0

CONTROL CANONICAL FORMS AND EIGENVALUE ASSIGNMENT BY FEEDBACK FOR A CLASS OF LINEAR HYPERBOLIC SYSTEMS*

B. M. N. CLARKE† AND D. WILLIAMSON‡

Abstract. Canonical forms are developed for a class of linear hyperbolic systems. They are then applied to solve the problem of eigenvalue assignment by distributed feedback and boundary control. The duality of this problem is demonstrated to one of eigenvalue assignment by boundary feedback of an adjoint system subject to distributed control. For both systems it is shown that by feedback, the set $\{\rho_j\}$, $j \in \mathbb{Z}$, can be assigned as eigenvalues of the closed loop system, subject to an asymptotic condition on the set $\{\rho_j\}$. The feedback control is explicitly characterized.

Analogous results are obtained for the problem of eigenvalue assignment by distributed feedback and distributed control.

1. Introduction.

We study the class of systems which can be described by a linear hyperbolic system in two dependent variables $\mathbf{y} = (y_1, y_2)^T$,

$$\frac{\partial \mathbf{y}}{\partial t} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \frac{\partial \mathbf{y}}{\partial x} + A(x)\mathbf{y}, \tag{1.1}$$

where $x \in [0, l]$, $t \in [0, \infty)$ and $A(x)$ is a continuous 2×2 matrix. It can be shown [5] that a wide class of interesting processes are described by an equation of the form (1.1). We impose initial and boundary conditions:

$$\mathbf{y}(x, 0) = \mathbf{y}_0(x), \tag{1.2}$$

$$(\alpha_0 + \beta_0, \alpha_0 - \beta_0)\mathbf{y}(0, t) = u(t), \tag{1.3a}$$

$$(\alpha_1 + \beta_1, \alpha_1 - \beta_1)\mathbf{y}(l, t) = 0, \tag{1.3b}$$

where $\mathbf{y}_0 \in L_2[0, l]$, $u \in L_2[0, T]$, $T > 0$, α_i and β_i are scalars (complex in general).

Under the stated conditions it is known [1], [4] that the initial-boundary value problem (IBVP) (1.1), (1.2), (1.3), is well posed and has a unique solution $\mathbf{y}(\cdot, t) \in L_2[0, l]$ which satisfies the inequality

$$\|\mathbf{y}(\cdot, t)\|^2_{L_2[0,l]} \leqq C\left(\|\mathbf{y}_0\|^2_{L_2[0,l]} + \int_0^T |u(s)|^2 \, ds\right), \qquad t \in [0, T]. \tag{1.4}$$

In our treatment of (1.1), (1.2), (1.3) we allow $\mathbf{y}_0 \in C_0^\infty[0, l]$ and $u \in C_0^\infty[0, T]$, which involves no loss of generality, since each solution $\mathbf{y}(\cdot, t) \in L_2[0, l]$ is the limit of smooth solutions resulting from such smooth data [9].

When $u(t) \equiv 0$, a necessary condition for there to exist solutions of (1.1), (1.3) of the form

$$\mathbf{y}(x, t) = e^{\lambda t}\hat{\boldsymbol{\phi}}(x), \qquad \hat{\boldsymbol{\phi}} \neq \mathbf{0},$$

* Received by the editors December 12, 1979, and in revised form June 3, 1980.
† School of Mathematics and Physics, Macquarie University, North Ryde, N.S.W. 2113, Australia.
‡ Systems and Control Department, University of New South Wales, Kensington, N.S.W. 2033, Australia.

is that $\hat{\boldsymbol{\phi}}(x)$ satisfy

$$\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \hat{\boldsymbol{\phi}}' + A(x)\hat{\boldsymbol{\phi}} = \lambda \hat{\boldsymbol{\phi}}, \qquad x \in [0, l], \tag{1.5}$$

$$(\alpha_0 + \beta_0, \alpha_0 - \beta_0)\hat{\boldsymbol{\phi}}(0) = 0, \qquad (\alpha_1 + \beta_1, \alpha_1 - \beta_1)\hat{\boldsymbol{\phi}}(l) = 0. \tag{1.6}$$

The values of λ for which (1.5), (1.6) has a nontrivial solution are the eigenvalues of the system (1.1), (1.3). The corresponding functions $\hat{\boldsymbol{\phi}}(x)$ are the eigenfunctions. It is known that the eigenvalues $\{\lambda_i\}$ are countable and the eigenfunctions $\{\hat{\boldsymbol{\phi}}_i(x)\}$ form a basis in $L_2[0, l]$, [5]. We will call the $\{\lambda_i\}$ and $\{\hat{\boldsymbol{\phi}}_i(x)\}$, the *open loop system eigenvalues and eigenfunctions.*

The system adjoint to (1.1), (1.3) is

$$\frac{\partial \mathbf{z}^*}{\partial t} = \frac{\partial \mathbf{z}^*}{\partial x}\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} - \mathbf{z}^* A(x), \tag{1.7}$$

$$\mathbf{z}^*(0, t)\begin{pmatrix} \alpha_0 - \beta_0 \\ \alpha_0 + \beta_0 \end{pmatrix} = 0, \qquad \mathbf{z}^*(l, t)\begin{pmatrix} \alpha_1 - \beta_1 \\ \alpha_1 + \beta_1 \end{pmatrix} = 0, \tag{1.8}$$

where * denotes conjugate transpose. If λ is an eigenvalue of (1.1), (1.3), then $-\bar{\lambda}$ is an eigenvalue of (1.7), (1.8), that is, $\mathbf{z}^*(x, t) = e^{-\lambda t}\hat{\boldsymbol{\psi}}^*(x)$ is a solution. The adjoint open loop system eigenfunctions $\{\hat{\boldsymbol{\psi}}_i(x)\}$ also form a basis for $L_2[0, l]$, the dual basis biorthogonal to the basis of open loop eigenfunctions $\{\hat{\boldsymbol{\phi}}_i(x)\}$. That is,

$$\int_0^l \hat{\boldsymbol{\psi}}_j^*(x)\hat{\boldsymbol{\phi}}_i(x)\, dx = \langle \hat{\boldsymbol{\phi}}_i, \hat{\boldsymbol{\psi}}_j \rangle = \delta_{ij}. \tag{1.9}$$

The eigenfunctions $\{\hat{\boldsymbol{\phi}}_j(x)\}$, $\{\hat{\boldsymbol{\psi}}_j(x)\}$ are said to form Riesz bases for $L_2[0, l]$.

Consider a control $u(t)$ in (1.3a) which is a linear function of the state $\mathbf{y}(x, t)$,

$$u(t) = \int_0^l \mathbf{g}^*(\xi)\mathbf{y}(\xi, t)\, d\xi = \langle \mathbf{y}(\cdot, t), \mathbf{g} \rangle, \tag{1.10}$$

for some $\mathbf{g} \in L_2[0, l]$. Thus, we consider the effect of a *distributed feedback control* applied at the boundary $x = 0$. The system is then described by (1.1), (1.2), (1.3b) and

$$(\alpha_0 + \beta_0, \alpha_0 - \beta_0)\mathbf{y}(0, t) = \int_0^l \mathbf{g}^*(\xi)\mathbf{y}(\xi, t)\, d\xi, \tag{1.11}$$

and we refer to it as the *closed loop system.* If we now again seek solutions of the form $\mathbf{y}(x, t) = e^{\rho t}\boldsymbol{\phi}(x)$ we find it necessary that $\boldsymbol{\phi}(x)$ satisfy,

$$\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \boldsymbol{\phi}' + A(x)\boldsymbol{\phi} = \rho \boldsymbol{\phi}, \tag{1.12}$$

$$(\alpha_0 + \beta_0, \alpha_0 - \beta_0)\boldsymbol{\phi}(0) = \int_0^l \mathbf{g}^*(\xi)\boldsymbol{\phi}(\xi)\, d\xi, \qquad (\alpha_1 - \beta_1, \alpha_1 + \beta_1)\boldsymbol{\phi}(l) = 0. \tag{1.13}$$

There exists a countable set $\{\rho_i\}$ for which (1.12), (1.13) has a nontrivial solution. The $\{\rho_i\}$ and corresponding solutions $\{\boldsymbol{\phi}_i(x)\}$ we refer to as *closed loop system eigenvalues and eigenfunctions.*

The question now arises as to what values $\{\rho_i\}$ can be assigned as closed loop eigenvalues by appropriate choice of $\mathbf{g}$.

The problem of eigenvalue assignment by feedback has been studied by Russell [6] for the case of distributed control, that is, where the system is described by

$$\frac{\partial \mathbf{y}}{\partial t} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \frac{\partial \mathbf{y}}{\partial x} + A(x)\mathbf{y} + \mathbf{h}(x)u(t), \qquad \mathbf{h} \in L_2[0, l], \tag{1.14}$$

$$(\alpha_0 + \beta_0, \alpha_0 - \beta_0)\mathbf{y}(0, t) = 0, \qquad (\alpha_1 + \beta_1, \alpha_1 - \beta_1)\mathbf{y}(l, t) = 0. \tag{1.15}$$

Russell expands $\mathbf{y}(x, t)$ in an eigenfunction series

$$\mathbf{y}(x, t) = \sum_{k \in \mathbb{Z}} w_k(t)\hat{\boldsymbol{\phi}}_k(x), \tag{1.16}$$

where $\{\hat{\boldsymbol{\phi}}_k(x)\}$ are the open loop system eigenfunctions and $\{w_k(t)\}$ are solutions of

$$\frac{dw_k}{dt}(t) = \lambda_k w_k(t) + h_k u(t), \qquad k \in \mathbb{Z}, \tag{1.17}$$

where $\{h_k\}$ are the coefficients of $\mathbf{h}$ relative to the basis $\{\hat{\boldsymbol{\phi}}_j(x)\}$. It is necessary for the analysis in [6] that the controllability condition

$$h_k \equiv \int_0^l \hat{\boldsymbol{\psi}}_k^*(x)\mathbf{h}(x)\, dx \neq 0, \qquad k \in \mathbb{Z}, \tag{1.18}$$

be satisfied. The expansion (1.16) in terms of open loop eigenfunctions is inappropriate for the boundary control system (1.1), (1.3), since only homogeneous boundary conditions (1.15) can be accommodated. We also feel that our treatment is more natural than that of Russell. In particular, our canonical form is merely another way of describing solutions of (1.1), (1.2), (1.3). The methods we develop are amenable to generalization to systems in more than two dependent variables.

An important difference between our methods and those of Russell [5], [6], is that we make no explicit use of the fact that the functions $\{e^{\lambda_i t}\}$ form a Riesz basis in $L_2[0, 2l]$. This property is a crucial part of the treatment in [6]. Essentially, our canonical form operates in the space domain whereas Russell's operates in the time domain and makes crucial use of the minimal interval for controllability.

We also give an explicit characterization of the required feedback function $\mathbf{g} \in L_2[0, l]$.

Consider the closed loop system (1.1), (1.3b), (1.11), and $\mathbf{z}(x, t)$ a sufficiently smooth function. Then integration by parts is justified in,

$$\begin{aligned}
&\int_0^T \int_0^l \mathbf{z}^* \left(\frac{\partial \mathbf{y}}{\partial t} - \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \frac{\partial \mathbf{y}}{\partial x} - A(x)\mathbf{y} \right) dx\, dt \\
&\quad = \int_0^l \mathbf{z}^* \mathbf{y}\, dx \Big|_{t=0}^{t=T} - \int_0^T \mathbf{z}^* \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \mathbf{y}\, dt \Big|_{x=0}^{x=l} \\
&\qquad - \int_0^T \int_0^l \left(\frac{\partial \mathbf{z}^*}{\partial t} - \frac{\partial \mathbf{z}^*}{\partial x} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} + \mathbf{z}^* A(x) \right) \mathbf{y}\, dx\, dt \\
&\quad = \int_0^l \mathbf{z}^* \mathbf{y}\, dx \Big|_{t=0}^{t=T} - \int_0^T \Big(\bar{z}_1(l, t) y_1(l, t) - \bar{z}_2(l, t) y_2(l, t) - \bar{z}_1(0, t) y_1(0, t) \\
&\qquad + \bar{z}_2(0, t) y_2(0, t) \Big)\, dt \\
&\qquad - \int_0^T \int_0^l \left(\frac{\partial \mathbf{z}^*}{\partial t} - \frac{\partial \mathbf{z}^*}{\partial x} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} + \mathbf{z}^* A(x) \right) \mathbf{y}\, dx\, dt
\end{aligned} \tag{1.19}$$

$$= \int_0^l \mathbf{z}^*\mathbf{y}\,dx \Big|_{t=0}^{t=T} - \int_0^T [(\alpha_1-\beta_1)\bar{z}_1(l,t)+(\alpha_1+\beta_1)\bar{z}_2(l,t)]\frac{y_1(l,t)}{\alpha_1-\beta_1}\,dt$$
$$+ \int_0^T [(\alpha_0-\beta_0)\bar{z}_1(0,t)+(\alpha_0+\beta_0)\bar{z}_2(0,t)]\frac{y_1(0,t)}{\alpha_0-\beta_0}\,dt$$
$$- \int_0^T \frac{\bar{z}_2(0,t)}{\alpha_0-\beta_0}\int_0^l \mathbf{g}^*\mathbf{y}\,dx\,dt - \int_0^T\int_0^l \left(\frac{\partial \mathbf{z}^*}{\partial t} - \frac{\partial \mathbf{z}^*}{\partial x}\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix} + \mathbf{z}^*A(x)\right)\mathbf{y}\,dx\,dt,$$

where we have used the boundary conditions for $\mathbf{y}$, (1.11), (1.3b), and assume $\alpha_0-\beta_0$, $\alpha_0+\beta_0$, $\alpha_1+\beta_1$, $\alpha_1-\beta_1$ are nonzero and finite.

From (1.19) we obtain the system adjoint to the closed loop system (1.1), (1.3b), (1.11),

$$\frac{\partial \mathbf{z}^*}{\partial t} - \frac{\partial \mathbf{z}^*}{\partial x}\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix} + \mathbf{z}^*A(x) + \frac{\bar{z}_2(0,t)}{\alpha_0-\beta_0}\mathbf{g}^*(x) = \mathbf{0}, \tag{1.20}$$

$$\mathbf{z}^*(0,t)\begin{pmatrix}\alpha_0-\beta_0\\ \alpha_0+\beta_0\end{pmatrix} = 0, \qquad \mathbf{z}^*(l,t)\begin{pmatrix}\alpha_1-\beta_1\\ \alpha_1+\beta_1\end{pmatrix} = 0. \tag{1.21}$$

Notice that the system (1.20), (1.21) adjoint to our closed loop control system is a system of the type (1.14), (1.15) subject to boundary-value feedback. That is,

$$\frac{\partial \mathbf{z}^*}{\partial t} - \frac{\partial \mathbf{z}^*}{\partial x}\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix} + \mathbf{z}^*A(x) + \bar{u}(t)\mathbf{g}^*(x) = \mathbf{0}, \tag{1.22}$$

where

$$u(t) = \frac{z_2(0,t)}{\alpha_0-\beta_0}. \tag{1.23}$$

If the closed loop system (1.1), (1.3b), (1.11) has a solution $\mathbf{y}(x,t) = e^{\rho t}\boldsymbol{\phi}(x)$, then the adjoint system (1.20), (1.21) has a solution $\mathbf{z}^*(x,t) = e^{-\rho t}\boldsymbol{\psi}^*(x)$. Thus the eigenvalue assignment problem could be just as easily posed for the adjoint closed loop system (1.20), (1.21). This is what, in fact, we do, since it turns out be slightly more convenient to treat the adjoint problem.

The eigenfunctions of the adjoint closed loop system satisfy

$$\boldsymbol{\psi}^{*\prime}\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix} - \boldsymbol{\psi}^*A(x) - \frac{\bar{\psi}_2(0)}{\alpha_0-\beta_0}\mathbf{g}^*(x) + \rho\boldsymbol{\psi}^* = \mathbf{0}, \tag{1.24}$$

$$\boldsymbol{\psi}^*(0)\begin{pmatrix}\alpha_0-\beta_0\\ \alpha_0+\beta_0\end{pmatrix} = 0, \qquad \boldsymbol{\psi}^*(l)\begin{pmatrix}\alpha_1-\beta_1\\ \alpha_1+\beta_1\end{pmatrix} = 0. \tag{1.25}$$

Let $\boldsymbol{\phi}(x)$ be a closed loop eigenfunction corresponding to an eigenvalue ρ_i and $\boldsymbol{\psi}(x)$ be

an adjoint closed loop eigenfunction corresponding to an eigenvalue $\rho_j \neq \rho_i$. Then,

$$\begin{aligned}\int_0^l \boldsymbol{\psi}^* \rho_i \boldsymbol{\phi}\, dx &= \int_0^l \boldsymbol{\psi}^* \left\{ \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \boldsymbol{\phi}' + A(x)\boldsymbol{\phi} \right\} dx \\ &\quad + \boldsymbol{\psi}^* \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \boldsymbol{\phi} \Big|_{x=0}^{x=l} \\ &\quad - \int_0^l \left\{ \boldsymbol{\psi}^{*\prime} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} - \boldsymbol{\psi}^* A(x) \right\} \boldsymbol{\phi}\, dx \\ &= [(\alpha_1 - \beta_1)\bar{\psi}_1(l) + (\alpha_1 + \beta_1)\bar{\psi}_2(l)] \frac{\phi_1(l)}{\alpha_1 - \beta_1} \\ &\quad - [(\alpha_0 - \beta_0)\bar{\psi}_1(0) + (\alpha_0 + \beta_0)\bar{\psi}_2(0)] \frac{\phi_1(0)}{\alpha_0 - \beta_0} \\ &\quad + \frac{\bar{\psi}_2(0)}{\alpha_0 - \beta_0} \int_0^l \mathbf{g}^* \boldsymbol{\phi}\, dx + \int_0^l \left(\rho_j \boldsymbol{\psi}^* - \frac{\bar{\psi}_2(0)}{\alpha_0 - \beta_0} \mathbf{g}^* \right) \boldsymbol{\phi}\, dx,\end{aligned}$$

where we have used the boundary conditions for $\boldsymbol{\phi}$, (1.3b), (1.11) and the differential equation (1.24). Finally we obtain

$$(\rho_i - \rho_j) \int_0^l \boldsymbol{\psi}^* \boldsymbol{\phi}\, dx = 0,$$

since $\rho_i \neq \rho_j$, $\int_0^l \boldsymbol{\psi}^* \boldsymbol{\phi}\, dx = 0$. When $i = j$ we scale the eigenfunctions so that

$$\int_0^l \boldsymbol{\psi}^*{}_i \boldsymbol{\phi}_i\, dx = 1.$$

In what follows we assume that the eigenvalues are of multiplicity one. Thus we have shown that the closed loop eigenfunctions $\{\boldsymbol{\phi}_i(x)\}$ and closed loop adjoint eigenfunctions $\{\boldsymbol{\psi}_j(x)\}$ form a biorthogonal set of functions such that

$$\langle \boldsymbol{\phi}_i, \boldsymbol{\psi}_j \rangle = \delta_{ij}.$$

It can be shown that as for the open loop case, the functions $\{\boldsymbol{\phi}_i\}$, $\{\boldsymbol{\psi}_j\}$ form Riesz bases for $L_2[0, l]$, and $0 < c < |\psi_{2i}(0)| < C < \infty$ for real constants c, C.

2. The control canonical form. For the moment we consider the system

$$\frac{\partial \mathbf{y}}{\partial t} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \frac{\partial \mathbf{y}}{\partial x} + \mathbf{f}(x, t), \tag{2.1}$$

subject to (1.2), (1.3), where, as before, $\mathbf{y}_0$, u, $\mathbf{f}$ are sufficiently smooth. Then we may assume a solution $\mathbf{y}(x, t)$ sufficiently smooth to justify use of the method of Laplace transforms in the t variable [4]. We use capital letters to denote the transformed variables. Then (2.1) implies that

$$s\mathbf{Y}(x, s) - \mathbf{y}_0(x) = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \frac{d\mathbf{Y}}{dx}(x, s) + \mathbf{F}(x, s). \tag{2.2}$$

where $\mathbf{Y}(x, s) = \mathscr{L}\{\mathbf{y}(x,t)\}$ and $\mathbf{F}(x, s) = \mathscr{L}\{\mathbf{f}(x, t)\}$. Transforming the boundary conditions (1.3), we obtain

$$(\alpha_0+\beta_0, \alpha_0-\beta_0)\mathbf{Y}(0, s) = U(s), \tag{2.3a}$$

$$(\alpha_1+\beta_1, \alpha_1-\beta_1)\mathbf{Y}(l, s) = 0. \tag{2.3b}$$

Equation (2.2) has a solution,

$$\mathbf{Y}(x, s) = \begin{bmatrix} e^{sx} & 0 \\ 0 & e^{-sx} \end{bmatrix} \mathbf{Y}(0, s) - \int_0^x \begin{bmatrix} e^{s(x-\xi)} & 0 \\ 0 & -e^{-s(x-\xi)} \end{bmatrix} \hat{\mathbf{F}}(\xi, s)\, d\xi, \tag{2.4}$$

where $\hat{\mathbf{F}}(x, s) = \mathbf{F}(x, s) + \mathbf{y}_0(x)$.

Substituting into the boundary conditions (2.3a, b) we obtain

$$\begin{bmatrix} \alpha_0+\beta_0 & \alpha_0-\beta_0 \\ (\alpha_1+\beta_1)e^{sl} & (\alpha_1-\beta_1)\, e^{-sl} \end{bmatrix} \mathbf{Y}(0, s)$$

$$= \begin{pmatrix} U(s) \\ (\alpha_1+\beta_1, \alpha_1-\beta_1) \int_0^l \begin{bmatrix} e^{s(l-\xi)} & 0 \\ 0 & -e^{-s(l-\xi)} \end{bmatrix} \hat{\mathbf{F}}(\xi, s)\, d\xi \end{pmatrix}.$$

Solving for $\mathbf{Y}(0, s)$, we obtain

$$\mathbf{Y}(0, s) = \frac{\begin{bmatrix} (\alpha_1-\beta_1)\, e^{-sl} & -(\alpha_0-\beta_0) \\ -(\alpha_1+\beta_1)\, e^{sl} & \alpha_0+\beta_0 \end{bmatrix}}{(\alpha_0+\beta_0)(\alpha_1-\beta_1)\, e^{-sl} - (\alpha_0-\beta_0)(\alpha_1+\beta_1)\, e^{sl}}$$

$$\cdot \begin{pmatrix} U(s) \\ \displaystyle\int_0^l [(\alpha_1+\beta_1)\, e^{s(l-\xi)}\hat{F}_1(\xi, s) - (\alpha_1-\beta_1)\, e^{-s(l-\xi)}\hat{F}_2(\xi, s)]\, d\xi \end{pmatrix}.$$

We define

$$\gamma = \left(\frac{\alpha_0+\beta_0}{\alpha_0-\beta_0}\right)\left(\frac{\alpha_1-\beta_1}{\alpha_1+\beta_1}\right), \tag{2.5}$$

which is neither zero or infinite since $\alpha_i+\beta_i$, $\alpha_i-\beta_i$ are not zero and are finite for $i = 0, 1$. Then,

$$(1-\gamma e^{-2ls})\mathbf{Y}(0, s)$$

$$= \begin{pmatrix} \dfrac{-\gamma}{\alpha_0+\beta_0}\, U(s)\, e^{-2ls} + \displaystyle\int_0^l e^{-s\xi}\hat{F}_1(\xi, s)\, d\xi - \left(\frac{\alpha_0-\beta_0}{\alpha_0+\beta_0}\right)\gamma \int_0^l e^{-s(2l-\xi)}\hat{F}_2(\xi, s)\, d\xi \\ \dfrac{U(s)}{\alpha_0-\beta_0} - \left(\dfrac{\alpha_0+\beta_0}{\alpha_0-\beta_0}\right) \displaystyle\int_0^l e^{-s\xi}\hat{F}_1(\xi, s)\, d\xi + \gamma \int_0^l e^{-s(2l-\xi)}\hat{F}_2(\xi, s)\, d\xi \end{pmatrix}.$$

On substituting into (2.4), we obtain

$$
\begin{aligned}
&(1-\gamma e^{-2ls})\mathbf{Y}(x,s)\\
&=\begin{pmatrix}
-\dfrac{\gamma U(s)}{\alpha_0+\beta_0}e^{s(x-2l)}+\displaystyle\int_x^l e^{s(x-\xi)}\hat{F}_1(\xi,s)\,d\xi+\gamma\int_0^x e^{s(x-\xi-2l)}\hat{F}_1(\xi,s)\,d\xi\\
\qquad -\left(\dfrac{\alpha_0-\beta_0}{\alpha_0+\beta_0}\right)\gamma\displaystyle\int_0^l e^{s(x+\xi-2l)}\hat{F}_2(\xi,s)\,d\xi\\
\dfrac{U(s)}{\alpha_0-\beta_0}e^{-sx}-\left(\dfrac{\alpha_0+\beta_0}{\alpha_0-\beta_0}\right)\displaystyle\int_0^l e^{-s(x+\xi)}\hat{F}_1(\xi,s)\,d\xi+\gamma\int_x^l e^{-s(x-\xi+2l)}\hat{F}_2(\xi,s)\,d\xi\\
\qquad +\displaystyle\int_0^x e^{-s(x-\xi)}\hat{F}_2(\xi,s)\,d\xi
\end{pmatrix}.
\end{aligned}
$$

An inverse Laplace transformation gives

$$
\begin{aligned}
&\mathbf{y}(x,t)-\gamma\mathbf{y}(x,t-2l)H(t-2l)\\
&=\begin{pmatrix}
-\dfrac{\gamma}{\alpha_0+\beta_0}u(t+x-2l)H(t+x-2l)\\
\dfrac{1}{\alpha_0-\beta_0}u(t-x)H(t-x)
\end{pmatrix}\\
&\quad+\begin{pmatrix}
\displaystyle\int_x^l f_1(\xi,t+x-\xi)H(t+x-\xi)\,d\xi+\gamma\int_0^x f_1(\xi,t+x-\xi-2l)H(t+x-\xi-2l)\,d\xi\\
\qquad -\left(\dfrac{\alpha_0-\beta_0}{\alpha_0+\beta_0}\right)\gamma\displaystyle\int_0^l f_2(\xi,t+x+\xi-2l)H(t+x+\xi-2l)\,d\xi\\
-\left(\dfrac{\alpha_0+\beta_0}{\alpha_0-\beta_0}\right)\displaystyle\int_0^l f_1(\xi,t-x-\xi)H(t-x-\xi)\\
\qquad +\gamma\displaystyle\int_x^l f_2(\xi,t-x+\xi-2l)H(t-x+\xi-2l)\,d\xi\\
\qquad +\displaystyle\int_0^x f_2(\xi,t-x+\xi)H(t-x+\xi)\,d\xi
\end{pmatrix}\\
&\quad+\begin{pmatrix}
\displaystyle\int_x^l \delta(t+x-\xi)y_{01}(\xi)\,d\xi+\gamma\int_0^x \delta(t+x-\xi-2l)y_{01}(\xi)\,d\xi\\
\qquad -\left(\dfrac{\alpha_0-\beta_0}{\alpha_0+\beta_0}\right)\gamma\displaystyle\int_0^l \delta(t+x+\xi-2l)y_{02}(\xi)\,d\xi\\
-\left(\dfrac{\alpha_0+\beta_0}{\alpha_0-\beta_0}\right)\displaystyle\int_0^l \delta(t-x-\xi)y_{01}(\xi)\,d\xi+\gamma\int_x^l \delta(t-x+\xi-2l)y_{02}(\xi)\,d\xi\\
\qquad +\displaystyle\int_0^x \delta(t-x+\xi)y_{02}(\xi)\,d\xi
\end{pmatrix},
\end{aligned}
\tag{2.6}
$$

where $H(\cdot)$, $\delta(\cdot)$ are the Heaviside and Dirac distributions respectively.

We define the 2×2 matrix distribution $k(x, \xi, t)$ where

$$
\begin{aligned}
k_{11}(x,\xi,t) &= \delta(t+x-\xi)+[\gamma\delta(t+x-\xi-2l)-\delta(t+x-\xi)]H(x-\xi),\\
k_{12}(x,\xi,t) &= -\left(\frac{\alpha_0-\beta_0}{\alpha_0+\beta_0}\right)\gamma\delta(t+x+\xi-2l),\\
k_{21}(x,\xi,t) &= -\left(\frac{\alpha_0+\beta_0}{\alpha_0-\beta_0}\right)\delta(t-x-\xi),\\
k_{22}(x,\xi,t) &= \gamma\delta(t-x+\xi)+[\delta(t-x+\xi)-\gamma\delta(t-x+\xi-2l)]H(x-\xi).
\end{aligned}
\tag{2.7}
$$

$k(x, \xi, t)$ has support $t\in[0, 2l]$ for $(x, \xi)\in[0, l]\times[0, l]$. Then after obvious changes of variable in (2.6),

$$
\begin{aligned}
&\mathbf{y}(x,t)-\gamma\mathbf{y}(x,t-2l)H(t-2l)\\
&=\begin{pmatrix}\dfrac{-\gamma}{\alpha_0+\beta_0}u(t+x-2l)H(t+x-2l)\\[2ex] \dfrac{1}{\alpha_0-\beta_0}u(t-x)H(t-x)\end{pmatrix}\\
&\quad+\begin{pmatrix}\displaystyle\int_{t+x-l}^{t}f_1(t-\tau+x,\tau)H(\tau)\,d\tau-\left(\frac{\alpha_0-\beta_0}{\alpha_0+\beta_0}\right)\gamma\int_{t+x-2l}^{t+x-l}f_2(2l-t+\tau-x,\tau)H(\tau)\,dt\\ \qquad\qquad+\gamma\displaystyle\int_{t-2l}^{t+x-2l}f_1(t-\tau+x-2l,\tau)H(\tau)\,d\tau,\\[2ex] -\left(\dfrac{\alpha_0+\beta_0}{\alpha_0-\beta_0}\right)\displaystyle\int_{t-x-l}^{t-x}f_1(t-\tau-x,\tau)H(\tau)\,dt+\gamma\int_{t-2l}^{t-x-l}f_2(x-t+\tau+2l,\tau)H(\tau)\,d\tau\\ \qquad\qquad+\displaystyle\int_{t-x}^{t}f_2(x-t+\tau,\tau)H(\tau)\,d\tau\end{pmatrix}\\
&\quad+\int_0^l k(x,\xi,t)\mathbf{y}_0(\xi)\,d\xi\\
&=\int_{t-2l}^{t}\mathbf{h}(x,t-\tau)u(\tau)H(\tau)\,d\tau+\int_{t-2l}^{t}\int_0^l k(x,\xi,t-\tau)\mathbf{f}(\xi,\tau)H(\tau)\,d\xi\,d\tau\\
&\quad+\int_0^l k(x,\xi,t)\mathbf{y}_0(\xi)\,d\xi,
\end{aligned}
\tag{2.8}
$$

where

$$
\mathbf{h}(x,\tau)=\begin{pmatrix}\dfrac{-\gamma}{\alpha_0+\beta_0}\delta(\tau+x-2l)\\[2ex] \dfrac{1}{\alpha_0-\beta_0}\delta(\tau-x)\end{pmatrix}.
\tag{2.9}
$$

Note that the support of $\mathbf{h}(x, \tau)$ is $\tau\in[0, 2l]$ for $x\in[0, l]$.

Now we let $\mathbf{f}(x, t) = A(x)\mathbf{y}(x, t)$ and arrive at the canonical form

$$\begin{aligned}\mathbf{y}(x, t) - \gamma\mathbf{y}(x, t-2l)H(t-2l) &= \int_{t-2l}^{t}\int_{0}^{l} k(x, \xi, t-\tau)A(\xi)\mathbf{y}(\xi, \tau)H(\tau)\,d\xi\,d\tau \\ &\quad + \int_{t-2l}^{t} \mathbf{h}(x, t-\tau)u(\tau)H(\tau)\,d\tau \\ &\quad + \int_{0}^{l} k(x, \xi, t)\mathbf{y}_0(\xi)\,d\xi.\end{aligned} \tag{2.10}$$

One thing immediately apparent from (2.10) is the natural way the interval $(0.2l]$ arises in the canonical form. As is well known $2l$ is the minimal time interval for boundary value controllability [2], [5], [6]. The canonical form (2.10) may present a way of proving results about boundary controllability independent of the properties of the functions $\{e^{\lambda_i t}\}$, but we leave such questions aside here.

For $0 \leqq t \leqq 2l$, (2.10) reduces to

$$\begin{aligned}\mathbf{y}(x, t) &= \int_{0}^{t}\int_{0}^{l} k(x, \xi, t-\tau)A(\xi)\mathbf{y}(\xi, \tau)\,d\xi\,d\tau \\ &\quad + \int_{0}^{t} \mathbf{h}(x, t-\tau)u(\tau)\,d\tau + \int_{0}^{l} k(x, \xi, t)\mathbf{y}_0(\xi)\,d\xi,\end{aligned} \tag{2.11}$$

and for $t \geqq 2l$

$$\begin{aligned}\mathbf{y}(x, t) - \gamma\mathbf{y}(x, t-2l) &= \int_{t-2l}^{t}\int_{0}^{l} k(x, \xi, t-\tau)A(\xi)\mathbf{y}(\xi, \tau)\,d\xi\,d\tau \\ &\quad + \int_{t-2l}^{t} \mathbf{h}(x, t-\tau)u(\tau)\,d\tau.\end{aligned} \tag{2.12}$$

In the special case when $A(x) \equiv 0$, (2.11), (2.12) define explicitly the solution of the IBVP (1.1), (1.2), (1.3),

$$\mathbf{y}(x, t) = \begin{cases} \displaystyle\int_{0}^{l} k(x, \xi, t)\mathbf{y}_0(\xi)\,d\xi + \int_{0}^{t} \mathbf{h}(x, t-\tau)u(\tau)\,d\tau, & 0 \leqq t \leqq 2l, \\ \displaystyle\gamma\mathbf{y}(x, t-2l) + \int_{t-2l}^{t} \mathbf{h}(x, t-\tau)u(\tau)\,d\tau, & t \geqq 2l. \end{cases}$$

Certain properties of the distribution $k(x, \xi, t)$ are quite straightforward to establish. It is the unique solution of

$$\frac{\partial k}{\partial t} - \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}\frac{\partial k}{\partial x} = 0,$$

for $(x, t) \in [0, l] \times [0, 2l]$, subject to

$$(\alpha_0 + \beta_0, \alpha_0 - \beta_0)k(0, \xi, t) = (0, 0),$$

$$(\alpha_1 + \beta_1, \alpha_1 - \beta_1)k(l, \xi, t) = (0, 0),$$

and

$$k(x, \xi, 0) = \delta(x - \xi)I.$$

Thus k is a certain fundamental solution of the principal part of the differential operator

$$L \equiv \frac{\partial}{\partial t} - \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}\frac{\partial}{\partial x} - A(x).$$

3. Eigenvalue assignment by boundary control. Since solutions of the form $\mathbf{y}(x, t) = e^{\rho t}\boldsymbol{\phi}(x)$ of our closed loop system are essentially steady state solutions, the initial time interval $[0, 2l]$ is of no consequence. Henceforth we deal with the canonical form for $t \geqq 2l$.

When we substitute the feedback control

$$u(t) = \int_0^l \mathbf{g}^*(\xi)\mathbf{y}(\xi, t)\, d\xi,$$

into (2.12) we arrive at (after a change of variable of integration)

$$\mathbf{y}(x, t) - \gamma\mathbf{y}(x, t-2l) = \int_0^{2l}\int_0^l \{k(x, \xi, \tau)A(\xi) + \mathbf{h}(x, \tau)\mathbf{g}^*(\xi)\}\mathbf{y}(\xi, t-\tau)\, d\xi\, d\tau. \tag{3.1}$$

On substituting $\mathbf{y}(x, t) = e^{\rho t}\boldsymbol{\phi}(x)$ we find that it is necessary that $\boldsymbol{\phi}(x)$ satisfy

$$(e^{2l\rho} - \gamma)\boldsymbol{\phi}(x) = \int_0^l [K(x, \xi, \rho)A(\xi) + \mathbf{H}(x, \rho)\mathbf{g}^*(\xi)]\boldsymbol{\phi}(\xi)\, d\xi, \tag{3.2}$$

where

$$K(x, \xi, \rho) = \int_0^{2l} k(x, \xi, \tau)\, e^{\rho(2l-\tau)}\, d\tau \tag{3.3}$$

$$= \begin{bmatrix} e^{\rho(x-\xi)}[e^{2l\rho} + (\gamma - e^{2l\rho})H(x-\xi)], & -\left(\dfrac{\alpha_0 - \beta_0}{\alpha_0 + \beta_0}\right)\gamma\, e^{\rho(x+\xi)} \\ -\left(\dfrac{\alpha_0 + \beta_0}{\alpha_0 - \beta_0}\right) e^{-\rho(x+\xi)+2l\rho}, & e^{-\rho(x-\xi)}[\gamma + (e^{2l\rho} - \gamma)H(x-\xi)] \end{bmatrix}, \tag{3.4}$$

and

$$\mathbf{H}(x, \rho) = \int_0^{2l} \mathbf{h}(x, \tau)e^{\rho(2l-\tau)}\, d\tau = \begin{pmatrix} \dfrac{-\gamma}{\alpha_0 + \beta_0}\, e^{\rho x} \\ \dfrac{e^{\rho(2l-x)}}{\alpha_0 - \beta_0} \end{pmatrix}. \tag{3.5}$$

It is important that $K(\cdot, \cdot, \rho) \in L_2([0, l]\times[0, l])$ for each ρ, which is clearly the case. In fact K is continuous in the square $(x, \xi) \in [0, l]\times[0, l]$ except on the diagonal line $x = \xi$, where it has a simple jump. Also we note $\mathbf{H}(\cdot, \rho) \in L_2[0, l]$.

Comparing the system (1.12), (1.13) for the closed loop eigenfunctions $\{\boldsymbol{\phi}_i(x)\}$ with that for the closed loop adjoint eigenfunctions $\{\boldsymbol{\psi}_j(x)\}$, (1.24), (1.25) we arrive at the corresponding equation

$$(e^{2l\rho} - \gamma)\boldsymbol{\psi}^*(x) - \int_0^l \left[\boldsymbol{\psi}^*(\xi)A(\xi) + \frac{\bar{\psi}_2(0)}{\alpha_0 - \beta_0}\mathbf{g}^*(\xi)\right]K(\xi, x, \rho)\, d\xi = \mathbf{0}. \tag{3.6}$$

Thus the closed loop system eigenfunctions and adjoint eigenfunctions are solutions of Fredholm integral equations of the second kind. These Fredholm equations are not of the standard type treated in most textbooks since the kernel depends on ρ in a nonstandard way. However, it has been shown [3a, b], [7], [8] that most of the usual results, including an alternative theorem, continue to hold for such equations.

We are now in a position to state our eigenvalue assignment problem definitively: Given a countable set of complex numbers $\{\rho_i\}$, $i \in \mathbb{Z}$, find $\mathbf{g} \in L_2[0, l]$ such that the integral equations (3.2) and (3.6) have nontrivial solutions if and only if $\rho = \rho_i$, $i \in \mathbb{Z}$.

Henceforth all integrations are over the interval $[0, l]$ unless otherwise indicated. We make the following assumption concerning the set $\{\rho_i\}$.

Assumption A.

(a) $\{e^{2l\rho_i} - \gamma\} \in l_2$.

(b) $e^{2l\rho_i} - \gamma = 0$ if and only if $\rho_i = \lambda_j$ for some $j \in \mathbb{Z}$.

(c) $\rho_i \neq \rho_j$ for $i \neq j$.

We will show that part (b) of Assumption A is equivalent to the invariance of controllability of the system (1.1), (1.3) (resp. (1.22), (1.21)) under the action of linear feedback (1.10) (resp. (1.23)). Subject to Assumption A we will show that the eigenvalues $\{\rho_i\}$ can be chosen at will.

At this point we will prove certain results which are of use in the sequel.

LEMMA 3.1. *Relative to the sesquilinear form*

$$[\boldsymbol{\phi}, \boldsymbol{\psi}] \equiv \int \boldsymbol{\psi}^*(x) A(x) \boldsymbol{\phi}(x)\, dx,$$

the adjoint of the open loop eigenfunction equation

$$(e^{2l\rho} - \gamma)\boldsymbol{\phi}(x) - \int K(x, \xi, \rho) A(\xi) \boldsymbol{\phi}(\xi)\, d\xi = \mathbf{0},$$

is

$$(e^{2l\rho} - \gamma)\boldsymbol{\psi}^*(x) - \int \boldsymbol{\psi}^*(\xi) A(\xi) K(\xi, x, \rho)\, d\xi = \mathbf{0}.$$

Proof.

$$\int \boldsymbol{\psi}^*(x) A(x) \{(e^{2l\rho} - \gamma)\boldsymbol{\phi}(x) - \int K(x, \xi, \rho) A(\xi) \boldsymbol{\phi}(\xi)\, d\xi\}\, dx$$

$$= \int \{(e^{2l\rho} - \gamma)\boldsymbol{\psi}^*(x) - \int \boldsymbol{\psi}^*(\xi) A(\xi) K(\xi, x, \rho)\, d\xi\} A(x) \boldsymbol{\phi}(x)\, dx. \qquad \square$$

THEOREM 3.1. *If ρ is not an open loop eigenvalue, then the unique solution of*

$$(e^{2l\rho} - \gamma)\boldsymbol{\psi}^*(x) - \int \boldsymbol{\psi}^*(\xi) A(\xi) K(\xi, x, \rho)\, d\xi = \int \hat{\mathbf{g}}^*(\xi) K(\xi, x, \rho)\, d\xi,$$

is given by

$$\boldsymbol{\psi}^*(x) = \int \hat{\mathbf{g}}^*(\xi) D(\xi, x, \rho)\, d\xi, \tag{3.7}$$

where $D(\xi, x, \rho)$ satisfies

$$(e^{2l\rho} - \gamma) D(\xi, x, \rho) - \int D(\xi, \eta, \rho) A(\eta) K(\eta, x, \rho)\, d\eta = K(\xi, x, \rho). \tag{3.8}$$

Proof. The existence and uniqueness are proved by standard methods, of which there are good treatments in [7], [8]. Direct substitution of (3.7) using (3.8) confirms the form of the solution. $\square$

LEMMA 3.2. *If ρ is not an open loop eigenvalue then $D(\xi, x, \rho)$ satisfies, in addition to* (3.8),

$$(e^{2l\rho} - \gamma) D(\xi, x, \rho) - \int K(\xi, \eta, \rho) A(\eta) D(\eta, x, \rho)\, d\eta = K(\xi, x, \rho). \tag{3.9}$$

Proof. It only remains to show that

$$\int D(\xi, \eta, \rho)A(\eta)K(\eta, x, \rho)\, d\eta = \int K(\xi, \eta, \rho)A(\eta)D(\eta, x, \rho)\, d\eta.$$

We temporarily suppress the dependence of D and K on ρ for reasons of clarity.

Multiply (3.8) on the left by $K(z, \xi)A(\xi)$ and integrate with respect to ξ,

$$(e^{2l\rho} - \gamma)\int K(z, \xi)A(\xi)D(\xi, x)\, d\xi - \int\int K(z, \xi)A(\xi)D(\xi, \eta)A(\eta)K(\eta, x)\, d\eta\, d\xi - \int K(z, \xi)A(\xi)K(\xi,x)\, d\xi = 0. \tag{3.10}$$

Similarly multiply (3.8) on the right by $A(x)K(x, z)$ and integrate with respect to x.

$$(e^{2l\rho} - \gamma)\int D(\xi, x)A(x)K(x, z)\, dx - \int\int D(\xi, \eta)A(\eta)K(\eta, x)A(x)K(x, z)\, d\eta\, dx - \int K(\xi, x)A(x)K(x, z)\, dx = 0.$$

Replacing (ξ, x, z) in the above equation by (z, ξ, x), we get

$$(e^{2l\rho} - \gamma)\int D(z, \xi)A(\xi)K(\xi, x)\, d\xi - \int\int D(z, \eta)A(\eta)K(\eta, \xi)A(\xi)K(\xi, x)\, d\eta\, d\xi - \int K(z, \xi)A(\xi)K(\xi,x)\, d\xi = 0. \tag{3.11}$$

Subtracting (3.10) from (3.11) and defining the matrix

$$M(\xi, x) \equiv \int D(\xi, \eta)A(\eta)K(\eta, x)\, d\eta - \int K(\xi, \eta)A(\eta)D(\eta,x)\, d\eta,$$

we obtain

$$(e^{2l\rho} - \gamma)M(z, x) - \int M(z, \xi)A(\xi)K(\xi,x)\, d\xi = 0. \tag{3.12}$$

Thus for fixed z, the rows of $M(z, x)$ satisfy

$$(e^{2l\rho} - \gamma)\boldsymbol{\psi}^*(x) - \int \boldsymbol{\psi}^*(\xi)A(\xi)K(\xi, x, \rho)\, d\xi = \mathbf{0}$$

and are identically zero for each z, since ρ is not an open loop eigenvalue.

Thus

$$M(\xi, x) \equiv 0, \qquad (\xi, x) \in [0, l]\times[0, l].$$ □

THEOREM 3.2. *If ρ is an open loop eigenvalue, then a necessary condition for the equation*

$$(e^{2l\rho} - \gamma)\boldsymbol{\psi}^*(x) - \int \boldsymbol{\psi}^*(\xi)A(\xi)K(\xi, x, \rho)\, d\xi = \mathbf{f}^*(x), \tag{3.13}$$

to have a solution is that

$$\int \mathbf{f}^*(x)A(x)\hat{\boldsymbol{\Phi}}(x)\, dx = 0, \tag{3.14}$$

for all solutions of

$$(e^{2l\rho}-\gamma)\hat{\boldsymbol{\phi}}(x)-\int K(x,\xi,\rho)A(\xi)\hat{\boldsymbol{\phi}}(\xi)\,d\xi=\mathbf{0}. \tag{3.15}$$

Proof. Let $\boldsymbol{\psi}(x)$ and $\hat{\boldsymbol{\phi}}(x)$ be solutions of (3.13) and (3.15) respectively. Then,

$$\begin{aligned}\int \mathbf{f}^*(x)A(x)\hat{\boldsymbol{\phi}}(x)\,dx &= \int [(e^{2l\rho}-\gamma)\boldsymbol{\psi}^*(x)\\ &\qquad -\int \boldsymbol{\psi}^*(\xi)A(\xi)K(\xi,x,\rho)\,d\xi]A(x)\hat{\boldsymbol{\phi}}(x)\,d\xi\\ &= \int \boldsymbol{\psi}^*(x)A(x)[(e^{2l\rho}-\gamma)\hat{\boldsymbol{\phi}}(x)-\int K(x,\xi,\rho)A(\xi)\hat{\boldsymbol{\phi}}(\xi)\,d\xi]\,dx\\ &= 0.\end{aligned}$$

□

It can be shown that the condition (3.14) is also sufficient for the existence of a solution of (3.13) when ρ is an open loop eigenvalue, [3a, b], [7], [8].

COROLLARY 3.1. *A necessary condition for the closed loop system adjoint equation*

$$(e^{2l\rho}-\gamma)\boldsymbol{\psi}^*(x)-\int \boldsymbol{\psi}^*(\xi)A(\xi)K(\xi,x,\rho)\,d\xi=\frac{\bar{\psi}_2(0)}{\alpha_0-\beta_0}\int \mathbf{g}^*(\xi)K(\xi,x,\rho)\,d\xi,$$

to have a solution when ρ is an open loop eigenvalue is

$$(e^{2l\rho}-\gamma)\int \mathbf{g}^*(x)\hat{\boldsymbol{\phi}}(x)\,dx=0,$$

for all solutions of (3.15).

Proof. From the above theorem a necessary condition is

$$\begin{aligned}0&=\int \mathbf{f}^*(x)A(x)\hat{\boldsymbol{\phi}}(x)\,dx\\ &=\int\int \mathbf{g}^*(\xi)K(\xi,x,\rho)A(x)\hat{\boldsymbol{\phi}}(x)\,d\xi\,dx\\ &=\int \mathbf{g}^*(x)\int K(x,\xi,\rho)A(\xi)\hat{\boldsymbol{\phi}}(\xi)\,d\xi\,dx\\ &=(e^{2l\rho}-\gamma)\int \mathbf{g}^*(x)\hat{\boldsymbol{\phi}}(x)\,dx.\end{aligned}$$

□

The following theorem is central to the analysis.

THEOREM 3.3. *If $\boldsymbol{\psi}(x)$ is a closed loop adjoint eigenfunction corresponding to a closed loop eigenvalue ρ, then it is necessary that*

$$\frac{\bar{\psi}_2(0)}{\alpha_0-\beta_0}[e^{2l\rho}-\gamma-\int \mathbf{g}^*(x)\mathbf{H}(x,\rho)\,dx]=\int \boldsymbol{\psi}^*(x)A(x)\mathbf{H}(x,\rho)\,dx.$$

Proof. There are a number of ways of obtaining this result and we choose the one which seems most direct. We return for the moment to the description of $\boldsymbol{\psi}(x)$ as a solution of the BVP (1.24), (1.25). It is easily seen that solutions of (1.24) also satisfy a

Volterra integral equation

$$\boldsymbol{\psi}^*(x) = \boldsymbol{\psi}^*(0)\begin{bmatrix} e^{-\rho x} & 0 \\ 0 & e^{\rho x} \end{bmatrix} + \int_0^x \left(\boldsymbol{\psi}^*(\xi)A(\xi) + \frac{\bar{\psi}_2(0)}{\alpha_0 - \beta_0}\mathbf{g}^*(\xi)\right) \cdot \begin{bmatrix} e^{-\rho(x-\xi)} & 0 \\ 0 & -e^{\rho(x-\xi)} \end{bmatrix} d\xi.$$

At $x = l$,

$$\begin{aligned} 0 &= \boldsymbol{\psi}^*(l)\begin{pmatrix} \alpha_1 - \beta_1 \\ \alpha_1 + \beta_1 \end{pmatrix} \\ &= \boldsymbol{\psi}^*(0)\begin{bmatrix} e^{-\rho l} & 0 \\ 0 & e^{\rho l} \end{bmatrix}\begin{pmatrix} \alpha_1 - \beta_1 \\ \alpha_1 + \beta_1 \end{pmatrix} \\ &\quad + \int \left(\boldsymbol{\psi}^*(\xi)A(\xi) + \frac{\bar{\psi}_2(0)}{\alpha_0 - \beta_0}\mathbf{g}^*(\xi)\right)\begin{bmatrix} e^{-\rho(l-\xi)} & 0 \\ 0 & -e^{\rho(l-\xi)} \end{bmatrix}\begin{pmatrix} \alpha_1 - \beta_1 \\ \alpha_1 + \beta_1 \end{pmatrix} d\xi \\ &= \bar{\psi}_1(0)\, e^{-\rho l}(\alpha_1 - \beta_1) + \bar{\psi}_2(0)\, e^{\rho l}(\alpha_1 + \beta_1) \\ &\quad + \int \boldsymbol{\psi}^*(\xi)A(\xi)\begin{bmatrix} e^{-\rho(l-\xi)} & 0 \\ 0 & -e^{\rho(l-\xi)} \end{bmatrix} \cdot \begin{pmatrix} \alpha_1 - \beta_1 \\ \alpha_1 + \beta_1 \end{pmatrix} d\xi \\ &\quad + \frac{\bar{\psi}_2(0)}{\alpha_0 - \beta_0}\int [\bar{g}_1(\xi)\, e^{-\rho(l-\xi)}(\alpha_1 - \beta_1) - \bar{g}_2(\xi)\, e^{\rho(l-\xi)}(\alpha_1 + \beta_1)]\, d\xi. \end{aligned} \tag{3.16}$$

From the boundary condition at $x = 0$ we obtain

$$\bar{\psi}_1(0) = -\left(\frac{\alpha_0 + \beta_0}{\alpha_0 - \beta_0}\right)\bar{\psi}_2(0),$$

and on substituting into (3.16),

$$\begin{aligned} 0 &= -\left(\frac{\alpha_0 + \beta_0}{\alpha_0 - \beta_0}\right)\bar{\psi}_2(0)\, e^{-\rho l}(\alpha_1 - \beta_1) + \bar{\psi}_2(0)\, e^{\rho l}(\alpha_1 + \beta_1) \\ &\quad + \int \boldsymbol{\psi}^*(\xi)A(\xi)\begin{pmatrix} e^{-\rho(l-\xi)}(\alpha_1 - \beta_1) \\ -e^{\rho(l-\xi)}(\alpha_1 + \beta_1) \end{pmatrix} d\xi \\ &\quad + \frac{\bar{\psi}_2(0)}{\alpha_0 - \beta_0}\int [\bar{g}_1(\xi)e^{-\rho(l-\xi)}(\alpha_1 - \beta_1) - \bar{g}_2(\xi)\, e^{\rho(l-\xi)}(\alpha_1 + \beta_1)]\, d\xi. \end{aligned}$$

We multiply throughout by $e^{\rho l}/(\alpha_1 + \beta_1)$ and use the definition of γ, (2.5) to obtain

$$\begin{aligned} 0 &= (-\gamma + e^{2l\rho})\bar{\psi}_2(0) - \bar{\psi}_2(0)\int [\bar{g}_1(\xi)\left(\frac{-\gamma e^{\rho\xi}}{\alpha_0 + \beta_0}\right) + \bar{g}_2(\xi)\frac{e^{\rho(2l-\xi)}}{\alpha_0 - \beta_0}]\, d\xi \\ &\quad + \frac{1}{\alpha_1 + \beta_1}\int \boldsymbol{\psi}^*(\xi)A(\xi)\begin{pmatrix} e^{\rho\xi}(\alpha_1 - \beta_1) \\ -e^{\rho(2l-\xi)}(\alpha_1 + \beta_1) \end{pmatrix} d\xi. \end{aligned}$$

Again using (2.5) and the definition of $\mathbf{H}(x, \rho)$, (3.5), we finally obtain

$$0 = (e^{2l\rho} - \gamma)\bar{\psi}_2(0) - \bar{\psi}_2(0)\int \mathbf{g}^*(\xi)\mathbf{H}(\xi, \rho)\, d\xi - (\alpha_0 - \beta_0)\int \boldsymbol{\psi}^*(\xi)A(\xi)\mathbf{H}(\xi, \rho)\, d\xi,$$

from which the theorem follows. □

Under Assumption A, we obtain from Theorems 3.1, 3.3 the condition

$$e^{2l\rho} - \gamma - \int \mathbf{g}^*(x)\mathbf{H}(x,\rho)\,dx = \int\int \mathbf{g}^*(\xi)D(\xi,x,\rho)A(x)\mathbf{H}(x,\rho)\,d\xi\,dx,$$

(3.17)

$$e^{2l\rho} - \gamma = \int \mathbf{g}^*(x)\left\{\mathbf{H}(x,\rho) + \int D(x,\xi,\rho)A(\xi)\mathbf{H}(\xi,\rho)\,d\xi\right\}dx.$$

LEMMA 3.3. *The function* $\boldsymbol{\phi}(x,\rho)$ *defined by*

$$\boldsymbol{\phi}(x,\rho) = \mathbf{H}(x,\rho) + \int D(x,\xi,\rho)A(\xi)\mathbf{H}(\xi,\rho)\,d\xi, \tag{3.18}$$

satisfies the Fredholm integral equation

$$(e^{2l\rho} - \gamma)\boldsymbol{\phi}(x,\rho) - \int K(x,\xi,\rho)A(\xi)\boldsymbol{\phi}(\xi,\rho)\,d\xi = (e^{2l\rho} - \gamma)\mathbf{H}(x,\rho). \tag{3.19}$$

Proof.

$$\int K(x,\xi,\rho)A(\xi)\boldsymbol{\phi}(\xi,\rho)\,d\xi$$

$$= \int K(x,\xi,\rho)A(\xi)\left[\mathbf{H}(\xi,\rho) + \int D(\xi,\eta,\rho)A(\eta)\mathbf{H}(\eta,\rho)\,d\eta\right]d\xi$$

$$= \int \left[K(x,\xi,\rho) + \int K(x,\eta,\rho)A(\eta)D(\eta,\xi,\rho)\,dn\right]A(\xi)\mathbf{H}(\xi,\rho)\,d\xi \qquad \text{(using (3.9))}$$

$$= \int (e^{2l\rho} - \gamma)D(x,\xi,\rho)A(\xi)\mathbf{H}(\xi,\rho)\,d\xi \qquad \text{(using (3.18))}$$

$$= (e^{2l\rho} - \gamma)(\boldsymbol{\phi}(x,\rho) - \mathbf{H}(x,\rho)).$$ □

We remark that Lemma 3.3 shows that the function (3.18) is a closed loop system eigenfunction scaled so that

$$\int \mathbf{g}^*(\xi)\boldsymbol{\phi}(\xi,\rho)\,d\xi = e^{2l\rho} - \gamma, \tag{3.20}$$

as can be seen by comparing (3.19) with (3.2).

When ρ is equal to an open loop eigenvalue, it is easy to show that the corresponding open loop eigenfunction is also unchanged and we set

$$\boldsymbol{\phi}(x,\rho) = \hat{\boldsymbol{\phi}}(x,\rho). \tag{3.21}$$

Now given a countable set $\{\rho_i\}$ satisfying Assumption A, the functions $\{\boldsymbol{\phi}(x,\rho_i)\}$ form a Riesz basis for $L_2[0,l]$ with biorthogonal basis $\{\boldsymbol{\psi}(x,\rho_i)\}$. We expand $\mathbf{g}$ in terms of the dual basis

$$\mathbf{g}(x) = \sum_{j\in\mathbb{Z}} g_j\boldsymbol{\psi}(x,\rho_j),$$

where

$$\begin{aligned}\bar{g}_j &= \int \mathbf{g}^*(x)\boldsymbol{\phi}(x,\rho_j)\,dx \\ &= e^{2l\rho_j} - \gamma,\end{aligned} \tag{3.22}$$

from (3.20). We finally state the following theorem.

THEOREM 3.4. *Let* $\{\rho_i\}$ *be a countable sequence of complex numbers satisfying Assumption* A. *Then there exists a* $\mathbf{g} \in L_2[0, l]$ *such that the eigenvalues of the closed loop system* (3.2) *and of the closed loop adjoint system* (3.6) *are precisely the set* $\{\rho_i\}$. *The unique feedback vector is*

$$\mathbf{g}(x) = \sum_{j \in \mathbb{Z}} (e^{2l\bar{\rho}_i} - \bar{\gamma})\boldsymbol{\psi}(x, \rho_j), \tag{3.23}$$

where $\{\boldsymbol{\psi}(x, \rho_j)\}$ *is the unique biorthogonal basis to* $\{\boldsymbol{\phi}(x, \rho_i)\}$.

Proof. The only point left to check is that $\mathbf{g} \in L_2[0, l]$, which is trivial since

$$\sum_{j \in \mathbb{Z}} |e^{2l\rho_i} - \gamma|^2 < \infty,$$

from Assumption A. □

It is probably an opportune time to discuss the implications of our Assumption A. The first part (a) is necessary in order that $\mathbf{g}$ should lie in $L_2[0, l]$. Part (b) of Assumption A ensures the invariance of controllability of the system (1.1), (1.3) (resp. (1.22), (1.21)) under the action of linear feedback

$$u(t) = \int \mathbf{g}^*(\xi)\mathbf{y}(\xi, t)\, d\xi + v(t), \tag{3.24}$$

or respectively,

$$u(t) = \frac{z_2(0, t)}{\alpha_0 - \beta_0} + v(t).$$

In the special case $A(x) = 0$, $e^{2l\rho} - \gamma = 0$ is a necessary and sufficient condition for ρ to be an open loop eigenvalue.

It has previously been observed [6], that the complex number γ may be changed at will by the inclusion of boundary feedback in the boundary control (1.3a), which thereupon assumes the form

$$(\alpha_0 + \beta_0, \alpha_0 - \beta_0)\mathbf{y}(0, t) = \int \mathbf{g}^*(x)\mathbf{y}(x, t)\, dx + (\alpha_2, \beta_2)\mathbf{y}(0, t),$$

or,

$$(\alpha_0 + \beta_0 - \alpha_2, \alpha_0 - \beta_0 - \beta_2)\mathbf{y}(0, t) = \int \mathbf{g}^*(x)\mathbf{y}(x, t)\, dx. \tag{3.25}$$

Thus γ is replaced by

$$\hat{\gamma} = \left(\frac{\alpha_0 + \beta_0 - \alpha_2}{\alpha_0 - \beta_0 - \beta_2}\right)\left(\frac{\alpha_1 - \beta_1}{\alpha_1 + \beta_1}\right),$$

which may be given any value by appropriate choice of (α_2, β_2).

For any *finite* set of distinct complex numbers $\{\rho_i^*\}$, it is possible to determine γ by boundary feedback, such that the finite set $\{\rho_i^*\}$ satisfies (b), (c) of assumption A. We augment the set $\{\rho_i^*\}$ such that the augmented set $\{\rho_i\}$ satisfies assumption A. Then by the above reasoning, we can assign the set $\{\rho_i\}$ as closed loop eigenvalues by distributed feedback.

We have proved the following corollary.

COROLLARY TO THEOREM 3.4. *Let* $\{\rho_i^*\}$ *be any finite set of complex numbers. Then there exists a vector* (α_2, β_2) *and a function* $\mathbf{g} \in L_2[0, l]$ *such that the closed loop system* (1.1), (1.3*a*), (3.25), *has the set* $\{\rho_i^*\}$ *as eigenvalues.*

4. Eigenvalue assignment by distributed control. In this section we obtain results analogous to those of the previous section for the system (1.14), (1.15) considered by Russell [6]. These results are obtained with very little extra effort, from the results of §2' and § 3.

Consider the closed loop system (1.14), (1.15) resulting from the distributed feedback control,

$$u(t) = \int \mathbf{g}^*(\xi)\mathbf{y}(\xi, t)\, d\xi. \tag{4.1}$$

Then the closed loop system is

$$\frac{\partial \mathbf{y}}{\partial t} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \frac{\partial \mathbf{y}}{\partial x} + A(x)\mathbf{y} + \mathbf{h}(x) \int \mathbf{g}^*(\xi)\mathbf{y}(\xi, t)\, d\xi, \tag{4.2}$$

$$(\alpha_0 + \beta_0, \alpha_0 - \beta_0)\mathbf{y}(0, t) = 0, \qquad (\alpha_1 + \beta_1, \alpha_1 - \beta_1)\mathbf{y}(l, t) = 0. \tag{4.3}$$

The system adjoint to the closed loop system (4.2), (4.3) is

$$\frac{\partial \mathbf{z}^*}{\partial t} - \frac{\partial \mathbf{z}^*}{\partial x} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} + \mathbf{z}^* A(x) + \int \mathbf{z}^*(\xi, t)\mathbf{h}(\xi)\, d\xi \mathbf{g}^*(x) = \mathbf{0}, \tag{4.4}$$

$$\mathbf{z}^*(0, t)\begin{pmatrix} \alpha_0 - \beta_0 \\ \alpha_0 + \beta_0 \end{pmatrix} = 0, \qquad \mathbf{z}^*(l, t)\begin{pmatrix} \alpha_1 - \beta_1 \\ \alpha_1 + \beta_1 \end{pmatrix} = 0, \tag{4.5}$$

which is of the same type as (4.1), (4.3).

We can immediately obtain the canonical form of (4.2), (4.3) from (2.8) by putting $u = 0$ and $\mathbf{f}(\xi, \tau) = A(\xi)\mathbf{y}(\xi, \tau) + \mathbf{h}(\xi) \int \mathbf{g}^*(\eta)\mathbf{y}(\eta, \tau)\, d\eta$ there.
We arrive at

$$\begin{aligned} &\mathbf{y}(x, t) - \gamma \mathbf{y}(x, t - 2l)H(t - 2l) \\ &\quad = \int_0^{2l} \int_0^{l} k(x, \xi, \tau)\{A(\xi)\mathbf{y}(\xi, t - \tau) + \mathbf{h}(\xi) \int_0^{l} \mathbf{g}^*(\eta)\mathbf{y}(\eta, t - \tau)\, d\eta\}\, d\xi\, d\tau. \end{aligned} \tag{4.6}$$

The eigenfunctions of (4.2), (4.3) satisfy

$$\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \boldsymbol{\phi}' + A(x)\boldsymbol{\phi} + \mathbf{h}(x) \int \mathbf{g}^*(\xi)\boldsymbol{\phi}(\xi)\, d\xi - \rho \boldsymbol{\phi} = \mathbf{0}, \tag{4.7}$$

$$(\alpha_0 + \beta_0, \alpha_0 - \beta_0)\boldsymbol{\phi}(0) = 0, \qquad (\alpha_1 + \beta_1, \alpha_1 - \beta_1)\boldsymbol{\phi}(l) = 0, \tag{4.8}$$

and the eigenfunctions of the adjoint closed loop system (4.4), (4.5) satisfy

$$\boldsymbol{\psi}^{*\prime} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} - \boldsymbol{\psi}^* A(x) - \int \boldsymbol{\psi}^*(\xi)\mathbf{h}(\xi)\, d\xi \mathbf{g}^*(x) + \rho \boldsymbol{\psi}^* = \mathbf{0}, \tag{4.9}$$

$$\boldsymbol{\psi}^*(0)\begin{pmatrix} \alpha_0 - \beta_0 \\ \alpha_0 + \beta_0 \end{pmatrix} = 0, \qquad \boldsymbol{\psi}^*(l)\begin{pmatrix} \alpha_1 - \beta_1 \\ \alpha_1 + \beta_1 \end{pmatrix} = 0. \tag{4.10}$$

The canonical form of the eigenfunction equation for (4.2), (4.3) can be obtained either by substituting $\mathbf{y}(x, t) = e^{\rho t}\boldsymbol{\phi}(x)$ in (4.6) for $t \geqq 2l$ or by comparing (4.7), (4.8) with the corresponding system for the boundary control case (1.1), (1.3b), (1.11) and its canonical form (3.2). In either case the equation which results is

$$(e^{2l\rho} - \gamma)\boldsymbol{\phi}(x) - \int [K(x, \xi, \rho)A(\xi) + \mathbf{H}(x, \rho)\mathbf{g}^*(\xi)]\boldsymbol{\phi}(\xi)\, d\xi = \mathbf{0}, \tag{4.11}$$

where now

$$\mathbf{H}(x,\rho)=\int K(x,\eta,\rho)\mathbf{h}(\eta)\,d\eta. \tag{4.12}$$

We note that (4.11) is formally identical to the corresponding equation for the case of boundary control (3.2), except for the different definition of $\mathbf{H}(x,\rho)$.

By our previous comment concerning the similarity of the adjoint system (4.9), (4.10) to (4.7), (4.8) we obtain the canonical form of (4.9), (4.10) as

$$(e^{2l\rho}-\gamma)\boldsymbol{\psi}^*(x)-\int \boldsymbol{\psi}^*(\xi)[A(\xi)K(\xi,x,\rho)+\mathbf{h}(\xi)\mathbf{G}^*(x,\rho)]\,d\xi=\mathbf{0}, \tag{4.13}$$

where

$$\mathbf{G}^*(x,\rho)=\int \mathbf{g}^*(\eta)K(\eta,x,\rho)\,d\eta. \tag{4.14}$$

Again we note the similarity between the boundary control case (3.6) and (4.13). The constant $\bar{\psi}_2(0)/(\alpha_0-\beta_0)$ is replaced by the constant $\int \boldsymbol{\psi}^*(\xi)\mathbf{h}(\xi)\,d\xi$.

We require the following counterpart of Theorem 3.3.

THEOREM 4.1. *If $\boldsymbol{\psi}(x)$ is a closed loop adjoint eigenfunction corresponding to a closed loop eigenvalue ρ, then*

$$\int \boldsymbol{\psi}^*(\eta)\mathbf{h}(\eta)\,d\eta[e^{2l\rho}-\gamma-\int \mathbf{g}^*(x)\mathbf{H}(x,\rho)\,dx]=\int \boldsymbol{\psi}^*(x)A(x)\mathbf{H}(x,\rho)\,dx. \tag{4.15}$$

Proof. Exactly as in Theorem 3.3, replacing $\bar{\psi}_2(0)/(\alpha_0-\beta_0)$ by $\int \boldsymbol{\psi}^*(\eta)\mathbf{h}(\eta)\,d\eta$. □

Applying Theorem 3.1 to (4.13) and substituting in (4.15) we obtain

$$\int \boldsymbol{\psi}^*(\eta)\mathbf{h}(\eta)\,d\eta[e^{2l\rho}-\gamma-\int \mathbf{g}^*(x)\mathbf{H}(x,\rho)\,dx]$$

$$=\int \boldsymbol{\psi}^*(\eta)\mathbf{h}(\eta)\,d\eta\int\int \mathbf{g}^*(x)D(x,\xi,\rho)A(\xi)\mathbf{H}(\xi,\rho)\,d\xi\,dx.$$

Using the condition that $\int \boldsymbol{\psi}^*(\eta)\mathbf{h}(\eta)\,d\eta=0$, if and only if ρ is an open loop eigenvalue,

$$e^{2l\rho}-\gamma=\int \mathbf{g}^*(x)\{\mathbf{H}(x,\rho)+\int D(x,\xi,\rho)A(\xi)\mathbf{H}(\xi,\rho)\,d\xi\}\,dx$$

$$=\int \mathbf{g}^*(x)\boldsymbol{\phi}(x,\rho)\,dx.$$

The analysis now proceeds in an identical manner to that of the boundary control case and we conclude with the following theorem.

THEOREM 4.2. *The results of Theorem* 3.4 *apply unchanged to the case of the distributed control system* (4.2), (4.3).

5. Concluding remarks. The first results on eigenvalue assignment for linear hyperbolic systems were obtained by Russell [5], who showed that a combination of distributed and boundary feedback could remove the perturbation in the eigenvalue positions due to the presence of $A(x)$. The later work [6] considered the class of systems treated here in § 4. Our Assumption A part (b) is absent from Russell's work. In the special case $A(x)=0$ it is redundant.

A decided advantage of the present results are that the required feedback is determined in terms of the original system state whereas in [6], the eigenvalues of a

system isomorphic to the original system are assigned and the computation of the corresponding feedback for the original system is nontrivial.

REFERENCES

[1] L. BERS, F. JOHN AND M. SCHECTER, *Partial Differential Equations*, Wiley (Interscience), New York, 1964.

[2] B. M. N. CLARKE, *Boundary controllability of linear symmetric hyperbolic systems*, J. Inst. Maths. Applic., 20, (1977), pp. 283–298.

[3] D. GRECO, *Su un problema ai limiti per un' equazione differenziale lineare ordinaria de secondo ordine*, Giorn. Mat. Battaglini, 78 (1948–9), pp. 216–237.

[3b] ———, *Gli sviluppi in serie di autosoluzioni in un problema ai limito relativo ad un' equazione differenziale lineare ordinaria del secondo ordine*, Giorn. Mat. Battaglini, 79 (1949–50), pp. 86–120.

[4] R. HERSH, *Mixed problems in several variables*, J. Math. Mech., 12, (1963), pp. 317–334.

[5] D. L. RUSSELL, *Control theory of hyperbolic equations related to certain questions in harmonic analysis and spectral theory*, J. Math. Anal. Appl., 40, (1972), pp. 336–368.

[6] ———. *Canonical forms and spectral determination for a class of hyperbolic distributed parameter control systems*, J. Math. Anal. Appl., 62, (1978), pp. 186–225.

[7] J. D. TAMARKIN, *On Fredholm integral equations whose kernels are analytic in a parameter.* Annals Math., 28(1927), pp. 127–152.

[8] F. TRICOMI, *Integral Equations*, Wiley Interscience, New York, 1957.

[9] C. H. WILCOX, *The domain of dependence inequality and initial-boundary value problems for symmetric hyperbolic systems*, Tech. Summary Rep. 333, Mathematics Research Center, Univ. of Wisconsin, 1962.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 6, November 1981

0363-0129/81/1906-0002 $01.00/0

ON SPECTRUM DISTRIBUTION OF COMPLETELY CONTROLLABLE LINEAR SYSTEMS*

SUN SHUN-HUA†

(*Translated by L. F. Ho*‡)

Abstract. This paper is concerned with the placement of the spectrum of the closed-loop operator $A+BK$ resulting from use of a linear feedback control law $u=Kx$ in the infinite dimensional linear control system $x'=Ax+Bu$. For a class of systems in Hilbert space with certain assumptions on the spectrum of the operator A, a complete characterization of the achievable spectra is obtained. The proofs are carried out in an operator-theoretic context.

1. Statement of the problem and main results. For the n-dimensional autonomous linear system

$$\frac{dx}{dt}=Ax(t)+Bu(t), \qquad x(0)=x_0, \tag{1.1}$$

where $x(\cdot)\in R_n$, $u(\cdot)\in R_r$, A, B are, respectively, $n\times n$, $n\times r$ constant matrices and R_n, R_r being n and r dimensional Euclidean spaces, we have the following familiar results [1]: in order that, given any n complex numbers $\lambda_1,\cdots,\lambda_n$, there should always exist an $r\times n$ complex matrix C such that the spectrum $\sigma(A+BC)=\{\lambda_1,\cdots,\lambda_n\}$, a necessary and sufficient condition is that

$$\text{Rank}\{B, AB,\cdots, A^{n-1}B\}=n,$$

or in other words that the system be completely controllable.

Does a property analogous to the one above hold for infinite dimensional spaces? Is it true that for an autonomous completely controllable linear system we can always choose a suitable feedback so that the closed-loop system has any preassigned spectrum? It seems impossible to answer the above question in general. In this paper, we will answer it in some special cases of practical importance.

We look at the autonomous linear control system in a Hilbert space H:

$$\frac{dx}{dt}=Ax+bu(t), \qquad x(0)=x_0, \tag{1.2}$$

and take the linear feedback

$$u(t)=\langle x(t), g\rangle \tag{1.3}$$

where $b, g\in H$ and $\langle\cdot,\cdot\rangle$ denotes the inner product in H. Then the closed-loop system would be

$$\frac{dx}{dt}=Ax+\langle x, g\rangle b, \qquad x(0)=x_0.$$

* Received by the editors June 11, 1980. This paper appeared originally in Chinese, in Acta Mathematica Sinica, 21 (1978), pp. 193–205.

† Department of Mathematics, Szechwan University, People's Republic of China. Present address, Department of Mathematics, Purdue University, West Lafayette, Indiana 47907.

‡ Department of Mathematics, University of Wisconsin-Madison, Madison, Wisconsin 53706. Partial support by the U.S. Air Force Office of Scientific Research under grant AFOSR 79-0018 is acknowledged.

We say that the linear operator A satisfies condition F if

(F_1) A is an unbounded spectral operator with discrete spectrum, its spectral decomposition being

$$A = \sum_{k=1}^{\infty} \lambda_k E(\lambda_k),$$

where $\lambda_k \neq \lambda_j$ (for all $k \neq j$) and $\dim E(\lambda_k) = 1$ ($k \geqq 1$). Without loss of generality [3], we may assume that the $E(\lambda_k)$, ($k \geqq 1$) are self-adjoint operators in H. The normalized eigenvector of A corresponding to $E(\lambda_k)$ will be denoted by $\phi_k (k \geqq 1)$,

(F_2) $$\inf_{\forall i \neq j} |\lambda_i - \lambda_j| = \delta > 0,$$

(F_3) $$\sup_{1 \leqq k < \infty} \sum_{\substack{j=1 \\ j \neq k}}^{\infty} \frac{1}{|\lambda_j - \lambda_k|^2} = \tau < \infty.$$

Our main results are

THEOREM 1.1. *Suppose that the operator A satisfies condition* F *and $b \in H$. Then for a given sequence of complex numbers $\Lambda \triangleq \{\nu_1, \nu_2, \cdots, \nu_n, \cdots\}$, in order that there should exist $g \in H$ such that the spectrum of the operator $A + \langle \cdot, g \rangle b$ satisfies (see remark* 1):

$$\sigma(A + \langle \cdot, g \rangle b) = \sigma_p(A + \langle \cdot, g \rangle b) = \Lambda,$$

a necessary and sufficient condition is that

(i) $$\langle \phi_k, b \rangle \neq 0 \qquad (k \geqq 1),$$

(ii) $$\sum_{k=1}^{\infty} \left| \frac{\lambda_k - \nu_k}{b_k} \right|^2 < \infty \quad (\textit{see remark } 2).$$

Remarks.

1) In this paper, we always look at the point spectrum $\sigma_p(\cdot)$ of "$\cdot$" as a suitably ordered complex sequence. The number of times that any complex number appears in the sequence is equal to the geometric multiplicity of that complex number in the spectrum of the operator "$\cdot$".

2) The precise meaning of condition (ii) is that there exists at least one rearrangement of the complex sequence Λ (still denoted by $\nu_1, \nu_2, \cdots, \nu_n, \cdots$) such that the inequality in (ii) holds.

THEOREM 1.2. *If the operator A satisfies condition* F *and there exists an index set J such that*

$$\begin{aligned} &\operatorname{Re} \lambda_k > 0, \qquad k \in J, \\ &\operatorname{Re} \lambda_k \leqq 0, \qquad k \notin J, \end{aligned} \quad \forall k \geqq 1,$$

then in order that there should exist g, $b \in H$ such that the system (1.2), *under feedback* (1.3), *is stable, a necessary and sufficient condition is that*

$$\sum_{k \in J} \operatorname{Re} \lambda_k < \infty.$$

2. Perturbation of spectral operators.

LEMMA 2.1. *Suppose the operator A satisfies condition* F. *Set $G = \langle \cdot, g \rangle b$ where $g, b \in H$. Then when $\lambda \notin \sigma(A)$ and $1 - \langle R(\lambda; A)b, g \rangle \neq 0$, we must have $\lambda \in \rho(A + G)$.*

Also

$$\text{(i)}\quad R(\lambda;A+G)-R(\lambda;A)=\frac{1}{1-\langle R(\lambda;A)b,g\rangle}R(\lambda;A)GR(\lambda;A),$$

$$\text{(ii)}\quad R(\lambda;A+G)-R(\lambda;A)-R(\lambda;A)GR(\lambda;A)$$
$$=\frac{\langle R(\lambda;A)b,g\rangle}{1-\langle R(\lambda;A)b,g\rangle}R(\lambda;A)GR(\lambda;A),$$

where $\rho(\cdot)$ *denotes the resolvent set of the operator* $\cdot$, *and* $R(\lambda;\cdot)=(\lambda I-\cdot)^{-1}$.

The proof follows from [4] and [5].

THEOREM 2.1. *Suppose that the operator* A *satisfies condition* F, $g, b\in H$, *then the operator* $A+\langle\cdot,g\rangle b$ *is a spectral operator with discrete spectrum. If* $\sigma_p(A+\langle\cdot,g\rangle b)=\{\nu_1,\nu_2,\cdots,\nu_n,\cdots\}$ *then there must exist a positive integer* n_0 *such that*

$$|\lambda_k-\nu_k|\leqq 6|b_kg_k|,\qquad k\geqq n_0. \tag{2.1}$$

Proof. The proof is divided into several steps.

Step 1. Denote $b_k=\langle\phi_k,b\rangle$, $g_k=\langle\phi_k,g\rangle$. Set

$$\varepsilon_l=\begin{Bmatrix}6\,|b_lg_l|\\ \varepsilon_l'\end{Bmatrix}\quad\text{when}\quad\begin{Bmatrix}|b_lg_l|\neq 0,\\ |b_lg_l|=0,\end{Bmatrix}\qquad l\geqq 1, \tag{2.2}$$

where ε_l' is an arbitrary but fixed positive number less than $\delta/2$. Suppose M is a positive number. Form the set

$$S_{M_\varepsilon}\triangleq\{\lambda\,||\lambda|<M\}\cup\bigcup_{k=1}^{\infty}\{\lambda\,||\lambda-\lambda_k|<\varepsilon_k\}, \tag{2.3}$$

and denote its complementary set by $\overline{S_{M_\varepsilon}}$. We claim that there exists a constant M_0 independent of the choice of ε_l' (as long as $\varepsilon_l'<\delta/2$) such that

$$\overline{S_{M_\varepsilon}}\subset\rho(A+\langle\cdot,g\rangle b),\qquad (M\geqq M_0). \tag{2.4}$$

First, by definition, for any $M>0$ we have $\overline{S_{M_\varepsilon}}\subset\rho(A)$. Second, because $g, b\in H$, we can choose k_0 sufficiently large, such that

$$\frac{2}{\delta}\sum_{k=k_0+1}^{\infty}|b_kg_k|<\frac{1}{6}. \tag{2.5}$$

Once the above k_0 is chosen, we could, by condition (F_3), always choose a positive number M_0 so that

$$\sum_{k=1}^{k_0}\frac{|b_kg_k|}{M-|\lambda_k|}<\frac{1}{6},\qquad (M\geqq M_0). \tag{2.6}$$

So for $\lambda\in\bar{S}_{M_\varepsilon}$, $(M\geqq M_0)$ we have

$$\begin{aligned}|\langle R(\lambda;A)b,g\rangle|&=\left|\sum_{k=1}^{\infty}\frac{\bar{b}_kg_k}{\lambda-\lambda_k}\right|\\ &\leqq\sum_{k=1}^{k_0}\left|\frac{b_kg_k}{\lambda-\lambda_k}\right|+\sum_{k=k_0+1}^{\infty}\left|\frac{b_kg_k}{\lambda-\lambda_k}\right|\\ &\leqq\sum_{k=1}^{k_0}\frac{|b_kg_k|}{M-|\lambda_k|}+\frac{|b_{l(\lambda)}g_{l(\lambda)}|}{\varepsilon_{l(\lambda)}}+\frac{2}{\delta}\sum_{k=k_0+1}^{\infty}|b_kg_k|\\ &<\frac{1}{2},\end{aligned} \tag{2.7}$$

here $l(\lambda)$ is determined by the equation

$$|\lambda-\lambda_{l(\lambda)}|=\inf_{k>k_0}|\lambda-\lambda_k|.$$

Equation (2.4) follows from Lemma 2.1 and inequality (2.7).

Step 2. Denote $C_l=\{\lambda\,||\lambda-\lambda_l|=\varepsilon_l\}$, $(l\geqq 1)$. From (2.4), (2.7) and condition (F_3) we know that when l is sufficiently large

$$C_l\subset\rho(A)\cap\rho(A+\langle\cdot\,,g\rangle b). \tag{2.8}$$

Let $C_{l,\delta}=\{\lambda\,||-\lambda_l|=\delta/2\}$. From the proof of (2.7) it is easy to see that there exists n_0 so that

$$\sup_{\lambda\in C_{l,\delta}}|\langle R(\lambda;A)b,g\rangle|<\tfrac{1}{2},\qquad l\geqq n_0 \tag{2.7'}$$

and by (2.5) we have $\varepsilon_l<\delta/2$, $l>k_0$. Also it is not hard to prove that

$$\frac{\langle R(\lambda;A)b,g\rangle}{1-\langle R(\lambda;A)b,g\rangle}R(\lambda;A)GR(\lambda;A)$$

is analytic in the annulus $\{\varepsilon_l<|\lambda-\lambda_l|<\delta/2\}$, $(l>\max\{n_0,k_0\})$. Therefore

$$\begin{aligned}\alpha_l&\triangleq\frac{1}{2\pi i}\oint_{c_l}\frac{\langle R(\lambda;A)b,g\rangle}{1-\langle R(\lambda;A)b,g\rangle}R(\lambda;A)GR(\lambda;A)\,d\lambda\\&=\frac{1}{2\pi i}\oint_{c_{l,\delta}}\frac{\langle R(\lambda;A)b,g\rangle}{1-\langle R(\lambda;A)b,g\rangle}R(\lambda;A)GR(\lambda;A)\,d\lambda.\end{aligned} \tag{2.9}$$

So by (2.7') we have

$$\begin{aligned}\|\alpha_l\|&\leqq\delta\sup_{\lambda\in C_{l,\delta}}\left\|\frac{\langle R(\lambda;A)b,g\rangle}{1-\langle R(\lambda;A)b,g\rangle}R(\lambda;A)GR(\lambda;A)\right\|\\&\leqq\delta\sup_{\lambda\in C_{l,\delta}}\|R(\lambda;A)GR(\lambda;A)\|\\&\leqq\delta\sup_{\lambda\in C_{l,\delta}}\left\{\left(\sum_{k=1}^{\infty}\left|\frac{b_k}{\lambda-\lambda_k}\right|^2\right)^{1/2}\times\left(\sum_{k=1}^{\infty}\left|\frac{g_k}{\lambda-\lambda_k}\right|^2\right)^{1/2}\right\}\\&\leqq\delta\left\{\left(\frac{4|b_l|^2}{\delta^2}+\sum_{k\neq l}\left|\frac{b_k}{|\lambda_l-\lambda_k|-\delta/2}\right|^2\right)^{1/2}\right.\\&\left.\times\left(\frac{4|g_l|^2}{\delta^2}+\sum_{k\neq l}\left|\frac{g_k}{|\lambda_l-\lambda_k|-\delta/2}\right|^2\right)^{1/2}\right\}\\&\leqq 4\delta\left(\frac{|b_l|^2}{\delta^2}+\sum_{k\neq l}\left|\frac{b_k}{\lambda_l-\lambda_k}\right|^2\right)^{1/2}\times\left(\frac{|g_l|^2}{\delta^2}+\sum_{k\neq l}\left|\frac{g_k}{\lambda_l-\lambda_k}\right|^2\right)^{1/2}\qquad(l>\max\{n_0,k_0\}).\end{aligned} \tag{2.10}$$

Hence

$$\sum_{l>\max(n_0,k_0)}\|\alpha_l\|\leqq 4\delta\sum_{l>\max\{n_0,k_0\}}\left(\frac{|b_l|^2}{\delta}+\sum_{k\neq l}\left|\frac{b_k}{\lambda_l-\lambda_k}\right|^2\right)^{1/2}\times\left(\frac{|g_l|^2}{\delta^2}+\sum_{k\neq l}\left|\frac{g_k}{\lambda_l-\lambda_k}\right|^2\right)^{1/2}$$

$$\leqq 4\delta\left(\sum_{l>\max\{n_0,k_0\}}\left(\frac{|b_l|^2}{\delta^2}+\sum_{k\neq l}\left|\frac{b_k}{\lambda_l-\lambda_k}\right|^2\right)\right)^{1/2}\times\left(\sum_{l>\max\{n_0,k_0\}}\left(\frac{|g_l|^2}{\delta^2}+\sum_{k\neq l}\left|\frac{g_k}{\lambda_l-\lambda_k}\right|^2\right)\right)^{1/2}$$

$$\leqq 4\delta\left(\frac{\|b\|^2}{\delta^2}+\sum_{k=1}^{\infty}\sum_{\substack{l>\max\{n_0,k_0\}\\ k\neq l}}\left|\frac{b_k}{\lambda_l-\lambda_k}\right|^2\right)^{1/2}\times\left(\frac{\|g\|^2}{\delta^2}+\sum_{k=1}^{\infty}\sum_{\substack{l>\max\{n_0,k_0\}\\ k\neq l}}\left|\frac{g_k}{\lambda_l-\lambda_k}\right|^2\right)^{1/2} \tag{2.11}$$

$$\leqq 4\delta\left(\frac{\|b\|^2}{\delta^2}+\tau\|b\|^2\right)^{1/2}\times\left(\frac{\|g\|^2}{\delta^2}+\tau\|g\|^2\right)^{1/2}$$

$$=4\delta\left(\frac{1}{\delta^2}+\tau\right)\|b\|\times\|g\|.$$

Similarly, one has

$$\beta_l\triangleq\frac{1}{2\pi i}\oint_{C_l}R(\lambda;A)GR(\lambda;A)\,d\lambda$$

$$=E(\lambda_l)GR^0(\lambda_l;A)+R^0(\lambda_l;A)GE(\lambda_l) \tag{2.12}$$

$$=\left\langle\sum_{\substack{k=1\\ k\neq l}}^{\infty}\frac{E(\lambda_k)}{\lambda_l-\lambda_k}\cdot,g\right\rangle E(\lambda_l)b+\langle E(\lambda_l)\cdot,g\rangle\sum_{\substack{k=1\\ k\neq l}}^{\infty}\frac{E(\lambda_k)b}{\lambda_l-\lambda_k},$$

Also,

$$\sum_{l>\max\{n_0,k_0\}}\|\beta_l\|<\infty. \tag{2.13}$$

Now we denote by E_l the sum inside C_l of the spectral measures $E'(\lambda)$ of the operator $A+\langle\cdot,g\rangle b$. From the proof of [4, Thm. 1 and Lemma 4], we know that when l is sufficiently large, E_l is a one-dimensional operator. In other words, there exists ν_l such that

$$E_l=E'(\nu_l)\text{ and }|\lambda_l-\nu_l|<\varepsilon_l\text{ when }l\text{ is sufficiently large.} \tag{2.14}$$

Step 3. For an arbitrary $M>0$, it is easy to see that $1-\langle R(\lambda;A)b,g\rangle=0$ has only finitely many singularities $\nu_1,\cdots,\nu_{r_M}$ in $|\lambda|\leqq M$ and the algebraic multiplicities $\rho_1,\cdots,\rho_{r_M}$ are also finite. So from Lemma 2.1 we know that

$$\sigma(A+\langle\cdot,g\rangle b)\cap\{\lambda\,||\lambda|\leqq M\}$$

$$\subset\{\underbrace{\nu_1,\cdots,\nu_1}_{\rho_1},\cdots,\underbrace{\nu_{r_M},\cdots,\nu_{r_M}}_{\rho_{r_M}}\}\cup(\sigma(A)\cap\{\lambda\,||\lambda|\leqq M\}). \tag{2.15}$$

Hence $\sigma(A)+\langle\cdot, g\rangle b)\cap\{\lambda\,|\,|\lambda|\leqq M\}$ has only finitely many points, and it is easy to see that all these points belong to $\sigma_p(A+\langle\cdot, g\rangle b)$.

Now by (2.14), (2.15) and an argument similar to the proof of [4, Thm. 1], one knows that $A+\langle\cdot, g\rangle b$ is a spectral operator with discrete spectrum and all but finitely many spectral points are simple. So there exists a positive integer m_0 such that

$$A+\langle\cdot, g\rangle b=\sum_{k=1}^{\infty}\nu_k E'(\nu_k)+N, \tag{2.16}$$

where $E'(\cdot)$ is the resolution of identity associated with $A+\langle\cdot, g\rangle b$, N commutes with $E'(\nu_k)$, $(k\geqq 1)$; $E'(\nu_k)N=0$, $(k>m_0)$; $\dim E'(\nu_k)H=1$, $(k>m_0)$; and N is a nilpotent operator on $\sum_{k=1}^{m_0}E'(\nu_k)H$.

Step 4. Let $M>\max_{1\leqq j\leqq m_0}|\nu_j|$ be a sufficiently large positive number. Construct the contour C_M as follows: C_M is formed by joining the circular arcs on $|\lambda|=M$ that do not intersect $\cup_{l=1}^{\infty}C_l$ and the circular arcs curving inwards of those C_l that intersect $|\lambda|=M$; i.e., C_M is as shown in Fig. 1. Here $\cup_{l=1}^{\infty}C_l$ intersects $|\lambda|=M$ on $C_{k_1},\cdots,C_{k_n}$. It is easy to see that for M sufficiently large the number of $C_{k_1},\cdots,C_{k_n}$ must be finite and C_M is a piecewise smooth simple closed contour.

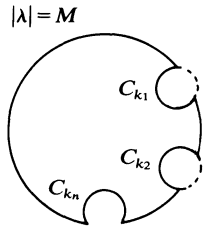


FIG. 1

When M is sufficiently large, noting (2.11) and (2.13) it is easy to prove that

$$\begin{aligned}&\left\|\frac{1}{2\pi i}\oint_{C_M}(R(\lambda;A+G)-R(\lambda;A))\,d\lambda\right\|\\&=\left\|\frac{1}{2\pi i}\oint_{C_M}\frac{1}{1-\langle R(\lambda;A)b, g\rangle}R(\lambda;A)GR(\lambda;A)\,d\lambda\right\|<1.\end{aligned} \tag{2.17}$$

Hence from [8, p. 29] one has

$$\sum_{\nu_j\text{ inside }C_M}\dim E'(\nu_j)H=\sum_{\lambda_j\text{ inside }C_M}\dim E(\lambda_j)H=n_M, \tag{2.18}$$

where n_M is the number of λ_j's inside C_M.

From (2.14) and (2.18) it follows immediately that we can suitably rearrange the spectrum $\nu_1,\nu_2,\cdots$ of $A+\langle\cdot, g\rangle b$ (still denoted by $\nu_1,\nu_2,\cdots$) and change the associated resolution of identity to a family of one-dimensional projections (still denoted by $E'(\nu_k)$), such that

$$A+\langle\cdot, g\rangle b=\sum_{k=1}^{\infty}\nu_k E'(\nu_k)+N,$$

where $\dim E'(\nu_k)=1$ $(k\geqq 1)$, and when l is sufficiently large one has

$$|\lambda_l-\nu_l|<\varepsilon_l,\qquad (l\text{ sufficiently large}). \tag{2.19}$$

Since the ε_l' in (2.2) is arbitrary, it follows from (2.14) (or (2.19)) that (2.1) is true.

In this way the proof of Theorem 2.1 is completed.

COROLLARY 2.1. *Under the conditions of Theorem* 2.1, *if* $b_k \neq 0$, $k \geqq 1$, *then we have*

$$\sum_{k=1}^{\infty} \left| \frac{\lambda_k - \nu_k}{b_k} \right|^2 < \infty. \tag{2.20}$$

Proof. This is a direct consequence of (2.1).

3. Proof of Theorem 1.1.

Necessity. By Corollary 2.1 we need only prove $b_k \neq 0$ $(k \geqq 1)$. We prove by contradiction. Without loss of generality assume $b_1 = 0$. By Theorem 2.1 $A + \langle \cdot, g \rangle b$ is a spectral operator for any given $g \in H$ and it is easy to prove that $(A + \langle \cdot, g \rangle b)^* = A^* + \langle \cdot, b \rangle g$. So $(A + \langle \cdot, g \rangle b)^* \phi_1 = A^* \phi_1 = \bar{\lambda}_1 \phi_1$; i.e., $\bar{\lambda}_1 \in \sigma_p((A + \langle \cdot, g \rangle)^*)$. Hence $\lambda_1 \in \sigma_p(A + \langle \cdot, g \rangle b)$. That is, for any complex sequence $\Lambda = \{\nu_1, \nu_2, \cdots, \nu_n, \cdots\}$, as long as $\nu_k \neq \lambda_1$, $(k \geqq 1)$, there does not exist $g \in H$ such that $\sigma_p(A + \langle \cdot, g \rangle b) = \Lambda$. This contradicts the hypothesis of the theorem. So we must have $b_1 \neq 0$, and hence the proof of the necessity part of Theorem 1.1 is complete.

Sufficiency. We prove sufficiency for the following three cases.

Case 1. $\Lambda = \{\nu_1, \nu_2, \cdots, \nu_n, \cdots\}$ satisfies the conditions of the theorem and (1) $\nu_i \neq \nu_j (\forall_i \neq j)$, (2) $\Lambda \cap \{\lambda_1, \lambda_2, \cdots, \lambda_n, \cdots\} = \varnothing$.

Denote $\alpha_k = (\lambda_k - \nu_k / \bar{b}_k)$, $k \geqq 1$. By assumption

$$\|\alpha\|^2 \triangleq \sum_{k=1}^{\infty} |\alpha_k|^2 < \infty. \tag{3.1}$$

Now we find $g = \sum_{k=1}^{\infty} g_k \phi_k$ in the following way:

$$\left.\begin{aligned} &\langle x_k, g \rangle = -1, \\ &x_k = (A - \nu_k I)^{-1} b \end{aligned}\right\}, \qquad (k \geqq 1). \tag{3.2}$$

That is

$$\sum_{j=1}^{\infty} \frac{\bar{b}_j g_j}{\lambda_j - \nu_k} = -1, \qquad k \geqq 1, \tag{3.3}$$

or equivalently

$$g_k + \sum_{\substack{j=1 \\ j \neq k}}^{\infty} \frac{\alpha_k \bar{b}_j}{\lambda_j - \nu_k} g_j = -\alpha_k, \qquad k \geqq 1. \tag{3.3'}$$

Now we consider the infinite matrix

$$T = (t_{ij}), \qquad t_{ij} = \begin{cases} \dfrac{\alpha_i \bar{b}_j}{\lambda_j - \nu_i}, & i \neq j, \\ 0, & i = j, \end{cases} \qquad i, j \geqq 1. \tag{3.4}$$

One has

$$\begin{aligned} \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} |t_{ij}|^2 &= \sum_{j=1}^{\infty} \sum_{\substack{k=1 \\ k \neq j}}^{\infty} \left| \frac{\alpha_k b_j}{\lambda_j - \nu_k} \right|^2 \\ &\leqq \gamma \|\alpha\|^2 \times \|b\|^2, \end{aligned} \tag{3.5}$$

where $\gamma = (\inf_{k \neq j} |\lambda_j - \nu_k|^2)^{-1}$. In virtue of assumption (2) and condition (F_2) one has $\gamma < \infty$. The infinite equation (3.3′) now becomes

$$(I+T)\begin{pmatrix} g_1 \\ g_2 \\ \vdots \\ g_n \\ \vdots \end{pmatrix} = -\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \\ \vdots \end{pmatrix}. \tag{3.6}$$

In order to solve (3.6) we need

LEMMA 3.1. *For arbitrary $2n$ distinct complex numbers $\nu_1, \cdots, \nu_n; \lambda_1, \cdots, \lambda_n$ we always have*

$$\det\left(\frac{1}{\nu_k - \lambda_j}\right)_{n\times n} = \frac{\prod_{j=1}^{n-1}\prod_{k=j+1}^{n}(\nu_k - \nu_j)(\lambda_j - \lambda_k)}{\prod_{k,j=1}^{n}(\nu_k - \lambda_j)}. \tag{3.7}$$

Proof. Think of $\nu_1, \cdots, \nu_n; \lambda_1, \cdots, \lambda_n$ as $2n$ complex variables. Then

$$f(\nu_1, \cdots, \nu_n; \lambda_1, \cdots, \lambda_n) \triangleq \prod_{k,j=1}^{n}(\nu_k - \lambda_j)\det\left(\frac{1}{\nu_k - \lambda_j}\right)_{n\times n} \tag{3.8}$$

is a polynomial in $\nu_1, \nu_2, \cdots, \nu_n, \lambda_1, \cdots, \lambda_n$. According to the properties of determinants $\det(1/(\nu_k - \lambda_j))_{n\times n}$ is zero when $\nu_k = \nu_j$, $k \neq j$. So $f(\nu_1, \cdots, \nu_n; \lambda_1, \cdots, \lambda_n)$ has a factor $(\lambda_j - \lambda_k)$ in it. Hence

$$\begin{aligned} &f(\nu_1, \cdots, \nu_n; \lambda_1, \cdots, \lambda_n) \\ &\quad = Q(\nu_1, \cdots, \nu_n; \lambda_1, \cdots, \lambda_n)\prod_{j=1}^{n-1}\prod_{k=j+1}^{n}(\nu_k - \nu_j)(\lambda_j - \lambda_k), \end{aligned} \tag{3.9}$$

where $Q(\nu_1, \cdots, \nu_n; \lambda_1, \cdots, \lambda_n)$ is a polynomial in $\nu_1, \cdots, \nu_n; \lambda_1, \cdots, \lambda_n$. By (3.8), $f(\nu_1, \cdots, \nu_n; \lambda_1, \cdots, \lambda_n)$ is a polynomial of degree $n \times n - n = n(n-1)$, and the power of $\prod_{j=1}^{n-1}\prod_{k=j+1}^{n}(\nu_k - \nu_j)(\lambda_j - \lambda_k)$ is $2\sum_{j=1}^{n-1}(n-j) = n(n-1)$. So $Q = \text{const}$. It is not hard to show that $Q \equiv 1$. That is

$$\prod_{k,j=1}^{n}(\nu_k - \lambda_j)\det\left(\frac{1}{\nu_k - \lambda_j}\right)_{n\times n} = \prod_{j=1}^{n-1}\prod_{k=j+1}^{n}(\nu_k - \nu_j)(\lambda_j - \lambda_k), \tag{3.10}$$

and this is just (3.7).

Now we prove that (3.6) has a unique solution $g \in H$, i.e., the unique solution

$$\begin{pmatrix} g_1 \\ g_2 \\ \vdots \end{pmatrix} \in l_2.$$

Denote by I_n the identity matrix of dimension $n \times n$, $T_n = (t_{ij})_{1 \leqq i,j \leqq n}$. By (3.7) we have

$$\begin{aligned} \det(I_n + T_n) &= \det\left(\frac{\alpha_k \bar{b}_j}{\lambda_j - \nu_k}\right)_{n\times n} \\ &= \prod_{k,j}^{n} \alpha_k \bar{b}_j \det\left(\frac{1}{\lambda_j - \nu_k}\right)_{n\times n} \\ &= \prod_{k,j=1}^{n} \alpha_k \bar{b}_j \frac{\prod_{j=1}^{n-1} \prod_{k=j+1}^{n} (\nu_k - \nu_j)(\lambda_j - \lambda_k)}{\prod_{k,j=1}^{n} (\lambda_k - \nu_j)} \\ &= \prod_{j=1}^{n-1} \prod_{k=j+1}^{n} \frac{(\nu_k - \nu_j)(\lambda_j - \lambda_k)}{(\lambda_k - \nu_j)(\lambda_j - \nu_k)}. \end{aligned} \tag{3.11}$$

Now consider the infinite product

$$\pi = \prod_{j=1}^{\infty} \prod_{k=j+1}^{\infty} \frac{(\nu_k - \nu_j)(\lambda_j - \lambda_k)}{(\lambda_k - \nu_j)(\lambda_j - \nu_k)}. \tag{3.12}$$

Because

$$\begin{aligned} \frac{(\nu_k - \nu_j)(\lambda_j - \lambda_k)}{(\lambda_k - \nu_j)(\lambda_j - \nu_k)} &= \left(1 - \frac{\alpha_k \bar{b}_k}{\lambda_k - \nu_j}\right) \cdot \left(1 + \frac{\alpha_k \bar{b}_k}{\nu_k - \lambda_j}\right) \\ &= 1 + \frac{\lambda_k - \nu_j - \nu_k + \lambda_j}{(\lambda_k - \nu_j)(\nu_k - \lambda_j)} \alpha_b \bar{b}_k - \frac{(\alpha_k \bar{b}_k)^2}{(\lambda_k - \nu_j)(\nu_k - \lambda_j)} \\ &= 1 + \frac{\alpha_j \bar{b}_j \alpha_k \bar{b}_k}{(\lambda_k - \nu_j)(\nu_k - \lambda_j)}. \end{aligned} \tag{3.13}$$

Therefore

$$\sum_{j=1}^{\infty} \sum_{k=j+1}^{\infty} \left| \frac{\alpha_j \bar{b}_j \alpha_k \bar{b}_k}{(\lambda_k - \nu_j)(\nu_k - \lambda_j)} \right| \leqq \gamma \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} |\alpha_j \bar{b}_j \alpha_k \bar{b}_k| \leqq \gamma \|\alpha\|^2 \times \|b\|^2, \tag{3.14}$$

where γ is the constant in (3.5). Therefore by familiar results (see for example [6, p. 17]) we have

$$\lim_{n\to\infty} \det(I_n + T_n) = \prod_{j=1}^{\infty} \prod_{k=j+1}^{\infty} \frac{(\nu_k - \nu_j)(\lambda_j - \lambda_k)}{(\lambda_k - \nu_j)(\lambda_j - \nu_k)} \neq 0. \tag{3.15}$$

Now by [7, Chapt. 9, § 17] we know that (3.9) has a unique solution

$$\begin{pmatrix} g_1 \\ g_2 \\ \vdots \end{pmatrix} \in l_2, \qquad \left(\text{i.e., } \sum_{k=1}^{\infty} |g_k|^2 < \infty\right),$$

or $g = \sum_{k=1}^{\infty} g_k \phi_k \in H$. And the g found in such a way obviously satisfies (3.2), and from (3.2) we have

$$(A + \langle \cdot, g \rangle b) x_k = \nu_k x_k, \qquad k \geqq 1. \tag{3.2'}$$

Since $\nu_k \neq \nu_j (\forall k \neq j)$ (3.2') says that

$$\sigma_p(A + \langle \cdot, g \rangle b) \supset \Lambda. \tag{3.16}$$

We must have

$$\sigma_p(A+\langle\cdot, g\rangle b)=\Lambda, \tag{3.17}$$

because otherwise there would exist complex numbers $\{\xi_1, \cdots, \xi_E\}$ where the positive $E \leqq \infty$, $\{\xi_1, \cdots, \xi_E\} \cap \Lambda = \varnothing$ and

$$\sigma_p(A+\langle\cdot, g\rangle b)=\Lambda \cup \{\xi_1, \cdots, \xi_E\}. \tag{3.18}$$

Denote $\Lambda \cup \{\xi_1, \cdots, \xi_E\} = \Lambda_E$. It follows from Theorem 2.1, the conditions in Theorem 1.1 and (3.16) that for $g \in H$ satisfying (3.16) there cannot exist $\{\xi_1, \cdots, \xi_E\}$, $(\notin \Lambda)$ such that both Λ and Λ_E satisfy (2.1). Hence $\{\xi_1, \cdots, \xi_E\} = \varnothing$; i.e., (3.17) holds.

In this way, we have finished the proof of the sufficiency part of Theorem 1.1 in Case 1.

Case 2. $\Lambda = \{\nu_1, \nu_2, \cdots, \nu_n, \cdots\}$ satisfies the conditions of the theorem and (1) $\nu_i \neq \nu_j$, for all $i \neq j$; (2) there exists a set J of positive integers such that

$$\nu_k = \begin{cases} \lambda_k, & k \notin J, \\ \neq \lambda_k, & k \in J, \end{cases} \qquad k \geqq 1.$$

Now we let

$$g = \sum_{k \in J} g_k \phi_k; \tag{3.19}$$

$g_k (k \in J)$ is determined as follows:

$$\begin{aligned} &\langle x_k, g\rangle = -1, \\ &x_k = (A-\nu_k)b^{-1}, \end{aligned} \qquad k \in J. \tag{3.20}$$

It is easy to see that this becomes, like (3.3),

$$\sum_{i \in J} \frac{b_j \bar{g}_j}{\lambda_j - \nu_k} = -1, \qquad k \in J. \tag{3.21}$$

The complex sequence $\Lambda_J = \bigcup_{k \in J} \{\nu_k\}$ and $\sigma_J(A) \triangleq \bigcup_{k \in j} \{\lambda_k\}$ appearing in (3.21) satisfies the conditions on Λ and $\sigma(A)$ in Case 1. Therefore (3.21) has a unique solution g_k, $(k \in J)$, $\sum_{k \in J} |g_k|^2 < \infty$. From (3.19), (3.20) and (3.21) we obviously have

$$\begin{aligned} &(A+\langle\cdot, g\rangle b)x_k = \nu_k x_k, \qquad k \in J, \\ &(A+\langle\cdot, g\rangle b)\phi_k = \lambda_k \phi_k, \qquad k \notin J. \end{aligned} \tag{3.21'}$$

This says that

$$\sigma_p(A+\langle\cdot, g\rangle b) \supset \Lambda. \tag{3.22}$$

Similarly, we have $\sigma_p(A+\langle\cdot, g\rangle b) = \Lambda$. This finishes the proof for Case 2.

Case 3. $\Lambda = \{\nu_1, \cdots, \nu_n, \cdots\}$ is any complex sequence which satisfies the condition in Theorem 1.1.

First, it follows from condition (ii) in the theorem that there exists a positive integer n_0 such that

$$\nu_i \neq \nu_j, \qquad i \neq j, \quad \forall i, j > n_0. \tag{3.23}$$

Now suppose

$$\Lambda \cap \{\lambda_1, \lambda_2, \cdots, \lambda_n, \cdots\} = \bigcup_{k \in J} \{\lambda_k\}, \tag{3.24}$$

where J is a set of integers.

Now form the complex sequence $\Lambda_J = \{\nu_1', \nu_2', \cdots, \nu_n', \cdots\}$ as follows:

$$\text{(3.25)} \qquad \begin{aligned} &(1) \quad \nu_k' \neq \nu_j', \qquad k \neq j, \quad \forall k, j, \\ &(2) \quad \nu_k' = \nu_k, \qquad k > n_0. \end{aligned}$$

It is easy to see not only that such a complex sequence exists, but also the choice of $\nu_k'(k \leqq n_0)$ is rather arbitrary. Hence Λ belongs to Case 2 and there exists $g' \in H$ such that

$$\text{(3.26)} \qquad \sigma_p(A + \langle \cdot, g' \rangle b) = \Lambda_J.$$

By Theorem 2.1 $A + \langle \cdot, g' \rangle b$ is a spectral operator with discrete spectrum, so it has [3] a bounded inverse linear operator ∇ such that $\nabla E'(\nu_k')\nabla^{-1}$ (denoted by F_k) is a self-adjoint operator in H. Here, $E'(\cdot)$ is the resolution of identity associated with $A + \langle \cdot, g' \rangle b$. Then

$$\text{(3.27)} \qquad \nabla(A + \langle \cdot, g' \rangle b)\nabla^{-1} = \sum_{k=1}^{\infty} \nu_k' F_k.$$

Because $\nu_k' \neq \nu_j'(k \neq j)$, we have $\dim E'(\nu_k)H = \dim F_k H = 1$, $k \geqq 1$. Let ψ_k be a vector in $F_k H$ of modulus 1. It is easy to see that ψ_k, $k \geqq 1$ form an orthonormal basis in H and are the eigenvectors belonging to ν_k' of the operator $A_1 \triangleq \nabla(A + \langle \cdot, g' \rangle b)\nabla^{-1}$. Denote $b' = \nabla b$. We claim that

$$\text{(3.28)} \qquad \langle \psi_k, b' \rangle \neq 0, \qquad k \geqq 1.$$

Indeed, let us assume that (3.28) is not true; then there must exist k_0 such that

$$\text{(3.29)} \qquad \langle \psi_{k_0}, b' \rangle = 0.$$

It follows from the proof of the sufficiency part of Theorem 1.1 (because it is not hard to show that A_1 still satisfies condition F) that no matter how we choose the $g \in H$, we must have

$$\text{(3.30)} \qquad \nu_{k_0}' \in \sigma_p(A_1 + \langle \cdot, g \rangle b') \quad \forall g \in H.$$

i.e.,

$$\text{(3.31)} \qquad \begin{aligned} \nu_{k_0}' &\in \sigma_p(A + \langle \cdot, g' + \nabla^* g \rangle b) \\ &= \sigma_p(\nabla(A + \langle \cdot, g' + \nabla^* g \rangle b)\nabla^{-1}) \\ &= \sigma_p(A_1 + \langle \cdot, g \rangle b'), \quad \forall g \in H. \end{aligned}$$

Because ∇ (hence ∇^*) has a bounded inverse, (3.31) reduces to Case 1 which we have already proved. This is because when we are given any Λ satisfying the conditions in Case 1 and such that $\nu_{k_0}' \notin \Lambda$, there must exist $g_\Lambda \in H$ such that

$$\text{(3.32)} \qquad \nu_{k_0}' \notin \Lambda = \sigma_p(A + \langle \cdot, g_\Lambda \rangle b).$$

If we let $g = \nabla^{*-1}(g_\Lambda - g')$, then (3.32) contradicts (3.31).

This says that (3.28) is true.

Now let $g'' = \sum_{k=1}^{n_0} g_k'' \psi_k$. Then $A_1 + \langle \cdot, g'' \rangle b'$ has, with respect to the basis ψ_k, $k \geqq 1$, the matrix representation

$$A_1 + \langle \cdot, g'' \rangle b \sim \begin{bmatrix} \nu_1' & & & & \\ & \nu_2' & & 0 & \\ & & \ddots & & \\ & & & \nu_{n_0}' & \\ & 0 & & & \nu_{n_0+1}' \\ & & & & & \ddots \end{bmatrix} + \begin{bmatrix} g_1'' b_1' & \cdots & g_{n_0}'' b_1' & \\ \vdots & & \vdots & \\ g_1'' b_{n_0}' & \cdots & g_{n_0}'' b_{n_0}' & 0 \\ \vdots & & \vdots & \end{bmatrix} \tag{3.33}$$

$$= \begin{bmatrix} p_{n_0} A_1 P_{n_0} + \langle \cdot, g'' \rangle p_{n_0} b' & \vdots & 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots & \vdots & \cdots\cdots\cdots\cdots\cdots \\ \langle \cdot, g'' \rangle (I - P_{n_0}) b' & \vdots & (I - P_{n_0}) A_1 (I - P_{n_0}) \end{bmatrix}.$$

Here $P_{n_0} = \sum_{k=1}^{n_0} F_k$, $b' = \sum_{k=1}^{\infty} b_k' \psi_k$. It follows from (3.33) that the spectrum of $A_1 + \langle \cdot, g'' \rangle b''$ is composed of the spectrum of $P_{n_0} A_1 P_{n_0} + \langle \cdot, g'' \rangle P_{n_0} b'$ in $P_{n_0} H$ and the spectrum of $(I - P_{n_0}) A_1 (I - P_{n_0})$ in $(I - P_{n_0}) H$.

On the other hand, from (3.28) and the simplicity (i.e., $\nu_k' \neq \nu_j'$, $(\forall i \neq j)$) of the spectrum of A_1 it follows immediately that

$$\text{Rank } (P_{n_0} b', P_{n_0} A_1 P_{n_0} b', \cdots, (P_{n_0} A_1 P_{n_0})^{n_0 - 1} P_{n_0} b'\} = n_0. \tag{3.34}$$

Hence by familiar results [1], we have that for any given complex numbers $\nu_1, \cdots, \nu_{n_0}$, there exists $g'' \in P_{n_0} H$ such that in $P_{n_0} H$ we have

$$\sigma_p(P_{n_0} A_1 P_{n_0} + \langle \cdot, g'' \rangle P_{n_0} b') = \{\nu_1, \cdots, \nu_{n_0}\}. \tag{3.35}$$

Hence for the $g'' = \sum_{k=1}^{n_0} g_k'' \psi_k (\in P_{n_0} H)$ determined from above we have

$$\sigma_p(A_1 + \langle \cdot, g'' \rangle b') = \{\nu_1, \cdots, \nu_{n_0}, \nu_{n_0+1}, \cdots\} = \Lambda. \tag{3.36}$$

Let $g = g' + \nabla^* g''$, we immediately get

$$\begin{aligned} \sigma_p(A + \langle \cdot, g \rangle b) &= \sigma_p(\nabla (A + \langle \cdot, g \rangle b \nabla^{-1}) \\ &= \sigma_p(A_1 + \langle \cdot, g'' \rangle b') = \Lambda, \end{aligned} \tag{3.37}$$

and (3.37) is exactly the proof for Case 3 of the sufficiency part of Theorem 1.1.

Hence the proof of Theorem 1.1 is complete.

4. Proof of Theorem 1.2.

Proof of sufficiency. Let $b = \sum_{k=1}^{\infty} b_k \phi_k$ where

$$b_k = \begin{cases} \sqrt{\text{Re } \lambda_k}, & k \in J, \\ \dfrac{1}{k}, & k \notin J. \end{cases} \tag{4.1}$$

Obviously $b \in H$ and $b_k \neq 0$, $(k \geqq 1)$. Let

$$\nu_k = \begin{cases} \lambda_k - 2 \operatorname{Re} \lambda_k, & k \in J, \\ \lambda_k - \dfrac{1}{2^k}, & k \notin J. \end{cases} \tag{4.2}$$

We have

$$\operatorname{Re} \nu_k < 0, \qquad k \geqq 1, \tag{4.3}$$

and

$$\sum_{k=1}^{\infty} \left| \frac{\lambda_k - \nu_k}{b_k} \right|^2 \leqq \sum_{k \notin J} \left(\frac{k}{2^k} \right)^2 + 4 \sum_{k \in J} \operatorname{Re} \lambda_k < \infty. \tag{4.4}$$

So by Theorem 1.1 there exists $g \in H$ such that

$$\sigma_p(A + \langle \cdot, g \rangle b) = \{\nu_1, \nu_2, \cdots, \nu_n, \cdots\}. \tag{4.5}$$

From (4.3), (4.5) and using an argument similar to that in [4], it follows that system (1.2) is stable under feedback (1.3).

Proof of necessity. Suppose that there exist $g, b \in H$ such that the system (1.2) is stable under feedback (1.3). By Theorem 2.1, $A + \langle \cdot, g \rangle b$ is an operator with a discrete spectrum. Let

$$\sigma_p(A + \langle \cdot, g \rangle b) = \{\nu_1, \nu_2, \cdots, \nu_n, \cdots\}. \tag{4.6}$$

According to the stability condition, we must have

$$\operatorname{Re} \nu_k \leqq 0, \qquad k \geqq 1. \tag{4.7}$$

So by (2.1) and (4.7) we have (without loss of generality we assume that (2.1) holds for all $k \in J$):

$$\begin{aligned} \sum_{k \in J} \operatorname{Re} \lambda_k &\leqq \sum_{k \in J} |\lambda_k - \nu_k| \\ &\leqq 6 \sum_{k \in J} |b_k g_k| \\ &\leqq 6 \|b\| \times \|g\|. \end{aligned} \tag{4.8}$$

Thus we get the proof of Theorem 1.2.

5. Some remarks on complete controllability and the distribution of the spectrum. We know that as long as $b_k \neq 0$, $k \geqq 1$ the system (1.2) is completely controllable [2], [5] under condition F. However, Theorem 1.1 asserts that it is not possible that for any arbitrary complex sequence $\Lambda = \{\nu_1, \nu_2, \cdots, \nu_n, \cdots\}$ (as long as it does not satisfy condition (ii) of Theorem 1.1) there exists $g \in H$ satisfying

$$\sigma_p(A + \langle \cdot, g \rangle b) = \Lambda.$$

This is intrinsically different from the case of finite dimensional spaces. When the space is finite dimensional for an autonomous completely controllable linear system, there must exist a suitable linear feedback such that the operator associated with the closed-loop system has any spectrum we have preassigned.

REFERENCES

[1] W. M. WONHAM, *On pole assignment in multi-input controllable linear systems*, IEEE Trans. Automat. Control, AC12 (1967), pp. 660–665.

[2] H. O. FATTORINI, *On complete controllability of linear systems*, J. Differential Equations, 3 (1967), pp. 391–402.

[3] P. WERMER, *Commuting spectral operations on Hilbert spaces*, Pacific J. Math., 4 (1954), pp. 355–361.

[4] J. T. SCHWARTZ, *Perturbation on spectral operators, and applications*, I., Pacific J. Math., 4 (1954), pp. 415–458.

[5] KWAN CHAO-CHIH AND WANG KANG-NING, *Sur la stabilization de la vibration élastique*, Scientia Sinica, 4 (1974), pp. 446–447.

[6] E. C. TITCHMARSH, *The Theory of Functions*, Oxford University Press, Oxford, 1939.

[7] A. C. ZAANEN, *Linear Analysis*, North-Holland, Amsterdam, 1956.

[8] I. C. GOKHBERG AND M. G. KREIN, *Introduction to the Theory of Linear Non-Self Adjoint Operators*, American Mathematical Society, Providence, RI, 1969.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 6, November 1981

0363-0129/81/1906-0003 $01.00/0

FINITE ELEMENTS AND TERMINAL PENALIZATION FOR QUADRATIC COST OPTIMAL CONTROL PROBLEMS GOVERNED BY ORDINARY DIFFERENTIAL EQUATIONS*

GOONG CHEN† AND WENDELL H. MILLS, JR.‡

Abstract. We use the finite element method to compute optimal controls of systems governed by linear ordinary differential equations with a quadratic performance index. As an application we use the penalty technique to solve terminal state optimal controllability problems. Numerical instabilities, which are common in the use of penalty, are minimized when the finite element method is applied to solve this problem. Convergence theorems are given and error and penalty parameter estimates are presented. Concrete examples for various situations are given to illustrate the theory.

Introduction. Given a finite dimensional linear control system

$$\text{(LC)}\qquad \begin{aligned} &\frac{dx(t;u)}{dt} = A(t)x(t;u)+f(t)+B(t)u(t), \qquad 0 \leqq t \leqq T,\\ &x(0;u) = x_0, \end{aligned}$$

where

$x(t)$ is the state of the system at time $t \in \mathbb{R}^n$,

x_0 is the initial state,

$u \in U_{\text{ad}}$ is an admissible control, $u(t) \in \mathbb{R}^m$,

$A(t) \equiv (a_{ij}(t))_{n\times n}$, $B(t) \equiv (b_{pq}(t))_{n\times m}$ are $n \times n$, $n \times m$ time-varying matrices,

$f \in L_n^2(0, T)$,

the *quadratic cost optimal control problem* is to minimize

$$(0.1)\qquad J(x_0, u) \equiv \int_0^T [|C_1(t)x(t;u)-z_1(t)|^2 + \langle N(t)u(t), u(t)\rangle]\, dt + \gamma |C_2x(T;u)-z_2|^2$$

with

$C_1(t) = (C_{ij}(t))_{l_1\times n}$, $z_1 \in L_{l_1}^2(0, T)$, $\gamma \geqq 0$,

C_2 is a constant $l_2 \times n$ matrix, $z_2 \in \mathbb{R}^{l_2}$,

$N(t)$ is a symmetric $m \times m$ time-varying matrix satisfying

$$(0.2)\qquad \langle Nu, u\rangle_{L_m^2(0,T)} \geqq \nu \|u\|^2_{L_m^2(0,T)}, \qquad \nu > 0.$$

In the early 1960's, Kalman and Bucy [12], [13] introduced the quadratic performance criteria which became a standard design technique for finite-dimensional linear systems. Consider the unconstrained case $U_{\text{ad}} \equiv L_m^2(0, T)$. The existence and uniqueness of the optimal control minimizing (0.1) follows from [14]. In (LC), let $u = \hat{u}$ denote the optimal control. One introduces the Lagrange multiplier $p(t)(\in \mathbb{R}^n)$ which

* Received by the editors December 26, 1979, and in revised form December 18, 1980.

† Department of Mathematics, Pennsylvania State University, University Park, Pennsylvania 16802. The work of this author was supported in part by the National Science Foundation under grant MCS 7822830 and in part by IRIA-LABORIA.

‡ Department of Mathematics, Pennsylvania State University, University Park, Pennsylvania 16802.

satisfies the adjoint system

$$\begin{aligned} &\frac{dp(t)}{dt} = -A^*(t)p(t) - C_1^*(t)C_1(t)[x(t;\hat{u}) - z_1(t)], \\ &p(T) = \gamma C_2^*[C_2 x(T;\hat{u}) - z_2]. \end{aligned} \tag{0.3}$$

This leads to a coupled system (LC), (0.3), which becomes a two-point boundary value problem. One may use an iterative shooting method, direct differencing, or decoupling into Riccati equations to solve for (x, p). The optimal control $\hat{u}$ is then obtained as

$$\hat{u}(t) = -N^{-1}(t)B^*(t)p(t), \qquad 0 \leqq t \leqq T. \tag{0.4}$$

This is perhaps the most often quoted numerical method in control literature.

The first mathematical study on computing the optimal control by another method using *finite elements* seems to be done by Bosarge and Johnson [3]. This and similar methods have been closely examined and used by many control theorists and numerical analysts to compute optimal controls for more complex problems—those with state and control constraints, nonlinear dynamics, etc. We mention [7], [8], [9], [11], [16] for a few such references on controlled ODE's. Comparatively much less literature has dealt with controlled PDE's; cf. [22].

Though the study of finite element applications to controlled ODE's has been carried out in many forms, many interesting questions still remain. The first main question in this paper deals with an important control *system* problem—that of terminal state controllability at a specified time T. In system design, attaining a prescribed target is often as important as the minimization of some cost functional. In this study, we wish to attain a given target while also minimizing an integral quadratic cost. Another question, which is perhaps more of numerical analysis in nature, is about the feasibility of a penalty technique applied to this terminal state problem. If one were to compute such a penalized problem via Riccati synthesis, the resulting equations can be seen (cf. (2.35, (2.36)) to be ill-conditioned. We find that using finite elements minimizes the effect of such instabilities while predicting quite accurate results.

In [3], Bosarge and Johnson's finite element algorithms are based upon Ritz–Trefftz's, a *dual* method. Minimization of a quadratic functional is reduced to the familiar form [14, Chap. 1] of solving

$$a(\hat{u}, v) = \theta(v) \quad \forall v \in V_h, \tag{0.5}$$

in a finite dimensional approximation space V_h. In § 1 of this paper, we also reduce our minimization problem to an equivalent one of this form, but without relying on a dual state. We proceed by a *primal* approach; a and θ in (1.4), (1.5) are very different from their counterparts in [3]. We prove a regularity theorem and derive optimal error estimates for the problem.

In § 2 we present the penalized problem (cf. (PQCCP)) as an application of § 1. We first prove the basic convergence Theorem 2.1 and then proceed to derive rates of convergence and error bounds for uncontrollable and controllable cases. The main results of this paper are Theorems 2.5, 2.6 and 2.7. The error bounds given in Theorem 2.7 are optimal.

In § 3 we apply our method to various examples. The stability and accuracy of the solutions comply exactly with the main error estimates and indicate that the technique is quite successful in producing accurate states and optimal controls.

1. Computations by the finite element method. Throughout this section we assume that the quadratic performance index $J(x_0, u)$ takes the form (0.1) and $U_{ad} = L^2_m(0, T)$. Other variants of $J(x_0, u)$ such as those in [4], [19] can also be treated without difficulty.

We let $\Phi(t, s)$ denote the $(n \times n)$ fundamental matrix solution satisfying

$$\frac{d}{dt}\Phi(t, s) = A(t)\Phi(t, s), \qquad 0 \leqq s \leqq t \leqq T,$$

$$\Phi(s, s) = I_{n\times n}.$$

For simplicity, $\Phi(t, 0)$ is denoted by $\Phi(t)$. We also use $F(t)$ to denote

$$F(t) = \int_0^t \Phi(t, s)f(s)\, ds.$$

Define a linear operator

$$\mathscr{L}_1 : L^2_m(0, T) \to L^2_n(0, T),$$

by

$$(\mathscr{L}_1 u)(t) = \int_0^t \Phi(t, s)B(s)u(s)\, ds. \tag{1.1}$$

Then, for any given $u \in L^2_m(0, T)$, the solution $x(t)$ of (LC) can be written as

$$x(t; u) = \Phi(t)x_0 + F(t) + (\mathscr{L}_1 u)(t), \qquad 0 \leqq t \leqq T. \tag{1.2}$$

Substituting (1.2) into (0.1) and using the calculus of variations, we know that in the unconstrained case the optimal control $\hat{u}$ is characterized by the variational equation

$$\tfrac{1}{2}J'(\hat{u}) \cdot v \equiv a(\hat{u}, v) - \theta(v) = 0 \quad \text{for all } v \in L^2_m(0, T), \tag{1.3}$$

where $a(\cdot, \cdot)$ is a symmetric bilinear form on $L^2_m(0, T) \times L^2_m(0, T)$ defined by

$$\begin{aligned} a(v_1, v_2) = {} & \langle Nv_1, v_2\rangle_{L^2_m(0,T)} + \langle C_1(\mathscr{L}_1 v_1), C_1(\mathscr{L}_1 v_2)\rangle_{L^2_{l_2}(0,T)} \\ & + \gamma\langle C_2(\mathscr{L}_1 v_1)(T), C_2(\mathscr{L}_1 v_2)(T)\rangle_{\mathbb{R}^{l_2}} \end{aligned} \tag{1.4}$$

and θ is a linear functional on $L^2_m(0, T)$ defined by

$$\begin{aligned} \theta(v) = {} & \langle z_1 - C_1[\Phi(\cdot)x_0 + F], C_1(\mathscr{L}_1 v)\rangle_{L^2_{l_1}(0,T)} \\ & + \gamma\langle z_2 - C_2[\Phi(T)x_0 + F(T)], C_2(\mathscr{L}_1 v)(T)\rangle_{\mathbb{R}^{l_2}}. \end{aligned} \tag{1.5}$$

LEMMA 1.1. *Let $A \in L^\infty_{n\times n}(0, T)$ and $B \in L^\infty_{n\times m}(0, T)$. If C_1, N, z_1 are L^2-functions, then there exist K_1, K_2, K_3, $K_4 > 0$ independent of u, v, γ such that*

$$a(u, u) \geqq \nu\|u\|^2_{L^2_m(0,T)} \qquad (\nu \text{ as in (0.2)}),$$

$$|a(u, v)| \leqq [K_1 + K_2\gamma]\|u\|\|v\|,$$

$$\|\theta\|_{[L^2_m(0,T)]'} \leqq K_3 + K_4\gamma.$$

Proof. Since the operator

$$\mathscr{L}_1 : v(\cdot) \to \int_0^t \Phi(t, s)B(s)v(s)\, ds$$

is Hilbert–Schmidt in $L^2_m(0, T)$, the result follows. □

THEOREM 1.2. *The equation*

$$a(\hat{u}, v) = \theta(v) \quad \text{(cf. (1.3))} \quad \forall v \in L^2_m(0, T) \tag{1.6}$$

has a unique solution $\hat{u}$ *which is the optimal control minimizing* (0.1) *subject to* (LC). *Furthermore,*

$$\|\hat{u}\|_{L^2_m(0,T)} \leqq \frac{K_3 + K_4\gamma}{\nu}.$$

Proof. This follows directly from Lemma 1.1 and the Lax–Milgram theorem [1]. □

From (1.6), the application of the finite element method is quite straightforward. We let $\hat{S}^h \subseteq L^2_m(0, T)$ be a finite dimensional approximating subspace and we pose (1.6) on this subspace:

$$\text{find } \hat{u}_h \in \hat{S}^h \text{ such that } a(\hat{u}_h, v) = \theta(v) \quad \forall v \in \hat{S}^h. \tag{1.7}$$

Let $\{\hat{\varphi}_1, \hat{\varphi}_2, \cdots, \hat{\varphi}_M\}$ be a basis for $\hat{S}^h$. Then (1.7) is equivalent to the matrix equation

$$K_h \hat{q}_h = \hat{g}_h, \tag{1.8}$$

where $K_h = [(K_h)_{ij}] = [a(\hat{\varphi}_i, \hat{\varphi}_j)]$ is an $M \times M$ symmetric positive definite matrix, $1 \leqq i, j \leqq M$.

$$\hat{g} = [(\hat{g}_h)_i] = [\theta(\hat{\varphi}_i)], \qquad 1 \leqq i \leqq M,$$

$$\hat{q}_h = [q_i], \qquad 1 \leqq i \leqq M, \quad \text{with } \hat{u}_h = \sum_{i=1}^{M} q_i \hat{\varphi}_i.$$

The linear matrix equation (1.8) can be solved by standard large order system solvers with iterative improvement. The only inputs are $A(t)$, $B(t)$, $f(t)$, $\Phi(t, s)$, $C_1(t)$, $z_1(t)$, $N(t)$, C_2 and z_2. A finite difference ODE solver may be used to obtain $\Phi(t, s)$, and high order quadrature to calculate K_h and $\hat{g}_h$. Examples in § 3 show that our computations produce extremely accurate results as compared with exact solutions.

THEOREM 1.3 (regularity). *Let* $k_1, k_2, k_3, k_4 \in \mathbb{Z}^+$ *be such that*

$$\left(\frac{d}{dt}\right)^{k_1} A \in L^\infty(0, T; \mathbb{R}^{n\times n}),$$

$$\left(\frac{d}{dt}\right)^{k_2} B \in L^\infty(0, T; \mathbb{R}^{m\times n}),$$

$$\left(\frac{d}{dt}\right)^{k_3} N^{-1} \in L^\infty(0, T; \mathbb{R}^{m\times m}),$$

$$\left(\frac{d}{dt}\right)^{k_4} C_1 \in L^\infty(0, T; \mathbb{R}^{l_1\times n}),$$

and let $z_1 \in H^{k_5}_{l_1}(0, T)$ *and* $f \in H^{k_6}_n(0, T)$. *Then the solution* (x, p) *of* (LC), (0.3) *and* (0.4) *is in* $H^{s_1+1}_n(0, T) \oplus H^{s_1+1}_n(0, T)$ *with* $s_1 = \min_{1\leqq i\leqq 6, i\neq 3}\{k_i\}$ *and the optimal control* $\hat{u}$ *is in* $H^{s_2}_m(0, T)$ *with* $s_2 = \min\{k_2, k_3, s_1 + 1\}$.

Proof. The two-point boundary value problem

$$\frac{d}{dt}\begin{bmatrix} x(t) \\ p(t) \end{bmatrix} = \begin{bmatrix} A(t) & -B(t)N^{-1}(t)B^*(t) \\ -C_1^*(t)C_1(t) & -A^*(t) \end{bmatrix}\begin{bmatrix} x(t) \\ p(t) \end{bmatrix} + \begin{bmatrix} f(t) \\ C_1^*(t)z_1(t) \end{bmatrix},$$

$$x(0) = x_0, \tag{1.9}$$

$$p(T) = \gamma C_2^*[C_2 x(T) - z_2]$$

has a solution $(x, p) \in H_n^1(0, T) \oplus H_n^1(0, T)$, provided that $A(t)$, $B(t)$, $C_1(t)$ and $N^{-1}(t)$ are in $L_j^\infty(0, T)$ for $j = n \times n$, $n \times m$, $n \times l_1$ and $m \times m$, respectively. Differentiating both sides of (1.9), one obtains higher regularity results for (x, p). The regularity of $\hat{u}$ follows from (0.4). □

Following [1, Chap. 4] we assume $\hat{S}^h \subseteq L_m^2(0, T)$ to be an $(r, 0)$-system. That is, for $u \in H_m^l(0, T)$, there exists $v \in \hat{S}^h$ such that

$$\|u - v\|_{L_m^2} \leqq K \cdot h^\mu \|u\|_{H_m^l}, \tag{1.10}$$

where $r > 0$, $\mu = \min(r, l)$ and K is a positive constant independent of u and h.

From now on throughout the rest of this paper, all of our error estimates are based upon a tacit assumption that *the fundamental matrix solution* $\Phi(t, s)$ (*cf.* (1.1)) *is exact,* thereby enabling the bilinear and linear forms *a and* θ *to be exact.*

THEOREM 1.4 (error estimates). *Let* $\hat{S}^h \subseteq L_m^2(0, T)$ *be an* $(r, 0)$*-system. Assume* $\hat{u} \in H_m^s(0, T)$ *is the optimal control solving* (LC), (0.1) *with* $s > 0$. *Let* $\hat{u}_h$ *be the finite element solution* (1.7). *Then for* $\mu = \min(r, s)$ *and some* $K_5 > 0$ (*independent of* h, γ, $\hat{u}$),

$$\|\hat{u}_h - \hat{u}\|_{L_m^2(0,T)} \leqq \left(\frac{K_1 + K_2\gamma}{\nu}\right)^{1/2} h^\mu \|\hat{u}\|_{H_m^s(0,T)}, \tag{1.11}$$

$$\sup_{[0,T]} |x(t; \hat{u}_h) - x(t; \hat{u})|_{\mathbb{R}^n} \leqq K_5 \left(\frac{K_1 + K_2\gamma}{\nu}\right)^{1/2} h^\mu \|\hat{u}\|_{H_m^s(0,T)}, \tag{1.12}$$

where K_1, K_2 *are the constants in Lemma* 1.1.

Proof. Since a is symmetric, we can apply [15, p. 51] and Lemma 1.1 to get

$$\|\hat{u}_h - \hat{u}\|_{L_m^2(0,T)} \leqq \left(\frac{K_1 + K_2\gamma}{\nu}\right)^{1/2} \inf_{v \in \hat{S}^h} \|v - \hat{u}\|_{L_m^2(0,T)}.$$

(1.11) then follows from (1.10). (1.12) follows from (1.11) and the estimate

$$\begin{aligned} |x(t; \hat{u}_h) - x(t; \hat{u})| &= \left| \int_0^t \Phi(t, s) B(s) [\hat{u}_h(s) - \hat{u}(s)] \, dx \right| \\ &\leqq K_5 \left(\frac{K_1 + K_2\gamma}{\nu}\right)^{1/2} h^\mu \|\hat{u}\|_{H_m^s(0,T)}. \quad \square \end{aligned}$$

§ 2. Penalization and finite element approximation of an optimal controllability problem with terminal condition. In this section, we apply the penalty method and § 1 to study a *quadratic cost controllability problem*

(QCCP). For given $x_1 \in \mathbb{R}^n$, find an optimal control $\hat{u} \in L_m^2(0, T)$ such that $x(t; \hat{u})$ solves (LC),

$$x(T; \hat{u}) = x_1 \tag{2.1}$$

and $\hat{u}$ minimizes the integral quadratic cost

$$J(x_0, u) = \int_0^T [|C(t)x(t; u) - z(t)|^2 + \langle N(t)u(t), u(t)\rangle] \, dt \tag{2.2}$$

over $L_m^2(0, T)$.

We can regard (2.1) as a constraint, augment (2.2) with a penalty term $(1/\varepsilon)|x(T; u) - x_1|^2$, and study the resulting *unconstrained* penalized problem:

(PQCCP). Find an optimal control $\hat{u}_\varepsilon \in L_m^2(0, T)$ such that $x(t; \hat{u}_\varepsilon)$ solves (LC) and the penalized cost

$$J_\varepsilon(x_0, u) = \int_0^T [|C(t)x(t;u) - z(t)|^2 + \langle N(t)u(t), u(t)\rangle]\, dt + \frac{1}{\varepsilon}|x(T; u) - x_1|^2$$

(2.3)

$$\equiv J(x_0, u) + \frac{1}{\varepsilon}|x(T; u) - x_1|^2, \qquad \varepsilon > 0$$

is minimized by $\hat{u}_\varepsilon$ over $L_m^2(0, T)$. This functional J_ε is the J in § 1 with $C_1(t) = C(t)$, $z_1(t) = z(t)$, $\gamma = 1/\varepsilon$, $C_2 = I_{n\times n}$, $z_2 = x_1 (l_1 = l, l_2 = n)$. This idea has been briefly mentioned in [2, p. 29], [6, p. 208]. It also falls into the category of [21, Algorithm 6.1] "pure increased" penalty technique. It is known that with the ordinary finite difference technique numerical instabilities will result when ε becomes small [9]. Here we show that by using our algorithm in § 1, we can analyze the parameters ε, h in such a way that such computations become feasible and errors can be minimized.

It is to be understood that this section (and this paper) is not intended as an exposition of general penalty techniques and their most abstract possible outcomes. Rather, our objectives are strictly those as outlined in the Introduction. Some convergence arguments presented here are standard proofs involving penalty and can probably be found under a more general setting elsewhere; see, e.g., [21]. We nevertheless present detailed proofs in the subsequent theorems so that *rates* of convergence can be carefully examined and *error analysis* can be made. This seems to be new and is the main thrust of this paper.

We first give two fundamental theorems of this paper, Theorems 2.1 and 2.2.

THEOREM 2.1. *Let $\hat{u}_\varepsilon$ denote the optimal control obtained from the penalized problem* (PQCCP). *Then there exists a unique $\hat{u} \in L_m^2(0, T)$ such that*

$$\hat{u}_\varepsilon \to \hat{u} \quad \text{strongly in } L_m^2(0, T).$$

The control $\hat{u}$ has the property that

$$J(x_0, \hat{u}) = \inf_{v \in W} J(x_0, v)$$

with

$$W = \{v \in L_m^2(0, T) \big| |x(T; v) - x_1| = \inf_{u \in L_m^2} |x(T; u) - x_1|\}.$$

That is to say, $\hat{u}$ is the unique control which makes $|x(T; u) - x_1|$ the smallest while making J small.

Proof. From now on, we write $J_\varepsilon(u)$, $J(u)$ instead of $J_\varepsilon(x_0, u)$, $J(x_0, u)$ when no ambiguity occurs.

Since $\hat{u}_\varepsilon$ minimizes J_ε, it is also the unique solution to

$$\inf_{u \in L_m^2} \varepsilon J_\varepsilon(u) = \inf_{u \in L_m^2} [\varepsilon J(u) + |x(T; u) - x_1|^2].$$

Choose a fixed element $v_0 \in W$. It steers $x(t)$ to the point $x(T; v_0)$ (unique!) closest possible to x_1, i.e.,

$$|x(T; v_0) - x_1| = \inf_u |x(T; u) - x_1| \geqq 0.$$

Such a v_0 is obviously nonunique.

Then we have

$$\varepsilon J(v_0)+|x(T;v_0)-x_1|^2 \geqq \varepsilon J_\varepsilon(\hat{u}_\varepsilon)=\varepsilon J(x_0,\hat{u}_\varepsilon)+|x(T;\hat{u}_\varepsilon)-x_1|^2$$
$$\geqq |x(T;v_0)-x_1|^2.$$

Letting $\varepsilon\downarrow 0$, we conclude that

$$\varepsilon J(\hat{u}_\varepsilon)\downarrow 0 \quad \text{and} \quad |x(T;\hat{u}_\varepsilon)-x_1|\downarrow|x(T;v_0)-x_1|.$$

Since

$$J(v_0)\geqq J(\hat{u}_\varepsilon),$$

we obtain a weakly convergent subsequence

$$\hat{u}_{\varepsilon_n}\to\hat{u} \quad \text{weakly in } L^2_m$$

for some weak limit $\hat{u}$. It is not difficult to see that this weak convergence is also *strong*, and that since *every* subsequence $\hat{u}_{\varepsilon_n}$ converges strongly to $\hat{u}$ in $L^2_m(0,T)$, we conclude $\hat{u}_\varepsilon\to\hat{u}$ strongly. □

Remarks. (1) The above theorem indicates that the methods of this paper will produce, as an application, a *stable* scheme for solving $\inf_{u\in L^2_m}|x(T,u)-x_1|^2$. Note that a direct solution to this problem by a gradient or conjugate gradient method will be unstable since its solution is nonunique.

(2) If (LC) is not controllable from x_0 to x_1, then

$$\inf_u |x(T;u)-x_1|^2=M>0,$$

so $J_\varepsilon(\hat{u}_\varepsilon)$ grows *unbounded*: $J_\varepsilon(\hat{u}_\varepsilon)\geqq M/\varepsilon\uparrow\infty$ as $\varepsilon\downarrow 0$. Nevertheless, the convergence

$$\hat{u}_\varepsilon\to\hat{u} \quad \text{strongly in } L^2_m$$

always holds for some $\hat{u}\in L^2_m$.

If the system (LC) is controllable from x_0 to x_1, for a given pair (x_0,x_1), then we have the following stronger result.

THEOREM 2.2. *The system* (LC) *is controllable from* x_0 *to* x_1 *for some* x_0, x_1 *if and only if* $J_\varepsilon(x_0,\hat{u}_\varepsilon)$ *is bounded from above by some* $M>0$ (*depending on* x_0, x_1). *If* (LC) *is controllable from* x_0 *to* x_1, *then the solution to the penalized problem* (PQCCP), $\hat{u}_\varepsilon$, *converges strongly to a control* $\hat{u}$ (*as* $\varepsilon\to 0$) *which solves* (QCCP). *Furthermore,*

$$|x(T;\hat{u}_\varepsilon)-x_1|=o(\sqrt{\varepsilon}) \quad \text{as } \varepsilon\downarrow 0. \tag{2.4}$$

Proof. If there is a control $\bar{u}$ steering x_0 to x_1, then

$$J(\bar{u})=J(\bar{u})+\frac{1}{\varepsilon}|x(T;\bar{u})-x_1|^2$$
$$=J_\varepsilon(\bar{u})\geqq\min_u J_\varepsilon(u)=J_\varepsilon(\hat{u}_\varepsilon). \tag{2.5}$$

Thus $J_\varepsilon(\hat{u}_\varepsilon)$ is bounded from above for all $\varepsilon\in\mathbb{R}^+$ by $M\equiv J(\bar{u})$.

Conversely, assume that $J_\varepsilon(\hat{u}_\varepsilon)$ is bounded from above by some $M>0$. Then

$$J(\hat{u}_\varepsilon)+\frac{1}{\varepsilon}|x(T;\hat{u}_\varepsilon)-x_1|^2\leqq M.$$

Hence

$$|x(T;\hat{u}_\varepsilon)-x_1|\leqq\sqrt{M}\sqrt{\varepsilon}\to 0 \quad \text{as } \varepsilon\to 0. \tag{2.6}$$

Because $\{\hat{u}_\varepsilon\}$ is bounded in $L^2_m(0, T)$, it contains a subsequence $\hat{u}_{\varepsilon_n} \to \hat{u}$ weakly convergent. From (2.6), we see that

$$x(T; \hat{u}) = \lim_{\varepsilon_n \downarrow 0} \left\{ \Phi(T)x_0 + F(T) + \int_0^T \Phi(T, s)B(s)\hat{u}_{\varepsilon_n}(s)\, ds \right\}$$

$$= \lim_{\varepsilon_n \downarrow 0} x(T; \hat{u}_{\varepsilon_n}) = x_1.$$

So $\hat{u}$ accomplishes controllability. (LC) is controllable from x_0 to x_1.

From (2.5), we have

$$J(\hat{u}) \geqq J_\varepsilon(\hat{u}_\varepsilon) \geqq J(\hat{u}_\varepsilon).$$

On the other hand, weak convergence of $\hat{u}_{\varepsilon_n} \to \hat{u}$ implies strong convergence of $\hat{u}_{\varepsilon_n} \to \hat{u}$ for every subsequence $\hat{u}_{\varepsilon_n}$. Also, for any $\bar{u}$ steering x_0 to x_1, we have

$$J(\bar{u}) \geqq \limsup_{\varepsilon} J_\varepsilon(\hat{u}_\varepsilon) = J(\hat{u}).$$

Hence $\hat{u}$ solves (QCCP).

From the fact that $J(\hat{u}_\varepsilon) \to J(\hat{u})$, we conclude

$$\frac{1}{\varepsilon}|x(T; \hat{u}_\varepsilon) - x_1|^2 \to 0,$$

so (2.4) follows. □

Theorem 2.2 formulates an equivalent condition for controllability depending on initial and terminal conditions only. In the next theorem we will formulate an equivalent condition for global controllability.

We first note [5] that the system (LC) is controllable from any initial state x_0 to an arbitrarily prescribed terminal state x_1 if and only if the system (LO)

$$\text{(LO)} \qquad \begin{aligned} \frac{dy}{dt} &= -A^*(t)y(t) \\ w(t) &= B(t)^*y(t) \end{aligned}$$

is observable. (LO) is observable if and only if

$$Z_N(T) \equiv \int_0^T \Phi(T, s)B(s)N(s)^{-1}B^*(s)\Phi^*(T, s)\, ds \tag{2.7}$$

is a positive definite matrix. For any x_0, x_1, the control

$$\hat{u}(t) = N(t)^{-1}B^*(t)\Phi^*(T, t)Z_N^{-1}[x_1 - \Phi(T)x_0 - F(T)], \qquad 0 \leqq t \leqq T \tag{2.8}$$

steers (LC) from x_0 to x_1 such that $\langle Nu, u\rangle_{L^2_m(0,T)}$ is minimal. This exact solution will be used for comparison in § 3, Example 1.

THEOREM 2.3 (Global uniform bounds). *The system* (LC) *is controllable from* x_0 *to* x_1 *for arbitrarily given* x_0, x_1 *if and only if there exists a positive constant* M *independent of* x_0, x_1 *such that*

$$J_\varepsilon(x_0, \hat{u}_\varepsilon) \leqq M|x_0 - \Phi(T)x_1 - F(T)|^2 + 4\|C\|^2\|\Phi(\cdot)x_0 + F(\cdot)\|^2_{L^2_n(0,T)} + 2\|z\|^2_{L^2_p(0,T)},$$

for all $\varepsilon > 0$.

Proof. By (2.8), the control

$$\bar{u}(t) \equiv N(t)^{-1}B^*(t)\Phi^*(T, t)Z_N^{-1}[x_1 - \Phi(T)x_0 - F(T)]$$

steers (LC) from x_0 to x_1. Therefore

$$J(\bar{u}) \geqq J_\varepsilon(\hat{u}_\varepsilon) \quad \text{for } \varepsilon \in \mathbb{R}^+.$$

But

$$\begin{aligned} J(x, \bar{u}) &= \int_0^T (|C(t)x(t; \bar{u}) - z(t)|^2 + \langle N(t)\bar{u}(t), \bar{u}(t)\rangle)\, dt \\ &\leqq \int_0^T [2|C(t)x(t; \bar{u})|^2 + 2|z(t)|^2 + \langle N(t)\bar{u}(t), \bar{u}(t)\rangle]\, dt. \end{aligned}$$

It is clear that

$$\langle Nu, u\rangle \leqq K_1 |x_0 - \Phi(T)x_0 - F(T)|^2$$

with

$$K_1 \equiv \left\| \int_0^T Z_N^{-1} * \Phi(T, t)B(t)B^*(t)\Phi(T, t)^* Z_N^{-1}\, dt \right\|.$$

Also, from (1.2),

$$\begin{aligned} 2\int_0^T |C(t)x(t, \bar{u})|^2\, dt &\leqq 4\left\{ \int_0^T \left[|C(t)(\Phi(t)x_0 + F(t))|^2 \right.\right. \\ &\quad + \left| C(t) \int_0^t \Phi(t, s)B(s)N^{-1}(s)B^*(s)\Phi^*(t, s) \right. \\ &\quad \left.\left. \cdot Z_N^{-1}(x_1 - \Phi(T)x_0 - F(T))|^2\, ds \right] dt \right\} \\ &\leqq 4\|C\|^2 \|\Phi(\cdot)x_0 + F(\cdot)\|^2_{L_n^2(0,T)} + K_2 |x_0 - \Phi(T)x_0 - F(T)|^2, \end{aligned}$$

where K_2 can be similarly defined.

Let $M \equiv K_1 + K_2$, and the proof is complete. □

The proofs of Theorems 2.1 and 2.2 do not indicate the *rate* of convergence of $\hat{u}_\varepsilon \to \hat{u}$. Indeed, such a rate of convergence is unknown to us if (LC) is not controllable from x_0 to x_1 as in the general case of Theorem 2.1. For Theorem 2.2, however, the following theorem gives the sharpest possible estimates.

THEOREM 2.4. *Assume the system* (LC) *is controllable for some given* x_0, $x_1 \in \mathbb{R}^n$. *Let* $\hat{u}$ *be the optimal control solving* (QCCP) *and let* $\hat{u}_\varepsilon$ *be the control obtained from the penalization* (PQCCP). *Then*

$$\|\hat{u}_\varepsilon - \hat{u}\|_{L_m^2(0,T)} \leqq K_6 |x(T; \hat{u}_\varepsilon) - x_1| \tag{2.9}$$

for some K_6 *depending on* f, z *only. Consequently,*

$$|x(T; \hat{u}_\varepsilon) - x_1| = O(\varepsilon) \quad \text{as } \varepsilon \downarrow 0, \tag{2.10}$$

$$\|\hat{u}_\varepsilon - \hat{u}\|_{L_m^2(0,T)} = O(\varepsilon) \quad \text{as } \varepsilon \downarrow 0, \tag{2.11}$$

$$\sup_{[0,T]} |x(t; \hat{u}_\varepsilon) - x(t; \hat{u})| = O(\varepsilon) \quad \text{as } \varepsilon \downarrow 0. \tag{2.12}$$

Proof. For *any* x_0, $x_1 \in \mathbb{R}^n$, f in (LC) and z in (2.2), the mapping

$$S_\varepsilon : \mathbb{R}^n \times \mathbb{R}^n \times L_n^2(0, T) \times L_l^2(0, T) \to L_m^2(0, T),$$

$$S_\varepsilon(x_0, x_1, f, z) \equiv \hat{u}_\varepsilon, \qquad \hat{u}_\varepsilon \text{ solves (PQCCP)}$$

is linear and continuous. This can be verified from the corresponding variational equation. Since

$$\lim_{\varepsilon \downarrow 0} S_\varepsilon(x_0, x_1, f, z) = \hat{u}, \tag{2.13}$$

$\hat{u}$ is the strong limit of $\hat{u}_\varepsilon$ in Theorem 2.1. By the Banach–Steinhaus theorem, we have

$$S_\varepsilon \to S \text{ pointwise}, \qquad S(x_0, x_1, f, z) \equiv \hat{u},$$

and S is a bounded linear operator from $\mathbb{R}^n \times \mathbb{R}^n \times L_n^2(0, T) \times L_l^2(0, T)$ into $L_m^2(0, T)$.

On the other hand, assuming (LC) is controllable from x_0 to x_1 for some given (x_0, x_1), for any $\varepsilon > 0$, we claim that

$$S(x_0, x(T; \hat{u}_\varepsilon), f, z) = \hat{u}_\varepsilon; \tag{2.14}$$

i.e., the optimal control steering x_0 to $x(T; \hat{u}_\varepsilon)$ minimizing (2.2) is $\hat{u}_\varepsilon$. This can be seen as follows. Let $\hat{v}(y)$ be the solution to

$$\text{(P1)} \qquad \begin{aligned} &\inf_{v \in L_m^2(0,T)} \left[J(x_0, v) + \frac{1}{\varepsilon} |y - x_1|^2 \right], \qquad y \in \mathbb{R}^n \text{ is given,} \\ &v \text{ steers (LC) from } x_0 \text{ to } y. \end{aligned}$$

Then $\hat{v}(y)$ also solves

$$\text{(P2)} \qquad \begin{aligned} &\inf J(x_0, v), \\ &v \text{ steers (LC) from } x_0 \text{ to } y, \end{aligned}$$

because in (P1) there is *no variation* in $(1/\varepsilon)|y - x_1|^2$. Hence $\hat{v}(y) = S(x_0, y, f, z)$. Choosing $y \equiv x(T; \hat{u}_\varepsilon)$, from (P2) we easily see that (2.14) holds.

Therefore, by the linear continuity of S,

$$\begin{aligned} \|\hat{u}_\varepsilon - \hat{u}\|_{L_m^2(0,T)} &= \|S(x_0, x(T; \hat{u}_\varepsilon), f, z) - S(x_0, x_1, f, z)\| \\ &\leqq K_6 |x(T; \hat{u}_\varepsilon) - x_1| \quad \text{for some } K_6 > 0 \text{ depending on } f, z \text{ only.} \end{aligned}$$

So (2.9) is proven. From (2.5), using an identity due to Hager [10], we have

$$\begin{aligned} \frac{1}{\varepsilon} |x(T; \hat{u}_\varepsilon) - x_1|^2 &\leqq J_\varepsilon(\hat{u}) - J(\hat{u}_\varepsilon) \\ &= J(\hat{u}) - J(\hat{u}_\varepsilon) \\ &= J'(\hat{u}_\varepsilon) \cdot (\hat{u} - \hat{u}_\varepsilon) + a(\hat{u} - \hat{u}_\varepsilon, \hat{u} - \hat{u}_\varepsilon) \\ &= 2[a(\hat{u}_\varepsilon, \hat{u} - \hat{u}_\varepsilon) - \theta(\hat{u} - \hat{u}_\varepsilon)] + a(\hat{u} - \hat{u}_\varepsilon, \hat{u} - \hat{u}_\varepsilon). \end{aligned} \tag{2.15}$$

Refer to (1.4), (1.5) for a, θ derived from the J of (2.2). The three terms on the right of (2.15) grow with an order of magnitude $\|\hat{u} - \hat{u}_\varepsilon\|$ (i.e., first order) when $\|\hat{u} - \hat{u}_\varepsilon\|$ is small and of magnitude $\|\hat{u} - \hat{u}_\varepsilon\|^2$ (i.e., quadratic order) when $\|\hat{u} - \hat{u}_\varepsilon\|$ is large. But it has been proven that $\|\hat{u} - \hat{u}_\varepsilon\| \to 0$. So, for ε sufficiently small, and K_7 independent of ε,

$$\begin{aligned} \frac{1}{\varepsilon} |x(T; \hat{u}_\varepsilon) - x_1|^2 &\leqq K_7 \|\hat{u} - \hat{u}_\varepsilon\| \\ &\leqq K_6 K_7 |x(T; \hat{u}_\varepsilon) - x_1| \qquad \text{(cf. (2.9)).} \end{aligned}$$

Hence

$$|x(T; \hat{u}_\varepsilon) - x_1| \leqq K_6 K_7 \varepsilon \quad \text{for } \varepsilon \text{ sufficiently small.}$$

So (2.10) is proven. (2.11) follows from (2.9) and (2.10). (2.12) follows from (1.2) and (2.11). □

Remarks. (1) In the paper by B. T. Polyak [17] the convergence rates (2.10), (2.11) have been obtained for a more general nonlinear (equality) constraint. Thus Theorem 2.4 becomes a special case of [17]. Because our proof will be used later in Theorem 2.7, we present it in a complete fashion as above. (2) By Theorem 1.2, the (PQCCP) solution u is characterized by the variational equation

$$a_\varepsilon(\hat{u}_\varepsilon, v) = \theta_\varepsilon(v) \quad \text{for all } v \in L^2_m, \tag{2.16}$$

with a_ε and θ_ε as in (1.4) and (1.5) and $\gamma = 1/\varepsilon$. From the Lax–Milgram theorem we have

$$a_\varepsilon(\hat{u}_\varepsilon, v) = \langle T_\varepsilon \hat{u}_\varepsilon, v\rangle_{L^2_m} = \theta_\varepsilon(v) = \langle g_\varepsilon, v\rangle_{L^2_m} \quad \forall v \in L^2_m,$$

where T_ε is a (symmetric, positive) invertible Fredholm integral operator and g_ε depends linearly on (x_0, x_1, f, z). Thus

$$\begin{aligned} &T_\varepsilon \hat{u}_\varepsilon = g_\varepsilon, \\ &\hat{u}_\varepsilon = T_\varepsilon^{-1} g_\varepsilon = S_\varepsilon(x_0, x_1, f, z). \end{aligned} \tag{2.17}$$

In the case that the system (LC) is globally controllable, we have the following asymptotic expansion which is an even stronger result than Theorem 2.4.

THEOREM 2.5 (Asymptotic expansion). *Assume that the system* (LC) *is controllable from x_0 to x_1 for arbitrarily given x_0, x_1. Let $\hat{u}_\varepsilon$ be the optimal control obtained from* (PQCCP). *Then we have the following asymptotic expansion*

$$\hat{u}_\varepsilon = \hat{u} + \varepsilon u_1 + \varepsilon^2 u_2 + \cdots + \varepsilon^n u_n + \cdots, \qquad \varepsilon > 0,$$

where $\hat{u}$ is the optimal control solving (QCCP).

Proof. We let $\mathscr{L}_1$ be defined as in (1.1) and define $\mathscr{L}_2 : L^2_m(0, T) \to \mathbb{R}^n$ by

$$\mathscr{L}_2 u = \int_0^T \Phi(T, s) B(s) u(s)\, ds.$$

Then one has

$$\mathscr{L}_1^* : L^2_n(0, T) \to L^2_m(0, T), \qquad \mathscr{L}_2^* : \mathbb{R}^n \to L^2_m(0, T),$$

$$(\mathscr{L}_1^* x)(t) = B^*(t) \int_t^T \Phi^*(s, t) x(s)\, ds, \qquad (\mathscr{L}_2^* z)(t) = B^*(t) \Phi^*(T, t) z.$$

From (2.17), we see that $\hat{u}_\varepsilon$ is the solution of

$$(N + \mathscr{L}_1^* \mathscr{L}_1)\hat{u}_\varepsilon + \frac{1}{\varepsilon} \mathscr{L}_2^* [x(T; \hat{u}_\varepsilon) - x_1] = \mathscr{L}_1^* h, \tag{2.18}$$

where

$$h(t) \equiv C^*(t)\{z_1(t) - C(t)[\Phi(t)x_0 + F(t)]\}.$$

We let

$$p_0 \equiv \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} [x(T; \hat{u}_\varepsilon) - x_1]. \tag{2.19}$$

This limit can be easily shown to exist because of (2.10). It is determinable from one of the transversality conditions [6].

Let $u_0 = \lim_{\varepsilon \downarrow 0} \hat{u}_\varepsilon$ as guaranteed by Theorem 2.2. Letting $\varepsilon \downarrow 0$ in (2.18), we get

$$(N + \mathscr{L}_1^* \mathscr{L}_1) u_0 + \mathscr{L}_2^* p_0 = \mathscr{L}_1^* h. \tag{2.20}$$

Furthermore, u_0 is the unique solution to (2.20) because $N + \mathscr{L}_1^* \mathscr{L}_1$ has a bounded inverse.

Next, let (u_1, p_1) solve

$$(N + \mathscr{L}_1^* \mathscr{L}_1) u_1 + \mathscr{L}_2^* p_1 = 0,$$

$$\mathscr{L}_2 u_1 = p_0.$$

This system has a unique pair of solutions

$$\begin{aligned} p_1 &= -[\mathscr{L}_2 (N + \mathscr{L}_1^* \mathscr{L}_1)^{-1} \mathscr{L}_2^*]^{-1} p_0, \\ u_1 &= -(N + \mathscr{L}_1^* \mathscr{L}_1)^{-1} \mathscr{L}_2^* p_1. \end{aligned} \tag{2.21}$$

Note here that the bracketed term in (2.21) has an inverse because the $Z_N(T)$ in (2.7) is invertible, i.e., because of global controllability.

We proceed inductively as follows: let (u_j, p_j) solve

$$\begin{aligned} &(N + \mathscr{L}_1^* \mathscr{L}_1) u_j + \mathscr{L}_2^* p_j = 0, \\ &\mathscr{L}_2 u_j = p_{j-1}, \end{aligned} \qquad j \geqq 1, \tag{2.22}$$

and let

$$\hat{u}_\varepsilon^N \equiv u_0 + \varepsilon u_1 + \cdots + \varepsilon^N u_N.$$

Multiplying (2.22) by ε^j and summing over j from 1 to N, we get

$$(N + \mathscr{L}_1^* \mathscr{L}_1) \hat{u}_\varepsilon^N + \mathscr{L}_2^* (p_0 + \varepsilon p_1 + \cdots + \varepsilon^{N-1} p_{N-1}) = \mathscr{L}_1^* h - \varepsilon^N \mathscr{L}_2^* p_N.$$

Using $p_j = \mathscr{L}_2 u_{j+1}$ from the second relation in (2.22), we get

$$(N + \mathscr{L}_1^* \mathscr{L}_1) \hat{u}_\varepsilon^N + \mathscr{L}_2^* \mathscr{L}_2 (u_1 + \varepsilon u_2 + \cdots + \varepsilon^{N-1} u_N) = \mathscr{L}_1^* h - \varepsilon^N \mathscr{L}_2^* p_N.$$

Equivalently, since $x(T; u_0) - x_1 = 0$ (u_0 solves (QCCP)),

$$(N + \mathscr{L}_1^* \mathscr{L}_1) \hat{u}_\varepsilon^N + \frac{1}{\varepsilon} \mathscr{L}_2^* [(x(T; u_0) - x_1) + \varepsilon \mathscr{L}_2 (u_1 + \varepsilon u_2 + \cdots + \varepsilon^{N-1} u_N)] = \mathscr{L}_1^* h - \varepsilon^N L_2^* p_N.$$

Subtracting (2.18) from the above, and using $x(T; u_0) - x(T; \hat{u}_\varepsilon) = \mathscr{L}_2 (u_0 - \hat{u}_\varepsilon)$, we obtain

$$(N + \mathscr{L}_1^* \mathscr{L}_1)(\hat{u}_\varepsilon^N - \hat{u}_\varepsilon) + \frac{1}{\varepsilon} \mathscr{L}_2^* \mathscr{L}_2 (\hat{u}_\varepsilon^N - \hat{u}_\varepsilon) = -\varepsilon^N \mathscr{L}_2^* p_N.$$

Forming the inner product of the above with $\hat{u}_\varepsilon^N - \hat{u}_\varepsilon$, we get

$$\begin{aligned} &\langle N(\hat{u}_\varepsilon^N - \hat{u}_\varepsilon), \hat{u}_\varepsilon^N - \hat{u}_\varepsilon \rangle + \|\mathscr{L}_1 (\hat{u}_\varepsilon^N - \hat{u}_\varepsilon)\|^2 + \frac{1}{\varepsilon} \|\mathscr{L}_2 (\hat{u}_\varepsilon^N - \hat{u}_\varepsilon)\|^2 \\ &\quad = -\varepsilon^N \langle p_N, \mathscr{L}_2 (\hat{u}_\varepsilon^N - \hat{u}_\varepsilon) \rangle \\ &\quad \leqq \frac{1}{2\varepsilon} \|\mathscr{L}_2 (\hat{u}_\varepsilon^N - \hat{u}_\varepsilon)\|^2 + \frac{\varepsilon^{2N+1}}{2} \|p_N\|^2, \end{aligned}$$

implying

$$\varepsilon^{-(2N+1)} \{[\langle N(\hat{u}_\varepsilon^N - \hat{u}_\varepsilon), \hat{u}_\varepsilon^N - \hat{u}_\varepsilon \rangle + \|\mathscr{L}_1 (\hat{u}_\varepsilon^N - \hat{u}_\varepsilon)\|^2] + 2^{-1} \varepsilon \|\mathscr{L}_2 (\hat{u}_\varepsilon^N - \hat{u}_\varepsilon)\|^2\} \leqq \tfrac{1}{2} \|p_N\|^2.$$

Therefore, for each N,

$$\varepsilon^{-N}\|\hat{u}_\varepsilon^N - \hat{u}_\varepsilon\| \leqq \left[\frac{1}{\nu}\varepsilon^{-2N}\langle N(\hat{u}_\varepsilon^N - \hat{u}_\varepsilon), \hat{u}_\varepsilon^N - \hat{u}_\varepsilon\rangle\right]^{1/2} \leqq \nu^{-1/2}\varepsilon^{1/2}2^{-1/2}\|p_N\| \to 0 \quad \text{as } \varepsilon \downarrow 0$$

proving the validity of the asymptotic expansion.

Remark. The procedures in the above expansion are standard; cf. [20, pp. 152–154], for example. The distinction between [20] and the work here seems to be that the operator in the bracketed term in (2.21) is not invertible in the nonglobally controllable case. It is not clear whether the asymptotic expansion is still possible in that case.

Now, we assume that $A(t)$, $B(t)$, $N^{-1}(t)$, $C(t)$, $z(t)$ and $f(t)$ are sufficiently smooth functions so that the regularity Theorem 1.3 can be applied to give positive s_1 and s_2 which are sufficiently large. We may now apply the finite element method (1.7) to approximate (PQCCP)'s variational equivalent (2.16).

THEOREM 2.6. *Let $\hat{S}^h$ be an $(r, 0)$-system. Assume that s: $0 < s \leqq \min(s_1, s_2)$ for some positive s_1, s_2 as mentioned above. Let $\hat{u}_\varepsilon \in H_m^s(0, T)$ be the solution to the penalty problem* (PQCCP) *and $\hat{u}$ be the L_m^2-limit of $\hat{u}_\varepsilon$ as in Theorem* 2.1. *Let $\hat{u}_{\varepsilon,h} \in \hat{S}^h$ be the finite element approximation* (1.7) *of the penalized problem: i.e.,*

$$J_\varepsilon(x_0, \hat{u}_\varepsilon) = \inf_{v \in L_m^2} J_\varepsilon(x_0, v), \qquad \hat{u} = \lim_{\varepsilon \downarrow 0} \hat{u}_\varepsilon$$

and

$$J_\varepsilon(x_0, \hat{u}_{\varepsilon,h}) = \inf_{v \in \hat{S}^n} J_\varepsilon(x_0, v).$$

Then there exist K_8, $K_9 > 0$ independent of ε, h such that for $\mu = \min(r, s)$,

$$\|\hat{u}_{\varepsilon,h} - \hat{u}\|_{L_m^2} \leqq \sigma_1(\varepsilon) + \frac{K_8}{(\nu\varepsilon)^{1/2}} h^\mu \|\hat{u}_\varepsilon\|_{H_m^s(0,T)}, \tag{2.23}$$

$$\sup_{[0,T]} |x(t; \hat{u}_{\varepsilon,h}) - x(t; \hat{u})| \leqq \sigma_2(\varepsilon) + \frac{K_9}{(\nu\varepsilon)^{1/2}} h^\mu \|\hat{u}_\varepsilon\|_{H_m^s(0,T)}, \tag{2.24}$$

where $\sigma_1(\varepsilon) \downarrow 0$, $\sigma_2(\varepsilon) \downarrow 0$ as $\varepsilon \downarrow 0$ depending on x_0, x_1.

Proof. By standard procedures we find

$$a_\varepsilon(\hat{u}_{\varepsilon,h} - \hat{u}_\varepsilon, v_h) = 0 \quad \forall v_h \in \hat{S}^h.$$

Hence, by Theorem 1.4, using $\gamma = 1/\varepsilon$, we get

$$\|\hat{u}_{\varepsilon,h} - \hat{u}_\varepsilon\| \leqq \frac{K_8}{(\nu\varepsilon)^{1/2}} h^\mu \|\hat{u}_\varepsilon\|_{H_m^s}.$$

Therefore, from Theorem 2.1 and

$$\|\hat{u}_{\varepsilon,h} - \hat{u}\| \leqq \|\hat{u}_{\varepsilon,h} - \hat{u}_\varepsilon\| + \|\hat{u}_\varepsilon - \hat{u}\|,$$

(2.23) follows. (2.24) is immediate from (2.23). □

If (LC) is controllable from x_0 to x_1, we have the following sharp asymptotic estimates:

THEOREM 2.7. *Assume that for given (x_0, x_1),* (LC) *is controllable from x_0 to x_1. Then the control $\hat{u}$ in Theorem* 2.4 *also belongs to $H_m^s(0, T)$ and we have, for some K_{10} independent of ε,*

$$\|\hat{u}_\varepsilon - \hat{u}\|_{H_m^s(0,T)} \leqq K_{10}\varepsilon, \tag{2.25}$$

and there exist constants $K_{11}, K_{12} > 0$ *independent of* ε, h *such that for* $\mu = \min(r, s)$

$$\|\hat{u}_{\varepsilon,h} - \hat{u}\|_{L^2_m(0,T)} \leqq K_{11}(\varepsilon + h^{\mu}\|\hat{u}\|_{H^s_m(0,T)}),$$

$$\sup_{[0,T]} |x^{(j)}(t; \hat{u}_{\varepsilon,h}) - x^{(j)}(t; \hat{u})| \leqq K_{12}(\varepsilon + h^{\mu}\|\hat{u}\|_{H^s_m(0,T)}), \qquad 0 \leqq j \leqq s, \tag{2.26}$$

for ε *sufficiently small.*

Proof. In order to consider (2.25), we first return to the proof of Theorem 1.3. As in that proof, we differentiate both sides of (1.9) k times, where k is an integer greater than s, keeping in mind that

$$\begin{aligned} &p(T) = p_{\varepsilon}(T) = \frac{1}{\varepsilon}[x(T; \hat{u}_{\varepsilon}) - x_1], \\ &C_1(t) \equiv C(t), \\ &z_1(t) \equiv z(t), \end{aligned} \tag{2.27}$$

for the current case. Since the k-times differentiated ordinary differential equation is still well-posed, the solution $(x_{\varepsilon}, p_{\varepsilon})$ is continuous with respect to initial and boundary data. From the estimate (2.10), we see that $p_{\varepsilon}(T)$ is continuous with respect to $\varepsilon \geqq 0$, and from (2.19)

$$\lim_{\varepsilon \downarrow 0} p_{\varepsilon}(T) = \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon}[x(T; \hat{u}_{\varepsilon}) - x_1] = p_0 = p(T; \hat{u}).$$

Therefore $(x_{\varepsilon}(t), p_{\varepsilon}(t))$ converges in the H^s-norm to $(x(t; \hat{u}), p(t; \hat{u}))$. Therefore (2.25) follows from (2.27) and (2.10).

Next, let $W = \{x(T; v) \in \mathbb{R}^n \mid v \in L^2_m(0, T)\}$, an affine subspace of $\mathbb{R}^n$. Note $x_1 \in W$. Since $\mathbb{R}^n$ is finite dimensional, $W = \{x(T, v_h) \in \mathbb{R}^n \mid v_h \in \hat{S}^h\}$ for all h sufficiently small. Define

$$S_{\varepsilon,h} : \mathbb{R}^n \times \mathbb{R}^n \times L^2_n(0, T) \times L^2_l(0, T) \to \hat{S}^h$$

by $S_{\varepsilon,h}(x_0, x_1, f, z) = \hat{u}_{\varepsilon,h}$ which solves the finite element approximation (PQCCP) over $\hat{S}^h$. Following the proof of Theorem 2.4, $S_{\varepsilon,h}$ is linear, continuous, and

$$\lim_{\varepsilon \downarrow 0} S_{\varepsilon,h}(x_0, x_1, f, z) = \hat{u}_h \equiv S_h(x_0, x_1, f, z)$$

is a bounded linear operator on $\mathbb{R}^n \times \mathbb{R}^n \times L^2_n \times L^2_l$. Also, $x_1 \in W$ implies $\hat{u}_h$ solves (QCCP) over $\hat{S}^h$ $(x(T; \hat{u}_h) = x_1)$ and $S_h(x_0, x(T; \hat{u}_{\varepsilon,h}), f, z) = \hat{u}_{\varepsilon,h}$. Hence

$$\|\hat{u}_{\varepsilon,h} - \hat{u}_h\| \leqq \|S_h\| \cdot |x(T; \hat{u}_{\varepsilon,h}) - x_1|, \tag{2.28}$$

and the same previous argument gives

$$|x(T; \hat{u}_{\varepsilon,h}) - x_1| \leqq K \cdot \|S_h\| \cdot \varepsilon.$$

Next, let $h(t)$ and p_0 be as defined in (2.19). Letting $\varepsilon \downarrow 0$ in (2.16), we get

$$a(\hat{u}, v) + \bar{\theta}_1(v) = \bar{\theta}_2(v), \tag{2.29}$$

where $a(\cdot, \cdot)$ is as defined in (1.4) with $\gamma = 0$, and

$$\bar{\theta}_1(v) \equiv -\langle h, \mathscr{L}_1 v \rangle, \qquad \bar{\theta}_2 \equiv -\langle p_0, \mathscr{L}_2 v \rangle.$$

Let $\tilde{n}_h$ be the unique element in $\hat{S}^h$ satisfying

$$a(\hat{u} - \tilde{u}_h, v_h) = 0 \quad \forall v_h \in \hat{S}^h. \tag{2.30}$$

Then $\tilde{u}_h$ has the property [15, p. 51] that

$$\|\hat{u} - \tilde{u}_h\| \leqq Ch^{\mu} \|\hat{u}\|_{H_m^s} \tag{2.31}$$

and hence

$$\|x(T; \hat{u}) - x(T; \tilde{u}_h)\| = \|\mathscr{L}_2(\hat{u} - \tilde{u}_h)\| \leqq Ch^{\mu} \|\hat{u}\|_{H_m^s}. \tag{2.32}$$

From (2.29), (2.30), we conclude that $\tilde{u}_h$ satisfies the variational equation

$$a(\tilde{u}_h, v_h) + \bar{\theta}_1(v_h) = \bar{\theta}_2(v_h) \quad \forall v_h \in \hat{S}^h$$

in $\hat{S}^h$. Thus $a(\tilde{u}_h, v_h) + \bar{\theta}_1(v_h) = 0$ for all v_h satisfying $\mathscr{L}_2 v_h = 0$. But this means that $J'(\tilde{u}_h) \cdot v_h = 0$ for all v_h satisfying $\mathscr{L}_2 v_h = 0$; i.e., $J(\tilde{u}_h + v_h) \geqq J(\tilde{u}_h)$ for all $\tilde{u}_h + v_h$ such that

$$\Phi(T)x_0 + F(T) + \mathscr{L}_2(\tilde{u}_h + v_h) = \Phi(T)x_0 + F(T) + \mathscr{L}_2 \tilde{u}_h = x(T; \tilde{u}_h).$$

In other words, we have $S_h(x_0, x(T; \tilde{u}_h), f, z) = \tilde{u}_h$. Therefore

$$\begin{aligned} \|\hat{u}_h - \tilde{u}_h\| &= \|S_h(x_0, x_1, f, z) - S_h(x_0, x(T; \tilde{u}_h), f, z)\| \\ &\leqq \|S_h\| \, \|x_1 - x(T; \tilde{u}_h)\| \\ &\leqq Ch^{\mu} \|S_h\| \, \|\hat{u}\|_{H_m^s} \end{aligned}$$

and

$$\|\hat{u} - \hat{u}_h\| \leqq \|\hat{u} - \tilde{u}_h\| + \|\tilde{u}_h - \hat{u}_h\| \leqq Ch^{\mu} \|\hat{u}\|_{H_m^s}. \tag{2.33}$$

Therefore, if S is the operator in the proof of Theorem 2.4, $S_h \to S$ pointwise as $h \to 0$. By the uniform boundedness principle, $\|S_h\| \leqq M$ independent of h. Combining this fact with (2.28) gives

$$\|\hat{u}_{\varepsilon,h} - \hat{u}_h\| \leqq K\varepsilon. \tag{2.34}$$

Combining (2.31) and (2.34) gives the conclusion of the theorem. □

Remarks. (1) Numerical experiments completely agree with the error estimates in Theorem 2.6. Existing examples of exact solutions of $\hat{u}_\varepsilon$ and $\hat{u}$ indicate that Theorem 2.6 gives the best possible estimates. (2) If one were to solve (PQCCP) by Riccati feedback synthesis, one would have the decoupled Riccati equations

$$\begin{aligned} &\mathbb{P}_\varepsilon' + \mathbb{P}_\varepsilon A + A^* \mathbb{P}_\varepsilon - \mathbb{P}_\varepsilon B N^{-1} B^* \mathbb{P}_\varepsilon = -C^* C \quad \text{on } [0, T], \\ &\mathbb{P}_\varepsilon(T) = \frac{1}{\varepsilon} I_{n \times n}, \end{aligned} \tag{2.35}$$

$$\begin{aligned} &r_\varepsilon' + A^* r_\varepsilon - \mathbb{P}_\varepsilon B N^{-1} B^* r_\varepsilon = -\mathbb{P}_\varepsilon f + C^* z, \\ &r_\varepsilon(T) = -\frac{1}{\varepsilon} x_1. \end{aligned} \tag{2.36}$$

Both are very ill-conditioned because of the terminal conditions.

3. Examples. We present three examples to illustrate the theory in § 2. Naturally, they are also examples for § 1. We let $\hat{S}^h \equiv \Pi S^h$ where S^h are continuous piecewise cubics, a (4, 0)-system. We choose $h = T/10$ and $\varepsilon = 10^{-4}$.

1. *Comparison with an exact solution.* Here we solve a 2×2 system ($n=m=2$) with only control cost minimized:

$$J(x_0, u)=\int_0^{\pi/2} |u(t)|^2\,dt,$$

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

$$x(0)=\begin{bmatrix}-1\\ 1\end{bmatrix},\quad x\left(\frac{\pi}{2}\right)=\begin{bmatrix}1\\ -1\end{bmatrix},\quad C=0,\quad z=f=0,$$

on the interval $[0, \pi/2]$. This has the exact (QCCP) solution (by (2.8))

$$u(t)=\frac{4}{\pi}\begin{bmatrix}\cos t\\ \sin t\end{bmatrix},\qquad x(t)=\begin{bmatrix}\left(\frac{4}{\pi}t-1\right)\cos t+\sin t\\ -\left(\frac{4}{\pi}t-1\right)\sin t+\cos t\end{bmatrix}.$$

The finite element solutions are given in Fig. 1. All solutions were found to be correct pointwise with a relative error of 10^{-4}.

2. *Example of an uncontrollable problem.* We look at a 2×1 system ($n=2$, $m=1$) which is not controllable.

$$A=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix},\quad B\begin{bmatrix}1\\ 0\end{bmatrix},\quad f=\begin{bmatrix}\cos 3t\\ \sin 5t\end{bmatrix},$$

$$x(0)=\begin{bmatrix}1\\ -1\end{bmatrix},\quad x(1)=\begin{bmatrix}-1\\ \frac{1}{2}\end{bmatrix},\quad N=I_{2\times 2},$$

$$C=I_{2\times 2},\quad z=\begin{bmatrix}\cos 4t\\ \sin 3t\end{bmatrix}\quad \text{on } [0, 1].$$

The finite element solutions are given in Fig. 2. We see that the second component of x does not attain its target of $x_1=\frac{1}{2}$. The finite element problem is a well-posed approximation to $\hat{u}_\varepsilon$ and $\hat{u}_\varepsilon$ does converge to some $\hat{u}\in L^2(0, 1)$, but $\hat{u}$ is not a solution to (QCCP). Refer to Theorems 2.1 and 2.6.

3. *Example of a larger problem with total cost minimized.* Here we solve the 4×3 system ($n=4$, $m=3$) on $[0, 2]$:

$$A=\begin{bmatrix}0 & 1 & 0 & 0\\ -1 & 0 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 2\end{bmatrix},\qquad B=\begin{bmatrix}t & e^{t^2} & t^2\\ 0 & t^2 & 0\\ 2t & 0 & t\\ t^2 & cos\ 3t & 4t\end{bmatrix},$$

$$x(0)=\begin{bmatrix}1\\ -1\\ 0\\ -2\end{bmatrix},\qquad x(T)=\begin{bmatrix}-1\\ \frac{1}{2}\\ -2\\ 1\end{bmatrix},$$

$$C=\begin{bmatrix}t^4 & \sin t & 4t^2-1 & 3e^{-t}\\ t^3 & t^2 & t & 1\end{bmatrix},\qquad N=\begin{bmatrix}3+t & \frac{1}{2} & 2\\ \frac{1}{2} & 4+t^2 & 0\\ 0 & 0 & 3+t^3\end{bmatrix},$$

$$f=\begin{bmatrix}\cos 2t\\ \sin 5t\\ e^t\\ \sin 10t\end{bmatrix},\qquad z=\begin{bmatrix}\cos 4t\\ \sin 3t\end{bmatrix}.$$

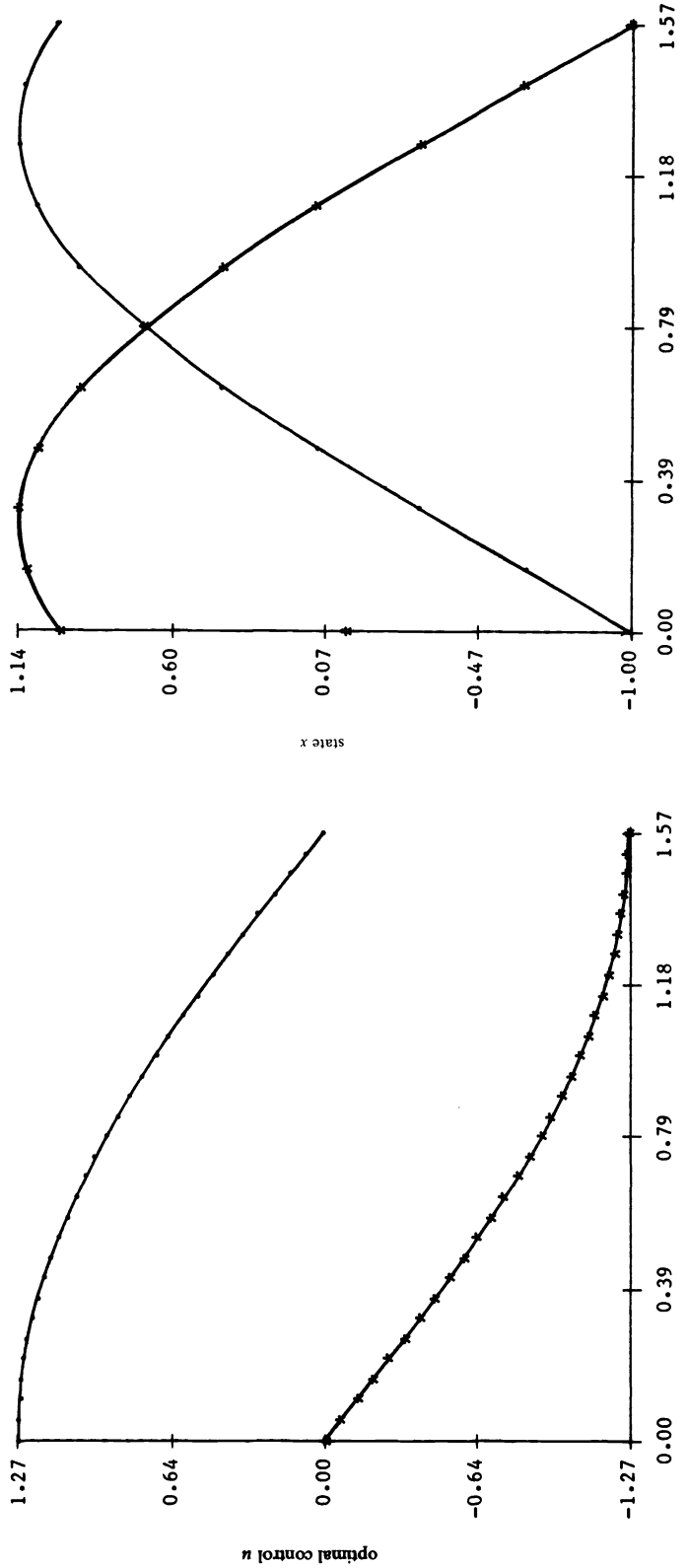


FIG. 1. *The 2 by 2 system: Example 1 solved by variational penalization using finite elements with* $S(4, 0)$ *cubic splines,* $\varepsilon = 0.0001$ *and* 10 *subintervals on* $(0, 1.5708)$.

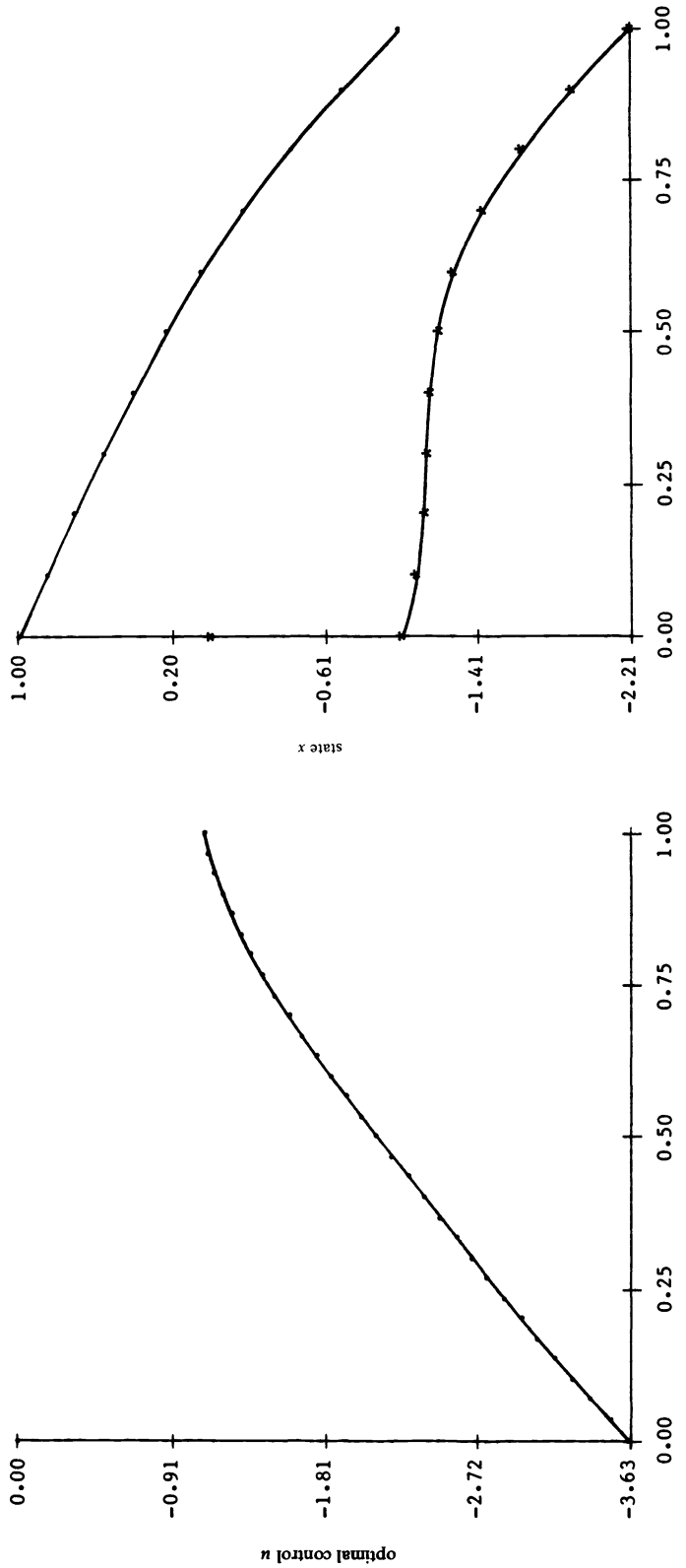
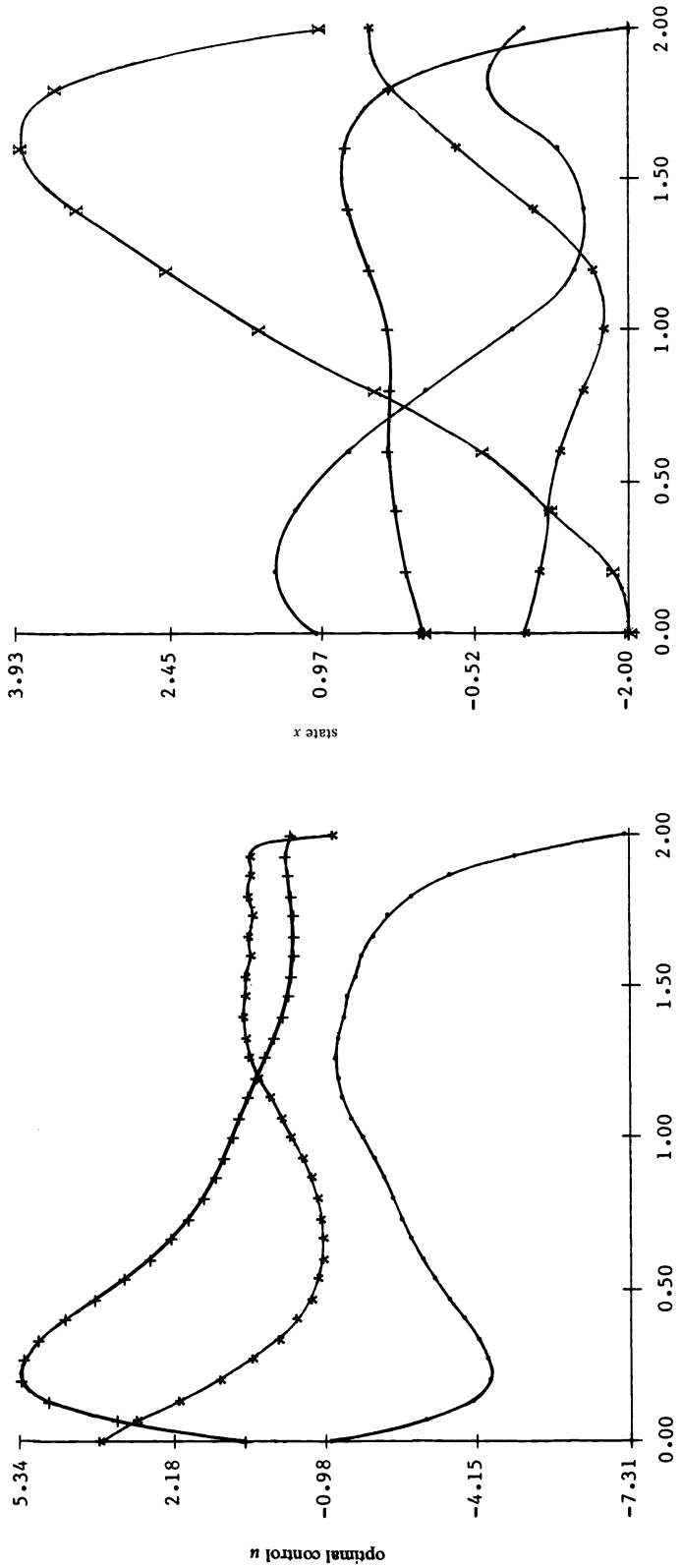


FIG. 2. *The 2 by 1 system: Example 2 solved by variational penalization using finite elements with* $S(4, 0)$ *cubic splines,* $\varepsilon = 0.0001$ *and* 10 *subintervals on* $(0, 1.000)$.



a. Optimal control $u(t)$ vs. t. $u_1 \cdots$; $u_2 ***$; $u_3 + + +$.

b. Displacement $x(u, t)$ vs. t. $x_1 \cdots$; $x_2 ***$; $x_3 + + +$; x_4 XXX.

FIG. 3. *The* 4 *by* 3 *system: Example* 3 *solved by variational penalization using finite elements with* $S(4, 0)$ *cubic splines,* $\varepsilon = 0.0001$ *and* 10 *subintervals on* $(0, 2.0000)$.

The finite element solutions are given in Fig. 3. Since all components of x attained the target, we conclude the problem is controllable with the optimal control as given.

From Fig. 3 one can also find that the control component u_2 is less smooth near the terminal time $t = T$. In keeping close to the distributed goal function $z(t)$, the trajectory must also attain the designated terminal state which may not be close to the terminal state of this goal function.

Remarks. (1) The program generating the above solutions uses standard Gaussian quadrature for $\int_0^t \Phi(t, s)B(s)\hat{\varphi}(s)\,ds$ calculation at equally spaced points t, and a high order Newton-Cotes formula for

$$\int_0^T \left\langle C(t)\int_0^t \Phi(t, s)B(s)\hat{\varphi}_i(s)\,ds,\, C(t)\int_0^t \Phi(t, s)B(s)\hat{\varphi}_j(s)\,ds \right\rangle dt$$

calculation. These additional computational errors are built into the finite element solution and do not decrease the asymptotic errors of Theorem 2.7 provided high order quadrature is used. The reader is referred to [1, p. 525] for a discussion of such matters.

(2) For each $\varepsilon > 0$, let $K_{\varepsilon,h}$ be the matrix K_h in (1.8). In practice, we find that the condition number of $K_{\varepsilon,h}$ to be $O(1/\varepsilon)$. For small ε this numerical instability will produce loss of accuracy in the solution. This loss may be completely recovered by *iterative refinement* procedures. The fact that we have only $O(1/\varepsilon)$ conditioning indicates that our penalty method produces minimized instabilities.

Acknowledgment. We thank Professor T. I. Seidman of the University of Maryland–Baltimore County for a discussion which motivated the proof of Theorem 2.4. We also wish to thank the referees for their suggestions, and particularly for the idea of sharpening the error estimates of Theorem 2.7.

REFERENCES

[1] I. BABUSKA AND A. K. AZIZ, in The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, A. K. Aziz, ed., Academic Press, New York, 1972, pp. 5–359.

[2] R. E. BELLMAN, *Introduction to the Mathematical Theory of Control Processes*, Vol. II, Academic Press, New York and London, 1971.

[3] W. E. BOSARGE AND O. G. JOHNSON, *Error bounds of high order accuracy for the state regulator problem via piecewise polynomial approximation*, SIAM J. Control, 9 (1971), pp. 15–28.

[4] R. W. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.

[5] G. CHEN AND W. H. MILLS, *Penalization and regularization of quadratic cost optimal control problems in a finite dimensional space*, Research Report #10, INRIA, Rocquencourt, France, March, 1980.

[6] S. J. CITRON, *Element of Optimal Control*, Holt, Reinhart and Winston, New York, 1969.

[7] J. W. DANIEL, *The Ritz-Galerkin method for abstract optimal control problems*, this Journal, 11 (1973), pp. 53–63.

[8] W. W. HAGER, *Rates of convergence for discrete approximations to unconstrained control problems*, SIAM J. Numer. Anal., 13 (1976), pp. 449–472.

[9] ———, *The Ritz-Trefftz method for state and control constrained optimal control problems*, SIAM J. Numer. Anal., 12 (1975), pp. 854–867.

[10] ———, *Inequalities and approximations*, Research Report 78–15, Dept. of Math., Carnegie-Mellon Univ., Pittsburgh, PA, 1978.

[11] W. W. HAGER AND S. K. MITTER, *Lagrange duality for convex control problems*, this Journal, 14 (1976), pp. 843–856.

[12] R. E. KALMAN, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana, 5 (1960), pp. 102–119.

[13] R. E. KALMAN AND R. S. BUCY, *New results in linear prediction and filtering theory*, J. Basic Eng., Trans. ASME Ser. D, 83 (1961), pp. 95–100.

[14] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, English translation, Springer-Verlag, New York, 1971.

[15] ———, *Topics in numerical analysis*, in Lecture Notes on the Finite Element Methods, Southeast Asia Mathematical Society, Penang, Malaysia, 1978.
[16] F. H. MATHIS AND G. W. REDDIEN, *Ritz-Trefftz approximations in optimal control*, this Journal 17 (1979), pp. 307–310.
[17] B. T. POLYAK, *The convergence rate of the penalty function method*, Zh. vychisl Mat. mat. fiz., 11 (1971), pp. 3–11.
[18] D. L. RUSSELL, *Quadratic performance criteria in boundary control of linear symmetric hyperbolic systems*, SIAM J. Control, 11 (1973), pp. 475–509.
[19] ———, *Mathematics of Finite Dimensional Control Systems*, Marcel-Dekker, New York, 1979.
[20] R. TEMAM, *Navier–Stokes Equations*, North Holland, Amsterdam, 1977.
[21] A. WIERZBICKI AND S. KURCYUSZ, *Projection on a cone, penalty functionals, and duality theory for optimal control problems*, this Journal, 15 (1977), pp. 25–56.
[22] I. LASIECKA, *Boundary control of parabolic systems; finite element approximation*, Appl. Math. Optim., 6 (1980), pp. 31–62.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 6, November 1981

0363-0129/81/1906-0004 $01.00/0

EXACT CONTROLLABILITY THEOREMS AND NUMERICAL SIMULATIONS FOR SOME NONLINEAR DIFFERENTIAL EQUATIONS*

GOONG CHEN†, WENDELL H. MILLS, JR.‡, AND GIOVANNI CROSTA§

Abstract. We study exact controllability problems for some nonlinear systems with linear controls. Our tools are contraction fixed point theorems and nonlinear semigroup properties. We show that under the assumptions of low order nonlinearity, reversibility and the existence of certain feedback controls, the nonlinear system is exactly controllable. The constructive aspect of the theory allows the application of numerical simulation. An analog-digital realization diagram is discussed. Accurate numerical schemes are developed and error estimates are presented with concrete examples to illustrate the theory.

Introduction. In this paper, we are concerned with the controllability problem of a nonlinear system

$$\text{(NCS)} \qquad \begin{aligned} &\frac{d}{dt}x(t,u) = A(x(t)) + Bu(t), \qquad t \geqq 0, \\ &x(0) = x_0 \in X, \end{aligned}$$

where

(0.1)
X = a Banach space with norm $\|\cdot\|$,
$x(t) \in X$ is the state of the system at time t,
$u \in U_{\text{ad}}$ is an admissible control, U_{ad} is the space of admissible controls,
$A: D(A) \subseteq X \to X$ is a single-valued nonlinear operator with dense domain $D(A)$, and A is the infinitesimal generator of a nonlinear semigroup,
B is a bounded linear operator from some space U into X.

The controllability problem for the system (NCS) is, for given $x_0, x_1 \in X$, to find a $u \in U_{\text{ad}}$ such that starting from x_0, the system is steered to x_1 at some $t = t_1$ (t_1 may be dependent upon x_0, x_1), i.e.,

$$x(t_1, u) = x_1.$$

The system (NCS) we are studying is autonomous, linear in the control variable and nonlinear only in the state variable. The most general form of a nonlinear control system would appear to be

$$\frac{d}{dt}x(t,u) = F(x(t), u(t), t), \tag{0.2}$$

where the defining relation F is nonautonomous and nonlinear in both state and control variables. In case (0.2) is a system governed by a nonlinear ODE, the controllability problem has been well studied by engineers and mathematicians; see, e.g., [2], [7], [10], [15], [16], [21], to mention a few. In the case of nonlinear PDE's, there is relatively less literature on this subject. We refer to [6], [8], [11], [12].

* Received by the editors October 29, 1979, and in revised form September 15, 1980.
† Department of Mathematics, Pennsylvania State University, University Park, Pennsylvania 16802.
The work of this author was supported in part by the National Science Foundation under grant MCS 7822830.
‡ Department of Mathematics, Pennsylvania State University, University Park, Pennsylvania 16802.
§ Via Dupré 5, 21013 Gallarate, Italy.

In this paper we make a study of the nonlinear controllability problem by a new approach. The basic tools we will use are some fixed point theorems for nonlinear mappings and nonlinear semigroup theory. If the system (NCS) is "reversible" and has some "good" feedback controllers, we can use a Lyapunov-type stability argument to obtain exact controllability. This is a unified theory for systems governed by nonlinear ordinary and partial differential equations. Basically, our paper is a generalization of D. L. Russell's "controllability via stabilizability" theorem [20] to nonlinear systems. Our theorems (except Theorem 2.3) are constructive. A realization block diagram (Fig. 1) for analog-digital simulation is discussed in § 1.2. Section 3 is devoted to the development of accurate numerical simulation schemes which are quite important and an integral part of this study. Examples are presented at the end of the paper.

The controls we obtain in § 2 are not unique. They depend on the choice of feedback controls and the injectivity of the operator B. A penalization technique as in [5] might be applied to achieve a unique optimal control.

1. Linear controllability via stabilizability and realization. In [20], Russell's controllability via stabilizability theorem was formulated for an autonomous ODE system with the terminal state to be controlled to 0. The fact that his proof in [20] can be immediately generalized to an infinite dimensional system is quite obvious, but here we present a slightly improved argument with an arbitrary terminal state. This refinement will be seen (§ 2) to be necessary for the controllability study of nonlinear evolution control systems. We also include a block diagram interpreting the realization of the mathematical argument. This diagram, with minor modification, also serves as realization for the proof of Theorem 2.1.

1.1. Linear controllability via stabilizability theorem of Russell.

THEOREM 1.1. *Let*

$$\text{(LCS)} \qquad \begin{aligned} &\frac{d}{dt}x(t) = Ax(t) + Bu(t), \qquad 0 \leqq t \leqq T, \\ &x(0) = x_0 \in X \end{aligned}$$

be a linear control system in X, where

$A: D(A) \to X$ *is the infinitesimal generator of a linear semigroup with dense domain* $D(A)$,

$B: U_{\text{ad}} \to X$ *is linear and bounded.*

If there exist bounded linear operators $K^+, K^-: X \to U_{\text{ad}}$ such that

$$A^+ \equiv A + BK^+, \qquad A^- \equiv A + BK^-$$

generate semigroups e^{A^+t}, e^{-A^-t} satisfying

$$\|e^{A^+t}\| \leqq Ce^{-kt}, \qquad t \geqq 0, \quad C, k > 0, \tag{1.1}$$

$$\|e^{-A^-t}\| \leqq Ce^{-kt}, \qquad t \geqq 0, \quad C, k > 0, \tag{1.2}$$

then for T sufficiently large we have the following: for each $x_0, x_1 \in X$, there exists a control $u \in C([0, T]; U)$ such that $x(t)$ is steered from x_0 to x_1 at $t = T$.

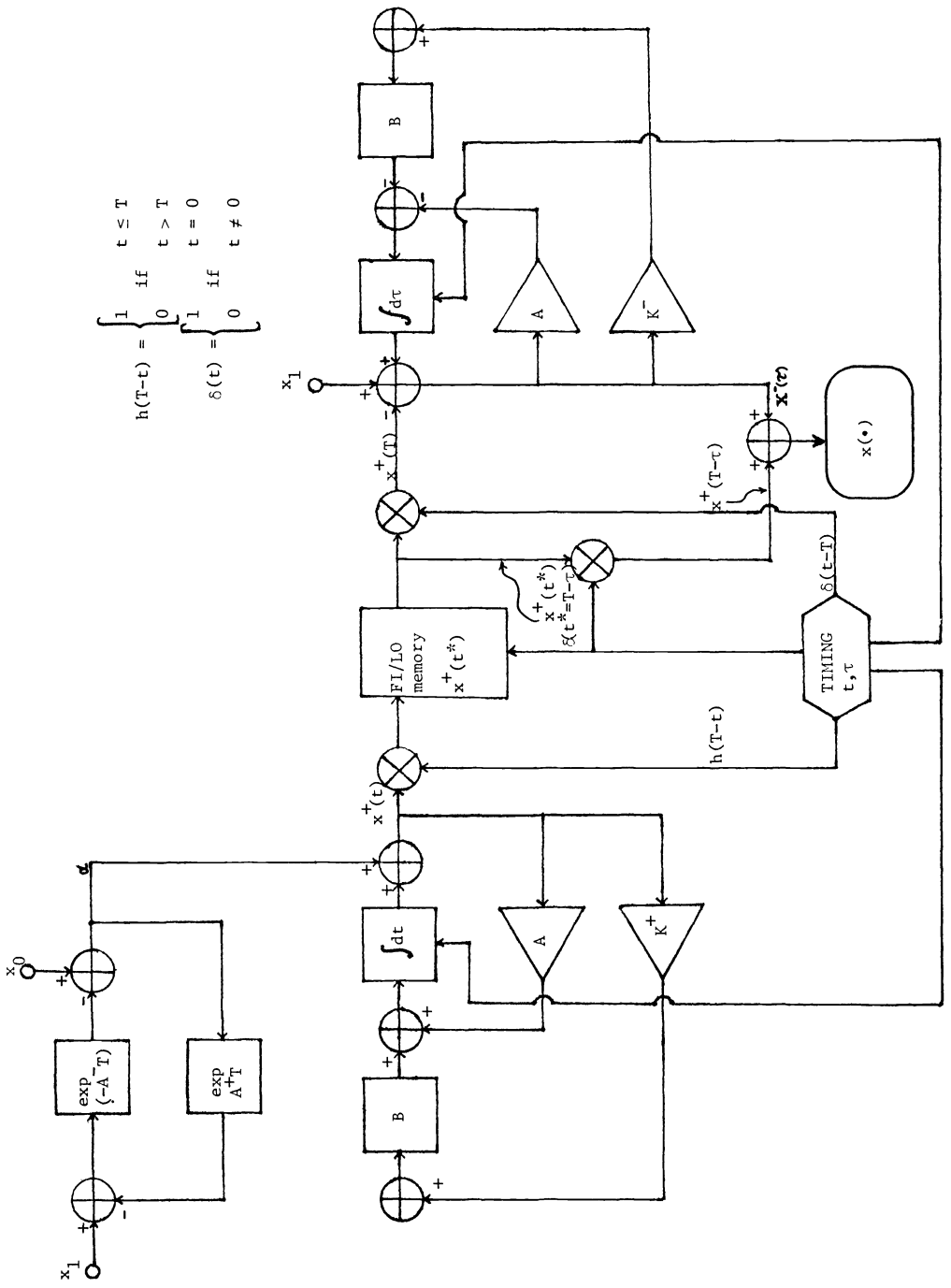

h(T-t) = 1 if t ≤ T, 0 if t > T
δ(t) = 1 if t = 0, 0 if t ≠ 0
x0
x1
exp (-A⁻T)
exp A⁺T
B
∫dt
A
K⁺
x⁺(t)
FI/LO memory
x⁺(t*)
x⁺(t*) δ(t*=T-τ)
x⁺(T)
x⁺(T-τ)
x(τ)
x(•)
TIMING t,τ
δ(t-T)
h(T-t)
B
∫dτ
A
K⁻

FIG. 1

Proof. Let $x^+(t)$ be the (generalized) solution of

$$\begin{aligned} &\frac{d}{dt}x^+(t) = Ax^+(t) + BK^+x^+(t) \equiv A^+x^+(t), \qquad 0 \leqq t \leqq T, \\ &x^+(0) = \alpha \in X, \end{aligned} \tag{1.3}$$

and let $x^-(t)$ be the solution of

$$\begin{aligned} &\frac{d}{dt}x^-(t) = Ax^-(t) + BK^-x^-(t) \equiv A^-x^-(t), \qquad 0 \leqq t \leqq T, \\ &x^-(T) = x_1 - x^+(T). \end{aligned} \tag{1.4}$$

Note that (1.4) is a backward equation. Then $x(t) \equiv x^+(t) + x^-(t)$ satisfies

$$\begin{aligned} &\frac{d}{dt}x(t) = Ax(t) + Bu(t), \qquad 0 \leqq t \leqq T, \quad \text{with } u \equiv K^+x^+ + K^-x^-, \\ &x(T) = x_1 \end{aligned} \tag{1.5}$$

with the initial condition

$$\begin{aligned} x(0) = x^+(0) + x^-(0) &= \alpha + \{e^{A^-(t-T)}[x_1 - x^+(T)]\}_{t=0} \\ &= \alpha + e^{-A^-T}[x_1 - e^{A^+T}\alpha] \qquad (\because x^+(T) = e^{A^+T}\alpha) \\ &= (I - e^{-A^-T}e^{A^+T})\alpha + e^{-A^-T}x_1. \end{aligned} \tag{1.6}$$

Therefore every initial state $x(0)$ of the form (1.6) can be steered to (1.5). Now, choose T large enough. By (1.1), (1.2)

$$\|e^{-A^-T}e^{A^+T}\| < 1.$$

So (1.6) is always solvable for any given $x(0) = x_0$, with solution

$$\alpha = (I - e^{-A^-T}e^{A^+T})^{-1}(x_0 - e^{-A^-T}x_1). \qquad \square$$

Remarks. (1) Theorem 1.1 remains valid if (LCS) contains an inhomogeneous forcing term $f(t)$:

$$\frac{d}{dt}x(t) = Ax(t) + Bu(t) + f(t).$$

See [5].

(2) For examples of controllability via stabilizability in an infinite dimensional space, we refer to [19], [3], [4].

1.2. Analog-digital simulations of Theorem 1.1. The realization of Theorem 1.1 is shown in Fig. 1. The only inputs are x_0 and x_1. α is computed via (1.6) and fed into the main analog-digital integrator circuit. System (1.3) with $x^+(0) = \alpha$ as initial condition must run first for T time units in order to yield $x^+(T)$. The data from system (1.3) are then stacked in an FI/LO (First-In-Last-Out) memory. At time $t = T$ system (1.4) starts running. A change of variable $T - t = \tau$ is needed to let τ run from 0 to T. Thus the backward equation (1.4) becomes a forward equation

$$\begin{aligned} &\frac{d}{d\tau}x^-(\tau) = -A^-x^-(\tau), \qquad 0 \leqq \tau \leqq T, \\ &x^-(0) = x_1 - e^{A^+T}\alpha. \end{aligned} \tag{1.4$'$}$$

As system (1.4)′ is running, the data stored in the memory are read out from the top and summed at each step with corresponding values of x^+. Thus we get a sequence of values of x starting from $x(T)$. It takes $2T$ time units for the whole process to be complete.

Due to the memory stack and time delay required, the circuit must be run on a hybrid computer to obtain the state $x(t)$. The individual flows x^+ and x^- may be read off to compute the control u. We note that it is the *linearity* of the operator A which enables the exact computation of α. In the nonlinear case of § 2, Fig. 1 must be further coupled with a digital diagram for iterative computation of α, $x(t)$ and $u(t)$. The nature of this coupling will be evident and will not appear here. In lieu of such analog-digital simulation we construct a fully digital numerical scheme in § 3.

2. Exact controllability for nonlinear systems. A nonlinear semigroup on a Banach space X is a function S with domain $\mathbb{R}^+ \times X$ and range in X satisfying

$$S(t_1, S(t_2, x)) = S(t_1 + t_2, x) \quad \forall x \in X, \quad t_1, t_2 \geqq 0, \tag{2.1}$$

$$S(0, x) = x, \tag{2.2}$$

$$\text{for every } x \in X,\ S(t, x) \text{ is continuous in } t \geqq 0, \tag{2.3}$$

$$S(t, x) \text{ is continuous in } x \text{ for every } t \geqq 0. \tag{2.4}$$

We also allow a nonlinear semigroup S to be defined on $\mathbb{R}^- \times X$ satisfying (2.1)–(2.4) with $t \leqq 0$.

For an autonomous nonlinear ordinary differential equation

$$\frac{d}{dt}x(t) = f(x(t)), \qquad t \geqq 0 \text{ or } t \leqq 0,$$

$$x(0) = x_0 \in \mathbb{R}^n,$$

$$f: \mathbb{R}^n \to \mathbb{R}^n \text{ is continuous,}$$

with global existence and uniqueness on $[0, \infty)$ or $(-\infty, 0]$, f is known to generate a nonlinear semigroup $S^+(t, x)$ on $\mathbb{R}^+ \times \mathbb{R}^n$ or $S^-(t, x)$ on $\mathbb{R}^- \times \mathbb{R}^n$ defined by

$$S^+(t, x_0) \equiv x(t), \quad t \geqq 0, \qquad S^-(t, x_0) \equiv x(t), \quad t \leqq 0,$$

or

$$x(0) = x_0, \qquad x(0) = x_0.$$

In the nonlinear evolutionary case

$$\frac{d}{dt}x(t) = F(x(t)), \qquad t \geqq 0,$$

$$x(0) = x_0 \in X, \quad X \text{ infinite dimensional,}$$

$$F: D(F) \subseteq X \to X \quad \text{a single-valued nonlinear operator with dense domain,}$$

it becomes much more difficult to prove the existence of such nonlinear semigroups generated by the Banach space-valued function F. Indeed, for many nonlinear evolution equations we only know [14] the existence and uniqueness of generalized solutions $x(t)$ which are in $L^\infty(0, T; X)$, i.e., the solution $x(t)$ may not be defined pointwise in t even after any modification on a set of measure zero in $[0, T]$. For such nonlinear equations, the study of exact controllability is impossible (or, must be defined in a different sense) since the state of the system at time t is not well defined. In the theorems

that follow we will require the existence of these semigroups for the F's under consideration.

Actually, the nonlinear operator F may be allowed to be multiple-valued. We refer the readers to [1], [9] for generation theorems for such maximal dissipative nonlinear operators.

We now consider the problem (NCS) with $U_{ad} \equiv L^2(0, \infty; U)$ or $U_{ad} \equiv L^\infty(0, \infty; U)$. The following assumptions require that the nonlinear operator A has only low order nonlinearity.

(H1) The nonlinear mapping $\mathcal{N}: D(A) \times D(A) \to X$ defined by

$$\mathcal{N}(v_1, v_2) = A(v_1 + v_2) - (Av_1 + Av_2)$$

can be extended uniquely into a continuous mapping on $X \times X$; i.e., for any $v_1, v_2 \in X$ and any sequences $\{v_1^{(m)}\}, \{v_2^{(n)}\} \subseteq D(A)$ such that $v_1^{(m)} \to v_1$, $v_2^{(n)} \to v_2$ in X, there exists a unique $z \in X$ such that

$$\lim_{\substack{m\to\infty \\ n\to\infty}} [A(v_1^{(m)} + v_2^{(n)}) - (Av_1^{(m)} + Av_2^{(n)})] = z \equiv \mathcal{N}(v_1, v_2).$$

The assumption (H1) is trivial if X is finite dimensional.

We also need

(H2) For any $v_1, v_2 \in X$, there exists a w (may be nonunique) $\in U$ such that

$$\mathcal{N}(v_1, v_2) = Bw$$

and w can be chosen to be continuously dependent upon v_1, v_2.

2.1. Exact controllability (I): global contractions at fixed time. The following theorem is a straightforward generalization of Theorem 1.1.

THEOREM 2.1. *Let the operators A and B of* (NCS) *satisfy* (H1) *and* (H2). *Assume that there exists $T > 0$ and two bounded linear operators $K^+, K^-: X \to U$ such that*

(i) *$A + BK^+$ generates a nonlinear semigroup Φ on $\mathbb{R}^+ \times X$ such that $\Phi(T, \cdot)$ is a strict contraction in X;*

(ii) *$A + BK^-$ generates a nonlinear semigroup Ψ on $\mathbb{R}^- \times X$ such that $\Psi(-T, \cdot)$ is a strict contraction in X.*

Then for each $x_0, x_1 \in X$, there exists a control $u \in U_{ad} \cap C([0, T]; U)$ such that $x(t)$, $u(t)$ satisfy (NCS) *and $x(t)$ is steered to x_1 at $t = T$. If B is injective, then such u is unique.*

Remark. The above is still a *fixed time* (T) controllability theorem.

Proof. It goes almost the same as that of Theorem 1.1, except that the linear semigroups e^{A^+t}, e^{-A^-t} are replaced by the nonlinear semigroups $\Phi(t, \cdot)$ and $\Psi(-t, \cdot)$.

Let $x^+(t)$ be the solution of

$$\frac{d}{dt}x^+(t) = A(x^+(t)) + BK^+x^+(t) \qquad (\equiv (A + BK^+)(x^+(t))), \qquad 0 \leqq t \leqq T,$$

$$x^+(0) = \alpha \in X.$$

Then

$$x^+(t) = \Phi(t, \alpha), \qquad 0 \leqq t \leqq T.$$

Let $x^-(t)$ be the solution of

$$\frac{d}{dt}x^-(t) = A(x^-(t)) + BK^-x^-(t) \qquad (\equiv (A + BK^-)(x^-(t))), \qquad 0 \leqq t \leqq T,$$

$$x^-(T) = x_1 - \Phi(T, \alpha).$$

Then

$$x^-(t) = \Psi(t-T, x^-(T)), \qquad 0 \leqq t \leqq T.$$

Therefore $x(t) \equiv x^+(t) + x^-(t)$ satisfies

$$\frac{d}{dt}x(t) = [A(x^+(t)) + BK^+x^+(t)] + [A(x^-(t)) + BK^-x^-(t)]$$

$$= A(x(t)) + Bu(t),$$

where

$$u(t) \equiv K^+x^+(t) + K^-x^-(t) - w(t),$$

with $w(t)$ satisfying

$$Bw(t) = \mathcal{N}(x^+(t), x^-(t)) \quad \text{by (H1), (H2),}$$

and the terminal condition for x is

$$x(T) = x^+(T) + x^-(T) = x_1. \tag{2.5}$$

The initial condition is

$$x(0) = x^+(0) + x^-(0) = \alpha + \Psi(-T, x_1 - \Phi(T, \alpha)). \tag{2.6}$$

So every initial state of the form (2.6) can be steered to x_1 at $t = T$. Note that if B is injective, w and hence u is unique. We want to show that initial states of the form (2.6) consist of all X.

For any $x_0 \in X$, consider the nonlinear mapping

$$\mathcal{T}: X \to X, \qquad \mathcal{T}(v) \equiv x_0 - \Psi(-T, x_1 - \Phi(T, v)).$$

We get

$$\begin{aligned}
&\|\mathcal{T}(v_1) - \mathcal{T}(v_2)\| \\
&\quad = \|\Psi(-T, x_1 - \Phi(T, v_1)) - \Psi(-T, x_1 - \Phi(T, v_2))\|, \\
&\quad \leqq c_1 \|[x_1 - \Phi(T, v_1)] - [x_1 - \Phi(T, v_2)]\| \qquad 0 \leqq c_1 < 1 \quad \text{(by assumption (ii))} \\
&\quad \leqq c_2 c_1 \|v_1 - v_2\|, \qquad 0 \leqq c_2 < 1 \quad \text{(by assumption (i)).}
\end{aligned}$$

So $\mathcal{T}$ is a strict contraction in X. By Banach's contraction mapping theorem, $\mathcal{T}$ has a unique fixed point α, $\alpha = \mathcal{T}(\alpha)$. Thus

$$x_0 = \alpha + \Psi(-T, x_1 - \Phi(T, \alpha)) \tag{2.7}$$

is always solvable. So u steers the nonlinear system (NCS) from x_0 to x_1 at $t = T$. From the way u was chosen, we easily see that $u \in C([0, T]; U)$. □

Example 1.

$$my''(t) + ky(t) + f(y(t)) + L = u(t),$$

where $m, k > 0$, L is a real constant, and

$$f: \mathbb{R}^1 \to \mathbb{R}^1 \text{ continuously differentiable such that } |f'(t)| \leqq c, \text{ with } c < k. \tag{2.8}$$

It can be written as a system

$$\frac{d}{dt}\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} x_2(t) \\ -\frac{k}{m}x_1(t) - \frac{1}{m}f(x_1(t)) - \frac{L}{m} \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{m} \end{bmatrix} u(t)$$

$$\equiv A\left(\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}\right) + Bu(t),$$

with

$$A\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} x_2 \\ -\frac{1}{m}(kx_1 + f(x_1) + L) \end{bmatrix}, \qquad B = \begin{bmatrix} 0 \\ \frac{1}{m} \end{bmatrix}.$$

We choose $K^+ = [0, -m\gamma]$ and $K^- = [0, m\gamma]$. We want to show that $A + BK^+$ and $A + BK^-$ generate semigroups satisfying the assumptions of Theorem 2.1.

Consider the semigroup generated by $A + BK^+$. Let $(x_{11}(t), x_{21}(t))$ and $(x_{12}(t), x_{22}(t))$ be the solutions of

$$\frac{d}{dt}\begin{bmatrix} x_{1i}(t) \\ x_{2i}(t) \end{bmatrix} = (A + BK^+)\begin{bmatrix} x_{1i}(t) \\ x_{2i}(t) \end{bmatrix} = \begin{bmatrix} x_{2i}(t) \\ -\frac{1}{m}(kx_{1i}(t) + \gamma x_{2i}(t) + f(x_{1i}(t)) + L) \end{bmatrix}, \qquad i = 1, 2$$

with initial conditions

$$\begin{bmatrix} a_1 \\ b_1 \end{bmatrix}, \qquad \begin{bmatrix} a_2 \\ b_2 \end{bmatrix}$$

respectively. We want to show that there exists $T > 0$ such that

$$\left\| \begin{bmatrix} x_{11}(T) - x_{12}(T) \\ x_{21}(T) - x_{22}(T) \end{bmatrix} \right\| \leqq c_1 \left\| \begin{bmatrix} a_1 - a_2 \\ b_1 - b_2 \end{bmatrix} \right\|, \quad \text{for some } c_1,\ 0 \leqq c_1 < 1 \quad \text{independent of } a_i, b_i.$$

Define $z(t) \equiv x_{11}(t) - x_{12}(t)$. Then $z'(t) = x_{21}(t) - x_{22}(t)$ and $z(t)$ satisfies

$$\begin{gathered} mz''(t) + \gamma z'(t) + kz(t) + [f(x_{11}(t)) - f(x_{12}(t))] = 0, \\ \begin{bmatrix} z(0) \\ z'(0) \end{bmatrix} = \begin{bmatrix} a_1 - a_2 \\ b_1 - b_2 \end{bmatrix}. \end{gathered} \tag{2.9}$$

Using $z'(t) + \lambda z(t)$ $(\lambda > 0)$ as multiplier to (2.9), we obtain

$$\begin{aligned} &\frac{d}{dt}\frac{1}{2}\{mz'^2 + (\lambda\gamma + k)z^2 + 2\lambda m z' z\} \\ &\quad + \{[\gamma - \lambda m]z'^2 + \lambda k z^2 + [z' + \lambda z][f(x_{11}(t)) - f(x_{12}(t))]\} = 0. \end{aligned} \tag{2.10}$$

But $f(x_{11}(t)) - f(x_{12}(t)) = f'(\theta(t))z(t)$, $\theta(t)$ lying between $x_{11}(t)$ and $x_{12}(t)$, so (2.10) can be written as

$$\frac{d}{dt}P(t) + Q(t) = 0,$$

with

$$\begin{aligned} P(t) &\equiv \tfrac{1}{2}\{mz'^2 + (\lambda\gamma + k)z^2 + 2\lambda m z' z\}, \\ Q(t) &\equiv [\gamma - \lambda m]z'^2 + \lambda[k + f'(\theta(t))]z^2 + f'(\theta(t))z' z. \end{aligned}$$

Now, we choose λ, γ such that

$$\gamma - \lambda m \geqq c, \quad \lambda(k-c) \geqq c, \quad \lambda\gamma + k \geqq 4\lambda^2 m.$$

Then for some constants $D_1, D_2 > 0$ (independent of (a_1, b_1), (a_2, b_2)) we have

$$D_1\{z^2(t) + [z'(t)]^2\} \geqq P(t) \geqq D_2\{z^2(t) + [z'(t)]^2\},$$
$$D_1\{z^2(t) + [z'(t)]^2\} \geqq Q(t) \geqq D_2\{z^2(t) + [z'(t)]^2\}.$$

Hence

$$0 = \frac{d}{dt}P(t) + Q(t) \geqq \frac{d}{dt}P(t) + \frac{D_2}{D_1}P(t),$$

$$D_2\{z^2(t) + [z'(t)]^2\} \leqq P(t) \leqq \exp\left(-\frac{D_2}{D_1}t\right)P(0)$$
$$\leqq D_1 \exp\left(-\frac{D_2}{D_1}t\right)\{z^2(0) + [z'(0)]^2\}.$$

Thus

$$\{z^2(t) + [z'(t)]^2\} \leqq \frac{D_1}{D_2}\exp\left(-\frac{D_2}{D_1}t\right)\{z^2(0) + [z'(0)]^2\}.$$

Choosing T large enough such that

$$c_1 \equiv \frac{D_1}{D_2}\exp\left(-\frac{D_2}{D_1}T\right) < 1,$$

we are done.

By reversing the sense of time $t \to T - t$, the proof above also works for $A + BK^-$.

One easily verifies that (H2) is satisfied. (H1) is trivial. So Example 1 is exactly controllable provided that T is large enough. □

Example 2. A PDE version of Example 1 is the nonlinear wave equation

$$\frac{\partial^2 w(x,t)}{\partial t^2} - \Delta w(x,t) + f(w(x,t)) = u(x,t), \qquad x \in \Omega, \quad t \geqq 0, \tag{2.11}$$

where $f: \mathbb{R} \to \mathbb{R}$ satisfies

(i) $f(w(\cdot)) \in L^2(\Omega)$ for all $w \in L^2(\Omega)$,

(ii) $\int |f(w_1) - f(w_2)|^2\, dx \leqq k^2 \int |w_1 - w_2|^2\, dx$, for all $w_1, w_2 \in L^2(\Omega)$, for some $k < \mu_1$, where μ_1 is the first eigenvalue of $(-\Delta)$.

The state space is $X \equiv H_0^1(\Omega) \oplus H^0(\Omega)$ with Ω a bounded domain in $\mathbb{R}^n$ with regular boundary. (2.11) can be written as a system

$$\frac{d}{dt}\begin{bmatrix} w(\cdot,t) \\ v(\cdot,t) \end{bmatrix} = \begin{bmatrix} v(\cdot,t) \\ \Delta w(\cdot,t) - f(w(\cdot,t)) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(\cdot,t)$$
$$\equiv A\left(\begin{bmatrix} w(\cdot,t) \\ v(\cdot,t) \end{bmatrix}\right) + Bu(\cdot,t),$$
$$D(A) = [H^2(\Omega) \cap H_0^1(\Omega)] \oplus H_0^1(\Omega).$$

We choose $K^+ = [0, -\gamma]$ and $K^- = [0, \gamma]$ for some $\gamma > 0$.

The fact that $A + BK^+$ and $A + BK^-$ generate nonlinear semigroups follows from [18, Thm. 4.4.2], since the mapping induced by f is globally Lipschitzian in $L^2(\Omega)$. The

semigroup property (2.1) follows from the uniqueness of solutions and the autonomy of $A+BK^+$ and $A+BK^-$.

The rest of the discussion is similar to that of Example 1. We use $\partial z/\partial t+\lambda z$ as multiplier: let $z(x,t)\equiv w_1^+(x,t)-w_2^+(x,t)$, where w_1^+ and w_2^+ are solutions of

$$\frac{\partial^2 w(x,t)}{\partial t^2}-\Delta w(x,t)+\gamma\frac{\partial w(x,t)}{\partial t}+f(w(x,t))=0$$

with initial states (φ_1,ψ_1) and (φ_2,ψ_2), respectively. Then z satisfies

$$\frac{\partial^2 z}{\partial t^2}-\Delta z+\gamma\frac{\partial z}{\partial t}+[f(w_1)-f(w_2)]=0.$$

Multiplying the above by $\partial z/\partial t+\lambda z$, integrating by parts and simplifying, we obtain

$$\frac{d}{dt}P(t)+Q(t)=0,$$

where

$$P(t)\equiv\frac{1}{2}\int_\Omega\left[|\nabla z|^2+\left(\frac{\partial z}{\partial t}\right)^2+\lambda\gamma z^2+2\lambda z\frac{\partial z}{\partial t}\right]dz,$$

$$Q(t)\equiv\int_\Omega\left\{(\gamma-\lambda)\left(\frac{\partial z}{\partial t}\right)^2+\lambda|\nabla z|^2+[f(w_1)-f(w_2)]\left(\frac{\partial z}{\partial t}+\lambda z\right)\right\}dx.$$

Then, since,

$$\left|\int[f(w_1)-f(w_2)]\frac{\partial z}{\partial t}\,dx\right|\leqq\frac{\varepsilon k^2}{2}\int z^2\,dx+\frac{1}{2\varepsilon}\int\left(\frac{\partial z}{\partial t}\right)^2dx$$

and

$$\left|\int[f(w_1)-f(w_2)]\lambda z\,dx\right|\leqq k\lambda\int z^2\,dx,$$

by choosing ε, γ and λ appropriately and using Poincaré's inequality we again find two positive constants D_1, D_2 such that

$$D_1\int\left[|\nabla z|^2+\left(\frac{\partial z}{\partial t}\right)^2\right]dx\geqq P(t)\geqq D_2\int\left[|\nabla z|^2+\left(\frac{\partial z}{\partial t}\right)^2\right]dx,$$

$$D_1\int\left[|\nabla z|^2+\left(\frac{\partial z}{\partial t}\right)^2\right]dx\geqq Q(t)\geqq D_2\int\left[|\nabla z|^2+\left(\frac{\partial z}{\partial t}\right)^2\right]dx.$$

One then obtains

$$\int_\Omega\left\{|\nabla(w_1^+(x,t)-w_2^+(x,t))|^2+\left[\frac{\partial}{\partial t}(w_1^+(x,t)-w_2^+(x,t))\right]^2\right\}dx$$

$$\leqq\frac{D_1}{D_2}\exp\left(-\frac{D_2}{D_1}t\right)\left\{\int_\Omega[|\nabla(\varphi_1-\varphi_2)|^2+|\psi_1-\psi_2|^2]\,dx\right\}.$$

So assumption (i) of Theorem 2.1 is verified. Assumption (ii) of Theorem 2.1 can be verified in a similar manner. Assumptions (H1) and (H2) are obviously satisfied in this case. So exact controllability is proved. □

2.2. Exact controllability (II): local contractions at nonfixed time. For many nonlinear evolution control systems, it is impossible to find feedback operators which

produce global contractions. However, for any given ball with arbitrary radius, it may still be possible to find feedback operators which produce strict contractions in that ball. In this case, exact controllability still holds according to the next theorem. The feedback gain becomes larger and control time becomes longer as the initial and terminal states are farther apart.

THEOREM 2.2. *Let the operators A and B of* (NCS) *satisfy* (H1) *and* (H2). *Assume that there is a constant c, $0<c<1$ such that for any $R>0$ there are feedback operators K^+, K^- and some $T>0$ (K^+, K^- and T in general depend on R) such that*

(i) *$A+BK^+$ generates a nonlinear semigroup Φ on $\mathbb{R}^+\times X$ such that $\Phi(T,\cdot)$ is a strict contraction in the ball $\mathbb{B}_R=\{x\in X\mid \|x\|\leqq R\}$ with Lipschitz constant $c<1$ independent of R.*

(ii) *$A+BK^-$ generates a nonlinear semigroup Ψ on $\mathbb{R}^-\times X$ such that $\Psi(-T,\cdot)$ is a strict contraction in $\mathbb{B}_R$ with Lipschitz constant $c<1$.*

(iii) *$\Phi(t,0)=0$, $\Psi(-t,0)=0$ for $t\in\mathbb{R}^+$.*

Then for any $x_0, x_1\in X$, there exists T (depending on x_0, x_1) such that $x(t)$ is steered to x_1 at $t=T$.

Proof. For any given $x_0, x_1\in X$, let $R\in\mathbb{R}^+$ be a number such that

$$R\geqq(1-c)^{-1}\max\{\|x_0\|, \|x_1\|\}.$$

For this R, choose K^+, K^- and $T>0$ satisfying the given assumptions. Then the argument follows along the same line as that in the proof of Theorem 2, except that for the nonlinear mapping

$$\mathcal{T}: X\to X,$$

$$\mathcal{T}(v)\equiv x_0-\Psi(-T, x_1-\Phi(T, v))$$

we have

$$\mathcal{T}: \mathbb{B}_R\to\mathbb{B}_R,$$

a strict contraction with Lipschitz constant c^2, because

$$\|\Phi(T, v)\|=\|\Phi(T, v)-\Phi(T, 0)\|\leqq cR,$$

$$\|\Psi(-T, x_1-\Phi(T, v))\|\leqq c[\|x_1\|+\|\Phi(T, v)\|]\leqq c[(1-c)R+cR]=cR,$$

$$\|\mathcal{T}(v)\|\leqq\|x_0\|+cR\leqq R\Rightarrow\mathcal{T}(v)\in\mathbb{B}_R$$

and

$$\|\mathcal{T}(v_1)-\mathcal{T}(v_2)\|\leqq c^2\|v_1-v_2\|.$$

So $\mathcal{T}$ has a unique fixed point in $\mathbb{B}_R$. □

Example 3 (A nonlinear hard spring).

$$\frac{d}{dt^2}y(t)+[\alpha_1 y(t)+\alpha_2 y^3(t)+\cdots+\alpha_n y^{2n-1}(t)]=u(t), \tag{2.12}$$

$$\alpha_1>0,\quad \alpha_k\geqq 0,\quad 2\leqq k\leqq n.$$

We have

$$\frac{d}{dt}x(t)=A(x(t))+Bu(t)$$

with

$$x(t)=\begin{bmatrix}x_1(t)\\ x_2(t)\end{bmatrix},\quad A\left(\begin{bmatrix}x_1\\ x_2\end{bmatrix}\right)=\begin{bmatrix}x_2\\ -\sum_{j=1}^{n}\alpha_j x_1^{2j-1}\end{bmatrix},\quad B=\begin{bmatrix}0\\ 1\end{bmatrix}.$$

Without the control $u(t)$, solutions of (2.12) are known to have (periodic) closed orbits in the phase space (x_1, x_2) with period

$$T = 4\int_0^a \frac{dy}{\sqrt{h - \sum_{k=1}^{n} (\alpha_k/k) y^{2k}}},$$

where a is the spring's maximum displacement and h is twice the total energy of the spring.

(2.12) can be written as

$$\frac{d^2}{dt^2} y(t) + [\alpha_1 + f(y(t))] y(t) = u(t),$$

$$f(y) = \sum_{j=2}^{n} \alpha_j y^{2j-2} \geqq 0.$$

Let $K^+ = [0, \gamma]$, $\gamma > 0$. Then by limiting all the initial states (a, b) to be in $\mathbb{B}_R = \{(a, b) | a^2 + b^2 \leqq R^2\}$, f' can be bounded by a positive constant uniformly for all initial states in $\mathbb{B}_R$. One then applies the same Lyapunoff stability type argument to show that $A + BK^+$ is a strict contraction in $\mathbb{B}_R$. Since the argument is similar to that in Example 4, we leave the details to the reader. □

Example 4. Consider the following nonlinear controlled PDE:

$$\frac{\partial^2 w(x, t)}{\partial t^2} - \Delta w(x, t) + \alpha_1 w(x, t) + \alpha_2 \left(\int_\Omega w^2(x, t)\, dx \right) w(x, t) = u(x, t),$$

$$x \in \Omega \subseteq \mathbb{R}^n, \qquad t \geqq 0, \quad \alpha_1 \geqq 0, \quad \alpha_2 > 0,$$

$$w(x, 0) = w_0(x) \in H_0^1(\Omega),$$

$$\frac{\partial w}{\partial t}(x, 0) = v_0(x) \in L^2(\Omega).$$

The nonlinear mapping $U: L^2(\Omega) \to L^2(\Omega)$ defined by

$$U(w) = \left(\int_\Omega w^2\, dx \right) w$$

can be verified to be locally Lipschitzian as follows:

$$\begin{aligned}
\left\| \left(\int w_1^2\, dx \right) w_1 - \left(\int w_2^2\, dx \right) w_2 \right\|_{L^2} &\leqq \left\| \left(\int w_1^2\, dx \right) w_1 - \left(\int w_1^2\, dx \right) w_2 \right\| \\
&\quad + \left\| \left(\int w_1^2\, dx \right) w_2 - \left(\int w_2^2\, dx \right) w_2 \right\|_{L^2} \\
&\leqq \|w_1\|^2 \|w_1 - w_2\| + \|w_2\| \, \|w_1 + w_2\| \, \|w_1 - w_2\| \\
&= [\|w_1\|^2 + \|w_2\| \, \|w_1 + w_2\|] \|w_1 - w_2\| \\
&\leqq 3R^2 \|w_1 - w_2\|_{L^2(\Omega)},
\end{aligned}$$

provided that

$$\|w_1\|_{L^2} \leqq R, \qquad \|w_2\|_{L^2} \leqq R$$

Again, we let $K^+ = [0, -\gamma]$ and $K^- = [0, \gamma]$ for some $\gamma > 0$. We want to show that $A + BK^+$ and $A + BK^-$ with

$$A \begin{bmatrix} w \\ v \end{bmatrix} \equiv \begin{bmatrix} v \\ \Delta w - \alpha_1 w - \alpha_2 (\int w^2\, dx) w \end{bmatrix}$$

indeed generate nonlinear semigroups on $\mathbb{R}^+ \times X$ and $\mathbb{R}^- \times X$, respectively. If one applies [18, Theorem 4.4.3], one can only obtain the local existence and uniqueness of solutions of

$$\frac{d}{dt}\begin{bmatrix} w(\cdot, t) \\ v(\cdot, t) \end{bmatrix} = (A + BK^+)\begin{bmatrix} w(\cdot, t) \\ v(\cdot, t) \end{bmatrix}, \qquad t \geqq 0. \tag{2.13}$$

Fortunately, we can do much better here. We prove the global existence, uniqueness and continuity of solutions as follows. We rewrite (2.13) as

$$\frac{\partial^2 w}{\partial t^2} - \Delta w + \gamma \frac{\partial w}{\partial t} + \alpha_1 w + \alpha_2 \left(\int w^2 \, dx \right) w = 0. \tag{2.14}$$

We find that

$$\frac{d}{dt}\left\{ \int \frac{1}{2}\left[\left(\frac{\partial w}{\partial t} \right)^2 + |\nabla w|^2 + \alpha_1 w^2 \right] dx + \frac{\alpha_2}{2}\left[\int w^2 \, dx \right]^2 \right\} = -\gamma \int \frac{\partial w^2}{\partial t} \, dx \leqq 0.$$

So if we use the Ritz–Galerkin method

$$\begin{aligned} &w_n = \sum_{i=1}^{n} g_i(t)\varphi_i, \quad \text{where } \{\varphi_i\} \text{ is a basis for } H_0^1(\Omega), \\ &\left\langle \frac{d^2}{dt^2} w_n - \Delta w_n + \gamma \frac{\partial w_n}{\partial t} + \alpha_1 w_n + \alpha_2 \left(\int w_n^2 \, dx \right) w_n, \varphi_i \right\rangle_{L^2(\Omega)} = 0, \qquad i = 1, \cdots, n, \end{aligned} \tag{2.15}$$

and take in (2.15) the combination of φ_i corresponding to (dw_n/dt), we obtain

$$\frac{d}{dt}\left[\frac{1}{2}\left\| \frac{dw_n}{dt} \right\|_{L^2}^2 + \frac{1}{2}\|\nabla w_n\|_{L^2}^2 + \frac{\alpha_1}{2}\|w_n\|_{L^2}^2 + \frac{\alpha_2}{2}\|w_n\|_{L^2}^4 \right] \leqq 0.$$

This gives

$$w_n \in \text{bounded set of } L^\infty(0, T; H_0^1(\Omega)),$$

$$\frac{dw_n}{dt} \in \text{bounded set of } L^\infty(0, T; L^2(\Omega)).$$

Then one applies the compactness lemma of Lions [14], [23] and extracts a strongly convergent sequence $w_n \to w$ in $L^p(0, T; L^2(\Omega))$ for any $p > 1$. For the nonlinearity in (2.14), the convergence of $w_n \to w$ is easily shown by its local Lipschitzian property. Hence a solution w of (2.14) exists in the sense that

$$\begin{aligned} &w \in L^\infty(0, T; H_0^1(\Omega)), \\ &\frac{\partial w}{\partial t} \in L^\infty(0, T; L^\infty(\Omega)) \end{aligned} \tag{2.16}$$

for initial state $(w_0, v_0) \in H_0^1(\Omega) \oplus L^2(\Omega)$.

But for this w, the function $f(t)$ defined by

$$f(t) \equiv \int w^2(x, t) \, dx$$

is a continuous scalar function. So any solution φ (of the linear equation)

$$\frac{\partial^2\varphi(x,t)}{\partial t^2}-\Delta\varphi(x,t)+\gamma\frac{\partial\varphi(x,t)}{\partial t}+\alpha_1\varphi(x,t)+\alpha_2 f(t)\varphi(x,t)=0,$$

$$\varphi(x,0)=\varphi_0(x)\in H_0^1(\Omega),$$

$$\frac{\partial\varphi}{\partial t}(x,0)=\psi_0(x)\in L^2(\Omega)$$

satisfies $(\varphi, \partial\varphi/\partial t)\in C([0,T]; H_0^1(\Omega)\oplus H^0(\Omega))$, and in particular for $(\varphi_0,\psi_0)=(w_0, v_0)$. So (2.16) is improved to $C([0,T]; H_0^1(\Omega)\oplus H^0(\Omega))$. Thus (2.1)–(2.4) follows.

To show that the assumptions of Theorem 2.2 are satisfied, we proceed as follows. From the same procedures in Example 2 we obtain

$$\frac{d}{dt}P(t)+Q(t)=0,$$

where

$$P(t)=\frac{1}{2}\int_\Omega\left[|\nabla z|^2+\left(\frac{\partial z}{\partial t}\right)^2+(\alpha_1+\lambda\gamma)z^2+2\lambda z\frac{\partial z}{\partial t}\right]dx,$$

$$Q(t)=\int_\Omega\left\{(\gamma-\lambda)\left(\frac{\partial z}{\partial t}\right)^2+\lambda|\nabla z|^2+\lambda\alpha_1 z^2+\alpha_2[U(w_1)-U(w_2)](w_1-w_2)\right.$$
$$\left.+\alpha_2[U(w_1)-U(w_2)]\frac{\partial z}{\partial t}\right\}dx.$$

Since

$$\int_\Omega[U(w_1)-U(w_2)](w_1-w_2)\,dx\geqq 0,$$

and if we choose γ, λ such that $\gamma>\lambda$ and γ is so large that

$$\left|\int_\Omega\alpha_2[U(w_1)-U(w_2)]\frac{\partial z}{\partial t}\,dx\right|\leqq 3R^2\alpha_2\int\left|z\frac{\partial z}{\partial t}\right|dx$$
$$\leqq\frac{1}{2}(\gamma-\lambda)\int\left(\frac{\partial z}{\partial t}\right)^2dx+\frac{1}{2}\int[\lambda|\nabla z|^2+\lambda\alpha_1 z^2]\,dx,$$

then we have

$$Q(t)\geqq D_R^{(1)}P(t)\quad\text{for some } D_R^{(1)}>0\quad\text{(depending on } R)$$

for all initial states $(\varphi_1,\psi_1), (\varphi_2,\psi_2)\in\mathbb{B}_R$. Thus again we have

$$\int_\Omega\left\{|\nabla(w_1^+(x,t)-w_2^+(x,t))|^2+\left[\frac{\partial}{\partial t}(w_1^+(x,t)-w_2^+(x,t))\right]^2\right\}dx$$
$$\leqq D_R^{(2)}\exp(-D_R^{(1)}t)\left\{\int_\Omega[|\nabla(\varphi_1-\varphi_2)|^2+|\psi_1-\psi_2|^2]\right\}dx$$

for some constant $D_R^{(2)}>0$ depending on R. Therefore, one easily verifies that the assumptions in Theorem 2.2 are satisfied. □

Another PDE version of Example 3 is

$$\frac{\partial^2 w(x,t)}{\partial t^2} - \nabla w(x,t) + \alpha_1 w(x,t) + \alpha_2 w^3(x,t) = u(x,t),$$

$$x \in \Omega \subseteq \mathbb{R}^n \qquad (n = 2 \text{ or } 3), \qquad t \geqq 0$$

which is a controlled wave equation of the Klein–Gordon type. The state space is $X = [H_0^1(\Omega) \cap L^4(\Omega)] \oplus H_0(\Omega)$. Unfortunately, our theorems do not apply to this equation. It is not clear whether this equation is exactly controllable or not. □

2.3. Exact controllability (III): stabilization. We give our last theorem of exact controllability below. This theorem works only in a finite dimensional space. The proof does not seem to be constructive.

THEOREM 2.3. *Let* (NCS) *be a finite dimensional system with* $X = \mathbb{R}^n$ *and* B *an* $n \times m$ *constant matrix. Assume that for any* $R > 0$ *there are* $m \times n$ *matrices* K^+, K^- *and some* $T > 0$ (K^+, K^-, T *depend on* R) *such that*

(i) $A + BK^+$ *generates a nonlinear semigroup* Φ *on* $\mathbb{R}^+ \times \mathbb{R}^n$ *such that*

$$\Phi(T, \cdot) : \mathbb{B}_R \to \mathbb{B}_{c \cdot R}, \qquad 0 \leqq c < 1 \ (c \text{ independent of } R). \tag{2.17}$$

(ii) $A + BK^-$ *generates a nonlinear semigroup* Ψ *on* $\mathbb{R}^- \times \mathbb{R}^n$ *such that*

$$\Psi(-T, \cdot) : \mathbb{B}_R \to \mathbb{B}_{c \cdot R}. \tag{2.18}$$

Then for any $x_0, x_1 \in \mathbb{R}$, *there exist* $T > 0$ *and a control* $u \in C([0, T]; \mathbb{R}^m)$ *such that* $x(t)$ *is steered from* x_0 *to* x_1 *at* $t = T$.

Proof. The nonlinear operator $\mathcal{T} : \mathbb{R}^n \to \mathbb{R}^n$ defined by

$$\mathcal{T}(v) \equiv x_0 - \Psi(-T, x_1 - \Phi(T, v))$$

maps $\mathbb{B}_R$ into $\mathbb{B}_R$ by (2.17), (2.18) if R is chosen such that

$$R \geqq (1 - c)^{-1} \max \{|x_0|, |x_1|\}.$$

By Brouwer's fixed point theorem, $\mathcal{T}$ has at least one fixed point $v : \mathcal{T}(v) = v$. □

Example 5. In Example 1, if $L = 0$ and $f(0) = 0$, then Theorem 2.3 can be applied to give exact controllability. The result is also weaker: control time T varies with different x_0's and x_1's.

Theorem 2.3 is also applicable to Example 3.

3. Numerical techniques and applications. We develop an accurate numerical method (Algorithm 3.1) to solve the control problem (NCS) for a control $u(t)$ and state $x(t; u)$ such that $x(T) = x_1$. The technique presented is based on the feedback theory of Theorems 2.1 and 2.2. In § 3.1 we analyze this algorithm for control problems governed by *ordinary* differential equations ($X = \mathbb{R}^n$). Applications in § 3.2 show such solutions are obtained with remarkable accuracy. For systems governed by *partial* differential equations (X infinite dimensional) the success of the algorithm is unknown. We comment on the delicacy of such problems in § 3.1.

We assume we have the existence of feedback operators K^+ and K^- as specified in Theorems 2.1 and 2.2. Thus, if the semigroups $\Phi(t, \cdot)$ and $\Psi(t, \cdot)$ on $\mathbb{R}^+ \times X$ and $\mathbb{R}^- \times X$ generated by $A + BK^+$ and $A + BK^-$ are strictly contractive on a ball $\mathbb{B}_R \subseteq X$, then the continuous operator $\mathcal{T} : X \to X$ defined by

$$\mathcal{T}(v) = x_0 - \Psi(-T, x_1 - \Phi(T, v)) \tag{3.1}$$

has a unique fixed point, $\alpha \in \mathbb{B}_R$. The state $x(t)$ and control $u(t)$ then satisfy

$$x(t) = x^+(t) + x^-(t), \qquad 0 \leqq t \leqq T, \tag{3.2}$$

with $x^+(t) = \Phi(t, \alpha)$ and $x^-(t) = \Psi(t-T, x_1 - \Phi(T, \alpha))$, and

$$u(t) = K^+x^+(t) + K^-x^-(t) - w(t), \qquad 0 \leqq t \leqq T, \tag{3.3}$$

with $Bw = A(x^+ + x^-) - (Ax^+ + Ax^-)$. Under the hypotheses of Theorem 2.1 or 2.2, (3.2) and (3.3) solve (NCS) with $x(T) = x_1$.

Our goal is to approximate (3.2) and (3.3) by using an iteration scheme to solve (3.1) for α. One such iteration scheme is

$$\alpha_0 \text{ given}, \qquad \alpha_{n+1} = \mathcal{T}(\alpha_n), \tag{3.4}$$

which is quite useful (indeed necessary) if X is an infinite dimensional Banach space. However if $X = \mathbb{R}^n$ then there are standard schemes available which may *converge faster* than (3.4). These are generally patterned [17] after

$$\begin{aligned} &\alpha_0 \in \mathbb{R}^n \text{ given}, \\ &\alpha_{n+1} = \alpha_n - J^{-1}(\alpha_n) \cdot (\mathcal{T}(\alpha_n) - \alpha_n). \end{aligned} \tag{3.5}$$

For example, if $F(x) = \mathcal{T}x - x$ and $J(\alpha_n) = (DF)(\alpha_n)$ then (3.5) is a second order Newton's method to solve $F(x) = 0$ [17].

Each of (3.4) and (3.5) (and all other one-point iteration schemes) are of the form $\alpha_{n+1} = G(\alpha_n, \mathcal{T})$, where G is an iteration function for solving $\mathcal{T}\alpha = \alpha$. Since $G(\alpha_n, \mathcal{T})$ requires the calculation of the semigroups Φ and Ψ, we must approximate these semigroups to make the iteration computationally feasible. Specifically, $x^+(t) = \Phi(t, v_1) \in X$ and $x^-(t) = \Psi(t-T, v_2) \in X$ are the solutions to

$$\frac{dx^+}{dt} = (A + BK^+)x^+, \qquad \frac{dx^-}{dt} = (A + BK^-)x^-,$$

$$x^+(0) = v_1, \qquad x^-(T) = v_2,$$

for $0 \leqq t \leqq T$, respectively. We approximate these solutions by an $O(h^p)$ differential equation solver. We make the following assumption which is valid for most numerical schemes to solve *ordinary* differential initial value problems.

(H3) Let $0 = t_0 < t_1 < \cdots < t_k = T$ be a partition of $[0, T]$ and $h = \max |t_{i+1} - t_i|$. If $y(t) \in X$ is the exact solution to $dy/dt = f(y(t))$, $y(t_0) = y_o \in X$, $t_0 \leqq t \leqq T$, and $y_h(t_i)$ is an approximate solution with $y_h(t_0) = y_0$, then $\|y(t_i) - y_h(t_i)\| \leqq C_d h^p$, for $0 < h \leqq h_d$, where C_d is independent of h_d, y, y_0 and depends on f, T and t_0.

For *partial* differential equation solvers, C_d above depends on y and y_0 as well, and h is a mesh length for a partition of $[0, T] \times \Omega$.

If this approximation is used in the iteration, then $\mathcal{T}(\alpha_n)$ becomes approximated by a perturbation, $\mathcal{T}_h(\alpha_{n,h})$. Taking all the above into account, we formulate the following iteration scheme.

ALGORITHM 3.1. Let K^+, K^-, T, x_0, x_1 be given, and B be injective.

(1) Let $\alpha_{0,h} \in X$.

(2) For $n = 0, 1, \cdots, N$

(1) Solve

$$\frac{dx_n^+}{dt} = (A + BK^+)(x_n^+),$$

$$x_0^+(0) = \alpha_{n,h}$$

for $x_{n,h}^+(t_i)$, $i = 0, \cdots, k$ by the differential solver.

(2) Solve

$$\frac{dx_n^-}{dt} = (A + BK^-)(x_n^-),$$

$$x_n^-(T) = x_1 - x_{n,h}^+(T)$$

for $x_{n,h}^-(t_i)$, $i = 0, 1, \cdots, k$ by the differential solver.

(3) Let $\mathcal{T}_h(\alpha_{n,h}) = x_0 - x_{n,h}^-(0)$.

(4) $\alpha_{n+1,h} = G(\alpha_{n,h}, \mathcal{T}_h)$.

(3) $x_{N,h}(t_i) = x_{N,h}^+(t_i) + x_{N,h}^-(t_i), \qquad i = 0, \cdots, k.$

(4) Solve $B\tilde{w}(t_i) = A(x_{N,h}^+(t_i) + x_{N,h}^-(t_i)) - [A(x_{N,h}^+(t_i) + A(x_{N,h}^-(t_i))]$ for $\tilde{w}(t_i)$, $i = 0, \cdots, k$.

(5) Let $u_{N,h}(t_i) = K^+ x_{N,h}^+(t_i) + K^- x_{N,h}(t_i) - \tilde{w}(t_i)$.

If each $G(\alpha_{n,h}, \mathcal{T}_h)$ is defined, then Algorithm 3.1 is well defined and computationally performable. To allow for package differential solvers, we note that h may be allowed to vary from step to step. Also, note that the iteration step 2.4 must use the *perturbed* operator, $\mathcal{T}_h$.

It remains to show under what conditions the algorithm is defined and successful.

3.1. Error estimates and convergence of Algorithm 3.1. Our main goal is to show the errors in $\|x_{N,h}(t_i) - x(t)\|$ and $\|u_{N,h}(t_i) - u(t)\|$, where $x_{N,h}$, $u_{N,h}$ are the iterates of Algorithm 3.1 and x, u are the exact solutions of Theorems 2.1 or 2.2. To do this we must formulate a theory for the iteration mapping G based on its application to the perturbed operator $\mathcal{T}_h$ of Algorithm 3.1. This theory will include the iterations (3.4) and (3.5).

Let $\mathcal{T}: X \to X$ such that $\mathcal{T}\alpha = \alpha$, D be an open set $\subset X$ with $\alpha \in D$, and $\mathcal{F}(X)$ be the set of all functions on X. Let $G: D \times \mathcal{F}(X) \to X$ be a mapping satisfying

(A1) $G(\alpha, \mathcal{T}) = \alpha$.

(A2) The iterates,

$$\beta_0 \in D, \qquad \beta_{n+1} = G(\beta_n, \mathcal{T})$$

converge to α for β_0 sufficiently close to α.

(A3) There exist an ε_1-neighborhood, $D_1 \subset D$, of α and a constant $C_1 > 0$ depending on D_1 such that $\|G(x_1, \mathcal{T}) - G(x_2, \mathcal{T})\| \leqq C_1 \|x_1 - x_2\|$ for all $x_1, x_2 \in D_1$.

(A4) There exist an ε_2-neighborhood, $D_2 \subset D$, of α and a constant $C_2 > 0$ depending on D_2 such that, for all $x \in D_2$ and $\mathcal{H} \in \mathcal{F}(X)$, $\|G(x, \mathcal{H}) - G(x, \mathcal{T})\| \leqq C_2 \|\mathcal{H}x - \mathcal{T}x\|$.

The iteration mapping for (3.4) is

$$G(x, \mathcal{H}) = \mathcal{H}_x \quad \text{with } D = X. \tag{3.6}$$

The following lemma applies to the operator $\mathcal{T}$ and fixed point α of Theorems 2.1 and 2.2.

LEMMA 3.2. *Let X be any Banach space and let $\mathcal{T}: X \to X$ be strictly contractive on $\mathbb{B}_R \subseteq X$ with $\mathcal{T}\alpha = \alpha \in \mathbb{B}_R$. Then the iteration mapping* (3.6) *satisfies* (A1)–(A4) *with the constant $C_1 < 1$ in* (A3). *Furthermore, if $\mathbb{B}_R = X$, (i.e., $R = \infty$) then $D_1 = D_2 = X$ in* (A3) *and* (A4).

Proof. (A1) is $G(\alpha, \mathcal{T}) = \mathcal{T}\alpha = \alpha$. (A2) is classical with $\beta_0 \in \mathbb{B}_R$. For (A3) let $D_1 \equiv$ ball about α with radius $R - |\alpha| = \varepsilon_1$. (If $\mathbb{B}_R = X$, $D_1 = X$.) Then $D_1 \subset \mathbb{B}_R$ and for any $x_1, x_2 \in D_1$, $\|G(x_1, \mathcal{T}) - G(x_2, \mathcal{T})\| = \|\mathcal{T}x_1 - \mathcal{T}x_2\| \leqq C_1\|x_1 - x_2\|$. For (A4) let $D_2 = X$. For $x \in D_2$, $\|G(x, \mathcal{T}) - G(x, \mathcal{H})\| = \|\mathcal{T}x - \mathcal{H}x\|$. □

The iteration mapping for (3.5) is as follows. Let D be a neighborhood of α and $J: D \to \mathbb{R}^{n \times n}$ be a linear function such that J and J^{-1} exist and are bounded on D. Then (3.5) is

$$G(x, \mathcal{H}) = x - J^{-1}(x) \cdot (\mathcal{H}x - x) \quad \text{for } x \in D, \quad \mathcal{H} \in \mathcal{F}(X). \tag{3.7}$$

LEMMA 3.3. *Let $\mathcal{T}: \mathbb{R}^n \to \mathbb{R}^n$ be contractive on $\mathbb{B}_R \subseteq \mathbb{R}^n$, $\mathcal{T}\alpha = \alpha \in \mathbb{B}_R$, and $\mathcal{T}$ be continuous at α. Assume J is a Lipschitz continuous on D and* (3.7) *satisfies* (A2). *Then* (3.7) *satisfies* (A1)–(A4).

Proof. (A1) is simply $G(\alpha, \mathcal{T}) = \alpha - J^{-1}(\alpha) \cdot (\mathcal{T}\alpha - \alpha) = \alpha$, since $J^{-1}(\alpha)$ exists. To show (A3) and (A4) we note that, for $x_1, x_2 \in D$ and $y_1, y_2 \in \mathbb{R}^n$,

$$\begin{aligned} J^{-1}(x_1) \cdot y_1 - J^{-1}(x_2) \cdot y_2 &= J^{-1}(x_1) \cdot y_1 - J^{-1}(x_1) \cdot y_2 + J^{-1}(x_1) \cdot y_2 - J^{-1}(x_2) \cdot y_2 \\ &= J^{-1}(x_1)(y_1 - y_2) - J^{-1}(x_1)[J(x_2) - J(x_1)] \cdot J^{-1}(x_2) \cdot y_2. \end{aligned}$$

Hence, since $x_1, x_2 \in D$,

$$\|J^{-1}(x_1) \cdot y_1 - J^{-1}(x_2) \cdot y_2\| \leqq C_1\|y_1 - y_2\| + C_2\|J(x_2) - J(x_1)\|_{R^{n \times n}} \cdot \|y_2\|. \tag{3.8}$$

To show (A3), let $D_1 \subset D \cap \mathbb{B}_R$ such that $\alpha \in D_1$ and $\mathcal{T}$ is continuous on D_1, $y_1 = \mathcal{T}x_1 - x_1$ and $y_2 = \mathcal{T}x_2 - x_2$ in (3.8). Then, for $x_1, x_2 \in D_1$, (3.8) and continuity of $\mathcal{T}$ give

$$\begin{aligned} \|G(x_1, \mathcal{T}) - G(x_2, \mathcal{T})\| &\leqq \|x_1 - x_2\| + \|J_{\mathcal{T}}^{-1}(x_1) \cdot y_1 - J_{\mathcal{T}}^{-1}(x_2) \cdot y_2\| \\ &\leqq \|x_1 - x_2\| + C_1\|\mathcal{T}x_1 - \mathcal{T}x_2\| + C_1\|x_1 - x_2\| \\ &\quad + C_2'\|x_2 - x_1\| \cdot \|Tx_2 - x_2\| \\ &\leqq C_1\|Tx_1 - Tx_2\| + C_3\|x_1 - x_2\|, \end{aligned}$$

the last term coming from the Lipschitz continuity of J. Finally, since $\mathcal{T}$ is contractive on $\mathbb{B}_R$, $\|\mathcal{T}x_1 - \mathcal{T}x_2\| \leqq \|x_1 - x_2\|$, giving (A3).

To show (A4), let $D_2 \subset D \cap \mathbb{B}_R$, $x_1 = x_2 = x \in D$, $y_1 = \mathcal{T}x - x$, $y_2 = \mathcal{H}x - x$ in (3.8). Then $\|G(x, \mathcal{T}) - G(x, \mathcal{H})\| = \|J^{-1}(x) \cdot y_1 - J^{-1}(x) \cdot y_2\| \leqq C_1\|\mathcal{T}x - \mathcal{H}x\|$. □

We now apply Lemmas 3.2–3.3 to Algorithm 3.1 for differential solvers satisfying (H3) and iteration mappings satisfying (A1)–(A4). We begin with a technical lemma which gives an estimate for the iterate $\alpha_{n,h}$ in Algorithm 3.1. It covers the situation in either Theorem 2.1 or 2.2.

LEMMA 3.4. *Let $R > 0$, x_0, x_1, K^+, K^-, T satisfy the hypotheses of Theorems* 2.1 *or* 2.2. *($R \equiv \infty$ in Theorem* 2.1.) *Let $\alpha \in \mathbb{B}_R$ be the fixed point of $\mathcal{T}$, $G: D \times \mathcal{F}(X) \to X$ satisfy* (A1)–(A4), *and the differential solver satisfy* (H3). *If $\alpha_{0,h}$ is chosen sufficiently close to α, then there exists $h_n > 0$ (depending on n) such that, for $n \geqq 0$,*

(1) *each iterate $\alpha_{n,h}$ of Algorithm* 3.1 *is defined and $\alpha_{n,h} \in \mathbb{B}_R \cap D$, and*

(2) $\|\alpha_{n,h} - \alpha\| \leqq 2C_dC_2((1 - C_1^n)/(1 - C_1))h^p + 0(\|\alpha_n - \alpha\|)$, *for $0 < h \leqq h_n$, where C_1 is the Lipschitz constant for G in* (A3), *C_2 the constant in* (A4), *and α_n are the iterates of* (A2).

Furthermore, if $\mathcal{T}$ is contractive on all of X (as in Theorem 2.1) and $D = D_1 = D_2 = X$ in (A3) *and* (A4), *then h_n is independent of n.*

Proof. Let $\mathbb{B}_R$ be the ball of contraction of $\mathcal{T}$. ($R = \infty$ if $\mathbb{B}_R = X$.)

Let $D_1 \subset D$ be the ε_1-neighborhood and $D_2 \subset D$ the ε_2-neighborhood of α from (A3), (A4), where D is the domain of G. Let $\varepsilon_3 = \min\{\varepsilon_1, \varepsilon_2, R - |\alpha|\}/2$ and $D_3 = \varepsilon_3$-neighborhood of α. By (A2), $\beta_{n+1} \equiv G(\beta_n, \mathcal{T}) \to \alpha$ for β_0 sufficiently close to α. Hence, there exists N such that $\|\beta_n - \alpha\| \leqq \varepsilon_3/2$, for $n \geqq N$. Choose $\alpha_{0,h} \equiv \alpha_0 \equiv \beta_N \in D_3$. Then $\|\alpha_n - \alpha\|_X \leqq \varepsilon_3/2$ for all $n \geqq 0$.

Now we claim that for each $n = 0, 1, 2, \cdots$ there exists $h_n > 0$ such that $\alpha_{n,h} \in D_3$ and $\|\alpha_{n,h} - \alpha_n\| \leqq 2C_2C_d((1 - C_1^n)/(1 - C_1))h^p$ for $0 < h \leqq h_n$. (C_2 the constant in (A4) and C_d the constant in (H3).)

The claim is true for $n = 0$ since $\|\alpha_{0,h} - \alpha_0\| = 0$. Assume the claim true for $k = n$. Then, since $\alpha_{n,h} \in D_3 \subset D$, $0 < h < h_n$, $G(\alpha_{n,h}, \cdot)$ is defined. Hence, for $0 < h < h_n$,

$$\begin{aligned} \|\alpha_{n+1,h} - \alpha_{n+1}\| &= \|G(\alpha_{n,h}, \mathcal{T}_h) - G(\alpha_n, \mathcal{T})\| \\ &\leqq \|G(\alpha_{n,h}, \mathcal{T}_h) - G(\alpha_{n,h}, \mathcal{T})\| + \|G(\alpha_{n,h}, \mathcal{T}) - G(\alpha_n, \mathcal{T})\| \\ &\leqq C_2\|\mathcal{T}_h\alpha_{n,h} - \mathcal{T}\alpha_{n,h}\| + C_1\|\alpha_{n,h} - \alpha_n\|, \end{aligned} \tag{3.9}$$

by (A4) and (A3). Next,

$$\begin{aligned} \|\mathcal{T}_h\alpha_{n,h} - \mathcal{T}\alpha_{n,h}\| &= \|x^-_{n,h}(0) - \Psi(-T, x_1 - \Phi(T, \alpha_{n,h}))\| \\ &\leqq \|x^-_{n,h}(0) - \Psi(-T, x_1 - x^+_{n,h}(T))\| \\ &\quad + \|\Psi(-T, x_1 - x^+_{n,h}(T)) - \Psi(-T, x_1 - \Phi(T, \alpha_{n,h}))\|. \end{aligned} \tag{3.10}$$

The first term is $\leqq C_d h^p$, $0 \leqq h \leqq h_d$ by (H3). Next, $\|x_1 - \Phi(T, \alpha_{n,h})\| \leqq \|x_1\| + c\|\alpha_{n,h}\| \leqq R$ by Theorem 2.2 where $c < 1$ is the semigroup contraction constant for Φ. So, $x_1 - \Phi(T, \alpha_{n,h}) \in \mathbb{B}_R$. Also, since $\alpha_{n,h} \in D_3$, $\|\alpha_{n,h} - \alpha\| < \varepsilon_3 \leqq (R - |\alpha|)/2$ giving $\|\alpha_{n,h}\| \leqq (R + |\alpha|)/2$ ($< R$). So

$$\begin{aligned} \|x_1 - x^+_{n,h}(T)\| &\leqq \|x_1\| + \|x^+_{n,h}(T)\| \\ &\leqq (1 - c)R + \|x^+_{n,h}(T) - \Phi(T, \alpha_{n,h})\| + \|\Phi(T, \alpha_{n,h})\| \\ &\leqq (1 - c)R + C_d h^p + c((R + |\alpha|)/2) \\ &= R + [C_d h^p - c((R - |\alpha|)/2)]. \end{aligned}$$

Choosing $\tilde{h}_{n+1} > 0$ such that $C_d\tilde{h}^p_{n+1} < c(R - |\alpha|)/2$ gives $x_1 - x^+_{n,h}(T) \in \mathbb{B}_R$, $0 < h \leqq \tilde{h}_{n+1}$. Thus, since Ψ contracts $\mathbb{B}_R$, (H3) gives

$$\begin{aligned} &\|\Psi(-T, x_1 - x^+_{n,h}(T)) - \Psi(-T, x_1 - \Phi(T, \alpha_{n,h}))\| \\ &\quad \leqq c\|x^+_{n,h}(T) - \Phi(T, \alpha_{n,h})\| \leqq C_d h^p, \qquad 0 \leqq h \leqq \tilde{h}_{n+1}. \end{aligned}$$

Hence, (3.10) is $\|\mathcal{T}_h\alpha_{n,h} - \mathcal{T}\alpha_{n,h}\| \leqq 2C_d h^p$, $0 \leqq h \leqq \tilde{h}_{n+1}$. Therefore, by (3.9) and the induction hypotheses,

$$\begin{aligned} \|\alpha_{n+1,h} - \alpha_{n+1}\| &\leqq 2C_2C_d h^p + C_1\left[2C_2C_d\left(\frac{1 - C_1^n}{1 - C_1}\right)h^p\right] \\ &= 2C_2C_d\left[\frac{1 - C_1^{n+1}}{1 - C_1}\right]h^p, \qquad 0 < h \leqq \min\{h_n, \tilde{h}_{n+1}\}. \end{aligned} \tag{3.11}$$

Now choose $\tilde{\tilde{h}}_{n+1}$ such that $2C_2C_d[(1 - C_1^{n+1})/(1 - C_1)]\tilde{\tilde{h}}^p_{n+1} < \varepsilon_3/2$. Then (3.11) holds for $0 < h \leqq h_{n+1} \equiv \min\{h_n, \tilde{h}_{n+1}, \tilde{\tilde{h}}_{n+1}\}$ and $\alpha_{n+1,h} \in D_3$. By induction, the claim is true.

Note if $\mathbb{B}_R = D = D_1 = D_2 = X$, then the claim is true for h_n independent of n. This completes the proof. □

THEOREM 3.5. *Let* R, x_0, x_1, T, K^+, *and* K^- *satisfy the hypotheses of Theorem* 2.1 *or* 2.2, *and* $x(t)$, $u(t)$ *be the exact solutions to* (NCS) *with* $x(T) = x_1$. *Let* $x_{n,h}(t_i)$, $u_{n,h}(t_i)$ *be the* n*th iterates of Algorithm* 3.1 *with* G *satisfying* (A1)–(A4), *and the differential solver satisfying* (H3). *Then, for* $\alpha_{0,h}$ *sufficiently close to* α, *there exists* C_3, $C_4 > 0$ *independent of* h, n *and* $h_n > 0$ *such that*

$$\|x(t_i) - x_{n,h}(t_i)\| \leqq C_3 h^p + C_4 \left(\frac{1 - C_1^N}{1 - C_1}\right) h^p + O(\|\alpha_n - \alpha\|) \tag{3.12}$$

and

$$\|u(t_i) - u_{n,h}(t_i)\| \leqq C_3 h^p + C_4 \left(\frac{1 - C_1^n}{1 - C_1}\right) h^p + O(\|\alpha_n - \alpha\|) + O(\|\mathcal{N}(x^+_{n,h}(t_i), x^-_{n,h}(t_i)) - \mathcal{N}(x^+(t_i), x^-(t_i))\|)$$

for $0 < h \leqq h_n$, *where* $\mathcal{N}$ *is as given in* (H1) *and* C_1 *the constant in* (A3).

Furthermore, if $\mathcal{T}$ *is contractive on all of* X *and* $D = D_1 = D_2 = X$ *in* (A3) *and* (A4), *then* h_n *is independent of* n.

Proof. Let h_n be as in the conclusion of Lemma 3.4. Note $x(t_i) = x^+(t_i) + x^-(t_i)$ and by Algorithm 3.1 $x_{n,h}(t_i) = x^+_{n,h}(t_i) + x^-_{n,h}(t_i)$. Hence, $\alpha_{n,h} \in \mathbb{B}_R$ gives $\|x^+_{n,h}(t_i) - x^+(t_i)\| \leqq \|x^+_{n,h}(t_i) - \Phi(t_i, \alpha_{n,h})\| + \|\Phi(t_i, \alpha_{n,h}) - \Phi(t_i, \alpha)\| \leqq C_d h^p + \|\alpha_{n,h} - \alpha\|_X$ by (H3) and the contraction Φ. By Lemma 3.4, it follows that for $0 < h \leqq h_n$,

$$\|x^+_{n,h}(t_i) - x^+(t_i)\| \leqq C_d h^p + 2C_d C_2 \left(\frac{1 - C_1^n}{1 - C_1}\right) h^p + O(\|\alpha_n - \alpha\|). \tag{3.13}$$

Similar results hold for $x^-_{n,h} - x^-$ giving the conclusion.

For the $u(t)$ estimate we have, by Algorithm 3.1 steps 4 and 5,

$$u_{n,h}(t_i) - u(t_i) = K^+(x^+_{n,h}(t_i) + x^+(t_i)) + K^-(x^-_{n,h}(t_i) - x^-(t_i)) + w(t_i) - \tilde{w}(t_i), \tag{3.14}$$

where, by the injectivity of B and (H2),

$$\tilde{w}(t_i) - w(t_i) = B^{-1}[\mathcal{N}(x^+_{n,h}(t_i), x^-_{n,h}(t_i)) - \mathcal{N}(x^+_{n,h}(t_i), x^-(t_i))].$$

The desired result follows from (3.13) and (3.14). □

COROLLARY 3.6. *Under the hypotheses of Theorem* 3.5, *if* $C_1^n h_n^p \to 0$ *as* $n \to \infty$, $h_n \to 0$, *then*

$$x_{n,h}(t_i) \to x(t_i) \quad \text{and} \quad u_{n,h}(t_i) \to u(t_i)$$

as $n \to \infty$, $h_n \to 0$.

Proof. The x convergence is clear. Since $\mathcal{N}(\cdot, \cdot)$ is continuous, the u convergence follows. □

An interesting corollary of the above is the following error estimate and convergence result concerning the iteration scheme (3.4).

THEOREM 3.7. *Let the hypotheses of Theorem* 3.5 *hold with the iteration function* (3.6) *in Algorithm* 3.1. *Then there exists an* $h_n > 0$ *such that*

$$\|x(t_i) - x_{n,h}(t_i)\| \leqq Ch^p + O(\|\alpha_n - \alpha\|)$$

for any $0 < h \leqq h_n$. *If* $\mathcal{T}$ *is strictly contractive on* X, *then* $x_{n,h}(t_i) \to x(t_i)$, $u_{n,h}(t_i) \to u(t_i)$ *as* $n \to \infty$ *and* $h \to 0$ (*independent of* n).

Proof. By Lemma 3.2 the constant $C_1 < 1$ in (A3). Hence, for $0 < h \leqq h_n$,

$$\|x(t_i) - x_{n,h}(t_i)\| \leqq C_3 h^p + C_4\left(\frac{1-C_1^n}{1-C_1}\right) h^p + O(\|\alpha_n - \alpha\|) \leqq C_5 h^p + O(\|\alpha_n - 2\alpha\|)$$

by Theorem 3.5. Next, by Lemma 3.2, if $\mathcal{T}$ contracts all of X, $D_1 = D_2 = X$ in (A3) and (A4). Hence by Theorem 3.5 h_n is independent of n, so the result holds. □

Remarks.

(1) The term $(1-C_1^n)/(1-C_1)h^p$ appearing in Theorem 3.5 is not surprising since the errors created by the differential solver are propagated by the iteration mapping G. The local Lipschitz constant C_1 for G in (A3) is usually not <1 in a neighborhood of a fixed point α. It will be <1 in some cone with vertex α, but there is no guarantee that the iterates of Algorithm 3.1 will lie in this cone. In any event, C_1 is close to 1, and hence the error in x and u will be dominated by h^p. Therefore, as Theorem 3.5 shows, the mesh h should be decreased as the iterations of Algorithm 3.1 proceed. This will counter the possible effects of this error term.

(2) The term $O(\|\alpha_n - \alpha\|)$ in Theorem 3.5 is the error in the exact iteration (A2), and is a function of the choice of G.

(3) The success or failure of Algorithm 3.1 applied to control problems governed by *partial* differential equations is unknown. Assumption (H3) must be modified to allow C_d to depend on the exact solution y and the initial conditions y_0. In this case the error in the differential solver is usually of the form

$$\|y(t_i) - y_h(t_i)\|_X \leqq C(T)h^p(\|y\|_{X_1} + \|y_0\|_{X_2}) \tag{3.15}$$

where X_1 and X_2 are higher order Sobolev spaces. Revisiting the proof of Lemma 3.4 will reveal a further error due to the propagation of initial condition errors. Furthermore, the partial differential equations are solved in pairs (forward and backward) with the final time solution of the first used in the initial conditions of the second. According to (3.15) this final time solution must be smooth enough to achieve the h^p accuracy. Otherwise accuracy will be lost. From these comments we conjecture the algorithm to be successful if the finite element method with smooth splines is used as the differential solver. This analysis will require further assumptions on the control problem and will not be discussed here.

3.2. Application of Algorithm 3.1. We apply our algorithm to two examples covered by the theory of Theorems 2.1 and 2.2. Both examples are differential control problems (NCS) with $X = \mathbb{R}^n$ and $A: \mathbb{R}^n \to \mathbb{R}^n$. We choose an iteration scheme of type (3.7) with $J(x) = D\mathcal{T}(x)$, (Jacobian of $\mathcal{T}$). Since the coefficients of the problem will be smooth, $\mathcal{T}$ will be twice Fréchet differentiable and hence D will be locally Lipschitz around α. Thus, by [17] the iteration (3.5) is *second order* convergent for α_0 sufficiently close to α. Hence, by Lemma 3.3 the results of Theorem 3.5 apply.

The differential scheme used is the automatic package DVERK [13] which is based on $O(h^5)$ and $O(h^6)$ Runge–Kutta schemes and satisfies (H3). In both examples the mesh h was decreased with increasing n automatically by DVERK to control local errors. Hence, an attempt was made to stay within the upper bounds in Theorem 3.5. The computing was performed on an IBM 370 Mod 3033.

Numerical example 1. We solve

$$y''(t) + y(t) + \tfrac{1}{20}\sin(5y) + 5 = u(t) \quad \text{on } [0, 1]$$

(Example 1 of § 2). This converts to

$$\frac{d}{dt}x = \begin{bmatrix} x_2 \\ -x_1 - \frac{1}{20}\sin(5x_1) - 5 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t),$$

$$x(0) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \qquad x(1) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

By § 2, the choice of $T = 1$, $K^+ = [0, -1]$, $K^- = [0, 1]$ satisfies the hypotheses of Theorem 2.1.

The results of Algorithm 3.1 appear in Fig. 2. The initial and ending conditions were obtained with a relative error of 10^{-13} with the control as given, a very accurate result. The CPU time involved was extremely small (3 seconds).

Numerical example 2. We couple together Examples 1 and 3 of § 2 into

$$\frac{dx}{dt} = \begin{bmatrix} x_2 \\ -x_1 - \frac{1}{15}\sin(5x_3) - 3 \\ x_4 \\ -x_3 - x_1 - x_1^3 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix},$$

$$x(0) = \begin{bmatrix} 0 \\ 0.4 \\ 0.3 \\ -0.1 \end{bmatrix}, \qquad x(2) = \begin{bmatrix} 0.2 \\ 0 \\ 0.1 \\ 0.2 \end{bmatrix}.$$

According to the theory of Examples 1 and 3 of § 2, $T = 2$, x_0, x_1 as given and K^+, K^- of the form

$$K^+ = \begin{bmatrix} 0 & -\gamma & 0 & 0 \\ 0 & 0 & 0 & -\gamma \end{bmatrix}, \qquad K^- = \begin{bmatrix} 0 & \gamma & 0 & 0 \\ 0 & 0 & 0 & \gamma \end{bmatrix}$$

will satisfy the hypotheses of Theorem 2.2 for γ large enough.

Algorithm 3.1 was performed for two cases, $\gamma = 2$ (Fig. 3) and $\gamma = 10$ (Fig. 4). Relative accuracy was achieved to 10^{-12}. The $\gamma = 2$ case required 5 seconds and $\gamma = 10$ 50 seconds.

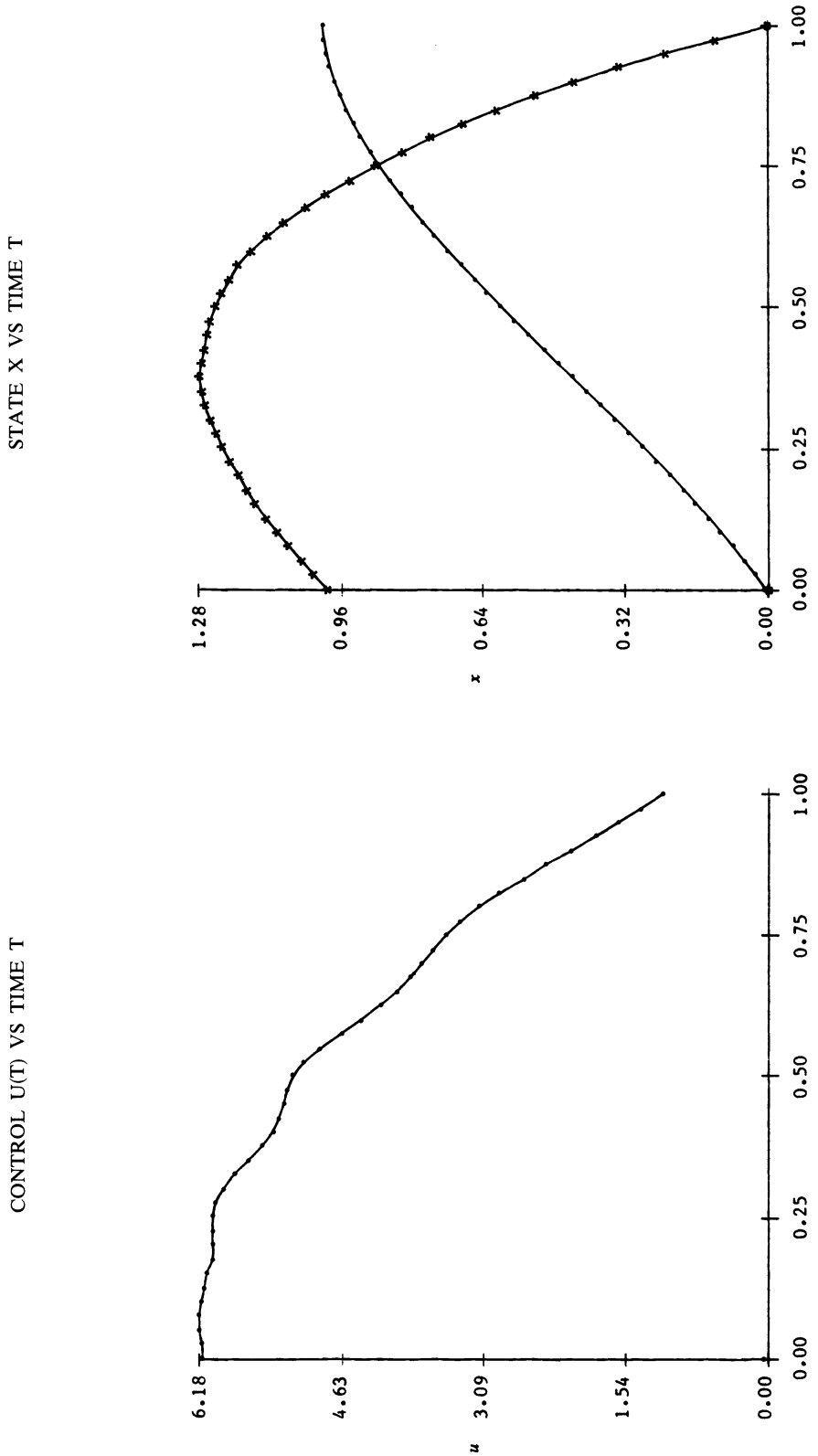
Remarks.

(1) The remarkable accuracy of the solutions of the above examples indicate Algorithm 3.1 to be a highly successful technique for solving nonlinear control problems covered by Theorems 2.1 and 2.2. In applying this algorithm the mesh should be controlled keeping in mind the error estimates of Theorem 3.5. As always, a measure of success is the accuracy attained in the end condition, $x_{n,h}(T) \approx x_1$.

(2) The storage required for Algorithm 3.1 is no more than that of solving two systems of differential equations. For most automatic system solvers, this is minimal. Hence, large order (NCS) problems may be solved.

(3) The time (operations) involved in Algorithm 3.1 is directly related to the choice of K^+ and K^-. If the linearization of $dx/dt = (A + BK)x$ is stiff ("large" K's), the time is increased in accordance with stiff ODE solvers. For moderate K^+, K^- (e.g., $\gamma = 2$ in numerical example 2) the solution is obtained quickly. In either case, the solution is obtained accurately.

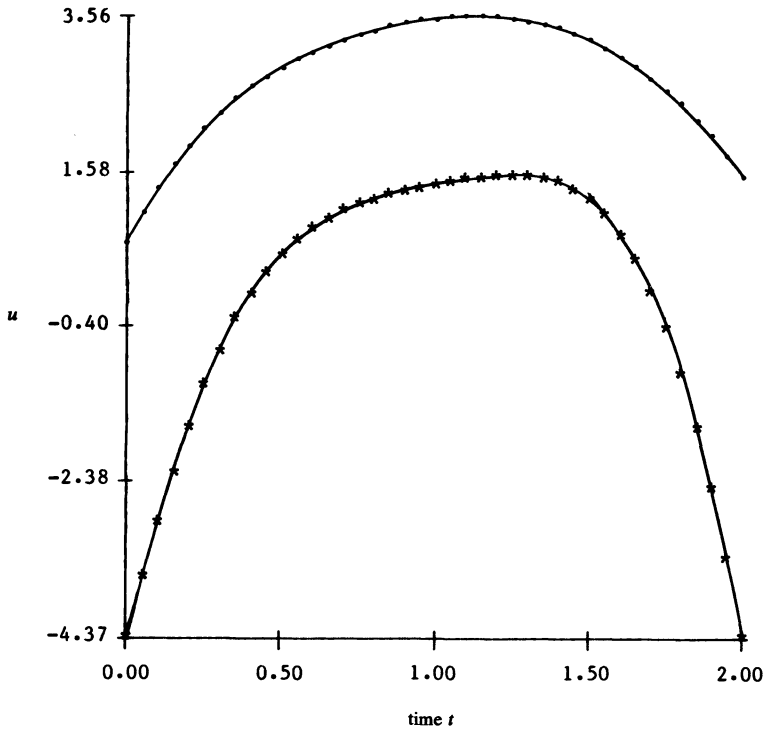
Acknowledgment. We wish to thank the referee for valuable comments concerning the application of Algorithm 3.1 to partial differential control problems.



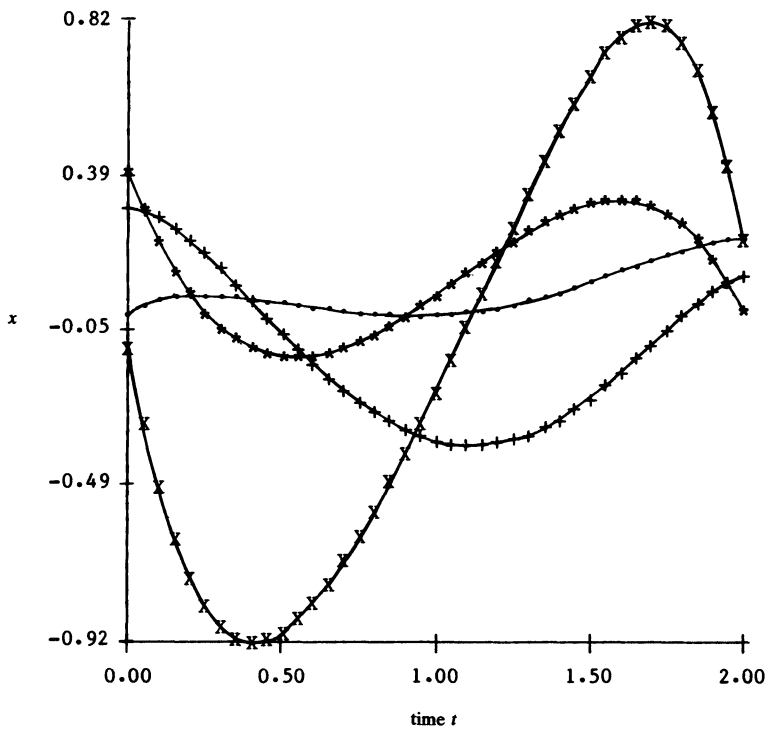
a. *Control* $u(t)$ *vs.* t. $u_1 \cdots$.

b. *State,* x *vs.* t. $x_1 \cdots ; x_2 ***$.

FIG. 2

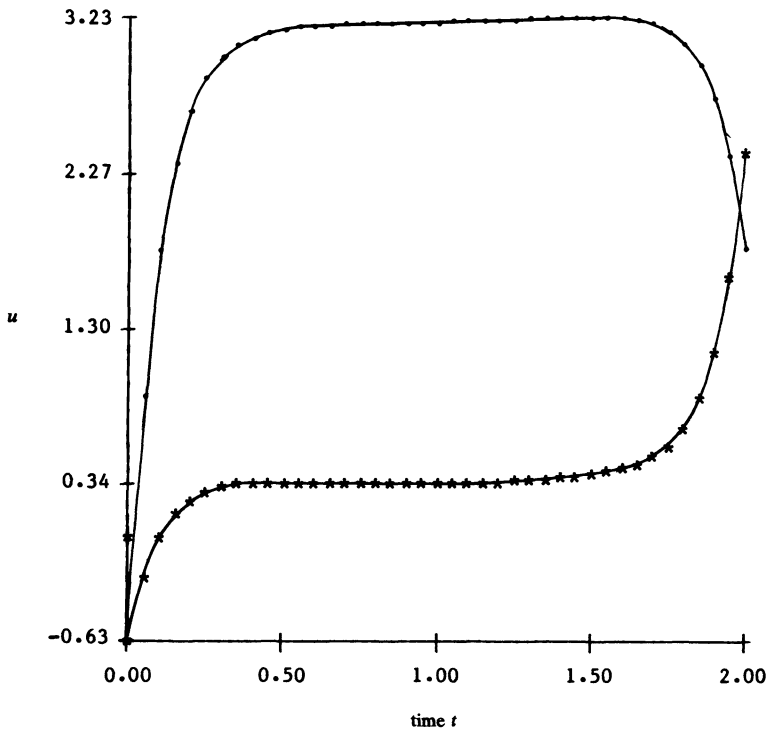


a. *Control* $u(t)$ *vs. t.* $u_1 \cdots$; u_2***.

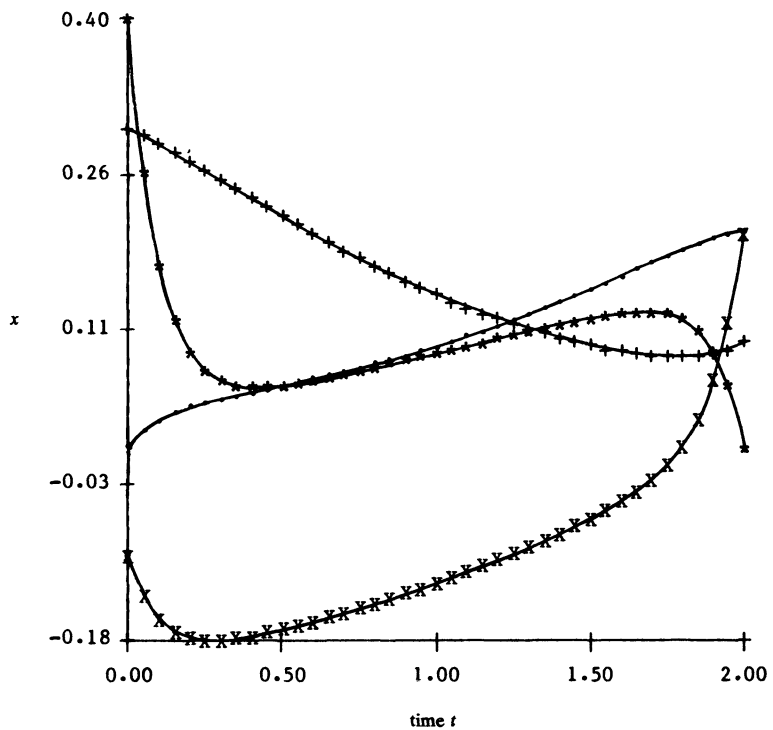


b. *State, x vs. t.* $x_1 \cdots$; x_2***; x_3+++; x_4*XXX*.

FIG. 3



a. *Control* $u(t)$ vs. *t.* $u_1 \cdots$; u_2***



b. *State, x vs. t.* $x_1 \cdots$; x_2***; x_3+++; x_4*XXX*.

FIG. 4

REFERENCES

[1] V. BARBU, *Nonlinear Semigroups and Differential Equations in Banach Spaces*, Noordhoff, Leyden, the Netherlands, 1976.

[2] B. R. BARMISH AND W. E. SCHMITENDORF, *Controlling a system to a target—Part 2: nonlinear system with a general target*, in Abstracts of papers presented at Canadian Optimization Days, May 1979, McGill University, Montreal.

[3] G. CHEN, *Energy decay estimate and exact boundary value controllability for the wave equation in a bounded domain*, J. Math. Pures Appl., 58 (1979), pp. 249–273.

[4] ———, *Control and stabilization for the wave equation in a bounded domain*, this Journal, 17 (1979), pp. 66–81.

[5] G. CHEN AND W. H. MILLS, *Penalization and regularization of quadratic cost controllability problems in a finite dimensional space*, INRIA Research Report #10, Rocquencourt, France, March 1980.

[6] W. C. CHEWNING, *Controllability of the nonlinear wave equation in several space variables*, this Journal, 14 (1976), pp. 19–25.

[7] E. N. CHUKWU, *Finite time controllability of nonlinear control processes*, this Journal, 13 (1975), pp. 807–816.

[8] M. A. CIRINA, *Boundary controllability of nonlinear hyperbolic systems*, this Journal, 7 (1969), pp. 198–212.

[9] M. G. CRANDALL, *Lecture notes in nonlinear functional analysis*, University of Wisconsin, spring 1977 (unpublished).

[10] D. L. ELLIOT, *A consequence of controllability*, J. Differential Equations, 10 (1974), pp. 364–370.

[11] H. O. FATTORINI, *Local controllability of a nonlinear wave equation*, Math. Systems Theory, 9 (1975), pp. 30–45.

[12] J. HENRY, *Thése d'etat*, Paris, June 1978.

[13] *International Mathematical and Statistical Library of Subprograms*, Library 1 Reference Manual, 2 Vol., IMSL Inc., Houston, TX, 1977.

[14] J. L. LIONS, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod-Gauthier-Villars, Paris, 1969.

[15] C. LOBRY, lecture notes (unpublished).

[16] D. L. LUKES, *Global controllability of nonlinear systems*, SIAM J. Control, 10 (1972), pp. 112–126.

[17] J. M. ORTEGA, *Numerical Analysis*, Academic Press, New York, 1972.

[18] A. PAZY, *Semigroup of Linear Operators and Applications to Partial Differential Equations*, Math. Dept. Lecture Notes, Vol. 10, Univ. Maryland, College Park, 1974.

[19] J. QUINN AND D. L. RUSSELL, *Asymptotic stability and energy decay rates for solutions of hyperbolic equations with boundary damping*, Proc. Royal Society Edinburgh, Ser. A., 77 (1977/78), pp. 97–127.

[20] D. L. RUSSELL, *Exact boundary value controllability theorems for wave and heat processes in star-complemented regions*, in Differential Games and Control Theory, Roxin, Liu, Sternberg, eds., Marcel Dekker, New York, 1974.

[21] A. S. C. SINHA, *Theory of cone methods for null-controllability of nonlinear control systems*, Proc. of the Fifteenth Annual Allerton Conference, Sept. 1977, pp. 114–119.

[22] W. A. STRAUSS, *The energy method in nonlinear partial differential equations*, Instituto de Mathematica Pura e Applicada, Brazil, 1969.

[23] L. TARTAR, *Evolution equations in infinite dimensions*, MRC Technical Summary Report #1485, University of Wisconsin, Madison, December, 1974.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 6, November 1981

0363-0129/81/1906-0005 $01.00/0

PARAMETER ESTIMATION AND IDENTIFICATION FOR SYSTEMS WITH DELAYS*

H. T. BANKS,† J. A. BURNS‡ AND E. M. CLIFF§

Abstract. Parameter identification problems for delay systems motivated by examples from aerodynamics and biochemistry are considered. The problem of estimation of the delays is included. Using approximation results from semigroup theory, a class of theoretical approximation schemes is developed and two specific cases ("averaging" and "spline" methods) are shown to be included in this treatment. Convergence results, error estimates, and a sample of numerical findings are given.

1. Introduction. The estimation of parameters in dynamical systems is an important scientific problem on which a number of contributions have been made in the engineering and mathematical literature (e.g., see [1], [23]). However, for systems with delays very little on identification is found in the engineering literature and essentially no theoretical convergence results are available for algorithms dealing with estimation of the delays themselves. One obvious difficulty (from both a practical and theoretical viewpoint) with such procedures is that solutions of delay systems are not in general differentiable with respect to the delays, and thus many common identification techniques (e.g., least squares gradient, maximum likelihood estimator, etc.) are not directly applicable.

In this paper we discuss a class of methods based on general approximation techniques for systems with delays. These approximation ideas have been considered earlier in the context of optimal control problems ([3], [4], [5], [6], [7], [9], [12], [18]), where they have proved quite useful. The use of such approximation ideas in connection with parameter estimation procedures was apparently first suggested in [11], and some preliminary theoretical results were stated in [7] and [13]. However, our presentation here is the first (to our knowledge) rigorous treatment of general theoretical aspects of these ideas.

While we do in § 7 below give a small sample of related numerical findings, the primary purpose of this paper is to present a theoretical foundation for the schemes we propose. A much more extensive discussion and a wider selection of numerical examples is presented in [8]. Our sample of numerical results in § 7 is included mainly to indicate that the procedures based on the schemes discussed are actually feasible.

The approximation ideas developed earlier in [5] and employed here are based on approximation results (the so-called Trotter–Kato theorem) from linear semigroup theory. In § 2 we formulate a class of identification problems for delay systems and show that they can be reformulated in an abstract setting so as to make use of the semigroup approximation theorem. A version of the Trotter–Kato results needed is given in § 3, while in § 4 we show how to use this theorem to insure convergence for a class of

* Received by the editors March 14, 1980, and in revised form September 26, 1980.

† Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. The research of this author was supported in part by the National Science Foundation under grant NSF-MCS79-05774, in part by the Air Force Office of Scientific Research under AF-AFOSR 76-3092C and in part by the U.S. Army Research Office under ARO-DAAG29-79-C-0161.

‡ Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061. The research of this author was supported in part by the U.S. Army Research Office under ARO-DAAG-29-78-G-0125.

§ Aerospace and Ocean Engineering Department, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061. The research of this author was supported in part by the U.S. Army Research Office under grant ARO-DAAG-29-78-G-0125.

identification schemes. We turn to the detailed development of particular schemes based on "averaging" (see [5]) and "spline" (see [10]) approximations in the subsequent two sections. Finally, a brief indication of numerical findings for these two particular schemes is given in § 7.

Notation used throughout the paper is completely standard. For example, $L_p^m(a, b) = L_p([a, b], R^m)$ denotes the usual Lebesgue spaces of R^m-valued "functions" on $[a, b]$ whose components are integrable when raised to the pth power. When $m = 1$, we shall suppress its appearance in the notation. $L_{p,\text{loc}}$ denotes the usually "locally" integrable function spaces. We shall use the symbol $|\cdot|$ to denote the norm of an element without distinguishing between different norms if the intended meaning is clear from the context. The space of functions with j continuous derivatives is denoted by $C^j(a, b)$. We shall also make use of the Sobolev spaces $W_p^{(j)}(a, b) = W_p^{(j)}([a, b], R^n)$ of R^n-valued absolutely continuous functions possessing $j-1$ absolutely continuous derivatives and jth derivatives that are in L_p.

In the remaining paragraphs of this introductory section, we turn to a discussion of examples which motivate the theoretical questions that are the focus of our attention in this paper.

1.1. Tubular reactor columns and delay system identification and control problems. Packed bed tubular enzyme reactors are very important in many areas of industrial and biological applications (potential uses involve purification or clarification of fruit juices, proteolytic treatment of beer, synthesis of essential amino acids, enzymatic biosynthesis—i.e., synthesis of antibiotics and steroids, etc.). These are column reactors (as depicted in Fig. 1.1) containing enzyme pellets (i.e., pellets in which an enzyme is insolubly bound), the enzyme being specific for a substrate S which is passed through the column. The substrate diffuses into the pellets where the enzyme catalyzes a reaction resulting in the product P.

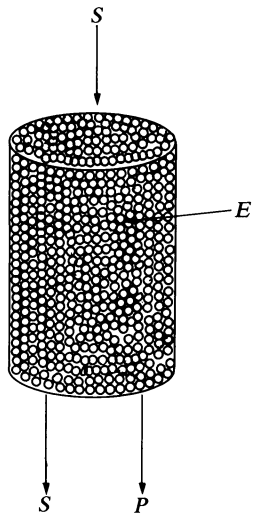


FIG. 1.1

We thus have enzymatically active particles or pellets in a convective flow region. Any model should embody important features of the system including (i) enzyme catalyzed reaction, (ii) metabolite (S or P) diffusion into, out of and inside of the pellets, and (iii) metabolite convection (and possibly diffusion) in the flow region in the column

exterior to the pellets. Extensive studies for both plug-flow (PF) models and diffusion-convection-reaction (DCR) models for the phenomena involved have been reported in the literature [15], [16]. These models can be formulated from first principles using transport equations of the form

$$\frac{\partial s}{\partial t}+c\frac{\partial s}{\partial x}=D_1\frac{\partial^2 s}{\partial x^2}+D_2\frac{\partial^2 s}{\partial y^2}+V,$$

where D_1, D_2 are diffusion coefficients, c is the convective flow velocity, x is the column axial direction, y the perpendicular direction (in a two-dimensional model), and V is a nonlinear reaction velocity approximation (e.g., $V=-\rho s/(1+s)$). The column in this case is approximated by a two-compartment (pellet phase and liquid phase) model as shown in Fig. 1.2.

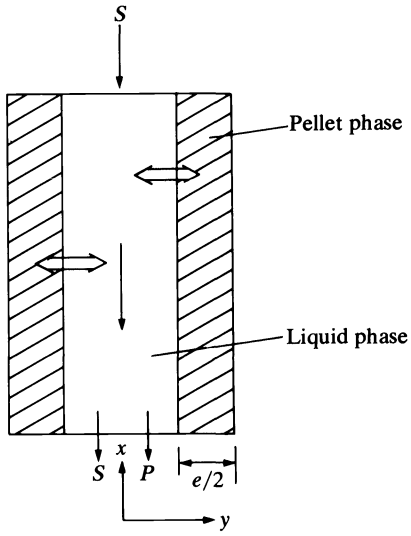


FIG. 1.2

PF models incorporate assumptions that one may ignore diffusion in both the pellet and solution (flow) regions. Careful investigation of these models reveal that they are of limited use in actual applications since it is found that certain kinetic *constants* must actually be allowed to vary (in an unpredictable manner) with the flow velocity in order to fit the models to experimental data. On the other hand, the DCR models were found to perform quite adequately when compared with the data. The main difficulty in employing the DCR models involves the rather lengthy calculations that must be made in carrying out identification and control procedures with these models. It is, therefore, desirable to have a model which in complexity and accuracy (hopefully similar to the DCR models with respect to the latter) is somewhere between the PF and DCR models and for which efficient numerical procedures are available.

A candidate for such a model has been proposed by J. P. Kernevez and his colleagues at Université de Technologie de Compiègne. It consists of n functional compartments for the column, each containing two subcompartments, one representing the pellet phase and the other the liquid phase. The two subcompartments in each compartment are connected by diffusion while the main compartments are connected via unidirectional transport (convective flow) between the liquid phase subcompartments (see Fig. 1.3).

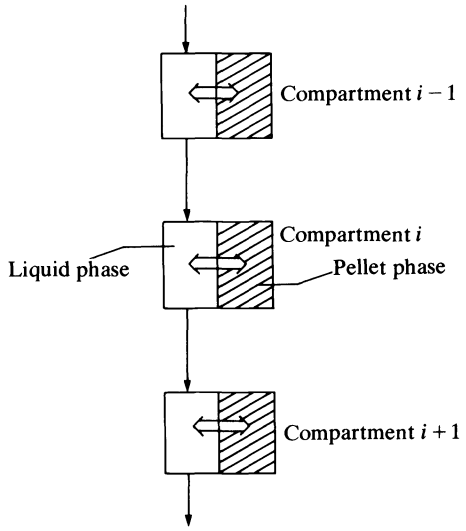


FIG. 1.3

Defining variables as follows (all are scaled and dimensionless):
$r_i(t)$ = substrate concentration in liquid phase in compartment i at time t,
$s_i(t)$ = substrate concentration in pellet phase in compartment i at time t,
$p_i(t)$ = product concentration in liquid phase in compartment i at time t,
$q_i(t)$ = product concentration in pellet phase in compartment i at time t,
one can write mass balance equations to obtain a model

$$\frac{dr_1}{dt} = -\alpha r_1(t) - \beta\{r_1(t) - s_1(t-\tau_1)\} + u(t),$$

$$\frac{dr_i}{dt} = \alpha\{r_{i-1}(t-\tau) - r_i(t)\} - \beta\{r_i(t) - s_i(t-\tau_1)\}, \qquad i > 1,$$

$$\frac{ds_i}{dt} = -\rho F(s_i(t)) + \nu\beta\{r_i(t) - s_i(t-\tau_1)\}, \qquad i \geqq 1,$$

$$\frac{dp_1}{dt} = -\alpha p_1(t) - \tilde{\beta}\{p_1(t) - q_1(t-\tau_2)\},$$

$$\frac{dp_i}{dt} = \alpha\{p_{i-1}(t-\tau) - p_i(t)\} - \tilde{\beta}\{p_i(t) - q_i(t-\tau_2)\}, \qquad i > 1,$$

$$\frac{dq_i}{dt} = \rho F(s_i(t)) + \nu\tilde{\beta}\{p_i(t) - q_i(t-\tau_2)\}, \qquad i \geqq 1.$$

Here F is a nonlinear reaction velocity term; the delays τ, τ_1, τ_2 are transport times between compartments $i-1$ and i, between pellet interior and liquid region for substrate, and between pellet interior and liquid region for product, respectively. The term u represents input of substrate to the liquid subcompartment of compartment 1. The parameters $\alpha, \beta, \rho, \nu, \tilde{\beta}$ are all related to biochemical and physical constants for the column configuration. For example, $\beta \cong ND_s\Sigma/eV$ where N is the "apparent" number of pellets per compartment, V = volume of liquid per compartment, D_s = coefficient of

diffusion for S within the pellet region, e = "thickness" of the model pellet region, and Σ = effective surface area per pellet.

Using data collected from a number of specific experiments performed with tracer, product and substrate inputs, one wishes to determine values of τ, τ_1, τ_2, ρ, β, $\tilde{\beta}$ so that the model describes accurately the operation of the column. Once this is done, the model then must be used to design (optimal) control procedures for the column.

We thus have classical identification and control problems for systems (let $x_i = (r_i, s_i, p_i, q_i)^T$)

$$\dot{x}_i(t) = A_0(\gamma)x_i(t) + A_1(\gamma)x_i(t-\tau_1) + A_2(\gamma)x_i(t-\tau_2) + f(\gamma, x_{i-1}(t-\tau), x_i(t), u(t)),$$

where the delays τ, τ_1, τ_2 and the vector parameter γ (involving only coefficients) are to be identified.

1.2. Identification problems for hereditary systems in unsteady aerodynamics. We consider next an interesting class of identification problems which arise in the study of unsteady aerodynamics (see [2]). Consider a thin, flat airfoil mounted on springs as shown in Fig. 1.4 in a region where we have fluid (air) flow with undisturbed stream velocity U (in the x-direction). Flow around the airfoil is disturbed and we

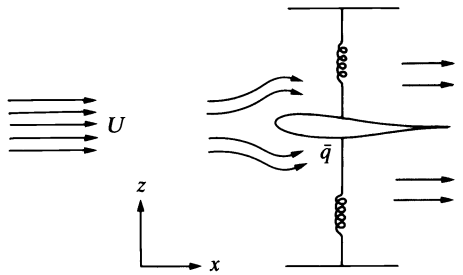


FIG. 1.4

assume it has velocity $\bar{q} = (u, w)$. Laws of conservation of mass and momentum lead to a system of partial differential equations for the fluid velocity components u and w. Assuming incompressible flow we have the continuity equation $\nabla \cdot \bar{q} = 0$. Elementary hydrodynamics also yield that curl $(\bar{q}) = 0$, from which we deduce the existence of a velocity potential φ so that $\bar{q} = \nabla\varphi$. The equation of continuity then becomes $\Delta\varphi = 0$. We restrict our considerations to small motions of the airfoil so that a linearized theory may be adopted. We assume that φ is given by

$$\varphi(x, z, t) = \tilde{\varphi}(x, z, t) + Ux,$$

where $\tilde{\varphi}$ is a disturbance potential. It follows that $\tilde{\varphi}$ must satisfy

$$\Delta\tilde{\varphi} = 0. \tag{1.1}$$

In addition one has the (flow tangency) boundary conditions

$$\frac{\partial\tilde{\varphi}}{\partial z}(x, 0, t) = w(x, 0, t) = w_a(x, t), \qquad -1 \leqq x \leqq 1, \tag{1.2}$$

where w_a is a given function describing the motion of the airfoil. We here assume that the airfoil is a thin plate located at $z = 0$, $-1 \leqq x \leqq 1$ as depicted in Fig. 1.5.

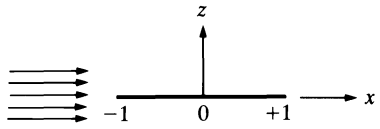


FIG. 1.5

After arguments involving a conformal mapping of the airfoil into the unit circle and the introduction of sources (elementary flows along radial lines) and vortices (elementary flows along concentric circles), one finds that a solution of (1.1), (1.2) for the disturbance potential $\tilde{\varphi}$ consists of an appropriate collection of sources distributed along the airfoil and any weighted combination of "compatible" vortex pairs. A "compatible" pair consists of one vortex on the airfoil at $r = r_1 < 1$ and an oppositely rotating one at $r = 1/r_1 > 1$. Compatible pairs induce a flow with finite angular momentum and with fluid velocity that is tangent to the airfoil. The required distribution of sources is uniquely defined by the airfoil motion ($w_a(x, t)$ in equation (1.2)) but the distribution of vortices in the wake given by a density function $\gamma_w(\xi, t)$ is as yet unknown. In lieu of γ_w we introduce a new function Γ termed the circulation. For brevity we shall "define" Γ by

$$\dot{\Gamma}\left(t - \frac{(\xi - 1)}{U}\right) = -\gamma_w(\xi, t), \tag{1.3}$$

with the boundary condition $\Gamma(-\infty) = 0$. This relationship reveals that vorticity in the wake at time t and position ξ was produced by a change in the circulation at an earlier time, i.e., an hereditary phenomenon is involved. In integrated form (using $\Gamma(-\infty) = 0$) this becomes

$$\Gamma(t) - \int_1^\infty \gamma_w(\xi, t)\, d\xi = 0. \tag{1.4}$$

To determine Γ (or γ_w) we impose an additional hypothesis, viz., finiteness of the fluid velocity at the trailing edge of the airfoil. Mathematically this is written

$$v(t) + \int_1^\infty \sqrt{\frac{\xi + 1}{\xi - 1}}\, \gamma_w(\xi, t)\, d\xi = 0. \tag{1.5}$$

Here v is the contribution to the velocity due to the source distribution. Subtracting (1.5) from (1.4) and using (1.3) we thus obtain

$$\Gamma(t) = v(t) + \int_1^\infty f(\xi)\dot{\Gamma}\left(t - \frac{(\xi - 1)}{U}\right) d\xi, \tag{1.6}$$

where $f(\xi) = \sqrt{(\xi + 1)/(\xi - 1)} - 1$. This finally is our model equation (see [28, p. 292]), the basis of hereditary models in unsteady aerodynamics.

A simple change of variables $\xi = 1 - \sigma$ in the integral in (1.6) yields the equation

$$\Gamma(t) = v(t) + \int_{-\infty}^0 \tilde{f}(\sigma)\dot{\Gamma}\left(t + \frac{\sigma}{U}\right) d\sigma, \tag{1.7}$$

where $\tilde{f}(\sigma) \equiv f(1 - \sigma)$. This is essentially a neutral functional differential equation with infinite memory. Among the numerous approximations made in the derivation of such a model is the expression for f,

$$f(\xi) = \sqrt{(\xi + 1)/(\xi - 1)} - 1. \tag{1.8}$$

It turns out that the transverse velocity component w exhibits a boundary layer phenomenon as sketched in Fig. 1.6, where r is the horizontal distance from the trailing edge of the airfoil. Thus, the expression for f in (1.8) is valid only for $\xi \gg 1$. To better

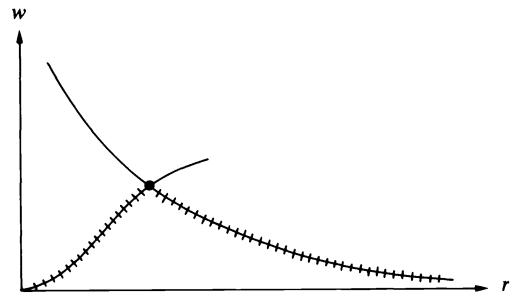


FIG. 1.6

approximate this phenomenon in (1.7) one might approximate $\tilde{f}$ by a function g having the form

$$g(\sigma;\alpha,\beta,\mu)=\begin{cases}\alpha\sigma, & -\mu\leqq\sigma\leqq 0,\\ \beta\tilde{f}(\sigma), & -\infty<\sigma<-\mu,\end{cases}$$

where it is understood that α, β must be chosen so that g is continuous at $\sigma=-\mu$. The model then is given by

$$\Gamma(t)=v(t)+\int_{-\infty}^{0} g(\sigma;\alpha,\beta,\mu)\dot{\Gamma}\left(t+\frac{\sigma}{U}\right)d\sigma. \tag{1.9}$$

Assuming smoothness of g, we formally integrate by parts the integral in (1.9) to obtain

$$\Gamma(t)=v(t)-\int_{-\infty}^{0}\dot{g}(\sigma)\Gamma\left(t+\frac{\sigma}{U}\right)d\sigma,$$

or, letting $s=t+\sigma/U$ in the integral and defining $G(\xi)=\dot{g}(U\xi)$, we have

$$\Gamma(t)=v(t)-\int_{-\infty}^{t} G(s-t)\Gamma(s)\,ds. \tag{1.10}$$

Equation (1.10) is a retarded FDE with infinite memory which, upon differentiation, yields the more familiar form

$$\dot{\Gamma}(t)=\dot{v}(t)-G(0)\Gamma(t)+\int_{-\infty}^{t}\dot{G}(s-t)\Gamma(s)\,ds.$$

In practice, one would often desire to replace the integral term by a finite integral

$$\int_{-\tau}^{t}\dot{G}(s-t)\Gamma(s)\,ds,$$

in which case one obtains an equation (taking $\Gamma(t)=x(t)$, observing that $G(0)=\dot{g}(0)=\alpha$ and identifying $\dot{v}(t)=u(t)$ as the input)

$$\dot{x}(t)=-\alpha x(t)+\int_{-\tau}^{t} U\ddot{g}(U[s-t];\alpha,\beta,\mu)x(s)\,ds+u(t). \tag{1.11}$$

An important identification problem then consists of making observations corresponding to an input $u(t) = \dot{v}(t)$ and using these to estimate the parameters α, β, μ, and τ so that the model yields a sufficiently accurate description of the aerodynamic phenomena under investigation.

2. The fundamental identification problem for delay systems. We consider in this paper n-vector systems of the form

$$\dot{x}(t) = L(q)x_t + B(\alpha)u(t), \qquad t \geqq 0, \tag{2.1}$$

with initial data

$$x(0) = \eta, \quad x_0 = \phi, \quad (\eta, \phi) \in R^n \times L_2^n(-r, 0) \tag{2.2}$$

and output

$$y(t) = C(\alpha)x(t) + D(\alpha)u(t). \tag{2.3}$$

We make the following definitions and assumptions about the operators and parameters in (2.1)–(2.3). There exists a fixed given $r > 0$ and compact convex set $\Omega \subset R^{\mu}$, and we define the compact convex set $Q \subset R^{\mu+\nu}$ by $Q \equiv \Omega \times \mathcal{H}$, where

$$\mathcal{H} = \{h = (r_1, r_2, \cdots, r_\nu) \in R^\nu \mid 0 \leqq r_i \leqq r_{i+1} \leqq r, i = 1, \cdots, \nu - 1\}.$$

For a function x we adopt the usual notation $x_t(\theta) = x(t + \theta)$. For a given element $q = (\alpha, h)$ in the *admissible parameter* set Q, we define the operators $L(q)$: $L_2^n(-r, 0) \to R^n$ of (2.1) by

$$L(q)\phi = \sum_{i=0}^{\nu} A_i(\alpha)\phi(-r_i) + \int_{-r_\nu}^{0} K(\alpha, \theta)\phi(\theta)\, d\theta, \tag{2.4}$$

where $r_0 \equiv 0$, and for each $\alpha \in \Omega$, $A_i(\alpha)$, $B(\alpha)$, $C(\alpha)$ and $D(\alpha)$ are $n \times n$, $n \times m$, $k \times n$ and $k \times m$ matrices respectively. We assume that the $n \times n$ matrix-valued function $\theta \to K(\alpha, \theta)$ is in $L_2(-r, 0)$, and that the functions A_i, B, C, D, $K(\ ,\cdot)$ are continuous in α.

Remark 1. In (2.4) one must give the proper interpretation to point evaluations in the event ϕ is only an L_2 "function". Since in (2.1) we are interested in integrals of the system, the usual interpretation is intended here (see [10] for a more detailed discussion).

We further assume that we are given an initial data set $\mathcal{S} \subset R^n \times L_2^n(-r, 0)$ which is closed, bounded and convex, and we define

$$\Gamma \equiv \mathcal{S} \times Q = \mathcal{S} \times \Omega \times \mathcal{H}$$

as our admissible initial data–parameter set. Elements γ in Γ will be denoted in one of several ways throughout our discussions below:

$$\gamma = (\eta, \phi, q) = (\eta, \phi, \alpha, h) = (\eta, \phi, \alpha, r_1, \cdots, r_\nu),$$

where $q = (\alpha, h) = (\alpha, r_1, \cdots, r_\nu)$. For each $\gamma = (\eta, \phi, q)$ in Γ we shall denote the output to (2.1)–(2.3) at time $t \geqq 0$ by $y = y(t; \gamma)$.

Identification of the system variables γ in (2.1)–(2.3) is based on input-output information. Given a piecewise continuous control input u defined on some time interval $[0, T]$, one samples the system at times $\{t_i\}$, $0 \leqq t_1 < t_2 < \cdots < t_M \leqq T$, to obtain observations $\{\hat{y}_i\}$, $\hat{y}_i \in R^k$, $i = 1, 2, \cdots, M$. One can then perform a least squares fit to data (or seek a maximum likelihood estimator for γ). Formally, we may state this as follows:

Problem. Given the input u and observations $\{\hat{y}_i\}$ at times $\{t_i\}$, find $\gamma^* = (\eta^*, \phi^*, q^*)$ in Γ which minimizes the fit error

$$J(\gamma) = \frac{1}{2}\sum_{i=1}^{M} |y(t_i; \gamma) - \hat{y}_i|^2. \tag{2.5}$$

Remark 2. Whenever $r_\nu^* < r$, one only needs ϕ^* defined on $[-r_\nu^*, 0]$ in order to obtain a solution to (2.1)–(2.3) (in practice, this is exactly what we shall obtain). However, we can view (η^*, ϕ^*) as an element of $\mathcal{S}$ by making a simple (arbitrary but definite) backward extension of ϕ^* to all of $[-r, 0]$.

2.1. An abstract formulation of the I.D. problem. Let $r > 0$ be fixed and given as in the previous section and define $Z \equiv R^n \times L_2^n(-r, 0)$. For $q = (\alpha, h) \in Q$ and $(\eta, \phi) \in Z$, define for $t \geqq 0$ the mappings $S(t; q): Z \to Z$ by

$$S(t; q)(\eta, \phi) = (x(t; \gamma), x_t(\gamma)),$$

where x is the solution to (2.1) with $u \equiv 0$ and $x_t(\theta) = x(t+\theta)$, $-r \leqq \theta \leqq 0$. It is easily verified that for each q, $\{S(t; q)\}_{t \geqq 0}$ is a strongly continuous semigroup of linear operators on Z. Furthermore, one finds [5] that the infinitesimal generator $\mathcal{A}(q)$, with domain

$$\mathcal{D}(\mathcal{A}(q)) = \mathcal{D} = \{(\eta, \phi) \in Z \mid \phi \in W_2^{(1)}(-r, 0), \eta = \phi(0)\},$$

is given by

$$\mathcal{A}(q)(\phi(0), \phi) = (L(q)\phi, \dot{\phi}).$$

We note that, for $q \in Q$, $\mathcal{D}(\mathcal{A}(q))$ does not depend on q itself. However, for $k > 1$, $\mathcal{D}(\mathcal{A}^k(q))$ *does* depend on q. For example, $\mathcal{D}(\mathcal{A}^2(q)) = \{(\phi(0), \phi) \mid \phi \in W_2^{(2)}(-r, 0), \dot{\phi}(0) = L(q)\phi\}$.

If we define the operators $\hat{B}(\alpha): R^m \to Z$ and $\hat{C}(\alpha): Z \to R^k$ by $\hat{B}(\alpha)u = (B(\alpha)u, 0)$ and $\hat{C}(\alpha)(\eta, \phi) = C(\alpha)\eta$, then the delay system (2.1)–(2.3) is formally equivalent to the abstract ordinary differential equation (ODE) system

$$\dot{z}(t) = \mathcal{A}(q)z(t) + \hat{B}(\alpha)u(t), \qquad t \geqq 0, \tag{2.6}$$

$$z(0) = (\eta, \phi), \tag{2.7}$$

$$y(t) = \hat{C}(\alpha)z(t) + D(\alpha)u(t). \tag{2.8}$$

As in the usual theory dealing with semigroups and abstract differential equations, a mild solution to (2.6)–(2.8) can be given by a variation of parameters formula. Specifically, (2.6)–(2.7) has the mild solution $z(t) = z(t; \gamma, u)$ given by

$$z(t) = S(t; q)(\eta, \phi) + \int_0^t S(t-\sigma; q)\hat{B}(\alpha)u(\sigma)\, d\sigma. \tag{2.9}$$

It is a happy circumstance that (2.9) is actually equivalent to (2.1)–(2.2) in a strong sense, as we now state precisely (for proof see [4], [5] or [6]).

THEOREM 2.1. *Let $x(\cdot; \gamma, u)$ denote the solution to* (2.1)–(2.2) *corresponding to $\gamma \in Z \times Q$ and $u \in L_{2,\mathrm{loc}}$. Then, for all $t \geqq 0$,*

$$z(t; \gamma, u) = (x(t; \gamma, u), x_t(\gamma, u)).$$

In view of the above equivalence results, the I.D. problem for (2.1)–(2.3) posed above can be reformulated in terms of an abstract I.D. problem. That is, given input u and observations $\{\hat{y}_i\}$ at times $\{t_i\}$, find $\gamma^* = (\eta^*, \phi^*, q^*)$ in Γ so as to minimize $J(\gamma)$ as

given in (2.5), where now $y(t)$ is given by (2.8) and (2.9) in place of (2.3). Whether the problem is formulated in terms of (2.8), (2.9) or (2.1)–(2.3), it is clear that we are dealing with I.D. problems involving infinite dimensional state systems. Formulation in the framework of the Hilbert space Z only emphasizes this, and is in no way an essential factor in the infinite dimensionality (and the associated difficulties) of the problem. Our main interests here are identification schemes that will result in computationally efficient algorithms. The approach we take is a classical one of the Ritz type. We shall choose a sequence of finite dimensional problems, each of which is defined on a finite dimensional state space X_N and approximates the original I.D. problem in Z. By appropriate choices of the sequence $\{X_N\}$ and the corresponding approximating problems, we hope to obtain a sequence of more easily solved problems with solutions $\gamma^N = (\eta^N, \phi^N, q^N)$ which converge to a solution γ^* of our original problem.

Fundamental to this endeavor is the convergence of the underlying approximating systems to the original system (2.9). Our formulation in a functional analytic framework will allow us to utilize abstract approximation theorems from semigroup theory (e.g., see [5]). The problems here, however, are a little different from the control problems of [5] where one chooses a sequence of *subspaces* $Z^N \subset Z$ on which to solve approximating control problems. The I.D. problems to be treated below pose some additional difficulties in that for each value of N, the "state" space changes. That is, the natural space for (2.9) with $q^N = (\alpha^N, r_1^N, \cdots, r_\nu^N)$ is $Z_N = R^n \times L_2^n(-r_\nu^N, 0)$ which, in addition to varying with N, is *not* a subspace of the original space $Z = R^n \times L_2^n(-r, 0)$. The approximating spaces X_N clearly should be chosen so that $X_N \subset Z_N$.

There are abstract approximation theorems (motivated by differencing schemes for partial differential equations and applications from probability theory) available in the literature in the case where $Z_N \not\subset Z$. For example, the original Lax, Trotter, Kato efforts [20], [26], [17] resulted in such theorems as did the later efforts of Kurtz [19]. However, all of these versions of the approximation results (and all others with which we are familiar) require the spaces X_N to approximate Z in the sense that there exist projection-like mappings $P_N : Z \to X_N$ which satisfy a norm convergence criterion $|P_N z|_{X_N} \to |z|_Z$ as $N \to \infty$. For the problems and approximations we shall discuss below such a criterion is not met (in general, one will *not* have $r_\nu^N \to r$, where r is the a priori chosen upper bound for the hereditary effects in the systems). We shall, therefore be obligated to state and prove an appropriate version of the abstract approximation results and this is done in the next section. The arguments used to establish this theorem are very similar to the standard ones found in the literature. One has a sequence of approximating infinitesimal generators (i.g.'s) A_N which converge in some sense to an i.g. A. This convergence is sufficient to imply convergence of the resolvents $R_\lambda(A_N)$ to $R_\lambda(A)$. These are the Laplace transforms of the corresponding semigroups $S_N(t)$, $S(t)$ respectively and their convergence is enough to guarantee the desired convergence $S_N(t) \to S(t)$. We make this more precise in the next section.

3. An abstract approximation theorem. Let Z and Z_N, $N = 1, 2, \cdots$, be Hilbert spaces with norms $|\cdot|$ and $|\cdot|_N$ respectively. Let X_N be a closed subspace of Z_N and $\pi_N : Z_N \to X_N$ be the canonical projection of Z_N onto X_N along $X_N^\perp$. Suppose $\mathcal{J}_N : Z \to Z_N$ is a mapping satisfying $\operatorname{Im}(\mathcal{J}_N) = Z_N$ and $|\mathcal{J}_N z|_N \leqq |z|$ for $z \in Z$. Finally define $P_N : Z \to X_N$ by $P_N = \pi_N \mathcal{J}_N$. (In our discussions for the I.D. problem above, $Z = R^n \times L_2^n(-r, 0)$, $Z_N = R^n \times L_2^n(-r_\nu^N, 0)$, X_N is an approximating space such as the AVE spaces of [5] or the spline spaces of [10]—these will be discussed fully below. Finally, $\mathcal{J}_N$ is the operator that takes $z = (\eta, \phi)$ in Z into $\tilde{z} = (\eta, \tilde{\phi})$ where $\tilde{\phi}$ is the restriction of ϕ to $[-r_\nu^N, 0]$. We note that in this case we would not expect to have $|P_N z|_N \to |z|$ for $z \in Z$ unless $r_\nu^N \to r$ and π_N itself has certain convergence properties.)

We adopt the following standard notation for the presentation of our fundamental approximation results. For a Hilbert space X, we write $B \in G(M, \beta)$ to mean $B : \mathscr{D}(B) \subset X \to X$ is the i.g. of a C_0-s.g. $\{T(t)\}$ satisfying $|T(t)| \leqq M e^{\beta t}$. We also denote the resolvent $(\lambda I - B)^{-1}$ by $R_\lambda(B)$ and recall that $R_\lambda(B)x = \int_0^\infty e^{-\lambda\sigma} T(\sigma)x \, d\sigma$.

THEOREM 3.1. *Let Z, Z_N, X_N, and P_N be given as above. Suppose for some M, β we have $A_N \in G(M, \beta)$ on X_N and $A \in G(M, \beta)$ on Z. Further suppose there exists $\mathscr{D} \subset \mathscr{D}(A)$, $\mathscr{D}$ dense in Z such that*

$$\text{(3.1)} \quad \begin{array}{ll} \text{(i)} & R_\lambda(A)\mathscr{D} \subset \mathscr{D} \text{ for } \operatorname{Re}\lambda > \beta, \\ \text{(ii)} & \text{for every } z \in \mathscr{D},\ |A_N P_N z - P_N A z|_N \to 0 \text{ as } N \to \infty. \end{array}$$

Then for every $z \in Z$

$$\text{(3.2)} \qquad |S_N(t)P_N z - P_N S(t) z|_N \to 0 \quad \text{as } N \to \infty,$$

and the convergence is uniform in t on compact intervals. Here A_N is the i.g. *for $S_N(t)$, A the* i.g. *for $S(t)$.*

Remark 3.1. Implicit in the statement and proof of Theorem 3.1 is the assumption that $P_N z \in \mathscr{D}(A_N)$ for every $z \in Z$. In our use of the theorem for I.D. schemes below, X_N will be finite dimensional and $\mathscr{D}(A_N) = X_N$. Indeed, we shall find A_N bounded with $S_N(t) = e^{A_N t}$.

Proof. Let λ be fixed throughout with $\operatorname{Re}\lambda > \beta$, so that $R_\lambda(A_N)$, $R_\lambda(A)$ exist. We first establish that for every $y \in Z$

$$\text{(3.3)} \qquad |R_\lambda(A_N)P_N y - P_N R_\lambda(A) y|_N \leqq \frac{M}{\operatorname{Re}\lambda - \beta} |(A_N P_N - P_N A) R_\lambda(A) y|_N.$$

From the definition of the resolvent operator we have for any operator B

$$R_\lambda(B)B = BR_\lambda(B) = \lambda R_\lambda(B) - I.$$

In particular,

$$R_\lambda(A_N)A_N P_N = \lambda R_\lambda(A_N)P_N - P_N,$$

$$P_N A R_\lambda(A) = \lambda P_N R_\lambda(A) - P_N,$$

so that

$$R_\lambda(A_N)A_N P_N R_\lambda(A) - R_\lambda(A_N)P_N A R_\lambda(A) = R_\lambda(A_N)P_N - P_N R_\lambda(A).$$

Hence for any $y \in Z$ we have

$$\begin{aligned} |R_\lambda(A_N)P_N y - P_N R_\lambda(A) y|_N &= |R_\lambda(A_N)[A_N P_N - P_N A] R_\lambda(A) y|_N \\ &\leqq \frac{M}{\operatorname{Re}\lambda - \beta} |(A_N P_N - P_N A) R_\lambda(A) y|_N, \end{aligned}$$

the last inequality following from the fact that $A_N \in G(M, \beta)$.

Next, for given $z \in \mathscr{D}$ where $\mathscr{D}$ is as in the hypotheses, define

$$F_N(\sigma) \equiv S_N(\sigma)P_N z - P_N S(\sigma) z.$$

Then from (i), (ii) and (3.3) we conclude that, for $\operatorname{Re}\lambda > \beta$,

$$|\mathscr{L}_\lambda[F_N]|_N \to 0 \quad \text{as } N \to \infty$$

where $\mathscr{L}_\lambda$ is the Laplace transform. We observe that from the bounds on S_N, P_N, S the sequence $\{F_N\}$ is uniformly exponentially bounded, i.e.,

$$|F_N(\sigma)| \leqq 2M e^{\beta\sigma} |z|.$$

Finally, since

$$\frac{d}{d\tau}P_N S(\tau)z = P_N S(\tau)Az$$

and

$$\frac{d}{d\tau}S_N(\tau)P_N z = S_N(\tau)A_N P_N z,$$

a simple quadrature reveals

$$F_N(\sigma) = \int_0^\sigma [S_N(\tau)A_N P_N z - P_N S(\tau)Az]\,d\tau, \tag{3.4}$$

and it follows that $\{F_N\}$ is a pointwise equicontinuous family on $[0, \infty)$. (From the convergence in (3.1)–(ii) and the bounds on S_N and S, one easily verifies that the integrand in (3.4) is uniformly exponentially bounded.) We are thus in a position to use a lemma due to Kurtz ([19, Lemma 2.11, p. 359]) to conclude that $|F_N(\sigma)|_N \to 0$ as $N \to \infty$, uniformly on compact intervals. (Actually, the lemma as stated by Kurtz requires uniform boundedness of $\{F_N\}$, but a careful inspection of his proof will convince the reader that this requirement can be replaced by uniform exponential boundedness as we have here.)

We thus obtain the desired convergence (3.2) at least for each $z \in \mathcal{D}$. But then standard density arguments (the triangle inequality, bounds for S_N, P_N, and the density of $\mathcal{D}$ in Z) can be employed to establish the convergence for all $z \in Z$.

Remark 3.2. We note that in the above theorem we could have hypothesized $A_N \in G(M, \beta)$ on Z_N instead of on X_N without altering the proof. However, in the applications we have in mind we wish to obtain invariancy of $S_N(t) = e^{A_N t}$ on X_N (the space where our approximating systems will be defined and used). Thus, if we posit $A_N \in G(M, \beta)$ on Z_N we must make the additional hypothesis Im $(A_N) \subset X_N \subset \mathcal{D}(A_N)$ in order to use the approximation result as we desire below.

Remark 3.3. One can clearly choose $X_N = Z_N$ (with π_N then the identity on Z_N or $Z_N = Z$ (and $\mathcal{J}_N$ the identity on Z) and obtain other versions of the approximation results. Again, our choice here is dictated by the application to be discussed below.

Remark 3.4. In the event one has $Z_N = X_N \subset Z$ and $P_N : Z \to Z_N$ satisfying $P_N z \to z$ for $z \in Z$, then the condition (3.1ii) can be replaced by $|A_N P_N z - Az| \to 0$ and the conclusion (3.2) by $|S_N(t)P_N z - S(t)z| \to 0$. This then is essentially the version of the approximation theorem that we employed in previous efforts dealing with control problems [4], [5].

COROLLARY 3.1. *Suppose the convergence in* (3.1ii) *is* $O(N^{-\delta})$ *whenever* z *has the form* $R_\lambda^2(A)y$, $S(t)R_\lambda(A)y$, *and* $S(t)R_\lambda^2(A)y$, $\lambda > \beta$, *for a given* $y \in Z$. *Suppose further that the constants in* $O(N^{-\delta})$ *are uniform in* t *in the latter two cases. Then the convergence in* (3.2) *is also* $O(N^{-\delta})$ *whenever* $z = R_\lambda^2(A)y$ *for this* y.

Proof. Using rather standard arguments [22, p. 87] one finds that

$$\frac{d}{d\sigma}[S_N(t-\sigma)R_\lambda(A_N)P_N S(\sigma)R_\lambda(A)x]$$

$$= S_N(t-\sigma)[P_N R_\lambda(A) - R_\lambda(A_N)P_N]S(\sigma)x$$

for arbitrary $x \in Z$ and $\lambda > \beta$. Hence, we have

$$\int_0^t S_N(t-\sigma)[P_N R_\lambda(A) - R_\lambda(A_N)P_N]S(\sigma)x\, d\sigma = R_\lambda(A_N)[P_N S(t) - S_N(t)P_N]R_\lambda(A)x. \tag{3.5}$$

Using (3.5) and (3.3) we have, for any $y \in Z$,

$$\begin{aligned}
&|[S_N(t)P_N - P_N S(t)]R_\lambda^2(A)y|_N \\
&\leqq |S_N(t)[P_N R_\lambda(A) - R_\lambda(A_N)P_N]R_\lambda(A)y|_N \\
&\quad + |R_\lambda(A_N)[S_N(t)P_N - P_N S(t)]R_\lambda(A)y|_N \\
&\quad + |[R_\lambda(A_N)P_N - P_N R_\lambda(A)]S(t)R_\lambda(A)y|_N \\
&\leqq M e^{\beta t}|[P_N R_\lambda(A) - R_\lambda(A_N)P_N]R_\lambda(A)y|_N \\
&\quad + \int_0^t |S_N(t-\sigma)||\{P_N R_\lambda(A) - R_\lambda(A_N)P_N\}S(\sigma)y|_N\, d\sigma \\
&\quad + |\{R_\lambda(A_N)P_N - P_N R_\lambda(A)\}S(t)R_\lambda(A)y|_N \\
&\leqq M e^{\beta t}\frac{M}{\lambda-\beta}|(P_N A - A_N P_N)R_\lambda^2(A)y|_N \\
&\quad + \int_0^t M e^{\beta(t-\sigma)}\frac{M}{\lambda-\beta}|(P_N A - A_N P_N)S(\sigma)R_\lambda(A)y|_N\, d\sigma \\
&\quad + \frac{M}{\lambda-\beta}|(P_N A - A_N P_N)S(t)R_\lambda^2(A)y|_N.
\end{aligned}$$

Thus, if y is chosen as in the hypothesis of the corollary, the conclusion follows immediately.

THEOREM 3.2. *Let $\mathcal{B} \subset \mathcal{D}(A^2)$ satisfy the following*:

(i) *For each $z \in \mathcal{B}$ there exists $k = k(z)$ such that*

$$|(A_N P_N - P_N A)z|_N \leqq \frac{k}{N^\delta}, \qquad N = 1, 2, \cdots.$$

(ii) *There exists $\mathcal{B}_1 \subset \mathcal{B}$ such that $z \in \mathcal{B}_1$ implies*

(a) $$S(t)z \in \mathcal{B}, \qquad 0 \leqq t \leqq T,$$

(b) $$S(t)(\lambda I - A)z \in \mathcal{B}, \qquad \lambda > \beta, \quad 0 \leqq t \leqq T,$$

and furthermore the constants guaranteed by (i) *for* (a), (b) *can be chosen independent of t. Then for $z \in \mathcal{B}_1$ we have that there exists $\tilde{k}(z)$ such that*

$$|[S_N(t)P_N - P_N S(t)]z|_N \leqq \frac{\tilde{k}(z)}{N^\delta}$$

for $0 \leqq t \leqq T$.

Proof. Let $z \in \mathcal{B}_1$ and define $y = (\lambda I - A)^2 z$. Then $R_\lambda^2(A)y = z$. Furthermore, by (ii) we have $S(t)R^2(A)y \in \mathcal{B}$ and $S(t)R_\lambda(A)y \in \mathcal{B}$ with the constants in the $O(N^{-\delta})$ estimates uniform in t. It follows that the hypotheses of Corollary 3.1 are satisfied and hence we reach the desired conclusion since $z = R_\lambda^2(A)y$.

4. Identification schemes for delay systems. Let π^0, π^1 be the coordinate projections of $Z = R^n \times L_2^n(-r, 0)$ onto R^n, $L_2^n(-r, 0)$ respectively. We recall that the I.D. problem of § 2 can be written

($\mathscr{P}$): Given input u and observations $\{\hat{y}_i\}$ at $\{t_i\}_{i=1}^M$, find $\gamma^* = (\eta^*, \phi^*, q^*)$ in Γ so as to minimize

$$J(\gamma) = \frac{1}{2} \sum_{i=1}^{M} |y(t_i; \gamma) - \hat{y}_i|^2$$

where y is the solution of (2.8), (2.9); that is,

$$\begin{aligned} y(t; \gamma) &= \hat{C}(\alpha) z(t; \gamma, u) + D(\alpha) u(t) \\ &= C(\alpha) \pi^0 z(t; \gamma, u) + D(\alpha) u(t) \\ &= C(\alpha) x(t; \gamma, u) + D(\alpha) u(t). \end{aligned}$$

Thus the identification problem can be viewed in a state-space Z, parameter-space Γ setting. This will lead to a sequence of approximate I.D. problems if we approximate Z by a sequence of spaces X_N.

Given approximation spaces $X_N(q)$ and semigroups $S_N(t)$ with i.g.'s $A_N(q) \in G(M, \beta)$ for $q \in Q$, let $P_N(q)\colon Z \to X_N(q)$ be the mappings as discussed in § 3. We next define

$$\Gamma_N = \bigcup_{q \in Q} (P_N(q)\mathscr{S} \times \{q\}).$$

For $\gamma^N = (z_0^N, q) \in \Gamma_N$, we then consider solutions $\tilde{z}^N(t; \gamma^N, u)$ of (2.9) with $S(t)(\eta, \phi)$ replaced by $S_N(t) z_0^N$ and $S(t-\sigma)$ replaced by $S_N(t-\sigma)P_N(q)$. The corresponding outputs are defined by $\tilde{y}^N(t; \gamma^N) = C(\alpha)\pi^0 \tilde{z}^N(t; \gamma^N, u) + D(\alpha)u(t)$. The approximate I.D. problems are:

($\mathscr{P}_N$): Given input u and observations $\{\hat{y}_i\}$, find $\bar{\gamma}^N \in \Gamma_N$ so as to minimize

$$J^N(\gamma^N) \equiv \sum_{i=1}^{M} |\tilde{y}^N(t_i; \gamma^N) - \hat{y}_i|^2.$$

Under reasonable and rather obvious continuity and compactness conditions (which will hold for the specific cases to be discussed subsequently under the assumptions invoked below), it is not difficult to establish existence of a solution $\bar{\gamma}^N = (\bar{z}_0^N, \bar{q}^N) = (\bar{z}_0^N, \bar{\alpha}^N, \bar{r}_1^N, \cdots, \bar{r}_\nu^N)$ to ($\mathscr{P}_N$).

Given a sequence $\{r_\nu^N\}$, $0 \leqq r_\nu^N \leqq r$, and closed $X_N \subset Z_N \equiv R^n \times L_2^n(-r_\omega^N, 0)$, we define the operator $\mathscr{J}_N\colon Z \to Z_N$ as the operator that truncates $\pi^1 z$ to the interval $[-r_\nu^N, 0]$ and then denote by $\mathscr{J}_N^\dagger$ the Moore–Penrose [21] pseudo-inverse $\mathscr{J}_N^\dagger\colon Z_N \to Z$. In this case, if $z = (\eta, \psi) \in Z_N$ then $\mathscr{J}_N^\dagger(\eta, \psi) = (\eta, \phi)$, where $\phi = \psi$ on $[-r_\nu^N, 0]$, $\phi = 0$ on $[-r, -r_\nu^N)$.

For $\gamma^N = (z_0^N, q^N) \in \Gamma_N$, we define

$$z^N(t; \gamma^N, u) \equiv S_N(t) z_0^N + \int_0^t S_N(t-\sigma) P_N \hat{B}(\alpha^N) u(\sigma)\, d\sigma \tag{4.1}$$

and

$$\begin{aligned} y^N(t; \gamma^N) &\equiv \hat{C}(\alpha^N) z^N(t; \gamma^N, u) + D(\alpha^N) u(t) \\ &= C(\alpha^N)\pi^0 z^N(t; \gamma^N, u) + D(\alpha^N) u(t), \end{aligned} \tag{4.2}$$

where $A^N(q^N)$ is the i.g. for $S_N(t)$ on $X_N = X_N(q^N)$.

THEOREM 4.1. *Suppose* $\bar{\gamma}^N = (\bar{z}_0^N, \bar{q}^N)$ *is a sequence of solutions to the problems* $(\mathscr{P}_N)$ *and that there exists* $\bar{\gamma} \in \Gamma$ *such that* $\bar{\gamma}^N \to \bar{\gamma}$ *in the sense* (a) $\bar{q}^N \to \bar{q}$ *in* $R^{\mu+\nu}$, (b) $\mathscr{J}_N^\dagger \bar{z}_0^N \to \bar{z}_0$ *in* Z. *Suppose further that* $A_N = A_N(\bar{q}^N)$, $A = A(\bar{q})$ *satisfy the conditions and hypotheses of Theorem* 3.1. *Then*

$$|P_N z(t; \bar{\gamma}, u) - z^N(t; \bar{\gamma}^N, u)|_N \to 0 \quad \text{as } N \to \infty,$$

uniformly in t *on compact intervals, where*

$$z(t; \bar{\gamma}, u) \equiv S(t; \bar{q})\bar{z}_0 + \int_0^t S(t-\sigma; \bar{q})\hat{B}(\bar{\alpha})u(\sigma)\, d\sigma. \tag{4.3}$$

Proof. From the hypotheses and Theorem 3.1 we have immediately that $|S_N(t)P_N z - P_N S(t)z|_N \to 0$, uniformly on compact intervals, for all $z \in Z$. Therefore

$$\begin{aligned}
|S_N(t)\bar{z}_0^N - P_N S(t)\bar{z}_0|_N &= |S_N(t)P_N\mathscr{J}_N^\dagger \bar{z}_0^N - P_N S(t)\bar{z}_0|_N \\
&\leqq |S_N(t)[P_N\mathscr{J}_N^\dagger \bar{z}_0^N - P_N\bar{z}_0]|_N + |S_N(t)P_N\bar{z}_0 - P_N S(t)\bar{z}_0|_N \\
&\leqq M e^{\beta t}|\mathscr{J}_N^\dagger \bar{z}_0^N - \bar{z}_0| + |S_N(t)P_N\bar{z}_0 - P_N S(t)\bar{z}_0|_N.
\end{aligned}$$

The first term approaches 0 by (b), as does the second from our preceding remark. Next, consider

$$\begin{aligned}
&\left|\int_0^t S_N(t-\sigma)P_N\hat{B}(\bar{\alpha}^N)u(\sigma)\, d\sigma - P_N\int_0^t S(t-\sigma)\hat{B}(\bar{\alpha})u(\sigma)\, d\sigma\right|_N \\
&\quad = \left|\int_0^t [S_N(t-\sigma)P_N\hat{B}(\bar{\alpha}^N)u(\sigma) - P_N S(t-\sigma)\hat{B}(\bar{\alpha})u(\sigma)]\, d\sigma\right|_N \\
&\quad \leqq \int_0^t |S_N(t-\sigma)P_N[\hat{B}(\bar{\alpha}^N) - \hat{B}(\bar{\alpha})]u(\sigma)|_N\, d\sigma \\
&\qquad + \int_0^t |[S_N(t-\sigma)P_N - P_N S(t-\sigma)]\hat{B}(\bar{\alpha})u(\sigma)|_N\, d\sigma.
\end{aligned}$$

The results of Theorem 3.1, the continuity of $\hat{B}$, and dominated convergence yield convergence of these terms to 0, uniformly in t. The desired conclusion follows immediately from these estimates.

COROLLARY 4.1. *Suppose* $P_N : Z \to X_N$ *satisfies*

$$\pi^0(P_N z) \to \pi^0 z \quad \text{in } R^n \quad \text{for each } z \in Z. \tag{4.4}$$

Then under the assumptions of Theorem 4.1 *we have*

$$y^N(t; \bar{\gamma}^N) \to y(t; \bar{\gamma}) \quad \text{for each } t.$$

Proof. Recall

$$\begin{aligned}
y(t; \bar{\gamma}) &= \hat{C}(\bar{\alpha})z(t; \bar{\gamma}, u) + D(\bar{\alpha})u(t) \\
&= C(\bar{\alpha})\pi^0 z(t; \bar{\gamma}, u) + D(\bar{\alpha})u(t),
\end{aligned}$$

while

$$y^N(t; \bar{\gamma}^N) = C(\bar{\alpha}^N)\pi^0 z^N(t; \bar{\gamma}^N, u) + D(\bar{\alpha}^N)u(t).$$

The claimed result follows at once from the result of Theorem 4.1,

$$|\pi^0 z^N(t; \bar{\gamma}^N, u) - \pi^0 P_N z(t; \bar{\gamma}, u)|_{R^n} \to 0,$$

and (4.4), which yields

$$|\pi^0 P_N z(t;\bar{\gamma},u)-\pi^0 z(t;\bar{\gamma},u)|_{R^n}\to 0.$$

We observe that in (4.1) we define z^N for initial data in X_N. However, one can define an analogue for initial data given in $\mathscr{S}$. In particular for fixed $\gamma=(z_0,q)\in\Gamma=\mathscr{S}\times Q$ we define

$$\tilde{z}^N(t;\gamma,u)\equiv S_N(t)P_N z_0+\int_0^t S_N(t-\sigma)P_N\hat{B}(\alpha)u(\sigma)\,d\sigma, \tag{4.5}$$

where $A_N=A_N(q)$, $A=A(q)$ are i.g.'s for S_N, S. If one then assumes that $A_N(q)$, $A(q)$ satisfy the hypotheses of Theorem 3.1 so that $|P_N S(t)z-S_N(t)P_N z|_N\to 0$, one can prove in almost exactly the same manner as that for Theorem 4.1 above that

$$|\tilde{z}^N(t;\gamma,u)-P_N z(t;\gamma,u)|_N\to 0$$

for $\gamma\in\Gamma$. Defining $\tilde{y}^N(t;\gamma)$ as in (4.2) except with $\tilde{z}^N(t;\gamma,u)$ of (4.5) in place of $z^N(t;\gamma^N,u)$, we have under hypothesis (4.4) the analogue of the results of Corollary 4.1:

$$\tilde{y}^N(t;\gamma)\to y(t;\gamma) \tag{4.6}$$

for each fixed $\gamma\in\Gamma$.

We make the following standing assumptions on $\mathscr{S}$, Q and the approxination operators P_N.

Assumption 4.1. Q and $\mathscr{S}$ are compact and furthermore any sequence $\{\gamma^N\}$, $\gamma^N\in\Gamma_N$ is sequentially compact in the following sense: For $\gamma^N=(z_0^N,q^N)\in\Gamma_N$, $\{\mathscr{J}_N^\dagger z_0^N\}$ has a limit point in $\mathscr{S}$.

THEOREM 4.2. *Suppose $\{\bar{\gamma}^N\}$ is a sequence of solutions of the approximate problems $(\mathscr{P}_N)$ under Assumption* 4.1. *Then there exist $\bar{\gamma}\in\Gamma$ and a subsequence $\{\bar{\gamma}^{N_k}\}$ such that $\bar{\gamma}^{N_k}\to\bar{\gamma}$ in the sense of Theorem* 4.1(a), (b). *If $A_N(\bar{q}^N)$, $A(\bar{q})$ satisfy the hypotheses of Theorem* 3.1, *then $\bar{\gamma}$ is a solution for the problem $(\mathscr{P})$.*

Proof. Since $\bar{\gamma}^N=(\bar{z}_0^N,\bar{q}^N)\in\Gamma_N$, defining $\bar{z}^N\equiv\mathscr{J}_N^\dagger\bar{z}_0^N$, we have that there exists a convergent subsequence, say $\{\bar{z}^{N_k}\}$, converging to some $\bar{z}_0$ in $\mathscr{S}$; i.e., $\mathscr{J}_{N_k}^\dagger\bar{z}_0^{N_k}\to\bar{z}_0$ in $\mathscr{S}$. From the compactness of Q, we have that $\{\bar{q}^{N_k}\}$ possesses a convergent subsequence with $\bar{q}^{N_{k_i}}\to\bar{q}$ for some $\bar{q}\in Q$. Defining $\bar{\gamma}=(\bar{z}_0,\bar{q})\in\Gamma=\mathscr{S}\times Q$, and reindexing we thus have a sequence $\{\bar{\gamma}^{N_i}\}$ that converges in the sense of Theorem 4.1a, b to $\bar{\gamma}$. Furthermore, it follows from Theorem 4.1, Corollary 4.1 and the remarks involving (4.5) and (4.6) that for *any* $\gamma=(z_0,q)\in\Gamma$ one has $J(\bar{\gamma})\leqq J(\gamma)$. First, we have

$$J(\bar{\gamma})=\lim_{N_i\to\infty}J^{N_i}(\bar{\gamma}^{N_i})\qquad\text{(Corollary 4.1 yields } y^{N_i}(t;\bar{\gamma}^{N_i})\to y(t;\bar{\gamma})).$$

But we find

$$\lim_{N_i\to\infty}J^{N_i}(\bar{\gamma}^{N_i})\leqq\lim_{N_i\to\infty}J^{N_i}((P_{N_i}z_0,q))$$

$$=\lim\left[\sum_{i=1}^M|y^{N_i}(t_i;(P_{N_i}z_0,q))-\hat{y}_i|^2\right].$$

But $y^{N_i}(t_i;(P_{N_i}z_0,q))$ given by (4.1) is exactly the same as $\tilde{y}^{N_i}(t_i;\gamma)$, $\gamma=(z_0,q)$, where $\tilde{y}^N$ is defined as in (4.5), (4.6), and hence the last term is the same as

$$\lim_{N_i\to\infty}\left[\sum_{i=1}^M|\tilde{y}^{N_i}(t_i;\gamma)-\hat{y}_i|^2\right]=J(\gamma).$$

Thus, $\bar{\gamma}$ is a solution for $(\mathscr{P})$.

We turn next to a discussion of particular schemes which fit into the theoretical framework developed above. Throughout our presentation we shall assume that we are given a sequence $\gamma^N = (\eta^N, \phi^N, q^N)$ when $q^N = (\alpha^N, h^N) = (\alpha^N, r_1^N, \cdots, r_\nu^N) \in Q$, with $0 < r_1^N < r_2^N < \cdots < r_\nu^N \leqq r$, and $q^N \to \bar{q} = (\bar{\alpha}, \bar{h}) = (\bar{\alpha}, \bar{r}_1, \cdots, \bar{r}_\nu) \in Q$. We recall that for the systems under discussion we have the operator $\mathscr{A} = \mathscr{A}(\bar{q})$ defined on $\mathscr{D} = \{(\phi(0), \phi) \mid \phi \in W_2^{(1)}(-r, 0)\}$ given by

$$\mathscr{A}(\bar{q})(\phi(0), \phi) = (L(\bar{q})\phi, D\phi),$$

where the operator L is defined in (2.4). Hereafter we shall use the notation $D\phi$ in place of $\dot{\phi}$ in contexts where confusion might arise otherwise.

We summarize for future reference the conditions that our approximating schemes must satisfy:

(4.7) X_N is a closed subspace of $Z_N = R^n \times L_2^n(-r_\nu^N, 0)$, π_N is the canonical projection of Z_N onto X_N, $P_N = \pi_N \mathscr{J}_N$ and $\pi^0(P_N z) \to \pi^0 z$ for all $z \in Z$.

(4.8) There exist constants M and β such that $\mathscr{A}_N = \mathscr{A}_N(q^N)$ and $\mathscr{A} = \mathscr{A}(\bar{q})$ are in $G(M, \beta)$ on X_N and Z respectively.

(4.9) There exists $\mathscr{D}_1 \subset \mathscr{D} = \mathscr{D}(\mathscr{A}(\bar{q}))$, $\mathscr{D}_1$ dense in Z, such that
(i) $R_\lambda(\mathscr{A}(\bar{q}))\mathscr{D}_1 \subset \mathscr{D}_1$ for $\lambda > \beta$
(ii) $|\mathscr{A}_N P_N z - P_N \mathscr{A} z|_N \to 0$ as $N \to \infty$ for $z \in \mathscr{D}_1$.

In our discussions below we shall refer to (4.8) as the *stability* condition while (4.9) will be called the *consistency* condition. Our first scheme will be based on the averaging approximations developed in some detail in [5] while the second scheme utilizes spline approximations as formulated in [10].

5. The averaging approximation scheme. This identification scheme is defined using the "averaging" type approximations as discussed in [4], [5], and many of the arguments to verify that conditions (4.7), (4.8), (4.9) are satisfied are only slight modifications of those found in [5]. Given $q^N = (\alpha^N, r_1^N, \cdots, r_\nu^N)$, we partition $[-r_\nu^N, 0]$ into subintervals $[t_j^N, t_{j-1}^N]$, where $t_j^N \equiv -jr_\nu^N/N$, $j = 0, 1, \cdots, N$. Let χ_j^N denote the characteristic function of $[t_j^N, t_{j-1}^N)$ for $j = 2, 3, \cdots, N$, with χ_1^N the characteristic function for $[t_1^N, t_0^N] = [-r_\nu^N/N, 0]$. Define, for $(\eta, \phi) \in Z$,

$$\phi_j^N \equiv \frac{N}{r_\nu^N} \int_{t_j^N}^{t_{j-1}^N} \phi(s)\, ds, \qquad j = 1, 2, \cdots, N, \tag{5.1}$$
$$\phi_0^N \equiv \eta.$$

We define the closed subspaces X_N of Z_N by

$$X_N \equiv \left\{(\eta, \psi) \in Z_N \mid \eta \in R^n, \psi = \sum_{j=1}^N v_j^N \chi_j^N, v_j^N \in R^n\right\}.$$

The projection π_N of Z_N onto X_N is then given by

$$\pi_N(\eta, \phi) = \left(\eta, \sum_{j=1}^N \phi_j^N \chi_j^N\right). \tag{5.2}$$

With these definitions it is immediately obvious that (4.7) is satisfied.

For the operator L given by (2.4), we define the approximating operator $L_N(q^N): X_N \to R^n$ by

$$L_N(q^N)\left(\eta, \sum_1^N v_i \chi_j^N\right) \equiv A_0(\alpha^N)\eta + \sum_{i=1}^\nu \sum_{j=1}^N A_i(\alpha^N) v_i \chi_j^N(-r_i^N) + \sum_{j=1}^N K_j^N(\alpha^N) v_j, \tag{5.3}$$

where

$$K_j^N(\alpha) \equiv \int_{t_j^N}^{t_{j-1}^N} K(\alpha, \theta)\, d\theta. \tag{5.4}$$

Next, we define $D_N : X_N \to L_2^n(-r_\nu^N, 0)$ by

$$D_N\Big(\eta, \sum_1^N v_j \chi_j^N\Big) \equiv \sum_{j=1}^N \frac{N}{r_\nu^N}\{v_{j-1} - v_j\}\chi_j^N, \tag{5.5}$$

where $v_0 \equiv \eta$. Finally, we define $\mathscr{A}_N(q^N) : X_N \to X_N$ by

$$\mathscr{A}_N(q^N)(\eta, \psi) \equiv (L_N(q^N)(\eta, \psi), D_N(\eta, \psi)). \tag{5.6}$$

The proof that $\mathscr{A}_N(q^N) \in G(M, \beta)$ on X_N for some M and β independent of N is essentially given in [5] (see pp. 183, 186). One first argues that there is an equivalent inner product $\langle \cdot, \cdot \rangle_{g^N}$ on X_N such that $\langle \mathscr{A}_N(q^N)z, z\rangle_{g^N} \leqq \beta(q^N)\langle z, z\rangle_{g^N}$ for all $z \in X_N$. As in [5] we define, for given $r_1^N, r_2^N, \cdots, r_\nu^N$, the index set $J^N = \{j_1^N, \cdots, j_{\nu-1}^N\}$, where j_i^N is the index such that $-r_i^N \in [t_{j_i^N}^N, t_{j_i^N-1}^N)$, $i = 1, 2, \cdots, \nu-1$, and $j_\nu^N \equiv N$. We next define numbers a_j^N by $a_N^N = 1$ and, for $j = N-1, N-2, \cdots, 1$,

$$a_j^N = \begin{cases} a_{j+1}^N + 1 & \text{if } j \in J^N, \\ a_{j+1}^N & \text{if } j \notin J^N. \end{cases}$$

Then define the nondecreasing piecewise constant weighting function g^N by $g^N(\theta) = a_j^N$, $t_j^N \leqq \theta < t_{j-1}^N$, $j = 1, 2, \cdots, N$. Finally, we take $Z_N(g^N)$ and $X_N(g^N)$ as the spaces Z_N, X_N, respectively with equivalent topology generated by the inner product

$$\langle(\eta, \phi), (\zeta, \psi)\rangle_{g^N} \equiv \langle \eta, \zeta\rangle_{R^n} + \int_{-r_\nu^N}^0 \phi\psi g^N. \tag{5.7}$$

If we then consider $(\eta, \psi) = (\eta, \sum_{j=1}^N v_j\chi_j^N) \in X_N$ (and define $v_0 \equiv \eta$), we find in a straightforward manner using estimates similar to those in [5, p. 186] that

$$\begin{aligned} \langle \mathscr{A}_N(q^N)(\eta, \psi), (\eta, \psi)\rangle_{g^N} \leqq & \left\{ |A_0(\alpha^N)| + \frac{1}{2}\sum_{i=1}^\nu |A_i(\alpha^N)|^2 \right\} |\eta|^2 \\ & + \frac{1}{2}\sum_{i=1}^\nu |v_{j_i}|^2 + \sum_{j=1}^N |K_j^N(\alpha^N)|\, |v_j|\, |\eta| \\ & + \sum_{j=1}^N \langle v_{j-1} - v_j, v_j\rangle a_j^N. \end{aligned} \tag{5.8}$$

Noting that, for $\psi = \sum v_i\chi_i^N$,

$$|(\eta, \psi)|_N^2 = |\eta|^2 + \sum_{j=1}^N \frac{r_\nu^N}{N}|v_j|^2,$$

we find

$$\begin{aligned} & \sum_{j=1}^N |K_j^N(\alpha^N)|\, |v_j|\, |\eta| \\ & = \sum_{j=1}^N \left| \left(\frac{N}{r_\nu^N}\right)^{1/2} \int_{t_j^N}^{t_{j-1}^N} K(\alpha^N, \theta)\, d\theta \right| |\eta| \left| \left(\frac{r_\nu^N}{N}\right)^{1/2} v_j \right| \\ & \leqq \sum_{j=1}^N \left\{ \frac{1}{2}|\eta|^2 \left(\left(\frac{N}{r_\nu^N}\right)^{1/2} \int_{t_j^N}^{t_{j-1}^N} K(\alpha^N, \theta)\, d\theta \right)^2 + \frac{1}{2}\frac{r_\nu^N}{N}|v_j|^2 \right\} \\ & \leqq \sum_{j=1}^N \left\{ \frac{1}{2}|\eta|^2 \left(\frac{N}{r_\nu^N}\right)(t_{j-1}^N - t_j^N) \int_{t_j^N}^{t_{j-1}^N} |K(\alpha^N, \theta)|^2\, d\theta + \frac{1}{2}\frac{r_\nu^N}{N}|v_j|^2 \right\} \end{aligned}$$

$$\leqq \frac{1}{2}|\eta|^2 \int_{-r_\nu^N}^{0} |K(\alpha^N, \theta)|^2 \, d\theta + \frac{1}{2} \sum_{j=1}^{N} \frac{r_\nu^N}{N} |v_j|^2$$

$$\leqq \left(\frac{1}{2} + \frac{1}{2} \int_{-r_\nu^N}^{0} |K(\alpha^N, \theta)|^2 \, d\theta\right) |(\eta, \psi)|_N^2.$$

Next observe that

$$\sum_{j=1}^{N} \langle v_{j-1} - v_j, v_j \rangle a_j^N \leqq \sum_{j=1}^{N} \left\{\frac{1}{2}|v_{j-1}|^2 + \frac{1}{2}|v_j|^2 - |v_j|^2\right\} a_j^N$$

$$= \sum_{j=1}^{N} \left\{\frac{1}{2}|v_{j-1}|^2 - \frac{1}{2}|v_j|^2\right\} a_j^N$$

$$= \frac{1}{2}|\eta|^2 a_1^N + \frac{1}{2} \sum_{j=1}^{N-1} (a_{j+1}^N - a_j^N)|v_j|^2 - \frac{1}{2}|v_N|^2 a_N^N$$

$$= \frac{1}{2}\nu|\eta|^2 - \frac{1}{2} \sum_{i=1}^{\nu} |v_{j_i}|^2.$$

Using these estimates in (5.8) we finally obtain

$$\langle \mathscr{A}_N(q^N)(\eta, \psi), (\eta, \psi) \rangle_{g^N} \leqq \beta(q^N)|(\eta, \psi)|_N^2,$$

where

$$\beta(q^N) \equiv |A_0(\alpha^N)| + \frac{1}{2} \sum_{i=1}^{\nu} |A_i(\alpha^N)|^2 + \frac{1}{2} \int_{-r_\nu^N}^{0} |K(\alpha^N, \theta)|^2 \, d\theta + \frac{\nu + 1}{2}.$$

From the continuity assumptions made in § 2 (see (2.4)) and the fact that $q^N \in Q$, Q compact, we have the existence of $\tilde{\beta}$ such that $\beta(q^N) \leqq \tilde{\beta}$ for all N. Finally, since the X_N and $X_N(g^N)$ norms are equivalent independent of N, one finds (again see p. 186 of [5]) $\mathscr{A}_N(q^N) \in G(M, \tilde{\beta})$ on X_N. Since $\mathscr{A}(\bar{q})$ is the i.g. for a C_0-semigroup it also satisfies the requirement $\mathscr{A}(\bar{q}) \in G(M_1, \beta_1)$ on Z for some M_1 and β_1. It follows that the stability condition (4.8) is satisfied for our averaging approximations.

We next consider the consistency criteria (4.9). We take $\mathscr{D}_1 \equiv \{(\phi(0), \phi) \mid \phi \text{ is } C^1 \text{ on } [-r, 0]\}$. Then clearly $\mathscr{D}_1$ is dense in Z and $\mathscr{D}_1 \subset \mathscr{D}(\mathscr{A}(\bar{q}))$. Furthermore, $R_\lambda(A(\bar{q}))\mathscr{D}_1 \subset \mathscr{D}(\mathscr{A}^2(\bar{q})) \subset \mathscr{D}_1$ (see § 2 above) so that (4.9i) is satisfied for this choice of $\mathscr{D}_1$.

It remains to establish (4.9ii). Given $z = (\phi(0), \phi)$ in $\mathscr{D}_1$ we observe that

$$P_N \mathscr{A}(\bar{q}) z = (L(\bar{q})\phi, \sum_{j=1}^{N} (D\phi)_j^N \chi_j^N), \tag{5.9}$$

where

$$(D\phi)_j^N \equiv \frac{N}{r_\nu^N} \int_{t_j^N}^{t_{j-1}^N} D\phi(s) \, ds = \frac{N}{r_\nu^N} [\phi(t_{j-1}^N) - \phi(t_j^N)],$$

while

$$\mathscr{A}_N(q^N) P_N z = \mathscr{A}_N(q^N)\left(\phi_0^N, \sum_{j=1}^{N} \phi_j^N \chi_j^N\right) = (L_N(q^N) P_N z, D_N P_N z) \tag{5.10}$$

where $L_N(q^N)$ and D_N are given by (5.3) and (5.5). In view of (5.9), (5.10) and (1.4), it

thus suffices to show

$$A_0(\alpha^N)\phi(0)+\sum_{i=1}^{\nu}\sum_{j=1}^{N}A_i(\alpha^N)\phi_j^N\chi_j^N(-r_i^N)+\sum_{j=1}^{N}K_j^N(\alpha^N)\phi_j^N \tag{5.11}$$
$$\xrightarrow[R^n]{} A_0(\bar{\alpha})\phi(0)+\sum_{i=1}^{\nu}A_i(\bar{\alpha})\phi(-\bar{r}_i)+\int_{-\bar{r}_\nu}^{0}K(\bar{\alpha},\theta)\phi(\theta)\,d\theta$$

and

$$\int_{-r_\nu^N}^{0}\left|\sum_{j=1}^{N}\left[\frac{N}{r_\nu^N}(\phi_{j-1}^N-\phi_j^N)-(D\phi)_j^N\right]\chi_j^N\right|^2\to 0 \tag{5.12}$$

as $N\to\infty$ (and $r_i^N\to\bar{r}_i$).

Consider (5.12) first and write this integral as

$$\int_{-r_\nu^N/N}^{0}\left|\frac{N}{r_\nu^N}\left(\phi_0^N-\phi_1^N-\left[\phi(0)-\phi\left(\frac{-r_\nu^N}{N}\right)\right]\right)\right|^2$$
$$+\sum_{j=2}^{N}\int_{t_j^N}^{t_{j-1}^N}\left|\frac{N}{r_\nu^N}(\phi_{j-1}^N-\phi_j^N-[\phi(t_{j-1}^N)-\phi(t_j^N)])\right|^2$$
$$=T_1^N+T_2^N.$$

Using an analogue of [5, (3.18), p. 177] with r replaced by r_ω^N, estimates similar to those of [5] yield

$$T_2^N\leqq r_\nu^N\left|\sup_{1\leqq j\leqq N}2\mathscr{E}_j^N\right|^2,$$

where as in [5] we define

$$\mathscr{E}_j^N\equiv\sup\{|\dot{\phi}(\theta)-\dot{\phi}(s)|\,|\,s,\theta\in[t_j^N,t_{j-1}^N]\}.$$

Use of the analogue of [5, (3.18)] in T_1^N allows us to write (after arguing in much the same manner as done in [5, p. 177])

$$T_1^N\leqq\frac{r_\nu^N}{N}\left\{\frac{1}{2}\left|\dot{\phi}\left(\frac{-r_\nu^N}{N}\right)\right|+\frac{1}{2}\mathscr{E}_1^N\right\}^2.$$

Since $\mathscr{E}_j^N\to 0$ as $N\to\infty$, uniformly in j, we conclude that (5.11) obtains.

We remark that if $\dot{\phi}(0)=0$ and $\phi\in W_\infty^{(2)}(-r,0)$, then

$$\left|\dot{\phi}\left(\frac{-r_\nu^N}{N}\right)\right|=\left|\dot{\phi}\left(\frac{-r_\nu^N}{N}\right)-\dot{\phi}(0)\right|\leqq\sup|\ddot{\phi}(\theta)|\frac{r_\nu^N}{N},$$

and, since $\mathscr{E}_j^N$ is $O(r_\nu^N/N)$—see [5, p. 178] we find that the convergence in (5.12) is $O(1/N^2)$ or that the convergence in the second component (L_2 component) of (4.9ii) is $O(1/N)$.

Returning to (5.11) and recalling that $A_i(\alpha^N)\to A_i(\bar{\alpha})$, we see that to establish (5.11), we only need show

$$\sum_{j=1}^{N}\phi_j^N\chi_j^N(-r_i^N)\to\phi(-\bar{r}_i),\qquad i=1,2,\cdots,\nu \tag{5.13}$$

and

$$\sum_{j=1}^{N}K_j^N(\alpha^N)\phi_j^N\to\int_{-\bar{r}_\nu}^{0}K(\bar{\alpha},\sigma)\phi(\sigma)\,d\sigma. \tag{5.14}$$

For ϕ in C^1 on $[-r, 0]$ we have, for $\theta \in [t_j^N, t_{j-1}^N)$,

$$|\phi_j^N - \phi(\theta)| = \left|\frac{N}{r_\nu^N}\int_{t_j^N}^{t_{j-1}^N} [\phi(\sigma)-\phi(\theta)]\,d\sigma\right| \leqq \frac{N}{r_\nu^N}\int_{t_j^N}^{t_{j-1}^N} |\dot\phi|_\infty|\sigma-\theta|\,d\sigma$$
$$\leqq \frac{N}{r_\nu^N}|\dot\phi|_\infty\left(\frac{r_\nu^N}{N}\right)^2 = |\dot\phi|_\infty\frac{r_\nu^N}{N}. \tag{5.15}$$

From this, it follows immediately that

$$\left|\sum \phi_j^N\chi_j^N - \phi\right|^2_{L_2(-r_\nu^N,0)} \leqq (|\dot\phi|_\infty)^2\left(\frac{r_\nu^N}{N}\right)^2. \tag{5.16}$$

For $j_i = j_i^N$ chosen so that $-r_i^N \in [t_{j_i}^N, t_{j_i-1}^N)$, we find using (5.15) that

$$|\textstyle\sum \phi_j^N\chi_j^N(-r_i^N) - \phi(-\bar r_i)| \leqq |\phi_{j_i}^N - \phi(-r_i^N)| + |\phi(-r_i^N)-\phi(-\bar r_i)|$$
$$\leqq |\dot\phi|_\infty\frac{r_\nu^N}{N} + |\dot\phi|_\infty|r_i^N - \bar r_i|,$$

and thus the convergence in (5.13) is ensured by the convergence $r_i^N \to \bar r_i$, with the order given by $1/N$ if $r_i^N \to \bar r_i$ is of this order.

Finally, in considering (5.14) we note that

$$\sum_{j=1}^N K_j^N(\alpha^N)\phi_j^N = \sum_{j=1}^N \int_{t_j^N}^{t_{j-1}^N} K(\alpha^N,\theta)\phi_j^N\,d\theta$$
$$= \int_{-r_\nu^N}^0 K(\alpha^N,\theta)\sum_{j=1}^N \phi_j^N\chi_j^N(\theta)\,d\theta,$$

and hence

$$\Delta^N \equiv \sum_{j=1}^N K_j^N(\alpha^N)\phi_j^N - \int_{-\bar r_\nu}^0 K(\bar\alpha,\sigma)\phi(\sigma)\,d\sigma$$
$$= \int_{-r_\nu^N}^0 K(\alpha^N,\sigma)\left[\textstyle\sum \phi_j^N\chi_j^N(\sigma)-\phi(\sigma)\right]d\sigma + \int_{-r_\nu^N}^0 K(\alpha^N,\sigma)\phi(\sigma)\,d\sigma$$
$$- \int_{-\bar r_\nu}^0 K(\bar\alpha,\sigma)\phi(\sigma)\,d\sigma.$$

We thus find

$$|\Delta^N| \leqq \left(\int_{-r_\nu^N}^0 |K(\alpha^N,\sigma)|^2\,d\sigma\right)^{1/2}\left(\int_{-r_\nu^N}^0 |\textstyle\sum \phi_j^N\chi_j^N - \phi|^2\right)^{1/2}$$
$$+ \left|\int_{-r_\nu^N}^0 K(\alpha^N,\sigma)\phi(\sigma)\,d\sigma - \int_{-\bar r_\nu}^0 K(\bar\alpha,\sigma)\phi(\sigma)\,d\sigma\right|.$$

The first term is $O(1/N)$ by (5.16) while standard estimates on the second term yield that it $\to 0$ since $K(\alpha^N,\cdot)\to K(\bar\alpha,\cdot)$ in L_2 and $r_\nu^N \to \bar r_\nu$. The order of convergence of the second term depends on that of these latter two. If $|r_\nu^N - \bar r_\nu| = O(1/N)$ and if the convergence $K(\alpha^N,\cdot)\to K(\bar\alpha,\cdot)$ in L_2 is $O(1/N)$, then Δ^N is $O(1/N)$ also.

In summary, we have established (4.7), (4.8) and (4.9) for the average approximation I.D. scheme. In doing so we have also shown that under certain circumstances, the

convergence in (4.9ii) is $O(1/N)$. In particular, if in $q^N \to \bar{q}$ we have

(5.17)
(a) the convergence $A_i(\alpha^N) \to A_i(\bar{\alpha})$ is $O(1/N)$,
(b) the convergence $K(\alpha^N, \cdot) \to K(\bar{\alpha}, \cdot)$ in $L_2(-r_\nu^N, 0)$ is $O(1/N)$,
(c) $r_i^N \to \bar{r}_i$ is $O(1/N)$, $i = 1, 2, \cdots, \nu$,
(d) $z = (\phi(0), \phi)$, $\phi \in C^1$ on $[-r, 0]$, $\dot{\phi}(0) = 0$,

then $|\mathscr{A}_N(q^N)P_N(\phi(0), \phi) - P_N\mathscr{A}(\bar{q})(\phi(0), \phi)|_N$ is $O(1/N)$.

Remark 5.1. We remark that the conditions (5.17a–c) clearly are *not* conditions that can be verified a priori when using the averaging scheme in practice. These error estimates merely provide information as to how well the scheme might perform when applied to specific I.D. problems. Note that the particular method (maximum likelihood estimator, least squares, etc.) chosen for determining q^N will obviously affect the rates of convergence in (a)–(c) above. Finally, if one drops the condition $\dot{\phi}(0) = 0$ from (5.17d) but retains all other conditions in (5.17), one finds the order in (4.9ii) is only $1/\sqrt{N}$.

Remark 5.2. Recalling the order estimates on (3.2) given in Theorem 3.2, we observe that one can easily find sets $\mathscr{B}$ and $\mathscr{B}_1$ to satisfy the hypothesis of that theorem in the case of the averaging based scheme. For example, to insure convergence of order $1/\sqrt{N}$, one can choose the sets $\mathscr{B} = \mathscr{D}(\mathscr{A}^2(\bar{q}))$ and $\mathscr{B}_1 = \mathscr{D}(\mathscr{A}^3(\bar{q})) = \{(\phi(0), \phi) | \phi \in W_2^{(3)}(-r, 0),\ \dot{\phi}(0) = L(\phi),\ \ddot{\phi}(0) = L(\dot{\phi})\}$. Then, under the assumptions (5.17a–c), one can without difficulty argue the claimed order results.

6. Spline-based approximation schemes. We discuss in this section an identification scheme based on spline approximations. While we shall present the details for a scheme based on first order splines, arbitrary order spline approximations may be utilized in a similar manner with only slight modifications in the arguments indicated below (see the theory developed in [10], on which all of our discussions here are based).

Given $q^N = (\alpha^N, r_1^N, \cdots, r_\nu^N) \to \bar{q} = (\bar{\alpha}, \bar{r}_1, \cdots, \bar{r}_\nu)$ as we have hypothesized previously, we partition each of the subintervals $[-r_k^N, -r_{k-1}^N]$, $k = 1, 2, \cdots, \nu$, into N equal subintervals to define the partition $\{t_j^N\}_{j=1}^{\nu N}$ of $[-r_\nu^N, 0]$, with

$$t_j^N \equiv -\frac{(j-(k-1)N)(r_k^N - r_{k-1}^N)}{N} + r_{k-1}^N, \tag{6.1}$$

$j = (k-1)N, \cdots, kN$, $k = 1, 2, \cdots, \nu$. We then define the finite dimensional subspace $X_N \subset Z_N$ by

$$X_N = \{(\phi(0), \phi) | \phi \text{ is a first order spline with knots at } \{t_j^N\}\}.$$

We define the weighting function g^N by

$$g^N(\theta) = \begin{cases} 1, & -r_\nu^N \leqq \theta < -r_{\nu-1}^N, \\ 2, & -r_{\nu-1}^N \leqq \theta < r_{\nu-2}^N, \\ \vdots & \\ \nu-1, & -r_2^N \leqq \theta < -r_1^N, \\ \nu, & -r_1^N \leqq \theta < 0 \end{cases}$$

and, as in § 5, denote by $Z_N(g^N)$ and $X_N(g^N)$ the spaces Z_N and X_N endowed with the equivalent topology generated by the weighted inner product (5.7). We then define $\pi_N : Z_N \to X_N$ (equivalently $\pi_N : Z_N(g^N) \to X_N(g^N)$) as the orthogonal projection of $Z_N(g^N)$ onto $X_N(g^N)$. Then (see [10, p. 509]) for ψ in Z_N, we have $\pi_N\psi = \hat{\psi}^N$, where $\hat{\psi}^N$ is the solution of the problem of minimizing $|\hat{\psi} - \psi|_{Z_N(g^N)}$ over $\hat{\psi} \in X_N$. The operator $P_N : Z \to X_N$ is defined as before by $P_N = \pi_N \mathscr{J}_N$.

We adopt the following notation. For any function ϕ that is defined pointwise on $[-r_\nu^N, 0]$, we write $\hat{\phi} = (\phi(0), \phi)$ and $\hat{\phi}_I^N = (\phi_I^N(0), \phi_I^N)$ where ϕ_I^N is the interpolating spline (with knots at $\{t_j^N\}_{j=1}^{\nu N}$) for ϕ on $[-r_\nu^N, 0]$. For the projections π_N defined above we shall write $\pi_N\hat{\phi} = \pi_N(\phi(0), \phi) = \hat{\phi}^N = (\phi^N(0), \phi^N)$.

For any $q^N \in Q$, we define the operator $\mathscr{A}(q^N): \mathscr{D}(\mathscr{A}(q^N)) \subset Z_N \to Z_N$, where $\mathscr{D}(\mathscr{A}(q^N)) = \{(\phi(0), \phi) \in Z_N | \phi \in W_2^{(1)}(-r_\nu^N, 0)\}$, by

$$\mathscr{A}(q^N)(\phi(0), \phi) = (L(q^N)\phi, D\phi).$$

More generally, for any $q \in Q$, we can define $\mathscr{A}(q): \mathscr{D} \subset Z \to Z$ by

$$\mathscr{A}(q)(\phi(0), \phi) = (L(q)\phi, D\phi).$$

Note that in this latter case $D\phi$ is defined on $[-r, 0]$, while in the former $D\phi$ is defined on $[-r_\nu^N, 0]$. However, in both cases the operators are essentially the same, and in the discussions below we shall abuse notation and speak of $\mathscr{A}$ as an operator defined either in Z_N or Z, depending on the context. We note that $X_N \subset \mathscr{D}(\mathscr{A}(q^N))$ so that $\mathscr{A}(q^N)$ is defined on all of X_N.

With the definitions above, we have immediately from [10, Lemma 2.3] that $\mathscr{A}(q^N)$ satisfies

$$\langle \mathscr{A}(q^N)z, z\rangle_{g^N} \leqq \omega(q^N)|z|_{g^N}^2, \qquad z \in \mathscr{D}(\mathscr{A}(q^N)), \tag{6.2}$$

where

$$\omega(q^N) \equiv \frac{\nu+1}{2} + |A_0(\alpha^N)| + \frac{1}{2}\sum_{i=1}^{\nu}|A_i(\alpha^N)|^2 + \frac{1}{2}\int_{-r_\nu^N}^{0}|K(\alpha^N, \theta)|^2\, d\theta.$$

We next define $\mathscr{A}_N(q^N): X_N \to X_N$ by

$$\mathscr{A}_N(q^N) = \pi_N\mathscr{A}(q^N)\pi_N. \tag{6.3}$$

In view of (6.2) and the fact that $\pi_N x = x$ for any $x \in X_N$, we find for every $x \in X_N$

$$\begin{aligned}\langle \mathscr{A}_N(q^N)x, x\rangle_{g^N} &= \langle \pi_N\mathscr{A}(q^N)x, x\rangle \\ &= \langle \mathscr{A}(q^N)x, x\rangle \leqq \omega(q^N)|x|_{g^N}^2.\end{aligned}$$

As noted in § 5, there exists β such that $\omega(q^N) \leqq \beta$ for all N and thus $\mathscr{A}_N(q^N) \in G(\tilde{M}, \beta)$ on $X_N(g^N)$ and hence $\mathscr{A}_N(q^N) \in G(M, \beta)$ on X_N for some M independent of N. Similar arguments establish that $\mathscr{A}(\bar{q}) \in G(M, \beta)$ on Z for M, β appropriately chosen and it thus follows that condition (4.8) is satisfied by the approximations (6.3).

We turn next to the consistency condition (4.9) and define $\mathscr{D}_1 \equiv \mathscr{D}(\mathscr{A}^3(\bar{q}))$. This set is dense in Z and it follows at once that $R_\lambda(\mathscr{A}(\bar{q}))\mathscr{D}_1 \subset \mathscr{D}_1$ so that (4.9)–(i) is satisfied. For $z = (\phi(0), \phi) \in \mathscr{D}_1$ we have

$$\mathscr{A}_N(q^N)P_N z = \pi_N(L(q^N)\phi^N, D\phi^N)$$

where $\hat{\phi}^N = P_N(\phi(0), \phi) = \pi_N\mathscr{J}_N(\phi(0), \phi)$, while

$$P_N\mathscr{A}(\bar{q})z = \pi_N\mathscr{J}_N(L(\bar{q})\phi, D\phi).$$

It thus follows that for $z \in \mathscr{D}_1$

$$|\mathscr{A}_N P_N z - P_N\mathscr{A}z|_N = |\pi_N(L(q^N)\phi^N - L(\bar{q})\phi, D(\phi^N - \phi))|_N, \tag{6.4}$$

where we now understand that $D(\phi^N - \phi)$ is to be taken as a function on $[-r_\nu^N, 0]$. Recalling that π_N is the orthogonal projection of $Z_N(g^N)$ onto $X_N(g^N)$ and that the norms of $Z_N(g^N)$ and Z_N are equivalent (with constants independent of N), i.e.,

$|\pi_N z|_{g^N} \leqq |z|_{g^N} \leqq m|z|_N$, we see that (6.4) allows us to establish condition (4.9ii) by verifying

$$|D(\phi^N - \phi)| \to 0 \quad \text{as } N \to \infty \tag{6.5}$$

and

$$|L(q^N)\phi^N - L(\bar{q})\phi|_{R^n} \to 0 \quad \text{as } N \to \infty, \tag{6.6}$$

where in (6.5) the norm can be that of $L_2(-r_\nu^N, 0)$ or L_2 with the weighting function g^N (see (5.7)).

To show (6.5) and (6.6) we shall make use of standard estimates from the theory of spline approximations. Specifically, from [24, Thm. 2.5] we find, upon considering the interval $[t_{kN}^N, t_{(k-1)N}^N]$, $k = 1, 2, \cdots, \nu$, which has mesh size $h = (r_k^N - r_{k-1}^N)/N$,

$$\int_{t_{kN}^N}^{t_{(k-1)N}^N} |D(\phi - \phi_I^N)|^2 \leqq \frac{1}{\pi^2}\left\{\frac{r_k^N - r_{k-1}^N}{N}\right\}^2 \int_{t_{kN}^N}^{t_{(k-1)N}^N} |D^2\phi|^2 \tag{6.7}$$

and

$$\int_{t_{kN}^N}^{t_{(k-1)N}^N} |\phi - \phi_I^N|^2 \leqq \frac{1}{\pi^4}\left\{\frac{r_k^N - r_{k-1}^N}{N}\right\}^4 \int_{t_{kN}^N}^{t_{(k-1)N}^N} |D^2\phi|^2. \tag{6.8}$$

Here ϕ_I^N is the interpolating spline for $\phi \in C^2[-r, 0]$ with knots at $\{t_j^N\}$. Denoting by $|\cdot|_{2,N}$ the norm in $L_2(-r_\nu^N, 0)$, we deduce from (6.7) and (6.8) the estimates

$$|D(\phi - \phi_I^N)|_{2,N} \leqq \frac{1}{\pi}\{\max_j |r_j^N - r_{j-1}^N|\}\frac{1}{N}|D^2\phi|_{2,N} \tag{6.9}$$

and

$$|\phi - \phi_I^N|_{2,N} \leqq \frac{1}{\pi^2}\{\max_j |r_j^N - r_{j-1}^N|\}^2\frac{1}{N^2}|D^2\phi|_{2,N}. \tag{6.10}$$

Denoting by $|\cdot|_{2,N,g^N}$ the weighted norm in $L_2(-r_\nu^N, 0)$, we easily argue for $\hat{\phi} = (\phi(0), \phi) \in \mathscr{D}_1$ (using the minimality properties associated with π_N)

$$\begin{aligned}|\phi^N - \phi|_{2,N} \leqq |\pi_N\hat{\phi} - \hat{\phi}|_N = |\hat{\phi}^N - \hat{\phi}|_N \leqq |\hat{\phi}^N - \hat{\phi}|_{g^N} \\ \leqq |\hat{\phi}_I^N - \hat{\phi}|_{g^N} = |\phi_I^N - \phi|_{2,N,g^N} \leqq \sqrt{\nu}|\phi_I^N - \phi|_{2,N}.\end{aligned}$$

From (6.10) we observe that this last quantity is $O(1/N^2)$, and hence so is $|\pi_N\hat{\phi} - \hat{\phi}|_N$. It follows immediately that $|\phi^N(0) - \phi(0)|_{R^n}$ is $O(1/N^2)$ also.

We remark that we have shown that $|\pi_N\hat{\phi} - \hat{\phi}|_N \to 0$ whenever $\hat{\phi} \in \mathscr{D}_1$. The density of $\mathscr{D}_1$ in Z and the boundedness of $\{\pi_N\}$ thus imply this convergence ($|\pi_N z - z|_N \to 0$) for all $z \in Z$ and the condition $\pi^0(P_N z) \to \pi^0 z$ for $z \in Z$ of (4.7) is satisfied.

We next consider the inequality

$$|D(\phi^N - \phi)|_{2,N} \leqq |D(\phi^N - \phi_I^N)|_{2,N} + |D(\phi_I^N - \phi)|_{2,N}, \tag{6.11}$$

and observe that the second term is $O(1/N)$ by (6.9). We employ the Schmidt inequality [24] to estimate the first term. Since both ϕ^N and ϕ_I^N are linear on each subinterval

$[t_j^N, t_{j-1}^N]$ we have

$$\int_{-r_\nu^N}^{0} |D(\phi^N - \phi_I^N)|^2 = \sum_{j=1}^{\nu N} \int_{t_j^N}^{t_{j-1}^N} |D(\phi^N - \phi_I^N)|^2$$

$$\leqq \sum_{j=1}^{\nu N} \frac{\mathcal{K}}{(t_{j-1}^N - t_j^N)^2} \int_{t_j^N}^{t_{j-1}^N} |\phi^N - \phi_I^N|^2$$

$$= \sum_{k=1}^{\nu} \frac{\mathcal{K}}{\left[\dfrac{(r_k^N - r_{k-1}^N)}{N}\right]^2} \int_{-r_k^N}^{-r_{k-1}^N} |\phi^N - \phi_I^N|^2 \tag{6.12}$$

$$\leqq \sum_{k=1}^{\nu} \frac{\mathcal{K}}{\left[\dfrac{(r_k^N - r_{k-1}^N)}{N}\right]^2} \left\{ \int_{-r_k^N}^{-r_{k-1}^N} |\phi^N - \phi|^2 + \int_{-r_k^N}^{-r_{k-1}^N} |\phi - \phi_I^N|^2 \right\}$$

$$= T_1^N + T_2^N.$$

Using (6.8) we obtain the estimate for T_2^N

$$T_2^N \leqq \sum_{k=1}^{\nu} \frac{\mathcal{K}}{\pi^4} \left\{ \frac{r_k^N - r_{k-1}^N}{N} \right\}^2 \int_{-r_k^N}^{-r_{k-1}^N} |D^2\phi|^2 \leqq \frac{\mathcal{K}}{\pi^4} \left(\frac{r}{N} \right)^2 |D^2\phi|_{2,N}.$$

To obtain the desired estimate on T_1^N we need an additional assumption on $\bar{q} = (\bar{\alpha}, \bar{r}_1, \cdots, \bar{r}_\nu)$. Specifically, we assume:

(6.13) there exists $\delta > 0$ such that $|\bar{r}_k - \bar{r}_{k-1}| \geqq \delta$, $k = 1, 2, \cdots, \nu$.

With the assumption we find (for N sufficiently large)

$$T_1^N \leqq \frac{\mathcal{K} N^2}{\left(\dfrac{\delta}{2}\right)^2} \int_{-r_\nu^N}^{0} |\phi^N - \phi|^2 \leqq \frac{4\mathcal{K} N^2}{\delta^2} |\phi^N - \phi|_{2,N}^2.$$

But our arguments above revealed that $|\phi^N - \phi|_{2,N}$ is $O(1/N^2)$, and thus T_1^N, like T_2^N, is $O(1/N^2)$. It follows from (6.12) that the first term in (6.11) is $O(1/N)$. We have thus, under the additional assumption (6.13), established (6.5).

Finally, we observe that, for $-r_\nu^N \leqq \theta \leqq 0$,

$$\phi^N(\theta) = \phi^N(0) + \int_0^\theta D\phi^N,$$

$$\phi(\theta) = \phi(0) + \int_0^\theta D\phi$$

and thus

$$|\phi^N(\theta) - \phi(\theta)| \leqq |\phi^N(0) - \phi(0)| + \int_{-r_\nu^N}^{0} |D\phi^N - D\phi|$$

$$\leqq |\phi^N(0) - \phi(0)| + \sqrt{r}\, |D(\phi^N - \phi)|_{2,N}.$$

But these last two terms are $O(1/N)$, uniformly in θ. It follows that $|\phi^N(-r_i^N) - \phi(-r_i^N)|$ is $O(1/N)$. Since ϕ is continuous and $q^N \to \bar{q}$ we find that $L(q^N)\phi^N \to L(\bar{q})\phi$ and thus (6.6) obtains.

Summarizing, we have shown that (4.7), (4.8), (4.9) (where π_N is now the projection of $Z_N(g^N)$ onto $X_N(g^N)$) hold for the first order spline-based scheme defined

by the operators in (6.3) under the assumption (6.13). Furthermore, if one inspects carefully the estimates given above, one finds that under the hypotheses (5.17a-c), the convergence in (4.9ii) is $O(1/N)$.

Remark 6.1. In the above estimates we chose $\mathscr{D}_1 = \mathscr{D}(\mathscr{A}^3(\bar{q}))$, so that ϕ in (6.4) (and the subsequent arguments) was in $W_2^{(3)}(-r, 0)$. To apply the needed estimates (e.g., [24, Thm. 2.5] and make the arguments above, it is actually sufficient to have ϕ in $W_2^{(2)}(-r, 0)$ (see [25, Thm. 21]). We thus could have just as easily chosen $\mathscr{D}_1 = \mathscr{D}(\mathscr{A}^2(\bar{q}))$ and arrived at the conclusions above, including the convergence rates obtained.

Remark 6.2. In light of the above remark, we may, in order to obtain that the approximating semigroups converge like $O(1/N)$ for the scheme developed here, choose $\mathscr{B} = \mathscr{D}(\mathscr{A}^2(\bar{q}))$ and $\mathscr{B}_1 = \mathscr{D}(\mathscr{A}^3(\bar{q}))$ in Theorem 3.2. Under assumptions (6.13) and (5.17a–c), one then can readily verify that the hypotheses of Theorem 3.2 are satisfied by the spline-based approximations.

7. Numerical results. In this section we present a brief summary of some numerical results for the identification problem obtained using the approximation schemes (AVE and SPLINE) outlined in the two previous sections. For a more detailed discussion of the numerical performance of the AVE and SPLINE schemes in identification and control problems, the reader can consult [8], where numerous examples, error analyses, etc. are presented. The summary given here, taken with the extensive numerical tests reported in [8], support our claims of efficacy and practical usefulness for these methods.

In order to generate the data for testing the algorithms, we select a "true" set of parameters $\gamma^* = (\eta^*, \phi^*, \alpha^*, r^*)$ (we take $\nu = 1$ and $r_1 = r$) and a control u, and use the method of steps [14] to solve for x on the interval $[0, T]$. In all of the examples presented below "data" were generated using $r^* = 1$ and $u = u_l$, where u_l is the unit step at $t = l$ defined by

$$u_l(t) = \begin{cases} 0, & t < l, \\ 1, & l \leqq t \end{cases}$$

and $0 < l < 1$. The final time of $T = 2$ was used. The observations $\hat{y}_i = y(t_i)$ were generated at 101 equally spaced time steps on $[0, T]$. It is possible to add noise to the "data" to produce "noisy observations" $\hat{y}(t) = y(t) + \nu(t)$, where, for example, $\nu(t) = \text{col}(\nu_1(t), \cdots, \nu_k(t))$ is a computer-simulated vector of normal random variables $\nu_i(t)$, each with zero mean and preset standard variation. This was done for some of the examples in [8] (we do not do it here), where one again finds that the algorithms perform quite well.

For each fixed N, the approximation problem $(\mathscr{P}_N)$ was solved using a maximum likelihood estimator (MLE). The resulting solutions are denoted $\bar{\gamma}_A^N$ and $\bar{\gamma}_S^N$ for the AVE and SPLINE schemes respectively. Since the MLE is an iterative procedure it is necessary to supply a startup value (i.e., an initial guess) for the parameters γ_A^N or γ_S^N. If β denotes an unknown parameter to be estimated (e.g., $\beta = \alpha$ or $\beta = r$), then $\beta^{N,I}$ will denote the estimate for $\bar{\beta}^N$ obtained after I iterations of the MLE applied to problem $(\mathscr{P}_N)$. The startup value is denoted by $\beta^{N,0}$.

Example 7.1. In this example we seek to estimate the initial data $(\eta, \phi) \in R \times L_2(-1, 0)$ and the coefficient of the delayed term in a simple scalar equation. The system is described by the equation

$$\dot{x}(t) = .05x(t) + a_1 x(t-1) + u_{.1}(t),$$

with (unknown) initial data

$$x(0)=\eta, \qquad x_0(s)=\phi(s), \qquad -1\leqq s<0$$

and output

$$y(t)=x(t).$$

Data were generated as described previously using the true values $\eta^*=1$, $\phi^*\equiv 1$, $a_1^*=-4.0$. For each $N=2$, 4, 8, 16 and 32, the approximating problem $(\mathscr{P}_N)$ was formulated as discussed in § 4. Thus, for AVE we seek the "parameter"

$$\bar{\gamma}_A^N=(\eta, \phi_1^N, \phi_2^N, \cdots, \phi_N^N, a_1),$$

where $(\eta, \phi_1^N, \phi_2^N, \cdots, \phi_N^N)$ are coordinates of the AVE projection of the initial data. Similarly, for SPLINE we seek the "parameter"

$$\bar{\gamma}_S^N=(\xi_0^N, \xi_1^N, \cdots, \xi_N^N, a_1),$$

where $(\xi_0^N, \xi_1^N, \cdots, \xi_N^N)$ are coordinates for the SPLINE projection of the initial data.

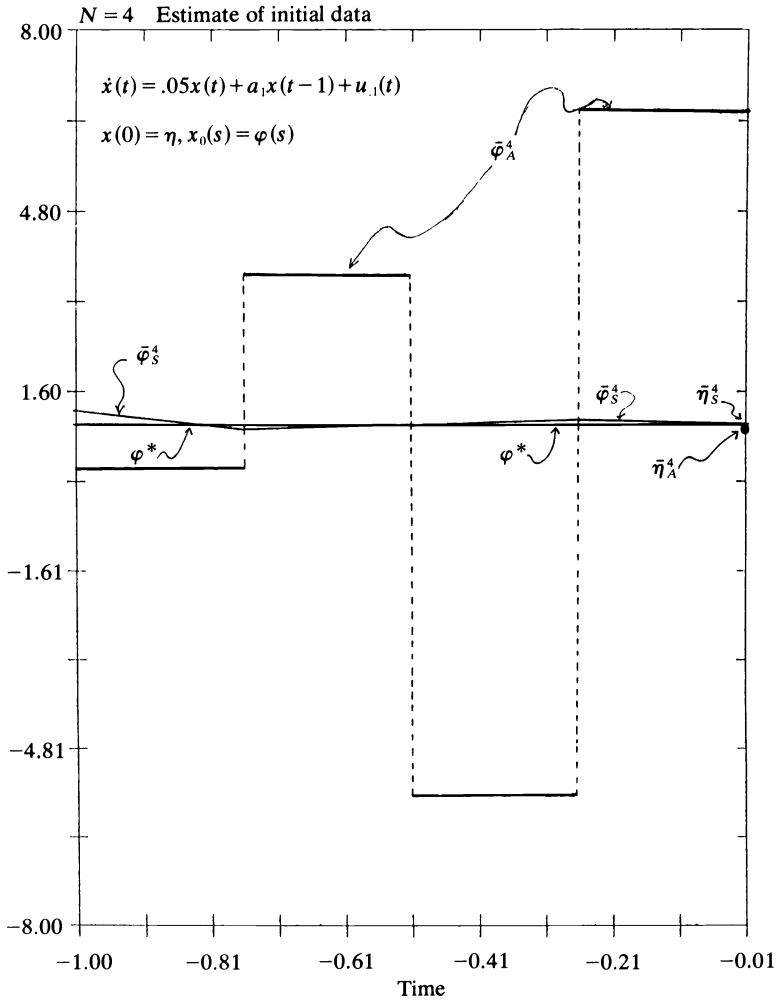


FIG. 7.1.1

TABLE 7.1.1

AVE			SPLINE		
N	$\bar{a}_1^N$	$\lvert z^*(0)-\bar{z}^N(0)\rvert$	N	$\bar{a}_1^N$	$\lvert z^*(0)-\bar{z}^N(0)\rvert$
2	−4.4103	2.08	2	−4.4382	.1595
4	−4.9924	4.53	4	−3.9381	.0867
8	−4.2651	41.76	8	−4.0031	.0287
16	did not converge		16	−4.0031	.0201
32	did not converge		32	−4.0001	.0386

The "startup" for $(\eta, \phi) \in R \times L_2(-1, 0)$ was taken as the zero initial data $(0, 0)$, whereas the "startup" for a_1 was chosen as $a_1^{N,0} = -3.0$. Table 7.1.1 provides an overview of the numerical findings. Because the initial data are in $R \times L_2(-1, 0)$ we have only displayed the Z-norm of the error and the estimated value for a_1^N. The comparison of the two schemes is quite striking; in particular, note the relative accuracies in estimating the initial data. Compared in Fig. 7.1.1 are graphs of the true initial data and the corresponding estimates produced by AVE and SPLINE for $N = 4$. It is apparent that (at least for the chosen "startup" values) the SPLINE procedure readily finds good estimates for the parameters, while the AVE scheme has considerable difficulty.

It is of some interest to compare the sequence of data fits generated as the MLE iteration procedure evolves. Figs. 7.1.2, 7.1.3 and 7.1.4 show the data matches from the AVE algorithm (with $N = 8$) for MLE iterations 0, 4 and 9, respectively. From the match at iteration 4 (Fig. 7.1.3) it might be deduced that the AVE scheme is in trouble. However, at iteration 9 the fit is quite good and Fig. 7.1.4 does not give any hint of the poor values of the parameters indicated in Table 7.1.1.

Figs. 7.1.5, 7.1.6 and 7.1.7 illustrate the SPLINE matches at iterations 0, 4 and 9, respectively. Again the iteration 4 matches indicate some difficulty while by iteration 9 the match is quite good. It happens that the SPLINE estimates of the parameters are excellent.

Although one cannot be certain, for the AVE scheme it does appear that the MLE procedure is converging to a local minimum of J^N. We suspect, however, that the problem $(\mathcal{P}_N)$ suffers a lack of identifiability. (See [8] for a further discussion of this matter.) The problem $(\mathcal{P}_N)$ for SPLINE seems to be much better behaved.

In order to further investigate identifiability for problems with unknown initial data, we made additional computations for this example using the same dynamics, changing only the initial data to

$$\eta = 1, \qquad \phi(s) = 1 + s, \qquad -1 \leqq s < 0.$$

Using the same start-ups as above, we found that SPLINE converged for all N values, whereas AVE *never* did. Results are summarized in Table 7.1.2.

Example 7.2. We consider an equation with a continuous (a constant function) kernel in which we wish to estimate the kernel, a system coefficient, and the time delay. The model is assumed to be of the form

$$\dot{x}(t) = a_1 x(t-r) + k \int_{-r}^{0} x(t+s)\, ds + u_{.1}(t),$$

with initial data

$$x_0(s) \equiv 1, \qquad -r \leqq s \leqq 0$$

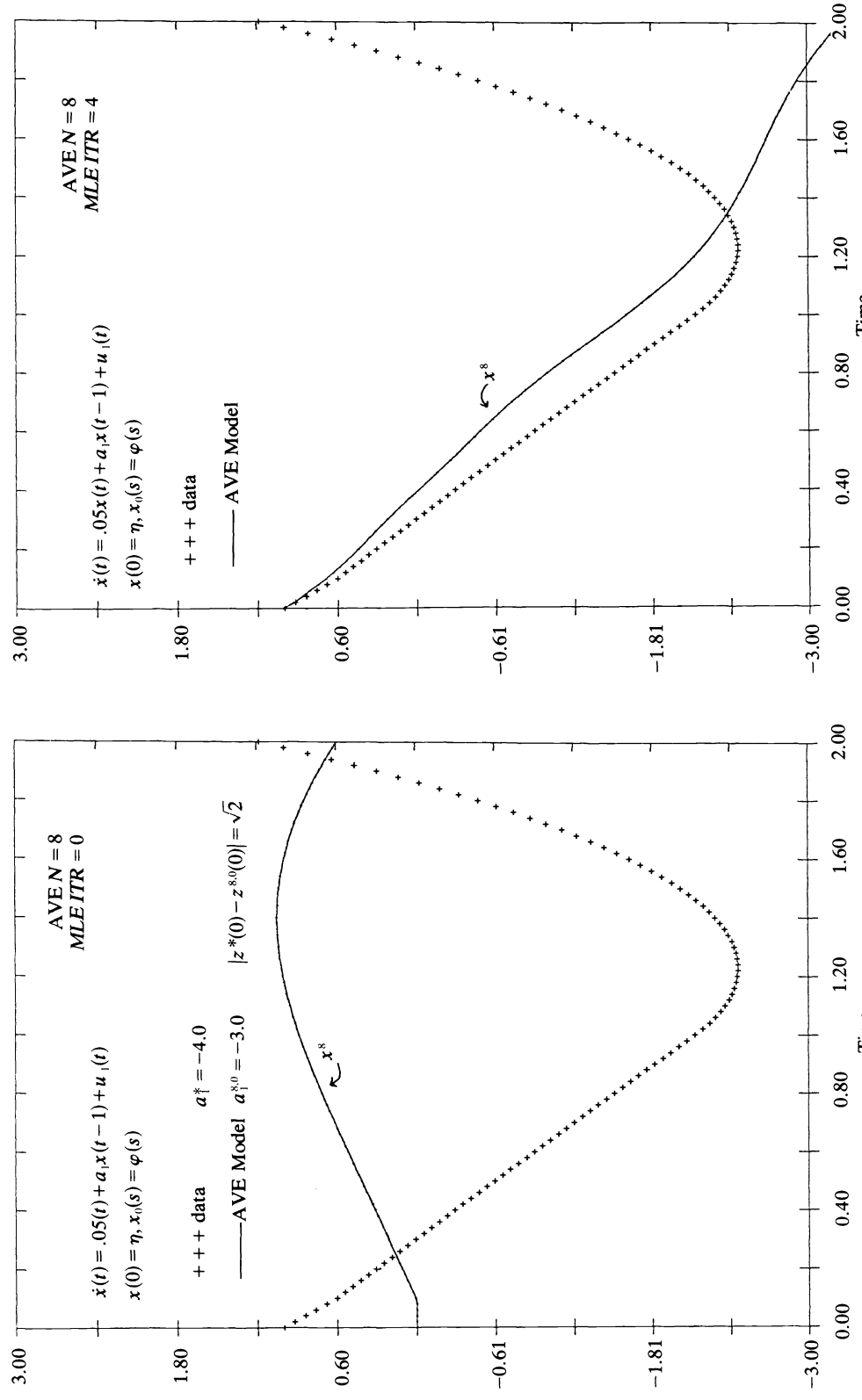


FIG. 7.1.3

FIG. 7.1.2

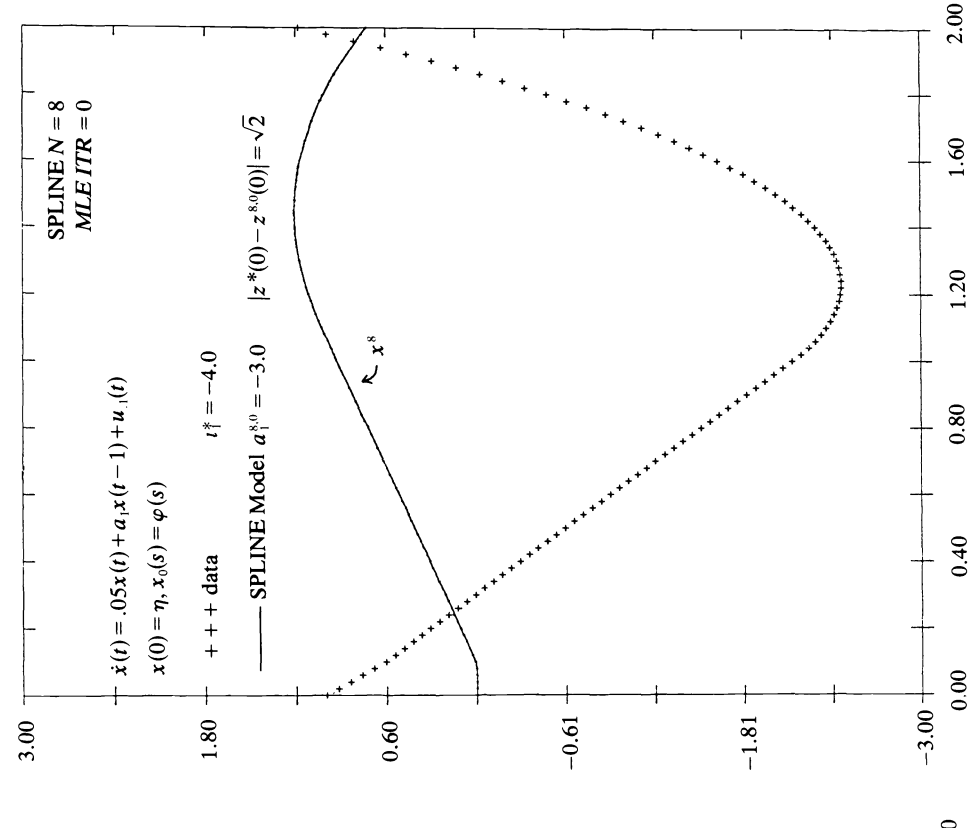


FIG. 7.1.5

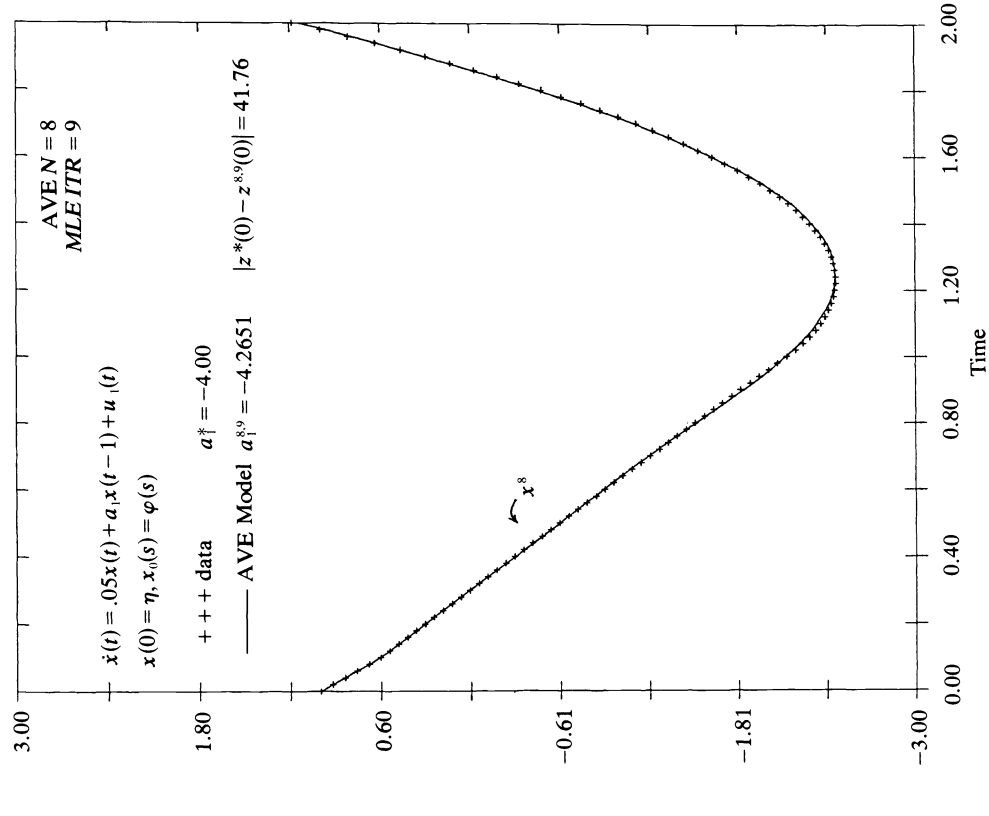


FIG. 7.1.4

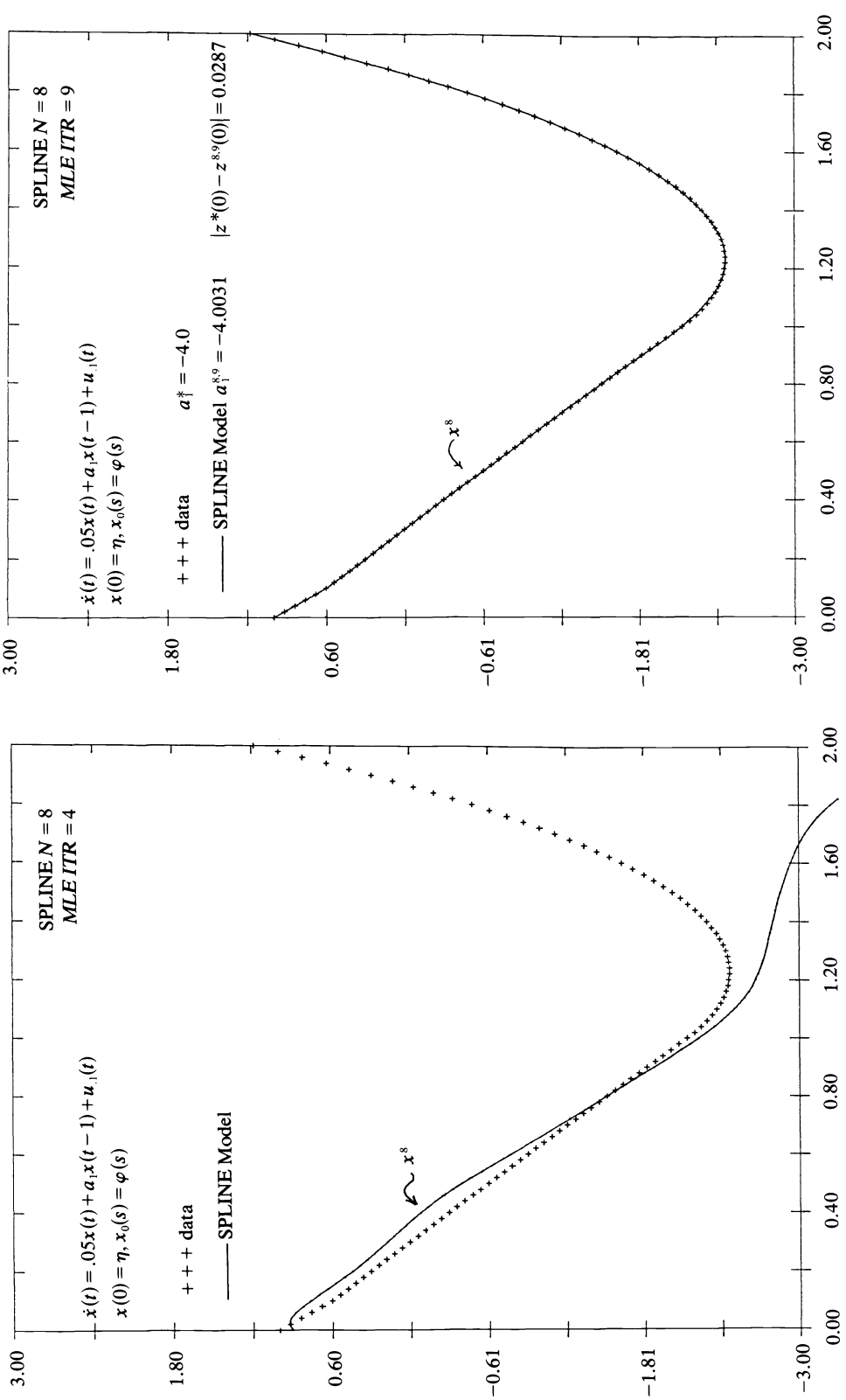


FIG. 7.1.6

FIG. 7.1.7

TABLE 7.1.2
Linear initial data

AVE			SPLINE		
N	$\bar{a}_1^N$	$\lvert z^*(0)-\bar{z}^N(0)\rvert$	N	$\bar{a}_1^N$	$\lvert z^*(0)-\bar{z}^N(0)\rvert$
2			2	−4.5201	.0563
4			4	−4.0975	.0318
8	did not converge		8	−4.0282	.0123
16			16	−4.0123	.0193
32			32	−4.0122	.0936

and output

$$y(t) = x(t).$$

The true parameters $a_1^* = -3.0$, $k^* = -1.0$ and $r^* = 1.0$ were estimated using startups of

$$a_1^{N,0} = -3.5, \quad k^{N,0} = -1.5, \quad r^{N,0} = 1.5.$$

Runs were made for $N = 2, 4, 8$ and 16. The MLE algorithm for the AVE scheme did not converge for $N = 2$ and 4. However, for $N = 8$ and 16 the AVE scheme converged but produced rather poor parameter estimates. The SPLINE scheme converged for each $N = 2$, 4, 8, 16 and for $N \geqq 4$ produced good parameter estimates. The numerical results for this problem are summarized in Tables 7.2.1 and 7.2.2, where $e_N \equiv \bar{\gamma}^N - \gamma^*$ is the error.

Figs. 7.2.1 through 7.2.4 compare the $N = 8$ AVE and SPLINE data fits. In particular, Figs. 7.2.1. and 7.2.2. show the $N = 8$ AVE start-up and converged data fits, respectively. Figs. 7.2.3 and 7.2.4 show similar results for the SPLINE procedure.

TABLE 7.2.1

AVE				
N	$\bar{r}^N$	$\bar{k}^N$	$\bar{a}_1^N$	$\lvert e_N\rvert$
2		did not converge		
4		did not converge		
8	.8802	.2182	−4.1641	2.0657
16	.9383	−.3806	−3.5535	1.2346
$\gamma^* =$	1.0000	−1.0000	−3.0000	

TABLE 7.2.2

SPLINE				
N	$\bar{r}^N$	$\bar{k}^N$	$\bar{a}_1^N$	$\lvert e_N\rvert$
2	.9100	−.4376	−3.4478	1.1002
4	.9896	−1.0087	−3.0580	.0071
8	1.0018	−1.0390	−2.9953	.0455
16	1.0042	−1.0410	−2.9841	.0611
$\gamma^* =$	1.0000	−1.0000	−3.0000	

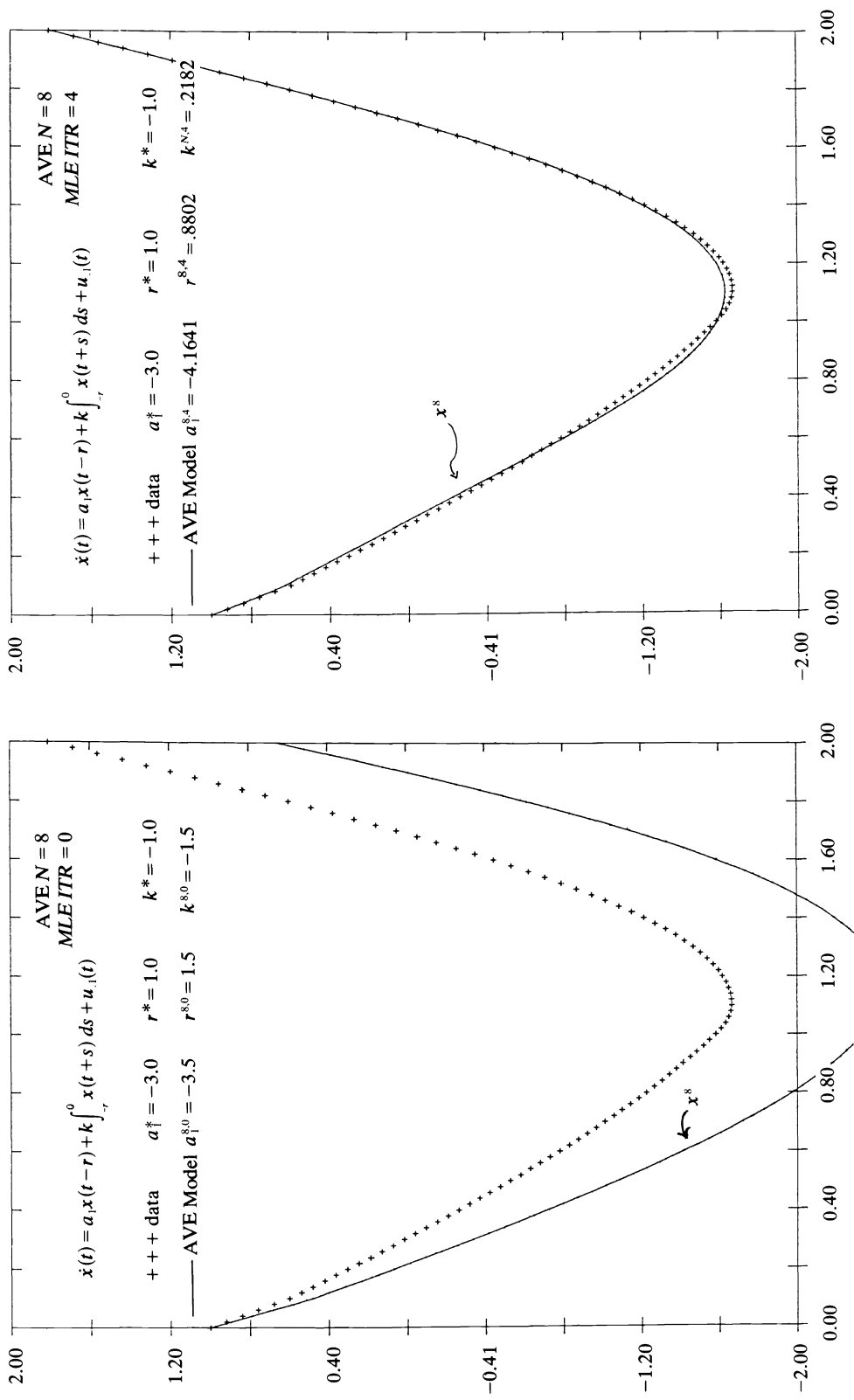


FIG. 7.2.1

FIG. 7.2.2

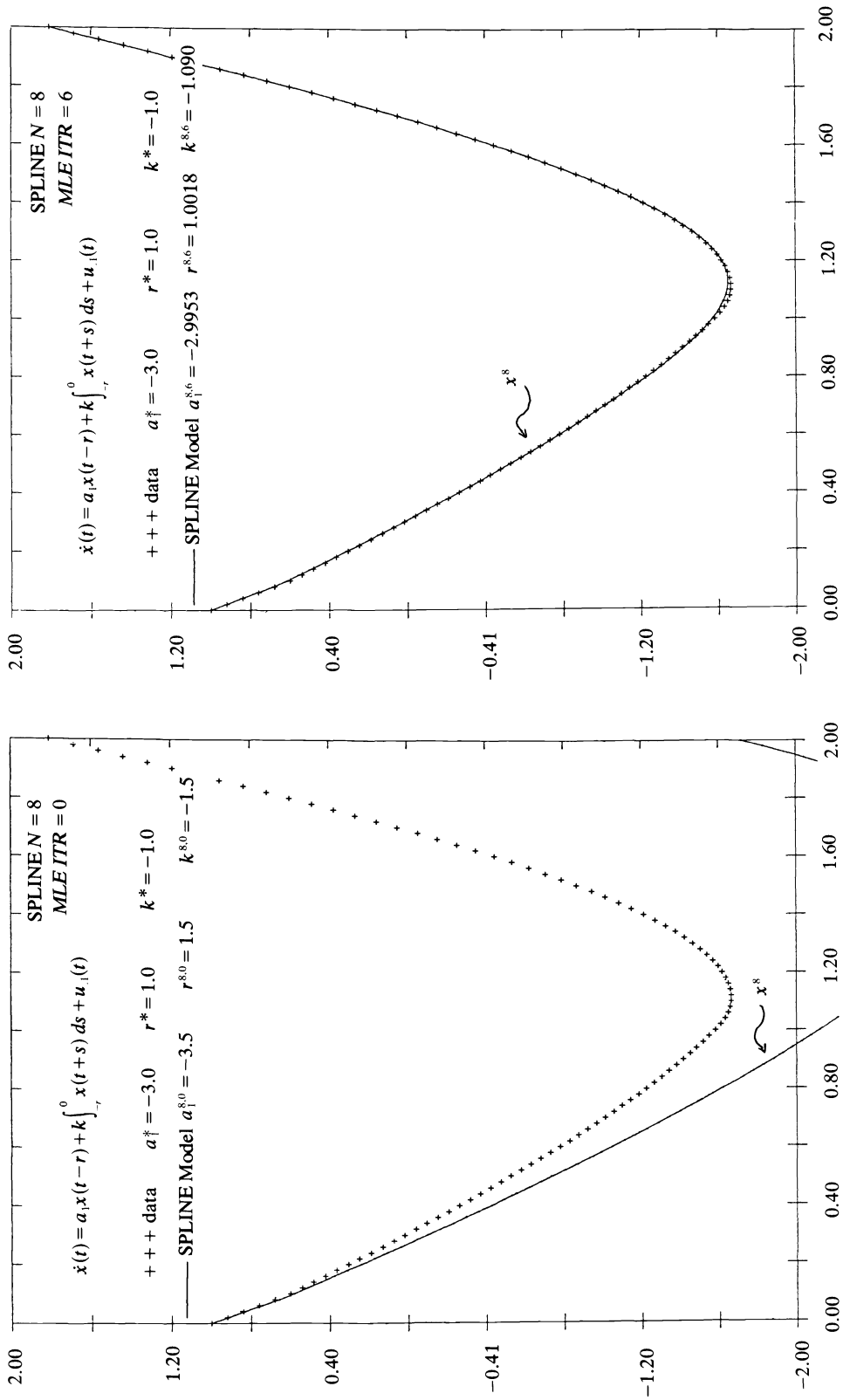


FIG. 7.2.3

FIG. 7.2.4

Example 7.3. In our final example we consider an oscillator with retarded damping and retarded restoring forces. We seek to estimate the coefficients of the delayed terms and the time delay itself. The system is governed by the equation

$$\ddot{x}(t)+16x(t)+a_0\dot{x}(t-r)+a_1x(t-r)=u_{.1}(t),$$

with initial data

$$x_0(s)\equiv 1, \qquad \dot{x}_0(s)\equiv 0, \qquad -r\leqq s\leqq 0$$

and output

$$y(t)=x(t).$$

This second order equation is equivalent to the two-dimensional system

$$\frac{d}{dt}\begin{bmatrix}x_1(t)\\x_2(t)\end{bmatrix}=\begin{bmatrix}0 & 1\\-16 & 0\end{bmatrix}\begin{bmatrix}x_1(t)\\x_2(t)\end{bmatrix}+\begin{bmatrix}0 & 0\\-a_1 & -a_0\end{bmatrix}\begin{bmatrix}x_1(t-r)\\x_2(t-r)\end{bmatrix}+\begin{bmatrix}0\\1\end{bmatrix}u_{.1}(t),$$

with initial condition

$$\begin{bmatrix}x_1\\x_2\end{bmatrix}_0(s)\equiv\begin{bmatrix}1\\0\end{bmatrix}, \qquad -r\leqq s\leqq 0,$$

and output

$$y(t)=[1 \quad 0]\begin{bmatrix}x_1(t)\\x_2(t)\end{bmatrix}.$$

The true parameters to be estimated are $a_0^*=10.0$, $a_1^*=-10.0$ and $r^*=1.0$. Startup values for each run were

$$a_0^{N,0}=11.0, \quad a_1^{N,0}=-9.0, \quad r^{N,0}=1.2.$$

Convergence results for this example are summarized in Tables 7.3.1 and 7.3.2. At $N=16$ the relative l_1 error ($|e_N|/|\gamma^*|$) for AVE is approximately 3.5%, while the $N=16$ SPLINE scheme produced a relative l_1 error of less than 1%.

Figs. 7.3.1 and 7.3.2 show the $N=4$ converged data fits for AVE and SPLINE, respectively. For $N\geqq 8$, the data fits are nearly perfect and are not shown.

TABLE 7.3.1

AVE

N	$\bar{a}_0^N$	$\bar{a}_1^N$	$\bar{r}^N$	$\lvert e_N\rvert$
2		did not converge		
4	54.5124	−9.1876	2.4190	46.7439
8	19.4941	−9.4927	1.3506	10.3520
16	10.6433	−9.9089	.9998	.7346
$\gamma^*=$	10.0000	−10.0000	1.0000	

TABLE 7.3.2

SPLINE

N	$\bar{a}_0^N$	$\bar{a}_1^N$	$\bar{r}^N$	$\lvert e_N\rvert$
2	9.2585	−10.5360	1.0908	1.3683
4	10.0927	−10.0619	1.0076	.1622
8	9.9724	−10.0177	1.0010	.0463
16	9.9811	−10.0108	1.0017	.0314
$\gamma^*=$	10.0000	−10.0000	1.0000	

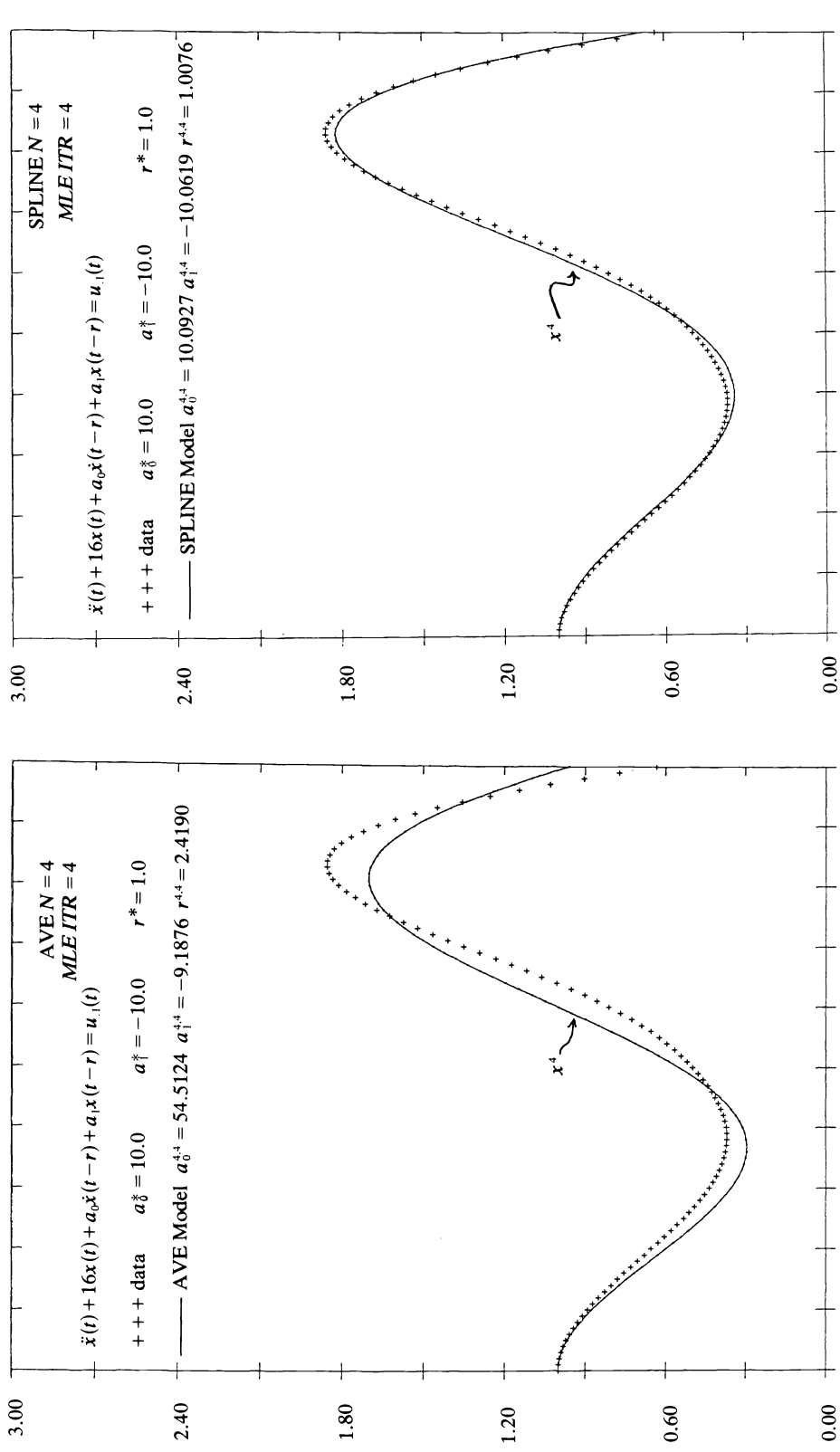


FIG. 7.3.1

FIG. 7.3.2

8. Concluding remarks. We close with an addendum of several remarks on questions and results that have arisen since the preceding part of this paper was written. First, the methods employed in this paper are also applicable to parameter identification problems for distributed parameter systems. In subsequent work (see [29]), both theoretical and numerical results have been obtained for linear and nonlinear equations of hyperbolic and parabolic type. Standard versions of the Trotter–Kato type approximation theorems (for example, the version due to Kurtz in [19]) are adequate for the partial differential equation problems treated to date.

With regard to the Trotter–Kato type theorem of § 3, we have subsequently learned that there is a version of the approximation results (which does not require $|P_N z| \to |z|$) due to Kurtz from which our Theorem 3.1 follows directly. Specifically, [30, Chapt. 1, Thm. 5.1] yields the results of our Theorem 3.1 in a rather straightforward manner. We gratefully acknowledge our fruitful conversations with Tom Kurtz and his willingness to provide us with a preliminary version of material from his forthcoming book upon learning of our own efforts and interests in these techniques.

Finally, we feel that additional comment might clarify the relationship between our numerical efforts on the simple examples of § 7 and the "real-world" motivating examples of § 1. The methods developed in this paper *are* applicable to the motivating examples and work is now in progress on both the column reactor problem (preliminary numerical results indicate that the methods should perform quite satisfactorily on these problems) and the unsteady aerodynamics problem where in each case one uses actual experimental data in computing estimates for the parameters. The focus of our efforts reported in this paper was the theoretical development and numerical testing of the techniques we have proposed. For the testing (both here and in [8]) we chose to use simple examples of types often encountered in applications for which "true" solutions could be easily obtained and used in comparing the techniques. Detailed treatments of the use of these techniques in conjunction with the motivating examples of § 1 will appear in future publications.

REFERENCES

[1] K. J. ÅSTRÖM AND P. EYKHOFF, *System identification—a survey*, Automatica, 7 (1971), pp. 123–162.

[2] A. V. BALAKRISHNAN, *Active control of airfoils in unsteady aerodynamics*, Appl. Math. Opt., 4 (1978), pp. 171–195.

[3] H. T. BANKS, *Approximation of nonlinear functional differential equation control systems*, J. Optim. Theory Appl., 29 (1979), pp. 383–408.

[4] H. T. BANKS AND J. A. BURNS, *An abstract framework for approximate solutions to optimal control problems governed by hereditary systems* in Proc. International Conference on Differential Equations (Univ. So. Calif., Sept., 1974), H. A. Antosiewicz, ed., Academic Press, New York, 1975, pp. 10–25.

[5] ———, *Hereditary control problems: numerical methods based on averaging approximations*, SIAM J. Control and Optimization, 16 (1978), pp. 169–208.

[6] ———, *Approximation techniques for control systems with delays*, Proc. Int'l Conference on Methods of Mathematical Programming, 1977, Polish Scientific Publishers, Warsaw, to appear.

[7] H. T. BANKS, J. A. BURNS AND E. M. CLIFF, *Spline-based approximation methods for control and identification of hereditary systems*, in International Symposium on Systems Optimization and Analysis, A. Bensoussan and J. L. Lions, eds., Lecture Notes in Control and Information Science, 14, Springer, Heidelberg, 1979, pp. 314–320.

[8] ———, *A comparison of numerical methods for identification and optimization problems involving control systems with delays*, Brown Univ. LCDS Tech. Rep. 79-7, Providence, RI, 1979.

[9] H. T. BANKS, J. A. BURNS, E. M. CLIFF AND P. R. THRIFT, *Numerical solutions of hereditary control problems via an approximation technique*, Brown Univ. LCDS Tech. Rep. 75-6, Providence, RI, 1975.

[10] H. T. BANKS AND F. KAPPEL, *Spline approximations for functional differential equations*, J. Differential Eq., 34 (1979), pp. 496–522.
[11] J. A. BURNS AND E. M. CLIFF, *On the formulation of some distributed system parameter identification problems*, Proc. AIAA Symposium on Dynamics and Control of Large Flexible Spacecraft, 1977, pp. 87–105.
[12] ———, *Methods for approximating solutions to linear hereditary quadratic optimal control problems*, IEEE Trans. Automat. Control, 23 (1978), pp. 21–36.
[13] ———, *Parameter identification for linear hereditary systems via an approximation technique*, in Information Linkage Between Applied Mathematics and Industry, Peter C. C. Wang, ed., Academic Press, New York, 1979, pp. 527–538.
[14] L. E. EL'SGOL'TS, *Introduction to the Theory of Differential Equations with Deviating Arguments*, Holden-Day, San Francisco, 1966.
[15] G. GELLF AND J. HENRY, *Experimental and theoretical study of diffusion, convection and reaction phenomena for immobilized enzyme systems*, in Analysis and Control of Immobilized Enzyme Systems, D. Thomas and J. P. Kernevez, eds., North-Holland/American Elsevier, New York, 1976, pp. 253–274.
[16] J. HENRY, *Contrôle d'un réacteur enzymatique à l'aide de modèles à paramètres distribués: quelques problèmes de contrôlabilité de systèmes paraboliques*, Thèses d'État, Université Paris VI, 1978.
[17] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.
[18] K. KUNISCH, *Approximation of optimal control problems for nonlinear hereditary systems of neutral type*, Ber. math.-stat. Sekt. Forschungszentrum, to appear.
[19] T. G. KURTZ, *Extensions of Trotter's operator semigroup approximation theorem*, J. Functional Anal., 3 (1969), pp. 354–375.
[20] P. D. LAX AND R. D. RICHTMYER, *Survey of the stability of linear finite difference equations*, Comm. Pure Appl. Math., 9 (1956), pp. 267–293.
[21] M. Z. NASHED, *Generalized inverses, normal solvability, and iteration for singular operator equations*, in Nonlinear Functional Analysis and Applications, L. B. Rall, ed., Academic Press, New York, 1971, pp. 311–359.
[22] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Math. Dept. Lecture Notes, Vol. 10, Univ. Maryland, College Park, 1974.
[23] A. P. SAGE AND J. L. MELSA, *System Identification*, Academic Press, New York, 1971.
[24] M. H. SCHULTZ, *Spline Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
[25] M. H. SCHULTZ AND R. S. VARGA, *L-splines*, Numer. Math., 10 (1967), pp. 345–369.
[26] H. F. TROTTER, *Approximations of semigroups of operators*, Pacific J. Math., 8 (1958), pp. 887–919.
[27] ———, *Approximation and perturbation of semigroups*, in Linear Operators and Approximation II, P. L. Butzer and B. Sz.-Nagy, eds., Birkhaüser-Verlag, Basel, 1974, pp. 3–21.
[28] T. VON KARMAN AND J. M. BURGERS, *General Aerodynamic Theory—Perfect Fluids, Vol. II, Aerodynamic Theory*, W. F. Durand, ed., Dover, New York, 1963.
[29] H. T. BANKS AND K. KUNISCH, *Parameter estimation techniques for nonlinear distributed parameter systems*, Proc. International Conference on Nonlinear Phenomena in Mathematical Sciences, University of Texas, Arlington, June 16–20, 1980, to appear.
[30] S. N. ETHIER AND T. G. KURTZ, *Markov Processes: Existence and Approximation*, John Wiley, New York, 1981, to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 19, No. 6, November 1981

0363-0129/81/1906-0006 $01.00/0

DISCRETE TIME STOCHASTIC ADAPTIVE CONTROL*

GRAHAM C. GOODWIN†, PETER J. RAMADGE‡ AND PETER E. CAINES¶

Abstract. This paper establishes global convergence of a stochastic adaptive control algorithm for discrete time linear systems. It is shown that, with probability one, the algorithm will ensure the system inputs and outputs are sample mean square bounded and the conditional mean square output tracking error achieves its global minimum possible value for linear feedback control. Thus, asymptotically, the adaptive control algorithm achieves the same performance as could be achieved if the system parameters were known.

1. Introduction. A key problem in stochastic control is the question of global convergence of adaptive control algorithms. By global convergence of a stochastic adaptive control algorithm we mean that for all initial system and algorithm states, the (conditional) mean square output tracking error is minimized, with probability one, and that this is achieved with a sample mean square bounded input sequence.

It is only recently that significant progress has been made on the global convergence of adaptive control algorithms. Feuer and Morse [1] and Morse [2] have treated a continuous time algorithm for deterministic systems. These results appear to be the most general to date for the single-input single-output continuous time case. In the discrete time case the present authors [3] have established global convergence for a class of adaptive control algorithms applied to multi-input multi-output deterministic linear systems.

Recently progress has also been made on the convergence of recursive algorithms used for parameter estimation. In [4], which deals with general estimation problems and [5], which in addition treats the adaptive control problem, Ljung presents general tools for the analysis of recursive algorithms. In [6] Solo has analyzed several recursive parameter estimation algorithms using a martingale approach. The works of both Ljung and Solo are important precursors to our subsequent analysis of stochastic adaptive algorithms.

To date the most extensive treatment of the problem of global convergence for stochastic recursive identification algorithms for discrete time linear systems appears in the recent work of Ljung [4], [5]. In these papers, the asymptotic properties of the algorithms are associated with the solutions of an ordinary differential equation. The analysis in [5] indicates that, for certain algorithms, a positive real condition on the system noise dynamics is required for convergence. However a question that remains unanswered in the above work concerns the boundedness of the system variables. It has been argued in [7] that a particular adaptive control algorithm, based on recursive least squares [8], has a sample mean square boundedness property. However the arguments in [7] are heuristic and as a result the question of global convergence of stochastic adaptive control algorithms remained unresolved in this set of papers.

In this paper we shall establish global convergence for a class of adaptive algorithms for stochastic linear systems. Subject to an inverse stability condition and a positive real condition (see, e.g. [13]), the system inputs and outputs will be shown to be

* Received by the editors January 8, 1979, and in final revised form November 1, 1980. This work was supported in part by the Australian Research Grants Committee, Joint Services Electronics Program under Contract N00014-75-C-0648, and in the case of the first author a Fullbright Grant and the Division of Applied Sciences, Harvard University.

† Department of Electrical Engineering, University of Newcastle, N.S.W. 2308, Australia.

‡ Department of Electrical Engineering, University of Toronto, Toronto, Ontario, Canada.

¶ Division of Applied Sciences, Harvard University, Pierce Hall, Cambridge, Massachusetts 02138; now with the Department of Electrical Engineering, McGill University, Montreal, P.Q., Canada, H3A 2A7.

sample mean square bounded and the (conditional) mean square output tracking error will be shown to converge to its global minimum value with probability one.

We shall first treat an algorithm for single-input single-output systems having correlated disturbances, since we believe this to be a case of prime interest. In a later section an analogous result for the multiple-input multiple-output case is presented.

2. Problem statement. In this paper we are concerned with the adaptive control of linear time-invariant finite dimensional systems which admit autoregressive moving average representations of the form

$$A(q^{-1})y(t) = q^{-d}[B(q^{-1})]u(t) + [C(q^{-1})]w(t), \qquad t \geqq 1, \tag{2.1}$$

where $\{y(t)\}, \{u(t)\}, \{w(t)\}$ denote s-dimensional output, r-dimensional input and s-dimensional disturbance sequences respectively. In (2.1) q^{-1} denotes the unit delay operator, $d \geqq 1$, $A(q^{-1})$ is a scalar polynomial in q^{-1} and $[B(q^{-1})]$, $[C(q^{-1})]$ are matrices whose ijth entries are the scalar polynomials $B_{ij}(q^{-1})$, $C_{ij}(q^{-1})$ respectively. Thus

$$\begin{aligned} A(q^{-1}) &= 1 + a_1 q^{-1} + \cdots + a_n q^{-n}, \\ B_{ij}(q^{-1}) &= (B_{ij})_0 + (B_{ij})_1 q^{-1} + \cdots + (B_{ij})_m q^{-m}, \\ C_{ij}(q^{-1}) &= (C_{ij})_0 + (C_{ij})_1 q^{-1} + \cdots + (C_{ij})_l q^{-l}, \end{aligned} \qquad \text{with } (C_{ij})_0 = \begin{cases} 1, & i = j, \\ 0, & \text{otherwise,} \end{cases}$$

Equation (2.1) is taken together with the initial condition $x_0 \triangleq \{y(0), y(-1), \cdots, y(1-k); u(1-d), \cdots, u(1-k); w(0), \cdots, w(1-k)\}$, where $k = \max\{n, m+d, l\}$.

As is shown in Appendix B, recursions of the ARMAX form (2.1) are equivalent to a large class of finite dimensional linear state space systems.

The process $\{x_0, w(1), w(2), \cdots\}$ is defined on the underlying probability space $(\Omega, \mathcal{F}, P)$, and we define $\mathcal{F}_0$ to be the σ-algebra generated by $\{x_0\}$. Further, for all $t \geqq 1$ $\mathcal{F}_t$ shall denote the σ-algebra generated by $\{x_0, w(1), w(2), \cdots, w(t)\}$. Clearly $\mathcal{F}_0 \subset \mathcal{F}_t \subset \mathcal{F}$ for all $t \geqq 0$. The distributions of the random variables $x_0, \cdots, (x_0, w(0), \cdots, w(t), \cdots$ are assumed mutually absolutely continuous with respect to Lebesgue measure.

The following independence and variance assumptions are made on the process w:

$$E\{w(t)|\mathcal{F}_{t-1}\} = 0 \quad \text{a.s.}, \quad t \geqq 1, \tag{2.2}$$

$$E\{w(t)w^T(t)|\mathcal{F}_{t-1}\} = Q \quad \text{a.s.}, \quad t \geqq 1, \tag{2.3}$$

with trace $Q < \infty$, and

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} \|w(t)\|^2 < \infty \quad \text{a.s.} \tag{2.4}$$

The feedback control actions $u(t)$ are assumed to be measurable with respect to the σ-algebra generated by $\{y(1), \cdots, y(t)\}$ together with $\{u(1), \cdots, u(t-1)\}$ for $t \geqq 2$ and by $\{y(1)\}$ for $t = 1$. Reasoning inductively, we see that for $t \geqq 1$ $u(t)$ is measurable with respect to the algebra generated by $\{y(1), \cdots, y(t)\}$, and we note that via (2.1) this algebra is in general smaller than $\mathcal{F}_t$.

The control problem we treat is an adaptive one because $u(t)$ is not permitted to be an explicit function of the coefficients of $A(q^{-1})$, $[B(q^{-1})]$, $[C(q^{-1})]$, Q, but only depends on these quantities through the observations $y(1), \cdots, y(t)$ and

$u(1), \cdots, u(t-1)$. In other words, the system coefficients are not known to the controller.

Our objective is to design a feedback control law to make $\{u(t)\}$ and $\{y(t)\}$ sample mean square bounded and, whenever it exists, to minimize the limit

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=d}^{N} E\{\|y(t)-y^*(t)\|^2 | \mathcal{F}_{t-d}\}, \tag{2.5}$$

where $\{y^*(t)\}$ is a bounded reference sequence.

3. Single-input single-output systems. For the single-input single-output case the system in (2.1) can be described by

$$A(q^{-1})y(t) = q^{-d}B(q^{-1})u(t) + C(q^{-1})w(t), \qquad t \geqq 1, \tag{3.1}$$

together with the initial condition x_0, where $\{y(t)\}, \{u(t)\}, \{w(t)\}$ denote the scalar output, input and disturbance sequences respectively, and $A(q^{-1}), B(q^{-1}), C(q^{-1})$ are polynomial functions of q^{-1} which we write as

$$\begin{aligned} A(q^{-1}) &= 1 + a_1 q^{-1} + \cdots + a_n q^{-n}, \\ B(q^{-1}) &= b_0 + b_1 q^{-1} + \cdots + b_m q^{-m}, \qquad b_0 \neq 0 \\ C(q^{-1}) &= 1 + c_1 q^{-1} + \cdots + c_l q^{-l}, \end{aligned}$$

respectively.

We shall make the following assumptions about the information set for the computation of the control actions and about the system:

(3A) d is known.

(3B) Upper bounds for n, m and l are known.

(3C) $C(z)$ and $B(z)$ have all zeros outside the closed unit circle.

In other words, (3A) and (3B) state that $u(t)$ may be an explicit function of d and integers bounding n, m and l from above. We observe that the assumption on $C(z)$ is without loss of generality in many circumstances, for instance, in the case where the noise process $w(t)$ is weakly stationary and the spectral density function of $C(q^{-1})w(t)$ has no zeros on the unit circle.

Our control objective is to minimize the sample mean of the sequence $E\{(y(t)-y^*(t))^2 | \mathcal{F}_{t-d}\}$, and we see that this last quantity may be written as

$$E\{(y(t)-y^*(t))^2 | \mathcal{F}_{t-d}\} = E\{v(t)^2 | \mathcal{F}_{t-d}\} + (E\{y(t) | \mathcal{F}_{t-d}\} - y_t^*)^2,$$

where $v(t)$ denotes $y(t) - E\{y(t)|\mathcal{F}_{t-d}\}$. Let us denote the second term on the right of this equation as $z(t-d)^2$, and let us suppose that the first takes the time-invariant value γ^2. Now we shall show that, if we allow $u(t-d)$ to be $\mathcal{F}_{t-d}$ measurable and to be a function of the system parameters, then it may be chosen so that $y_t^* = E\{y(t)|\mathcal{F}_{t-d}\}$ i.e. so that $z^2(t-d) = 0$. Since $u(t-d)$ in our problem is constrained to be measurable with respect to a sigma algebra contained in $\mathcal{F}_{t-d}$, we see that a lower bound for (2.5) is the constant quantity γ^2. In this paper we demonstrate that our adaptive control algorithm minimizes (2.5) by showing that the lower bound γ^2 is achieved. This in turn is performed by showing that our algorithm results in

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=0}^{N} (z(t))^2 = 0 \quad \text{a.s.}$$

The fact that $E\{v(t)^2|\mathcal{F}_{t-d}\}=\gamma^2$ is proven as follows. As in Appendix C, factor $C(q^{-1})$ as

$$C(q^{-1})=F(q^{-1})A(q^{-1})+q^{-d}G(q^{-1}),$$

substitute in (2.1) to obtain

$$C(q^{-1})(y(t)-F(q^{-1})w(t))=q^{-d}G(q^{-1})y(t)+q^{-d}F(q^{-1})B(q^{-1})u(t) \tag{3.2}$$

together with the initial condition x_0, and write this in the equivalent form

$$y(t)-\sum_{k=0}^{d-1}F_k w(t-k)=\sum_{k=0}^{t-d}M_{t-d-k}y(k)+\sum_{k=0}^{t-d}N_{t-d-k}u(k)+L_t x_0, \tag{3.3}$$

where the sequences $\{M_0, M_1, \cdots\}$, $\{N_0, N_1, \cdots\}$ and $\{L_0, L_1, \cdots\}$ are scalar sequences in the case under discussion and in all cases are functions of the polynomials appearing in (2.1). Taking the conditional expectation of both sides of this last equation with respect to the σ-algebra $\mathcal{F}_{t-d}$, using (2.2) and recalling that x_0 is a generator of $\mathcal{F}_t$ for $t\geqq 0$ we obtain

$$E\{y(t)|\mathcal{F}_{t-d}\}=\sum_{k=0}^{t-d}M_{t-d-k}y(k)+\sum_{k=0}^{t-d}N_{t-d-k}u(k)+L_t x_0 \tag{3.4}$$

for $t\geqq d$.

This shows immediately that

$$v(t)\triangleq y(t)-E\{y(t)|\mathcal{F}_{t-d}\}=\sum_{k=0}^{d-1}f_k w(t-k) \quad \text{a.s.}$$

for $t\geqq d$, where we have replaced F_k by f_k for convenience in this scalar case. Further, by (2.5),

$$\begin{aligned}E\{v(t)^2|\mathcal{F}_{t-d}\}&=E\left\{\left(\sum_{k=0}^{d-1}f_k w(t-k)\right)^2\middle|\mathcal{F}_{t-d}\right\}\\&=\sigma^2\sum_{k=0}^{d-1}f_k^2 \quad \text{a.s.}\end{aligned} \tag{3.5}$$

for all $t\geqq d$. We may denote this quantity by γ^2, by virtue of its time-invariant nature. For convenience in the scalar case under consideration we have replaced Q in (3.5) by γ^2.

Now, just as (3.2) and (3.3) were equivalent, so (3.5) may be replaced by the equivalent expression

$$\begin{aligned}C(q^{-1})(E\{y(t+d)|\mathcal{F}_t\})&=C(q^{-1})(y(t+d)-v(t+d))\\&=\alpha(q^{-1})y(t)+\beta(q^{-1})u(t), \qquad t\geqq 0,\end{aligned} \tag{3.6}$$

plus the initial condition x_0, where $\alpha(q^{-1})=G(q^{-1})$ and $\beta(q^{-1})=F(q^{-1})B(q^{-1})$. From this compact form it is apparent that if $u(t)$ is permitted to be $\mathcal{F}_t$ measurable and to be a function of the system parameters then it can be chosen to make $E\{y(t+d)|\mathcal{F}_t\}$ take on any preassigned value. This is the basis of the minimum variance control algorithm of Åström—which inspires our adaptive control strategy—and it justifies our earlier statement that a lower bound for (2.5) is γ^2.

4. A single-input single-output algorithm. Here we shall consider a simple algorithm for stochastic adaptive control. This algorithm uses a stochastic approximation iteration [5] to estimate a set of control law parameters, $\hat{\theta}(t)$. The input, $u(t)$ will then be computed as a function of the current control law parameter estimates and a current estimate of the system state, $\phi(t)$. The algorithm for the case $d=1$ is:

UNIT DELAY ALGORITHM.

$$\hat{\theta}(t)=\hat{\theta}(t-1)+\frac{\bar{a}}{r(t-1)}\phi(t-1)[y(t)-\phi(t-1)^T\hat{\theta}(t-1)], \qquad \bar{a}>0, \quad t\geqq k+1, \tag{4.1}$$

$$r(t-1)=r(t-2)+\phi(t-1)^T\phi(t-1), \qquad r(k-1)=1, \tag{4.2}$$

$$\phi(t)^T\hat{\theta}(t)=y^*(t+1), \tag{4.3}$$

where $k=\max(n, m+1), l)$ and where $\phi(t)$ is given by

$$\begin{aligned}\phi(t-1)^T=[&y(t-1),\cdots,y(t-n),u(t-1),\cdots,u(t-m),\\ &-y^*(t-1),\cdots,-y^*(t-l)].\end{aligned} \tag{4.4}$$

Equations (4.1) and (4.2) constitute the recursive parameter estimator. The choice of the scalar gain $\bar{a}$ will be discussed presently. Equation (4.3) defines a feedback control law. Here and in the sequel we assume that the initial inputs $\{u(1),\cdots,u(k)\}$ and the initial parameter estimates are arbitrarily chosen. The feedback law (4.3) is explicitly given by

$$\begin{aligned}u(t)=\frac{-1}{\hat{\theta}_{n+1}(t)}[&\hat{\theta}_1(t)y(t)+\cdots+\hat{\theta}_n(t)y(t-n+1)\\ &+\hat{\theta}_{n+2}(t)u(t-1)+\cdots+\hat{\theta}_{n+m}(t)u(t-m+1)\\ &-y^*(t+1)-\hat{\theta}_{n+m+1}(t)y^*(t)\cdots-\hat{\theta}_{n+m+l}(t)y^*(t-l+1)], \quad t\geqq k+1,\end{aligned} \tag{4.5}$$

Employing our assumptions on the distribution of the initial conditions and the noise process $\{w(n)\}$ it may be verified inductively that division by zero is a zero probability event. Since all the results in this paper are almost sure (a.s.) results no data-dependent strategy involving the value of $\bar{a}$ is required to avoid this occurrence. This is in contrast to the deterministic case [3].

Alternatively one may make a random choice of $\bar{a}$, independent of past observations at each instant t, by use of an absolutely continuous distribution with respect to Lebesgue measure on $[\varepsilon, \bar{a}]\ 0<\varepsilon<\bar{a}<\infty$. This also clearly guarantees the probability of a zero division is zero and, further, the analysis in this paper covers this modified version of the algorithm with only minor modifications.

It is perhaps worth remarking that the choice of $\phi(t)$ can be motivated by the following observations:

As shown in § 3,

$$C(q^{-1})[y(t+d)-v(t+d)]=G(q^{-1})y(t)+F(q^{-1})B(q^{-1})u(t), \qquad t\geqq 1,$$

with the initial condition x_0. Then, subtracting $C(q^{-1})y^*(t+d)$ from both sides of this equation, we have

$$\begin{aligned}&C(q^{-1})[y(t+d)-y^*(t+d)-v(t+d)]\\ &\quad=G(q^{-1})y(t)+F(q^{-1})B(q^{-1})u(t)-C(q^{-1})y^*(t+d), \qquad t\geqq 1,\end{aligned}$$

which can be written in the form

$$C(q^{-1})[e(t+d)-v(t+d)]=\phi(t)^T\theta_0-y^*(t+d), \qquad t\geqq 1, \tag{4.6}$$

again with x_0, where $e(t+d)=y(t+d)-y^*(t+d)$ is the tracking error and θ_0 is a vector of system parameters. It is evident that if θ_0 was known, $e(t+d)$ would achieve its optimal value $v(t+d)$ if the feedback law is given via $\phi(t)^T\theta_0=y^*(t+d)$. Equation (4.3) is the adaptive analogue of this relation for $d=1$.

To avoid confusion, we stress that equation (4.5) is not anticipative since the desired output $y^*(t+d)$ is known or is computable at time t.

For the general delay case, there are a number of possible modifications of the previous algorithm. For the case $d>1$ but with $C(q^{-1})=1$, we shall analyze the following algorithm:

MULTIPLE RECURSION ALGORITHM.

$$\hat\theta(t)=\hat\theta(t-d)+\frac{\bar a}{\bar r(t-d)}\phi(t-d)[y(t)-\phi(t-d)^T\hat\theta(t-d)], \tag{4.7}$$

$$\bar a>0, \quad t\geqq k+d,$$

$$\bar r(t-d)=\bar r(t-2d)+\phi(t-d)^T\phi(t-d), \tag{4.8}$$

$$t\geqq d+k, \quad \bar r(\tau)=1, \quad \tau=k-d,\cdots,k-1,$$

$$\phi(t)^T\hat\theta(t)=y^*(t+d) \tag{4.9}$$

where

$$\phi(t-d)^T=[y(t-d),\cdots,y(t-d-n+1),u(t-d), \tag{4.10}$$

$$u(t-m-d+1),-y^*(t-1),\cdots,-y^*(t-l+1)].$$

We note that (4.7) to (4.10) actually represent d-interlaced recursions each of which is similar to the unit delay algorithm. The method of interlaced algorithms was first introduced in [3].

We have also recently shown [14] that the following algorithm is globally convergent for the general delay ($d\geqq 1$), colored noise case:

$$\hat\theta(t)=\hat\theta(t-d)+\frac{\bar a}{r(t-d)}\phi(t-d)[y(t)-\phi(t-d)^T\hat\theta(t-d)], \qquad \bar a>0, \quad t\geqq k+d,$$

$$r(t-d)=r(t-d-1)+\phi(t-d)^T\phi(t-d), \qquad r(k-1)=1,$$

$$\phi(t)^T\hat\theta(t)=y^*(t+d),$$

$$\phi(t-d)^T=[y(t-d),\cdots,y(t-d-n+1),u(t-d),\cdots,$$

$$u(t-d-m+1),-y^*(t-1),\cdots,-y^*(t-l+1)].$$

The analysis of the above algorithm is similar to the unit delay and multiple recursion algorithms and thus we will not present the proof here. Details are given in [14].

In the next section we will analyze the unit delay and multiple recursion stochastic adaptive control algorithms introduced above.

5. Analysis of SISO algorithms. The convergence properties of the algorithms introduced in the previous section will be analyzed using a variant of the martingale convergence theorem (see Appendix A). Other results in Appendix A will also be used in the analysis which follows.

THEOREM 5.1. *Let assumptions* 3A, 3B *and* 3C *hold for the system* (2.1), *and assume* $r = s = 1$ *and* $d = 1$. *Further assume that,*

$$\left[C(z) - \frac{\bar{a}}{2}\right] \tag{5.1}$$

is strictly positive real and that the unit delay algorithm (4.1)–(4.4) *is used. Then, with probability one for any initial parameter estimate* $\hat{\theta}(k)$,

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} y(t)^2 < \infty, \tag{5.2}$$

(1)

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} u(t)^2 < \infty, \tag{5.3}$$

(2)

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} E\{(y(t) - y^*(t))^2 | \mathcal{F}_{t-1}\} = \gamma^2, \tag{5.4}$$

(3)

where γ^2 *is the minimum possible mean square control error achievable with any causal feedback. (This includes feedback designed using the true system parameters.)*

Proof. Part 1. In this first section we will establish an important property of the algorithm. In the analysis to follow we take $t \geqq k+1$, and note that all the required initial conditions have been specified. Define

$$e(t) = y(t) - y^*(t) = y(t) - \phi(t-1)^T \hat{\theta}(t-1) \tag{5.5}$$

using (4.3). Let $\tilde{\theta}(t) = \hat{\theta}(t) - \theta_0$ where θ_0 was defined in equation (4.6). Then (4.1) can be written as

$$\tilde{\theta}(t) = \tilde{\theta}(t-1) + \frac{\bar{a}}{r(t-1)} \phi(t-1)^T e(t).$$

Let $V(t) = \tilde{\theta}(t)^T \tilde{\theta}(t)$. Then

$$V(t) = V(t-1) + \frac{2\bar{a}}{r(t-1)} \tilde{\theta}(t-1)^T \phi(t-1)(e(t) - v(t)) + \frac{2\bar{a}}{r(t-1)} \tilde{\theta}(t-1)^T \phi(t-1) v(t)$$

$$+ \frac{\bar{a}^2}{r(t-1)^2} \phi(t-1)^T \phi(t-1)[(e(t) - v(t))^2 + 2v(t)(e(t) - v(t)) + v(t)^2]$$

where $v(t)$ was defined in equation (3.3). Now let

$$b(t-1) = -\tilde{\theta}(t-1)^T \phi(t-1) \tag{5.6}$$

and

$$z(t-1) = e(t) - v(t). \tag{5.7}$$

Note from (4.6) that $e(t) - v(t)$ is $\mathcal{F}_{t-1}$ measurable. Then

$$E[V(t)|\mathcal{F}_{t-1}] = V(t-1) - \frac{2\bar{a}}{r(t-1)} b(t-1) z(t-1)$$

$$+ \frac{\bar{a}^2}{r(t-1)^2} \phi(t-1)^T \phi(t-1) z(t-1)^2$$

$$+ \frac{\bar{a}^2}{r(t-1)^2} \phi(t-1)^T \phi(t-1) \gamma^2 \quad \text{a.s.}$$

So, noting

$$\frac{\phi(t-1)^T\phi(t-1)}{r(t-1)} \leqq 1,$$

we have

$$\begin{aligned} E[V(t)|\mathcal{F}_{t-1}] \leqq V(t-1) &- \frac{2\bar{a}}{r(t-1)}\left\{b(t-1)-\frac{(\bar{a}+\rho)}{2}z(t-1)\right\}z(t-1) \\ &- \rho\bar{a}\frac{z(t-1)^2}{r(t-1)} + \frac{\bar{a}^2}{r(t-1)^2}\phi(t-1)^T\phi(t-1)\gamma^2 \quad \text{a.s.} \end{aligned} \tag{5.8}$$

where ρ is a small positive constant chosen so that

$$\left[C(z)-\frac{\bar{a}+\rho}{2}\right]$$

is positive real. The existence of such a ρ is assured by the strict positive real condition (5.1).

Now let

$$h(t-1) = b(t-1) - \frac{(\bar{a}+\rho)}{2}z(t-1) \tag{5.9}$$

and recalling equations (4.6) and (4.3) we have

$$\begin{aligned} C(q^{-1})[z(t-1)] &= \phi(t-1)^T\theta_0 - y^*(t) \\ &= -\phi(t-1)^T\tilde{\theta}(t-1) \\ &= b(t-1). \end{aligned}$$

Hence

$$h(t-1) = \left[C(q^{-1}) - \frac{\bar{a}+\rho}{2}\right]z(t-1). \tag{5.10}$$

Equation (5.8) can now be written as

$$\begin{aligned} E[V(t)|\mathcal{F}_{t-1}] \leqq V(t-1) &- \frac{2\bar{a}}{r(t-1)}h(t-1)z(t-1) - \frac{\rho\bar{a}z(t-1)^2}{r(t-1)} \\ &+ \frac{\bar{a}^2}{r(t-1)^2}\phi(t-1)^T\phi(t-1)\gamma^2 \quad \text{a.s.} \end{aligned} \tag{5.11}$$

Since we intend to use Lemma A.3 from Appendix A we now define

$$S(t) = 2\bar{a}\sum_{j=1}^{t} h(j-1)z(j-1) + K, \qquad 0 < K < \infty, \tag{5.12}$$

and note that condition (5.1) of the theorem statement together with Lemma A.4 of Appendix A ensure $S(t) \geqq 0$ for some K, $0 < K < \infty$.

Now define the nonnegative random variable

$$Z(t) = V(t) + \frac{S(t)}{r(t-1)}. \tag{5.13}$$

So

$$E[Z(t)|\mathcal{F}_{t-1}]=E[V(t)|\mathcal{F}_{t-1}]+\frac{S(t)}{r(t-1)}$$

$$\leqq V(t-1)+\frac{S(t-1)}{r(t-1)}-\frac{\rho\bar{a}z(t-1)^2}{r(t-1)}$$

$$+\frac{\bar{a}^2}{r(t-1)^2}\phi(t-1)^T\phi(t-1)\gamma^2 \quad \text{a.s.},$$

where we have used (5.11).

Next, since $r(t-2)\leqq r(t-1)$, we obtain

$$E[Z(t)|\mathcal{F}_{t-1}]\leqq V(t-1)+\frac{S(t-1)}{r(t-2)}-\frac{\rho\bar{a}z(t-1)^2}{r(t-1)}+\frac{\bar{a}^2}{r(t-1)^2}\phi(t-1)^T\phi(t-1)\gamma^2$$

$$=Z(t-1)-\frac{\rho\bar{a}z(t-1)^2}{r(t-1)}+\frac{\bar{a}^2}{r(t-1)^2}\phi(t-1)^T\phi(t-1)\gamma^2 \quad \text{a.s.}$$

By Lemma A.2 of Appendix A

$$\sum_{j=k+1}^{\infty}\frac{\phi(j-1)^T\phi(j-1)}{r(j-1)^2}<\infty.$$

So applying Lemma A.3 of Appendix A yields

$$Z(t)\to Z \quad \text{a.s. with } E\{Z\}<\infty,$$

and

$$\sum_{t=k+1}^{\infty}\frac{\rho\bar{a}z(t-1)^2}{r(t-1)}<\infty \quad \text{a.s.}$$

Now since $\rho\bar{a}\neq 0$ we conclude

$$\sum_{t=1}^{\infty}\frac{z(t)^2}{r(t)}<\infty \quad \text{a.s.}$$

Thus using Kronecker's Lemma [9, p. 117] we have

$$\lim_{N\to\infty}\frac{N}{r(N)}\frac{1}{N}\sum_{t=1}^{N}z(t)^2=0 \quad \text{a.s.} \tag{5.14}$$

This characterizes an important property of the recursive parameter identifier (4.1), (4.2) and control law (4.3). It now remains to be shown that this property ensures the theorem conclusions (5.2), (5.3) and (5.4).

Part 2. Here we show that the condition (5.14) is a sufficient condition to ensure the conclusions of the theorem statement. Firstly using assumption 3C and Lemma A.5 of Appendix A and (2.4), it follows that there exists an N' such that

$$\frac{1}{N}\sum_{t=k+1}^{N}u(t)^2\leqq\frac{K_1}{N}\sum_{t=k}^{N}y(t+1)^2+K_2 \quad \text{for } N>N' \quad \text{a.s.} \tag{5.15}$$

and hence, using the definition of $r(N)$ and $\phi(t)$, that

$$\frac{r(N)}{N}\leqq\frac{K_3}{N}\sum_{t=k}^{N}y(t+1)^2+K_4 \quad \text{for } N>N' \quad \text{a.s.} \tag{5.16}$$

Now by definition $e(i) \triangleq y(i) - y^*(i)$ so

$$z(i-1) \triangleq e(i) - v(i) = y(i) - y^*(i) - v(i),$$

where $v(i)$ is defined via (3.3), and

$$y(i) = z(i-1) + y^*(i) + v(i)$$

with $|y^*(t)| < M < \infty$; hence

$$\frac{1}{N}\sum_{i=1}^{N} y(i+1)^2 \leqq \frac{3}{N}\sum_{i=1}^{N} z(i)^2 + M_2 + \frac{3}{N}\sum_{i=1}^{N} v(i+1)^2.$$

But now since

$$\limsup_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} w(i)^2 < \infty \quad \text{a.s.},$$

it follows that

$$\limsup_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} v(i)^2 < \infty \quad \text{a.s.}$$

Hence there exists an N'' such that

$$\frac{1}{N}\sum_{i=1}^{N} y(i+1)^2 \leqq \frac{3}{N}\sum_{i=1}^{N} z(i)^2 + M_3, \quad \text{for } N \geqq N'' \quad \text{a.s.} \tag{5.17}$$

Hence using (5.16) and (5.17) we have

$$\frac{r(N)}{N} \leqq \frac{C_1}{N}\sum_{i=1}^{N} z(i)^2 + C_2 \quad \text{for } N > \bar{N} \quad \text{a.s.} \tag{5.18}$$

with $0 < C_1 < \infty$, $0 < C_2 < \infty$ and $\bar{N} = \max(N', N'')$.

Having established the above bounds we are now in a position to prove the required result. We proceed using a sample path analysis on the set of paths of measure one that satisfy (5.18).

First let us assume that the sequence

$$\frac{1}{N}\sum_{i=1}^{N} y(i+1)^2$$

is not bounded. Then it follows by the definition of $r(N)$ that

$$\limsup_{N\to\infty} \frac{r(N)}{N} = \infty \tag{5.19}$$

and using (5.18)

$$\limsup_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} z(i)^2 = \infty. \tag{5.20}$$

Let

$$\bar{z}(N) = \frac{1}{N}\sum_{i=1}^{N} z(i)^2. \tag{5.21}$$

Then from (5.18)

$$\left(\frac{r(N)}{N}\right)^{-1}\frac{1}{N}\sum_{i=1}^{N} z(i)^2=\left(\frac{r(N)}{N}\right)^{-1}\bar{z}(N)\geqq\frac{\bar{z}(N)}{C_1\bar{z}(N)+C_2}\quad\text{for } N>\bar{N}.$$

But since

$$\limsup_{N\to\infty}\bar{z}(N)=\infty$$

there exists a subsequence $\{N_K\}$ such that

$$\lim_{K\to\infty}\bar{z}(N_K)=\infty.$$

Then

$$\liminf_{K\to\infty}\left[\frac{r(N_K)}{N_K}\right]^{-1}\frac{1}{N_K}\sum_{i=1}^{N_K} z(i)^2\geqq\frac{1}{C_1}$$

which contradicts (5.14). Thus our assumption that

$$\frac{1}{N}\sum_{i=1}^{N} y(i+1)^2$$

was not bounded in N was false.

Thus

$$\frac{1}{N}\sum_{i=1}^{N} y(i+1)^2$$

is bounded in N. So from (5.16)

$$\limsup_{N\to\infty}\frac{r(N)}{N}<\infty,$$

and hence

$$\liminf_{N\to\infty}\frac{N}{r(N)}>\frac{1}{K}>0.$$

Then from (5.14) we have that

$$\lim_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N} z(t)^2=0\quad\text{a.s.}\tag{5.22}$$

but

$$\begin{aligned} z(i-1)&=e(i)-v(i)\\ &=y(i)-y^*(i)-v(i).\end{aligned}$$

Now $z(i-1)$ is $\mathcal{F}_{i-1}$ measurable since $y^*(i)$ and $y(i)-v(i)=E\{y(i)|\mathcal{F}_{i-1}\}$ are $\mathcal{F}_{i-1}$ measurable. Thus

$$E\{(y(i)-y^*(i))^2|\mathcal{F}_{i-1}\}=z(i-1)^2+\gamma^2 \quad \text{a.s.}$$

and so from (5.22) we obtain

$$\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}E\{(y(i+1)-y^*(i+1))^2|\mathcal{F}_i\}=\gamma^2 \quad \text{a.s.}$$

which completes the proof. □

The positive real condition of Theorem 5.1 is to be expected. Similar conditions have been noted previously for recursive parameter estimation schemes [5], [6]. It is interesting that in the case presented here, the positive real condition is a function of the algorithm gain constant, $\bar{a}$. The weakest condition is obtained for $\bar{a}$ small, though this may affect other properties such as the convergence rate.

The next theorem examines the multiple recursion algorithm (4.7) to (4.9) for the case $d\geqq 1$, $C(q^{-1})=1$.

THEOREM 5.2. *Let assumptions* 3A, 3B, 3C *hold for the system* (2.1), *and assume* $r=s=1$, $d\geqq 1$ *and* $C(q^{-1})=1$. *When the multiple recursion algorithm,* (4.7) *to* (4.9), *is used with* $0<\bar{a}<2$ *then with probability one*

(5.23) (1) $$\limsup_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}y(t)^2<\infty,$$

(5.24) (2) $$\limsup_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}u(t)^2<\infty,$$

(5.25) (3) $$\lim_{N\to\infty}\frac{1}{N}\sum_{t=d}^{N}E\{(y(t)-y^*(t))^2|\mathcal{F}_{t-d}\}=\gamma^2,$$

where γ^2 *is the minimum possible mean square control error achievable with any causal feedback.*

Proof. We first establish the property corresponding to (5.14). Throughout the following analysis we take $t\geqq k+d$, and again note that all required initial conditions have been specified. As before define

$$\begin{aligned}e(t)&=y(t)-y^*(t)\\&=-\phi(t-d)^T\tilde{\theta}(t-d)+v(t)\end{aligned}$$

and

$$V(t)=\tilde{\theta}(t)^T\tilde{\theta}(t).$$

We shall analyze each of the interlaced algorithms separately. Therefore consider

$$V(t)=V(t-d)+\frac{2\bar{a}}{\bar{r}(t-d)}\phi(t-d)^T\tilde{\theta}(t-d)e(t)+\frac{\bar{a}^2}{\bar{r}(t-d)^2}\|\phi(t-d)\|^2e(t)^2.$$

Let

$$b(t-d)=-\phi(t-d)^T\tilde{\theta}(t-d).$$

Then

$$V(t)=V(t-d)-\frac{2\bar{a}}{\bar{r}(t-d)}b(t-d)[b(t-d)+v(t)]$$
$$+\frac{\bar{a}^2}{\bar{r}(t-d)^2}\|\phi(t-d)\|^2[b(t-d)^2+2b(t-d)v(t)+v(t)^2],$$

$$E[V(t)|\mathcal{F}_{t-d}]\leqq V(t-d)-\frac{\bar{a}}{\bar{r}(t-d)}[2-\bar{a}]b(t-d)^2+\frac{\bar{a}^2}{\bar{r}(t-d)^2}\|\phi(t-d)\|^2\gamma^2.$$

The sequences

$$\{V(i+nd)\},\ \left\{\frac{b(i+nd)}{\bar{r}(i+nd)}\right\}\left\{\frac{\|\phi(i+nd)\|^2}{\bar{r}(i+nd)^2}\right\}$$

are adapted to the increasing sequence of σ-algebras $\mathcal{F}_{i+nd}$ for $1\leqq i\leqq d$. Hence we can use Lemmas A.2 and A.3 of Appendix A to conclude

$$\bar{a}(2-\bar{a})\sum_{n=0}^{\infty}\frac{b(i+nd)^2}{\bar{r}(i+nd)}<\infty \quad \text{a.s.} \quad \text{for } i=1,\cdots,d. \tag{5.26}$$

Since $0<\bar{a}<2$, summing (5.26) over $1\leqq i\leqq d$ we have

$$\sum_{t=1}^{\infty}\frac{b(t)^2}{\bar{r}(t)}<\infty \quad \text{a.s.} \tag{5.27}$$

Now, as before define

$$r(t)=r(t-1)+\phi(t)^T\phi(t), \qquad r(0)=1. \tag{5.28}$$

It follows from (5.28) and (4.8) that

$$r(t)=1+\sum_{j=1}^{t}\phi(j)^T\phi(j)\geqq 1+\sum_{i=0}^{[t/d]}\phi(t-id)^T\phi(t-id)=\bar{r}(t). \tag{5.29}$$

Hence from (5.27)

$$\sum_{t=1}^{\infty}\frac{b(t)^2}{r(t)}<\infty \quad \text{a.s.} \tag{5.30}$$

Now we have

$$b(t)=-\phi(t)^T\tilde{\theta}(t)=\phi(t)^T\theta_0-\phi(t)^T\hat{\theta}(t)$$
$$=y(t+d)-v(t+d)-y^*(t+d)$$

which corresponds to $z(t)$ in the proof of Theorem 5.1.

Thus applying Kronecker's lemma to (5.27) gives

$$\lim_{N\to\infty} \frac{N}{r(N)} \cdot \frac{1}{N} \sum_{t=1}^{N} z(t)^2 = 0 \quad \text{a.s.} \tag{5.31}$$

The remainder of the proof then follows that of Part 2 of Theorem 5.1. □

6. An alternative algorithm. As discussed in [3], there are alternative adaptive control algorithms to the one discussed in the previous section. One possibility is to factor the modulus of the leading coefficient of $u(t)$ from (3.2). Thus for $d = 1$, $t \geqq k+1$, and using the initial condition x_0, we have

$$C(q^{-1})[y(t+1) - v(t+1)] = |\beta_0|[\alpha'(q^{-1})y(t) + (\operatorname{Sgn}\beta_0)u(t) + \beta'(q^{-1})u(t)]. \tag{6.1}$$

Subtracting $C(q^{-1})y^*(t+1)$ from both sides gives

$$\begin{aligned} &C(q^{-1})[e(t+1) - v(t+1)] \\ &\quad = |\beta_0|[\alpha'(q^{-1})y(t) + (\operatorname{Sgn}\beta_0)u(t) + \beta'(q^{-1})u(t)] - C(q^{-1})y^*(t+1) \\ &\quad = |\beta_0|\left[\alpha'(q^{-1})y(t) + (\operatorname{Sgn}\beta_0)u(t) + \beta'(q^{-1})u(t) + \frac{1}{|\beta_0|} C(q^{-1})y^*(t+1)\right] \\ &\quad = |\beta_0|[\phi(t)^T\theta_0 + (\operatorname{Sgn}\beta_0)u(t)] \end{aligned} \tag{6.2}$$

where

$$\phi(t)^T = [y(t), \cdots, y(t-n+1), u(t-1), \cdots, u(t-m), y^*(t+1), \cdots, y^*(t-l)].$$

The above considerations motivate the following algorithm:

$$\hat{\theta}(t) = \hat{\theta}(t-1) + \bar{a}\phi(t-1)\frac{1}{r(t-1)}[y(t) - y^*(t)], \tag{6.3}$$

$$r(t) = r(t-1) + \phi(t-1)^T\phi(t-1), \qquad r(0) = 1, \tag{6.4}$$

$$u(t) = -(\operatorname{Sgn}\beta_0)\phi(t)^T\hat{\theta}(t). \tag{6.5}$$

As before, equations (6.3), (6.4) consist of a recursive parameter estimator and (6.5) defines a feedback control law.

We then have the following theorem.

THEOREM 6.1. *Subject to assumptions* 3A, 3B *and* 3C *if the algorithm* (6.3) *to* (6.5) *is applied to the system* (2.1), *with* $r = s = 1$, $d = 1$ *then, provided*

(i) *the sign of* β_0 *is known and*

(ii) $[C(z) - \frac{1}{2}\bar{a}|\beta_0|]$ *is strictly positive real,* (6.6)

it follows that with probability one that

$$\text{(1)} \qquad \limsup_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} y(t)^2 < \infty, \tag{6.7}$$

(6.8) (2) $$\limsup_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}u(t)^2<\infty,$$

(6.9) (3) $$\lim_{N\to\infty}\frac{1}{N}\sum_{t=d}^{N}E\{[y(t)-y^*(t)]^2|\mathcal{F}_{t-d}\}=\gamma^2,$$

where γ^2 is the minimum possible mean square control error achievable with any causal feedback.

Proof. The proof follows that of Theorem 5.1 with the following correspondences:

(6.10) $$e(t)=y(t)-y^*(t),$$

(6.11) $$b(t)=-\phi(t)^T\tilde{\theta}(t),$$

(6.12) $$z(t)=e(t+1)-v(t+1),$$

(6.13) $$h(t)=b(t)-\left(\frac{\bar{a}+\rho}{2}\right)z(t),$$

(6.14) $$C(q^{-1})z(t)=|\beta_0|b(t).$$ □

It will be noted that to employ the above algorithm it is necessary to know the sign of β_0 and the positive real condition is weakest when $\bar{a}|\beta_0|$ is small. The approach of factoring $|\beta_0|$ from (6.1) is related to the procedure used by Åström and Wittenmark in [8] where a fixed estimate of β_0 was used.

7. Multiple-input multiple-output systems. In the multiple-input multiple-output case the system output is described by

(7.1) $$A(q^{-1})y(t)=q^{-d}[B(q^{-1})]u(t)+[C(q^{-1})]w(t)$$

where $[M(q^{-1})]$ denotes a matrix whose ij-th entry is the scalar polynomial $M_{ij}(q^{-1})$. In (7.1), $\{y(t)\}$, $\{u(t)\}$, $\{w(t)\}$ denote the s, r and s component vectors of output, input and disturbance sequences, respectively. d denotes a pure time delay. As in the previous sections, we take $t\geqq k+d$ and note that all required initial conditions have been specified.

The following assumptions will be made about the system:

(7A) The number of inputs r, equals the number of outputs s;

(7B) d is known;

(7C) Upper bounds for the orders of all scalar polynomials appearing in $\{A(q^{-1}), [B(q^{-1})]$ and $[C(q^{-1})]\}$ are known;

(7D) $$\det[B(z)]\neq 0, \qquad |z|\leqq 1,$$
$$\det[C(z)]\neq 0, \qquad |z|\leqq 1.$$

It is shown in Appendix C, that (7.1) can be manipulated into the following prediction form:

(7.2) $$\bar{c}(q^{-1})\{y(t+d)-v(t+d)\}=[\alpha(q^{-1})]y(t)+[\beta(q^{-1})]u(t)$$

where, as in Appendix C,

(7.3) $$v(t)=[F(q^{-1})]w(t)$$

and $\bar{c}(q^{-1})$ is the scalar polynomial given by $\det[C(q^{-1})]$. Let

$$E[v(t+d)v(t+d)^T|\mathcal{F}_t]=\Gamma. \tag{7.4}$$

The control objective is to achieve with probability one

$$\limsup_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}\|y(t)\|^2<\infty, \tag{7.5}$$

$$\limsup_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}\|u(t)\|^2<\infty, \tag{7.6}$$

$$\lim_{N\to\infty}\frac{1}{N}\sum_{t=d}^{N}E\{[y_i(t)-y_i^*(t)]^2|\mathcal{F}_{t-d}\}=\Gamma_{ii},\quad i=1,\cdots,s, \tag{7.7}$$

where $\{y^*(t)\}$ is a bounded reference vector sequence and Γ_{ii} is the minimum mean square tracking error for causal linear feedback.

THE MIMO ALGORITHM. We shall begin with a closer analysis of equation (7.2).

Define $[\alpha_{i.}(q^{-1})]$ to be the ith row of $[\alpha(q^{-1})]$ and similarly for $[\beta_{i.}(q^{-1})]$. Then we have

$$\bar{c}(q^{-1})(y_i(t+d)-v_i(t+d))=[\alpha_{i.}(q^{-1})]y(t)+[\beta_{i.}(q^{-1})]u(t). \tag{7.8}$$

Now subtracting $\bar{c}(q^{-1})y_i^*(t+d)$ from both sides of the above equation yields

$$\bar{c}(q^{-1})(y_i(t+d)-y_i^*(t+d)-v_i(t+d))=[\alpha_{i.}(q^{-1})]y(t)+[\beta_{i.}(q^{-1})]u(t)-\bar{c}(q^{-1})y_i^*(t+d)$$

or

$$\bar{c}(q^{-1})(e_i(t+d)-v_i(t+d))=\phi_i(t)^T\theta_i^0-y_i^*(t+d) \tag{7.9}$$

where $e_i(t+d)=y_i(t+d)-y_i^*(t+d)$,

$$\phi_i(t)^T=(y(t)^T,y(t-1)^T,\cdots,u(t)^T,u(t-1)^T,\cdots,y_i^*(t+d-1),\cdots)$$

and θ_i^0 is a vector of system parameters.

The above considerations motivate the following algorithm:

$$\hat{\theta}_i(t)=\hat{\theta}_i(t-1)+\frac{\bar{a}}{r_i(t-1)}\phi_i(t-1)(y_i(t)-\phi_i(t-1)^T\hat{\theta}_i(t-1)), \tag{7.10}$$

$$r_i(t-1)=r_i(t-2)+\phi_i(t-1)^T\phi_i(t-1), \tag{7.11}$$

$$\phi_i(t)^T\hat{\theta}_i(t)=y_i^*(t+1) \tag{7.12}$$

for $i=1,\cdots,r$, and where we have taken d to be 1.

Now, as in the scalar case, (7.12) is an implicit definition of a feedback control law. Assumption 7D ensures that the set of simultaneous equations (7.12), $i=1,\cdots,r$ can be uniquely solved for the vector $u(t)$.

Analogously to the single-input single-output case we have the following theorem:

THEOREM 7.1. *Subject to assumptions* 7A *through* 7D; *if the algorithm* (7.10)–(7.12) *is applied to the system* (2.1), *with* $r = s > 1$, $d = 1$ *and if*

$$\left[\bar{c}(z) - \frac{\bar{a}}{2}\right] \quad \textit{is strictly positive real,} \tag{7.13}$$

then with probability one

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} \|y(t)\|^2 < \infty, \tag{7.14}$$

(1)

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} \|u(t)\|^2 < \infty, \tag{7.15}$$

(2)

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=d}^{N} E\{(y_i(t) - y_i^*(t))^2 | \mathcal{F}_{t-d}\} = \Gamma_{ii}, \qquad i = 1, \cdots, s, \tag{7.16}$$

(3)

where Γ_{ii} *is the ii-th element of* Γ.

Proof. Part 1. The proof proceeds as for Theorem 5.1, Part 1 with the following correspondences for $i = 1, \cdots, s$:

$$\begin{aligned}
e(t) \equiv e_i(t) &= y_i(t) - y_i^*(t) \\
&= y_i(t) - \phi_i(t-1)^T \hat{\theta}_i(t-1); \\
b(t) \equiv b_i(t) &= -\phi_i(t)^T \tilde{\theta}_i(t); \\
z(t) \equiv z_i(t) &= e_i(t+1) - v_i(t+1); \\
h(t) \equiv h_i(t) &= b_i(t) - \left(\frac{\bar{a} + \rho}{2}\right) z_i(t);
\end{aligned}$$

$$\begin{aligned}
\bar{c}(q^{-1}) z(t-1) \equiv \bar{c}(q^{-1}) z_i(t-1) &= \phi_i(t-1)^T \theta_i^0 - y_i^*(t) \\
&= -\phi_i(t-1)^T \tilde{\theta}_i(t-1) \\
&= b_i(t-1).
\end{aligned}$$

This leads to the following important set of properties of the algorithm (7.10) to (7.12)

$$\lim_{N\to\infty} \frac{N}{r_i(N)} \frac{1}{N} \sum_{t=1}^{N} z_i(t)^2 = 0 \quad \text{a.s.}, \quad i = 1, \cdots, s. \tag{7.17}$$

Part 2. As before we proceed to show that the above condition is a sufficient condition to ensure the conclusions of the theorem.

In the following we use a sample path analysis. Firstly, using assumption 7C and Lemma A.5 and (2.4) it follows that there exists an N' such that

$$\frac{1}{N} \sum_{t=1}^{N} \|u(t)\|^2 \leqq \frac{K_1}{N} \sum_{t=0}^{N} \|y(t+1)\|^2 + K_2, \qquad N > N',$$

and hence that

$$\frac{r_i(N)}{N} \leqq \frac{K_3}{N} \sum_{t=0}^{N} \|y(t+1)\|^2 + K_4, \qquad N > N'. \tag{7.18}$$

Also, there exists an N'' such that

$$\frac{1}{N} \sum_{k=1}^{N} y_i(k+1)^2 \leqq \frac{3}{N} \sum_{k=1}^{N} z_i(k)^2 + M_3, \qquad N \geqq N''. \tag{7.19}$$

Hence, using (7.18) and (7.19)

$$\frac{r_i(N)}{N} \leqq \frac{C_1}{N} \max_{1 \leqq j \leqq m} \sum_{k=1}^{N} z_j(k)^2 + C_2, \quad \text{for } N > \bar{N} \tag{7.20}$$

and $0 < C_1 < \infty$, $0 < C_2 < \infty$, $i = 1, \cdots, s$.

We next assume that

$$\frac{1}{N} \sum_{k=1}^{N} \|y(k+1)\|^2$$

is not bounded in N. Hence, there exists at least one j, $1 \leqq j \leqq s$ such that

$$\frac{1}{N} \sum_{k=1}^{N} z_j(k)^2$$

is unbounded by (7.19). Thus from (7.20), there exists a subsequence $\{N_k\}$ and an integer l, $1 \leqq l \leqq s$ such that

$$\frac{r_l(N_k)}{N_k} \leqq \frac{C_1}{N_k} \sum_{t=1}^{N_k} z_l(t)^2 + C_2 \quad \text{for } N_k > \bar{N}$$

and

$$\lim_{k \to \infty} \frac{1}{N_k} \sum_{t=1}^{N_k} z_l(t)^2 = \infty.$$

Defining

$$\bar{z}_l(N_k) = \frac{1}{N_k} \sum_{t=1}^{N_k} z_l(t)^2,$$

then along the subsequence $\{N_k\}$:

$$\left(\frac{r_l(N_k)}{N_k}\right)^{-1} \frac{1}{N_k} \sum_{t=1}^{N_k} z_l(t)^2 \geqq \frac{\bar{z}_l(N_k)}{C_1 \bar{z}_l(N_k) + C_2} \quad \text{for } N_k > \bar{N}.$$

But, since $\lim_{k \to \infty} \bar{z}_l(N_k) = \infty$, it follows that

$$\liminf_{k \to \infty} \left(\frac{r_l(N_k)}{N_k}\right)^{-1} \frac{1}{N_k} \sum_{t=1}^{N_k} z_l(t)^2 \geqq \frac{1}{C_1},$$

which contradicts (7.17). Thus our assumption that

$$\frac{1}{N}\sum_{t=1}^{N}\|y(t)\|^2$$

was not bounded in N was false.

The remainder of the proof is straightforward and follows the proof of Theorem 5.1. □

8. Conclusions. The paper has analyzed several discrete time stochastic adaptive control algorithms and has shown that, under suitable conditions, they will be globally convergent. The algorithms have a simple structure and are applicable to both single-input single-output systems and multi-input multi-output systems. These results are believed to constitute the first complete and rigorous analysis of any stochastic adaptive control algorithm of this type.

Appendix A. The following technical results will be called upon in our analysis of adaptive stochastic control algorithms.

LEMMA A.1. *Consider the asymptotically stable n-th order time invariant linear system*:

$$x(t+1) = Ax(t) + Bz(t),$$
$$h(t) = Cx(t) + Dz(t)$$

with $h(t)$, $z(t)$ the $s \times 1$ output and $r \times 1$ input vectors, respectively and $x(t)$ the $n \times 1$ state vector.

There exist constants C_1 and C_2 which are independent of N such that

$$\sum_{t=1}^{N}\|h(t)\|^2 \leqq C_1 \sum_{t=0}^{N}\|z(t)\|^2 + C_2 \quad \textit{for all } N \in \mathbb{N},$$

$0 < C_1 < \infty$ $0 \leqq C_2 < \infty$.

Proof. Consider

$$x(t+1) = Ax(t) + Bz(t), \qquad x(0) = x_0,$$
$$h(t) = Cx(t) + Dz(t).$$

Then

$$h(t) = cA^t x_0 + Dz(t) + \sum_{j=1}^{t} cA^j Bz(t-j).$$

So

$$\|h(t)\|^2 \leqq 3\left\{\|C\|^2\|A^t\|^2\|x_0\|^2 + \|D\|^2\|z(t)\|^2 + \left[\sum_{j=1}^{t}\|C\|\,\|A\|^j\|B\|\,\|z(t-j)\|\right]^2\right\}$$
$$\leqq K_1\lambda^{2t} + K_2\|z(t)\|^2 + K_3\left[\sum_{j=1}^{t}\lambda^j\|z(t-j)\|\right]^2$$

where we have used the fact that if A is asymptotically stable then $\|F\|^j \leqq K\lambda^j$, $0 \leqq \lambda < 1$

and $0 \leqq K < \infty$ [11, p. 174]. Thus

$$\|h(t)\|^2 \leqq K_1\lambda^{2t} + K_2\|z(t)\|^2 + K_3\left[\sum_{j=1}^{t} \lambda^{j/2}\lambda^{j/2}\|z(t-j)\|\right]^2$$

$$\leqq K_1\lambda^{2t} + K_2\|z(t)\|^2 + K_3 \sum_{j=1}^{t} \lambda^j \sum_{j=1}^{t} \lambda^j\|z(t-j)\|^2.$$

So

$$\sum_{t=1}^{N} \|h(t)\|^2 \leqq K_4 + K_2 \sum_{t=1}^{N} \|z(t)\|^2 + K_5 \sum_{t=1}^{N} \sum_{j=1}^{t} \lambda^j\|z(t-j)\|^2.$$

Introducing $\tau = t-j$

$$\sum_{t=1}^{N} \|h(t)\|^2 \leqq K_4 + K_2 \sum_{t=1}^{N} \|z(t)\|^2 + K_5 \sum_{\tau=0}^{N-1} \sum_{t=\tau}^{N} \lambda^{t-\tau}\|z(\tau)\|^2$$

$$\leqq K_4 + K_2 \sum_{t=1}^{N} \|z(t)\|^2 + K_6 \sum_{\tau=0}^{N-1} \|z(\tau)\|^2$$

$$\leqq C_1 + C_2 \sum_{t=0}^{N} \|z(t)\|^2. \qquad \square$$

LEMMA A.2. *Let $\{\sigma(t)\}$ be a real n-vector sequence. Define $r(t) = r(t-1) + \sigma(t)^T\sigma(t)$, with $r_0 = 1$. Then*

$$\sum_{t=1}^{\infty} \frac{\sigma(t)^T\sigma(t)}{r(t)^2} < \infty.$$

Proof.

$$\frac{\sigma(t)^T\sigma(t)}{r(t)^2} \leqq \frac{\sigma(t)^T\sigma(t)}{r(t)r(t-1)}$$

$$= \frac{r(t)-r(t-1)}{r(t)r(t-1)}$$

$$= \frac{1}{r(t-1)} - \frac{1}{r(t)}.$$

Hence

$$\sum_{1}^{\infty} \frac{\sigma(t)^T\sigma(t)}{r(t)^2} \leqq \sum_{1}^{\infty} \frac{1}{r(t-1)} - \frac{1}{r(t)} \leqq \frac{1}{r_0} < \infty. \qquad \square$$

LEMMA A.3. (martingale convergence theorem). *Let $\{T_n\}$, $\{\alpha_n\}$, $\{\beta_n\}$ be sequences of nonnegative random variables adapted to an increasing sequence of σ-algebras $\mathcal{F}_n$ such that*

$$E[T_n|\mathcal{F}_{n-1}] \leqq T_{n-1} - \alpha_{n-1} + \beta_{n-1}.$$

If $\sum_1^\infty \beta_n < \infty$, a.s., then T_n converges almost surely to a finite random variable T and $\sum_1^\infty \alpha_n < \infty$ a.s.

Proof. See also Neveu [8], Solo [5].

$$E[T_n|\mathcal{F}_{n-1}] \leqq T_{n-1} + \beta_{n-1}.$$

Then following [8, p. 33], we conclude $T_n \to T$ a.s. with T a nonnegative finite random variable.

Now define

$$Z_n = T_n + \sum_{l=1}^{n-1} \alpha_l.$$

So

$$E[Z_n|\mathcal{F}_{n-1}] \leqq T_{n-1} - \alpha_{n-1} + \sum_{i=1}^{n-1} \alpha_i + \beta_{n-1}$$

$$= Z_{n-1} + \beta_{n-1}.$$

Thus again using [8, p. 33] we have $Z_n \to Z$ a.s. where Z is a nonnegative a.s. finite random variable.

Thus

$$\sum_{1}^{\infty} \alpha_n < \infty \quad \text{a.s.} \qquad \square$$

LEMMA A.4 (positive real lemma). *Consider the following minimal state space model*

$$x(t+1) = Ax(t) + Bz(t), \qquad x(0) = x_0,$$
$$h(t) = Cx(t) + Dz(t).$$

Then if the complex function

$$Z(z) = C[zI - A]^{-1}B + D$$

is positive real:

(a) *there exist matrices P, L, W,*

$$P > 0$$

such that

$$A^T PA - P = -LL^T,$$

$$A^T PB = C^T - LW,$$

$$W^T W = D + D^T - B^T PB;$$

(b)

$$2\sum_{i=1}^{t-1} z(i)^T h(i) + x_0^T P x_0 \geqq 0.$$

Proof. (a) See Hitz and Anderson [12].

(b) The results of part (a) will be used. Consider

$$\begin{aligned}
x(n)^T Px(n) &= (Ax(n-1)+Bz(n-1))^T P(Ax(n-1)+Bz(n-1)) \\
&= x(n-1)^T A^T PAx(n-1)+2x(n-1)^T A^T PBz(n-1) \\
&\quad + z(n-1)^T B^T PBz(n-1) \\
&= x(n-1)^T Px(n-1)-x(n-1)^T LL^T x(n-1)+2x(n-1)^T[C^T-LW]z(n-1) \\
&\quad -z(n-1)^T[D+D^T-W^TW]z(n-1) \\
&= x(n-1)^T Px(n-1)-[L^T x(n-1)+Wz(n-1)]^T[L^T x(n-1)+Wz(n-1)] \\
&\quad +2[h(n-1)-Dz(n-1)]^T z(n-1)+z(n-1)^T[D+D^T]z(n-1) \\
&= x(n-1)^T Px(n-1)-[L^T x(n-1)+Wz(n-1)]^T[L^T x(n-1)+Wz(n-1)] \\
&\quad +2h(n-1)^T z(n-1).
\end{aligned}$$

Thus summing from 1 to N we have

$$0 \leqq x(N)^T Px(N) \leqq 2\sum_{t=1}^{N-1} h(t)^T z(t)+x(0)^T Px(0). \qquad \square$$

LEMMA A.5. *Consider the system* (2.1), (2.2), *Subject to assumption* 3C *or* 7C

$$\frac{1}{N}\sum_{t=1}^{N}\|u(t)\|^2 \leqq \frac{K_1}{N}\sum_{t=1}^{N}\|y(t+1)\|^2+\frac{K_2}{N}\sum_{t=1}^{N}\|w(t+1)\|^2+\frac{K_3}{N}.$$

Proof. In view of assumption 3C or 7C $u(t)$ can be considered as the output of an asymptotically stable linear system with inputs $\{y(t)\}$ and $\{w(t)\}$. Hence

$$\begin{aligned}
x(t+1) &= Ax(t)+B_1 v_1(t)+B_2 v_2(t), \\
u(t) &= Cx(t)+D_1 v_1(t)+D_2 v_2(t)
\end{aligned}$$

where

$$v_1(t) = y(t+d)$$

and

$$v_2(t) = w(t+d).$$

Now using superposition, Lemma A.1 and the Schwarz inequality, the result follows. $\square$

Appendix B. State space and ARMA representations. Consider a time-invariant stochastic state space system for which there exists the following representation:

$$x(t+1) = Ax(t)+Bu(t)+Kw(t), \tag{B.1}$$

$$y(t) = Cx(t)+w(t) \tag{B.2}$$

where $x(t)$ is an $n\times 1$ state vector sequence and $\{y(t)\}$ is the s component output sequence and $\{u(t)\}$ is the r component input sequence respectively. The s component sequence $\{w(t)\}$ is a stochastic process defined on an underlying probability space $(\Omega, \mathcal{A}, P)$.

Let the matrix A have the following characteristic polynomial

$$p(\lambda) = \lambda^n + a_1\lambda^{n-1}+\cdots+a_n. \tag{B.3}$$

It is evident from (B.1) that

$$x(t+k)=A^k x(t)+\sum_{i=1}^{K} A^{i-1}\{Bu(t+k-i)+Kw(t+k-i)\}. \tag{B.4}$$

Then using the Cayly–Hamilton theorem gives

$$\begin{aligned} x(t+n)&=(-a_1A^{n-1}-a_2A^{n-2}\cdot\cdots-a_nI)x(t)\\ &\quad+\sum_{i=1}^{n} A^{i-1}\{Bu(t+n-i)+Kw(t+n-i)\}. \end{aligned} \tag{B.5}$$

Using (B.4)

$$\begin{aligned} x(t+n)&=[-a_1x(t+n-1)\cdot\cdots-a_nx(t)]+\sum_{i=1}^{n} A^{i-1}\{Bu(t+n-i)+Kw(t+n-i)\}\\ &\quad+\sum_{j=1}^{n-1} a_{n-j}\sum_{i=1}^{j} A^{i-1}\{Bu(t+j-i)+Kw(t+j-i)\}. \end{aligned} \tag{B.6}$$

Hence (A.2)

$$\begin{aligned} y(t+n)&=\sum_{j=1}^{n} a_j y(t+n-j)+\sum_{i=1}^{n} CA^{i-1}Bu(t+n-i)\\ &\quad+\sum_{j=1}^{n-1} a_{n-j}\sum_{i=1}^{j} CA^{i-1}Bu(t+j-i)+w(t+n)+\sum_{j=1}^{n} a_j w(t+n-j)\\ &\quad+\sum_{i=1}^{n} CA^{i-1}Kw(t+n-i)+\sum_{j=1}^{n-1} a_{n-j}\sum_{i=1}^{j} CA^{i-1}Kw(t+j-i). \end{aligned} \tag{B.7}$$

Equation (B.7) is of the form:

$$\begin{aligned} A(q^{-1})y(t)&=\begin{bmatrix} q^{-d_{11}}B_{11}(q^{-1}) & \cdots & q^{-d_{1r}}B_{1r}(q^{-1})\\ \vdots & & \vdots\\ q^{-d_{s1}}B_{s1}(q^{-1}) & \cdots & q^{-d_{sr}}B_{sr}(q^{-1}) \end{bmatrix}u(t)\\ &\quad+\begin{bmatrix} C_{11}(q^{-1}) & \cdots & C_{1s}(q^{-1})\\ \vdots & & \vdots\\ C_{s1}(q^{-1}) & \cdots & C_{ss}(q^{-1}) \end{bmatrix}w(t) \end{aligned} \tag{B.8}$$

where $A(q^{-1})$, $B_{ij}(q^{-1})$, $C_{ik}(q^{-1})$ $(i=1,\cdots,s;\ j=1,\cdots,r;\ k=1,\cdots,s)$ denote sclar polynomials in the unit delay operator q^{-1}. If we let

$$d=\min_{\substack{1\leqq i\leqq s\\ 1\leqq j\leqq r}} d_{ij},$$

(B.8) is then of the form of equation (2.1).

We remark that the model (B.1), (B.2) might be viewed as the result of constructing a state estimation filter for a state space system with respect to past data. If such a filter achieves stationary ergodic behavior, it takes the form of (B.1), (B.2) with the conditions (2.2)–(2.4) on the innovation process $\{w(t)\}$ automatically satisfied.

Appendix C. D-step ahead prediction form. We use the notation $C(q^{-1})w(t)$ to denote the linear operation $C(q^{-1})=1+C_1q^{-1}+\cdots C_lq^{-l}$ on the sequence $\{w(t)\}$ where

$$c_jq^{-j}\{w(t)\}=\{c_jw(t-j)\}.$$

A. Single-input single-output systems. Consider the following system description:

$$A(q^{-1})y(t) = q^{-d}B(q^{-1})u(t) + C(q^{-1})w(t). \tag{C.1}$$

From the division algorithm [10, p. 200], $C(q^{-1})$ may be written as $F(q^{-1})A(q^{-1}) + q^{-d}G(q^{-1})$ where $F(q^{-1}) = f_0 + f_1q^{-1} + \cdots + f_{d-1}q^{-d+1}$ and $G(q^{-1}) = g_0 + g_1q^{-1} + \cdots + g_{n-1}q^{-n+1}$.

Hence operating on (C.1) by $F(q^{-1})$ yields

$$F(q^{-1})A(q^{-1})y(t) = q^{-d}F(q^{-1})B(q^{-1})u(t) + F(q^{-1})C(q^{-1})w(t)$$

or

$$(C(q^{-1}) - q^{-d}G(q^{-1}))y(t) = q^{-d}F(q^{-1})B(q^{-1})u(t) + F(q^{-1})C(q^{-1})w(t).$$

Rearranging this gives

$$C(q^{-1})(y(t) - F(q^{-1})w(t)) = q^{-d}G(q^{-1})y(t) + q^{-d}F(q^{-1})B(q^{-1})u(t)$$

or

$$C(q^{-1})(y(t+d) - F(q^{-1})w(t+d)) = G(q^{-1})y(t) + F(q^{-1})B(q^{-1})u(t). \tag{C.2}$$

B. Multiple-input multiple-output systems. Consider the system of equations (B.8). Let

$$d = \min_{\substack{1 \le i \le s \\ 1 \le j \le r}} d_{ij}$$

and define

$$q^{-d}[B(q^{-1})] = \begin{bmatrix} q^{-d_{11}}B_{11}(q^{-1}) & \cdots & q^{-d_{1r}}B_{1r}(q^{-1}) \\ \vdots & & \\ q^{-d_{s1}}B_{s1}(q^{-1}) & \cdots & q^{-d_{sr}}B_{sr}(q^{-1}) \end{bmatrix}. \tag{C.3}$$

Then (B.8) can be written as

$$A(q^{-1})y(t) = q^{-d}[B(q^{-1})]u(t) + [C(q^{-1})]w(t) \tag{C.4}$$

where the notation $[M(q^{-1})]$ denotes the matrix whose ij-th entry is the scalar polynomial $M_{ij}(q^{-1})$.

From the division algorithm [10, p. 200], $c_{ij}(q^{-1})$ is operationally equivalent to $A(q^{-1})F_{ij}(q^{-1}) + q^{-d}G_{ij}(q^{-1})$; $i, j = 1, \cdots, s$. Let $[\mathscr{C}(q^{-1})] = Adj[C(q^{-1})]$. Then operating on (C.4) by $[F(q^{-1})][\mathscr{C}(q^{-1})]$ yields

$$\begin{aligned} &A(q^{-1})[F(q^{-1})][\mathscr{C}(q^{-1})]y(t) \\ &\quad = q^{-d}[F(q^{-1})][\mathscr{C}(q^{-1})][B(q^{-1})]u(t) + [F(q^{-1})][\mathscr{C}(q^{-1})][C(q^{-1})]w(t) \end{aligned}$$

or

$$\begin{aligned} &[[C(q^{-1})] - q^{-d}[G(q^{-1})]][\mathscr{C}(q^{-1})]y(t) \\ &\quad = q^{-d}[F(q^{-1})][\mathscr{C}(q^{-1})][B(q^{-1})]u(t) + [F(q^{-1})][\mathscr{C}(q^{-1})][C(q^{-1})]w(t) \end{aligned}$$

or

$$\begin{aligned} &\bar{c}(q^{-1})y(t) - q^{-d}[G(q^{-1})][\mathscr{C}(q^{-1})]y(t) \\ &\quad = q^{-d}[F(q^{-1})][\mathscr{C}(q^{-1})][B(q^{-1})]u(t) + \bar{c}(q^{-1})[F(q^{-1})]w(t) \end{aligned} \tag{C.5}$$

where

$$[C(q^{-1})][\mathscr{C}(q^{-1})] = [\mathscr{C}(q^{-1})][C(q^{-1})] = \bar{c}(q^{-1})I$$

with

$$\bar{c}(q^{-1}) = \det[C(q^{-1})].$$

Rearranging (C.5) gives

$$\bar{c}(q^{-1})(y(t) - [F(q^{-1})]w(t)) = q^{-d}[G(q^{-1})][\mathscr{C}(q^{-1})]y(t) + q^{-d}[F(q^{-1})][\mathscr{C}(q^{-1})][B(q^{-1})]u(t). \tag{C.6}$$

This is of the form:

$$\bar{c}(q^{-1})(y(t+d) - v(t+d)) = [\alpha(q^{-1})]y(t) + [\beta(q^{-1})]u(t) \tag{C.7}$$

where $\bar{c}(q^{-1})$ is a scalar polynomial and $[\alpha(q^{-1})]$ and $[\beta(q^{-1})]$ are matrices of scalar polynomials.

REFERENCES

[1] A. FEUER AND S. MORSE, *Adaptive control of single-input single-output linear systems*, IEEE Trans. Auto. Control, AC-23 (1978), pp. 557–570.

[2] A. S. MORSE, *Global Stability of Parameter Adaptive Control Systems*, S & IS Report 7902 R, Yale University, March 1979.

[3] G. C. GOODWIN, P. J. RAMADGE AND P. E. CAINES, *Discrete time multi-variable adaptive control*, IEEE Trans. Auto. Control., AC-25 (1980), pp. 449–456.

[4] L. LJUNG, *Analysis of recursive stochastic algorithms*, IEEE Trans. Auto. Control, AC-22 (1977), pp. 551–575.

[5] ———, *On positive real transfer functions and the convergence of some recursive schemes*, IEEE Trans. Auto. Control, AC-22 (1977), pp. 539–551.

[6] V. SOLO, *Time series recursions and stochastic approximation*, Ph.D. dissertation, The Australian National University, Sept. 1978.

[7] L. LJUNG AND B. WITTENMARK, *On a stabilizing property of adaptive regulators*, Reprint IFAC Symposium on Identification (Tbilisi, USSR), 1976.

[8] K. J. ÅSTRÖM AND B. WITTENMARK, *On self-tuning regulators*, Automatica, 9 (1973), pp. 195–199.

[9] K. L. CHUNG, *A Course in Probability Theory*, Harcourt Brace and World, New York, 1978.

[10] D. BURTON, *Introduction to Modern Abstract Algebra*, Addison-Wesley, Reading, MA, 1967.

[11] J. NEVEU, *Discrete Parameter Martingales*, North Holland, Amsterdam, 1975.

[12] J. L. WILLEMS, *Stability Theory of Dynamical Systems*, John Wiley, New York, 1970.

[13] K. L. HITZ AND B. D. O. ANDERSON, *Discrete positive real functions and their applications in system stability*, Proc. IEEE, 116 (1969), pp. 153–155.

[14] G. C. GOODWIN AND K. S. SIN, *Stochastic adaptive control: The general delay-colored noise case*, Technical Report 7904, Dept. of Electrical Engineering, University of Newcastle, Australia, March, 1979.